

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ НАУЧНОЕ УЧРЕЖДЕНИЕ
«ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР ИНФОРМАЦИОННЫХ И
ВЫЧИСЛИТЕЛЬНЫХ ТЕХНОЛОГИЙ»

На правах рукописи

КОЛПАКОВ ФЕДОР АНАТОЛЬЕВИЧ

Компьютерное моделирование биологических систем
и анализ биомедицинских данных

1.5.8 – Математическая биология, биоинформатика

Диссертация на соискание ученой степени
доктора биологических наук

Научный консультант:
член-корреспондент РАН,
профессор, д.б.н. Поройков В.В.

Новосибирск 2023

ОГЛАВЛЕНИЕ

Введение	8
Актуальность проблемы	8
Цель и задачи работы	9
Положения, выносимые на защиту	10
Научная новизна	10
Теоретическая значимость диссертационной работы	13
Практическая значимость диссертационной работы	13
Апробация работы	15
Публикации	16
Личный вклад автора	17
Структура и объем диссертации	17
Технические замечания	17
Глава 1. Обзор литературы	18
1.1 Системная биология	18
1.1.1 Понятие системной биологии	18
1.1.2 Экспериментальные методы системной биологии	20
1.1.3 Воспроизводимость и стандарты в системной биологии	20
1.1.3.1 Воспроизводимость моделей биологических систем	20
1.1.3.2 COMBINE	24
1.1.3.3 SBML	25
1.1.3.4 SBGN	27
1.1.3.5 MIASE, SED-ML	27
1.2 Методы моделирования биологических систем	28
1.2.1 Математические методы	28
1.2.2 Визуальное моделирование	30
1.2.3 Модульное моделирование	33
1.2.4 Агентное моделирование	35
1.2.5 Цифровые двойники и виртуальные организмы	36
1.3 Программные комплексы для моделирования биологических систем	38
1.3.1 Модульное моделирование	39
1.3.2 Агентное моделирование	40
1.3.3 BioSimulators	42
1.4 Анализ биомедицинских данных	44
1.4.1 Установка и запуск программ	44
1.4.2 Реестры программ	45

1.4.2.1 Dockstore	45
1.4.2.2 Biocontainers	46
1.4.2.3 ELIXIR, bio.tools	47
1.4.2.4 Bioconda	47
1.4.3 Сценарии анализа данных	48
1.4.3.1 CWL	48
1.4.3.2 WDL	51
1.4.3.3 Nextflow	52
1.4.2.4 Galaxy	54
1.4.4 Jupyter	55
1.4.5 SoS	57
1.4.6 R, Bioconductor, RStudio	57
1.5 Графическое представление биологических данных	58
1.5.1 Диаграммы метаболических путей и пути передачи сигнала в клетке	59
1.5.2 Геномный браузер	60
1.5.3 Выравнивания и филогенетические деревья	63
1.5.4 Графическое представление 2D структур	66
1.5.5 Графическое представление 3D структур	66
1.6 Обзор предшествующих работ автора	67
1.7. Заключение по обзору литературы	70
Глава 2. ПК BioUML	71
2.1 Интерфейс пользователя	71
2.2 Архитектура ПК BioUML	73
2.3 Используемые технологии	74
2.3.1 OSGi – модульная архитектура	74
2.3.2 Технология JavaBeans	75
2.3.3 Технология BeanExplorer	75
2.3.4 Динамические свойства	76
2.3.5 Доступ и поиск информации в БД	77
2.3.6 Библиотека графических объектов	77
2.3.7 Git, GitLab	77
2.3.8 Docker	77
2.3.9 Jupyter notebook, hub	78
Глава 3. Технология создания и использования модульных моделей биологических систем	80
3.1 Мета-модель	80
3.2 Объектно-ориентированное представление биологических данных	81

3.3 Визуальное моделирование	82
3.3.1 Графическая нотация	82
3.3.2 Исходная графическая нотация ПК BioUML	85
3.3.3 Расширенная нотация SBGN	85
3.3.4 Моделирование физиологических систем	85
3.3.5 Артериальное дерево	88
3.4 Модульное моделирование	89
3.5 Агентное моделирование	92
3.6 Текстовое представление модели, его синхронизация с графическим представлением	92
3.7 Автоматическая генерация кода	93
3.7.1 Разбор математических выражений	94
3.8 Численное моделирование	95
3.9 Тестирование численного моделирования	97
3.10 Определение параметров моделей	98
3.11 Использование практик из разработки программного обеспечения	100
3.12 Тестирование моделей	101
3.13 Создание математических моделей сложных биологических систем эволюционным путем	102
Глава 4. Анализ и графическое представление биомедицинских данных	103
4.1 Интерфейс для анализа биомедицинских данных	105
4.2 Сценарии анализа данных	106
4.2.1 WDL	106
4.2.2 Nextflow	109
4.2.3 Интеграция с Galaxy и Galaxy Toolshed	110
4.3 Геномный браузер	113
Глава 5. База данных GTRD	115
5.1 Постановка задачи	115
5.2 Методы и подходы	116
5.2.1 Отбор видов эукариот	117
5.2.2 Аннотация данных	117
5.2.3 Единый словарь клеточных линий и типов	121
5.2.4 Единообразная обработка NGS данных при помощи сценариев	122
5.2.5 Контроль качества данных	124
5.2.6 Мета-кластеры – интеграция данных ChIP-seq экспериментов	126
5.2.7 Мета-анализ для наиболее достоверного выявления регуляторных районов	128
5.2.8 Мастер-сайт и мастер-трек	129
5.2.9 Структура БД	129

5.2.10	Веб-интерфейс	130
5.3	Использование БД GTRD	131
5.3.1	Построение цистрома	131
5.3.2	Ресурсы, построенные на основе данных GTRD	134
5.3.3	Предсказание уровня экспрессии генов	134
5.4	Обсуждение	135
5.4.1	Сравнение с другими БД по ChIP-seq экспериментам	135
5.4.2	Полнота покрытия ТФ	135
5.4.3	Полнота покрытия ССТФ	136
Глава 6.	Моделирование сложных биологических систем	140
6.1	Модульная модель апоптоза	140
6.1.1	Эволюционное создание моделей	141
6.1.2	Планирование	142
6.1.3	Приемочные тесты	143
6.1.4	Кодирование	144
6.1.5	Итерация 1	144
6.1.6	Итерация 2	146
6.1.7	Итерация 3	146
6.1.8	Итоговая модель	149
6.2	Модульная модель регуляции генной экспрессии при физической нагрузке в скелетных мышцах	152
6.2.1	Модель энергетического метаболизма в клетках скелетной мышцы	152
6.2.2	Регуляция экспрессии генов через Ca ²⁺ -зависимый сигнальный путь	157
6.2.3	Итоговая модульная модель	162
6.3	Модели COVID-19	166
6.3.1	Эпидемиологические модели	166
6.3.1.1.	Первая версия модели	166
6.3.1.2.	Вторая версия модели	167
6.3.1.3.	Третья версия модели	169
6.3.2	Агентная модель	173
6.3.3	Мультимасштабная эпидемиологическая модель	175
6.3.4	Взаимодействие вируса и организма хозяина	177
6.4	Модели регуляции артериального давления у человека	177
6.4.1	Модель Гайтона	179
6.4.2	Резервуарная модель	179
6.4.3	Модель долговременной почечной регуляции	186
6.4.4	Гидродинамическая модель артериальной системы	187

6.4.5 Комплексная модель регуляции артериального давления	188
Глава 7. Цифровой двойник пациента	190
7.1 Концепция цифрового двойника пациента	190
7.2 Генерация виртуального пациента	192
7.3 Генерация виртуальной популяции	194
7.4 Моделирование действия антигипертензивных препаратов	196
7.4.1 Алискирен	198
7.4.2 Лозартан	203
7.4.3 Бисопролол	205
7.4.4 Эналаприл	206
7.4.5 Амлодипин	207
7.4.6 Гидрохлоротиазид	208
7.5 Валидация модели на основе клинических данных	209
7.5.1 Генерация популяции виртуальных пациентов	210
7.5.2 Моделирование АГП терапии	211
7.5.3 Сравнение с результатами клинических исследований	211
7.6 Персонализация параметров модели виртуального пациента	212
7.7 Моделирование лечения пациента	216
7.8 Валидация прогноза результата лечения артериальной гипертензии	218
7.9 Обсуждение	220
Глава 8. Другие разработки на основе ПК BioUML	223
8.1 Биоинформатическая платформа для одномолекулярного секвенирования ДНК	223
8.2 Подсистемы для прототипа «Национальной базы генетической информации»	226
8.3 Платформа geneXplain	226
8.4 Платформа Genome Enhancer	227
8.5 Платформа Sirius-web	229
8.5.1 Постановка задачи	229
8.5.2 Накапливаемые данные	230
8.5.3 Архитектура	231
8.5.4 Использование	233
8.6 Платформа u-science	233
Заключение	236
Выводы	238
Список сокращений и условных обозначений	239
Список литературы	246
Свидетельства о регистрации программ для ЭВМ	281

Список основных публикаций по теме диссертации	282
Благодарности	290
Приложение 1.1 Высокопроизводительные технологии для сбора численных данных	292
Приложение 1.2 Краткое описание формата SED-ML	295
Приложение 2.1 Основные этапы развития ПК BioUML	297
Приложение 2.2 Основные модули (plug-ins) ПК BioUML	300
Приложение 2.3 Java библиотека для доступа и поиска информации в БД	304
Приложение 2.4 Java библиотека графических объектов	307
Приложение 3.1 Основные Java интерфейсы для задания графической нотации	311
Приложение 3.2 Исходная графическая нотация ПК BioUML	315
Приложение 3.3 Расширение и синхронизация SBGN Process Diagram и Antimony	321
Приложение 5.1 Перспектива GTRD для ПК BioUML	329
Приложение 5.2 Веб-интерфейс БД GTRD на платформе BeanExplorer EE	334
Приложение 5.3 Сравнение GTRD с другими БД по ChIP-SEQ экспериментам	340
Приложение 6.1 Экспериментально подтвержденные РСТФ в промоторных областях генов NR4A2, NR4A3 и PPARGC1A	342
Приложение 6.2 Результаты численных расчетов второй версии модели для эпидемиологии COVID-19	344
Приложение 6.3 Результаты численных расчетов третьей версии модели для эпидемиологии COVID-19	349
Приложение 6.4 Результаты численных расчетов агентной модели для эпидемиологии COVID-19	350
Приложение 7.1 Список оптимизируемых параметров для генерации виртуальных пациентов с АГ	351
Приложение 8.1 Пример сценария для биоинформатической платформы для анализа данных одномолекулярного секвенирования	354
Приложение 8.2 Подсистемы для прототипа «Национальной базы генетической информации»	357
Приложение 8.3 Платформа geneXplain	362
Приложение 8.4 Платформа Genome Enhancer	366
Приложение 8.5 Примеры использования платформы Sirius-web	370
Приложение 8.6 Апробация платформы u-science в проекте "Супер С-тау фабрика"	379
Приложение 8.7 Акты внедрения результатов диссертационной работы: ПК BioUML и разработанных на его основе компьютерных систем	384

ВВЕДЕНИЕ

Актуальность проблемы

С началом XXI века активное развитие получила системная биология – междисциплинарная наука, исследующая сложные взаимодействия в живых системах (Kitano H., 2002a, b, c). В системной биологии предполагается определенный цикл проведения исследований, состоящий из теории, аналитического или компьютерного моделирования для формулировки гипотез о системе, экспериментальной проверки гипотез, и затем использование полученных данных для описания клеточных процессов, позволяющее улучшить компьютерную модель или уточнить теорию (Covert et al., 2004; Kholodenko et al., 2005). Поскольку конечной целью является построение портретной модели взаимодействий в сложной системе, экспериментальные методики, которые используются в системной биологии, должны быть максимально детализированными. Для этого применяются такие экспериментальные подходы как транскриптомика, метаболомика, протеомика и другие высокопроизводительные омиксные технологии, используемые для получения численных данных. Таким образом, успех в развитии высокопроизводительных автоматизированных биомедицинских технологий привел к ренессансу моделирования биологических систем на молекулярно-клеточном уровне.

Используя данный подход, было построено множество моделей, как на уровне отдельных клеток, так и на уровне отдельных органов, физиологических систем и целых организмов. Однако, при построении больших моделей биологических систем выявились ограничения и области, которые не охвачены существующими подходами в системной биологии:

- создание сложных моделей – до сих пор остается искусством, доступным немногим исследователям;
- итерационная разработка моделей – хотя такой подход подразумевается в цикле исследований, предлагаемом системной биологией, на самом деле он используется редко. По-видимому, основной причиной такой ситуации является недостаточное развитие соответствующей технологии и инструментария;
- объединение и повторное использование ранее созданных моделей – как правило, использование ранее созданных моделей в качестве составных блоков при построении композиционных моделей вызывает существенные трудности, поскольку исходные модели не были предназначены для этого;
- тестирование и отладка моделей – данные процессы не формализованы и не автоматизированы;
- персонализация параметров моделей – настройка параметров модели под конкретный биологический объект (например, настройка параметров модели для определенного пациента на основе имеющихся индивидуальных клинических данных);

– использование численных данных, полученных с помощью высокопроизводительных технологий – системно-биологический подход предполагает применение этих данных для верификации и возможного последующего уточнения модели. Однако, только в небольшом количестве работ такие данные и результаты их анализа были использованы для построения моделей, например для создания полногеномной модели контроля генной экспрессии в фибробластах мыши (Schwanhausser B. et al., 2011), автоматической генерации полной модели клетки *Mycoplasma genitalium* на основе специально созданной базы данных (БД) (Karr et al., 2012, 2014; Maritan et al., 2022), для крупномасштабной механистической модели *E. coli* (Macklin et al., 2020). Инструмент для моделирования должен быть интегрирован с инструментом для анализа омиксных данных.

Таким образом, для эффективного создания и использования сложных модульных моделей биологических систем требуется развитие особого подхода и создание соответствующего инструментария (специализированные программные комплексы и БД).

Цель и задачи работы

Целью данной работы является разработка технологии и инструментария (программный комплекс и БД) для эффективного построения и использования сложных модульных моделей биологических систем в интеграции с анализом геномных/омиксных данных.

Для достижения этой цели были поставлены и решены следующие задачи:

1. Разработать технологию для эффективного создания и повторного использования сложных модульных моделей биологических систем.
2. Разработать программный комплекс (ПК), реализующий весь инструментарий, необходимый для успешного использования предложенной технологии.
3. Включить в состав разрабатываемого комплекса программы и методы для анализа и графического представления генетических данных.
4. Разработать БД для интеграции широкого круга экспериментальных NGS (Next Generation Sequencing) данных по регуляции транскрипции для построения моделей регуляции транскрипции.
5. Продемонстрировать на практически важных задачах возможности применения разработанной технологии и программного комплекса для создания сложных моделей биологических систем.
6. Разработать технологию построения цифрового двойника пациента и продемонстрировать ее применимость на примере оптимизации выбора лекарственной терапии для лечения артериальной гипертензии (АГ).

Положения, выносимые на защиту

1. Разработана эффективная технология моделирования для итерационного создания модульных моделей сложных биологических систем, имеющих иерархическую структуру.
2. Разработан программный комплекс BioUML – Biological Universal Modeling Language (ПК BioUML, <https://biouml.org>), реализующий инструментарий, необходимый для успешного использования предложенной технологии. В состав разработанного ПК также включены алгоритмы, программы, сценарии для анализа и графического представления биомедицинских данных.
3. Для построения моделей регуляции транскрипции разработана БД GTRD (Gene Transcription Regulation Database, <https://gtrd.biouml.org>), которая обеспечивает единообразную аннотацию, контроль качества, обработку и интеграцию большого объема экспериментальных NGS данных по регуляции транскрипции. На основе этих данных, используя методы машинного обучения разработаны модели для предсказания уровня транскрипции генов для наиболее экспериментально исследованных клеточных линий.
4. Используя разработанную технологию и ПК BioUML построены новые модульные модели сложных биологических систем:
 - наиболее полная модульная модель апоптоза;
 - иерархическая модель регуляция генной экспрессии в скелетных мышцах при физической нагрузке;
 - модели эпидемиологического процесса COVID-19;
 - модульная модель регуляции артериального давления у человека;
 - модель терапии АГ основными классами лекарственных препаратов.
5. Разработана технология построения цифрового двойника пациента. Показана ее применимость на примере оптимизации выбора лекарственной терапии для лечения АГ.

Научная новизна

1. Разработана оригинальная технология моделирования для итерационного создания, тестирования и использования сложных модульных моделей биологических систем. Новизной этой технологии являются:
 - универсальная мета-модель для представления широкого круга моделей биологических систем;
 - расширение языка Antimony для текстового представления биологических моделей в формате SBML (Systems Biology Markup Language) и их графического представления в расширенной графической нотации SBGN (Systems Biology Graphical Notation).
2. Разработан специализированный ПК BioUML реализующий весь инструментарий,

необходимый для успешного использования предложенной технологии для итерационного создания модульных моделей биологических систем. Уникальными особенностями ПК BioUML являются:

- интеграция методов анализа данных высокопроизводительных технологий с возможностями построения моделей сложных биологических систем (визуальное моделирование, автоматическая генерация моделей);
- поддержка широкого круга подходов к моделированию биологических систем: обычные дифференциальные уравнения с задержкой переменных по времени и мгновенными событиями, стохастическое моделирование, генерация моделей на основе правил (rule based modeling), одномерные модели кровотока на основе уравнений в частных производных, анализ метаболических потоков (flux balance analysis) и метод линейного программирования, мультиагентное моделирование (agent based modelling). Отдельно стоит подчеркнуть, что в рамках разработанной нами технологии все эти методы можно одновременно использовать при построении модульных моделей, в этом случае для их численного моделирования их динамики используется мультиагентное моделирование.

Стоит особо отметить, что по результатам независимых сравнений (Maggioli et al., 2019), ПК BioUML признан единственным в мире симулятором биологических моделей, который проходит все тесты на правильность численного моделирования биологических систем SBML Test Suite Core v3.3.0, а также является самым быстрым симулятором.

3. Для графического представления геномных данных разработан новый геномный браузер, который по функционалу сопоставим с ведущими геномными браузерами (Ensembl genome browser, UCSC genome browser, JBrowse, Integrative Genomics Viewer, Integrated Genome Browser, NCBI Genome Data Viewer), а также имеет ряд уникальных возможностей: интерактивное объединение и пересечение треков с возможностью комбинации их отображения (объединение, только общие области, различия), отображение мастер-треков для БД GTRD.

4. Создана новая БД GTRD, которая в настоящее время является крупнейшей в России БД по генетической информации (общий объем, включая исходные NGS данные, составляет 600+ ТБ) и одной из крупнейших в мире по регуляции транскрипции (по количеству ChIP-seq и ChIP-ехо экспериментов). При создании БД GTRD были предложены следующие новые идеи/подходы для интеграции данных ChIP-seq экспериментов и достоверного выявления соответствующих районов связывания транскрипционных факторов (РСТФ):

- единообразная аннотация и обработка NGS данных при помощи сценариев;
- мета анализ результатов 4-х разных алгоритмов для выявления РСТФ: MACS/MACS2 (Zhang et al., 2008), SISSRs (Narlikar and Jothi, 2012), GEM (Guo et al., 2012) и PICS (Zhang et al., 2011) для наиболее достоверного выявления РСТФ;

- кластеры – объединение по разным экспериментальным условиям (клеточные типы, воздействие и т.п.) выявленных районов связывания для заданного ТФ и алгоритма (см. выше);
- мета-кластеры – объединение таких кластеров по используемым алгоритмам, что позволяет получить уникальный набор районов связывания для заданного ТФ по всему геному;
- ранжирование РСТФ (rank aggregation) для выявления наиболее достоверных районов;
- мастер-трек – новая концепция для интеграции всех результатов анализа для заданного ТФ.

На основе полученных мета-кластеров РСТФ для 3-х клеточных линий построена оригинальная модель регуляции транскрипции, которая позволяет предсказать уровень экспрессии соответствующих стартов инициации транскрипции из БД FANTOM5.

5. Используя разработанную технологию и ПК BioUML построены новые модульные модели сложных биологических систем:

- наиболее полная модульная модель апоптоза, которая включает: 13 модулей, 286 белков и их комплексов, 684 реакции и 719 параметров;
- модульная агентная модель регуляции артериального давления у человека (2 агента (сердечно-сосудистая система и почка), 20 модулей, 25 обыкновенных дифференциальных уравнений (ОДУ), 160 алгебраических выражений, 160 переменных, 132 параметра, 10 дискретных событий). Она была калибрована и валидирована на большом количестве экспериментальных данных, соответствующих физиологическим параметрам человека в норме и патологии. Данная модель стала первой модульной моделью в БД Biomodels (ID – MODEL2202160001);
- с использованием агентной модели регуляции артериального давления у человека была разработана модель терапии АГ основными классами лекарственных препаратов, как по отдельности так и их комбинациями. Она была валидирована по наибольшему количеству клинических исследований, по сравнению со сходными моделями;
- модульная модель энергетического метаболизма в клетках скелетной мышцы;
- математическая модель, связывающая Ca^{2+} -зависимый сигнальный путь с регуляцией экспрессии генов в клетках скелетной мышцы человека. При построении этой модели использовались экспериментально подтвержденные сайты связывания ТФ CREB, FOS и JUN семейств в промоторных областях генов NR4A2, NR4A3 и PPARGC1A из БД GTRD;
- представленные выше две модели работы скелетных мышц были объединены в единую модульную модель, что впервые позволило построить математическую модель, описывающую изменение экспрессии отдельных генов в ответ на физическую нагрузку;

- модель эпидемиологического процесса COVID-19. Впервые для описания процесса заражения использовалось дифференциальное уравнение с задержкой, параметры которого определялись путем подгонки кривой инфицирования к реальным данным независимо от других параметров модели. Это позволяет с одной стороны более точно описать динамику инфицирования, а с другой – уменьшить количество параметров, определяемых в ходе решения обратной задачи, что делает эти параметры более идентифицируемыми.

6. Впервые была разработана технология построения цифрового двойника пациента и показана ее применимость на примере оптимизации выбора лекарственной терапии для лечения АГ. При использовании созданной модели регуляции артериального давления невозможно однозначно идентифицировать большое количество параметров этой модели по ограниченному данному из истории болезни пациента, данных клинических и лабораторных наблюдений. Для решения этой проблемы предложен принципиально новый подход, когда для каждого пациента строится популяция виртуальных пациентов (200-500), которые совпадают с заданным пациентом по наблюдаемым клиническим параметрам, но имеют различные значения внутренних скрытых параметров модели. Выбор оптимальной лекарственной терапии производится путем моделирования лечения различными антигипертензивными препаратами всей такой популяции и анализом однородности отклика отдельных пациентов.

Теоретическая значимость диссертационной работы

1. В диссертационной работе разработана новая технология для итерационного создания, тестирования и использования модульных моделей биологических систем.
2. При создании БД GTRD разработана новая технология объединения данных из различных NGS экспериментов в виде кластеров, мета-кластеров и мастер-треков.
3. Разработан ряд новых модульных моделей сложных биологических систем.
4. Разработана технология построения цифрового двойника пациента.

Практическая значимость диссертационной работы

1. ПК BioUML можно рассматривать как полнофункциональную отечественную биоинформатическую платформу, что особенно важно для обеспечения технологического суверенитета Российской Федерации в этой области. В этом качестве ПК BioUML используется в ряде отечественных проектов:

- В ходе реализации работ по контракту № 43-03/21/44/152 от 13.12.2021 "Исследование, обоснование и выбор программных решений для визуализации генетических данных и обеспечения инструментов для работы с генетической информацией в «Национальной базе генетической информации» на основе ПК BioUML созданы 2 подсистемы: "Геномный

браузер" и "Рабочие окружения".

- Для создания биоинформатической платформы в рамках договора № 33.2022 на выполнение научно-исследовательских работ по теме: "Разработка и испытания биоинформатической платформы для хранения, анализа и графического представления данных, полученных при одномолекулярном секвенировании ДНК".
2. Sirius-web (<https://sirius-web.org/>) – это информационная платформа на основе ПК BioUML для проектов, связанных с анализом данных и моделированием для образовательных и научных проектов и активно используемая в Научно-технологическом университете «Сириус» для проведения исследований в области наук о жизни.
 3. На основе ПК BioUML созданы коммерческие программы для визуализации и анализа геномных/омиксных данных:
 - платформа Genome Enhancer (<https://genexplain.com/genexplain-platform>) – это специализированный вариант ПК BioUML для анализа омиксных данных в автоматическом режиме (Kel et al., 2011), изначально разработанный ООО "Биософт.Ру" при поддержке гранта Фонда "Сколково";
 - платформа geneXplain (<https://genexplain.com/genexplain-platform>) – это коммерческий вариант ПК BioUML, содержащий множество готовых сценариев для анализа омиксных данных, с использованием информации из БД TRANSFAC и TRANSPATH. Она разрабатывается компаниями geneXplain GmbH (Вольфенбютель, Германия) и ООО "Биософт.Ру" для анализа омиксных данных (Kel et al., 2011).
 4. ПК BioUML используется как основной инструмент в курсах по системной биологии, проводимых в Новосибирском государственном университете и научно-технологическом университете "Сириус".
 5. Разработанная технология создания модульных моделей биологических систем и реализующий ее ПК BioUML позволили перейти на более высокий уровень сложности создания таких моделей (сотни и тысячи переменных и уравнений), а также упростили и ускорили, с точки зрения создателей моделей, этот процесс.
 6. Созданная БД GTRD является высоко востребованной, широко используемой и цитируемой: три публикации в специализированных выпусках Nucleic Acids Research 2017, 2019 и 2021 года набрали в совокупности 400+ цитирований по версии Semantic Scholar (<https://www.semanticscholar.org/>), включая цитирования в журналах Nature и Science. На основе GTRD коллективом под руководством проф., чл.корр. РАН Макеева В.Ю. и д.б.н. Кулаковского И.В. были созданы ресурсы:
 - HOCOMOCO – HOMO sapiens COMprehensive MOdel COLLECTION – коллекция мотивов для сайтов связывания транскрипционных факторов (ССТФ) человека и мыши

- (<https://hocomoco11.autosome.ru/>);
- ADASTRA – Allelic Dosage-corrected Allele-Specific human Transcription factor binding sites (<https://adastra.autosome.ru/>) – коллекция данных по аллель-специфичному связыванию факторов транскрипции в геноме человека;
 - ANANASTRA – ANnotation and enrichment ANALysis of SNP sets with allele-specific TRAnscription factor binding (<https://ananastra.autosome.ru/>) – коллекция данных по влиянию SNP на аллель-специфичное связывание факторов транскрипции в геноме человека;
 - версия цистрома для человека и мыши (<https://doi.org/10.1186/s13104-018-3856-x>).

Также информация из БД GTRD была использована при создании широко известных международных БД:

- BaMM motif – Bayesian Markov Models (BaMMs) – для распознавания ССТФ (<https://bammotif.soedinglab.org/>);
- mSigDB – Molecular Signatures DataBase, C3: regulatory target gene sets (<https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp#C3>).

Результаты диссертационной работы внедрены в научные исследования и образовательный процесс ряда организаций, включая: Институт аналитического приборостроения РАН, Институт белка РАН, Институт медико-биологических проблем РАН, Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирский государственный университет, Казанский федеральный университет, Научно-технологический университет «Сириус» (акты внедрения приведены в приложении 8.7).

Апробация работы

Результаты работы были представлены и обсуждены на следующих российских и международных конференциях: “International Conference on Systems Biology” – ICSB-2002 (Стокгольм, Швеция), ICSB-2004 (Гайдельберг, Германия), ICSB-2006 (Йокохама, Япония), ICSB-2007 (Лонг Бич, США), ICSB-2008 (Гетеборг, Швеция), ICSB-2010 (Эдинбург, Великобритания), ICSB-2011 (Гейдельберг, Германия); "Биоинформатика регуляции и структуры генома (BGRS)", г. Новосибирск: BGRS-1998, BGRS-2000, BGRS-2002, BGRS-2004, BGRS-2006, BGRS-2008, BGRS/SB-2010, BGRS/SB-2012, BGRS/SB-2014, BGRS/SB-2016, BGRS/SB-2018, BGRS/SB-2020, BGRS/SB-2022); 7th Noordwijkerhout Symposium on Pharmacokinetics, Pharmacodynamics and Systems Pharmacology (Нордвейкерхаут, Нидерланды, 2013); International conference “High-throughput sequencing in genomes – 2013”, г. Новосибирск; German Conference on Bioinformatics, GCB’2003, Нойхерберг, Германия, 2003; Московская международная конференция “Биотехнологии и медицина”, Москва, 2006; «Молекулярная онкология», Всероссийская конференция с международным участием, Новосибирск, 2008;

Международная Московская конференция по компьютерной молекулярной биологии (МССМБ), г. Москва, в 2005 и 2011 гг.; международной конференции по современным проблемам математики, информатики и биоинформатики, посвященной 100-летию со дня рождения член-корреспондента АН СССР Алексея Андреевича Ляпунова (Новосибирск, 2011); международный семинар «From virtual cell to virtual human and virtual patient» (Новосибирск, 2012); международной конференции по биомедицинской инженерии и компьютерных технологиях SIBIRCON (Новосибирск, 2015); XVI Российская конференция «Распределенные информационно-вычислительные ресурсы. Наука – цифровой экономике», DICR-2017 (Новосибирск, 2017); NETTAB – 17th International Workshop on Network Tools and Applications in Biology: Methods, Tools and Platforms for Personalized Medicine in the Big Data Era. – (Палермо, Италия, 2017); Workshop on BIOINFORMATICS OF GENE REGULATION. 30 Years TRANSFAC (Геттинген, Германия, 2018); Международный семинар "LIFE OF GENOMES 2018" (Казань, 2018); XVII Международная конференция «Распределенные информационно-вычислительные ресурсы: Цифровые двойники и большие данные», DICR-2019 (Новосибирск, 2019); Всероссийская мультikonференция с международным участием "Биотехнология – медицине будущего" (Новосибирск, 2019); VII Съезд Вавиловского общества генетиков и селекционеров, посвященный 100-летию кафедры генетики СПбГУ, и ассоциированные симпозиумы (Санкт-Петербург, 2019); Международный конгресс "Биотехнология: состояние и перспективы развития" (Москва, 2019-02-27); X Всероссийская конференция «Новые подходы к изучению проблем физиологии экстремальных состояний» (Москва, 2021); Международный семинар «Сочинский диалог» «Диалоги о геномике: лучшие практики лабораторий РФ и Европы» (Сочи, 2021); IX Международный форум технологического развития «Технопром-2022» (Новосибирск, 2022); VI Сеченовский Международный Биомедицинский Саммит (SIBS-2022, Москва); Школа "Цифровая кардиология" (Москва, 2022); Международная конференция «Прикладная математика иммунологии и вирусологии» (Сочи, 2022); Sensors and their medical applications (Huawei, он-лайн конференция, 2022); VII Всероссийской научно-практической конференция по Искусственному интеллекту в здравоохранении и системам поддержки принятия врачебных решений ITM-AI (он-лайн конференция, 2023).

Публикации

Материалы диссертационной работы отражены в 75 научных публикациях, включая: 34 публикации в журналах Q1 и Q2 Web of Science/Scopus, 12 публикаций в журналах Q3 и Q4, 3 главы в монографиях. Издано 1 учебное пособие. Получено 7 свидетельств на регистрацию программ и баз данных для ЭВМ.

Личный вклад автора

Автором предложена новая технология для итерационного создания, тестирования и использования сложных модульных моделей биологических систем, лично автором сформулированы соответствующие идеи и подходы.

В 2002 г. лично автором была создана первая версия ПК BioUML. В дальнейшем развитии этой платформы под постоянным руководством автора приняло участие более 30 человек (см. раздел "Благодарности"). Также автором постоянно осуществлялись разработка и развитие архитектуры ПК BioUML, начиная от настольного приложения и заканчивая облачным веб-приложением. Автором лично написан начальный код ключевых Java библиотек и Java интерфейсов: для доступа к данным (ru.biosoft.access), для визуализации данных (ru.biosoft.graphics), для объектно-ориентированного представления данных (biouml.standard, ru.biosoft.bsa), мета-модель для представления биологических моделей (biouml.model), для расположения вершин графов на плоскости (ru.biosoft.graph) и ряд других.

Под руководством автора была создана БД GTRD, при этом автором лично сформулированы ключевые идеи и подходы.

Под руководством автора, используя ПК BioUML, были разработаны приведенные в диссертационной работе модульные модели биологических систем. Автор также отбирал существующие математические модели, которые стали отправной точкой для построения соответствующих модульных моделей.

Автором была разработана технология построения цифрового двойника пациента.

На основе архитектуры и кода ПК BioUML автором предложена архитектура универсальной научной платформы u-science – для графического представления и анализа данных из других предметных областей.

Также автором разработана архитектура и ключевые идеи платформы Sirius-web.

Структура и объем диссертации

Диссертационная работа состоит из введения, 9 глав, заключения и списка литературы содержащего 438 ссылок. Работа изложена на 395 страницах (включая 291 страницы основного текста и 24 приложения), содержит 120 рисунков, 20 таблиц.

Технические замечания

Нумерация рисунков и таблиц дана согласно подразделам диссертации.

Номера приложений соответствуют номерам глав диссертации.

Шрифтом Courier New выделены названия интерфейсов, классов и пакетов разработанных библиотек на языке Java, например интерфейс View.

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

1.1 Системная биология

1.1.1 Понятие системной биологии

С завершением расшифровки геномов многих организмов, включая геном человека, исследователи переходят к следующей фазе – изучение и моделирование организмов и их частей как сложных нелинейных систем. Проблема состоит в том, что для того чтобы понять как функционируют биологические системы, недостаточно разложить их на составляющие и изучить каждую из них (этот этап в существенной части уже пройден и громадный массив экспериментальных данных накоплен в различных БД), необходимо понять как эти составляющие взаимодействуют друг с другом, образуя целую систему – этот этап становится наиболее актуальным в последнее время.

«Любой, уважающий себя, редукционист также является реконструкционистом. Это значит, что если вы спросите редукциониста, в чем состоит его цель, то выясните, что она состоит в том, чтобы разобрать сложную систему на ее элементарные составляющие для того чтобы охарактеризовать их, и когда это будет сделано, использовать полученные знания для реконструкции и понимания целой системы, с которой было начато исследование. Однако проблема состоит в том, что фаза реконструкции в этой программе выполняется редко. Отчасти это объясняется тем, что все внимание поглощает задача исследования молекулярных элементов. Предстоит еще очень много сделать в этом направлении, но выполнение этой задачи (реконструкции) обещает еще большие успехи.

Парадоксально, но в зените ее успехов, слабость этой парадигмы становится все более очевидной. Ясно, что как отмечалось выше, скоро мы будем иметь полные каталоги всех составляющих для некоторых организмов. Тем не менее, мы относительно мало знаем об интегрированной системе, образующей живую клетку, или о том как она ответит на новые изменения в окружающей среде или специфические изменения в ее составе. Вкратце, наши знания все еще фрагментарны и носят описательный характер, мы почти не понимаем «принципов дизайна», которые управляют биологическими системами. Причина неудачи при реконструкции более фундаментальна, чем уделение внимания другим задачам. Парадигма редукционизма сама по себе не способна справиться с данной задачей, необходим радикально другой, комплементарный «системный» подход.» (Savageau, 1991).

Осознание данной проблемы привело к активному развитию, начиная с 2000-х годов, нового научного направления – системной биологии, которая является междисциплинарной наукой о жизни, направленной на изучении сложных взаимодействий в живых системах,

использующей новый подход в биологии: холизм вместо редукционизма (Kitano H., 2002a,b,c).

В самом общем понимании задачами системной биологии являются исследование и моделирование свойств сложных биологических систем, которые нельзя объяснить суммой свойств ее составляющих. Пионером системной биологии часто считают Людвиг Фон Берталанфи, создателя общей теорией систем. В 1937 на семинаре по философии в Чикагском университете он высказал идею наличия общих закономерностей при взаимодействиях большого, но не бесконечного числа физических, биологических и социальных объектов. Однако первые его публикации на эту тему появились только после войны. Основной идеей общей теории систем является признание изоморфизма законов, управляющих функционированием системных объектов (von Bertalanffy, 1950, 1951, 1962).

Согласно Википедии, первая работа по системной биологии, как самостоятельной дисциплине, была представлена системным теоретиком Михайло Месарович в 1966 году на международном симпозиуме в Институте технологии в Кливленде (США, штат Огайо) под названием «Системная теория и биология» (Mesarovic, 1968).

Сейчас системная биология предлагает определенный цикл исследований (рисунок 1.1.1), состоящий из теории, аналитического или компьютерного моделирования для формулировки гипотез о системе, экспериментальной проверки, и затем использование полученных данных для описания клетки или клеточных процессов для улучшения компьютерной модели или теории (Covert et al., 2004; Kholodenko et al., 2005).

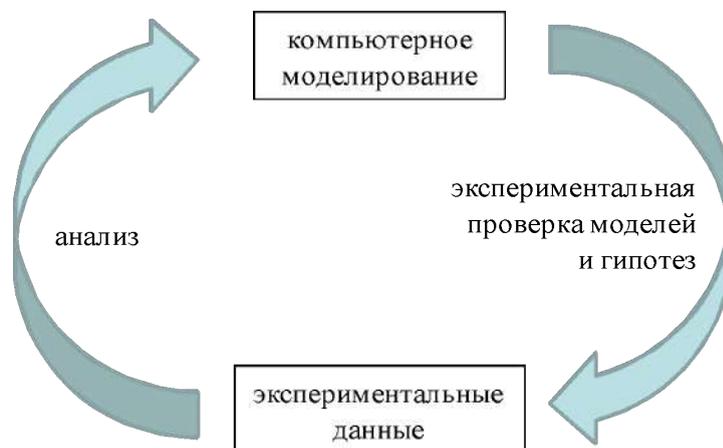


Рисунок 1.1.1 – Основной цикл исследований в системной биологии

Поскольку конечной целью является построение портретной модели взаимодействий в сложной системе, экспериментальные методики, которые используются в системной биологии, должны быть максимально детализированными. Для этого используются высокопроизводительные технологии для сбора численных данных, основные из которых описаны ниже.

1.1.2 Экспериментальные методы системной биологии

Для верификации создаваемых моделей системная биология работает с различными типами экспериментальных данных, описывающих как отдельные составляющие, так и систему в целом. В качестве исходной информации для формулировки гипотез и выводов используются данные, полученные в других областях биологии: биохимии, биофизики, молекулярной биологии.

Тем не менее, существует ряд методов (приложение 1.1), прочно ассоциируемых с системной биологией – это высокопроизводительные автоматизированные потоковые технологии для сбора большого количества численных данных (high throughput technologies, omics технологии). Прочная связь системной биологии и омиксных технологий объясняется тем, что они идеально взаимодополняют друг друга в цикле исследований, представленном на рисунке 1.1.1

До появления омиксных технологий экспериментальные данные в клеточной и молекулярной биологии были в основном представлены наблюдениями и измерениями различных качественных и количественных признаков, например скорость роста культуры клеток или их гибель в ответ на воздействие различных факторов, изменение структуры цитоскелета. Экспериментальные методы, например, вестерн-блот и ПЦР позволяли измерить количества отдельных белков и РНК. Данная ситуация привела к высказыванию Аарона Кацера: “Биологи делятся на два класса: экспериментаторы, которые наблюдают вещи, которые нельзя объяснить, и теоретики, которые объясняют вещи, которые нельзя наблюдать” (Oster, 1983).

С другой стороны, объем численных данных, полученных в ходе омиксных экспериментов, составляет десятки и сотни тысяч измерений. Для их анализа и интерпретации нужен системный подход и соответствующий инструментарий.

1.1.3 Воспроизводимость и стандарты в системной биологии

Математические и вычислительные модели являются ключевой технологией в системной биологии. Прогресс в этой области зависит от повторяемости и воспроизводимости их свойств и поведения. Для этого важным требованием является набор четких стандартов для спецификации модели и ее распространения (Vieira and Laubenbacher, 2022).

1.1.3.1 Воспроизводимость моделей биологических систем

Достоверность модели важна в системной биологии, независимо от того, используются ли модели для проведения экспериментов, разработки лекарств или оптимизации лечения пациентов. Повторяемость (replicability) и воспроизводимость (reproducibility) напрямую

связаны с достоверностью (*credibility*) и могут быть проблемой как в экспериментальных, так и вычислительных исследованиях (Vieira and Laubenbacher, 2022).

Даже для моделей, основанных на системах дифференциальных уравнений, воспроизводимость далеко не гарантирована, как было показано в исследовании, основанном на БД BioModels (Tiwari et al., 2021). Для этого группа ученых, работающих над БД BioModels, проанализировала 455 кинетических моделей, опубликованных в 152 рецензируемых журналах (рисунок 1.1.2). Большинство этих моделей были вручную воссозданы с нуля в формате SBML для оценки воспроизводимости.

Около половины опубликованных моделей системной биологии не могут быть воспроизведены напрямую (рисунок 1.1.2.a). Около 12% (рисунок 1.1.2.b) моделей могут быть воспроизведены с эмпирическими исправлениями или при поддержке автора (среди авторов-переписчиков, с которыми они связались, ответили менее 30%). Остальные 37% остались не воспроизводимыми из-за отсутствующих значений параметров, отсутствующих начальных концентраций, непоследовательной структуры модели или другой отсутствующей информации.

Также авторы исследовали зависимость возможности воспроизвести модель от года публикации (рисунок 1.1.2.c) и размера модели по количеству объектов (рисунок 1.1.2.d). Как видно из рисунка 1.1.2.d, чем больше размер модели, тем ниже шансы ее воспроизвести.

Модели из многих журналов по биологическим наукам не удалось воспроизвести (рисунок 1.1.2.e, приведены журналы, где было опубликовано две и более моделей, использованных в исследовании), что выявило общую проблему в процессе рецензирования соответствующих статей.

В работе Tiwari et al. (2021) сформирована система показателей для оценки воспроизводимости, состоящая из 8 вопросов (список 1.1.1), которые помогла бы другим ученым воспроизвести результаты моделирования модели с разумными усилиями. За каждый ответ "да" начисляется 1 балл. Все 8 вопросов не всегда применимы, но авторы настоятельно рекомендуют, чтобы модель получила как минимум 4 балла.

На рисунке 1.1.3 представлены результаты исследования зависимости предложенной системы показателей и воспроизводимости моделей (Tiwari et al., 2021). Для этого авторы отобрали из 455 моделей случайным образом 110 моделей (45 невозпроизводимых и 65 воспроизводимых, воспроизводимые модели включают как модели, воспроизводимые непосредственно ($n = 46$), так и модели, воспроизводимые с дополнительными усилиями ($n = 19$)) и оценили их с использованием предложенной системы показателей. Тест хи-квадрат показал значительную связь между оценкой ≥ 4 и воспроизводимостью модели ($P = 0,0015$), что

подтверждает рекомендацию авторов, чтобы модель получила как минимум 4 балла (Tiwari et al., 2021).

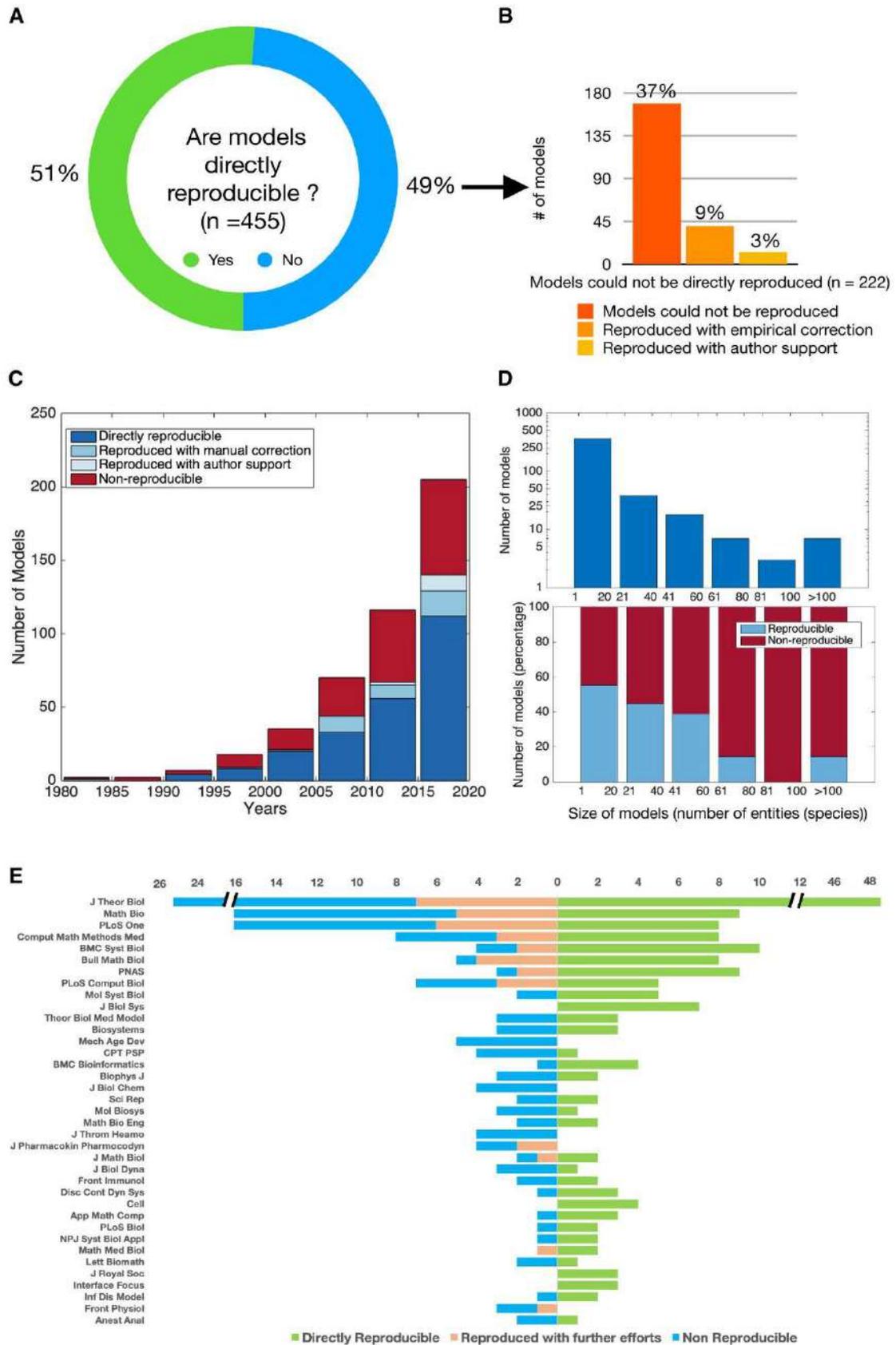


Рисунок 1.1.2 – Результаты исследования воспроизводимости моделей группой ученых из BioModels (Tiwari et al., 2021). См. описание в тексте

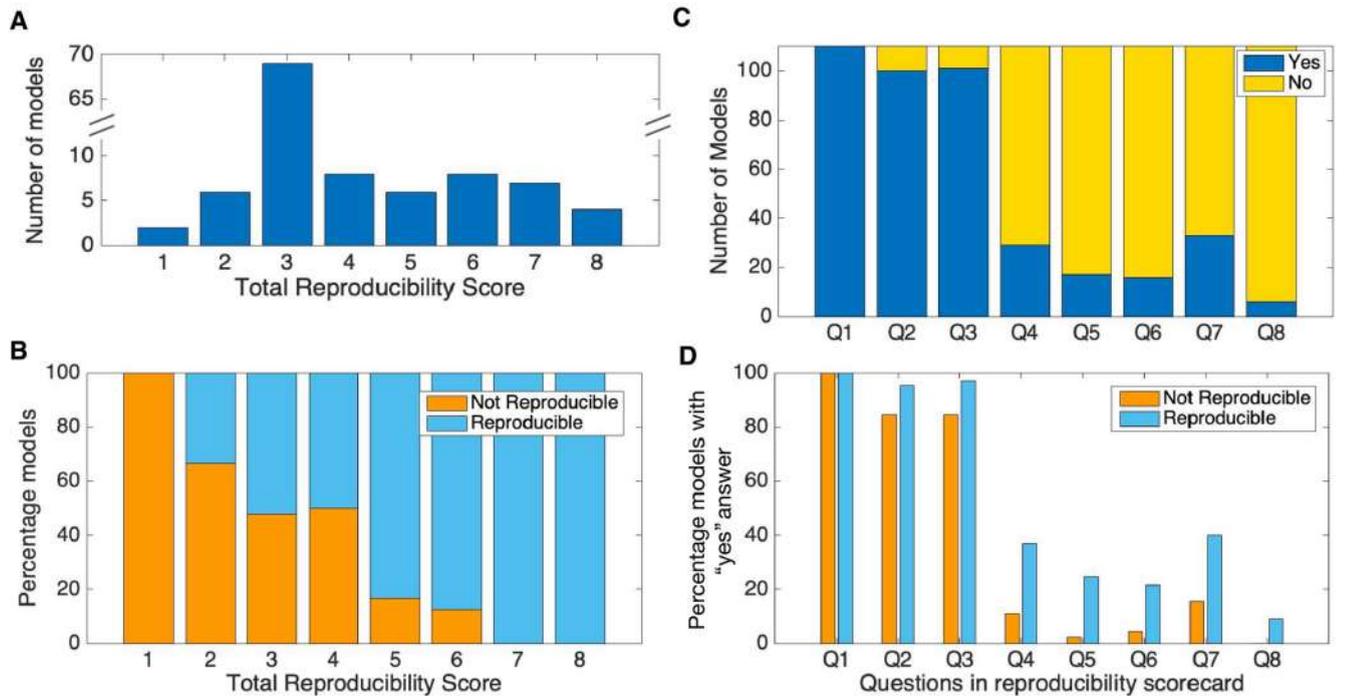


Рисунок 1.1.3 – Система показателей для воспроизводимости биологических моделей (Tiwari et al., 2021). (A) Распределение общей оценки воспроизводимости по моделям*. (B) Процент невоспроизводимых и воспроизводимых моделей в зависимости от набранных баллов. (C) Распределение баллов, набранных моделями по каждому вопросу. (D) Сравнение баллов, набранных невоспроизводимыми и воспроизводимыми моделями по каждому вопросу. Q1–Q8 – это вопросы, представляющие 8 баллов в оценочной таблице воспроизводимости

Это и другие исследования (например, Niarakis et al., 2022; Porubsky et al., 2020) подчеркивают важность использования стандартов и определенных практик при создании, отладке и публикации моделей, чтобы они могли быть воспроизведены и повторно использованы другими учеными.

Список 1.1 – Система показателей (рейтинг) для воспроизводимости биологических моделей (Tiwari et al., 2021)

1. Описаны ли математические выражения в рукописи/дополнительных материалах?
2. Перечислены ли параметры и начальные уровни сущностей (в виде таблицы) в рукописи/дополнительных материалах?
3. Описаны ли условия моделирования, включая программное обеспечение/среду программирования, алгоритм, изменения параметров/концентраций/состояний и любая нормализация данных под каждым рисунком моделирования или прилагаются в качестве дополнительных материалов?
4. Является ли код(ы) модели для математического выражения и симуляции общедоступным?

5. Доступны ли коды моделей в стандартных форматах, таких как SBML, архив COMBINE, SED-ML, и проверены ли они синтаксически?
6. Депонированы ли коды моделей в соответствующей открытой БД моделей?
7. Хорошо ли задокументированы коды моделей для однозначной идентификации объектов/переменных модели (с дополнительной аннотацией реакций, математических выражений, событий, условий и т.д. при необходимости)?

Являются ли модели в стандартных форматах, таких как SBML и COMBINE Archive, семантически обогащенными, то есть аннотированы контролируруемыми словарями, такими как Gene Ontology и ChEBI, и ресурсами БД, такими как Gene Ontologies?

8. Публично ли публикуются численные результаты вместе с кодами моделей?

1.1.3.2 COMBINE

COMBINE (<https://co.mbine.org/>) – COmputational Modeling in BIology' NEtwork (Waltemath et al., 2020) – это организация, координирующая разработку стандартов в системной и синтетической биологии. Она была основана в 2009 году и с тех пор поддерживает и координирует разработку соответствующих стандартов по всему миру, модерирует дискуссии, разрабатывает и реализует стратегии распространения стандартов; а также предлагает центральную точку доступа к спецификациям стандартов и поддерживающих их библиотек.

На рисунке 1.1.4 представлены стандарты COMBINE, в разбивке на 3 группы: базовые стандарты, связанные стандарты и стандарты-кандидаты.

Ниже мы рассмотрим основные стандарты, связанные с системной биологией. Стоит отметить, что автор диссертационной работы принимал непосредственное участие в разработке этих стандартов.

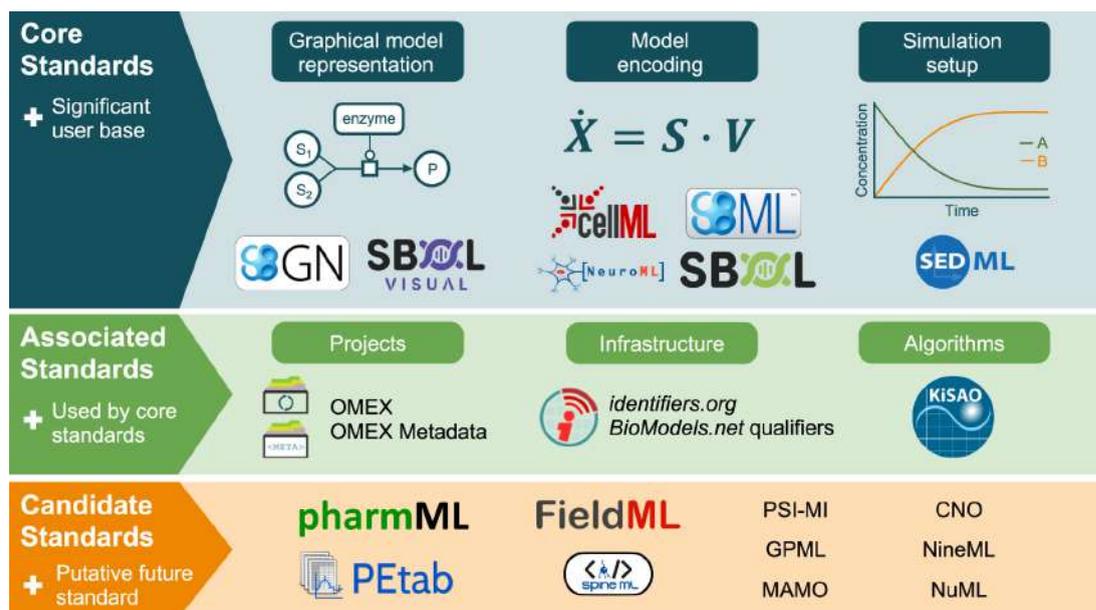


Рисунок 1.1.4 – Обзор стандартов COMBINE (Schreiber et al., 2021)

1.1.3.3 SBML

SBML (www.sbml.org; Hucka et al., 2001, 2003, 2019; Keating et al., 2020) – System Biology Markup Language – стандарт для представления математических моделей биологических систем. Его использование обеспечивает возможности для воспроизведения и переиспользования моделей широкого круга биологических систем. Этот формат поддерживает более 300 различных программ (<https://en.wikipedia.org/wiki/SBML>).

Изначально (Hucka et al., 2001) был разработан для представления моделей, представленных в виде набора биохимических реакций, которые могут протекать в различных компартментах. В дальнейшем для SBML level 3 (Keating et al., 2020) была предложена модульная архитектура (рисунок 1.1.5).

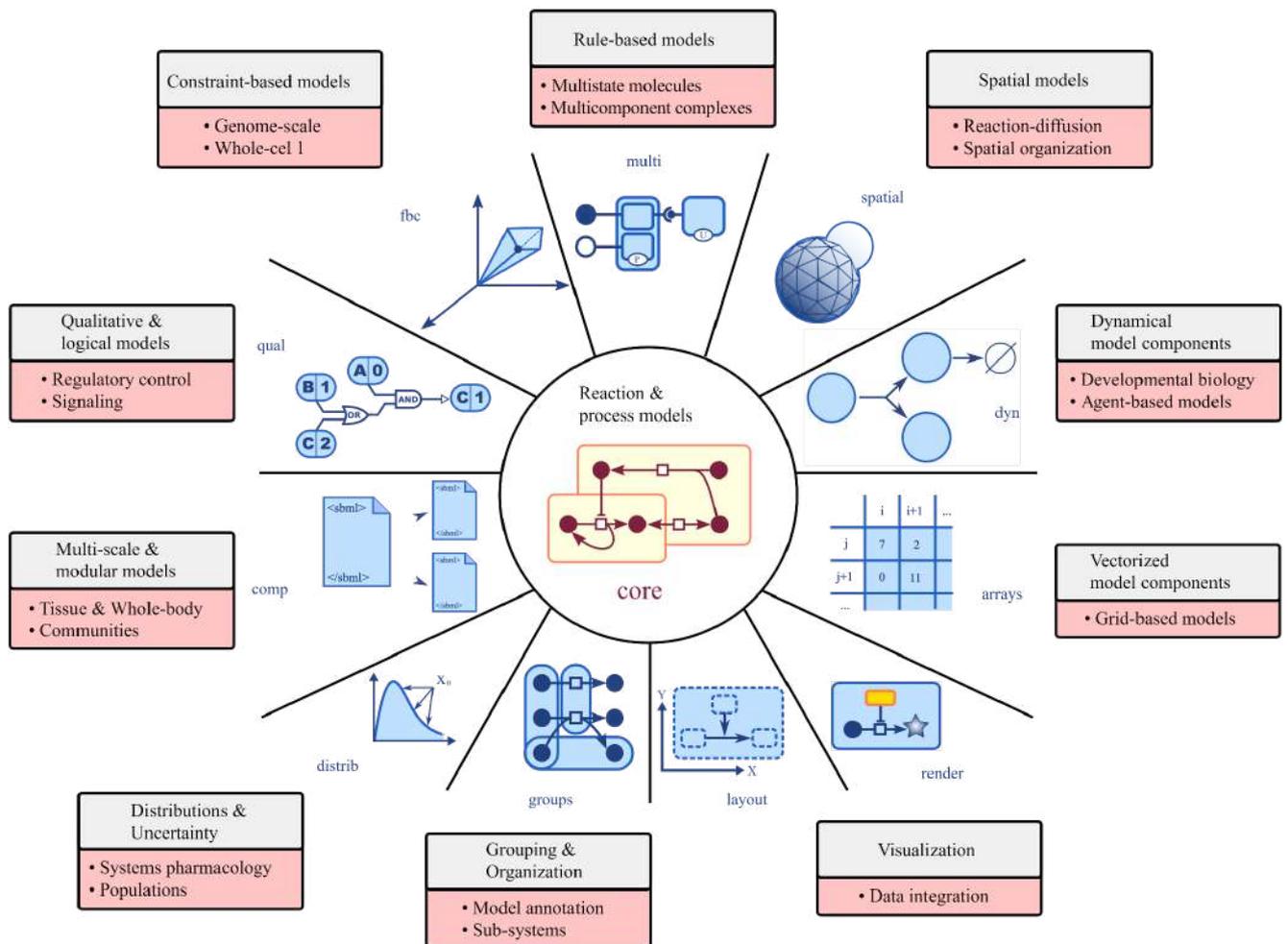


Рисунок 1.1.5 – Модульная архитектура формата SBML level 3 (Keating et al., 2020)

Для этого было выделено:

- ядро формата SBML level 3 core – лучше всего подходит для представления моделей на основе биохимических реакций или набора математических формул (ОДУ, алгебраических выражений, дискретных событий);

- набор расширений (пакетов), которые расширяют ядро с помощью конструкций, подходящих для других типов моделей, включая модели на основе ограничений, модели реакции-распространения, модели логических сетей, модели на основе правил и многое другое.

Модель, выраженная с использованием SBML уровня 3, может объявлять, какие расширения она использует. В зависимости от используемых расширений, пользователю может потребоваться разное программное обеспечение, поддерживающее соответствующие пакеты, что создает определенные проблемы – автору неизвестно хотя бы одна программа, которая поддерживает все разработанные пакеты.

Более того, в силу сложности SBML level 3 core version 2 (Hucka et al., 2019), его спецификация занимает 182 страницы, возникают 2 проблемы:

- полная поддержка спецификации – многие известные программы, не поддерживают отдельные конструкции спецификации, например, уравнения с задержкой по времени;
- правильность численного моделирования – часто результаты численного моделирования, выполненные в разных программах дают различные результаты.

Для решения этих проблем был разработан специальный набор тестов – SBML semantic test suite (Hucka et al., 2022; <https://synonym.caltech.edu/software/sbml-test-suite/>). Версия 3.4.0 содержит 1821 тестовую модель, которые покрывают все конструкции:

- SBML level 3 core version 2;
- расширение 'comp' для представления иерархических модульных (Smith et al., 2015);
- расширение 'fbc' для представления потоковых моделей (flux balance analysis; Olivier and Bergmann, 2018).

Для каждой модели приведены:

- NNNNN-sbml-INvN.xml – файлы модели в SBML форматах разных версий (level 1 version 2 - level 3 version 2);
- NNNNN-model.m – текстовое описание теста;
- NNNNN-settings.txt – файл, с параметрами настройки для численного моделирования (интервал моделирования, шаг по времени, относительная и абсолютная погрешность, список переменных, значения которых будут сравниваться с приведенным в тесте правильным результатом);
- NNNNN-results.csv – текстовый файл с правильными (образцовыми) результатами расчета.

Также разработан набор тестов для стохастического моделирования (Evans et al., 2008). В этом случае для каждой модели необходимо проводить расчет минимум 1000 раз, а в тестах для

сравнения приводятся правильные средние значения и стандартные отклонения в заданных временных точках.

1.1.3.4 SBGN

В 2008 году международным консорциумом была предложена графическая нотация SBGN – Systems Biology Graphical Notation (Le Novère N., et al. 2008; <https://sbgn.github.io/>) для формального описания структуры и функционирования биологических систем при помощи 3 типов диаграмм:

- Activity flows – диаграмма активностей показывает поток информации в ходе работы биологических путей, при этом многие детали опускаются;
- Process descriptions – диаграмма процессов предназначена для описания биологических путей, как правило, на основе графического представления соответствующих биохимических реакций;
- Entity relationships – диаграмма взаимодействий описывает взаимодействия между биологическими процессами без учета временного аспекта.

Графическая нотация SBGN не предназначена для визуального моделирования, и поэтому в ней отсутствуют многие необходимые для этого компоненты – уравнения, события, состояния, функции и т.п. Для устранения этого недостатка было предложено расширение диаграмм типа Process descriptions, описанное в разделе 3.3.3.

1.1.3.5 MIASE, SED-ML

Однако информации только о модели недостаточно для ее эффективного повторного использования. Сложные численные алгоритмы и сценарии моделирования, используемые в современной системной биологии, затрудняют воспроизведение результатов численного моделирования. Для решения этой задачи научным сообществом было разработано 2 стандарта:

- MIASE – Minimum Information About a Simulation Experiment (Waltemath et al., 2011a) описывает минимальный набор информации, которая должна быть предоставлена, чтобы сделать описание имитационного эксперимента доступным для других. Он включает в себя список используемых моделей и их модификаций, все применяемые процедуры моделирования и их порядок, обработку численных результатов и их представление.
- SED-ML – Simulation Experiment Description Markup Language (Waltemath et al., 2011b; <https://sed-ml.org/>) – обеспечивает XML формат для представления всей информации, указанной в MIASE, необходимой для воспроизведения вычислительных экспериментов и их результатов (графики и т.п.).

Краткое описание формата SED-ML приведено в приложении 1.2.

1.2 Методы моделирования биологических систем

1.2.1 Математические методы

Для моделирования биологических систем используются различные математические методы и подходы (таблица 1.2.1). Эти методы можно разделить на непрерывные и дискретные:

- непрерывные – основой модели биологической системы является набор дифференциальных уравнений, в модели также могут использоваться алгебраические уравнения;
- дискретные – модель биологической системы представляет собой набор дискретных состояний и правил перехода между ними.

С точки зрения воспроизводимости результаты методы можно разделить на:

- детерминированные – при разных запусках модели с одними и теми же параметрами и начальными значениями получается один и тот же результат;
- стохастические – модель биологической системы описывается как некоторый вероятностный процесс, поэтому при численном моделировании биологической системы, при разных запусках программы мы можем получить разный результат.

Кроме того, существуют гибридные методы, которые могут объединять несколько описанных выше подходов. Например, часть модели описывается набором дифференциальных уравнений, а другая часть – набором дискретных состояний и переходами между ними, причем при переходе между состояниями могут изменяться как параметры и начальные значения в системе дифференциальных уравнений, так и вся система уравнений. Другой вариант – часть системы с малым количеством молекул моделируется с использованием стохастического подхода, а другая часть – с большим количеством молекул – с использованием ОДУ.

Наиболее часто используемым подходом для моделирования на данный момент в системной биологии являются гибридные модели, на основе системы ОДУ и событий, которые задаются условием наступления события и набором переменных и значений параметров дифференциальных уравнений, которые должны быть изменены при его наступлении. Поэтому такой подход был принят за основу стандарта SBML.

Таблица 1.2.1 – Основные методы моделирования динамики биологических систем

Методы	Примеры моделей (ссылки)
<i>непрерывные, детерминированные</i>	
обыкновенные дифференциальные уравнения, алгебро-дифференциальные уравнения	Bliss et al., 1982; Goldbeter, 1995; Mestl et al., 1995a, b; Endy et al., 1997; Koh et al., 1998; Carrier, Keasling, 1999; Smolen et al., 2000; Chen et al., 2000, 2004; Novak et al., 2001; De Jong, 2002
обыкновенные	Landahl, 1969; Mahaffy, 1984; Smolen et al., 2000;

дифференциальные уравнения с запаздывающим аргументом	MacDonald, 2008
поточковые методы, constraints-based models	Covert, Palsson, 2003; Price et al., 2003; Covert et al., 2004
<i>дискретные, детерминированные</i>	
Булевы сети	Weisbuch, 1986; Kauffman, 1993; Somogyi, Sniegoski, 1996; Huang, 1999; Shmulevich, Kauffman, 2004
обобщенные логические сети	Thomas, d'Ari, 1990; Thieffry, Thomas, 1995; Thomas et al., 1995; Sanchez et al., 1997; Mendoza et al., 1999
сети Петри	Reddy et al., 1993; Hofestadt, Meineke, 1995; Peleg et al., 2002; Moore, Hahn, 2003
<i>стохастические</i>	
стохастическое моделирование химической кинетики	Gillespie, 1976; Gibson, Bruck, 2000; Kierzek et al., 2001; Li et al., 2008
вероятностные Булевы сети	Shmulevich et al., 2002a,b
стохастические сети Петри	Goss, Peccoud, 1998
<i>гибридные</i>	
гибридные сети Петри	Matsuno et al., 2000; Doi et al., 2004
обобщенный пороговый метод моделирования	Tchuraev, 1991, Prokudina et al., 1991; Чураев, Галимзянов, 2001; Tchuraev, Galimzyanov, 2003
нейронные сети и их обобщения	Taylor, Beiu, 1996; Hanai, Honda H, 2004

Для формализации алгоритмов моделирования и анализа биологических моделей, описания характеристик и параметров этих методов была разработана специальная онтология: – KiSAO – The Kinetic Simulation Algorithm Ontology (Courtot et al., 2011; <https://www.ebi.ac.uk/ols/ontologies/kisao>). Она состоит из 3 основных разделов (рисунок 1.2.1):

- алгоритмы моделирования;
- характеристики алгоритмов моделирования;
- параметры алгоритмов моделирования.

Версия 2.30 (последнее обновление, 12 октября 2021 г.) содержит 510 терминов.

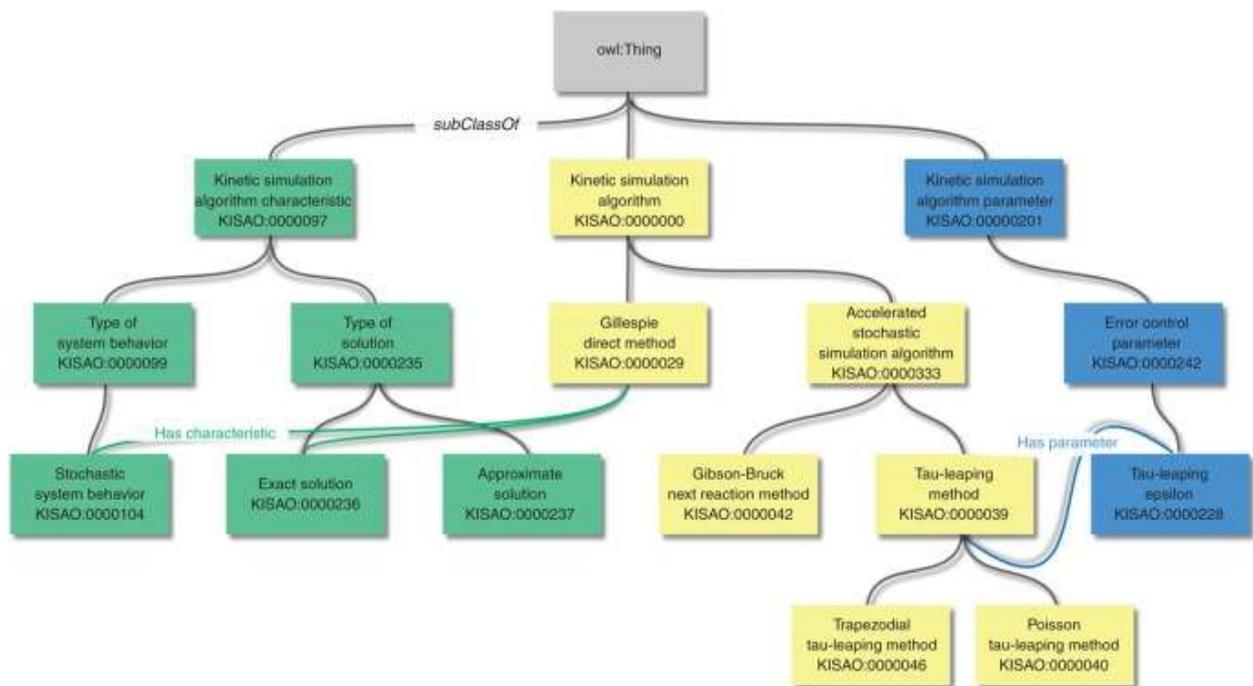


Рисунок 1.2.1 – Основные разделы онтологии KiSAO (Courtot et al., 2011). Элементы разделов связаны друг с другом при помощи отношений "Has characteristic" и "Has parameter"

1.2.2 Визуальное моделирование

Задача моделирования сложных систем может быть существенно упрощена с точки зрения исследователя путем представлении моделей систем в виде диаграмм и использования компьютерных программ, поддерживающих визуальное моделирование.

Визуальное моделирование – это формальное графическое представление систем и процессов в виде диаграмм и последующее моделирование их динамики на основе такого графического представления.

Необходимой составной частью визуального моделирования является графическая нотация (визуальный синтаксис, графический язык) которая позволяет формально и полно описать модель (Lee, 2001). Графическая нотация может быть представлена несколькими типами диаграмм, позволяющими описать разные аспекты структуры и функционирования сложных систем с разным уровнем формальности и детализации. Далее на основе такого формального графического представления системы (диаграммы) компьютерная программа автоматически генерирует код для численного моделирования динамики этой системы.

Данный подход был с успехом применен для моделирования сложных физических, электротехнических и программных систем. Разработано множество программных пакетов для визуального моделирования таких систем (Бенькович и др., 2002; Lee, 2001; и другие), наиболее известным из которых является MATLAB/Simulink (The MathWorks, Inc., <http://www.mathworks.com/products/simulink>).

Рассмотрим в качестве одного из примеров (рисунок 1.2.2), классическую SIR модель распространения инфекционных заболеваний, выполненную в инженерном пакете SimInTech (Petuhoff, 2020), имеющем графическую нотацию, сходную с MATLAB/Simulink.

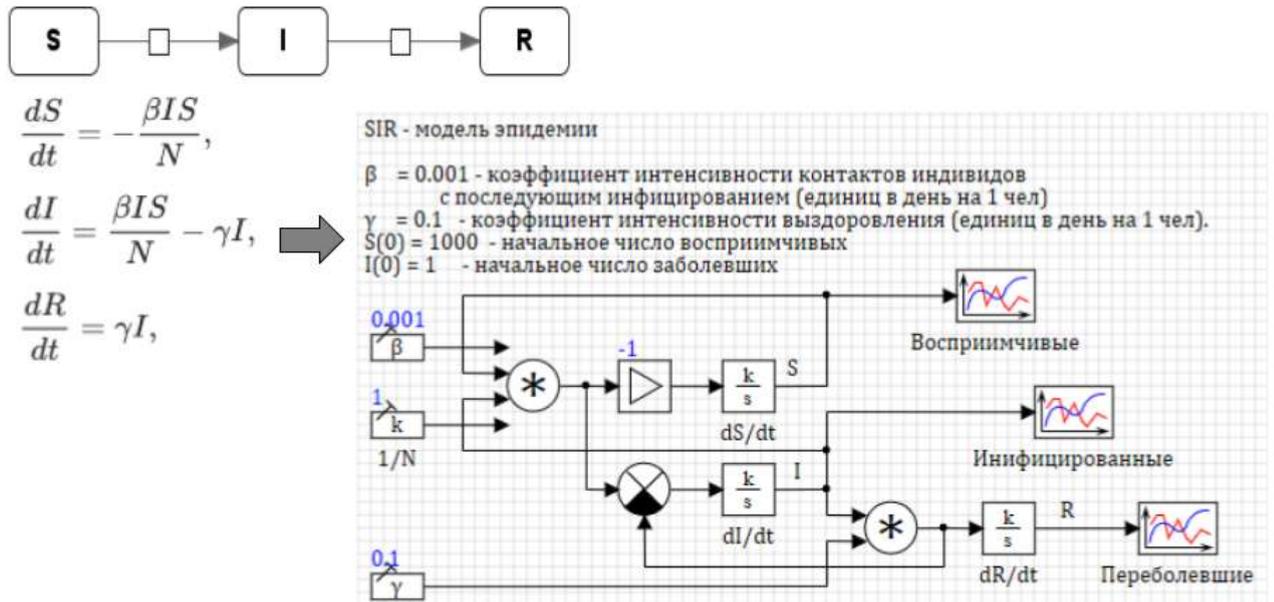
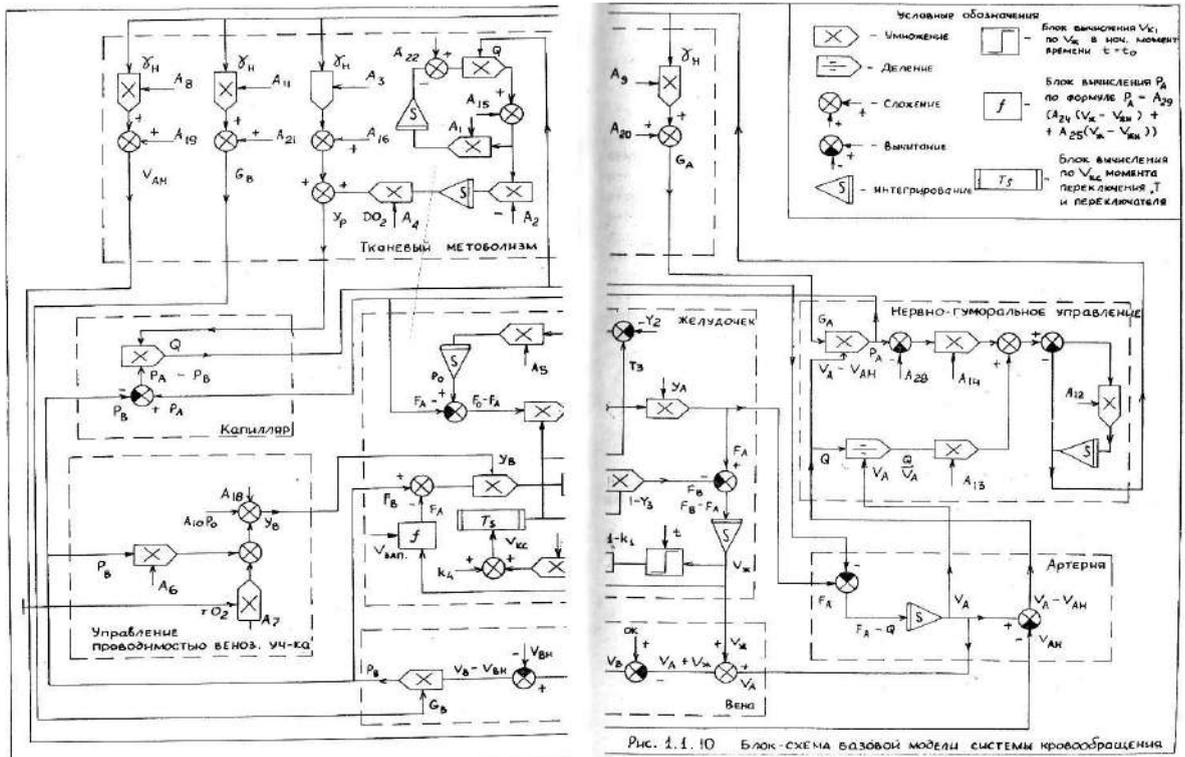
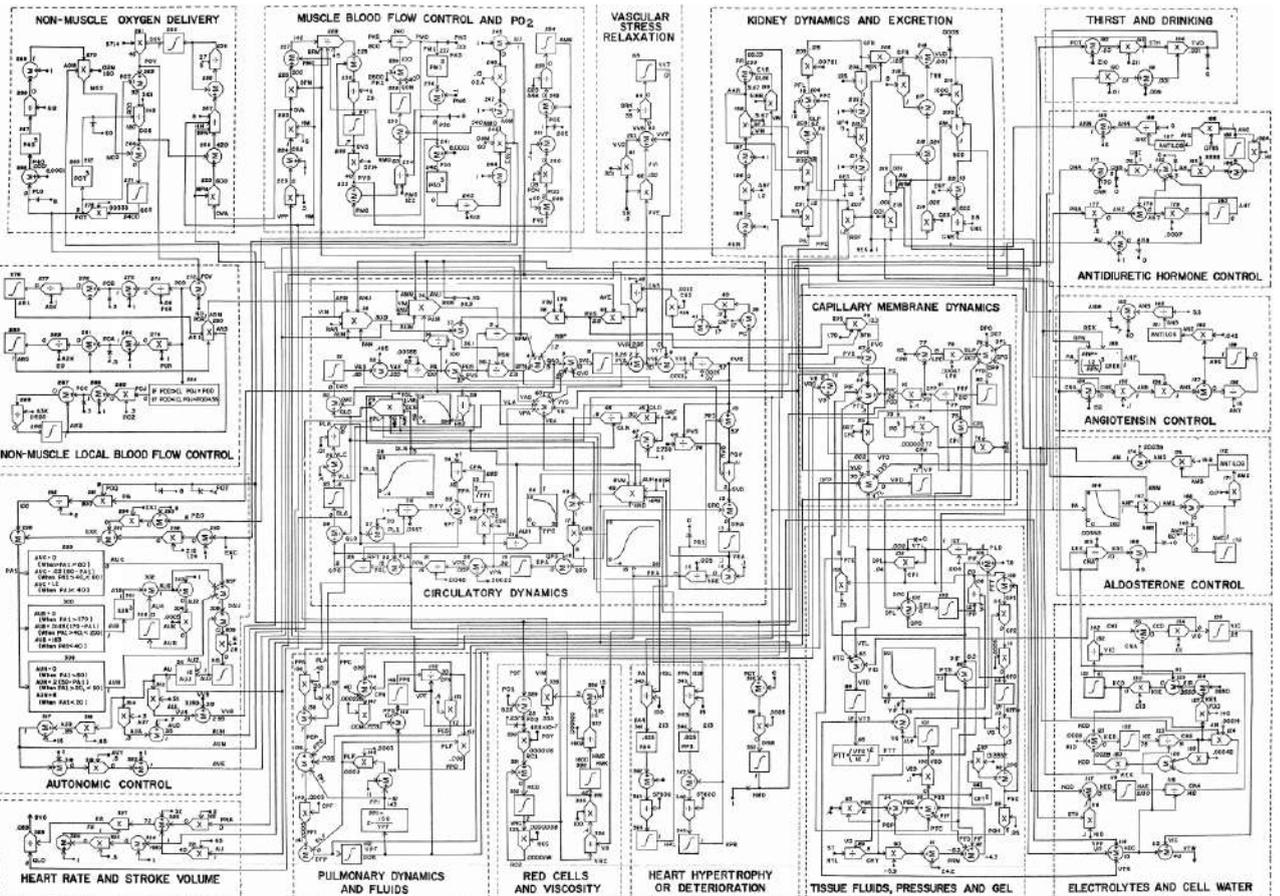


Рисунок 1.2.2 – SIR модель распространения инфекционных заболеваний, выполненную в инженерном пакете SimInTech (Petuhoff, 2020). Сверху – представление в виде графа реакций, слева – соответствующая система дифференциальных уравнений, справа снизу – представление модели в инженерном пакете SimInTech

На рисунке 1.2.3 представлены еще два примера моделей, когда инструменты, разработанные для моделирования физических и электротехнических систем, были применены для моделирования биологических систем.



а)



б)

Рисунок 1.2.3 – Использование графической нотации для электротехнических систем для моделирования биологических систем; а) модель кровообращения (Солодяников Ю.В., 1994); б) модель кровообращения и регуляции артериального давления Гайтона (Guyton, 1972)

Как видно из этих примеров, графическая нотация для моделирования физических и электротехнических систем не подходит для моделирования биологических систем, поскольку:

- теряется биологическая семантика биологического процесса;
- формулы, описывающие скорость биологических процессов, могут быть достаточно сложными, и в таком представлении они очень сложно воспринимаются человеком.

В разделе 1.1.3.4 была описана графическая нотация SBGN. Однако, она не предназначена для визуального моделирования, и поэтому в ней отсутствуют многие необходимые для этого компоненты – уравнения, события, состояния, функции и т.п.

Поэтому, необходимо создание графической нотации для модульного моделирования сложных биологических систем. Наиболее рациональным выглядит расширение графической нотации SBGN, путем добавления в нее необходимых для этого компонентов – уравнений, событий, состояний, функций и т.п.

1.2.3 Модульное моделирование

Принцип управления сложными системами известен еще со времен Древнего Рима – «Divide et Impera» («Разделяй и властвуй»). При моделировании сложных систем необходимо ее разделять на все меньшие и меньшие подсистемы, каждую из которых можно создать и совершенствовать независимо.

Модульный подход к математическому моделированию сложных биологических систем заключается в разделении системы на части, для каждой из которых модель может быть построена отдельно. Модель всей системы строится как комбинация моделей подсистем (модулей). Взаимодействие блоков моделируется связями между переменными модулей, которые указывают на пути передачи сигналов между ними. Модули представляют собой математические модели различных подсистем, также они могут содержать вложенные модульные модели, составляя иерархическую структуру.

Основные преимущества модульного подхода:

- в упрощении восприятия модели человеком – на каждом иерархическом уровне модель представляется в виде сравнительно небольшого количества блоков (как правило, не более 20);
- в упрощении работы с моделью – модули могут редактироваться по отдельности;
- итерационном/эволюционном развитии модели – модули могут заменяться на более сложные или проработанных;
- возможность использовать различные математические формализмы – как дальнейшее развитием модульного подхода мы можем рассматривать агентное моделирование, которое, в свою очередь, позволяет использовать модули (агенты) с использованием

различных математических формализмов (ОДУ, УЧП, клеточные автоматы, стохастические модели, и т.д.).

Отдельным вопросом является следующий – как правильно разбить сложную биологическую систему на модули? Насколько это известно автору – единственно-правильного ответа на этот вопрос не существует.

Однако, блочно-модульный принцип является почти очевидным правилом организации биологических систем (Ратнер 1992; Ратнер и др., 1985), поскольку выявление любой структурно-функциональной подсистемы свидетельствует в его пользу. Альтернативным является распределенный способ организации, когда многочисленные компоненты системы одинаковы и полифункциональны, т.е. функции системы распределены между многими дублирующими компонентами системы. В этом случае изменение числа компонент влияет лишь на интенсивность функции. Таковы многокомпонентные химические растворы, смеси. В реальных системах можно видеть использование обеих альтернатив, с преобладанием одной из них. В биологических системах преобладание блочно-модульного принципа организации присуще структурированным системам – клеткам, организмам, а преобладание распределенного – сообществам, популяциям, экосистемам (Ратнер, 1992).

Существует два общих подхода к численным расчетам модульных моделей (Hernandez et al., 2009; Vangheluwe, 2000; Киселев, 2016):

- 1) Трансформация модульной модели в модель пригодную для численных расчетов одним из стандартных методов. При этом, если модули выполнены в разных математических формализмах, необходимо преобразовать их в единый формализм. Пример разработанного на этих принципах подхода (в предположении одинакового формализма модулей) можно найти, например, в работе Vangheluwe, 2000. Плюсами данного подхода является строгое математическое обоснование численных расчетов модели. Минусом – жесткие ограничения на используемые математические формализмы модулей.
- 2) Подход, использующий принципы агентного моделирования. В рамках данного подхода, расчеты для каждого модуля производятся отдельно, при этом обеспечивается корректный обмен информацией между ними. Минусом подхода является недостаток математического обоснования, в частности остаются открытыми вопросы существования и единственности решения, однако подход является универсальным и позволяет объединять в рамках одной модульной модели модули только на основе соответствия их интерфейсов, ничего не зная о внутренней структуре модулей. Пример создания модели на основе данного подхода см. в работе Hernandez et al., 2009.

Другим важным вопросом является то, каким образом модульная модель может взаимодействовать с элементами подмоделей. Элементы подмоделей (уравнения, химические

реакции, переменные и т.п.) могут быть явно доступны для других частей модели. Другой вариант – по умолчанию все элементы недоступны извне модуля. Для того чтобы элемент стал доступен, модуль должен определить дополнительный элемент: порт. Совокупность таких портов определяет интерфейс модуля. Первый подход более гибкий, в то время как второй позволяет создавать более контролируемые и хорошо определенные модульные модели.

Если модели подсистем изначально разработаны с целью объединения в модульную модель, они могут определять свои интерфейсы для взаимодействия с другими подмоделями в рамках модульной модели. В этом случае, создание модульной модели сводится к выбору нужных подмоделей и установке нужных связей между их интерфейсами.

Обсуждение включения синтаксиса для описания модульных моделей в язык SBML продолжается с момента создания языка в 2000 году, за это время было предложено несколько вариантов от различных авторов. Официальная спецификация для расширения языка SBML (SBML “comp”) была опубликована в 2013 году (Smith et al., 2013). Одновременно был опубликован набор тестов для этого расширения.

Также вопросы модульного моделирования биологических систем подробно рассматриваются в кандидатской диссертации Киселева И.Н. (Киселев, 2016), выполненной под руководством автора этой диссертационной работы.

1.2.4 Агентное моделирование

Агентное моделирование – раздел имитационного моделирования, которое подразумевает представление системы в виде набора взаимодействующих сущностей способных самостоятельно принимать решения – агентов. Общее поведение системы складывается из взаимодействия агентов. Агентное моделирование связано с модульным в том смысле, что оно тоже моделирует систему как набор связанных подсистем.

Не существует общепринятого определения, что такое агент. Из практических соображений, принимают, что агент должен обладать следующими свойствами (Macal and North, 2009; Киселев, 2016):

- агент должен быть автономным, действовать независимо от других агентов за исключением ряда определенных взаимодействий;
- агент должен быть определен как отдельная сущность со своим набором характеристик;
- агент должен взаимодействовать с другими агентами;
- часто агенты рассматриваются как существующие в некоторой внешней среде.

Существуют различные классификации агентов, в частности агенты могут быть (Wooldridge, 2009):

- реактивные – такие агенты реагируют на сигнал от внешней среды или других агентов, и генерирует определенную реакцию на этот сигнал;
- адаптивные или агенты с состоянием – такие агенты хранят внутреннее состояние и под воздействием внешнего сигнала меняют его. Действие агента зависит от того как изменилось внутреннее состояние;
- проактивные – такие агенты имеют некоторую глобальную цель и действующими так, чтобы достичь этой цели. Это свойство может быть реализовано путем задания функции оценки текущего состояния и корректировки действий агента с целью улучшения текущего состояния;
- обучающиеся, эволюционирующие, рассуждающие и др.

Агентное моделирование в основном используется для моделирования систем, поведение которых на глобальном уровне слишком сложно или невозможно формализовать: моделирование социальных процессов (Bonabeau E., 2002), фондовых рынков (Chan, 2001), предсказания распространения инфекционных заболеваний (Perez and Dragicevic, 2009) и т.д.

Метод также хорошо подходит для изучения сложных биологических систем, например, для взаимодействия большого количества клеток разного типа. Агенты могут создаваться, уничтожаться или каким-то образом изменяться (например, переходить из одной фазы клеточного цикла в другую) в процессе расчетов.

В данный момент разрабатывается расширение “dyn” для языка SBML, предназначенное, в первую очередь, для создания моделей клеточных популяций. Оно позволяет задавать правила, по которым в процессе численных расчетов должны динамически удаляться и создаваться новые элементы модели, а также задавать пространственные координаты объектам и возможность им перемещаться (Stevens and Myers, 2013).

1.2.5 Цифровые двойники и виртуальные организмы

Цифровой двойник — это цифровая (виртуальная) модель любых объектов, систем, процессов или людей. Она достаточно точно воспроизводит форму и действия оригинала и синхронизирована с ним. Цифровой двойник нужен, чтобы смоделировать, что будет происходить с оригиналом в тех или иных условиях.

Впервые концепция цифрового двойника была предложена профессором Мичиганского университета Майклом Гривсом (Michael Grieves) еще в 2002 г. Первоначально она называлась «моделью зеркальных пространств» (Grieves, 2017). Далее он определяет цифрового двойника следующим образом: «Цифровой двойник представляет собой набор виртуальных информационных конструкций, который полностью описывает потенциальное или фактическое промышленное изделие: от его атомарных функций до геометрии. В идеальных условиях вся

информация, которую можно получить от изделия, может быть получена от его цифрового двойника» (Grieves, 2017).

Цифровой двойник содержит три основные части (Grieves, 2014):

- физический продукт в реальном пространстве;
- виртуальный продукт в виртуальном пространстве;
- данные и информацию, которые объединяют виртуальный и физический продукт.

Таким образом, математические модели сложных биологических систем могут лечь в основу соответствующих цифровых двойников.

Термин «виртуальный пациент» используется десятки лет в разных контекстах: медицинском обучении, электронной истории болезни, клинических исследованиях. В каждом контексте есть свое понимание, что же такое «виртуальный пациент» (Kononowicz et al., 2015).

В данной работе мы будем использовать термин "виртуальный пациент" – как часть цифрового двойника реального пациента, который представлен математической моделью физиологии и биохимии человека, причем параметры этой модели подобраны так, чтобы максимально соответствовать наблюдаемым клиническим данным для заданного реального пациента (Lehrach, 2016).

Термин "виртуальный" также активно используется при построении моделей, как на уровне отдельных клеток, так и на уровне целых органов, физиологических систем и целых организмов:

- **VCell** – Virtual Cell – виртуальная клетка (Blinov et al., 2017; <https://vcell.org>) – это комплексная платформа для моделирования биологических систем клетки, которая построена на центральной БД и распространяется в виде веб-приложения.
- **The Virtual Liver** – виртуальная печень (Drasdo et al., 2014; <https://www.h-its.org/projects/virtual-liver/>) – междисциплинарный исследовательский проект, финансируемый Федеральным министерством образования и исследований Германии.
- **The Virtual Brain** – виртуальный мозг (Sanz Leon et al., 2013; <https://www.thevirtualbrain.org/>) – модель мозга пациента, которая может достаточно точно воспроизводить результаты ЭЭГ, МЭГ, фМРТ. за счет уменьшения сложности в миллион раз с помощью методов, подобных статистической физике.
- **The Virtual Physiological Rat Project** – проект физиологии виртуальной крысы (Beard et al., 2012; <https://virtualrat.org/>) – проект сосредоточен на системной биологии сердечно-сосудистых заболеваний – понимании того, как фенотипы болезни, проявляющиеся в масштабе всего организма, возникают в результате взаимодействий систем на молекулярном, клеточном, тканевом и органном уровнях.

- **The Virtual Physiological Human** – физиология виртуального человека (Coveney et al., 2011; <http://physiomeproject.org/about/the-virtual-physiological-human>) – европейская инициатива, которая фокусируется на методологической и технологической структуре, которая позволит проводить совместные исследования человеческого тела как единой сложной системы.
- **Virtual patient** – виртуальный пациент (<http://virtual-patient.com/>) – интерактивное программное приложение Web 2.0, для обучения и оценки студентов медицинских учебных заведений.

1.3 Программные комплексы для моделирования биологических систем

В настоящее время в мире существуют тысячи программ и программных комплексов, предназначенных для моделирования биологических систем.

Также существует большое количество программных продуктов для создания и работы с моделями биологических систем. Список продуктов, поддерживающих язык SBML включает в себя более чем 300 программных продуктов (<https://en.wikipedia.org/wiki/SBML>). Однако большая их часть не поддерживает модульный подход. Причина этого в том, что SBML сравнительно недавно был расширен синтаксисом для описания модульных моделей. Среди них и наиболее популярные программы для работы с SBML моделями:

CellDesigner (Matsuoka et al., 2014; <https://www.celldesigner.org/>) – графический редактор SBML моделей, используя графическую нотацию SBGN Process Diagram. Стоит отметить, что основой SBGN Process Diagram является графическая нотация, предложенная Хироаки Китано для CellDesigner, и которая впоследствии, с небольшими уточнениями была принята как стандарт. CellDesigner поддерживает моделирование и сканирование параметров за счет интеграции с SBML ODE Solver, SBML Simulation Core и Copasi.

COPASI (Bergmann et al., 2017; <http://copasi.org/>) – это программное приложение для моделирования и анализа биохимических сетей и их динамики. Оно поддерживает модели в стандарте SBML и может численно моделировать динамику биологических систем, используя ОДУ и алгоритм стохастического моделирования Гиллеспи. COPASI также включает набор методов анализа и оценки параметров.

VANTED (Rohn et al., 2012; <http://vanted.org>) – популярный инструмент визуализации SBML-моделей в виде SBGN-диаграмм. Он позволяет пользователям создавать и редактировать биологические сети, а также сопоставлять экспериментальные данные с этими сетями.

1.3.1 Модульное моделирование

Ниже приводится краткий обзор программных продуктов для системной биологии, поддерживающих концепцию модульности. Некоторые из них уже не поддерживаются и не развиваются, но представляют исторический интерес.

ProMoT (Mirschel et al., 2009) – один из первых инструментов для моделирования, поддерживающих модульный подход. Он представляет объектно-ориентированный язык, способный описывать модульные модели, использующие ОДУ-формализм. На данный момент сайт проекта (<http://www2.mpi-magdeburg.mpg.de/projects/promot/>) не доступен.

iBioSim (Myers et al., 2009, Watanabe et al., 2018; <https://github.com/MyersResearchGroup/iBioSim>) – сейчас позиционируется как инструмент автоматизированного проектирования для синтетической биологии, т.е. для моделирования, анализа и проектирования генетических систем. Его также можно использовать для моделирования метаболических сетей, клеточных сигнальных путей и других биологических и химических систем, он также включает поддержку моделирования и визуализации для многоклеточных и пространственных моделей. iBioSim поддерживает все основные конструкции SBML 13v1, за исключением некоторых типов быстрых реакций, а также поддерживает пакеты для иерархической композиции моделей (comp), отображения диаграмм (layout), баланса метаболических потоков (fbc) и массивы (arrays) .

JigCell Model Connector (Jones et al., 2018) – программа для создания иерархических моделей. Он использует ту же концепцию портов для соединения модулей, как и BioUML (Kutumova et al., 2012). Выходной порт (output) связан с внутренним компонентом, который будет отправлять значение в соответствующий входной порт (input). Порт эквивалентности (equivalence port) связан с внутренним компонентом, который будет получать и отправлять значения из другого соответствующего порта эквивалентности. Сейчас это часть проекта Copasi (http://copasi.org/Projects/JigCell_Model_Connector).

TinkerCell (Chandran et al., 2009; Chandran and Sauro, 2012; <http://www.tinkercell.com/>), подобно iBioSim, поддерживает как модульность моделей, так и визуальное представление. Благодаря собственной модульной структуре, TinkerCell позволяет подключать дополнительные пакеты для численных расчетов и анализа. Последние обновления датируются 2012 годом.

Существуют также специализированные библиотеки, для проведения численных расчетов, поддерживающих модульные SBML модели:

M2SL – Multi-formalism Modeling and Simulation Library (Hernandez et al., 2009) – поддерживает модели с множественным формализмом используя ко-симуляционный подход, не поддерживает визуальное моделирование и стандарты SBML и CellML. Используя этот подход была воспроизведена модель Гайтона (Guyton, 1972). Веб-сайт проекта сейчас не доступен.

libRoadRunner (Welsh et al., 2023; <https://github.com/sys-bio/roadrunner>) – это библиотека, которую можно использовать в других программах, через API-интерфейсы на языках C++, C и Python или в интерактивном режиме через интерфейс Python или Julia. libRoadRunner использует специальный JIT-компилятор, созданный на основе широко используемой среды JIT-компилятора LLVM. Он компилирует указанные SBML модели непосредственно в машинный код, что делает его достаточно быстрым для моделирования больших моделей. libRoadRunner поддерживает большую часть спецификации SBML (за исключением задержки и нелинейных алгебраических уравнений), а также несколько расширений SBML, включая пакет для иерархической композиции моделей (comp). Он предлагает несколько численных решателей, как для детерминированного так и для стохастического моделирования, а также инструменты для стационарного анализа, анализа чувствительности, устойчивости и структурного анализа.

The Systems Biology Simulation Core Library (Panchiwala et al., 2022; <https://github.com/draeger-lab/SBSCL>) – Java библиотека для численного моделирования, сложных биологических систем, используя ОДУ, стохастические дифференциальные уравнения и анализ на основе ограничений метаболических потоков. Она поддерживает стандарты SBML, включая пакеты comp и dba, SED-ML и архивы COMBINE.

1.3.2 Агентное моделирование

Ниже приведен краткий обзор наиболее популярных инструментов для агентного моделирования общего назначения. В силу общности подхода агентного моделирования, эти программы с успехом применяются для моделирования и биологических систем, например, модели "хищник-жертва", где хищники и жертвы распределены по некоторой территории, для моделирования распространения инфекционных заболеваний и т.п.

Swarm (Minar et al. 1996; <http://www.swarm.org/>) — это платформа для агентных моделей, которая включает концептуальную основу для проектирования, описания и проведения экспериментов с агентными моделями. Она является одним из первых инструментов для создания агентных моделей и была разработана в 1994 г. в институте Санта Фе. На данный момент проект не активен, последняя версия, Swarm 2.4 вышла в 2009 году.

Repast Suite (North et al., 2013; <https://repast.github.io/>) — это семейство платформ с открытым исходным кодом для моделирования на основе агентов, которые непрерывно разрабатывались более 20 лет:

- Repast Symphony — позволяет создавать модели с помощью кода на языке Java;
- Repast for High Performance Computing — версия для высокоэффективных распределенных вычислений на кластерах и суперкомпьютерах, реализована на языке C++;

- Repast for Python — версия для языка Python.

NetLogo (Wilensky and Rand, 2015; <https://ccl.northwestern.edu/netlogo/>) — по-видимому, самая популярная платформа для агентного моделирования. На данный момент Google Scholar's содержит более 38 600 ссылок на NetLogo. Она предоставляет графический интерфейс и специализированный язык программирования NetLogo для создания агентных моделей.

MASON — Multi-Agent Simulator Of Networks (Luke et al., 2005; <https://github.com/eclab/mason>) — это Java библиотека для дискретно-событийного многоагентного моделирования. Она также содержит дополнительный набор инструментов визуализации результатов в 2D и 3D.

Ascape (Parker, 2001; <https://ascape.sourceforge.net>) — Java платформа для агентного моделирования общего назначения. Название Ascape отражает две фундаментальные концепции, лежащие в его основе: "agents" и «scapes» — территория, где живут эти агенты. Термин «скейп» происходит от Sugarscape (1), где он относится к территории, на которой живут агенты (Inchiosa and Parker, 2002). С 2010 года проект не обновляется.

Jade – Java Agent DEvelopment Framework (Bellifemine et al., 2003; <https://jade.tilab.com/>) – фреймворк с открытым исходным кодом для многоагентных распределенных систем. Он поддерживает набор графических инструментов для отладки и развертывания на кластере – расчет модели на основе Jade можно распределить между несколькими компьютерами.

AnyLogic (Карпов, 2006; <https://www.anylogic.ru>) – отечественный коммерческий программный комплекс с закрытым кодом на языке Java для имитационного моделирования. Он предоставляет возможности для создания дискретно-непрерывных и имитационных моделей, в основном в областях управления бизнес-процессами, производством, финансами и социальной динамикой. AnyLogic предоставляет графический интерфейс для создания моделей, при этом поведение агентов задается пользователем путем написания кода на языке Java.

В работе (Bajracharya and Duboz, 2013) одна и та же модель распространения заразных заболеваний была реализована в трех разных программных продуктах, включая NetLogo и Repast. Показано, что результаты численных расчетов значительно отличаются. Авторы предполагают, что причина может быть в том, что пользователю не всегда предоставлен полный контроль над тем, как проходят численные расчеты – различные программные продукты используют различные стратегии управления агентами, не всегда эти стратегии в достаточной степени документированы. Таким образом, необходимо с осторожностью относиться к результатам расчетов выполненных с помощью агентного моделирования (Киселев, 2016).

Существуют попытки интегрировать агентное моделирование и модели биологических путей в формате SBML. В данном случае, каждая клетка, является агентом, чье поведение определяется внутриклеточной моделью, представленной в формате SBML.

Например, такой подход реализован в сценарии EPISIM (Sütterlin et al., 2013). В нем внутреннее состояние каждой клетки описывается SBML моделью и для численного расчета таких моделей используется программа COPASI. Для моделирования ткани, состоящей из клеток, как агентов, используется программа MASON для агентного моделирования. Таким образом, клеточные состояния, такие как пролиферация и дифференцировка, могут быть гибко связаны с генно-регуляторными или биохимическими сетями.

Более подробный обзор существующих методов интеграции SBML моделей для моделирования динамики клеточных популяций приведен в работе Matzko et al., 2023.

1.3.3 BioSimulators

Несмотря на описанные выше стандарты форматов COMBINE (SBML, SED-ML и другие), существенными проблемами совместного использования и повторного использования моделей биологических систем являются различные формализмы моделирования, алгоритмы моделирования, форматы моделей, инструменты моделирования и репозитории для различных типов моделей и биологических систем. Это разнообразие методов, форматов, репозитория и инструментов затрудняет, особенно для неспециалистов, поиск моделей и инструментов моделирования для их повторного использования. Кроме того, существующие репозитории моделей имеют ограниченные возможности для совместного использования связанных ресурсов, таких как обучающие данные, эксперименты по моделированию и интерактивные визуализации их результатов.

Проект BioSimulators (Shaikh et al., 2022; <https://biosimulators.org>) пытается решить эти проблемы, создав реестр программ для моделирования, а также набор стандартов и соглашений для описания и запуска программ для моделирования биологических систем (рисунок 1.3.1):

- формат для описания программ – позволяет описывать сами программы (авторы, лицензия, ссылки на документацию, Docker образ и т.п.), поддерживаемые алгоритмы моделирования (их параметры, типы данных и значения по умолчанию), форматы моделей, поддерживаемые инструментом моделирования. Формат также позволяет исследователям описывать параметры каждого алгоритма, их типы данных и значения по умолчанию;
- соглашения для запуска программ из командной строки – они обеспечивают, что запускаемые программы поддерживают согласованный синтаксис для запуска расчетов и сохраняют полученные результаты в согласованных директориях и форматах;
- соглашения о Docker образах — формализуют входные точки (entry points) для запуска программ, а также описания мета-данных;
- формат SED-ML – должен использоваться для описания экспериментов по моделированию;

- формат `sedml:report` – для представления отчетов о результатах моделирования;
- рекомендации по визуализации данных с помощью Vega. Vega (<https://vega.github.io>) – это формат для описания графического представления различных графиков и карт. В этих рекомендациях описывается, как использовать Vega для визуализации результатов экспериментов по моделированию, зафиксированных в отчетах SED-ML;
- рекомендации по использованию формата метаданных OMEX для аннотирования данных проектов по моделированию (архивы COMBINE/OMEX);
- формат журналов выполнения проектов моделирования (архивы COMBINE/OMEX) – позволяет программам сообщать о статусе и результатах выполнения расчетов.

BioSimulators описывает 54 программы для 13 форматов моделей и 91 алгоритм моделирования, а также интерфейс для запуска 21-ой из этих программ (Shaikh et al., 2022).

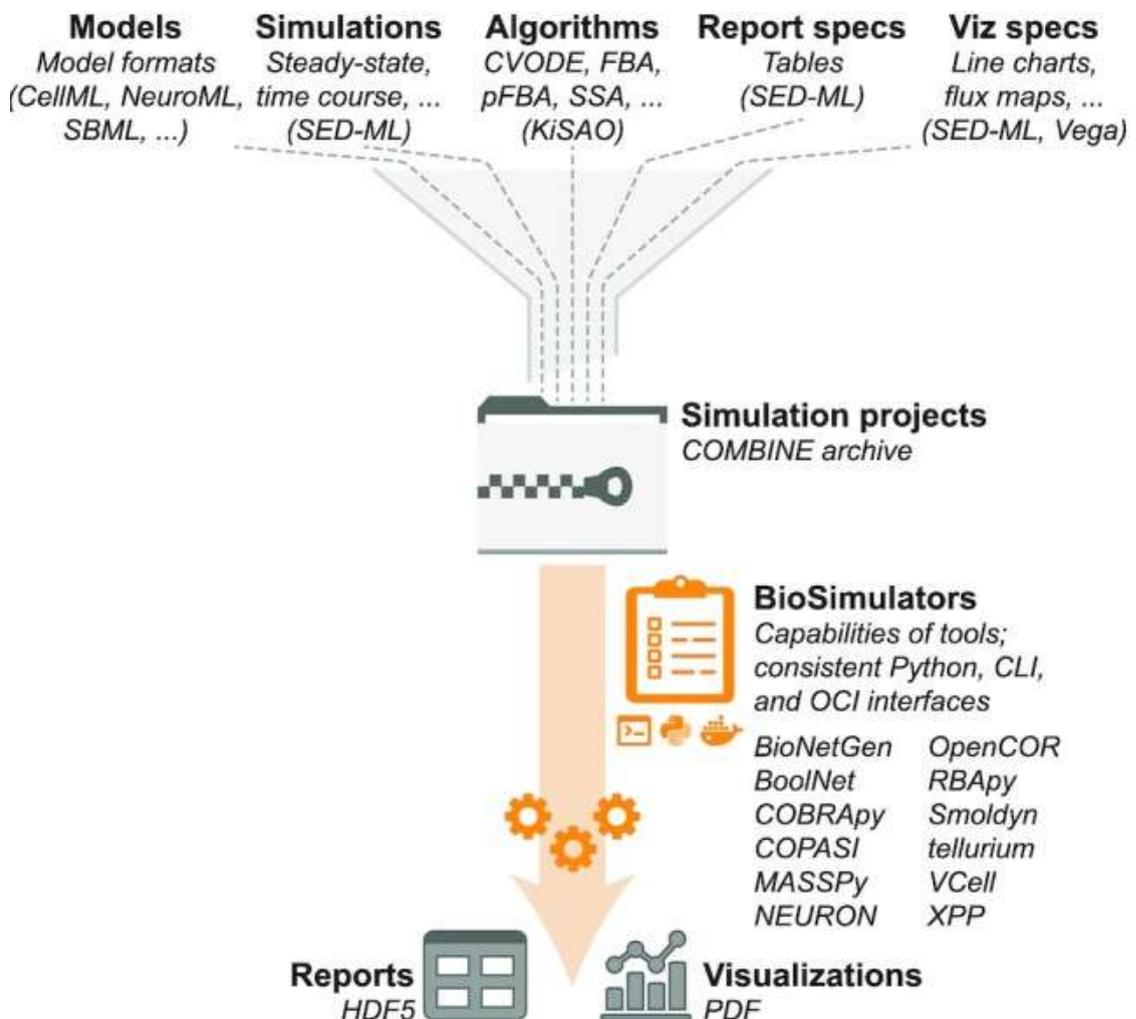


Рисунок 1.3.1 – Основные стандарты и подходы проекта BioSimulators (Shaikh et al., 2022)

1.4 Анализ биомедицинских данных

Как отмечалось во введении, полный цикл исследований в системной биологии, включает в себя не только моделирование биологических систем, но и анализ соответствующих экспериментальных данных. Как правило, это генетические и омиксные данные. Однако, рассматриваемые ниже подходы применимы и ко многим другим типам биомедицинских данных.

1.4.1 Установка и запуск программ

Установка программ вместе со всеми необходимыми программными компонентами может быть достаточно сложной задачей. Для упрощения этой задачи готовый к выполнению код программ, вместе со всеми необходимыми программными модулями и другими программами упаковывается в специальные контейнеры.

Контейнер – это пакет программного обеспечения, который содержит все зависимости (код, среду выполнения, конфигурацию и системные библиотеки) и может выполняться на различных операционных системах, взаимодействуя с ней при помощи стандартных интерфейсов.

Контейнеры более «легкие» и компактные, чем традиционные виртуальные машины, они очень быстро запускаются, используют общую операционную систему. Их можно создавать гораздо быстрее, чем экземпляры виртуальных машин.

Docker (<https://www.docker.com>) – наиболее часто используемая технология контейнеризации, которая позволяет упаковать в образ (image) различные программы и данные, так что затем их можно запустить на любом сервере или компьютере. Для хранения таких образом используется специальное хранилище (Docker registry), откуда такой образ может быть легко извлечен по его идентификатору и далее запущен на сервере. DockerHub (<https://hub.docker.com/>) – самое крупное и популярное хранилище Docker образов.

С точки зрения работы Docker его основой является служебный демон dockerd, который управляет работой контейнеров. Важно отметить, что он работает только при наличии привилегий суперпользователя, что вызывает сложности с безопасностью работы Docker в инфраструктуре высокопроизводительных кластеров (Тахтеев, 2022).

Singularity (<https://sylabs.io/>) – другая технология контейнеризации, которая изначально разрабатывалась под архитектуру высокопроизводительных кластеров. Ее ключевые особенности:

- неизменяемый однофайловый формат образов контейнеров, поддерживающий криптографические подписи и шифрование;

- интеграция устройств поверх изоляции, позволяющая нативно использовать GPU, вычислительную сеть и параллельные файловые системы;
- простая и эффективная модель безопасности: пользователь и группа внутри контейнера совпадают с теми, с которыми контейнер был запущен.

Singularity позволяет использовать образы Docker как непосредственно для выполнения, так и в качестве базового образа в инструкциях по сборке Singularity (Тахтеев, 2022).

В качестве альтернативного контейнеризации подхода часто используется **Conda** (<https://docs.conda.io/>) – система управления пакетами с открытым исходным кодом и система управления средой, которая работает в Windows, macOS и Linux. Изначально она была создана для программ Python, но может упаковывать и распространять программное обеспечение для любого языка. Conda позволяет легко устанавливать и обновлять различные версии бинарных программных пакетов и любых необходимых библиотек, подходящих для компьютера пользователя из репозитория программного обеспечения. Наиболее известным таким репозиторием является Anaconda (<https://anaconda.org/>).

1.4.2 Реестры программ

Реестры программ – централизованные списки существующих программ для анализа биомедицинских данных существенно упрощают поиск программ для решения заданной задачи. Кроме того, реестры помогают решить проблему с поисками подходящей версии одной и той же программы.

1.4.2.1 Dockstore

Dockstore (Yuen et al., 2021; <https://dockstore.org/>) – это платформа с открытым исходным кодом для публикации, обмена и поиска программ и сценариев, используемых в биоинформатике.

Она состоит из следующих компонентов (рисунок 1.4.1):

- реестр программ для анализа биомедицинских данных. Каждая программа имеет :
 - дескриптор, который содержит мета-информацию о программе, описывает ее параметры и командную строку для запуска. Для написания файлов дескрипторов используются языки CWL или WDL (см. раздел 1.4.3);
 - файл или ссылки на git репозиторий, где описывается построение соответствующего Docker образа;
 - ссылка на репозиторий, где хранится Docker образ программы.
- реестр сценариев для анализа биомедицинских данных на одном из поддерживаемых языков для написания сценариев: CWL, WDL, Nextflow или Galaxy(см. раздел 1.4.3);

- Docstore CLI (Command Line Interface) – программа для запуска программ и сценариев из реестров, описанных выше.

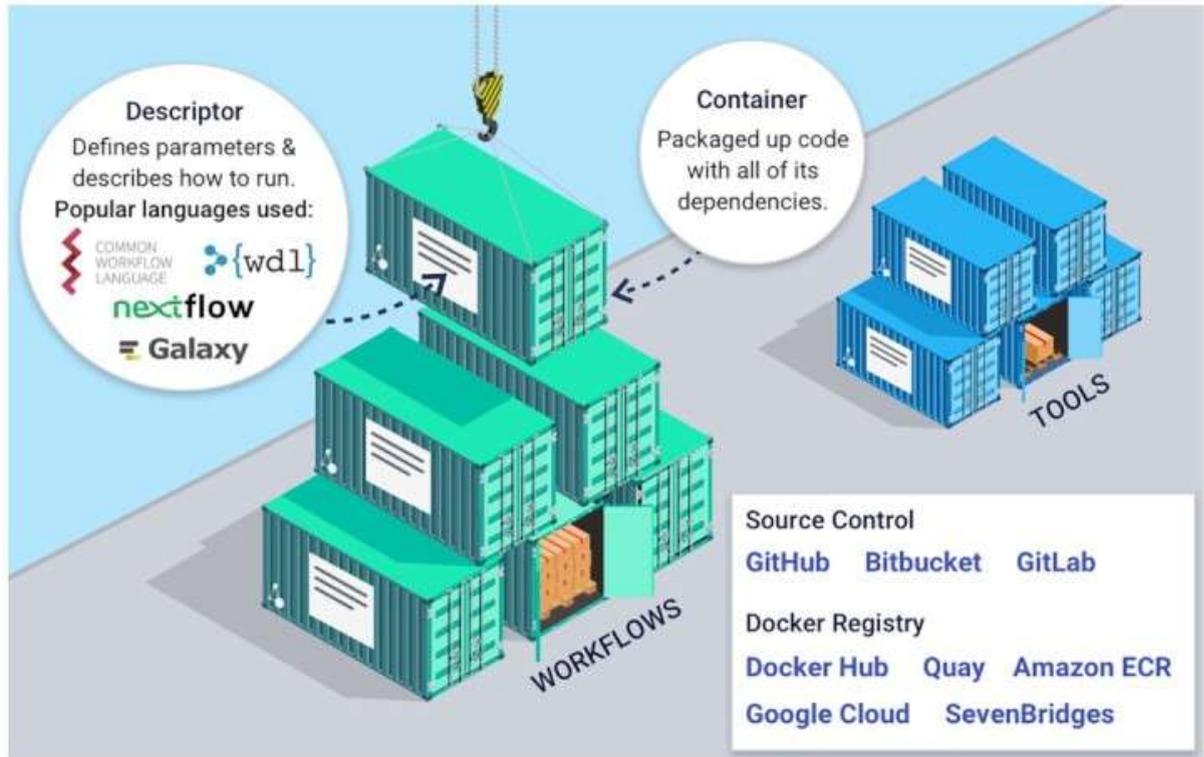


Рисунок 1.4.1 – Основные идеи Dockstore (Yuen et al., 2021)

1.4.2.2 Biocontainers

Biocontainers (da Veiga Leprevost et al., 2017; <https://biocontainers.pro/>) – это реестр контейнеров и сценариев для анализа биомедицинских данных. Этот проект также предоставляет инфраструктуру и основные рекомендации для создания, управления и распространения программ в виде пакетов (например, conda) и контейнеров (Docker, Singularity).

Пользователи и разработчики могут использовать инфраструктуру BioContainers, взаимодействуя через страницу учетной записи GitHub. Все Dockerfile-контейнеры находятся в свободном доступе. Контейнеры можно получить через интерфейс командной строки Docker или загрузив Dockerfile непосредственно из GitHub.

Основные компоненты Biocontainers:

- Docker Containers – список рецептов (Dockerfile) для автоматического создания контейнеров в BioContainers;
- Контейнеры на основе Conda – список рецептов Conda для автоматической сборки сначала пакета conda, а затем контейнера Docker;

- реестр контейнеров, готовых к использованию;
- спецификации – набор правил и соглашений для создания контейнеров.

1.4.2.3 ELIXIR, bio.tools

Панъевропейский проект ELIXIR (<https://elixir-europe.org/>) координирует и развивает ресурсы, связанные с науками о жизни, чтобы исследователям было легче находить, анализировать и обмениваться данными, обмениваться опытом и внедрять передовой опыт.

В рамках этого проекта развиваются соответствующие платформы (<https://elixir-europe.org/platforms/>):

- Compute Platform – обеспечивает инфраструктуру и сервисы, с помощью которых исследователи по всей Европе смогут получать доступ, хранить, передавать и анализировать большие объемы данных науки о жизни.
- Data Platform – определяет ключевые ресурсы данных по всей Европе и поддержку связей между данными и литературой, например, упрощая переход от научной статьи к набору данных, на котором она была основана.
- Tools Platform – помогает исследователям найти подходящие программы для анализа их данных.
- Interoperability Platform – устанавливает общеевропейские стандарты, которые можно использовать для описания данных для наук о жизни. Это упрощает сравнение и анализ различных наборов данных.
- Training Platform – помогает ученым и разработчикам найти нужные учебные курсы и пройти их.

bio.tools (Ison et al., 2019; <https://bio.tools/>) – это составная часть Tools Platform, которая предоставляет реестр программного обеспечения и БД по всем областям биомедицинских исследований, что помогает исследователям находить, использовать и цитировать ресурсы, которые им нужны в их повседневной работе.

Каждой записи в bio.tools присваивается удобочитаемый уникальный идентификатор, основанный на имени ресурса, например, biotools:signalp. Эти идентификаторы обеспечивают постоянную ссылку на соответствующую веб-страницу с описание программы и обеспечивают интеграцию bio.tools с другими ресурсами.

1.4.2.4 Bioconda

Bioconda (Grüning et al., 2018; <https://bioconda.github.io/>) – это дистрибутив программного обеспечения для биоинформатики, реализованный как канал для универсального менеджера пакетов Conda. Она включает более 9 тысяч пакетов.

1.4.3 Сценарии анализа данных

Как правило, анализ биомедицинских и других научных данных включает в себя множество шагов, поскольку данные обрабатываются и анализируются с использованием ряда независимых программ. Чтобы автоматизировать этот процесс, обеспечить возможность повторного запуска процесса такого анализа используются сценарии анализа данных (workflows), которые формализуют процесс обработки.

Сценарии анализа данных могут быть написаны как в виде скриптов (например, Bash) или на высокоуровневом языке программирования (например, Python). Однако, такой императивный подход имеет ряд недостатков:

- сложность переноса таких сценариев в "облако" или на другой компьютер, поскольку, как правило, такие сценарии подразумевают предварительную установку и настройку необходимых программ;
- сложность задач в таких сценариях по нескольким серверам или в "облаке";
- сложность или невозможность возобновления сценария после ошибки.

Поэтому для решения этих проблем используется другой подход (декларативный):

- сценарии пишутся на специализированных декларативных языках, которые описывают сценарий как направленный граф из связанных задач и результатов их выполнения. Для каждой задачи определяются: программа для выполнения, ее входные и выходные данные и параметры;
- специализированные программы/платформы (workflow management systems) обеспечивают выполнение таких сценариев, как на локальном компьютере пользователя, так и в "облаке" или компьютерном кластере, обеспечивая эффективное распараллеливание выполнения задач.

Ниже кратко рассмотрим наиболее часто используемые языки для декларативного определения сценариев и компьютерные программы/платформы для их выполнения.

1.4.3.1 CWL

CWL — Common Workflow Language (Crusoe et al., 2022; <https://www.commonwl.org/>) – это открытый стандарт для описания сценариев анализа и запуска различных инструментов. Он специально разработан с возможностью использования в различных программных и аппаратных окружениях, как на персональных компьютерах, так и кластерах и в облачных сервисах. CWL используется в различных областях, таких как биоинформатика, астрономия, физика высоких энергий и машинное обучение.

CWL описывает команды выполняемые в командной строке, набор параметров для этих команд, а также последовательности выполнения этих команд.

CWL основан на текстовых форматах YAML (YAML Ain't Markup Language, <https://yaml.org/>) и json (<https://www.json.org/>), которые не являются языками разметки и разработаны так, чтобы быть более лаконичными и легко читаемыми для человека чем XML. В основе стандарта лежит набор пар ключ-значение. Для более сложных случаев используются иерархические структуры, в которых значениями для ключей могут быть другие наборы ключ-значение.

Основными ключами простого CWL-документа являются:

- class — специфицирует описана ли здесь одна команда (CommandLineTool) или сценарий (Workflow);
- baseCommand — имя выполняемой команды;
- inputs — набор параметров команды. Для каждого параметра устанавливается набор его атрибутов: тип (числовой, строка, булев), позиция, префикс и т.д.;
- outputs — набор файлов из числа сгенерированных выполненной командой, которые должны рассматриваться как результат исполнения CWL-файла и копируются в результирующую директорию. Могут задаваться с помощью маски (например, *.txt).
- stdout — имя файла для записи лога выполняемой команды

Простейший CWL-файла описывает одну программу и набор параметров которые должны быть заданы для нее. На рисунке 1.4.2 описано выполнение команды

```
tar - extract -file <FILE>
```

с последующим копированием файла hello.txt в результирующую папку. Путь до архива <FILE> должен быть задан отдельно.

```
#!/usr/bin/env cwl-runner

cwlVersion: v1.0
class: CommandLineTool
baseCommand: [tar, --extract]
inputs:
  tarfile:
    type: File
    inputBinding:
      prefix: --file
outputs:
  example_out:
    type: File
    outputBinding:
      glob: hello.txt
```

Рисунок 1.4.2 – Пример CWL-файла, описывающего одну программу – распаковку архива командой tar с обязательным аргументом –extract, одним параметром пути к архиву и одним результатом выполнения – файлом hello.txt, который должен содержаться в исходном архиве

CWL также позволяет описывать сценарии последовательного выполнения нескольких команд, которые используют в качестве параметров результаты выполнения предыдущих команд. Такой сценарий имеет единый список входных параметров, значения которых задаются в отдельном файле также как и для единичной команды, описанной в CWL. Язык также позволяет создавать вложенные сценарии.

Ключи, используемые для спецификации сценария:

- `steps` — содержит все шаги сценария;
- `run` — ссылается на CWL-файл, который описывает отдельного шаг сценария;
- `in` — задает параметры отдельного шага сценария. Значениями могут быть ссылки на глобальные параметры данного сценария;
- `out` — результат отдельного шага.

Детали того как именно будет выполнена команда или сценарий должны быть описаны внутри CWL-файла. Таким образом если пользователь хочет, например, чтобы сценарий был выполнен в конкретном докер-контейнере, эти настройки необходимо указывать в CWL-файле.

Пример сценария представлен на рисунке 1.4.3 В данном сценарии на первом шаге происходит распаковка архива, а во втором компиляция файла полученного из архива с помощью `compile.cwl`. Входными параметрами сценария является архив и имя файла который должен быть извлечен. Результатом является скомпилированный файл.

Эталоном реализации стандарта CWL служит программа `cwl-tool` (<https://github.com/common-workflow-language/cwltool>). С ее помощью CWL-файл может быть интерпретирован в команду для командной строки с учетом значений всех параметров из отдельного файла, после чего созданная команда выполняется. `cwl-tool` выполняется через командную строку следующим образом:

```
cwl-runner cwl.yml job.yml
```

Здесь `cwl.yml` – это CWL-файл, а `job.yml` – файл с значениями параметров.

В БД <https://view.commonwl.org/> собрано около 3000 различных сценариев на языке CWL. Список программных продуктов, поддерживающих стандарт CWL приведен на странице проекта (<https://www.commonwl.org/#Implementations>).

```
#!/usr/bin/env cwl-runner

cwlVersion: v1.0
class: Workflow
inputs:
  tarball: File
  name_of_file_to_extract: string

outputs:
  compiled_class:
    type: File
    outputSource: compile/classfile

steps:
  untar:
    run: tar-param.cwl
    in:
      tarfile: tarball
      extractfile: name_of_file_to_extract
    out: [extracted_file]

  compile:
    run: arguments.cwl
    in:
      src: untar/extracted_file
    out: [classfile]
```

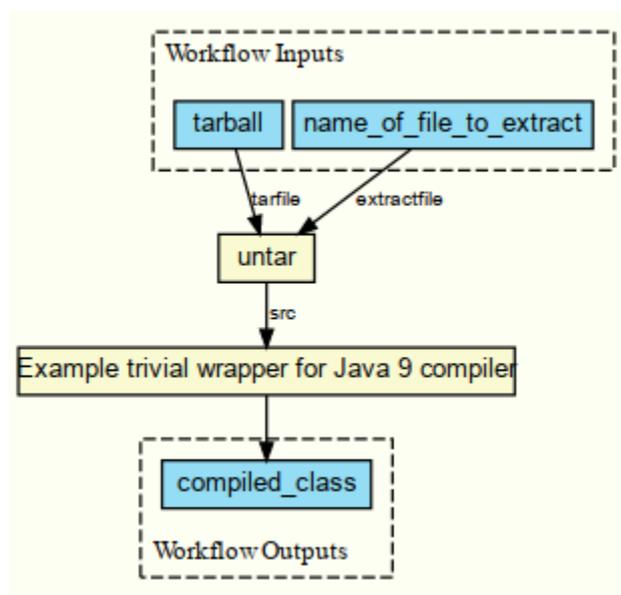


Рисунок 1.4.3 – а) пример сценария, описанного с помощью CWL;
б) представление этого сценария в виде графа

1.4.3.2 WDL

WDL (Workflow Description Language, <https://openwdl.org/>) – это специализированный, человекочитаемый язык для описания сценариев обработки данных. Он позволяет описать программы, доступные как docker образы, как задачи WDL, объединять их сценарии, в рабочие процессы и распараллеливать их выполнение. Пример такого сценария приведен на рисунке 1.4.4. Полная спецификация языка WDL приведена по адресу: <https://github.com/openwdl/wdl/blob/main/versions/1.0/SPEC.md>

Для выполнения сценариев на языке WDL используется программы:

- Cromwell (<https://github.com/broadinstitute/cromwell>)
- miniwdl (<https://github.com/chanzuckerberg/miniwdl>)
- dxCompiler (<https://github.com/dnanexus/dxCompiler>).

```

task hello {
  input {
    String pattern
    File in
  }

  command {
    egrep '${pattern}' '${in}'
  }

  runtime {
    docker: "broadinstitute/my_image"
  }

  output {
    Array[String] matches = read_lines(stdout())
  }
}

workflow wf {
  call hello
}

```

- task – выполняемая задача под названием «hello»
- input – входные параметры для задачи (здесь шаблон строки и входящий файл)
- command – описание формирование командной строки для выполнения соответствующей задачи;
- runtime – позволяет задать контейнер, содержащий соответствующую программу для выполнения;
- output – выходные параметры задачи и способ их получения;
- workflow – собственно сам сценарий анализа данных. Здесь он состоит из одной задачи «hello».

Рисунок 1.4.4 – Пример сценария на языке WDL

1.4.3.3 Nextflow

Nextflow (Di Tommaso et al., 2017; <https://www.nextflow.io>) – среда и язык для создания и выполнения сценариев, управляемых данными. Nextflow обеспечивает масштабируемые и воспроизводимые сценарии анализа данных с использованием контейнеров.

Nextflow имеет свой предметно-ориентированный язык для написания сценариев – Nextflow DSL (Domain Specific Language), который упрощает реализацию и развертывание сложных параллельных и реактивных сценариев в облаках и кластерах. Nextflow DSL – это расширение языка программирования Groovy, который, в свою очередь, является надмножеством языка программирования Java. Groovy можно рассматривать как Python для Java, поскольку он упрощает написание кода и более доступен.

На практике скрипт Nextflow создается путем объединения различных процессов. Каждый процесс может быть написан на любом скриптовом языке, который может выполняться платформой Linux (Bash, Perl, Ruby, Python и т. д.).

Процессы выполняются независимо и изолированы друг от друга, то есть они не имеют общего (доступного для записи) состояния. Единственный способ связи – через асинхронные очереди, которые в Nextflow называются каналами. Любой процесс может определять один или несколько каналов как входных так и выходных. Взаимодействие между этими процессами и, в

конечном итоге, сам поток выполнения конвейера неявно определяется этими объявлениями ввода и вывода.

Пример сценария на языке Nextflow приведен на рисунке 1.4.5. В этом примере определены два процесса. Порядок их выполнения не определяется тем фактом, что процесс `blastSearch` предшествует `extractTopHits`. Вместо этого, поскольку первый процесс определяет канал `top_hits_ch` в своих выходных объявлениях, а процесс `extractTopHits` определяет канал в своем входном объявлении, устанавливается связь. Такое связывание через каналы означает, что `extractTopHits` ожидает вывода `blastSearch`, а затем запускается реактивно, когда канал имеет содержимое.

```
// Declare syntax version
nextflow.enable.dsl=2
// Script parameters
params.query = "/some/data/sample.fa"
params.db = "/some/path/pdb"

process blastSearch {
  input:
    path query
    path db
  output:
    path "top_hits.txt"

  """
  blastp -db $db -query $query -outfmt 6 > blast_result
  cat blast_result | head -n 10 | cut -f 2 > top_hits.txt
  """
}

process extractTopHits {
  input:
    path top_hits

  output:
    path "sequences.txt"

  """
  blastdbcmd -db $db -entry_batch $top_hits > sequences.txt
  """
}

workflow {
  def query_ch = Channel.fromPath(params.query)
  blastSearch(query_ch, params.db) | extractTopHits | view
}
```

Рисунок 1.4.5 – Пример сценария на языке Nextflow

Nextflow также предоставляет среду для выполнения сценариев, которые могут выполняться как локально, так и на компьютерном кластере или в "облаке". Для этого

поддерживаются различные планировщики задач (например, Linux SLURM, PBS Works, Torque, HTCondor) и облачные платформы (Amazon Web Services, Google Cloud Platform, Kubernetes). Среда Nextflow предоставляет информацию для отслеживания и визуализации исполнения сценариев в нескольких видах: текстовый лог выполнения, HTML отчет выполнения, трассировка выполнения, визуализация графа исполнения, хронология выполнения, HTTP WebLog.

1.4.2.4 Galaxy

Galaxy (Galaxy Community, 2022; <https://galaxyproject.org>) – это биоинформатическая платформа с открытым исходным кодом на языке Python для научных сценариев, интеграции, хранения и публикации данных. Ее цель – сделать вычислительную биологию доступной для ученых-исследователей, не имеющих опыта компьютерного программирования или системного администрирования. Хотя изначально Galaxy был разработан для исследований в области геномики, он в значительной степени не зависит от предметной области и теперь используется в качестве общей системы сценариев в биоинформатике.

Интерфейс Galaxy (рисунок 1.4.6) разделен на три панели, на которых отображаются доступные анализы, главное окно для запуска анализов и история наборов данных.

Анализы сгруппированы по категориям, и их можно искать. Когда анализ выбран, он отображается в главном окне, где задаются его входные данные и параметры, и анализ запускается. Когда пользователь запускает анализ, его выходные наборы данных добавляются на панель истории.

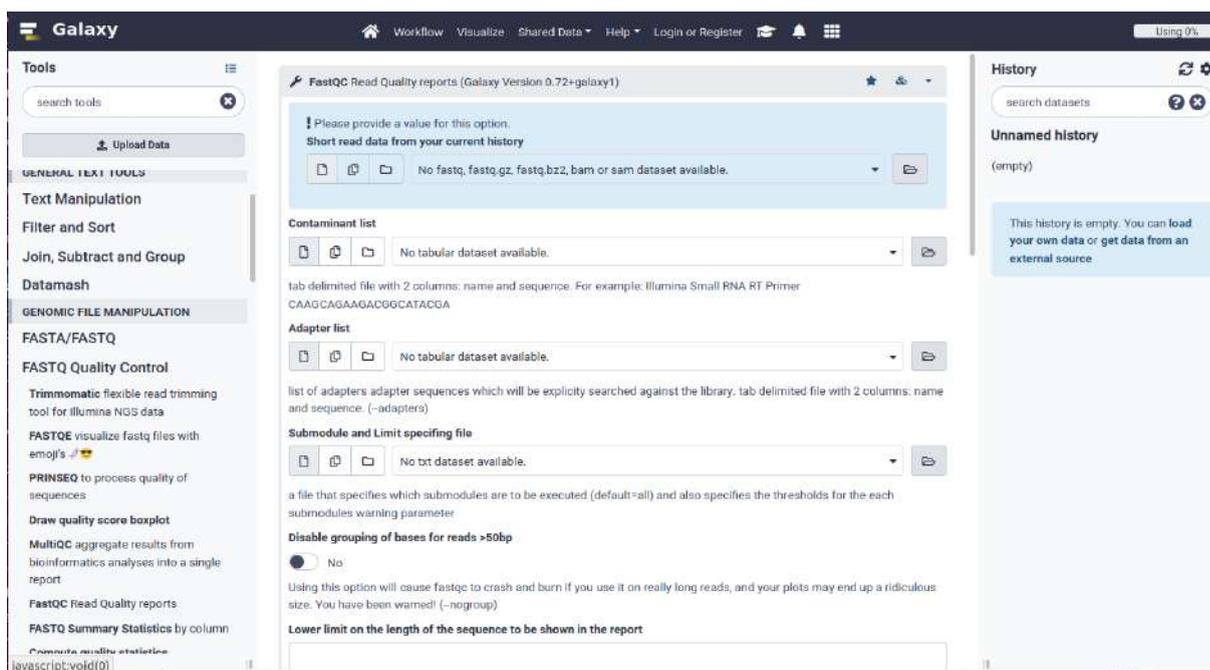


Рисунок 1.4.6 – Интерфейс Galaxy для запуска программы для анализа данных

При помощи встроенного графического редактора (рисунок 1.4.7) анализы можно добавлять в сценарий, а также связывать друг с другом, так что выходные данные одного анализа становятся входными данными других анализов.

Сценарий задает параметры анализов, а также поддерживаются общие действия, такие как сокрытие промежуточных результатов, изменение форматов данных и переименование данных. Сценарии Galaxy позволяют автоматизировать и многократно выполнять большие комплексные анализы. После создания, сценарии могут использоваться как обычные анализы: к ним можно получить доступ и запустить из основного интерфейса Galaxy.

Galaxy ToolShed – ресурс для распространения galaxy анализов и сценариев. Этот сервер функционирует как «Магазин приложений» для серверов Galaxy, где разработчики и администраторы Galaxy могут размещать, совместно использовать и устанавливать galaxy анализы и сценарии. На момент написания он содержит более 8000 анализов.

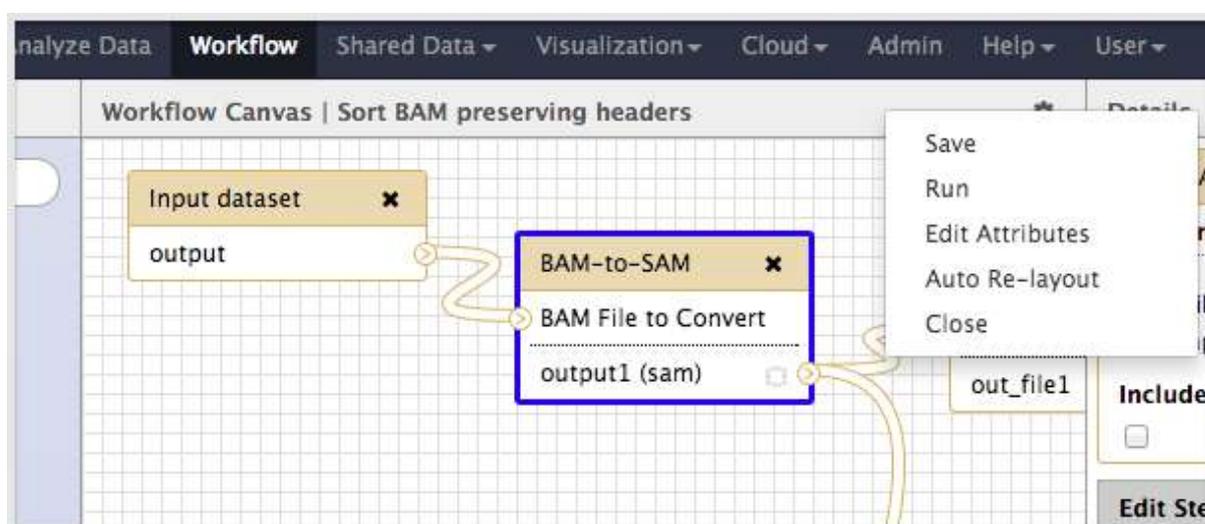


Рисунок 1.4.7 – Графическое редактирование сценария в платформе Galaxy

1.4.4 Jupyter

Помимо описанного выше воспроизводимого способа анализа биомедицинских данных при помощи сценариев, активно используется интерактивный анализ данных при помощи Jupyter блокнотов.

Jupyter Notebook (<https://jupyter.org/>) — это веб-приложение с открытым исходным кодом, которое позволяет создавать и совместно использовать документы, состоящие из различных типов ячеек (cell): программный код, графическое представление данных и описательный текст (рисунок 1.4.8). Jupyter Notebook поддерживает использование множества языков программирования. Само его название является сокращением от 3 языков программирования: **Julia**, **Python** и **R**. Для этого используется понятие ядро (kernel) – для каждого ноутбука на сервере запускается соответствующее ядро, которое собственно и выполняет анализ и визуализацию данных на соответствующем языке программирования,

взаимодействуя с веб-интерфейсом (блокнотом) через заданный API и передачу данных в JSON формате. Разработаны ядра для более чем 20-ти языков программирования: <https://gist.github.com/chronitis/682c4e0d9f663e85e3d87e97cd7d162>

a) We explore the Lorenz system of differential equations:

$$\begin{aligned}\dot{x} &= \sigma(y - x) \\ \dot{y} &= \rho x - y - xz \\ \dot{z} &= -\beta z + xy\end{aligned}$$

Let's change (σ, β, ρ) with ipywidgets and examine the trajectories.

б) [4]: `w=interactive(solve_lorenz,sigma=(0.0,50.0),rho=(0.0,50.0))`
w

в) [4]:

sigma	<input type="range" value="10.00"/>	10.00
beta	<input type="range" value="2.63"/>	2.63
rho	<input type="range" value="28.00"/>	28.00

г)

Рисунок 1.4.8 – Интерфейс Jupyter блокнота позволяет использовать различные типы ячеек (cell): а) текстовое описание и документация в формате Markdown; б) код на различных языках программирования; в) графические виджеты для настройки параметров вычислений; г) графическое представление результатов расчета

JupyterHub (<https://jupyter.org/hub>) — это многопользовательское приложение которое устанавливается на сервере или в "облаке". Он обеспечивает авторизацию и управляет отдельной средой Jupyter для каждого пользователя. Поэтому его удобно использовать в компьютерном классе для студентов или научно-исследовательской группе.

Binder (<https://mybinder.org/>) – позволяет создать документ Jupyter Notebook из заданного Git репозитория. Для этого в репозитории располагается специальный файл, который описывает как создать Docker образ из этого репозитория, используя `jupyterrepo2docker`. После этого собранный Docker образ разворачивается и запускается на одном из серверов, предоставляемых проектом Binder, и пользователь может работать с соответствующим Jupyter Notebook в интерактивном режиме. На создание контейнера уходит некоторое время, а после периода бездействия контейнер удаляется. Изменения, вносимые студентами в контейнерную версию, не затрагивают оригинальный репозиторий и могут рассматриваться как временные.

1.4.5 SoS

SoS – Script of Scripts (<https://vatlab.github.io/sos-docs/>) – эта среда позволяет объединить анализ биомедицинских данных как при помощи сценариев, так и в интерактивном режиме. С технической точки зрения (рисунок 1.4.9), SoS это специальное ядро (kernel), которое позволяет:

- выполнять интерактивный анализ данных, передавая команды другим ядрам для разных языков программирования (R, Python, Julia, MATLAB и т.д.);
- выполнять сценарии анализа данных, в том числе и в пакетном режиме, используя собственную систему управления сценариями, которая позволяет использовать Docker образы с различными программами анализа данных, а также распараллелить запуск этих Docker образов по компьютерному кластеру или "облаку", используя различные менеджеры управления задачами.

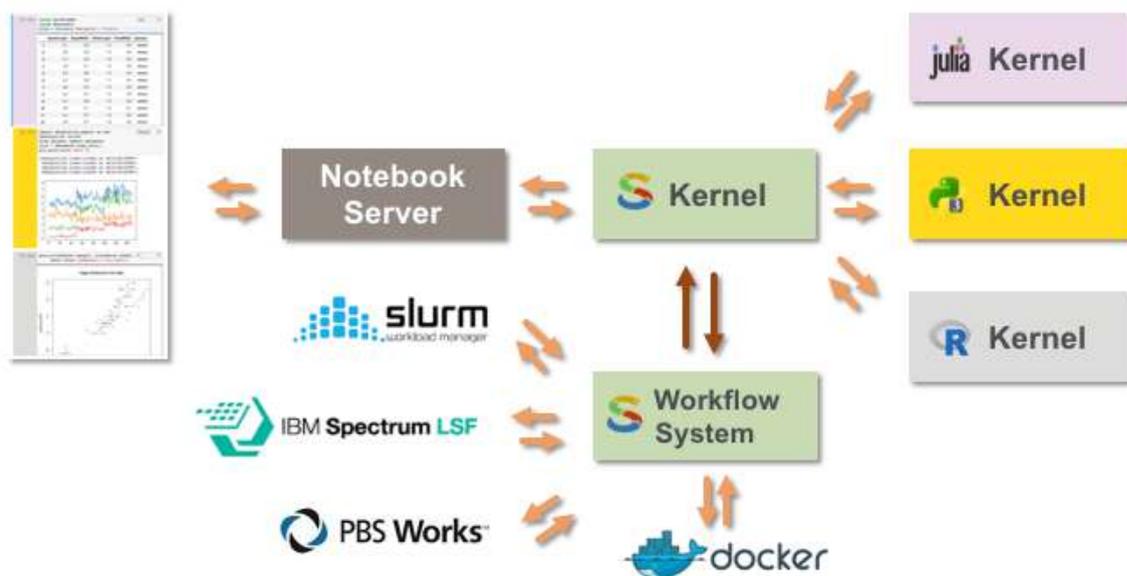


Рисунок 1.4.9 – Архитектура SoS (<https://vatlab.github.io/sos-docs/>)

1.4.6 R, Bioconductor, RStudio

R (<https://www.r-project.org/>) — язык программирования для статистической обработки данных и работы с графикой, а также свободная программная среда вычислений с открытым

исходным кодом в рамках проекта GNU. Он широко используется как статистическое программное обеспечение для анализа данных и фактически стал стандартом для статистических программ.

CRAN – The Comprehensive R Archive Network (<https://cran.r-project.org/web/packages/>) – это сеть ftp и веб-серверов по всему миру, на которых хранятся идентичные актуальные версии кода и документации для R. На данный момент она содержит более 19 000 пакетов (<https://cran.r-project.org/web/packages/>).

Bioconductor (Huber et al., 2015; <https://www.bioconductor.org/>) – это проект, направленный на разработку, создание и распространение бесплатного программного обеспечения с открытым исходным кодом для воспроизводимого анализа биомедицинских данных, в первую очередь, омиксных данных. Он включает более 1000 пакетов (<https://www.bioconductor.org/packages/release/bioc/>).

RStudio (<https://posit.co/>) – наиболее популярная IDE для работы с R. Она доступна в 2 вариантах: RStudio Desktop, в которой программа выполняется на локальной машине как обычное приложение; и RStudio Server, в которой предоставляется доступ через браузер к RStudio установленной на удаленном Linux-сервере.

1.5 Графическое представление биологических данных

Рассмотренные выше программы для статистического анализа данных обеспечивают широкий набор методов для графического представления полученных результатов: графики, диаграммы, тепловые карты и т.п. Однако, при анализе генетических данных необходимы и специализированные способы представления таких данных. Основными из них являются:

- графы (диаграммы) метаболических путей и пути передачи сигнала в клетке. Одной из важных задач при графическом представлении таких графов – автоматическое упорядочивание расположение вершин и ребер графа на плоскости (graph layout). На такие диаграммы также могут быть нанесены экспериментальные данные: цветом или контуром у вершин может быть отображена экспрессия соответствующих генов или количество метаболитов, толщина и цвет ребер графа могут отражать скорость соответствующих метаболических реакций;
- геномный браузер – позволяет представить в графическом виде геномные последовательности, их аннотацию, функциональные сайты, а также результаты высокопроизводительного секвенирования, например, выравнивание прочтений на референсный геном, плотность таких прочтений вдоль геномов;
- визуализация выравниваний и филогенетические деревья также являются необходимым компонентом для графического представления эволюционного анализа генетических данных;

- визуализация 2-х мерных структур химических веществ (структурных формул);
- визуализация 3-х мерных структур – необходима для графического представления пространственной структуры: результаты рентгеноструктурного анализа макромолекул, докинг малых молекул в активные центры ферментов и т.п.

Ниже будут кратко рассмотрены эти способы представления результатов анализа биологических данных и примеры соответствующих программ. В данной работе не рассматривается графическое представление и анализа ряда медицинских данных, в первую очередь связанных с 3-мерными структурами органов: результаты рентгенографии, магнитно-резонансной томографии, ультразвуковые исследования и т.п.

1.5.1 Диаграммы метаболических путей и пути передачи сигнала в клетке

KEGG Pathways (Kanehisa et al., 2021; <https://www.genome.jp/kegg/pathway.html>) является основной БД по метаболическим путям для многих видов организмов. На рисунке 1.5.1 приведена диаграмма цикла Кребса. Как пример нанесения на диаграмму дополнительных данных красным цветом подсвечены собственно метаболиты и реакции, входящие в этот цикл.

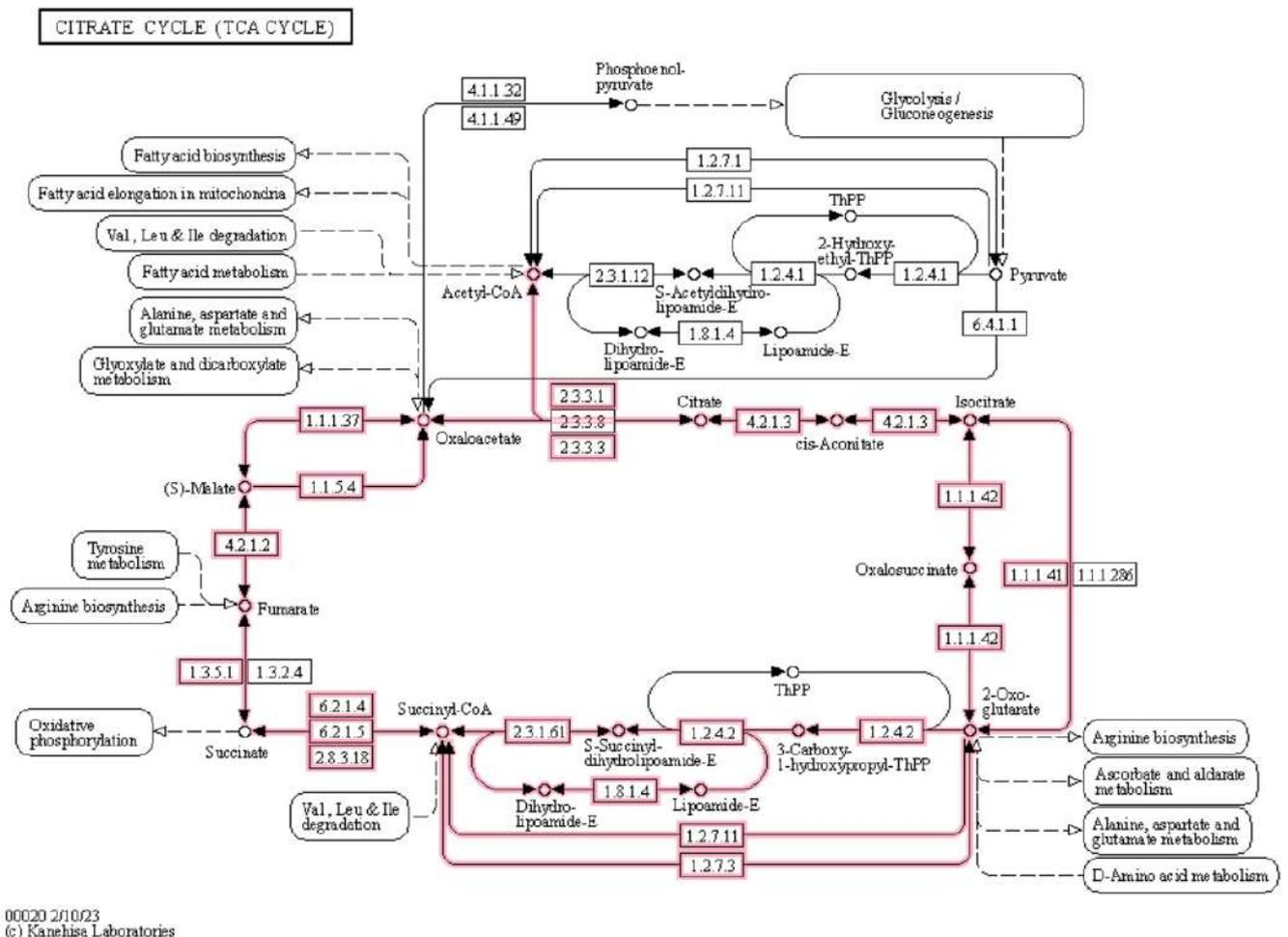


Рисунок 1.5.1 – Диаграмма цикла Кребса из БД KEGG Pathways. Красным подсвечены собственно метаболиты и реакции, входящие в цикл Кребса

Reactome (Fabregat et al., 2018; <https://reactome.org/>) является одной из основных БД по передаче сигнала в клетке. Пример такой диаграммы представлена на рисунке 1.5.2, используется графическая нотация SBGN Process Diagram, описанная выше.

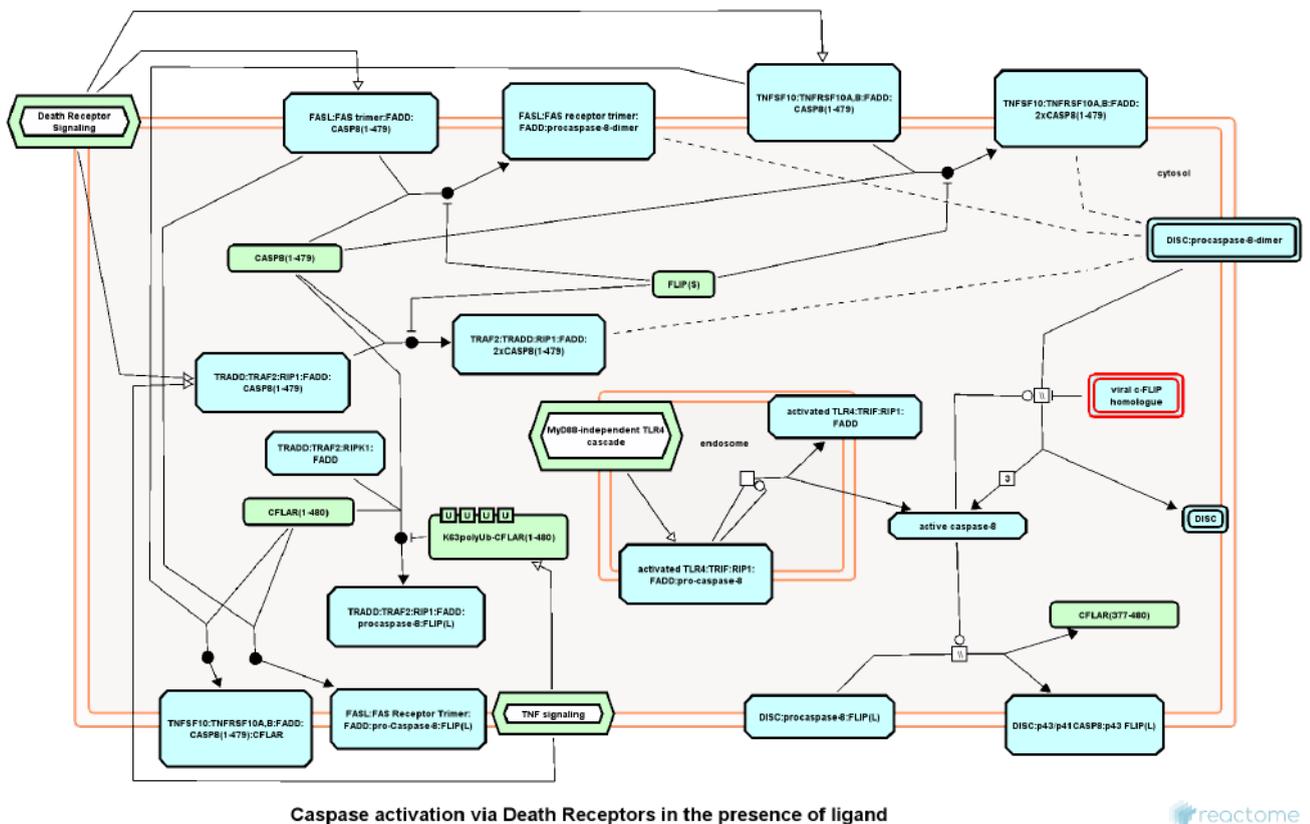


Рисунок 1.5.2 – Пример диаграммы пути передачи сигнала в клетке из БД Reactome. Используется графическая нотация SBGN Process Diagram

1.5.2 Геномный браузер

Геномный браузер – это программное решение, позволяющее визуализировать имеющиеся геномы и их аннотации в различном масштабе. Многие геномные браузеры поддерживают изменение масштаба от целого кариотипа до конкретного участка последовательности генома в текстовом виде. Аннотированные элементы отображаются в виде полос (треков) относительно референсного генома. Визуализация трекв может настраиваться пользователем под его нужды в зависимости от функциональных возможностей конкретного геномного браузера. Геномные браузеры используются как для визуализации общих референсных геномов и доступных для них аннотационных трекв, так и для визуализации пользователем собственных данных по геномной аннотации.

Наиболее популярные геномные браузеры:

Ensembl Genome Browser (<https://www.ensembl.org/index.html>) – один из первых в мире геномных браузеров, был разработан в 2002 году. Используется в проекте Ensembl для

визуализации геномной информации для множества геномов позвоночных. Он поддерживает визуализацию геномных данных в области сравнительной геномики, эволюции, вариации последовательностей и регуляции транскрипции (рисунок 1.5.3).

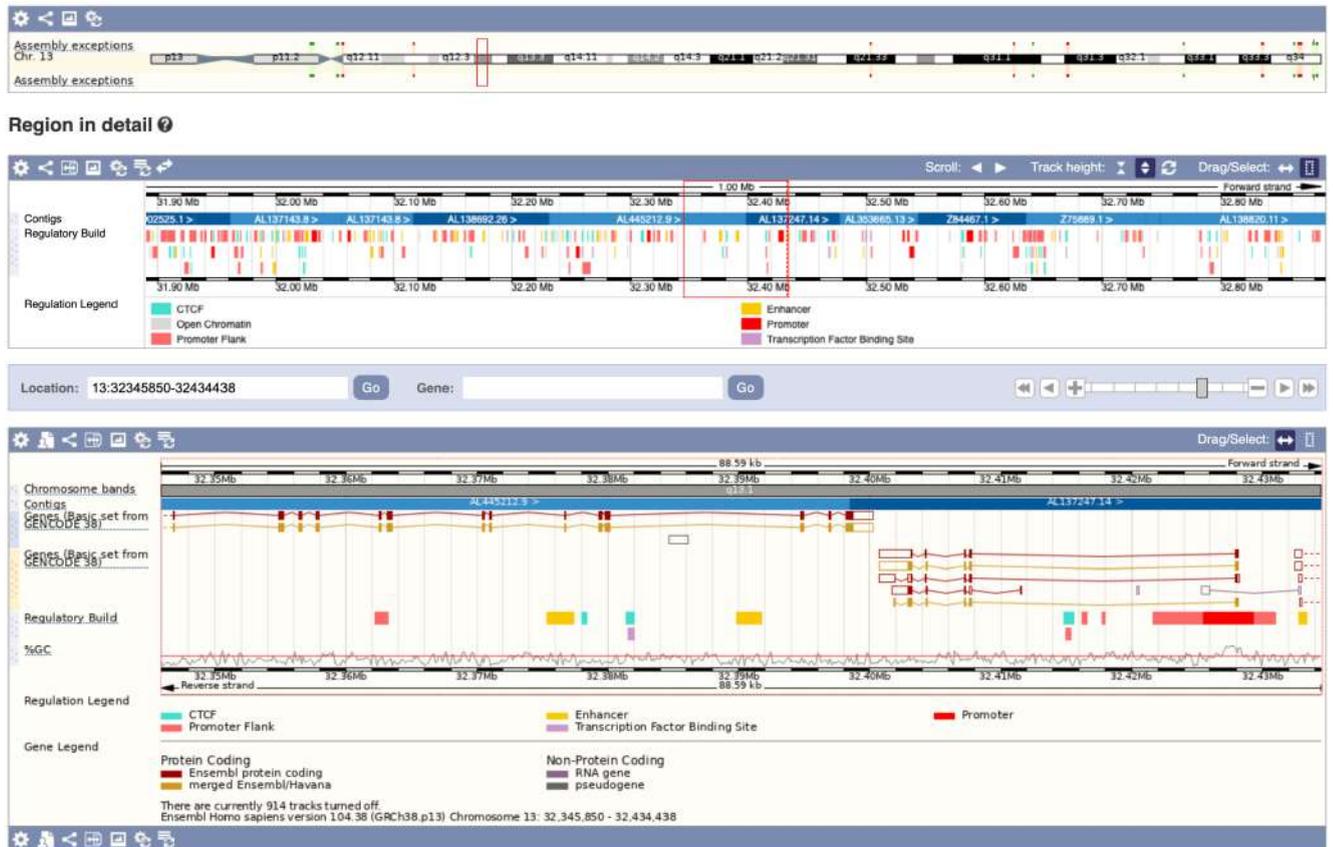


Рисунок 1.5.3 – Пример отображения генетической информации и связанной метаинформации на геномном браузере Ensembl

UCSC Genome browser – University of California Santa Cruz (<http://genome.ucsc.edu/index.html>) – прототип UCSC геномного браузера был выпущен одновременно с первой сборкой генома человека в 2000 году. В то время он был доступен только как общедоступный веб-интерфейс. Браузер постоянно совершенствовался, чтобы включать в себя широкий набор эукариотических организмов, а также новые источники и форматы данных. С 2014 года браузер генома USCS также существует как локальное приложение с широкими возможностями совместного использования.

NCBI Genome Data Viewer (<https://www.ncbi.nlm.nih.gov/genome/gdv>) – был разработан Национальным Центром Биотехнологической Информации США (NCBI) для отображения их внутренней аннотации RefSeq. Он интегрирован с другими ресурсами NCBI, такими как Gene, dbGaP и BLAST, а также связан с компонентой для изучения последовательностей NCBI Sequence Viewer. NCBI Genome Data Viewer доступен пользователям в виде веб-сервера, он

позволяет проводить исследования и анализ как аннотированных NCBI сборок генома эукариот, так и собственных последовательностей пользователя

IGV – Integrative Genomics Viewer (<https://software.broadinstitute.org/software/igv/>) – одна из самых популярных программ для просмотра генома, разработанная Broad Institute. Он поддерживает широкий спектр типов данных, включая выравнивание ридов, геномные аннотации, данные экспрессии, генетические вариации и т.п. С 2018 года существует веб-версия IGV.

JBrowse (<https://jbrowse.org/>) – JBrowse был первым геномным веб-браузером с архитектурой клиент-сервер, основанной на JavaScript, CSS и HTML5. Он входит в проект Generic Model Organism Database (GMOD), крупный вклад в его разработку вносит Evolutionary Software Foundation. Веб-приложение также доступно как отдельное приложение через оболочку Electron (десктопная версия JBrowse).

Как правило, геномные браузеры поддерживает высокую масштабируемость отображаемой геномной информации: система позволяет осуществлять визуализацию просматриваемой информации от уровня всего кариотипа до максимального зума на уровне отдельных нуклеотидов (рисунок 1.5.4).



Рисунок 1.5.4 – Схематичное представление отображения информации на разных уровнях зума в геномном браузере Ensembl

1.5.3 Выравнивания и филогенетические деревья

Визуализация множественного выравнивания родственных нуклеотидных или белковых последовательностей и соответствующих филогенетических деревьев является необходимым компонентом для графического представления эволюционного анализа генетических данных. Поэтому эта функция входит в множество специализированных программ. На рисунке 1.5.5 представлен редактор множественных выравниваний, входящий в отечественный программный пакет UGENE (Okonechnikov et al., 2012; <http://ugene.net>).

Обширный список программ для визуализации множественного выравнивания приведены на сайте: https://en.wikipedia.org/wiki/List_of_alignment_visualization_software.

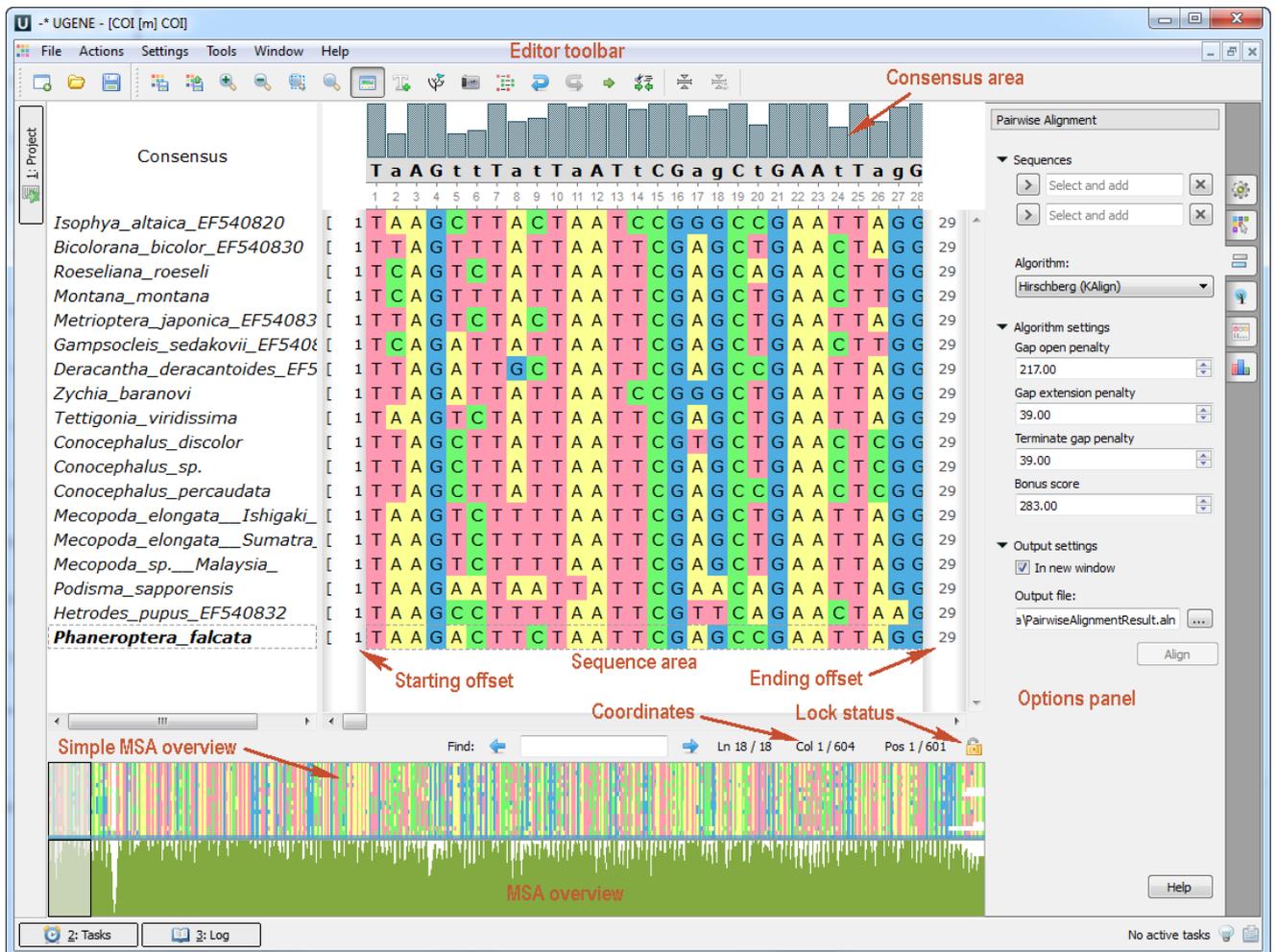


Рисунок 1.5.5 – Просмотр и редактирование множественного выравнивания

последовательностей в программе UGENE

(<https://doc.ugene.net/wiki/display/UM/Alignment+Editor+Components>)

Также существуют специализированные JavaScript библиотеки, которые позволяют создавать специализированные веб-приложения или встраивать соответствующий редактор в другие веб-приложения. Одним из наиболее быстрых и функциональных таких веб-приложений

является AlignmentViewer2.0 (Reguant et al., 2020; <https://github.com/sanderlab/AlignmentViewer2.0>), приведенный на рисунке 1.5.6.

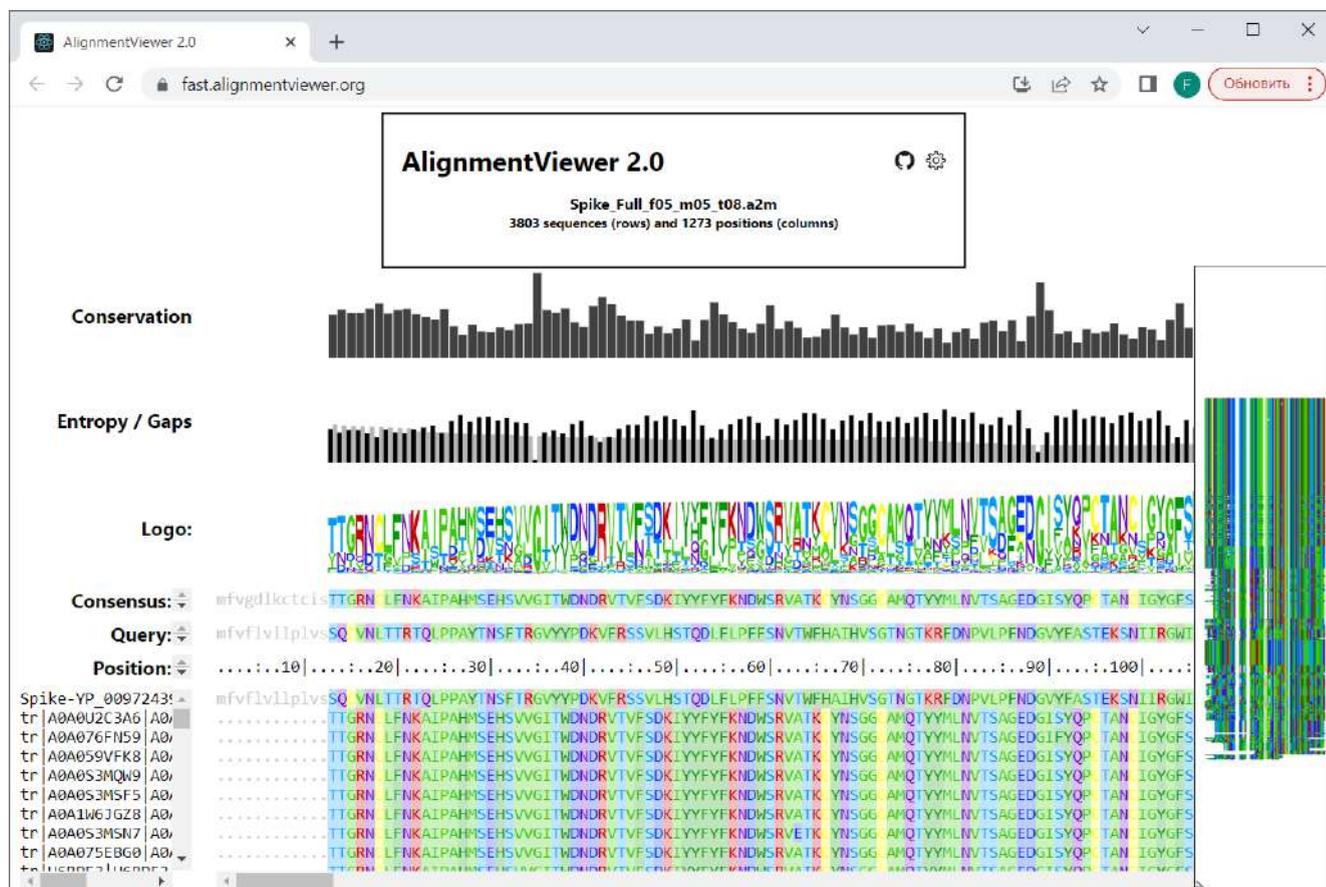


Рисунок 1.5.6 – Просмотр множественного выравнивания последовательностей в веб-приложении AlignmentViewer2.0

Также существует список программ для визуализации филогенетических деревьев: https://en.wikipedia.org/wiki/List_of_phylogenetic_tree_visualization_software

Для визуализации филогенетических деревьев существуют JavaScript библиотеки, которые позволяют создавать специализированные веб-приложения или встраивать соответствующий редактор в другие веб-приложения. На рисунке 1.5.7 приведен пример веб-приложения, которое интегрирует функционал библиотеки phyloTree.js (Shank et al., 2018; <http://phyloTree.hyphy.org>) для просмотра филогенетического дерева и библиотеки PV (Biasini et al., 2015; <https://github.com/biasmv/pv>) для просмотра трехмерной структуры соответствующего белка.

Также существуют специализированные библиотеки для интерактивной визуализации филогенетических деревьев в Jupyter блокнотах. Наиболее часто используемой такой библиотекой является ETE Toolkit (Huerta-Cepas et al., 2010; <http://etetoolkit.org/>).

Описанные выше библиотеки phylotree.js и ETE Toolkit были использованы для создания прототипа компонента "Рабочие пространства" для Национальной Базы Генетической Информации.

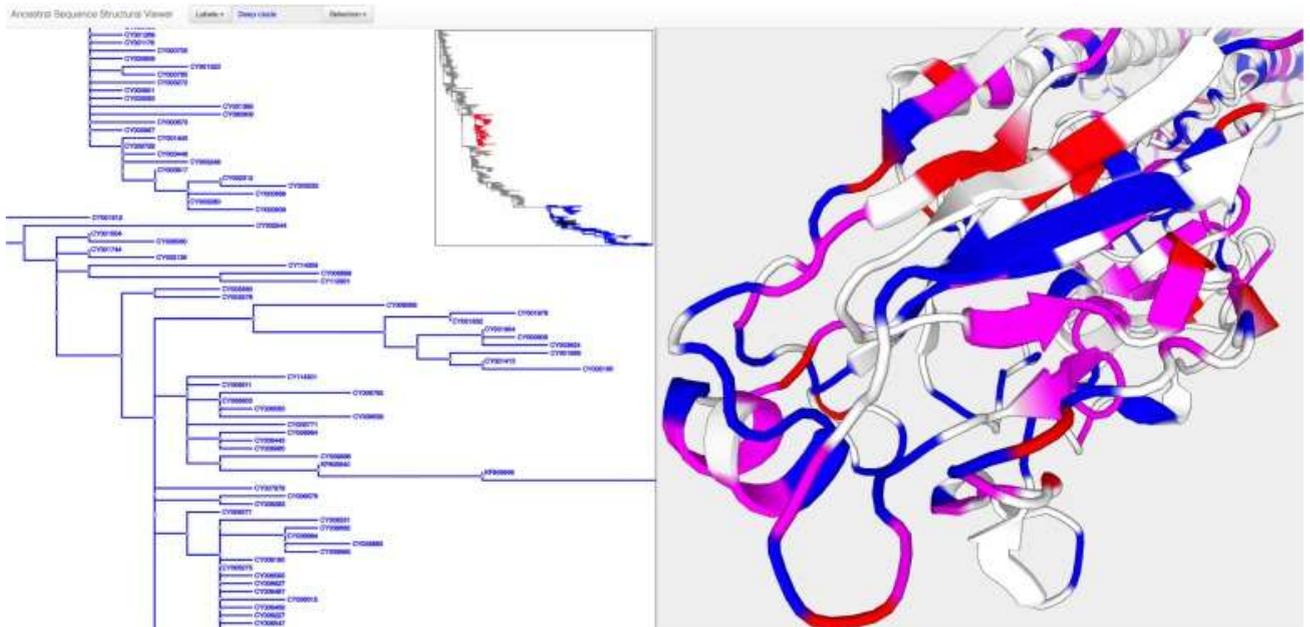


Рисунок 1.5.7 – Интерактивный просмотр филогенетического дерева совместно с наложением различных замен на 3D структуру белка (гемагглютинин вируса гриппа) при помощи библиотек phylotree.js и PV (Shank et al., 2018)

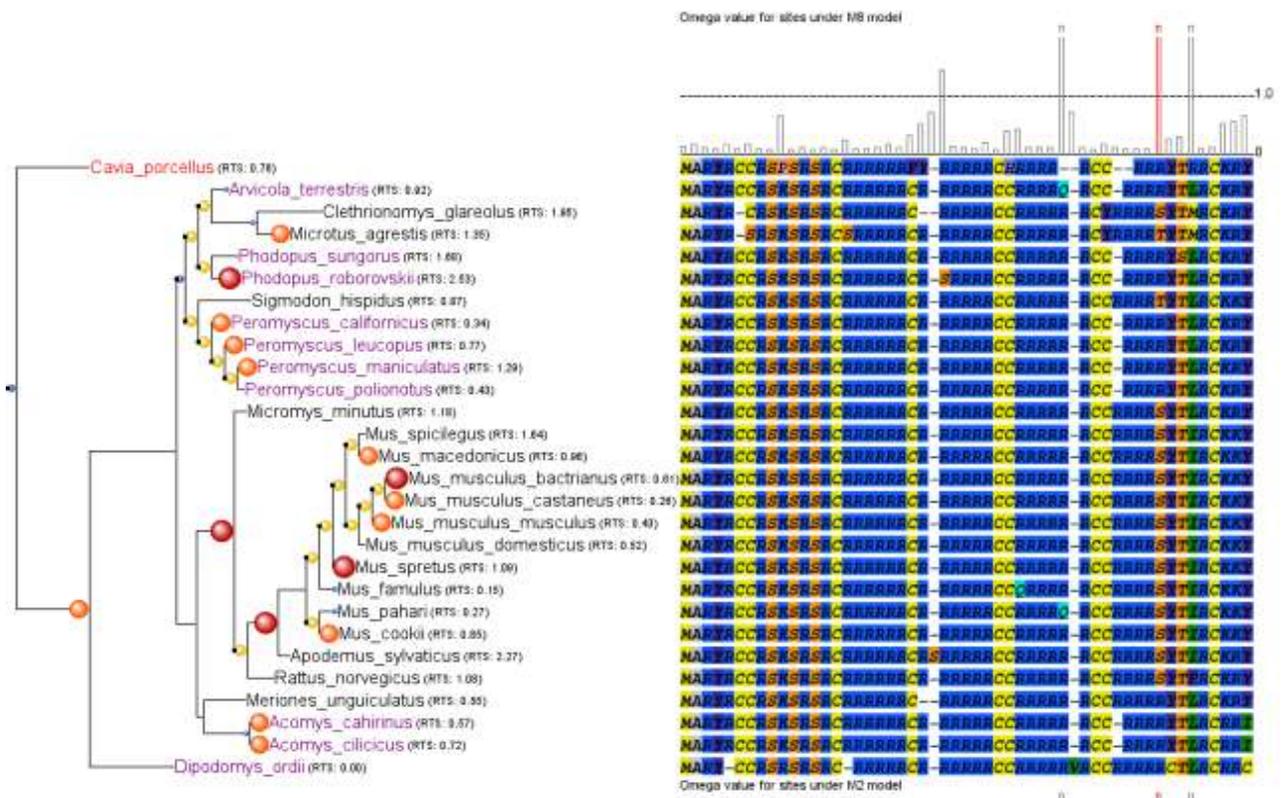


Рисунок 1.5.8 – Совместная визуализация филогенетического дерева и соответствующего множественного выравнивания в ETE Toolkit (<http://etetoolkit.org/static/img/gallery/omega400x400.png>)

1.5.4 Графическое представление 2D структур

Визуализация структурных формул, а также их редактирование является важной задачей для хемо- и биоинформатики. Отметим несколько наиболее популярных Java программ и библиотек для этой задачи, которые были интегрированы в ПК BioUML.

- **CDK** – The Chemistry Development Kit (Steinbeck et al., 2006; <https://cdk.github.io/>) – открытая коллекция Java библиотек для хемо- и биоинформатики.
- **JChemPaint** (<https://jchempaint.github.io/>) – это редактор и средство просмотра 2D-химических структур, разработанный с использованием Chemistry Development Kit (рисунок 1.5.9).

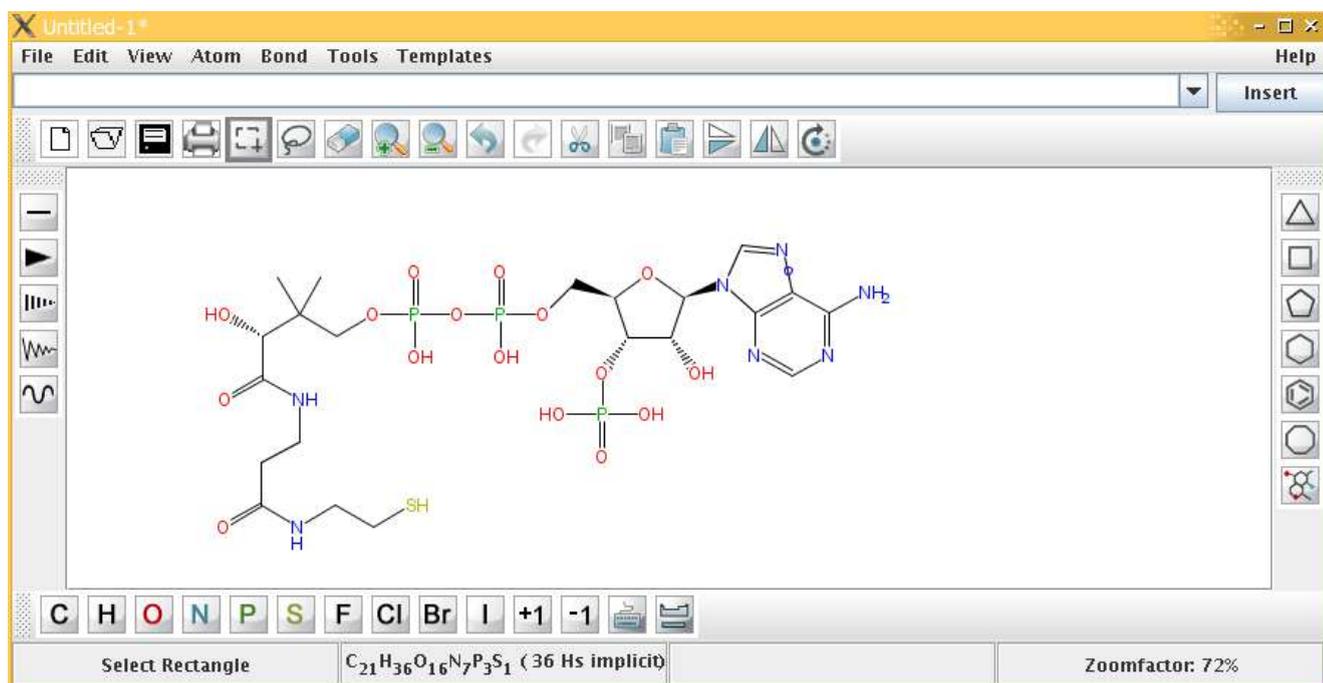


Рисунок 1.5.9 – Интерфейс графического редактора структурных формул JChemPaint (<https://jchempaint.github.io/>)

1.5.5 Графическое представление 3D структур

Визуализация 3-х мерных структур – необходима для графического представления пространственной структуры макромолекул, докинга малых молекул в активные центры ферментов и т.п. Наиболее популярные программы для этих задач:

- **RasMol** (<https://sourceforge.net/projects/openrasmol/>; <http://www.openrasmol.org>) – компьютерная программа, написанная на C, и предназначенная для визуализации молекул, в первую очередь белков и нуклеиновых кислот. Система команд, разработанная для программы ("язык" RasMol'a) используется в других программах, например, в Jmol.

Первая версия программы RasMol была создана Роджером Сэйлом в начале 90-х годов. Последняя версия вышла в 2009 г.

- **JMol** (Hanson, 2010; <https://jmol.sourceforge.net>) — Java программа для просмотра 3-х мерных структур макромолекул. Раньше она часто использовалась внутри веб-браузеров как Java applet. Jmol постепенно вытесняет RasMol, поскольку почти столь же проста в освоении, охватывает всю функциональность RasMol и обладает некоторыми дополнительными возможностями.
- **JSMol** (<https://jmol.sourceforge.net>) — в связи с отменой поддержки Java апплетов в современных веб-браузерах было разработано практически идентичное JMol приложение на JavaScript, которое может запускаться в современных веб-браузерах, используя HTML5.
- **PyMOL** (DeLano, 2002; <https://github.com/schrodinger/pymol-open-source>; <https://pymol.org>) — программа с открытым исходным кодом, поддерживаемая и распространяемая Schrödinger, Inc., для просмотра 3-х мерных структур макромолекул, написанная на Python.
- **VMD** — visual molecular dynamics (Humphrey et al., 1996; Hsin et al., 2008; <http://www.ks.uiuc.edu/Research/vmd/>) – программа для визуализации, анимации и анализа больших биомолекулярных систем с использованием трехмерной графики и встроенных сценариев, написанная на C++.
- **Chimera** (Pettersen et al., 2004; <http://www.cgl.ucsf.edu/chimera/>) — расширяемая программа для интерактивной визуализации и анализа молекулярных структур и связанных данных, включая выравнивание последовательностей.

JMol как библиотека была интегрирована в ПК BioUML для просмотра 3-х мерных структур макромолекул в 2011 г. После отмены поддержки Java апплетов в современных веб-браузерах, в веб-версии ПК BioUML используется JSMol.

1.6 Обзор предшествующих работ автора

В 1994–2000 г.г. во время работы автора в лаборатории теоретической генетики Института Цитологии и Генетики СО РАН под руководством академика РАН, профессора, д.б.н. Колчанова Н.А. был разработан ряд программ и подходов, послуживших заделом для разработки ПК BioUML.

Первой была программа автора, написанная на C++ для ОС Windows, для визуализации ССТФ в промоторных районах генов (рисунок 1.6.1) в ПК FUNSITE (Kel et al., 1995).

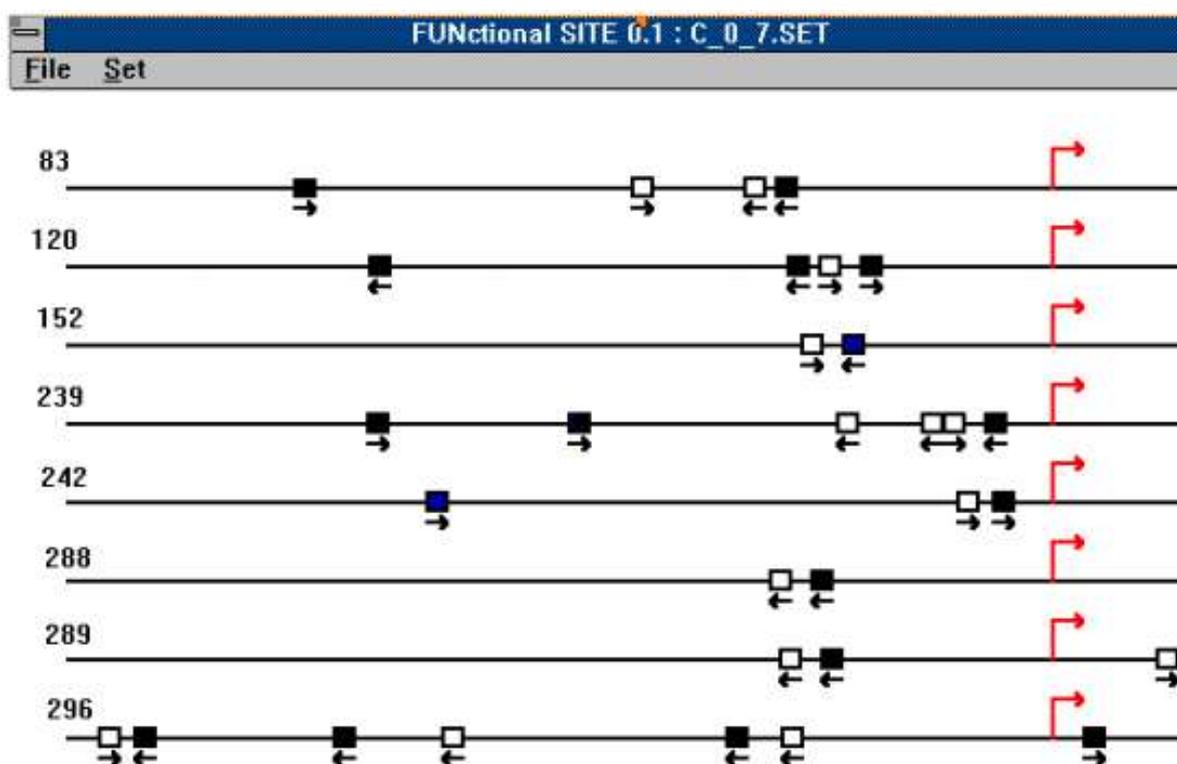


Рисунок 1.6.1 – Визуализация данных в ПК FUNSITE: локализация потенциальных композиционных элементов, образованных ССТФ AP-1 (белый) и RAR (черный) в промоторных последовательностях (от -500 до +100). Гены (номер из EPD), например, 83 – ген островного амилоидного полипептида человека (Kel et al., 1995)

Далее автором была начата разработка компьютерной системы MGL – Molecular Genetics Language – инструмент для составления выборок, графического представления и анализа регуляторных геномных последовательностей (Колпаков и Бабенко, 1997). Одна из ключевых идей MGL, написанного на также на C++ для ОС Windows, была разработка собственного языка и его интерпретатора, для решения указанных выше задач (рисунок 1.6.2). Также в ходе этой работы была создана первая версия C++ библиотеки для графического представления сложных биологических объектов. Интересно отметить некоторое сходство в названиях между MGL и BioUML.

В 1996-1997 г. основные классы MGL были переведены с языка C++ на язык Java и были созданы MGL сервер и клиент (апплет TRRD Viewer) для визуализации регуляторных районов генов из БД TRRD (Heinemeyer et al., 1998, Kolchanov et al., 1999) (рисунок 1.6.3).

Java библиотеки, разработанные для TRRD Viewer, позволили буквально за 3 месяца создать первую версию системы GeneNet, в которой были реализованы новые принципы формального описания и графического представления генных сетей, реализованные в системе GeneNet (Kolpakov et al., 1998, Kolpakov and Ananko, 1999) (рисунок 1.6.4).

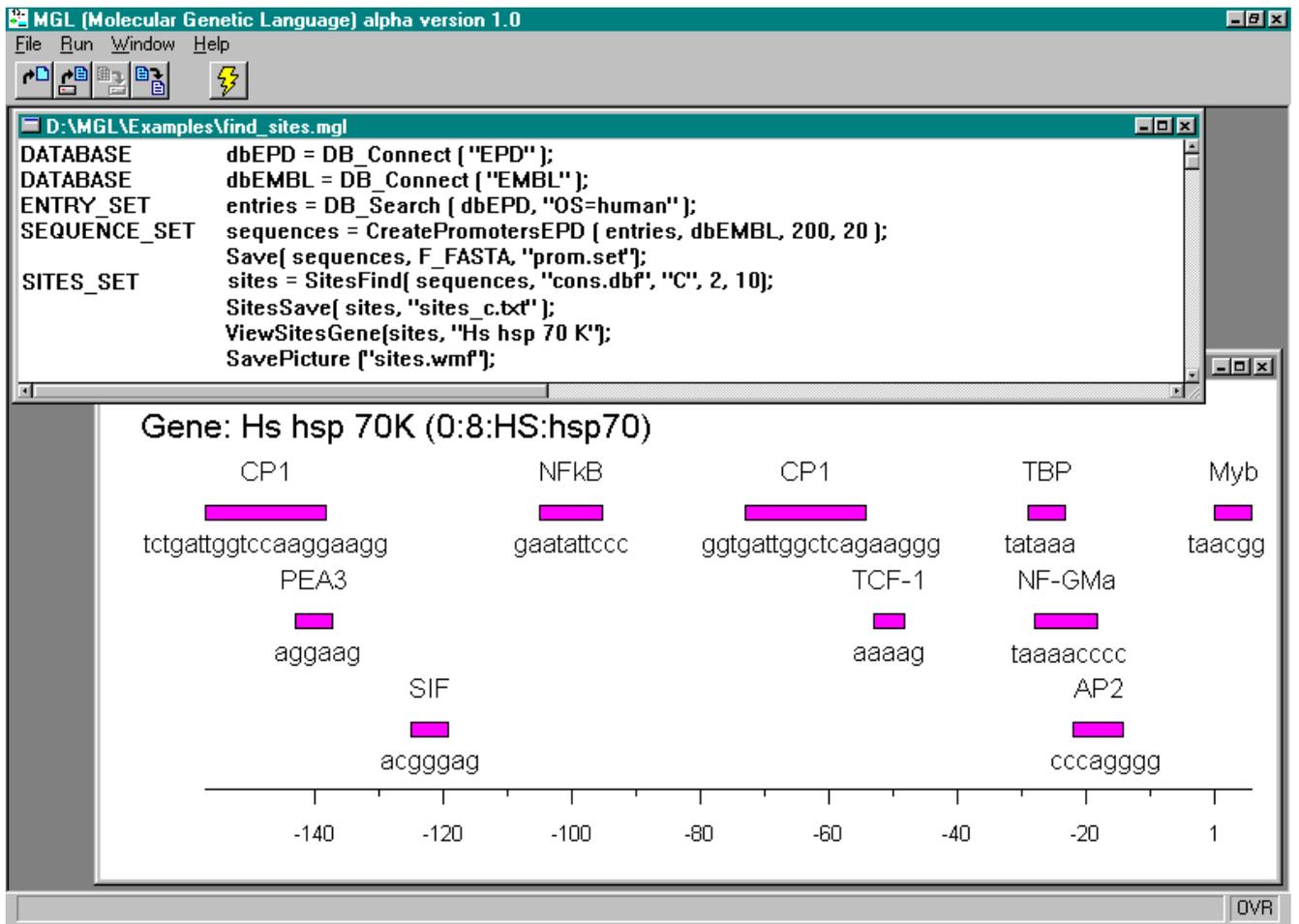


Рисунок 1.6.2 – Графический интерфейс MGL, верхнее окно – MGL скрипт для формирования выборки и графического представления регуляторного района гена hsp70 человека; нижнее окно – графическое представление регуляторных районов этого гена

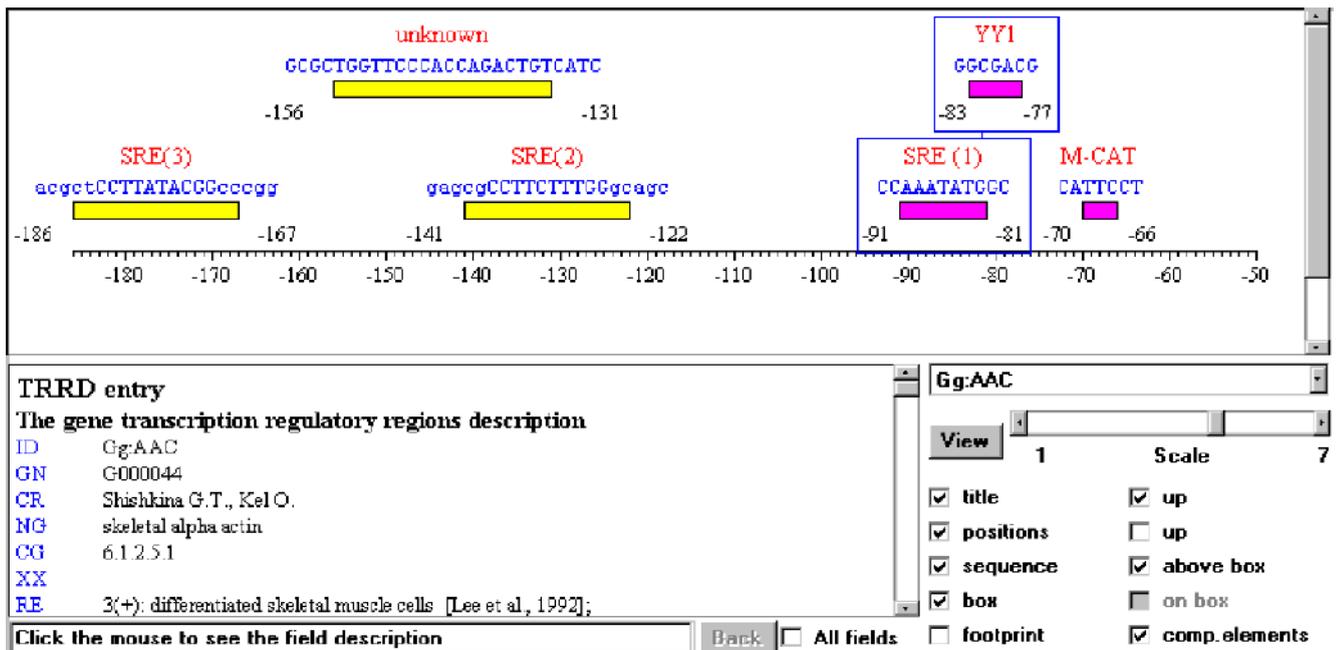


Рисунок 1.6.3 – Апплет TRRD Viewer для визуализации регуляторных районов генов из БД TRRD (Heinemeyer et al., 1998).

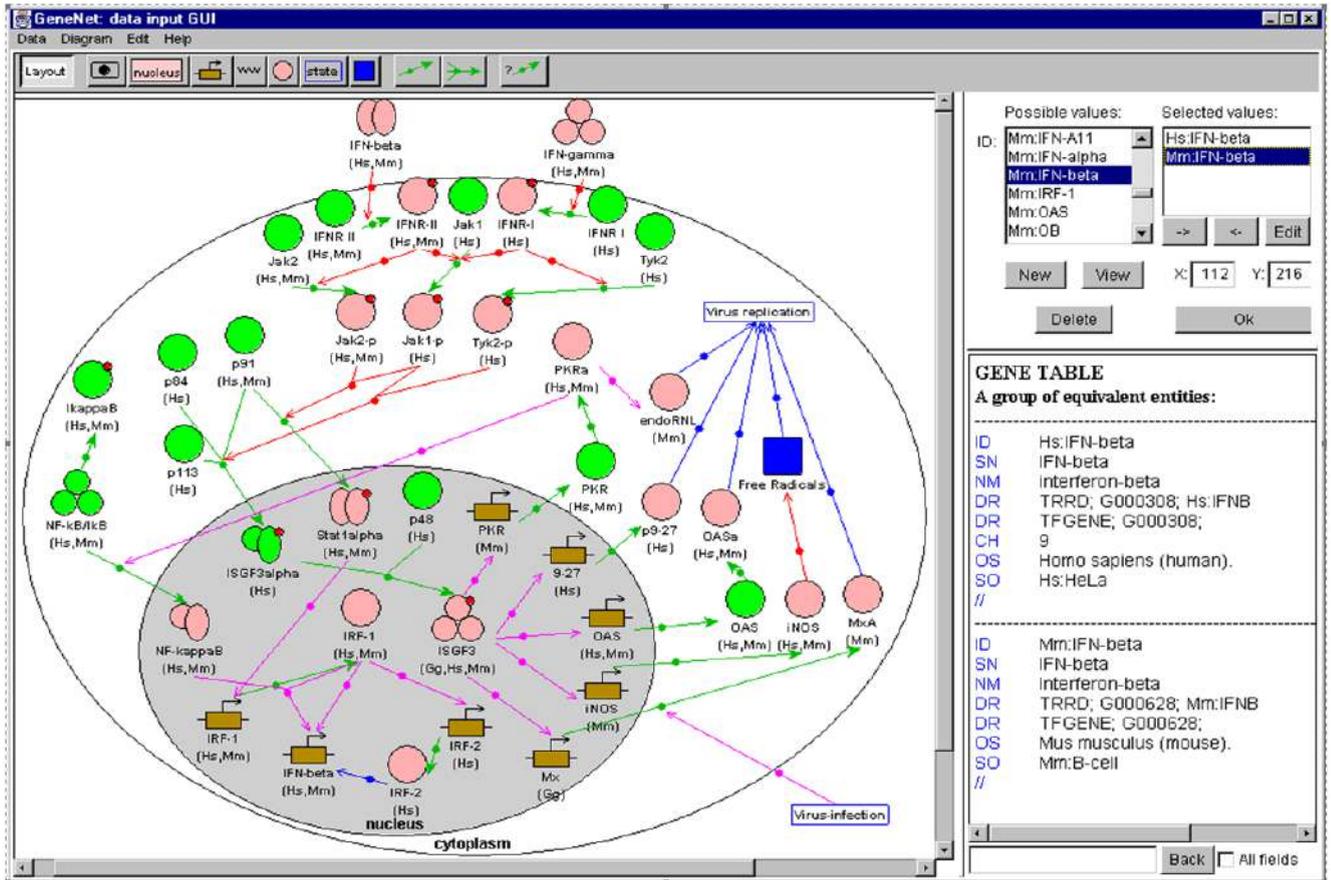


Рисунок 1.6.4 – Java апплет для создания и редактирования диаграмм генных сетей (Kolpakov and Ananko, 1999)

Работа автора руководителем проекта в коммерческой компании WebSci Technologies Inc., США дала ему новый взгляд на разработку программного обеспечения и позволила освоить подходы, используемые для разработки коммерческого программного обеспечения. Как результат, в ПК BioUML появились концепции репозитория и инспектора свойств.

1.7. Заключение по обзору литературы

Обзор литературы подтверждает актуальность целей и задач, сформулированных в разделе «Введение»: для эффективного создания и использования сложных модульных моделей биологических систем требуется развитие особого подхода и создание соответствующего инструментария (специализированные программные комплексы и БД).

Чтобы обеспечить многократное использование описанных в диссертационной работе математических моделей и воспроизводимость полученных результатов численных расчетов в обзоре литературы рассмотрены современные стандарты в области системной биологии, которые использованы в диссертации.

Также рассмотрены современные подходы и программы для анализа и графического представления биомедицинских данных, которые использованы для реализации соответствующего функционала в ПК BioUML.

ГЛАВА 2. ПК BioUML

ПК BioUML разрабатывается с 2001 года по настоящее время под руководством автора в ООО «Биософт.Ру» и Федеральном исследовательском центре информационных и вычислительных технологий.

Ниже приведена некоторая статистика, показывающая сложность ПК BioUML:

- период разработки 2001 – 2023 годы;
- затрачено на разработку 100+ человеко-лет;
- объем кода (не включая сторонние библиотеки):
 - Java – 6 199 файлов, объем 36.8 Мб, ~1.2 млн. строк кода
 - JavaScript – 74 файла, 1.5 Мб, ~50 000 строк кода
- количество модулей – 194.

Учитывая огромный размер кода и выполняемых функций ПК BioUML, разработанных за 22 года, невозможно подробно описать его в данной главе. В приложении 2.1 представлены основные этапы развития ПК BioUML с ссылками на соответствующие публикации, а также указаны основные соавторы, которые внесли вклад в разработку соответствующего функционала.

В данной главе приведен краткий обзор интерфейса пользователя ПК BioUML, его архитектуры и основных компонентов. В главе 3 описан функционал, связанный с моделированием сложных биологических систем, а в главе 4 – функции ПК BioUML для анализа генетических и других омиксных данных. В главе 5 продемонстрировано использование ПК BioUML для создания и визуализации информации из БД GTRD. В главе 6 представлено использование ПК BioUML для создания модульных моделей сложных биологических систем. В главе 7 рассмотрена технология построения цифрового двойника пациента на основе созданных моделей. В главе 8 приведено описание других программных платформ, созданных на основе ПК BioUML.

2.1 Интерфейс пользователя

С точки зрения пользователя ПК BioUML – это интегрированный, расширяемый, свободно доступный ПК для формального описания, графического представления, моделирования сложных биологических систем и анализа биомедицинских данных (Kolpakov et al., 2019).

ПК BioUML доступен для пользователя в 2 вариантах:

1. BioUML workbench – Java приложение, которое устанавливается на компьютер пользователя;

2. BioUML сервер – в этом случае ПК BioUML устанавливается на сервере или в "облаке", а пользователь может использовать обычный веб-браузер.

Интерфейс ПК BioUML представлен на рисунке 2.1.1 Он состоит из следующих основных компонентов: меню и панель инструментов, панель репозитория, главное окно документа, вкладки/части документа и инспектор свойств.

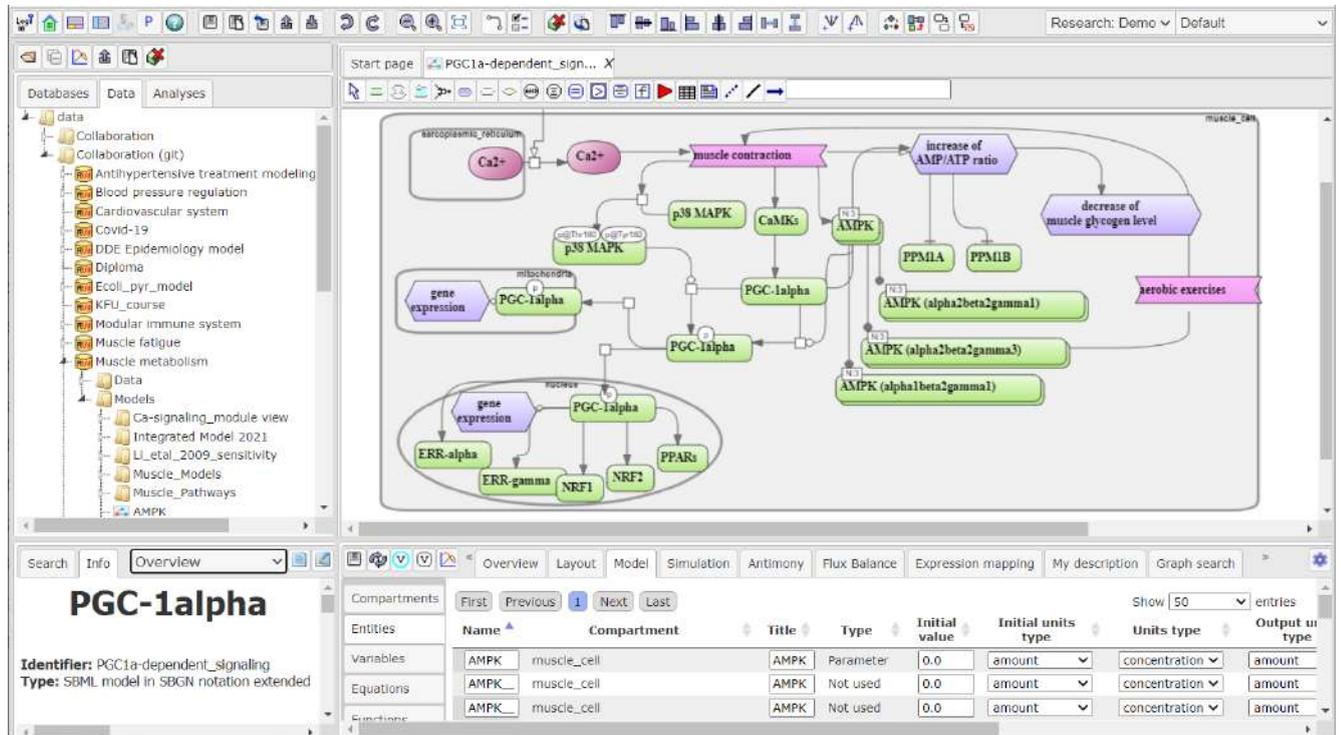


Рисунок 2.1.1 – Интерфейс ПК BioUML: верхняя левая панель – репозиторий проектов; верхняя правая панель – документ для редактирования, здесь фрагмент диаграммы регуляции экспрессии гена PGC-1alpha в графической нотации SBGN; левая нижняя панель – инспектор свойств для просмотра и редактирования

Кратко рассмотрим несколько ключевых понятий/концепций, используемых в представленном интерфейсе:

- *репозиторий* – представляет собой совокупность биологических БД и данных пользователя. Каждая БД определяет типы содержащихся в ней данных (белки, гены, молекулы, реакции и т.д.), доступные типы диаграмм для представления данных, отображение элементов БД на элементы диаграммы, а также механизм поиска элементов БД. В репозитории могут содержаться как локальные БД, так и загруженные с сервера. Новые БД могут быть созданы по шаблону, либо импортированы пользователем. В виде репозитория также представляются проекты пользователя и доступные методы для анализа данных.
- *документ* – некоторые элементы репозитория имеют достаточно сложную структуру и

могут быть открыты в виде отдельного документа. Каждый документ определяет способ отображения и редактирования элемента БД. Документ может определять набор дополнительных вкладок (view parts) для работы с отдельными аспектами элемента.

2.2 Архитектура ПК BioUML

Архитектура текущей версии платформы BioUML представлена на рисунке 2.2.1 (Kolpakov et al., 2022). В разделе 2.3 будут кратко рассмотрены основные программные технологии и платформы, используемые как компоненты ПК BioUML.

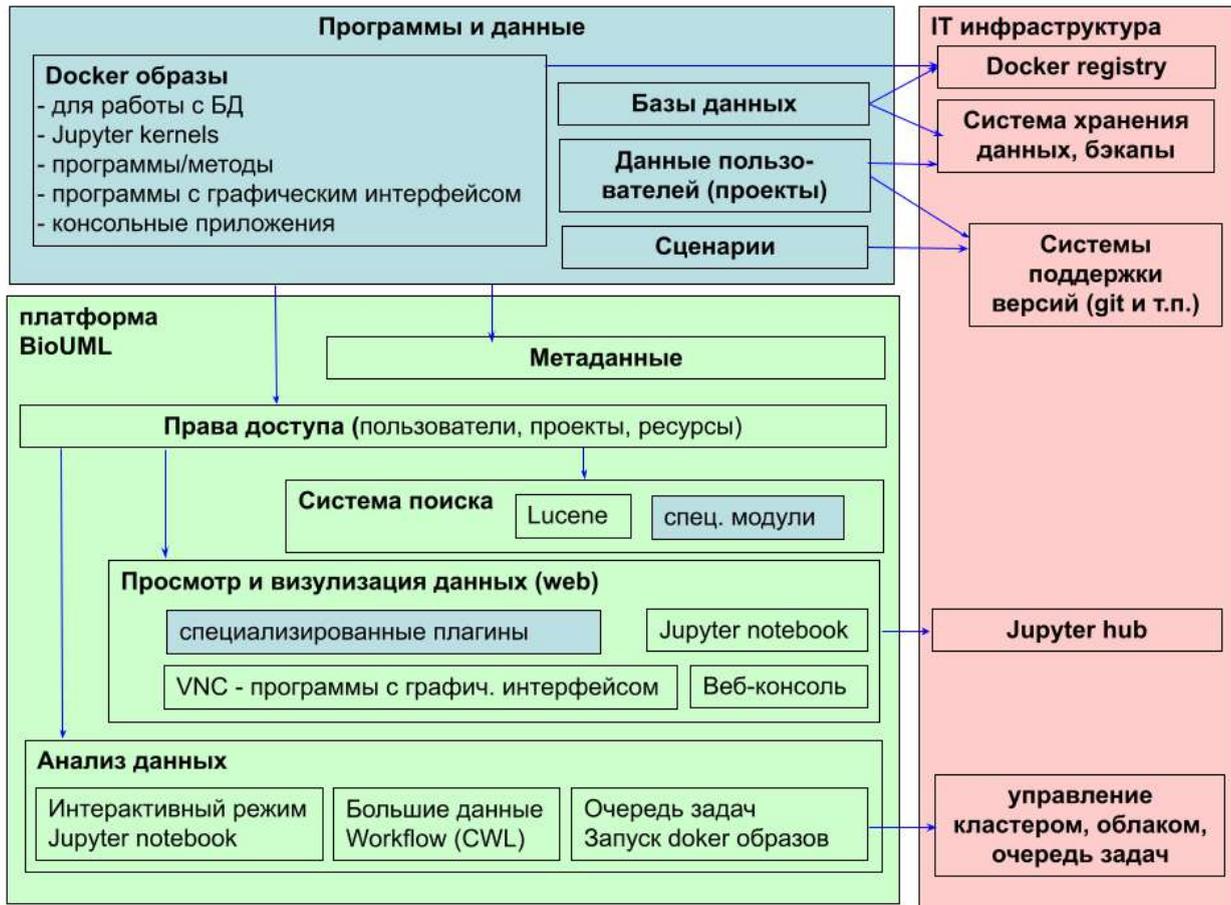


Рисунок 2.2.1 – Архитектура ПК BioUML

Предлагаемая архитектура направлена на решение следующих задач:

- следование принципам FAIR – Findable, Accessible, Interoperable, and Re-usable – находимости, доступности, совместимости и переиспользования;
- воспроизводимость результатов исследований (reproducible research);
- совместную работу над проектами (collaborative research);
- удобный пользовательский веб-интерфейс для поиска и анализа биомедицинских данных, а также их графического представления.

2.3 Используемые технологии

В данном разделе будут кратко рассмотрены основные программные технологии и платформы, используемые как компоненты ПК BioUML.

2.3.1 OSGi – модульная архитектура

В основе архитектуры ПК BioUML лежит модульный подход на основе спецификации динамической модульной системы и сервисной платформы для Java-приложений OSGi (Open Services Gateway Initiative), разрабатываемой консорциумом OSGi Alliance (<https://www.osgi.org/>). В качестве реализации OSGi спецификации используется Equinox OSGi, разрабатываемый сообществом Eclipse (<http://www.eclipse.org/equinox/>).

Приложение состоит из модулей (plug-ins), которые взаимодействуют друг с другом через определенные соглашения/интерфейсы (extension points) (рисунок 2.3.1). Например, один модуль может объявить точку расширения для экспорта диаграмм, а другие модули обеспечат экспорт диаграмм в разных форматах. Как правило, каждый модуль состоит из:

- jar файлов, содержащих выполняемый Java код и вспомогательные ресурсы (иконки, тексты сообщений, и т.п.);
- вспомогательные файлы – тексты лицензий, документация и т.п.;
- plugin.xml – XML файл, который описывает модуль:
 - зависимости от других модулей;
 - входящие в состав модуля библиотеки (jar файлы) и их доступность для других модулей;
 - extensions points – точки расширения в которые другие модули могут добавлять свою функциональность;
 - extensions – функциональность, добавляемая к другим модулям.

На данный момент ПК BioUML включает 194 модуля (приложение 2.2).

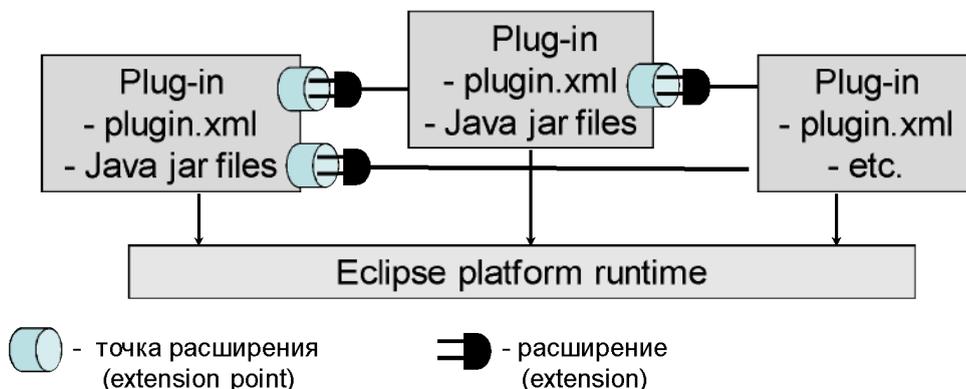


Рисунок 2.3.1 – Основные компоненты модульной архитектуры на основе стандарта OSGi

2.3.2 Технология JavaBeans

Технология JavaBeans разработана компанией Sun Microsystems Inc. Согласно спецификации (Sun Microsystems Inc, 1997), JavaBean – это многократно используемый программный компонент, который можно настраивать визуально. Для этого технология JavaBeans обеспечивает следующие возможности:

- интроспекция – возможность получить список свойств объекта;
- настраиваемость – возможность читать и устанавливать значения свойств объекта;
- поддержка "событий" – обеспечивает взаимодействие объекта JavaBean с программой и другими объектами;
- сохраняемость – возможность сохранить в поток вывода данных и восстановить значение всех свойств объекта из потока ввода данных.

2.3.3 Технология BeanExplorer

Технология BeanExplorer была разработана компанией DevelopmentOnTheEdge.com при непосредственном участии автора диссертационной работы и состоит из 3 частей:

- 1) расширение технологии JavaBeans – дополнительная мета-информация о свойствах JavaBean, динамические свойства, различные редакторы, часто встречающихся типов свойств;
- 2) внутренняя модель представления объектов JavaBean и динамических свойств в виде дерева;
- 3) набор инспекторов свойств, которые позволяют просматривать и редактировать информацию об объекте в различном виде:
 - для Java приложения – в виде редактора свойств, диалога, таблицы;
 - для веб приложения – в виде HTML формы, таблицы, или форматированного текста.

В данной работе технология BeanExplorer используется для просмотра и редактирования биологических данных, представленных в объектно-ориентированном виде. Для этого при помощи механизма интроспекции, встроенного в язык Java, для любого экземпляра Java класса, следующего соглашениям JavaBeans и/или использующего набор динамических свойств, технология BeanExplorer позволяет автоматически сгенерировать в момент выполнения программы элемент пользовательского интерфейса, позволяющий просматривать и редактировать все свойства соответствующего объекта (рисунок 2.3.2).

Подробная информация о технологии BeanExplorer доступна по адресу:
http://www.beanexplorer.com/beanexplorer_overview.shtml

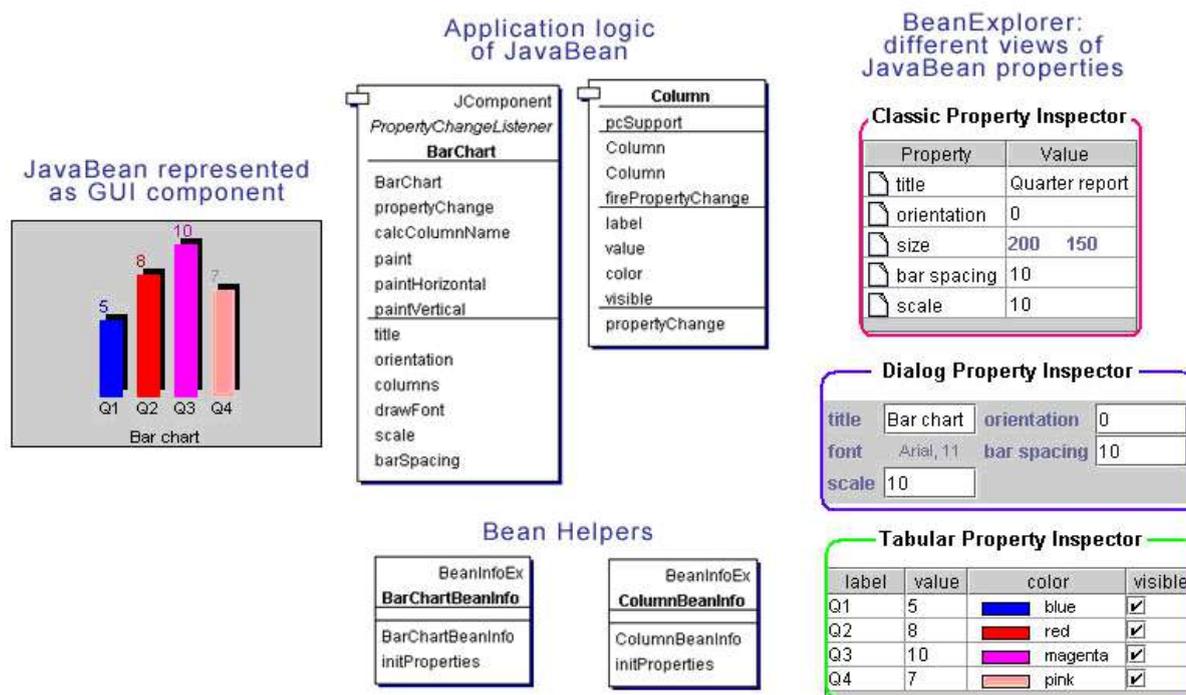


Рисунок 2.3.2 – BeanExplorer – автоматическая генерация компонентов пользовательского интерфейса (справа) на основе модели данных (в центре)

2.3.4 Динамические свойства

Для отображения информации из различных биологических БД в набор predefined Java классов без потери информации, а также для добавления свойств, определенных пользователем, используется концепция динамических свойств, реализованная в системе BeanExplorer, разработанной ООО "Девелопмент групп" (<http://beanexplorer.com>).

В работе автора (Колпаков, 2011) подробно описана данная технология. Здесь кратко опишем основные классы и понятия:

- `DynamicProperty` – определяет динамическое свойство как объект, состоящий из типа свойства (`type`), его значения (`value`) и мета-информации о свойстве, представляемой согласно спецификации JavaBeans в виде объекта типа `java.beans.PropertyDescriptor`.
- `DynamicPropertySet` – интерфейс, который определяет контейнер для динамических свойств. Динамические свойства могут быть добавлены и удалены из такого контейнера во время выполнения программы. Интерфейс `DynamicPropertySet` поддерживает механизм оповещения об изменении значений свойств, принятый в технологии JavaBeans. Для этого в классах, реализующих данный интерфейс, используется стандартный вспомогательный класс `java.beans.PropertyChangeSupport`.
- `DynamicPropertySetSupport` – одна из реализаций интерфейса `DynamicPropertySet`, где в качестве контейнера для динамических свойств используется `java.util.Vector`, а также для быстрого поиска свойства по его имени используются 3 хэш-таблицы (по `name`, `name` в верхнем регистре и `displayName`).
- `DynamicPropertySetAsMap` – реализация интерфейса `DynamicPropertySet`, где в качестве контейнера для динамических свойств используется `java.util.TreeMap`.

В таком контейнере свойства упорядочены по имени свойства. Таким образом, данный контейнер обеспечивает быстрый поиск свойства по имени, а с другой стороны, используется меньше памяти, по сравнению с `DynamicPropertySetSupport`.

2.3.5 Доступ и поиск информации в БД

Для доступа и поиска информации в биологических БД используется разработанная автором библиотека Java классов (Колпаков, 2011), обеспечивающая унифицированный интерфейс для работы с гетерогенными БД, представление информации из них в объектно-ориентированном виде, а также универсальную полнотекстовую поисковую систему. Данная библиотека кратко описана в приложении 2.3., базовые классы данной библиотеки свободно доступны в Git репозитории по адресу: <https://github.com/Biosoft-ru/access>.

2.3.6 Библиотека графических объектов

Библиотека графических объектов, разработанная автором, обеспечивает графические примитивы, функции и компоненты для визуального представления биологических данных. Кратко она описана в приложении 2.4, более подробно – в работе автора (Колпаков, 2011). Данная библиотека свободно доступна в Git репозитории по адресу: <https://github.com/Biosoft-ru/graphics>

2.3.7 Git, GitLab

Git (<https://git-scm.com>) – система контроля версий – система для хранения информации с поддержкой сохранения истории и совместной работы, стандарт де-факто для всех информационных систем.

ПК BioUML может использовать Git для хранения проектов пользователей по созданию и анализу моделей, а также для анализа биомедицинских данных. Для этого реализован специальный тип `DataCollection`, который позволяет интегрировать Git проекты в репозиторий ПК BioUML.

GitLab (<https://about.gitlab.com>) – веб-интерфейс для работы с git со множеством дополнительных функций (документирование, автоматическая сборка и т.п.). Используя возможности GitLab пользователь может просматривать и редактировать Git проекты по созданию и анализу моделей и геномных данных. GitLab является частью платформы Sirius-web (см. раздел 8.3).

2.3.8 Docker

Технология Docker (<https://www.docker.com>) позволяет упаковать в образы (image) различные программы и данные, так что затем их можно запустить на любом сервере или компьютере. Для хранения таких образом используется специальное хранилище (Docker registry), откуда такой образ может быть легко извлечен по его идентификатору и далее запущен на сервере.

ПК BioUML собирается как набор Docker образов, которые могут быть далее развернуты в "облаке" или на отдельном сервере, используя Docker Swarm (<https://docs.docker.com/engine/swarm/>).

Также ПК BioUML поддерживает использование Docker образов для работы со следующими типами данных:

- БД (например, Ensembl) могут быть упакованы в Docker образы. Это упрощает их установку;
- программы и платформы для анализа данных;
- ядра (kernels) для Jupyter hub.

Для каждого Docker образа существует метка (label), которая содержит мета-информацию об этом образе. Мы расширили этот подход, чтобы при динамическом подключении Docker образа к ПК BioUML в репозитории отображалась информация, о том какие программы и/или ядра Jupyter представлены на этом образе. Для этого в метку в формате json добавлен список CWL файлов для всех программ, входящих в данный образ и/или список ядер Jupyter.

2.3.9 Jupyter notebook, hub

Схема интеграция Jupyter notebook/hub в ПК BioUML представлена на рисунке 2.3.4 Jupyter ядра (kernel) хранятся в виде docker образов в специализированном хранилище. Из него они могут быть установлены в ПК BioUML:

- в процессе инсталляции и конфигурации ПК BioUML;
- при подключении хранилища docker образов к ПК BioUML.

По умолчанию в состав платформы входят 4 Jupyter ядра:

- SoS – Script of Scripts – специальное ядро, позволяющее использовать несколько ядер (для разных языков программирования) в одном блокноте Jupyter, а также специализированный язык сценариев. Для более подробной информации смотрите документацию на сайте разработчика – <https://vatlab.github.io/sos-docs/>;
- Python 3 (BioUML) – ядро для выполнения скриптов Python (версия 3.9);
- R (BioUML) – ядро для выполнения скриптов R;
- JS (BioUML node) – ядро для выполнения скриптов на языке JavaScript. Для работы этого ядра запускается специально сконфигурированная версия (BioUML node), которая может взаимодействовать с основной версией ПК BioUML через REST API. Для выполнения JavaScript используется библиотека Rhino, которая позволяет выполнять JavaScript код внутри Java машины. Благодаря такому подходу, оператор может использовать

внутренние функции Платформы. Также для работы с платформой существует множество JavaScript функций и объектов.

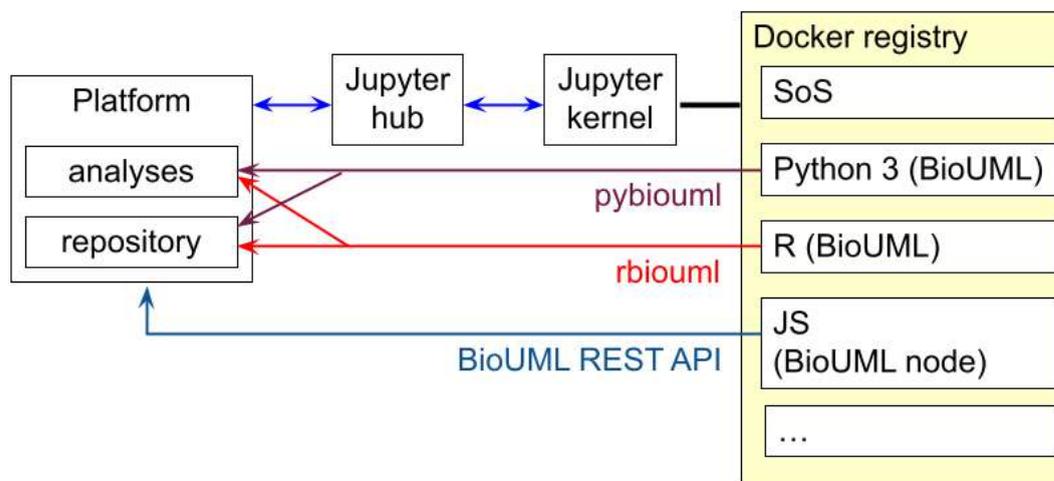


Рисунок 2.3.3 – Схема интеграция Jupyter notebook/hub в ПК BioUML

Для доступа к репозиторию и методам анализа данных Платформы из скриптов R и Python 3 используются библиотеки:

- **rbiouml** (<https://cran.r-project.org/web/packages/rbiouml/index.html>) – библиотека для работы с BioUML с помощью написания кода на языке программирования R;
- **pybiouml** (<https://pypi.org/project/pybiouml/>)- библиотека для работы с BioUML с помощью написания кода на языке программирования Python 3.

На данный момент обе библиотеки включены в ядра для Python и R соответственно, и любой пользователь может воспользоваться ими при работе в Jupyter Notebook. Данные библиотеки позволяют выполнять широкий спектр задач, начиная от импортирования и экспортирования данных и заканчивая запуском различных сценариев и анализов, при этом используя лишь Jupyter Notebook. Более того, эти библиотеки позволяют работать с данными с помощью языков программирования R и Python непосредственно в самом Jupyter Notebook без необходимости перемещать или преобразовывать данные какими-либо другими способами.

ГЛАВА 3. ТЕХНОЛОГИЯ СОЗДАНИЯ И ИСПОЛЬЗОВАНИЯ МОДУЛЬНЫХ МОДЕЛЕЙ БИОЛОГИЧЕСКИХ СИСТЕМ

3.1 Мета-модель

В данном разделе рассматривается разработанная автором мета-модель, используемая для формального представления структуры биологических систем.

Мета-модель («модель модели» биологической системы) обеспечивает формализм для комплексного описания, графического представления и численного моделирования широкого круга биологических систем. Она состоит из трех взаимосвязанных уровней описания сложной системы (рисунок 3.1.1):

- 1) граф – структура системы описывается как компарментализованный граф;
- 2) семантика – на этом уровне описывается биологическая (или другая) семантика каждого объекта. Для этого указывается тип объекта, его основные свойства и/или приводятся ссылки на соответствующие записи в БД;
- 3) математическая модель – переменные, формулы, уравнения, события, состояния и переходы. Например, вершины графа могут быть переменными или состояниями системы, а ребра графа могут быть уравнениями, описывающими изменения этих переменных или условными переходами между состояниями.

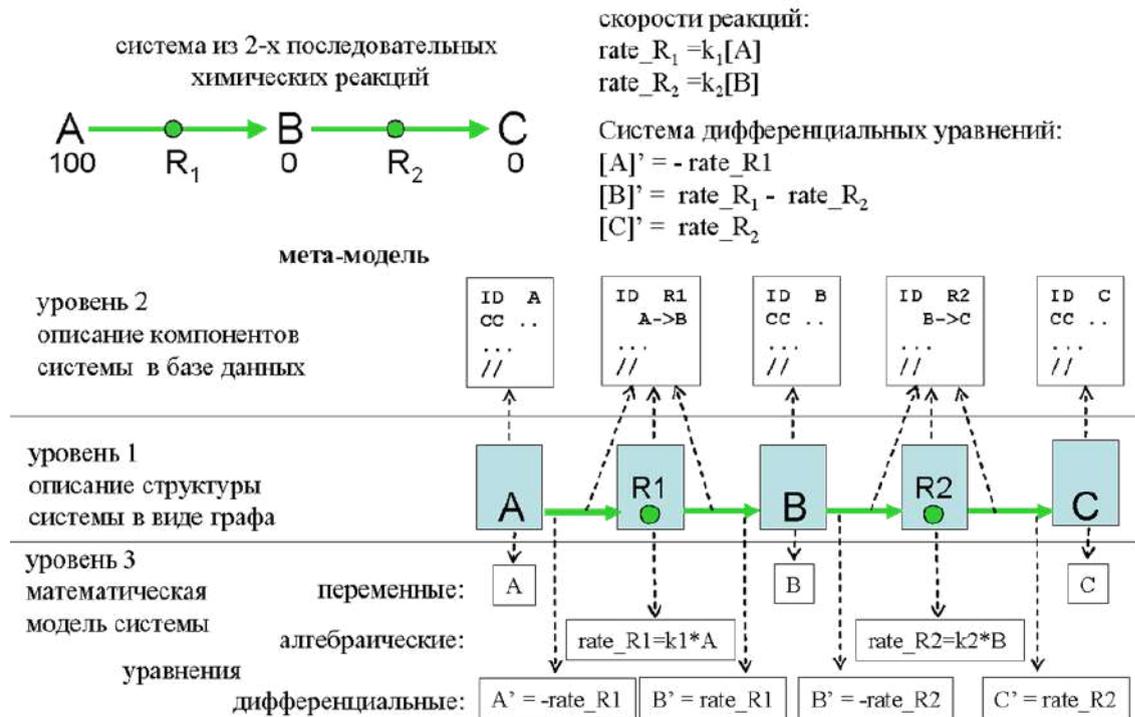


Рисунок 3.1.1 – Пример отображения простой системы, состоящей из двух последовательных химических реакций и ее представление по уровням мета-модели

Информация из широко используемых БД по биологическим путям и моделям (KEGG, TRANSPATH, Reactome, BioModels и другие), а также данные в форматах SBML и BioPAX могут быть полностью представлены в терминах разработанной мета-модели. Для хранения данных в терминах мета-модели разработан специальный XML формат – DML – Diagram Markup Language.

Информация о каждом биологическом объекте представляется в виде объекта Java, следуя соглашениям технологии JavaBeans. Соответствующий Java класс определяет основные свойства биологического объекта, а дополнительная информация о биологическом объекте представляется в виде набора динамических свойств, связанных с конкретным экземпляром соответствующего Java класса. Подробно эти Java классы представлены в работе (Колпаков, 2011).

Кроме того предложенная мета-модель используется для представления сценариев анализа данных, а также может быть применена и к другим предметным областям (см. раздел 8.6 как пример).

3.2 Объектно-ориентированное представление биологических данных

В мета-модели каждый элемент графа, как правило, связан с некоторой биологической сущностью (ген, последовательность, функциональный сайт, РНК, белок, низкомолекулярное вещество, химическая реакция, компартмент, процесс и т.п.). Для описания свойств этих биологических объектов автором разработана библиотека Java классов, подробно описанная в работе (Колпаков, 2011).

Таким образом, информация о каждом биологическом объекте представляется в виде объекта Java, следуя соглашениям технологии JavaBeans и BeanExplorer. Это позволяет использовать универсальный инспектор свойств, предоставляемый технологией Bean Explorer для просмотра и редактирования свойств соответствующих биологических объектов.

Первая реализация данного подхода была осуществлена автором в системе MGL в 1994-1998 г. (Колпаков, Бабенко, 1997). В дальнейшем этот подход был реализован в системе GeneNet (Kolpakov et al., 1998; Kolpakov, Ananko, 1999).

Однако, разработанная библиотека Java классов фиксирует в виде статических полей только основные свойства соответствующего биологического объекта. В различных биологических БД может содержаться дополнительная информация о биологических объектах, не покрываемая статическими полями разработанной библиотеки классов. Поэтому для отображения информации из БД в объекты Java без потери информации, а также для добавления свойств, определенных пользователем, используется концепция динамических свойств, реализованная в системе BeanExplorer и описанная в разделе 2.3.3

3.3 Визуальное моделирование

ПК BioUML упрощает создание сложных биологических моделей путем визуального моделирования:

- пользователь создаёт и редактирует модель биологической системы в виде диаграммы;
- на основе созданной диаграммы ПК BioUML генерирует Java код;
- сгенерированный Java код компилируется и используется соответствующими решателями для численного моделирования динамики исходной системы.

Основные компоненты ПК BioUML для визуального моделирования, и их взаимодействия друг с другом показаны на рисунке 3.3.1. Данная архитектура позволяет численно моделировать широкий спектр биологических систем, используя:

- алгебро-дифференциальные уравнения, также поддерживаются уравнения с задержкой переменной по времени, кусочно-непрерывные функции, события (гибридные модели), состояния и переходы;
- уравнения в частных производных (на данный момент только 1D для моделирования гемодинамики);
- стохастическое моделирование;
- модульные модели;
- мультиагентные модели (agent based modeling);
- модели, на основе правил (rule-based models, Harris et al., 2016). В этом случае описывается набор правил для взаимодействия компонентов системы в зависимости от их состояния. Далее по этим правилам строится возможный граф реакций.

3.3.1 Графическая нотация

Необходимой составной частью визуального моделирования является графическая нотация (визуальный синтаксис, графический язык) которая позволяет формально и полно описать модель (Lee, 2001). Она может быть представлена несколькими типами диаграмм, позволяющими описать разные аспекты структуры и функционирования сложных систем с разным уровнем формализации и детализации. В разработанном автором подходе основными компонентами формального определения графической нотации являются:

- 1) типы вершин и ребер – какие типы биологических (и математических) объектов и их взаимодействий могут быть показаны на диаграмме;
- 2) правила отрисовки объектов – как биологические объекты должны быть представлены на диаграмме, учитывая их свойства. Например, РНК может быть представлена в виде волнистой линии, белок может быть представлен в виде круга, а маленькие круги на нем могут представлять различные модификации аминокислотных остатков (фосфорилирование, убиквитинирование и т.п.);
- 3) семантические правила и ограничения – используются для обеспечения семантической целостности диаграммы в процессе ее создания и редактирования. Например, если

пользователь удаляет вещество с диаграммы, то все реакции, в которых оно участвует должны быть тоже удалены.

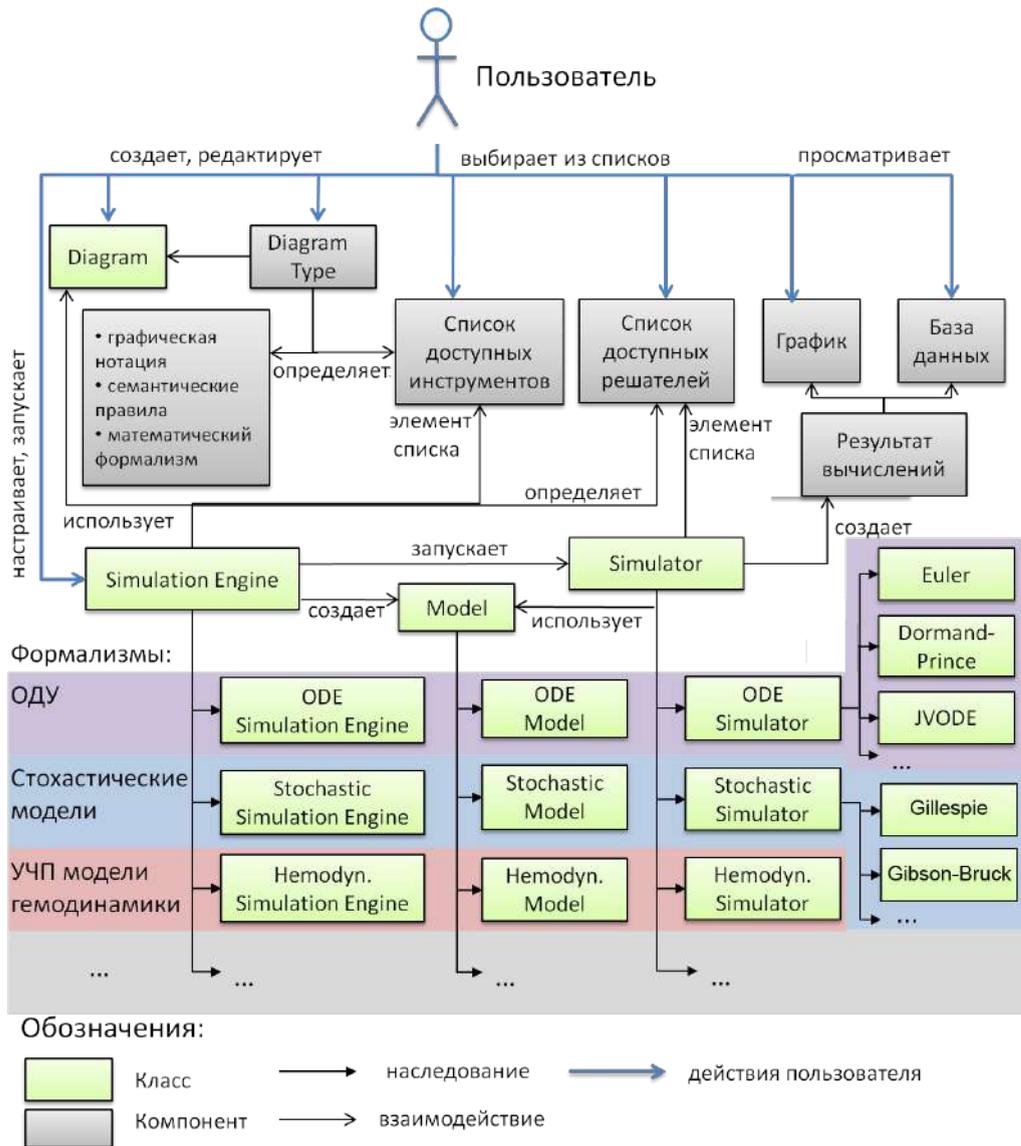


Рисунок 3.3.1 – Основные компоненты архитектуры для визуального моделирования в ПК BioUML (Киселев, 2016)

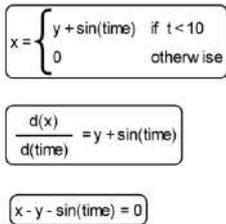
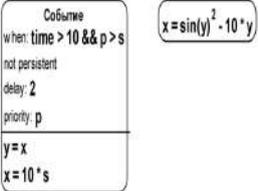
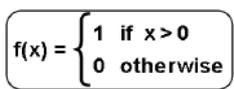
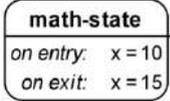
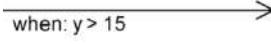
Для формального определения графической нотации автором была разработана библиотека Java классов и интерфейсов (пакет `biouml.model`), подробно описанная в работе (Колпаков, 2011). Краткое описание основных Java интерфейсов приведено в приложении 3.1.

В настоящее время в ПК BioUML используются следующие графические нотации (приведены только основные) и соответствующим им типы моделей:

- SBML model in SBGN notation – основной тип моделей, используется расширенная SBGN нотация;
- SBML 'comp' model in SBGN notation – модульная модель в расширенной SBGN нотации (таблица 3.3.1);

- Mathematical model – используется для моделирования физиологии организмов, основные компоненты модели: алгебраические и дифференциальные уравнения и события.
- Arterial tree – используется для моделирования кровотока в основных артериях и артериолах;
- BioNetGen model – модели, на основе правил (rule-based models, Harris et al., 2016). В этом случае описывается набор правил для взаимодействия компонентов системы в зависимости от их состояния. Далее по этим правилам строится возможный граф реакций;
- Agent-based model – агентная модель.

Таблица 3.3.1 – Графическая нотация математических элементов диаграмм BioUML, соответствующее обозначение в SBGN и элемент языка SBML

BioUML-нотация	SBGN	SBML	Описание
	-	Rule: scalar rate algebraic	Уравнение присваивание дифференциальное уравнение алгебраическое уравнение
	-	Event	Мгновенное событие , заключающееся в скачкообразном изменении значений переменных модели при выполнении определенного условия (например, в определенный момент времени или при заданном соотношении между переменными).
	-	Function	Заданная пользователем функция , которая может использоваться в уравнениях или реакциях модели.
	-	Event	Состояние системы, характеризуемое скачкообразным изменением переменных модели при входе в состояние и выходе из него. Вместе с переходами математически эквивалентно набору мгновенных событий.
	-	Event	Переход из одного состояния в другое при выполнении определенного условия.

3.3.2 Исходная графическая нотация ПК BioUML

Исходно, в ПК BioUML было определено 5 типов диаграмм (Колпаков, 2011) для формального описания и моделирования биологических систем, преимущественно на клеточном и организменном уровне:

1. **Семантическая сеть** – описывает концепции (признаки, процессы или состояния биологических систем), отношения между ними, а также их связь с другими компонентами молекулярно-биологических систем (например, демонстрирует, что определенный ген связан с определенным признаком).
2. **Метаболический путь** – описывает биологические пути, задействованные в метаболических процессах. Основными компонентами таких диаграмм являются вещества, белки, реакции и компартменты.
3. **Генная сеть** – данный тип диаграмм, близок к графической нотации, используемой в системе GeneNet. Основные компоненты – гены, белки, вещества, реакции, регуляторные события и компартменты.
4. **Генная сеть + семантическая сеть** – это гибридный тип диаграмм, когда для описания отдельных звеньев генной сети не хватает экспериментальных данных, то они могут быть описаны только на семантическом уровне (например, когда показано, что какие-то процессы связаны с определенным геном, но механизм их не известен).
5. **Математическая модель** – расширение графической нотации генной сети, где с элементами графа ассоциированы переменные и уравнения. Также добавляются специальные обозначения для вершин графа соответствующих уравнениям, функциям, событиям, состояниям и переходам, что позволяет описать дискретное поведение системы, например, смену фаз клеточного цикла.

Данная нотация является развитием графической нотации, разработанной для системы GeneNet (Kolpakov et al., 1998). Ее краткое описание приведено в приложении 3.2.

3.3.3 Расширенная нотация SBGN

В связи с возникновением графической нотации SBGN, предназначенной для представления биологических путей (метаболические пути, пути передачи сигнала в клетке, генные сети) был осуществлен переход на использование этой графической нотации, совместно с форматом SBML. Для этого, используя XML тэги <annotation> в формате SBML сохраняется дополнительная информация (типы и расположение вершин графа, заголовки, цвета и т.п.).

3.3.4 Моделирование физиологических систем

Характерной чертой биологических систем, которую необходимо учитывать при математическом моделировании, является существенная структурная сложность, обусловленная многоуровневым иерархическим строением организма. Указанную иерархию можно выразить следующей цепочкой: организм – системы организма – органы – клетки – клеточные органеллы – молекулярные структуры. На каждом уровне этой цепочки действуют свои законы, тесно

связывающие элементы одного уровня друг с другом. Так на уровне систем организма можно выделить кровеносную, дыхательную, пищеварительную, нервную и другие системы. Все эти системы взаимодействуют друг с другом, образуя систему более высокого уровня – живой организм.

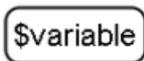
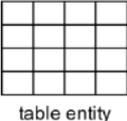
Принципы модульного подхода, реализованные в рамках ПК BioUML и описанные в разделе 3.4, хорошо согласуются с иерархической структурой биологических систем и позволяют создавать модели различных элементов и различных уровней этой структуры и объединять всё в рамках одной модели.

Важно заметить, что многие принципы и законы, действующие в физиологических системах не имеют на сегодняшний день полного теоретического обоснования и представляют собой аппроксимацию огромного количества экспериментальных данных, выраженную уравнениями или графиками зависимости одних параметров от других.

Визуально, модели таких систем представляют собой системы дифференциальных и алгебраических уравнений. Если модель достаточно большая – становится трудно отследить зависимости между этими уравнениями. Для упрощения работы с такими моделями в ПК BioUML реализован тип диаграммы “Mathematical model” (Киселев, 2016).

Такая диаграмма может содержать следующие элементы: уравнения, события, состояния, переходы между состояниями, примечания, блоки и порты. Помимо этого добавлены дополнительные элементы, их графическая нотация представлена в таблице 3.3.2.

Таблица 3.3.2 – Графическая нотация дополнительных элементов модели физиологических процессов

Графическая нотация	Описание
	Переменная – соответствует свойству или характеристике объекта (длина или упругость сосуда, объем крови и т.п.)
	Зависимость между уравнениями: - красный – положительная; - синий – отрицательная; - черный – неопределенная.
	Табличный элемент – задает зависимость между двумя переменными в виде таблицы. По таблице автоматически генерируется кубический сплайн, который представляется в модели в виде правила присваивания (Киселев, 2016).
Выходной порт  Входной порт  Контактный порт 	Порт (входной, выходной, контактный) – указывает на переменные, значения которых должны быть переданы в модель извне или переданы из модели наружу (см. модульное моделирование).

Зависимости между уравнениями генерируются автоматически. Например, заданы два уравнения, где в 1-ом уравнении, переменная v_2 влияет на значение переменной v_1 :

$$v_1 = f_1(v_2, \dots)$$

$$v_2 = f_2(x, y, z)$$

Тогда переменные v_2 и в v_1 связаны:

- положительной связью, если увеличение v_2 вызывает увеличение v_1 ;
- отрицательной связью, если увеличение v_2 вызывает уменьшение v_1 ;
- неопределенной связью – изменение v_2 не влияет на v_1 ,

при фиксированных остальных значениях и заданных начальных значениях всей модели. Связи не меняют математический смысл диаграммы, их назначение – наглядное представление графа связи между уравнениями модели.

Диаграмма физиологических процессов может быть представлена в 2 видах:

- полный вариант (рисунок 3.3.2.а) – все уравнения модели явно представлены на диаграмме;
- краткий вариант (рисунок 3.3.2.б) – уравнения представлены только именем переменной, рассчитываемой в данном уравнении.

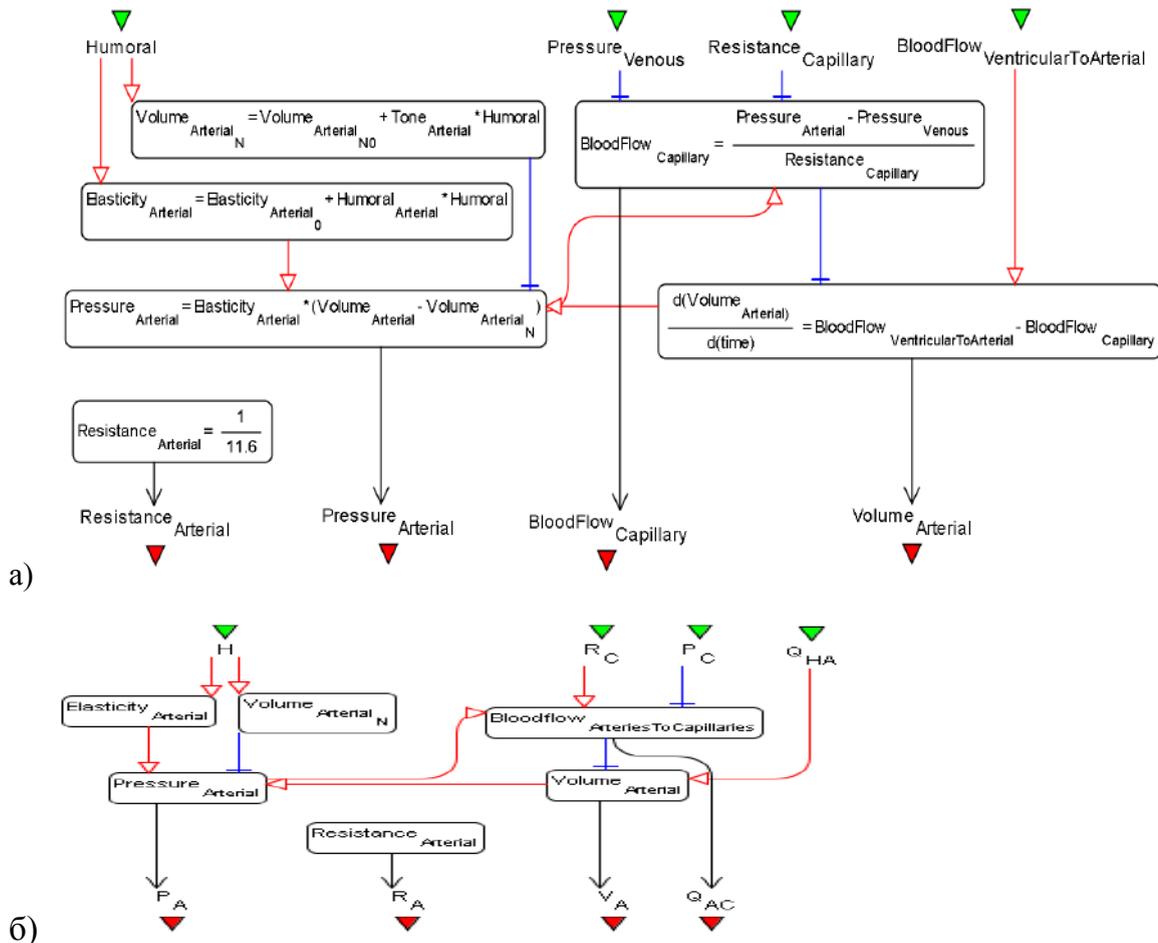


Рисунок 3.3.2 – Модуль артериальной системы, реализованная в виде диаграммы типа “Mathematical model”, а) полный вариант; б) краткий вариант

3.3.5 Артериальное дерево

Отдельный тип диаграмм "Arterial tree" в ПК BioUML специально разработан для работы с моделями тока крови по сосудистому дереву. Модель представляет бинарное дерево состоящее из точек разветвления (вершины) и сосудов (ребра). Конфигурацию сосудов в дереве и их свойства задает пользователь. Графическая нотация элементов дерева представлена в таблице 3.3.3.

Краткое описание моделирование гемодинамики с использованием ПК BioUML приведено в главе 6, подробное описание – в работе Бибердорф и др., 2015.

Таблица 3.3.3 – Графическая нотация диаграммы типа “Arterial tree”

Графическая нотация	Описание
	точка разветвления сосудов
	сосуд – характеризуется следующими свойствами: название, длина, площадь поперечного сечения, коэффициент жесткости.
	сердце – вершина, через которую кровь втекает в дерево.
$x = \begin{cases} y + \sin(\text{time}) & \text{if } t < 10 \\ 0 & \text{otherwise} \end{cases}$ $\frac{d(x)}{d(\text{time})} = y + \sin(\text{time})$ $x - y - \sin(\text{time}) = 0$	уравнения: <ul style="list-style-type: none"> - присваивание - дифференциальное уравнение - алгебраическое уравнение
	контрольная точка – выводит значение параметров сосудов на график во время численных расчетов. Пользователь определяет: <ul style="list-style-type: none"> - сосуд, которому соответствует точка, - область сосуда, в которой измеряется значение: начало сосуда, середина или конец, - тип переменной: давление, полное давление, поток крови, скорость потока, сопротивляемость или площадь сечения сосуда.
Выходной порт   Входной порт Контактный порт 	порт (входной, выходной, контактный) – указывает на переменные, значения которых должны быть переданы в модель извне или переданы из модели наружу.

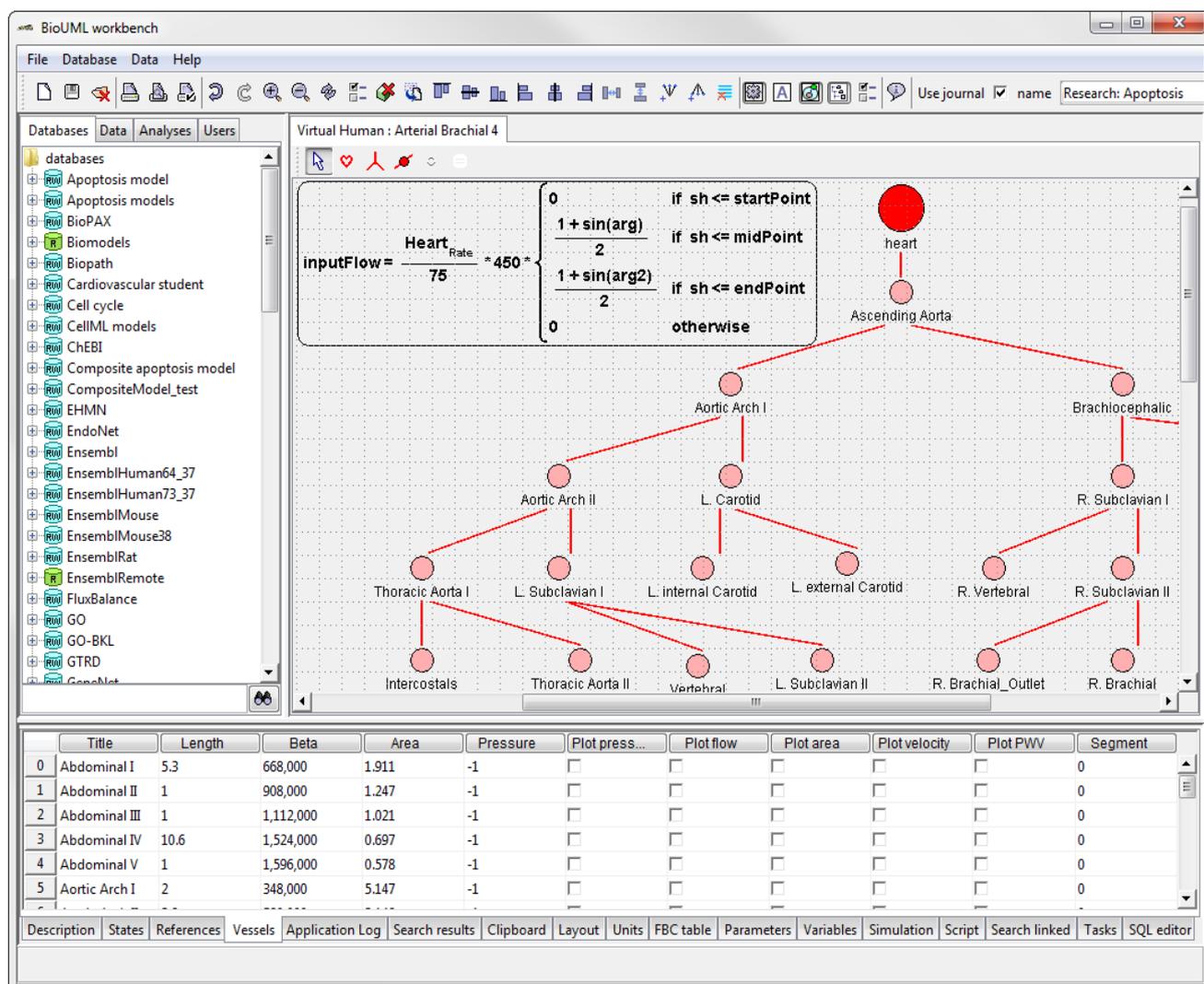


Рисунок 3.3.3 – Пользовательский интерфейс визуального создания диаграммы типа “Arterial tree”. Графическая нотация приведена в таблице 3.3

3.4 Модульное моделирование

Модульный подход к математическому моделированию сложных биологических систем заключается в разделении системы на части, для каждой из которых модель может быть построена отдельно. Модель всей системы строится как комбинация моделей подсистем (модулей). Взаимодействие блоков моделируется связями между переменными модулей, которые указывают на пути передачи сигналов между ними. Модули представляют собой математические модели различных подсистем, также они могут содержать вложенные модульные модели, составляя иерархическую структуру.

Ключевой аспект при создании модульных моделей – четкая спецификация интерфейса модуля для взаимодействия с другими частями модели. В ПК BioUML для этого используется понятие портов (Kiselev and Kolpakov, 2012, 2013;). В зависимости от направления передачи сигнала выделяется 3 типа портов и связей между модулями (на рисунке 3.4.1 представлена

аналогия между портами модуля биологической модели и звуковой карты для компьютера):

- вход – значение переменной подается на вход модуля и не может быть изменено модулем;
- выход – значение переменной полностью определяется (вычисляется) внутри модуля и подается на вход в неизменном виде другим модулям;
- контакт – разделяемая переменная, соединенные модули могут вносить изменения (приращение или убывание) в значение соответствующей переменной. Например, количество АТФ может изменяться в нескольких модулях модели работы скелетных мышц (в одних производится, в других – потребляется). В этом случае такие модули соединяются через порты типа контакт.

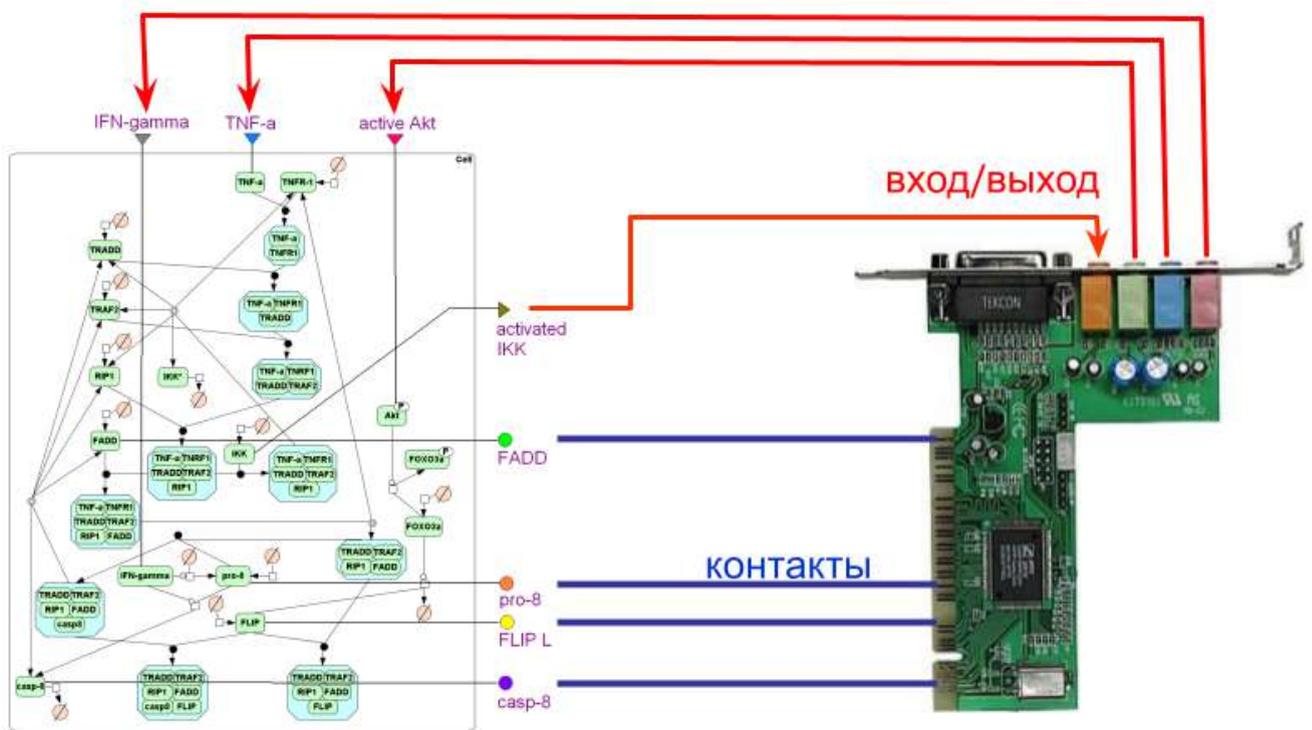


Рисунок 3.4.1 – Аналогия между портами модуля биологической модели и звуковой карты для компьютера

Соответственно для одной и той же переменной порты могут соединяться двумя типами связей (рисунок 3.4.2):

- направленная – для портов "вход – выход";
- ненаправленная – для портов типа "контакт".

Графическая нотация для представления модульных моделей представлена в таблице 3.4.1. В ней добавлено несколько вспомогательных элементов: усреднитель, переключатель, константа, график и шина.

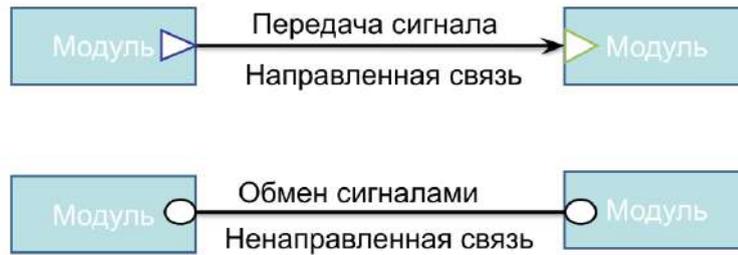
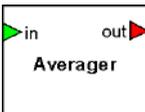
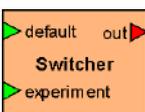


Рисунок 3.4.2 – Типы связей между модулями

Таблица 3.4.1 – Элементы модульной модели

Графическая нотация	Название	Описание
	Подмодель	Модуль, содержащий математическую модель: <ul style="list-style-type: none"> • Модульная модель, • SBML (SBML-SBGN) модель, • Математическая модель, созданная в BioUML. Выходные и выходные переменные задаются портами.
	Усреднитель	Модуль, рассчитывающий скользящее среднее значение входного сигнала и подающий его на выход.
	Переключатель	Модуль, в зависимости от условия, подающий на выход один из двух сигналов, поступающих на вход.
	Константа	Модуль, подающий на выход константное значение.
	График	Модуль, принимающий сигнал и выводящий его на график.
	Шина	Переменная модульной модели. Несколько шин могут соответствовать одной переменной.
	Входной порт	Порт, определяющий входную переменную подмодели.
	Выходной порт	Порт, определяющий выходную переменную подмодели.
	Контактный порт	Порт, определяющий разделяемую переменную подмодели.
	Направленная связь	Связь, означающая передачу сигнала из одного модуля в другой.
	Ненаправленная связь	Связь, означающая обмен сигналами между модулями.

3.5 Агентное моделирование

Агентное моделирование – раздел имитационного моделирования, которое подразумевает представление системы в виде набора взаимодействующих сущностей способных самостоятельно принимать решения – агентов. Общее поведение системы складывается из взаимодействия агентов. Агентное моделирование связано с модульным в том смысле, что оно тоже моделирует систему как набор связанных подсистем.

Не существует общепринятого определения, что такое агент. Из практических соображений, принимают, что агент должен обладать следующими свойствами (Macal and North, 2009; Киселев, 2016):

- агент должен быть автономным, действовать независимо от других агентов за исключением ряда определенных взаимодействий;
- агент должен быть определен как отдельная сущность со своим набором характеристик;
- агент должен взаимодействовать с другими агентами;
- часто агенты рассматриваются как существующие в некоторой внешней среде.

Агентное моделирование в ПК BioUML является дальнейшим развитием модульного моделирования и использует ту же графическую нотацию. Основное отличие от модульной модели – каждый модуль может иметь собственный решатель (simulation engine) и шаг по времени. Таким образом, в рамках агентной модели можно объединить как разные подходы к моделированию динамики биологических систем (например, системы дифференциальных уравнений и стохастическое моделирование), так и модули с существенно различным шагом по времени. Например, в модели регуляции артериального давления для модуля "сердца" шаг по времени – доли секунды, для модуля "почка" – минуты и часы.

3.6 Текстовое представление модели, его синхронизация с графическим представлением

Выше было описано, что визуальное представление моделей биологических систем в виде графов существенно упрощает их восприятие человеком. Однако, для других сценариев работы более удобным может быть текстовое представление моделей. В идеале одна и та же модель должна иметь два представления: в виде диаграммы и в виде текста. При этом в процессе редактирования модели должна происходить синхронизация между этими представлениями. Это и было реализовано в ПК BioUML.

В качестве декларативного текстового представления был выбран язык Antimony (Smith et al., 2009; <https://tellurium.readthedocs.io/en/latest/antimony.html>), который позволяет представить в удобном человеко-читаемом виде SBML модели. Далее в этот язык был добавлен механизм аннотаций, при помощи которых для каждого элемента модели можно задавать

дополнительные свойства, необходимые для его графического представления в расширенной графической нотации SBGN Process Diagrams (приложение 3.3).

Также был разработан механизм автоматической синхронизации между представлением биологических моделей в графическом и текстовом виде. Для этого каждому элементу описанной выше мета-модели сопоставляется вершина дерева синтаксического разбора соответствующего текстового представления модели. Поток событий при редактировании диаграммы транслируется в соответствующие изменения текстовой модели и наоборот. Наиболее сложным для обработки является добавление новых элементов:

- при добавлении элемента в тексте мы должны разумным образом расположить новый элемент графа на диаграмме;
- при добавлении элемента на диаграмме мы должны разумным образом расположить новый фрагмент текста в текстовом представлении модели.

Для решения этой задачи был разработан ряд эвристик.

3.7 Автоматическая генерация кода

На основе формального описания биологической системы в терминах разработанной мета-модели автоматически генерируется Java код. Генерируемый код зависит от численного метода и типа диаграммы, используемой для дальнейших численных расчетов. На данный момент ПК BioUML поддерживает генерацию следующих типов кода для соответствующих численных решателей:

- алгебро-дифференциальные уравнения, также поддерживаются уравнения с задержкой переменной по времени, кусочно-непрерывные функции, события (гибридные модели), состояния и переходы;
- уравнения в частных производных (на данный момент только 1D для моделирования гемодинамики);
- стохастическое моделирование;
- стохастические дифференциальные уравнения;
- агентные модели (agent based modeling).

Также есть несколько препроцессоров, которые осуществляют преобразование диаграммы к некоторому стандартному виду, подходящему для генерации кода. На данный момент есть следующие препроцессоры:

- превращение модульной модели в "плоскую";
- обработка быстрых реакций, заменяя соответствующие дифференциальные уравнения на алгебраические, в соответствии с теоремой Тихонова;
- раскрытие правил (rule-based models, Harris et al., 2016) и генерация плоской модели.

Для генерации каждого Java кода для каждого типа модели используются

соответствующие шаблоны на языке Velocity (<https://velocity.apache.org/>). Стоит особо подчеркнуть, что для очень больших SBML моделей (10 000+ компонентов и реакций) разработан специальный шаблон. Это связано с тем, что Java ограничивает размер (64 Кб) выполняемой функции и объявлений переменных.

3.7.1 Разбор математических выражений

Одним из ключевых этапов генерации кода является разбор математических выражений, представленных в виде текста (для этого определена соответствующая грамматика) или в формате MathML (поддерживается подмножество, определенное в стандартах CellML и SBML level 2-3).

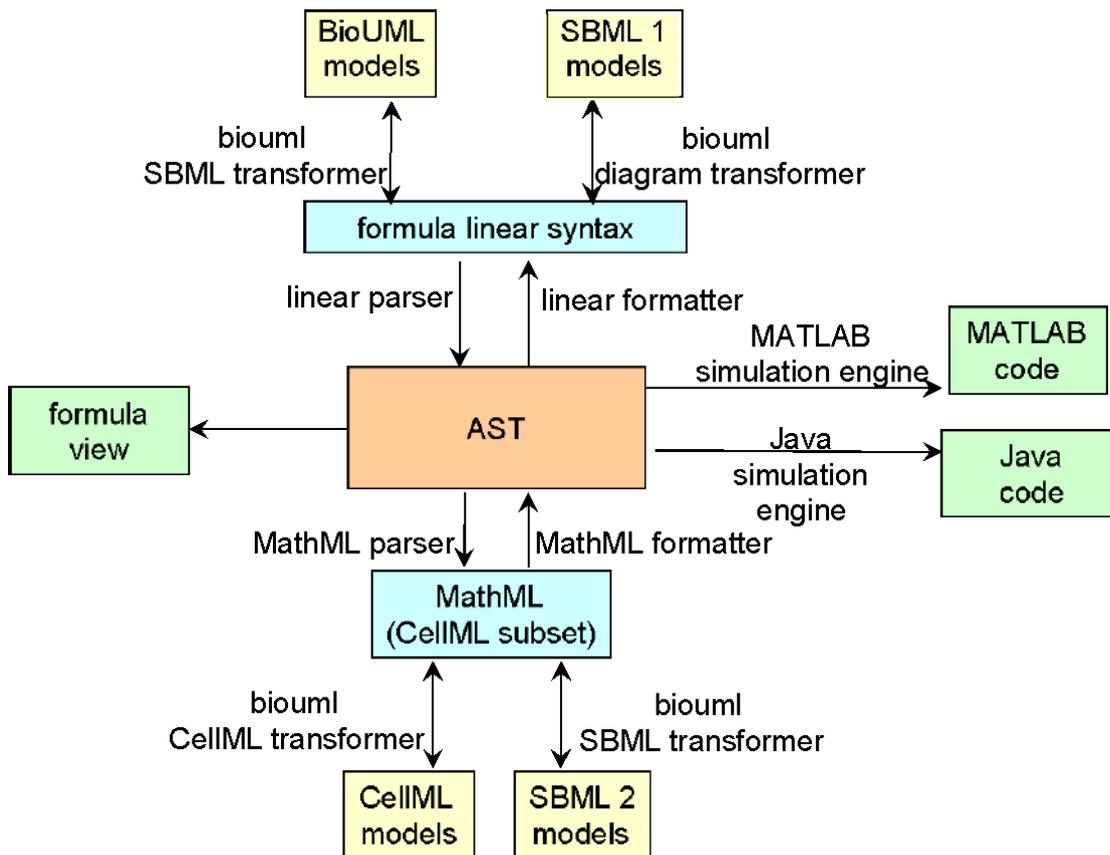


Рисунок 3.7.1 – Разбор и генерация математических выражений на основе абстрактного синтаксического дерева

Для этого нами разработан следующий подход (рисунок 3.7.1):

- абстрактное синтаксическое дерево (AST) – является внутренним и основным представлением математических выражений;
- Parser – осуществляет синтаксический разбор и семантический анализ математических выражений, результатом разбора является AST. Для разбора текстовых математических выражений определена соответствующая грамматика и при помощи

технологии JavaCC (Copeland, 2007; <https://javacc.github.io/javacc/>) сгенерирован соответствующий парсер. Для разбора выражений MathML написан парсер, который отображает DOM дерево, построенное на основе XML файла в соответствующее AST.

- `ParserContext` – содержит список переменных и функций, заданных в мета-модели. Эта информация необходима для семантической проверки выражений. Например, при разборе уравнения, задающего скорость реакции, проверяется, что переменные, соответствующие компонентам реакции, определены в модели (диаграмме), а необъявленные константы заносятся в общий список переменных.
- `Formatter` – генерирует по заданному AST выражению текстовую строку на заданном языке программирования, текстовое математическое выражение или фрагмент MathML. Более подробно этот подход описан в работе автора (Колпаков, 2011).

3.8 Численное моделирование

ПК BioUML поддерживает 6 основных типов решателей для численного моделирования динамики биологических систем:

- 1) алгебро-дифференциальные уравнения с дискретными событиями и задержкой по времени;
- 2) стохастические дифференциальные уравнения;
- 3) стохастические модели;
- 4) анализ стационарных потоков (Flux Balance Analysis, или FBA);
- 5) моделирования кровотока в основных артериях и артериолах;
- 6) агентное моделирование.

Для численного решения систем ОДУ реализованы следующие решатели:

- Euler – классический метод Эйлера;
- Dormand-Prince – классический численный решатель, использующий вложенные схемы Рунге-Кутты 4 и 5 порядка для контроля точности решения и автоматического управления длиной шага (Dormand, Prince, 1980);
- Imex – метод, одновременно использующий явную и неявную схемы семейства Рунге-Кутты (Patterson, 2002);
- Jvode – портированный на язык Java (Киселев, 2016) программный пакет CVODE (Cohen, Hindmarsh, 1996; Hindmarsh et al., 2005). Он реализует метод Адамса-Мултона или метод обратной производной (BDF – Backward Differential Formula);
- Radau5 – неявный метод Рунге-Кутты 5-го порядка для решения жестких задач (Hairer and Wanner, 1996).

Для решения алгебраических уравнений используются методы Ньютона (Канторович,

1949), Левенберга-Марквардта (Levenberg, 1944; Marquardt, 1963), Нелдера-Мида (Nelder and Mead, 1965), а также метод сопряженных градиентов (Hestenes and Stiefel, 1952).

Для учета задержки по времени автоматически определяется список необходимых переменных и для них хранится история на глубину максимальной задержки. Значения в требуемой временной точке получаются методом линейной интерполяции и экстраполяции.

Для решения стохастических дифференциальных уравнений реализован метод Эйлера-Маруяма (Maruyama, 1955).

Стохастическое моделирование – для того чтобы учесть случайные процессы, оказывающие влияние на динамику системы, применяются стохастические методы. Это особенно важно, когда в реакции участвует малое количество реагентов. В ПК BioUML реализованы следующие методы:

- Gillespie – классический алгоритм расчетов для стохастических моделей (Gillespie, 1977);
- Gibson-Bruck – алгоритм, который позволяет ускорить вычисления по сравнению с алгоритмом Gillespie (Gibson and Bruck, 2000);
- Tau-leaping – приближенный алгоритм, вычисляющий сколько реакций произойдет за определенный интервал времени и выполняющий сразу несколько реакций. Алгоритм позволяет добиться существенного выигрыша в скорости расчетов, но страдает от потери точности (Puchalka and Kierzek, 2004);
- Maximal TimeStep – гибридный алгоритм, разделяющий реакции на медленные (задействующие малое количество молекул) и быстрые (задействующие большое количество молекул) (Puchalka and Kierzek, 2004). Медленные реакции симулируются точными методами (Gillespie или Gibson-Bruck), быстрые – алгоритмом Tau-leaping.

Анализ стационарных потоков (Orth et al., 2010) использует методы линейного программирования для оценки потока, проходящего через метаболическую сеть. Для этого вводится целевая линейная функция, которую необходимо максимизировать/минимизировать, при этом не выйдя за рамки допустимого решения. При этом подходе используются минимальные данные о кинетических параметрах реакций (стехиометрические коэффициенты и границы, определяющие скорость реакции).

Для моделирования кровотока в основных артериях и артериолах реализован специальный численный решатель, подробно описанный Бибердорф и др., 2015.

Агентное моделирование используется для численных расчетов модульных моделей, когда алгоритм генерации плоской модели не может быть применен. Численные расчеты для каждого агента проводятся отдельно с использованием подходящего численного метода (например, метода Рунге-Кутты для системы ОДУ или алгоритма типа Gillespie для стохастической модели). Координация агентов и обмен сообщениями между ними обеспечивается

специальным планировщиком (рисунок 3.8.1). Обмен сообщений производится в соответствии с установленными в модели связями между агентами и заключается в однонаправленной или двунаправленной передаче изменения значения интерфейсных переменных агентов со времени предыдущего обмена сообщениями. Такой подход позволяет объединять математические модели произвольного типа в одну агентную модель.

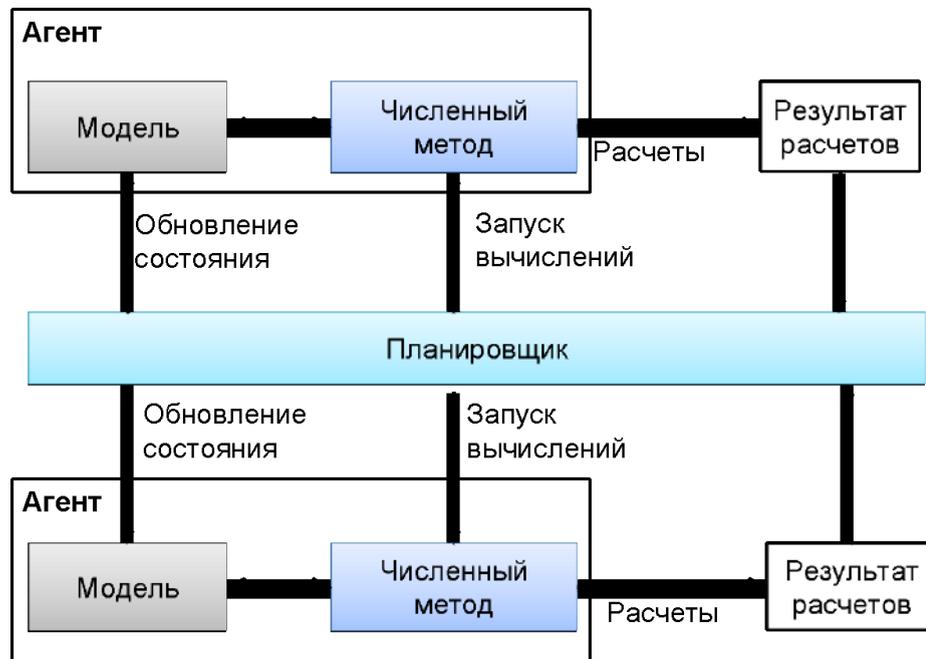


Рисунок 3.8.1 – Реализация агентного моделирования в ПК BioUML (Киселев, 2016)

При использовании агентного моделирования в диссертации Киселева И.Н., выполненной под руководством автора, (Киселев, 2016) показано, что для частного случая, когда все агенты содержат системы ОДУ, такой подход эквивалентен решению объединенной системы ОДУ методом Эйлера. При этом шаг численного метода равняется временному промежутку между сеансами обмена сообщениями между агентами.

В текущей версии ПК BioUML рассматривается наиболее простая схема, которая не учитывает возможность агентов перемещаться в пространстве, динамически рождаться или умирать при выполнении некоторого условия. Эти возможности могут быть добавлены в будущем, в частности в соответствии со спецификацией расширения SBML “dyn”.

3.9 Тестирование численного моделирования

Во время дискуссий по формату SBML обсуждался вопрос, что различные программы по численному моделированию динамики биологических систем, при использовании одних и тех же SBML файлов выдают различные численные результаты. Это объясняется как сложностью

самого формата SBML, так и необходимостью реализации численных решателей алгебро-дифференциальные уравнения с дискретными событиями и задержкой по времени. Для решения этой проблемы консорциумом SBML был разработан набор тестов, покрывающий все основные конструкции формата SBML (<https://synonym.caltech.edu/software/sbml-test-suite/>).

По результатам независимых сравнений (Maggioli et al., 2019), ПК BioUML единственный в мире симулятор биологических моделей, который полностью проходит все тесты на правильность численного моделирования биологических систем SBML Test Suite Core v3.3.0.

Корректность работы численного моделирования для стохастического моделирования была проверена на наборе тестов DSMTS (Discrete stochastic models test suite), доступном по адресу <https://github.com/darrenjw/dsmts>. Результаты тестирования показали, что точные методы не проходят только один тест (результат отличается менее чем на 5%), в то время как другой известный пакет для стохастической симуляции FERN (Erhard et al., 2008) не проходит два теста из 36.

3.10 Определение параметров моделей

Необходимым шагом при разработке моделей является решение обратной задачи – нахождение параметров модели на основе исходных экспериментальных данных. Для этого в ПК BioUML (рисунок 3.10.1) реализованы следующие оптимизационные методы (Кутумова, 2012; Kutumova et al., 2013a):

- SRES – эволюционная стратегия на основе стохастического ранжирования (Runarsson and Yao, 2000);
- MOCell – клеточный генетический алгоритм (Nebro et al., 2009);
- PSO – метод роя частиц (Sierra and Coello, 2005);
- glbSolve – детерминированный метод глобальной оптимизации (Björkman and Holmström, 1999);
- ASA – адаптивный метод имитации отжига (Ingber, 1996).

Важной особенностью модуля оптимизации в ПК BioUML является возможность одновременно оптимизировать параметры для нескольких экспериментов. Для этого используются:

- типы параметров
 - глобальный – значение параметра одно, и его значение оптимизируется по всем экспериментам одновременно;

- локальный – значение параметра оптимизируется индивидуально для каждого эксперимента;
- концепция состояния модели – позволяет для каждого эксперимента задать отличающиеся значения начальных условий, значения констант, а также позволяет изменить набор уравнений в ОДУ.

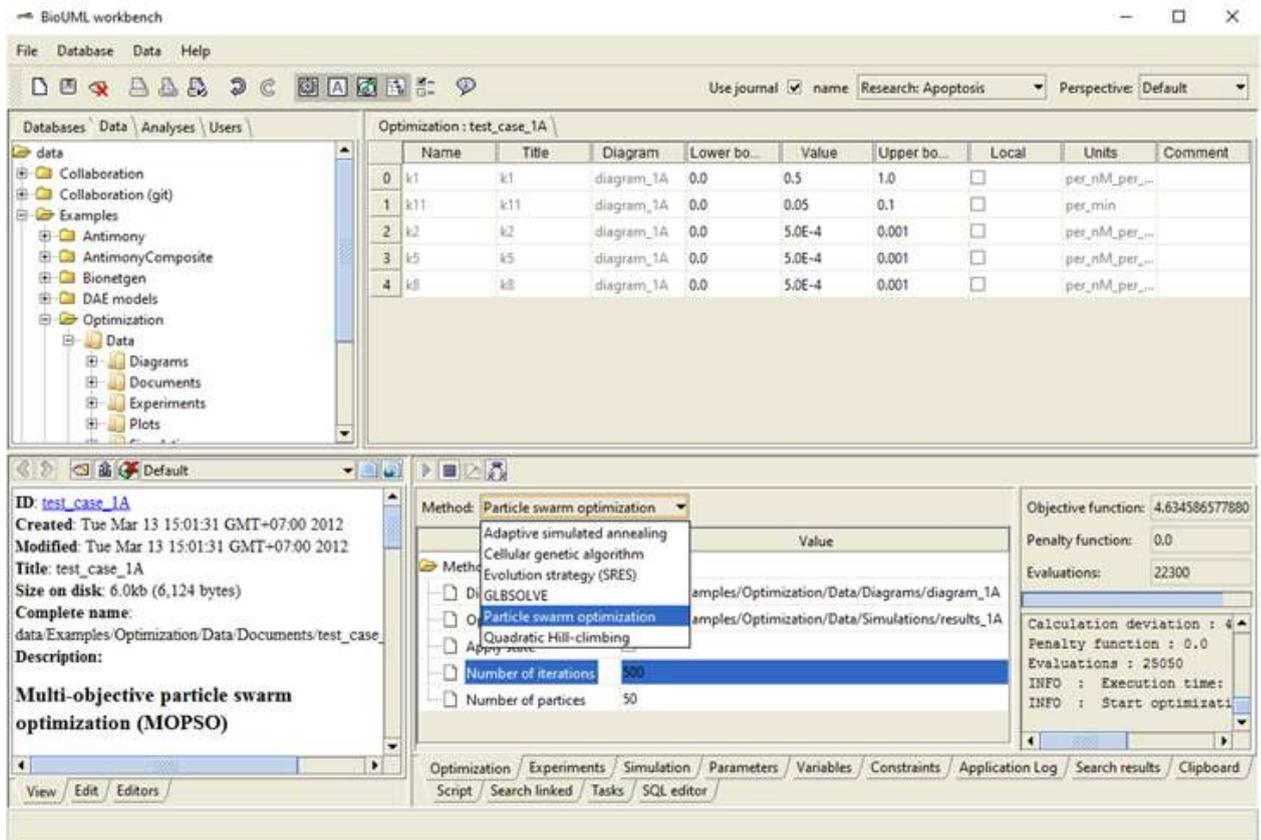


Рисунок 3.10.1 – Интерфейс пользователя для оптимизации параметров на основе экспериментальных данных: верхняя правая панель – набор параметров модели для оптимизации, их начальные значения и интервал допустимых значений; нижняя правая панель – выбор и настройка параметров оптимизационного метода

Другим важным вопросом при определении параметров модели является вопрос – можно ли идентифицировать значение параметра модели по заданным экспериментальным данным. Для ответа на этот вопрос в ПК BioUML реализован численный анализ идентифицируемости параметров (Raue et al., 2010) для выявления принципиально идентифицируемых, неидентифицируемых и частично идентифицируемых параметров.

3.11 Использование практик из разработки программного обеспечения

Методология разработки программного обеспечения активно развивается и достигла больших успехов. Наиболее часто сейчас для этого используются гибкие методики разработки (agile software development) — обобщающий термин для целого ряда подходов и практик, основанных на ценностях Манифеста гибкой разработки программного обеспечения (<https://www.agilealliance.org/agile101/>). Одна из таких технологий – экстремальное программирование (XP, eXtreme Programming). Мы постарались адаптировать предложенный в этой технологии практики для разработки моделей сложных биологических систем (Таблица 3.11.1; Колпаков и др., 2022).

Таблица 3.11.1 – Основные практики XP технологии и их адаптация к разработке моделей

Методология XP	Адаптация к разработке моделей
Планирование	
- истории пользователей	
набор коротких описаний для задания требований к программе	требования к модели – набор экспериментальных данных и феноменов, которые она должна воспроизводить
- небольшие частые релизы	
- проект разбит на итерации	Для каждой итерации задан набор экспериментальных данных/феноменов, которые должны быть воспроизведены
Кодирование	
- код должен быть написан в соответствии с согласованными стандартами	использование стандартов SBGN, SBML, Antimony
- сначала написание тестов	
- частая интеграция написанного кода вместе	
- коллективная собственность	использование git репозитория для коллективного владения кодом. Интеграция с git в ПК BioUML.
Тестирование	
Модульные тесты	Адаптация идеи JUnit тестов в ПК BioUML.

3.12 Тестирование моделей

В соответствии с предложенной технологией необходимой частью инструментария является система автоматического тестирования моделей и их модулей. В данной работе мы попытались адаптировать идеологию JUnit (<https://junit.org>) для тестирования биологических моделей. Для этого для каждой модели создается набор приемочных тестов (рисунок 3.12.1.а). В таблице 3.12.1 приведены возможные тесты, их параметры, и что именно проверяется. Пользователь может добавлять в набор отдельные тесты, конфигурировать их параметры (рисунок 3.12.1.б), и запускать набор приемочных тестов. Результат каждого теста будет показан в колонке "Status". Так же для удобства работы мы добавили специальную колонку "Plot", которая позволяет открыть график для переменной, указанной в тесте. На графике приведены полученные результаты численного моделирования (рисунок 3.12.1.в), а также экспериментальные значения (если были указаны в тесте) для соответствующей переменной.

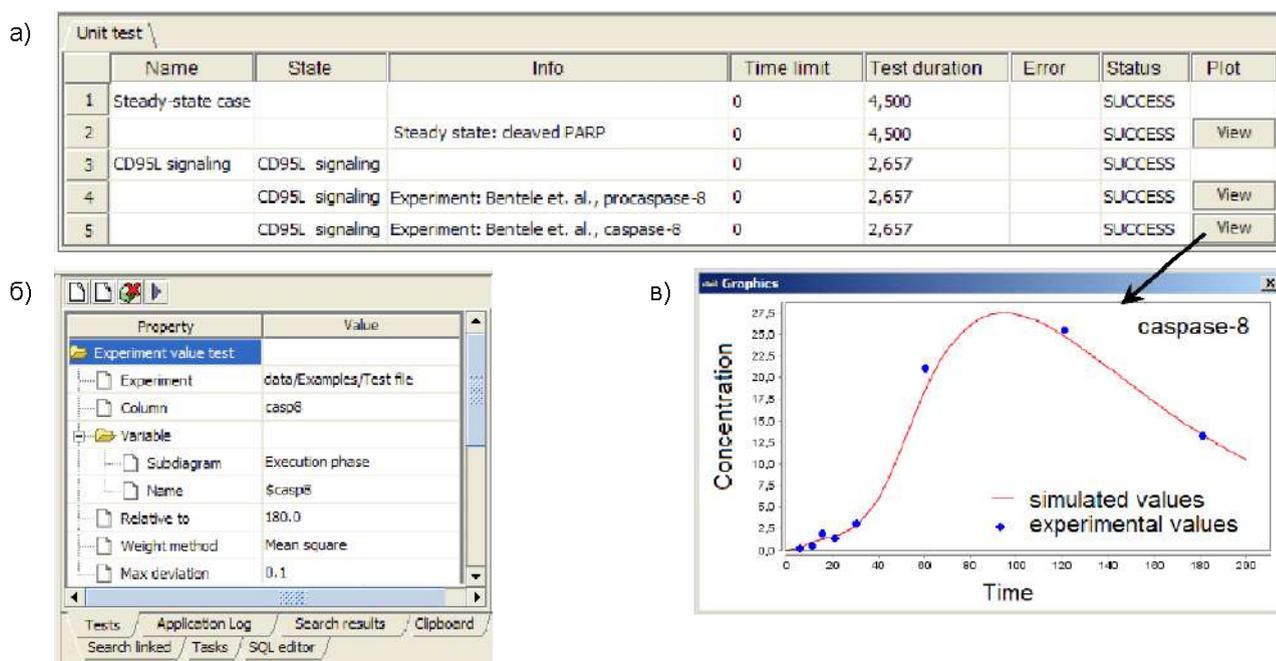


Рисунок 3.12.1 – Реализация агентного моделирования в ПК BioUML; а – приемочный тест, состоящий из набора простых тестов, б – параметры отдельного теста; в – результаты численного расчета и экспериментальные значения для соответствующей переменной

Таблица 3.12.1 – Типы тестов значений переменных модели

Тип теста	Параметры	Проверка
для всех тестов	x – имя переменной [t ₁ , t ₂] – интервал времени z – состояние модели	
1. стационарное значение	x* - значение переменной ε – максимально возможное отклонение	x(t) - x* < ε, t ∈ [t ₁ , t ₂]

2. допустимый интервал	x_{\min} – минимально допустимое значение переменной x_{\max} – минимально допустимое значение переменной	$x_{\min} \leq x(t) \leq x_{\max}, t \in [t_1, t_2]$
3. допустимое отклонение от экспериментальных данных	$x^*(t_i), i \in [1, n]$ – экспериментальные значения переменной x во временных точках $1, \dots, n$ ε – максимально возможное отклонение $f(x_1, x_2)$ – весовая функция, учитывающая расстояние между рассчитанными и экспериментальными значениями ε – максимально возможное отклонение	$\frac{\sum_{i=1}^n f(x(t_i) - x^*(t_i))}{n} < \varepsilon$

3.13 Создание математических моделей сложных биологических систем эволюционным путем

Разработка математических моделей сложных биологических систем вряд ли возможна с нуля. Такие модели можно разрабатывать только поэтапно, с постепенным усложнением. Этот процесс подобен эволюционному (табл. 3.13.1). Например, самая современная модульная модель сердечно-сосудистой и почечной систем человека для регуляции артериального давления (Kutumova, 2021) все еще содержит часть исходной модели Гайтона для общей регуляции кровообращения (Guyton, 1972), которой уже более 50 лет.

Платформа Sirius-web (раздел 8.3), которая сочетает в себе возможности ПК BioUML для моделирования и анализа данных, и GitLab для хранения, управления и документирования проектов, создают такую среду, на которой модели могут развиваться подобно живым существам.

Таблица 3.13.1 – Основные эволюционные процессы у живых существ и их аналогии при построении сложных математических моделей

Живые существа	Математические модели биологических систем
размножение	копирование из репозитория
жизнь	возможность запустить модель и воспроизвести опубликованные результаты
мутагенез	редактирование модели
рекомбинации	разбиение сложных моделей на модули, и создание из этих модулей новых
естественный отбор	прохождение тестов, соответствие экспериментальным данным

ГЛАВА 4. АНАЛИЗ И ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ БИМЕДИЦИНСКИХ ДАННЫХ

В обзоре литературы (раздел 1.4) были описаны основные инструменты, используемые для анализа биомедицинских данных. Следующий этап – поместить их в единый контекст, где они могут быть легко найдены и использованы. Для этого нужно формализовать, единообразно описать и структурировать существующие данные, программы и методы для их анализа и визуализации. Предложенная для этого архитектура и используемые программные решения были кратко описаны в разделах 2.2 и 2.3.

ПК BioUML обеспечивает широкие возможности для графического представления и анализа биомедицинских данных:

- **интеграция со многими биологическими БД;**
- **полнотекстовый поиск и визуализация результатов;**
- **анализ данных:**
 - собственные методы (200+, в основном анализ регуляторных районов генов);
 - интеграция с R/Bioconductor;
 - интеграция с Galaxy (и через нее с другими программами);
 - интеграция с Jupyter Notebook/Lab/Hub;
 - интеграция с Docker для использования различных методов анализа данных;
 - интеграция с CDK для хемоинформатики;
- **сценарии анализа данных:**
 - графическое представление и редактирование сценариев анализа данных;
 - поддержка языков сценариев CWL, WDL, Nextflow, Galaxy;
 - распараллеливание выполнения задач сценария по "облаку" или кластеру;
 - набор готовых сценариев для анализа NGS данных;
- **визуализация данных:**
 - редактирование и просмотр диаграмм биологических путей и других диаграмм;
 - встроенный геномный браузер;
 - различные графики, тепловые карты и т.п. (через интеграцию с R, Jupyter);
 - визуализация филогенетических деревьев, используя JavaScript библиотеку phylotree.js и ETE Toolkit (внутри Jupyter Notebook);
 - визуализация и редактирование структурных формул, используя CDK и JChemPaint;
 - визуализация трехмерных структур макромолекул и химических веществ, используя JMol и JSMol;
- **воспроизводимые исследования (reproducible research)** – для этого используются:

- журнал исследований – записываются все шаги по анализу данных;
- для каждого результата записываются мета-данные – в ходе какого анализа и с какими параметрами был получен данный результат;
- сценарии анализа данных и Jupyter notebooks обеспечивают возможность воспроизвести результаты;
- **совместные исследования (Collaborative research):**
 - портал Biosotore – создание проектов, управление правами доступа к совместным данным;
 - совместное редактирование диаграмм и сценариев анализа данных (подобно Google docs)
- **возможность локальной установки платформы и в "облаке"**. Две версии платформы:
 - настольное Java приложение;
 - серверное Java приложение, в качестве клиента выступает обычный веб-браузер.

На основе ПК BioUML создано несколько специализированных решений для анализа биомедицинских данных (см. главу 8):

- платформа geneXplain;
- платформа Genome Enhancer;
- биоинформатическая платформа для хранения, анализа и графического представления данных, полученных при одномолекулярном секвенировании ДНК, разработанная в рамках НИР "Разработка и испытания биоинформатической платформы для хранения, анализа и графического представления данных, полученных при одномолекулярном секвенировании ДНК";
- геномный браузер и программные решения для анализа генетических данных в Национальной базе генетической информации, разработанные в рамках НИР "Исследование, обоснование и выбор программных решений для визуализации генетических данных и обеспечения инструментов для работы с генетической информацией в «Национальной базе генетической информации» (контракт №43-03/21/44/152 от 13.12.2021);
- биоинформатическая платформа, интегрирующая доступные данные и полученные экспериментальные данные для предсказания некодирующих регуляторных участков генома для генетического редактирования с целью изменения активности заданного гена, разработанная в рамках проекта «Регуляторная транскриптомика для оценки потенциала модификации генома и изучения регенерации в животноводстве».

Данные программные решения используют ядро и архитектуру ПК BioUML и отличаются только набором используемых БД, программ и сценариев анализа биомедицинских данных и документацией.

В рамках диссертационной работы невозможно подробно описать все программы и сценарии, входящие в ПК BioUML. Например, документация по методам и сценариям анализа биомедицинских данных для платформы geneXplain занимает, более 500 страниц:

https://genexplain-platform-documentation.readthedocs.io/_/downloads/en/latest/pdf/

4.1 Интерфейс для анализа биомедицинских данных

На рисунке 4.1.1 показан интерфейс ПК BioUML для анализа биомедицинских данных. Сначала пользователь может создать проект и загрузить исходные данные секвенирования. После этого он может выбрать одну из перспектив (рабочее пространство) для анализа данных или группу методов или сценариев. Также возможен поиск методов анализа и сценариев по ключевым словам.

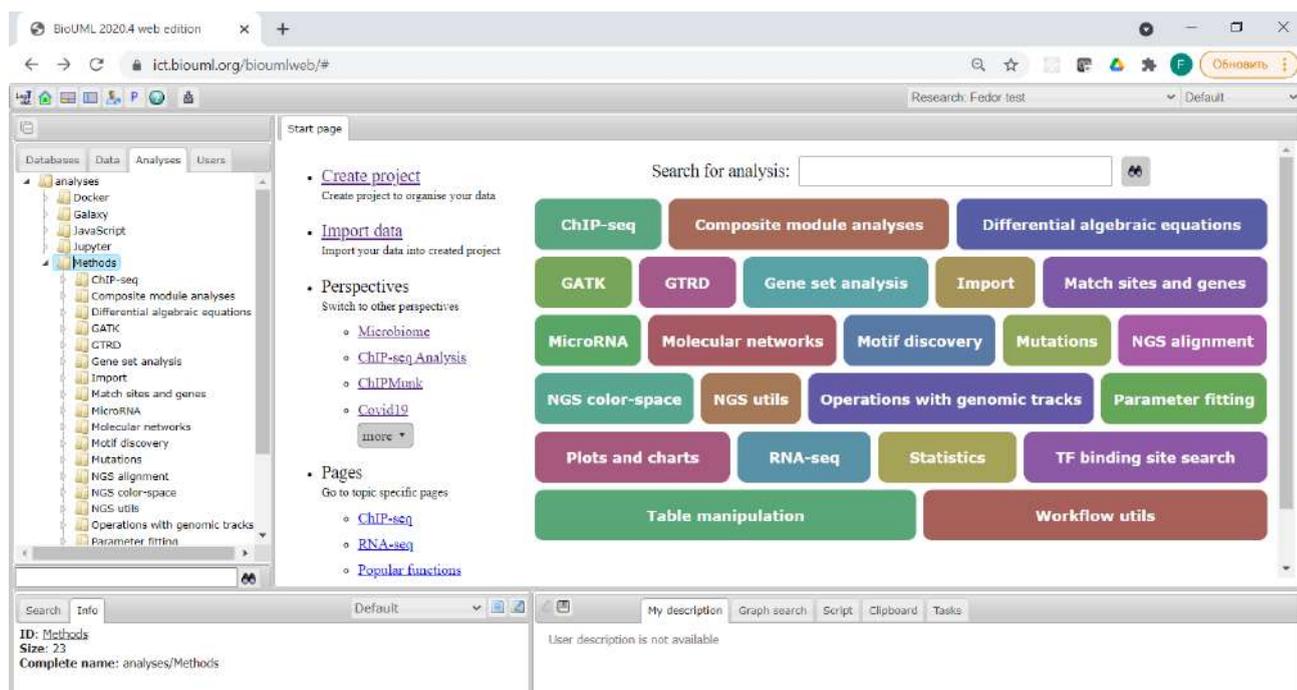


Рисунок 4.1.1 – Интерфейс ПК BioUML для анализа биомедицинских данных

На рисунке 4.1.2 представлен интерфейс ПК BioUML для запуска сценария анализа данных. В верхней части экрана расположена форма, в которой пользователь должен указать параметры для запуска сценария. В нижней части экрана представлена диаграмма сценария. Входные параметры из формы на ней представлены зелеными прямоугольниками. Синие прямоугольники – методы анализа. Желтые шестиугольники – промежуточные результаты.

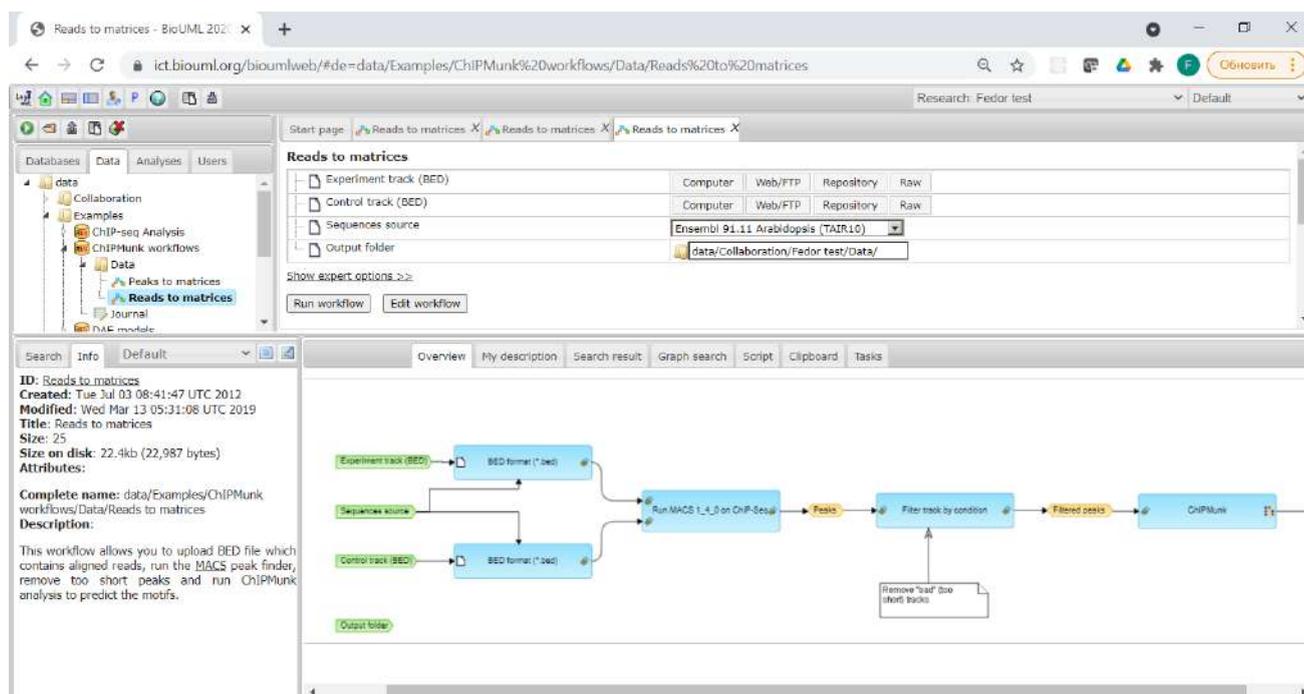


Рисунок 4.1.2 – Интерфейс ПК BioUML для запуска сценария анализа данных

4.2 Сценарии анализа данных

4.2.1 WDL

В ПК BioUML реализована возможность описания и запуска сценариев на языке WDL. Для этого используется плагин `biouml.plugins.wdl`. Скрипт WDL можно создать с помощью действия “New WDL script” в контекстном меню. Редактирование скрипта осуществляется во встроенном редакторе (FCKEditor) с подсветкой синтаксиса (рисунок 4.2.1). Также WDL скрипт отображается в графическом виде во вкладке “WDL diagram” (рисунок 4.2.2). Графическое представление WDL скрипта создается автоматически посредством генерации графа выполнения на основе скрипта, где вершинами являются задачи сценария и переменные скрипта, а ребра отражают передачу данных. Для визуализации сгенерированного графа применяется алгоритм иерархического упорядочивания.

Чтобы запустить WDL скрипт используется действие “Run WDL” в контекстном меню (рисунок 4.2.3). При этом открывается форма для заполнения параметров запуска скрипта. Параметры запуска можно задать двумя способами: в виде json файла – как при запуске WDL из командной строки, или заполнив форму параметров. Форма параметров создается автоматически на основе анализ WDL скрипта и более удобна при заполнении параметров вручную. Для исполнения WDL скриптов используется система управления сценариями

Cromwell (<https://cromwell.readthedocs.io/en/stable/>) в виде отдельного внутреннего сервиса. Взаимодействие с сервисом Cromwell происходит по REST API.

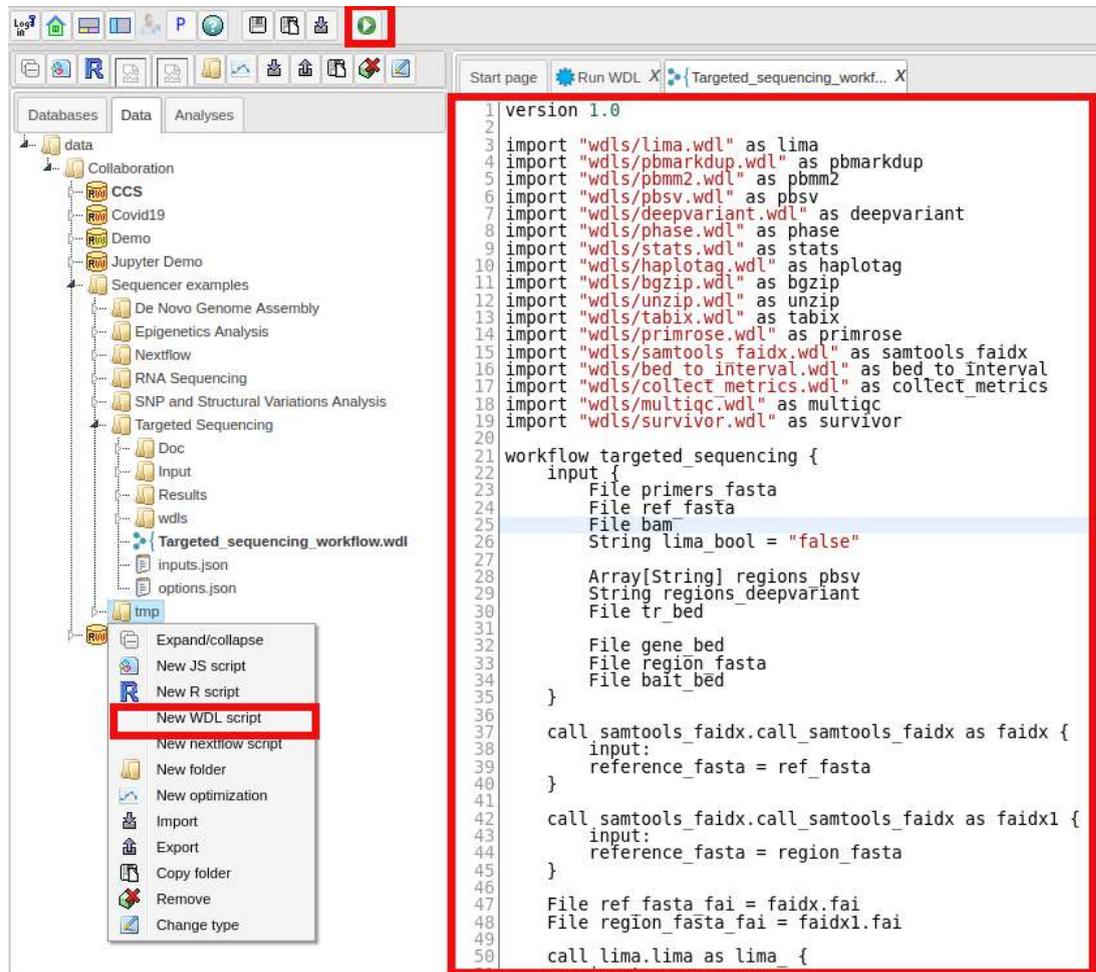


Рисунок 4.2.1 – Интерфейс для создания и редактирования WDL скриптов

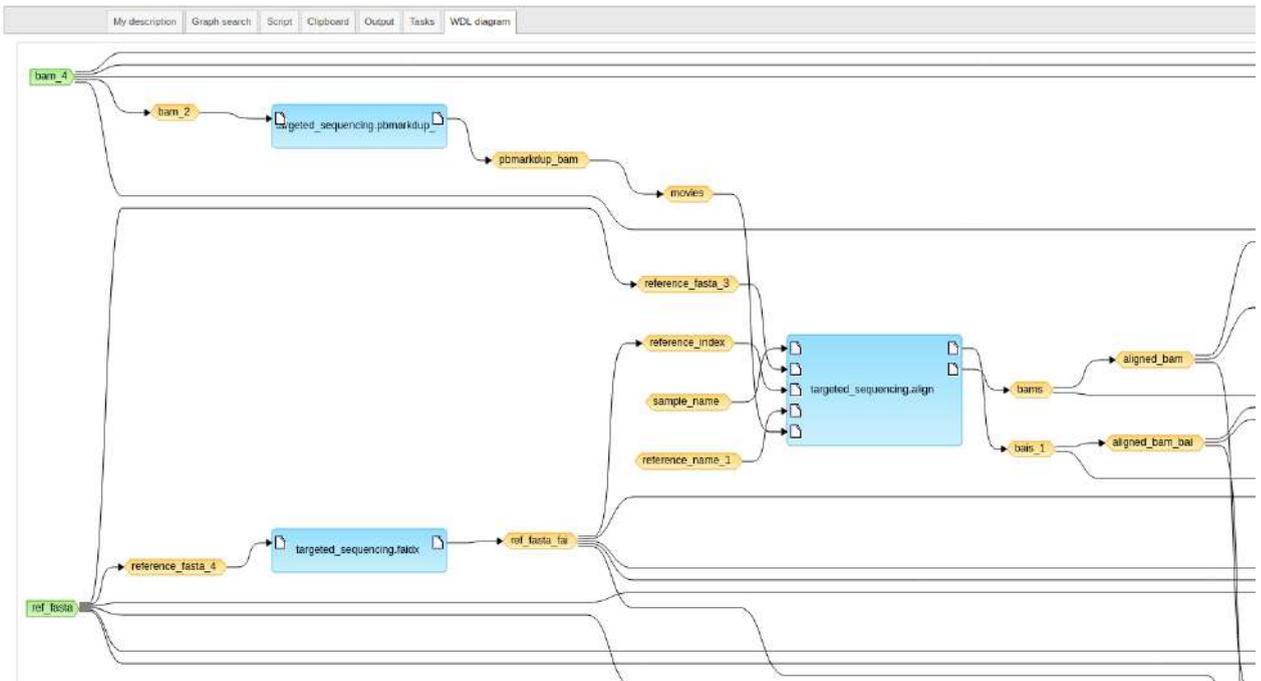


Рисунок 4.2.2 – Представление WDL скрипта в графическом виде

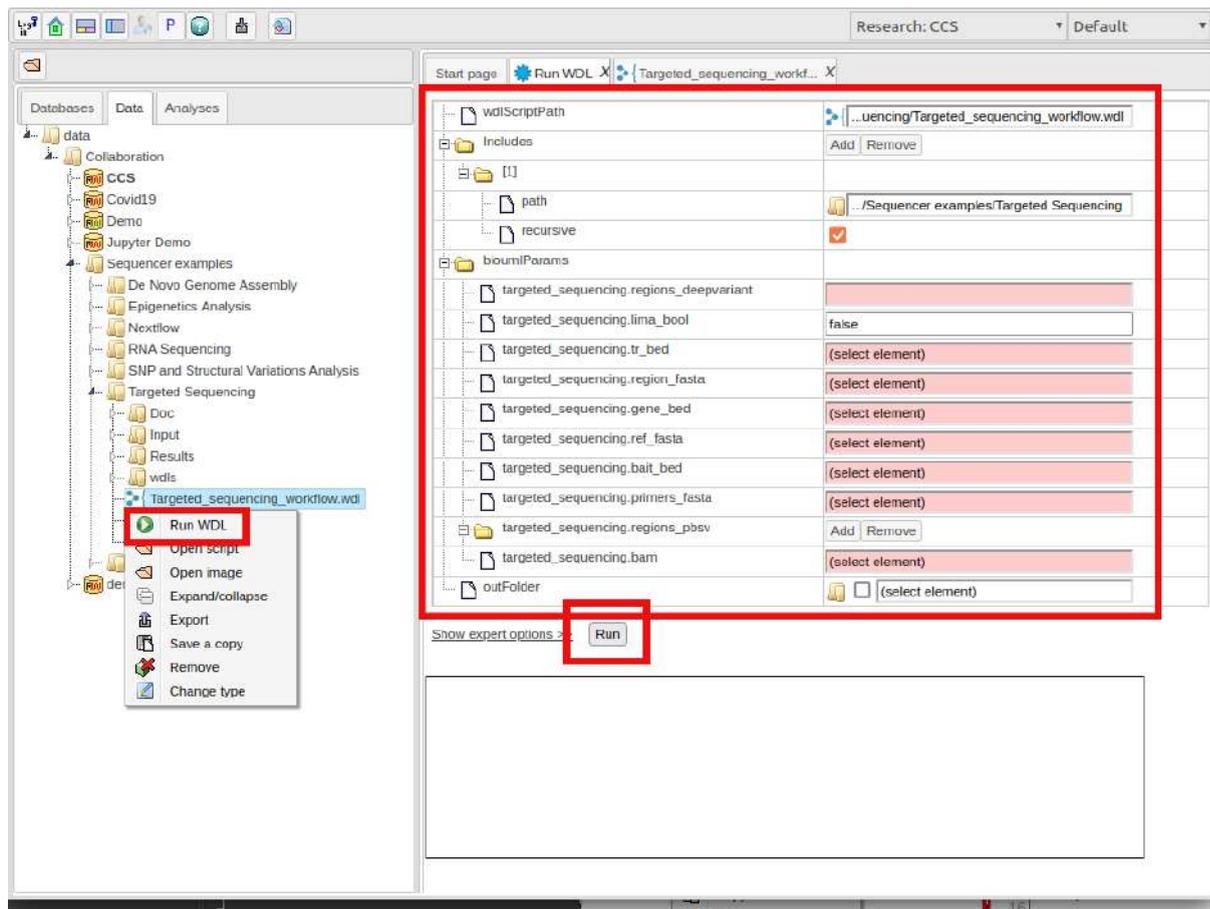


Рисунок 4.2.3 – Интерфейс для запуска WDL скриптов и задания параметров запуска

Помимо запуска комплексных сценариев в ПК BioUML доступен запуск отдельных программ, входящих в данный сценарий. На рисунке 4.2.4 приведен интерфейс запуска программы pbmm2. В данном случае программа реализована в виде Docker-образа, и процесс её запуска, а также набор входных/выходных параметров, описаны в виде WDL-скрипта.

Таким образом, программы и сценарии, описанные на языке WDL, могут быть использованы внутри сценариев ПК BioUML.

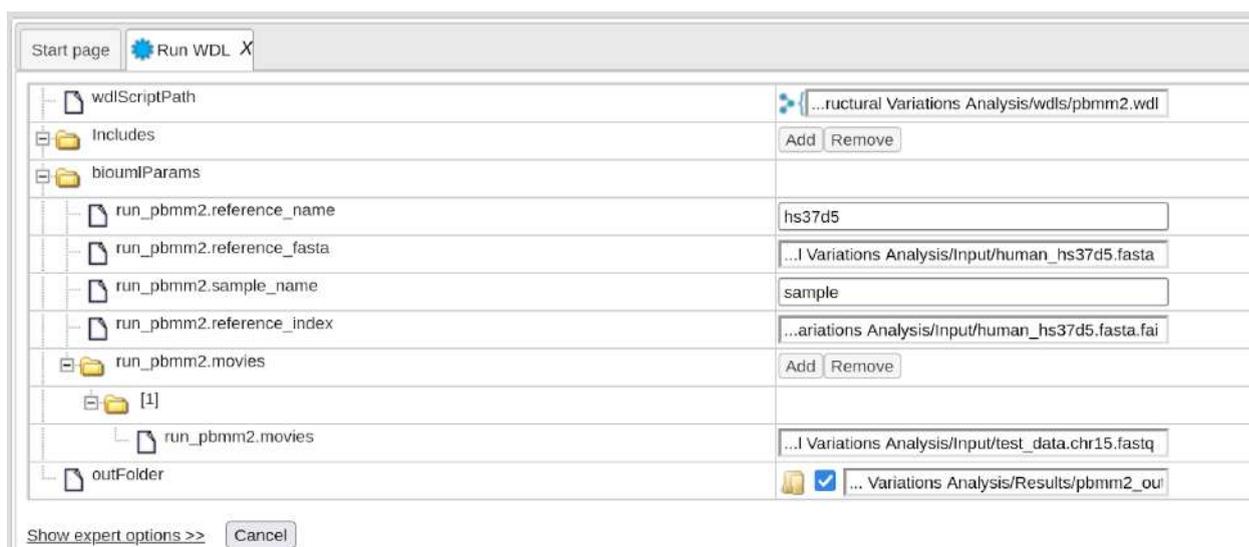


Рисунок 4.2.4 – Интерфейс запуска программы pbmm2, реализованной в виде Docker-образа и описанной в виде WDL-скрипта, на платформе BioUML

4.2.2 Nextflow

В ПК BioUML реализована возможность описания и запуска сценариев на языке Nextflow. Скрипт Nextflow можно создать с помощью действия “New nextflow script” в контекстном меню. Редактирование скрипта осуществляется во встроенном редакторе с подсветкой синтаксиса. Запуск скрипта осуществляется кнопкой “Run” на панели инструментов. Логи выполнения отображаются во вкладке “Output” (рисунок 4.2.5). В Nextflow скрипте можно ссылаться на файлы из репозитория, переменная “\$baseDir” ссылается на директорию где расположен скрипт. Результирующие файлы автоматически импортируются в папку \$baseDir/work. Запуск скриптов Nextflow осуществляется в изолированной среде с использованием контейнеров Docker.

The screenshot displays the Nextflow web interface. On the left, a file tree shows a project structure with a 'Nextflow' folder. A context menu is open over this folder, with 'New nextflow script' highlighted. The main area shows a Nextflow DSL2 script for splitting and reversing FASTA files. The script includes parameters, two processes (splitSequences and reverse), and a workflow definition. Below the script, the execution output is shown, indicating successful completion of both processes and the resulting output 'TTTTTTAAAAA'.

```

1 nextflow.enable.dsl = 2
2
3 params.in = "$baseDir/data/sample.fa"
4
5
6 /*
7  * Split a fasta file into multiple files
8  */
9 process splitSequences {
10     input:
11     path 'input.fa'
12
13     output:
14     path 'seq_*'
15
16     """
17     awk '/^>/{f="seq_"++d} {print > f}' < input.fa
18     """
19 }
20
21 /*
22  * Reverse the sequences
23  */
24 process reverse {
25
26     input:
27     path x
28
29     output:
30     stdout
31
32     """
33     cat $x | rev
34     """
35 }
36
37 /*
38  * Define the workflow
39  */
40 workflow {
41     splitSequences(params.in) \
42     | reverse \
43     view
44 }
45

```

Search Info Default My description Graph search Script Clipboard Output

ID: [Nextflow](#)
Size: 1
Complete name: data/Collaboration/Sequencer examples/Nextflow

```

NEXTFLOW ~ version 22.12.0-edge
Launching 'my_script.Nextflow' [pensive_fourier] DSL2 - revision: 654ef83e7

executor > local (1)
[20/e5c01a] process > splitSequences [ 0%] 0 of 1
[- ] process > reverse -

executor > local (2)
[20/e5c01a] process > splitSequences [100%] 1 of 1 ✓
[1b/6a984d] process > reverse [100%] 1 of 1 ✓
1>
TTTTTTAAAAA

```

Рисунок 4.2.5 – Интерфейс для работы с Nextflow скриптами

4.2.3 Интеграция с Galaxy и Galaxy Toolshed

Для использования анализов Galaxy осуществлена интеграция на уровне REST API с Galaxy сервером (рисунок 4.2.6). Посредством REST API Galaxy осуществляется получение списка анализов, получение описания анализа и его параметров, запуск анализа, отслеживание статуса задач, передача информации о загруженных файлах. Обмен файлами с Galaxy сервером осуществляется посредством общей директории на сетевом диске.

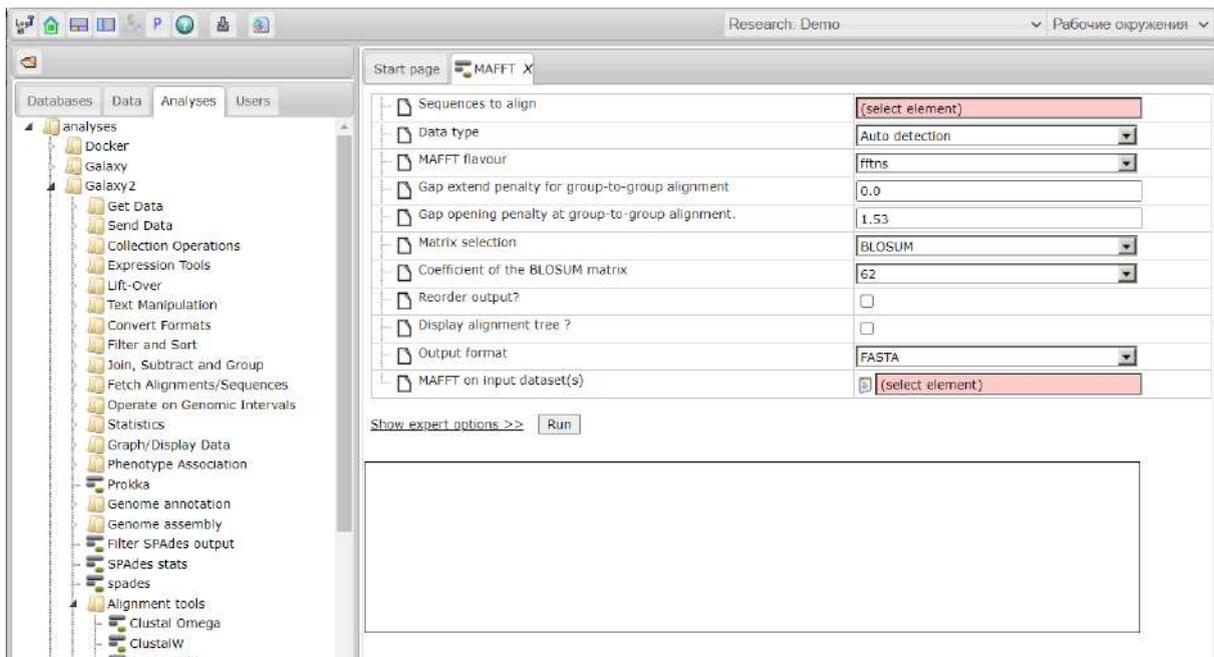


Рисунок 4.2.6 – Интерфейс доступа к анализам Galaxy. Показан список анализов, Параметры анализа MAFFT и его описание

Для администрирования Galaxy сервера была осуществлена интеграция на уровне REST API с Galaxy сервером (рисунок 4.2.7). Посредством REST API Galaxy осуществляется получение списка источников репозиториях из Galaxy Toolshed, классификации репозиториях по темам, получение списка репозиториях из заданной категории, получение информации о репозитории, поиск по репозиториях, установка и удаление репозиториях (рисунок 4.2.8).

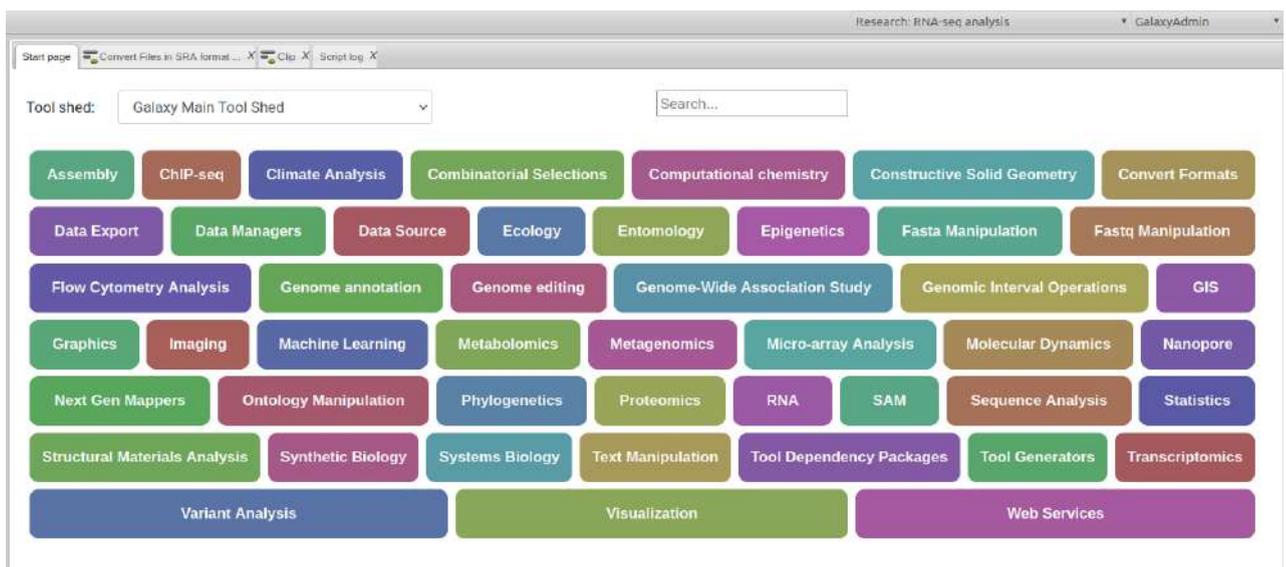


Рисунок 4.2.7 – Интерфейс для администрирования Galaxy. Показан Tool Shed, классификация репозиториях по темам, поиск по репозиториях

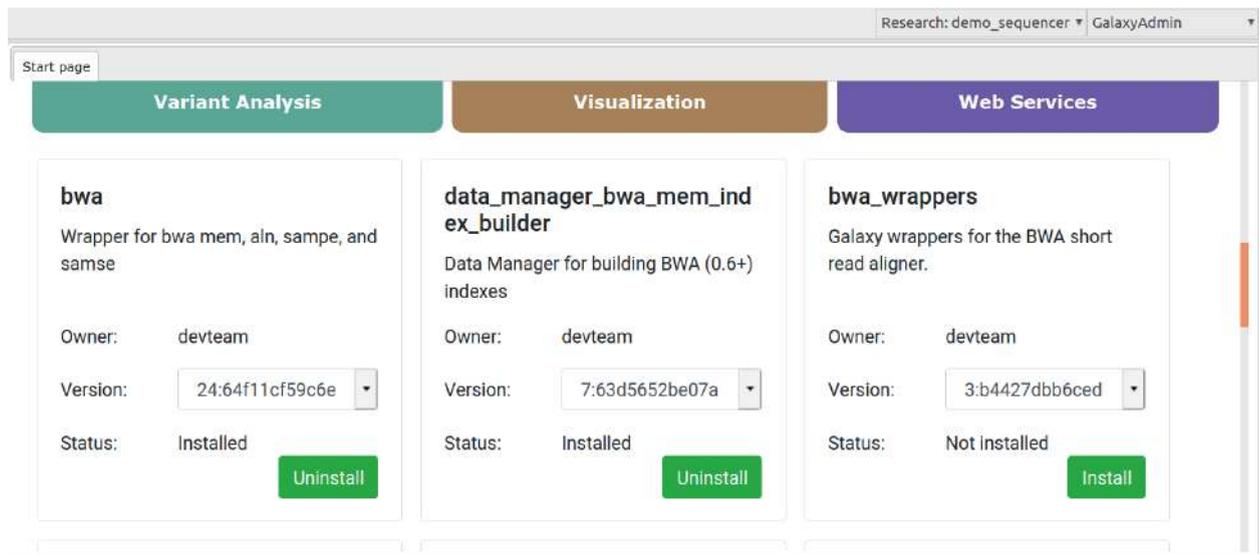


Рисунок 4.2.8 – Интерфейс установки/удаления анализов/репозитория из Galaxy Toolshed на платформе BioUML

После установки анализа из Galaxy Toolshed он появляется в списке анализов во вкладке “Analysis”, как показано на рисунке 4.2.6 Для всех анализов в ПК BioUML доступен универсальный графический интерфейс установки начальных параметров и запуска (рисунок 4.2.9). Таким образом на платформе осуществляется поддержка запуска программ и сценариев, размещенных в Galaxy Toolshed.

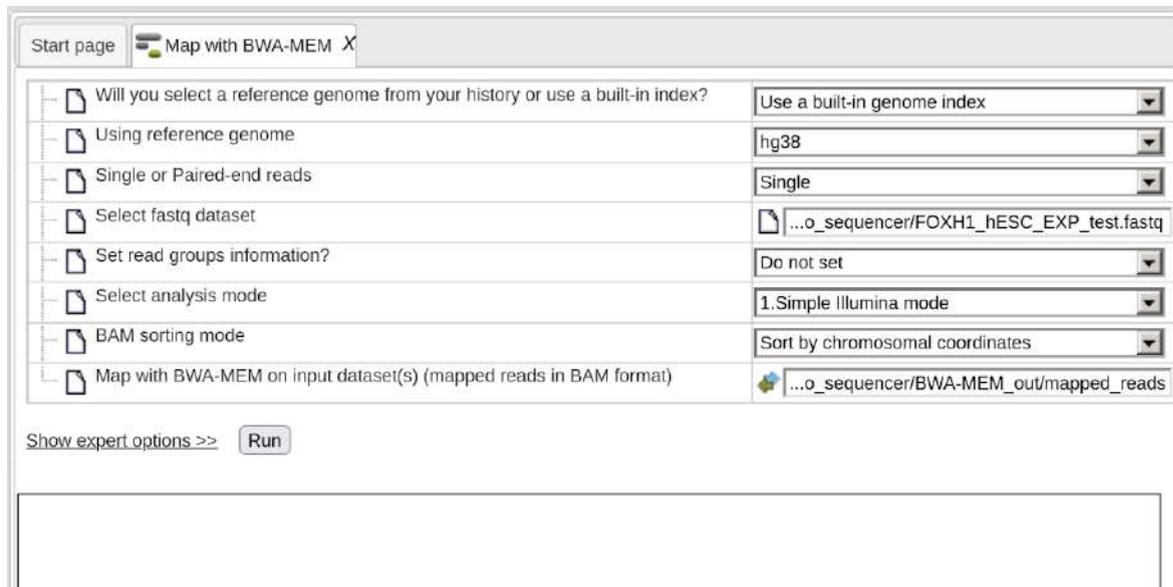


Рисунок 4.2.9 – Интерфейс анализа “Map with BWA-MEM”, установленного на платформу BioUML из Galaxy Toolshed

4.3 Геномный браузер

На рисунке 4.3.1 представлен интерфейс геномного браузера, входящего в состав ПК BioUML. Структурная и функциональная аннотация генетических данных в базе представляется в виде отдельных треков. Например, для установленной версии БД Ensembl генома человека в ней содержатся следующие треки:

- Karyotype – с информацией о кариотипе, включая данные о номерах хромосом, плечах, локусах, G-бэндинге;
- GC content – визуализация GC состава ДНК: красным цветом обозначаются нуклеотиды А и Т, а синим – нуклеотиды G и С, проставлены номера хромосом и позиций в них;
- Genes – с информацией о генах: названия, ID в БД Ensembl и RefSeq, их расположение, количество экзонов и транскриптов, GC-состав;
- Repeats – с информацией о повторах: номера хромосом, позиций в них, класс повтора и его тип;
- Variations – с информацией об однонуклеотидных вариантах (SNV): ID сайта, название варианта вида rs..., тип (полиморфизм, инсерция, делеция, и т.д.), вариант нуклеотида в аллеле;
- Transcripts – с информацией о транскриптах : ID сайта, тип (кодирует белок или нет), номер хромосомы, позиции в ней, координаты экзонов, Ensembl ID и координаты транслируемых частей транскрипта.

Для просмотра соответствующей генетической информации в графическом виде необходимо выбрать базу EnsemblHuman, открыть папку Tracks и выбрать треки для отображения нужной информации.

Более подробно работа с геномным браузером описана в руководстве оператора для ПК BioUML:

https://docs.google.com/document/d/1ubx4ShM_uZQh21hPXZbZrHmZmJ2K47SznP3MhaQcCU/

Архитектура и детали реализации геномного браузера описаны в работе Valeev et al., 2013.

The screenshot displays the BioUML genome browser interface. The browser address bar shows the URL: `ict.biouml.org/bioumlweb/#de=databases/EnsemblHuman85_38/Sequences/chromosomes%20GRCh38&pos=1:6614-46081`. The main window shows a genomic track for chromosome 14, with the position set to 1:6614-46081. The track displays various genomic features, including genes (e.g., TAR1, ERVL-E-int, MLT1E1A-int, L1MA8), repeats, and variations. The interface includes a navigation sidebar on the left, a search bar at the top, and a bottom panel with search and track management options.

Search Info Default

My description Search result Graph search Script Clipboard Tasks Sites **Tracks** Track finder

ID: Sequences
Size: 1
Complete name: databases/EnsemblHuman85_38/Sequences

Track	Visible	Options
ExtendedGeneTrack	<input checked="" type="checkbox"/>	Options
Repeats	<input checked="" type="checkbox"/>	Options
Variations	<input checked="" type="checkbox"/>	Options

Рисунок 4.3.1 – Интерфейс геномного браузера, входящего в состав платформы BioUML

ГЛАВА 5. БАЗА ДАННЫХ GTRD

5.1 Постановка задачи

Понимание механизмов регуляции транскрипции у эукариот является одной из центральных задач современной биологии. В регуляции транскрипции задействовано множество компонентов, основными из которых являются: структура хроматина (открытый или закрытый), его модификации (метилирование и ацетилирование гистонов) и пространственная структура, эпигенетические модификации ДНК (прежде всего метилирование), а также регуляторные РНК, в первую очередь микроРНК и энхансерная РНК.

Современные методы высокопроизводительных исследований производят огромные массивы информации по этим ключевым факторам регуляции транскрипции, в частности:

- РСТФ – основные экспериментальные методы: ChIP-seq, ChIP-exo, ChIP-nexus;
- доступность хроматина – DNase-seq, ATAC-seq, FAIRE-seq;
- гистоновые модификации – ChIP-seq;
- сайты метилирования – WGBS, RRBS, methyl array;
- сайты инициации транскрипции – CAGE-seq;
- уровень экспрессии соответствующих генов (CAGE, RNA-seq).

Эти данные были получены как отдельными лабораториями, так и большими международными консорциумами:

- ENCODE – энциклопедия (регуляторных) ДНК элементов (Luo et al., 2020);
- FANTOM 5 – функциональная аннотация и уровень экспрессии функциональных ДНК участков (Kawaji et al., 2017);
- Roadmap Epigenomics Project – данные по эпигеномике и метилированию ДНК (Roadmap Epigenomics Consortium., 2015).

Исходные данные таких экспериментов представлены в специальных хранилищах данных, основными из которых являются SRA (Leinonen et al., 2010) и GEO (Clough and Barrett, 2016).

Однако эти данные недостаточно интегрированы друг с другом, что существенно затрудняет их совместное использование как для понимания механизмов регуляции транскрипции, так и для решения практических задач – например, предсказание возможных эффектов одиночных нуклеотидных замен (SNV) в регуляторных районах генов.

Основные проблемы при интеграции таких данных:

- данные были получены на разных клеточных линиях и типах, под воздействием различных факторов в разных временных точках;

- следствием первой проблемы является недостаточность экспериментальных данных для заданной клеточной линии (типа) при различных экспериментальных воздействиях;
- разные процедуры анализа данных одного типа (например, ChIP-seq экспериментов по связыванию ТФ);
- наличие разных методов компьютерной обработки экспериментальных данных, при этом нет однозначно лучшего выбора. Например, для поиска РСТФ наилучшие результаты, как правило, показывают методы MACS2 (Zhang et al., 2008) и GEM (Guo et al., 2012), при этом количество найденных РСТФ и их размеры могут существенно отличаться (Thomas et al., 2014; Kolmykov et al., 2019, 2020);
- для данных одного и того же типа существуют варианты экспериментальных методов, которые невозможно единообразно обработать. Например, для идентификации ССТФ используется ChIP-seq в вариантах single-end, paired-end, ChIP-exo, ChIP-nexus и для каждого из них необходимо использовать свой метод для выявления пиков выравниваний;
- надежность и достоверность данных – сравнительный анализ и контроль качества данных (например в проекте ENCODE, БД ReMap) показывает, что данные отдельных экспериментов существенно отличаются от сходных экспериментов (Kolmykov et al., 2019).

Таким образом, необходимо создание специализированной БД для интеграции большинства омиксных данных по регуляции транскрипции эукариот.

5.2 Методы и подходы

В основе создания БД GTRD лежат 4 основные идеи:

- 1) единообразная аннотация, контроль качества и обработка большого объема экспериментальных NGS данных по регуляции транскрипции;
- 2) одновременное использование нескольких методов анализа NGS данных и последующий мета-анализ с целью наиболее достоверного выявления регуляторных элементов;
- 3) интеграция результатов из различных типов NGS экспериментов в контексте заданного клеточного типа и условий;
- 4) автоматический анализ интегрированных NGS данных по регуляции транскрипции, вышеописанными методами, и сохранение результатов в БД GTRD. Данный подход повторяет идеологию проекта ENSEMBL, которая объединяет как экспериментальные данные по аннотации генома, так и компьютерные предсказания.

Таким образом, GTRD должна стать аналогом ENSEMBL, но в плане регуляции транскрипции. Ниже мы рассмотрим каждый из этих пунктов подробнее.

5.2.1 Отбор видов эукариот

Первая версия БД GTRD включала данные для человека и мыши для которых имеется максимальное количество NGS данных по регуляции транскрипции (Yevshin et al., 2017). Далее, на основе количества доступных ChIP-seq экспериментов в БД GEO (рисунок 5.2.1) были отобраны еще 7 видов организмов для включения в БДGTRD (Yevshin et al., 2017). В 2022 был начат сбор данных для курицы *Gallus gallus*.

Таким образом, на данный момент в БД GTRD собираются данные для 10 видов организмов: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Gallus gallus*, *Danio rerio*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* и *Arabidopsis thaliana*.

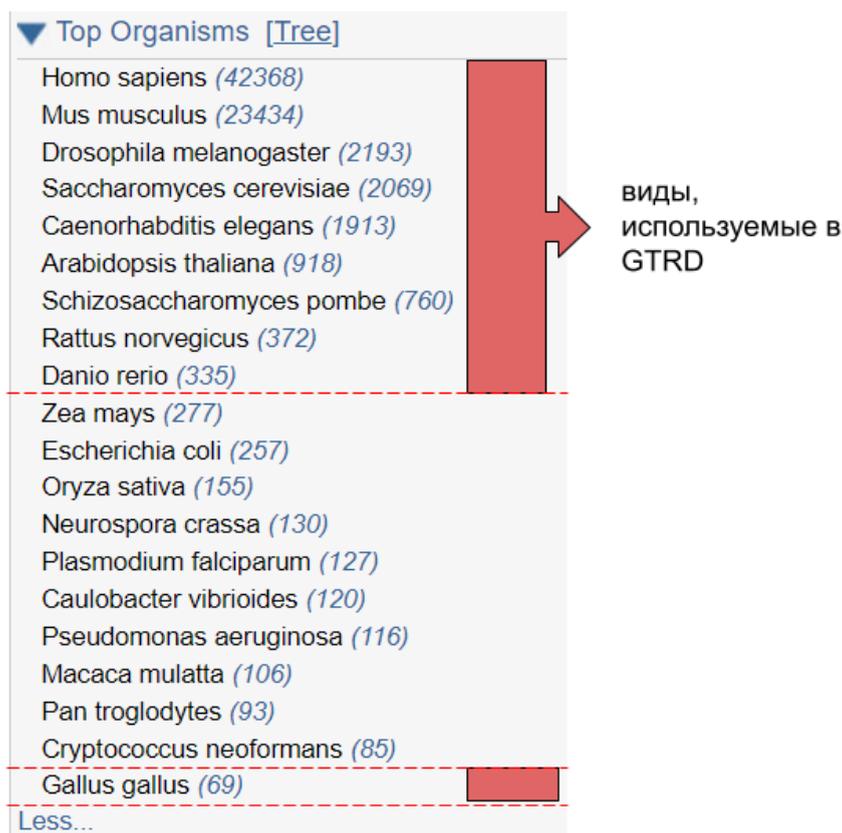


Рисунок 5.2.1 – Количество ChIP-seq экспериментов, найденных в БД GEO (<https://www.ncbi.nlm.nih.gov/gds/?term=ChIP-seq>, дата 2022.06.12)

5.2.2 Аннотация данных

Рабочий процесс для единообразной аннотации и анализа NGS данных по регуляции транскрипции в БД GTRD представлен на рисунке 5.2.2.

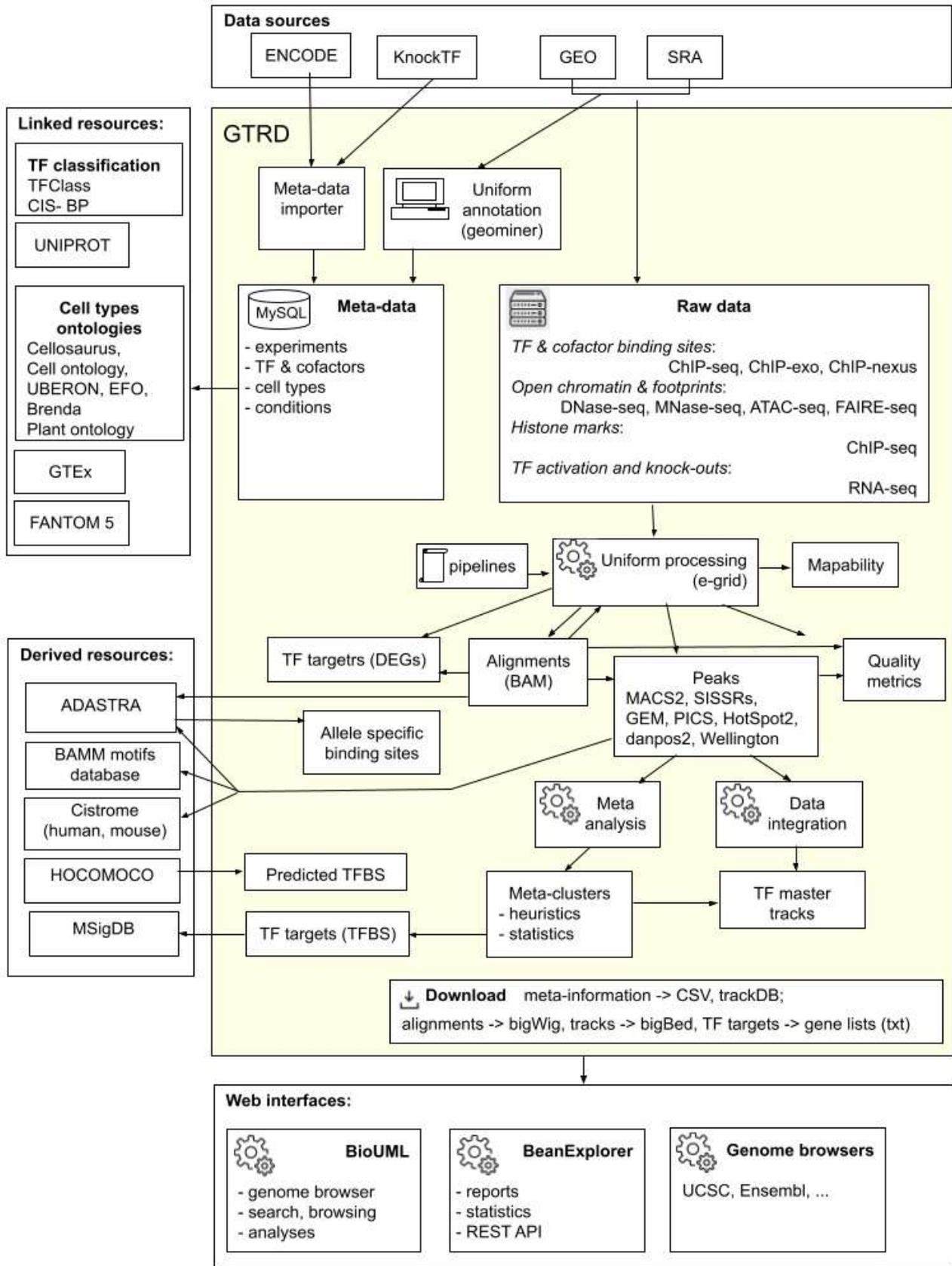


Рисунок 5.2.2 – Текущий рабочий процесс для единообразной аннотации и анализа NGS данных в БД GTRD (Kolmykov et al., 2021)

Таблица 5.2.1 – Статистика БД GTRD (версия 23.07) по объему аннотированных NGS данных по регуляции транскрипции

	Транскрипционные факторы и кофакторы		Гистоновые модификации и ChIP-seq	Открытый хроматин				Метилирование	Экспрессия при модификации транскрипционных факторов	
	ChIP-seq	ChIP-exo		DNase-seq	ATAC-seq	MNase-seq	FAIRE-seq		WGBS	RNA-seq
<i>Mus musculus</i>	21 456	134	16 284	928	15 368	607	144	85	5 326	1 018
<i>Homo sapiens</i>	29 797	2 021	22 987	1 898	14 190	247	346	117	3 664	1 816
<i>Rattus norvegicus</i>	469	2	680	12	165	4	0	0	20	55
<i>Gallus gallus</i>	107	0	259	15	116	0	1	0	0	119
<i>Drosophila melanogaster</i>	3 680	51	2 056	47	767	331	90	0	108	0
<i>Caenorhabditis elegans</i>	2 058	0	868	94	127	46	0	0	0	0
<i>Arabidopsis thaliana</i>	1 106	0	2 162	14	285	143	22	1	12	0
<i>Saccharomyces cerevisiae</i>	3 482	1 120	1 862	9	192	1358	102	0	8	0
<i>Danio rerio</i>	122	0	597	0	433	7	4	0	17	0
<i>Schizosaccharomyces pombe</i>	670	13	919	0	6	174	0		4	0
Всего (версия 23.07)	62 947	3 341	48 674	3 017	31 649	2 944	709	203	9159	3 008

Таблица 5.2.2 – Статистика по единому словарю клеточных линий и типов в БД GTRD и его сопоставление со специализированными БД клеточных линий и типов

Species	Cell Ontology (CL)	UBER-anatomy ontology (UBERON)	Cellosaurus (CVCL)	Brenda tissue ontology (BTO)	Experimental factor ontology (EFO)	Plant ontology (PO)	Unmatched
<i>Arabidopsis thaliana</i>	1	5	0	105	0	21	4
<i>Caenorhabditis elegans</i>	1	90	0	3	0	0	1
<i>Danio rerio</i>	13	55	2	10	1	0	1
<i>Drosophila melanogaster</i>	33	188	14	29	13	0	3
<i>Homo sapiens</i>	1 155	332	1375	528	839	0	23
<i>Mus musculus</i>	1 422	788	168	322	115	0	53
<i>Rattus norvegicus</i>	29	51	12	15	1	0	2
<i>Saccharomyces cerevisiae</i>	0	323	0	15	0	0	0
<i>Schizosaccharomyces pombe</i>	0	188	0	0	0	0	0
Итого	2 654	2 020	1 571	1 027	969	21	87

На первом этапе производится сбор информации о доступных экспериментах (мета-данные) данных по регуляции транскрипции – ChIP-seq, ChIP-exo, DNase-seq, MNase-seq, ATAC-seq, FAIRE-seq, Histone-ChIP-seq, RNA-seq, CAGE, а также данных по метилированию. Статистика по объему аннотированных NGS данных для версии 23.07 БД GTRD приведена в таблице 5.2.1.

В зависимости от источника данных (рисунок 5.2.2), применялись разные подходы: хорошо структурированная информация из проекта ENCODE собиралась автоматически (программно), в то время как для аннотации данных из GEO была создана специальная программа Geominer, позволяющая аннотировать данные с участием человека. Данная программа принимает на вход метаданные, описывающие серию экспериментов (GSE), в формате MINiML (MIAME Notation in Markup Language) и предоставляет полученную информацию аннотатору посредством графического интерфейса. Помимо этого Geominer имеет набор полей, заполняемых аннотатором на основании имеющейся об эксперименте информации, обеспечивающих единообразие аннотации экспериментов. При описании экспериментальных данных в GTRD, помимо клеточного типа или ткани, важно учитывать и формализовать и другие параметры: условия обработки, стадию развития организма, генотип и др. Для решения данной задачи были введены дополнительные ключи, которые формализуются в виде набора "ключ-значение" и привязываются к соответствующим экспериментам в GTRD.

Далее все данные проходили контроль качества и единым образом обрабатывались, используя платформу BioUML и собственное распределенное вычислительное решение e-grid, позволяющее выполнять параллельную обработку данных на нескольких вычислительных узлах (Kolmykov et al., 2021).

5.2.3 Единый словарь клеточных линий и типов

Для понимания регуляции транскрипции необходима интеграция различных типов NGS данных по клеточным типам (линиям) и экспериментальным условиям. Для этого необходимо:

- построение единого словаря клеточных линий и онтологии экспериментальных условий;
- привязка словаря клеточных линий к существующим онтологиями клеточных линий и типов Cell Ontology (Diehl et al., 2016), UBER-anatomy ontology (Mungall et al., 2012), Cellosaurus (Bairoch, 2018), Brenda tissue ontology (Gremse et al., 2010), Experimental factor ontology (Malone et al., 2010), Plant ontology (Cooper et al., 2013);
- использование этого словаря и онтологии экспериментальных условий при аннотации всех экспериментальных данных в БД GTRD.

Статистика по единому словарю клеточных линий и типов в БД GTRD и его сопоставление со специализированными БД клеточных линий и типов приведены в таблице 5.2.2.

В процессе аннотации данных, для каждого нового клеточного типа, добавляемого в БД, заполняются два поля:

- Source – поле, которое характеризует источник тканей, из которого клеточный тип был выделен в ходе эксперимента, либо же характеризует к какой анатомической системе относится данный клеточный тип.
- Cell type – описывает максимально близко клеточный тип, на котором проводился эксперимент. Все эти поля заполняются с помощью уникальных идентификаторов из различных специализированных БД, которые подробно описаны в приведенных выше работах.

В настоящее время 99% клеточных линий в БД GTRD имеют привязку к конкретной анатомической системе или типу ткани и лишь 87 (1%) не сопоставлены с какой-либо системой (т.е. параметр source при аннотации указан как "other"). Из этих 87, 29 (менее 1% от всех клеточных типов и тканей в базе) не имеют каких-либо уникальных идентификаторов как по параметру source, так и по параметру cell type, что связано, в первую очередь, с недостаточным качеством описания экспериментов в карточке в БД GEO.

Экспериментальные факторы также добавляются в БД в процессе аннотации по принципу ключ:значение, где ключом является тип фактора, а значение, описание этого фактора в эксперименте. На основе этой информации был сгенерирован словарь экспериментальных факторов.

Используя сопоставленные клеточные типы GTRD и БД Ensembl, ENCODE, FANTOM 5, и GTEx, были сопоставлены различные типы экспериментов из этих БД, но проведенные на одних и тех же клеточных линиях. Таким образом, появилась возможность интегрировать эксперименты по регуляции транскрипции между этими БД. Данное сопоставление позволило облегчить интеграцию данных для проведения различных анализов. Детальное описание работ по этому пункту приведено в публикациях Kulyashov et al., 2020a, b.

5.2.4 Единообразная обработка NGS данных при помощи сценариев

Были реализованы сценарии обработки данных NGS экспериментов различного типа:

- ChIP-seq – картирование РСТФ и различных типов модификаций гистонов на геном;
- ChIP-exo и ChIP-nexus – картирование РСТФ;
- DNase-seq, ATAC-seq, MNase-seq и FAIRE-seq – картирование на геноме областей открытого хроматина, а также локализации отдельных нуклеосом;

- RNA-seq – исследование уровня экспрессии соответствующих генов, идентификация генов-мишеней ТФ;
- WGBS (whole genome bisulfite sequencing) и RRBS (reduced representation bisulfite sequencing) – выявления паттернов метилирования ДНК;
- CAGE-seq (кэп-анализ экспрессии генов) – определения специфических стартов транскрипции (TSS), уровней их экспрессии и идентификации промоторных и энхансерных районов.

Для идентификации РСТФ для ChIP-seq экспериментов используются методы MACS2 (Zhang et al., 2008), GEM (Guo et al., 2012), SISSRs (Narlikar and Jothi, 2012) и PICS (Zhang et al., 2011). Для разных подтипов экспериментальных данных (например, ChIP-seq в вариантах single-end, paired-end, ChIP-exo, ChIP-nexus) разработаны соответствующие варианты сценариев, сопоставлены результаты и разработан алгоритм объединения таких данных.

Был разработан и валидирован алгоритм объединения сайтов инициации транскрипции (TSS) и энхансеров из аннотированных и единообразно обработанных CAGE-seq экспериментов, как друг с другом так и с FANTOM5. Это позволило создать стабильную систему идентификаторов TSS и энхансеров вида `rn6_4589.1`, которые неизменны при обработке и добавлении в БД GTRD новых CAGE-seq экспериментов. В результате обработки новых экспериментов добавляются только новые TSS и энхансеры, а также уточняются границы ранее выявленных TSS и энхансеров. В случае изменения координат сайта изменяется только версия идентификатора, а в случае слияния и разбиения сайтам назначается новый идентификатор и создается соответствующая запись в истории сайта (Pintus et al., 2021).

Для обработки рассматриваемых типов данных составленные сценарии были реализованы для:

- eGrid – система управления распределенными вычислениями, которая была разработана для контроля за распределенной обработкой NGS данных, входящих в состав GTRD, на кластере из 14 серверов (каждый – 2 процессора Intel® Xeon® CPU X5650, 32-48 GB RAM). Сценарии реализованы в виде программ, написанных на языке Java, и необходимых для обработки NGS данных программных пакетов.
- платформы BioUML – разработка и использование таких сценариев были описаны в главе 4. Пример сценария для обработки ATAC-seq данных приведен на рисунке 5.2.3.

При составлении сценариев обработки описанных выше типов NGS данных были учтены рекомендации, разработанные консорциумом ENCODE. Реализованные сценарии анализа NGS данных описаны в публикациях Kolmykov et al. 2020a и Pintus et al., 2021.

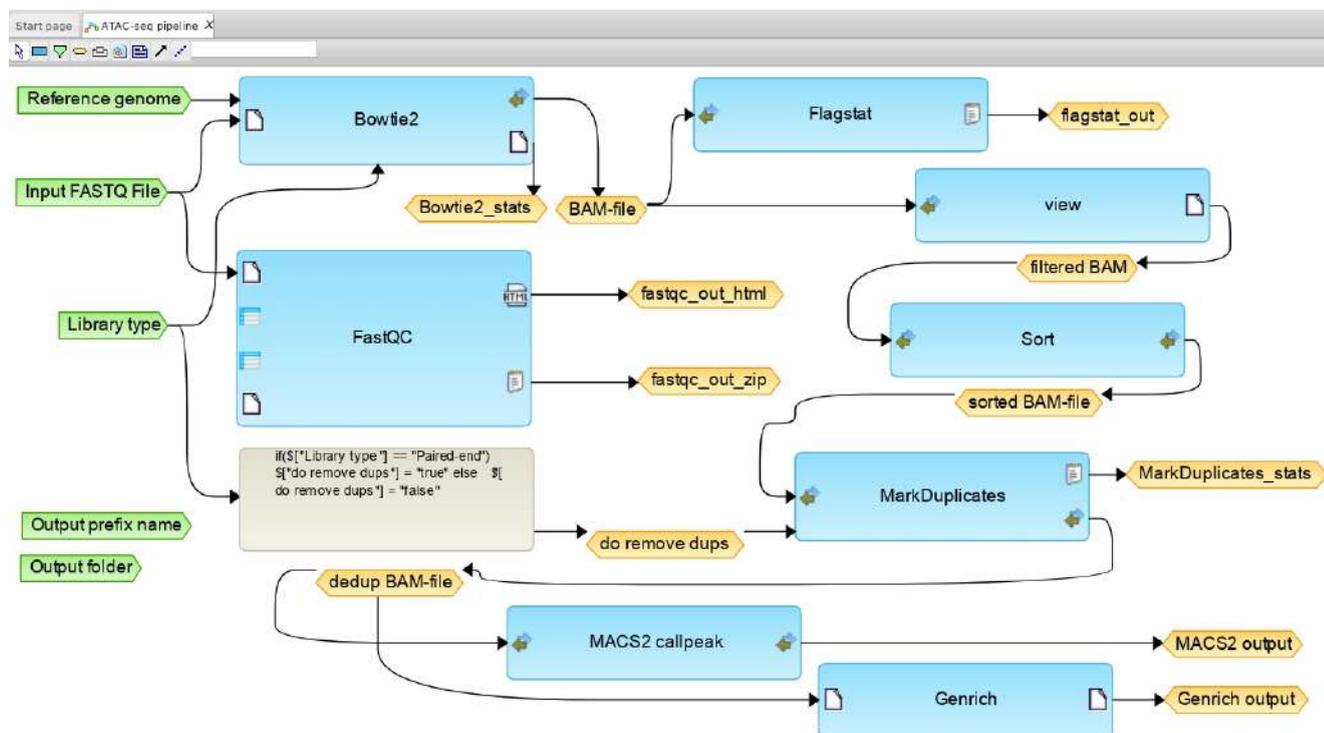


Рисунок 5.2.3 – Пример сценария для ПК BioUML для обработки ATAC-seq данных

5.2.5 Контроль качества данных

Для анализа качества NGS данных использовался стандартный для решения данной задачи подход. Качество “сырых” данных оценивается при помощи программы FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). На следующем этапе производится оценка качества выравнивания исходных прочтений. Помимо расчета базовых статистик выравнивания (samtools flagstat) производится оценка сложности библиотеки (NRF, PBC1 и PBC2) и оценка отношения сигнал-шум при помощи метода кросс-корреляции (NSC и RSC) (Landt et al., 2012). По завершению стадии определения РСТФ производится вычисление доли прочтений, попавших в полученные РСТФ (FRiP).

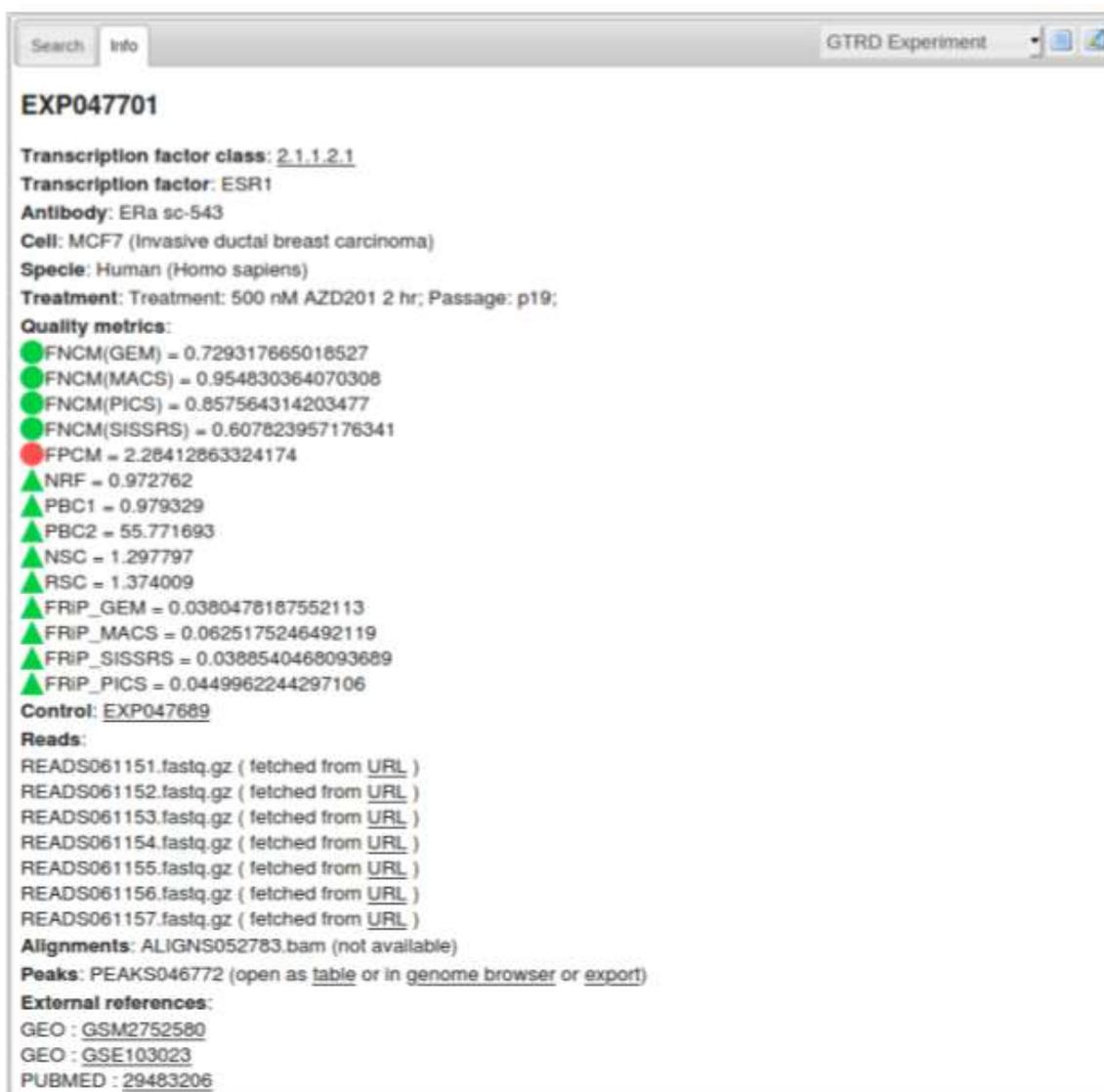
Для оценки доли ложно-выявленных и ложно-невыявленных РСТФ, полученных на основании обработки ChIP-seq экспериментов при помощи 4 разных методов, используемых в процессе обработки исходных ChIP-seq данных в GTRD (GEM, MACS2, PICS и SISSRs), нами были разработаны и применены новые метрики: FPCM и FNCM (Kolmykov et al., 2019).

Метрика FPCM (False Positive Control Metric) разработана для контроля ложно-выявленных РСТФ. Данная метрика выведена аналитически на основе двух предположений:

- 1) неизвестное число подлинных РСТФ является случайной величиной с распределением Пуассона;

- 2) большая часть ложно-выявленных РСТФ является «сиротами» (т.е. не имеет пересечений с другими РСТФ, полученными другими методами для конкретного эксперимента). Если значение FPCM превышает пороговое значение (2.0 или 3.0), то рекомендуется удалять такие РСТФ.

Метрика FNСM (False Negative Control Metric) разработана для контроля ложно-невыявленных РСТФ. Данная метрика применяется для каждого метода. Эта метрика определяется как отношение количества РСТФ, выявленных конкретным методом, и ожидаемым количеством подлинных РСТФ. При этом ожидаемое количество подлинных РСТФ оценивалось на основе использования следующих четырех известных оценок: оценка Чао (Chao, 1987), оценка Ланумтинга-Бонинга (Lanumteang, Bohning, 2011), оценка Зельтермана (Zelterman, 1988) и оценка методом максимального правдоподобия (McCrea, Morgan, 2014).



EXP047701

Transcription factor class: [2.1.1.2.1](#)
Transcription factor: ESR1
Antibody: ERa sc-543
Cell: MCF7 (Invasive ductal breast carcinoma)
Specie: Human (Homo sapiens)
Treatment: Treatment: 500 nM AZD201 2 hr; Passage: p19;

Quality metrics:

- FNСM(GEM) = 0.729317665018527
- FNСM(MACS) = 0.954830364070308
- FNСM(PICS) = 0.857564314203477
- FNСM(SISSRS) = 0.607823957176341
- FPCM = 2.28412863324174
- ▲ NRF = 0.972762
- ▲ PBC1 = 0.979329
- ▲ PBC2 = 55.771693
- ▲ NSC = 1.297797
- ▲ RSC = 1.374009
- ▲ FRIP_GEM = 0.0380478187552113
- ▲ FRIP_MACS = 0.0625175246492119
- ▲ FRIP_SISSRS = 0.0388540468093689
- ▲ FRIP_PICS = 0.0449962244297106

Control: [EXP047689](#)

Reads:

- READS061151.fastq.gz (fetched from [URL](#))
- READS061152.fastq.gz (fetched from [URL](#))
- READS061153.fastq.gz (fetched from [URL](#))
- READS061154.fastq.gz (fetched from [URL](#))
- READS061155.fastq.gz (fetched from [URL](#))
- READS061156.fastq.gz (fetched from [URL](#))
- READS061157.fastq.gz (fetched from [URL](#))

Alignments: [ALIGNS052783.bam](#) (not available)

Peaks: [PEAKS046772](#) (open as [table](#) or in [genome browser](#) or [export](#))

External references:

- GEO : [GSM2752580](#)
- GEO : [GSE103023](#)
- PUBMED : [29483206](#)

Рисунок 5.2.4 – Пример представления результатов контроля качества ChIP-seq при показе описания ChIP-seq эксперимента в платформе BioUML

Детально этот алгоритм представлен в работе Kolmykov et al., 2019. Данные метрики были рассчитаны для всех обработанных в БД GTRD ChIP-seq экспериментов. Пользователю они представляются в графическом виде при просмотре информации о РСТФ (мета-кластере, см. ниже) в платформе BioUML (рисунок 5.2.4).

5.2.6 Мета-кластеры – интеграция данных ChIP-seq экспериментов

Начиная с первой публично-доступной версии GTRD (Yevshin et al., 2017) мы используем следующий сценарий, чтобы интегрировать данные о РСТФ из разных ChIP-seq экспериментов для заданного ТФ, обработанных разными алгоритмами поиска РСТФ. Шаги этого сценария (рисунок 5.2.5) кратко описаны ниже.

1. Приведение к единому формату данных – для дальнейшей унифицированной обработки данных использовался формат данных FASTQ. Данные ChIP-seq, извлеченные из SRA в формате .sra, были преобразованы с использованием инструментария SRA (<http://www.ncbi.nlm.nih.gov/books/NBK158900>).
2. Выравнивание прочтений — мы использовали Bowtie2 (версия 2.2.3) (Langmead et al., 2012) для выравнивания прочтений ChIP-seq с эталонными геномами *Homo sapiens* (GRCh38), *Mus musculus* (GRCm38), *Drosophila melanogaster* (dm6), *Caenorhabditis elegans* (WBcel235), *Danio rerio* (GRCz11), *Rattus norvegicus* (Rnor_6.0), *Arabidopsis thaliana* (TAIR10), *Saccharomyces cerevisiae* (R64), *Schizosaccharomyces pombe* (ASM294v2) и *Gallus gallus* (GRCg6a). В основном использовались параметры по умолчанию, за исключением того, что у нас было фиксированное случайное начальное число (--seed 0) для воспроизводимости, ввод-вывод с отображением памяти (--mm) для более эффективного использования памяти и восемь потоков на процесс Bowtie2 (-p 8). Полученные выравнивания были преобразованы в файлы .bam, затем отсортированы и проиндексированы с помощью SAMtools v1.0 (Li H. et al., 2009).
3. Поиск РСТФ – мы использовали четыре разных метода, упомянутых выше: MACS, SISR, GEM и PICS. Эти четыре разных программы были использованы, поскольку они основаны на различных алгоритмах и учитывают различные аспекты данных ChIP-seq. Контрольные эксперименты использовались, когда они были доступны.
4. Кластеры РСТФ – РСТФ, рассчитанные для одного и того же ТФ и метода выявления РСТФ, но разных условий эксперимента (например, клеточная линия, обработка и т. д.), были объединены в кластеры. Поскольку ширина РСТФ (пиков) может значительно различаться при использовании разных методов, мы использовали только центры РСТФ, найденные соответствующей программой, и рассчитывали ширину РСТФ на основе его предполагаемой ширины и вариаций оцененных центров пиков.

5. Мета-кластеры – кластеры для одного и того же ТФ, выявленные разными методами выявления РСТФ, объединялись в мета-кластеры. Алгоритмы, использованные для построения кластеров и мета-кластеров последовательно развивались и будут описаны ниже.
6. Предсказанные ССТФ – для их поиска использовалась библиотека весовых матриц HOCOMOCO (Kulakovskiy et al., 2018).

Рисунок 5.2.5 демонстрирует шаги 1–5 для построения одного мета-кластера для ТФ USF1. Как видно, построенный мета-кластер хорошо соответствует как предсказанным ССТФ, так и реальным ССТФ, описанным в БД TRANSFAC (Matys et al., 2006). Таким образом, мета-кластеры можно рассматривать как невырожденный набор регуляторных районов, содержащих ССТФ.

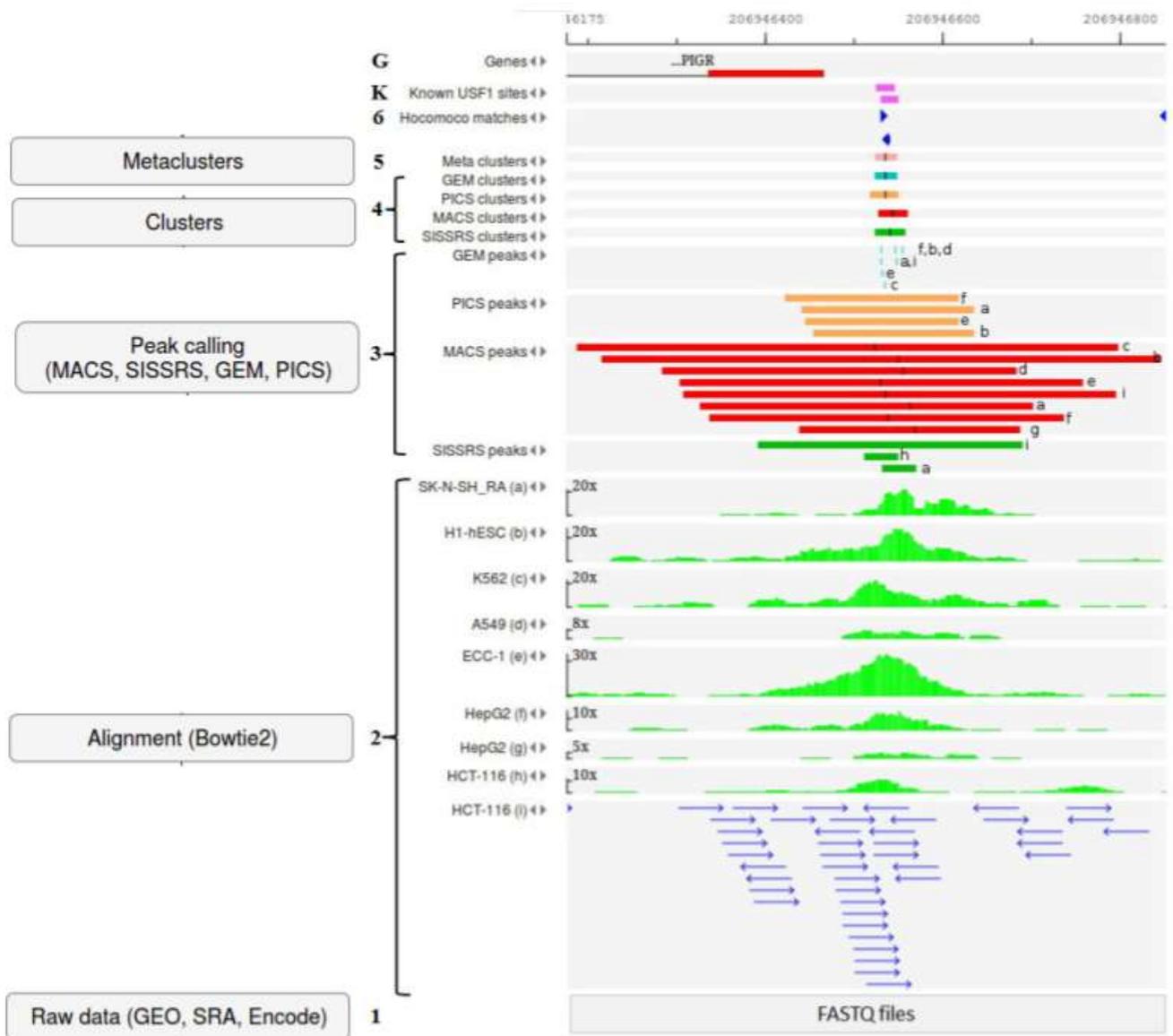


Рисунок 5.2.5 – Шаги сценария для интеграции данных о пиках из разных ChIP-seq экспериментов для заданного ТФ, обработанных разными алгоритмами поиска пиков (см. пояснение в тексте)

5.2.7 Мета-анализ для наиболее достоверного выявления регуляторных районов

Для выявления наиболее достоверных РСТФ на основе мета-анализа результатов ChIP-seq анализа всех экспериментов из БД GTRD для заданного ТФ был разработан новый метод METARA – MEtA Analysis of ChIP-seq datasets through the Rank Aggregation (рисунок 5.2.6, Kolmykov et al., 2019, 2020b). Выявленные при его помощи РСТФ, встречающиеся в нескольких экспериментах одновременно, называются мета-кластерами. Подробное описание метода METARA и его основные свойства опубликованы в статьях Kolmykov et al., 2019, 2020b, а также в Приложении к статье Kolmykov et al., 2021 (разделы: 1. Novel algorithm for meta-cluster construction; 2. Two ways for selection of the most reliable meta-clusters).

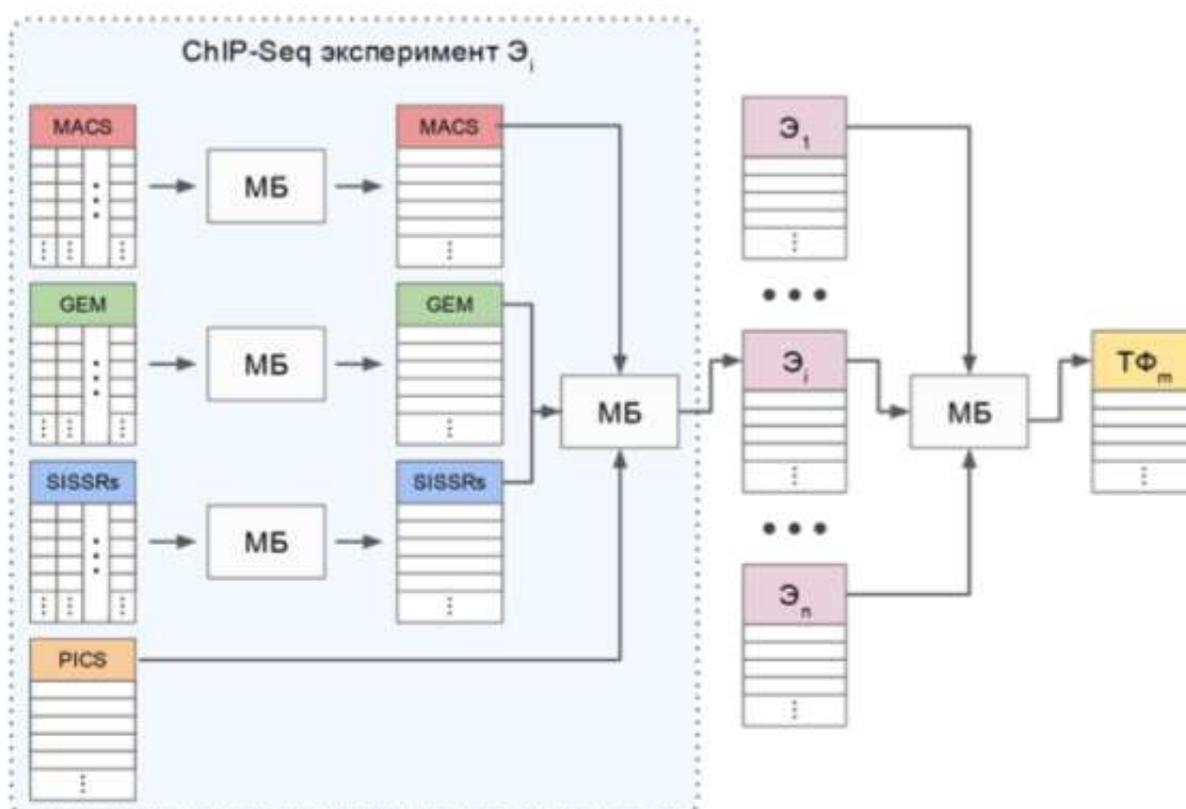


Рисунок 5.2.6 – Схема метода METARA для построения мета-кластеров. МБ – метод Борда

Этот метод представляет собой трехэтапное применение метода коллективного выбора (Rank Aggregation Approach, RA-approach):

- 1-й этап – метод Борда применяется для ранжирования РСТФ, найденных каждым методом (MACS, SISRrs, GEM и PICS), на основании различных характеристик качества, присваиваемых соответствующим методом (например, p-value, “fold enrichment” и др.);
- 2-й этап – полученные и упорядоченные на основании полученных весов списки РСТФ также подаются на вход методу Борда;

- 3-й этап – полученный список РСТФ группируется с другими подобными списками, полученными из экспериментов по рассматриваемому ТФ, и затем обрабатываются используемым методом коллективного выбора.

Как окончательный результат применения метода METARA, для заданного ТФ идентифицируется множество мета-кластеров РСТФ, при этом каждому мета-кластеру приписывается свой вес (мета-скор, RA-score). Полученные мета-скоры дают возможность упорядочить мета-кластера в порядке их надежности.

Было показано, что наиболее надежные мета-кластера намного сильнее насыщены ССТФ, чем наименее надежные мета-кластера (Kolmykov et al., 2020). Для поиска ССТФ использовались весовые матрицы из БД НОСОМОСО.

С использованием метода двухкомпонентной смеси нормальных распределений к мета-скорам было показано, что имеется возможность выявлять оптимальный порог, чтобы отфильтровывать менее надежные мета-кластера.

METARA также был применен для определения границ районов открытого хроматина. Для этого, в отличие от определения границ РСТФ оказалось лучше объединять, а не пересекать границы исходных пиков.

5.2.8 Мастер-сайт и мастер-трек

Мастер-сайт является дальнейшим расширением концепции мета-кластера РСТФ. Мета-кластер определяет границы РСТФ (область связывания, серая рамка). Мастер-сайт для такой области объединяет всю информацию из GTRD, включая:

- кластеры ChIP-seq пиков в разных клеточных линиях;
- найденные в последовательности мета-кластера при помощи весовых матриц из БД НОСОМОСО соответствующие мотивы;
- перекрывающиеся области открытого хроматина;
- пересекающиеся области гистоновых модификаций.

Геномный браузер платформы BioUML позволяет отображать/скрывать различные сведения о мастер-сайте (рисунок 5.2.7).

Мастер-сайты хранятся в специальном типе треков в bigBed формате. Для этого последняя колонка для каждого сайта в этом формате содержит поле в формате JSON, в котором хранится вся дополнительная информация о мастер-сайте.

5.2.9 Структура БД

Первые версии БД GTRD использовали СУБД MySQL для хранения метаданных и полученных результатов анализа NGS данных (за исключением BAM-файлов). По мере увеличения объема данных это стало узким местом, и сейчас БД MySQL используется только

для хранения метаданных. Результаты анализа сохраняются непосредственно в файловой системе в виде файлов в формате bigBed, что дает ряд преимуществ:

- обработка файлов bigBed идет значительно эффективнее в параллельном режиме;
- требования к пространству для хранения значительно ниже за счет внутреннего сжатия bigBed;
- файлы bigBed можно напрямую загружать из БД GTRD без дополнительного преобразования и напрямую визуализировать в браузерах генома UCSC или Ensembl.



Рисунок 5.2.7 – Пример представления мастер-сайта в платформе BioUML (см. описание в тексте)

5.2.10 Веб-интерфейс

Веб-интерфейс для работы с БД GTRD разработан в двух взаимодополняющих вариантах:

- специальная перспектива в ПК BioUML (приложение 5.1), которая содержит специальную стартовую страницу с основными методами для работы с GTRD, а также скрывает другие нерелевантные для этого методы анализа, пункты меню, БД и т.п.
- веб-интерфейс на основе Bean Explorer Enterprise Edition (приложение 5.2) – представляет собой более классический веб-интерфейс, который позволяет просматривать информацию из БД GTRD в виде таблиц, специальных представлений, а также искать данные, используя заранее predeterminedенные фильтры.

5.3 Использование БД GTRD

БД GTRD является крайне востребованной, широко используемой и цитируемой: три публикации в специализированных выпусках *Nucleic Acids Research* 2017, 2019 и 2021 года набрали в совокупности 400+ цитирования по версии Semantic Scholar (<https://www.semanticscholar.org/>), включая статьи в журналах *Nature* и *Science*.

5.3.1 Построение цистрома

Под цистромом понимают множество всех ССТФ для заданного организма (Liu et al., 2011). Было сделано 2 попытки построения цистрома человека и мыши на основе информации из БД GTRD.

В первом варианте, Vorontsov et al., 2018, использовали пики для 599 ТФ у человека и 349 ТФ у мыши, идентифицированные в ChIP-seq экспериментах, используя методы MACS, GEM, SISSRs и PICS. Используя подход, описанный в работе Kulakovskiy et al., 2018, некоторые РСТФ были исключены как ненадежные. Затем, используя BEDTools 2.26.0 (Quinlan et al., 2010), были проанализированы пересечения для РСТФ из разных экспериментов. Полученные регионы были классифицированы по четырем категориям надежности следующим образом:

- А (наивысшая надежность, экспериментальная и техническая воспроизводимость) – эта группа регионов, состоящие из перекрывающихся пиков, обнаруженных не менее чем в двух наборах экспериментальных данных и не менее чем в двух наборах пиков, выдаваемых соответствующими методами, т.е. подтвержденных не менее чем двумя экспериментами и не менее чем двумя методами;
- В (высокая надежность, экспериментальная воспроизводимость) – регионы, подтвержденные как минимум двумя экспериментами;
- С (средняя надежность, техническая воспроизводимость) – регионы, поддерживаемые как минимум двумя методами;
- D (ограниченная надежность) – все остальные регионы.

Второй вариант связан с развитием алгоритма METARA. На его основе был разработан алгоритм IMETARA (Kolmykov et al., 2022), который позволяет инкрементально находить или уточнять мета-кластера РСТФ по мере поступления и анализа новых данных в БД GTRD.

Согласно схеме (рисунок 5.3.1), сначала метод IMETARA метод применяется только к вновь поступившим множествам РСТФ, а затем вновь полученные мета-кластера внедряются в уже существующее множество мета-кластеров. Важно отметить, что такое внедрение проводится на основе пересчета новых мета-скоров уже существовавших мета-кластеров. Таким образом, IMETARA метод добавляет только часть новых (ранее не существующих) мета-кластеров, а в уже существовавших мета-кластерах, при необходимости, производится

только пересчет мета-скоров и уточнение их геномных границ. Это позволяет создать стабильную систему идентификаторов РСТФ (мета-кластеров), которые будут стабильны при обработке новых данных. Такая схема разработанного метода позволяет пользователям БД GTRD сохранить преимущество результатов при переходе от анализа данных из старой версии БД GTRD к анализу данных из новой версии.

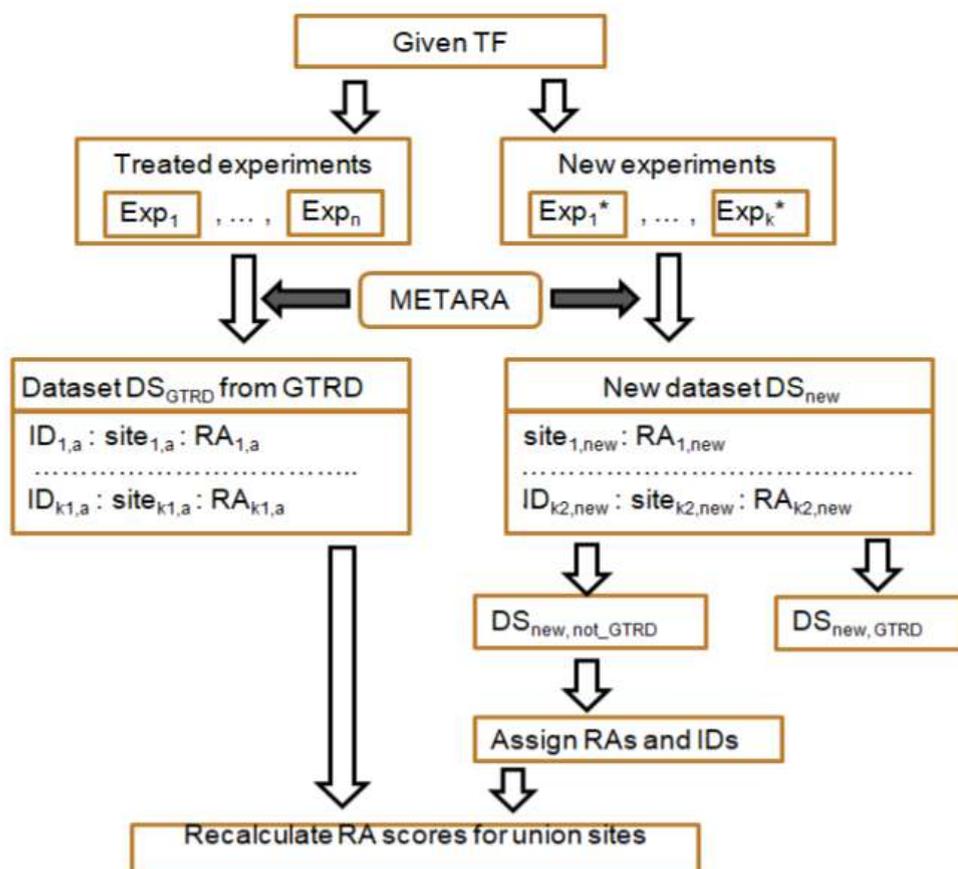


Рисунок 5.3.1 – Схема метода IMETARA для построения мета-кластеров (цистррома)

В построенном цистроме человека каждому РСТФ (мета-кластеру) был дан уникальный стабильный идентификатор, что позволит исследователям ссылаться на них так, как они сейчас ссылаются, например, на SNP. Идентификатор РСТФ имеет вид:

`ms.Uniprot_name.n.vn`, например, `ms.AHR_HUMAN.31.v1`

- `ms` – означает что это мастер-сайт БД GTRD;
- `Uniprot_name` – названия ТФ в БД Uniprot;
- `n` – порядковый номер РСТФ;
- `vn` – номер версии РСТФ.

По мере поступления новых данных в GTRD геномные границы этого сайта могут быть пересмотрены, в этом случае изменится только версия, и идентификатор примет вид `ms.AHR_HUMAN.31.v2`.

Полученный цистром человека доступен для просмотра в геномном браузере BioUML на сайте <https://gtrd.biouml.org/> в разделе `databases/GTRD/Data/bigBed/Homo sapiens/Master tracks`, а также доступен для скачивания (<http://gtrd.biouml.org:8888/egrid/mastertracks>).

Следует отметить, что пользователей БД GTRD в ряде случаев может интересовать более упрощенная схема оценки надежности РСТФ вместо численных мета-скоров. Для таких случаев имеющееся распределение мета-скоров для каждого РСТФ предварительно разбивается на три группы надежности:

- 1) наиболее надежные РСТФ;
- 2) кандидаты (умеренно надежные РСТФ);
- 3) ненадежные РСТФ.

Для решения этой задачи использовался известный алгоритм разложения имеющегося распределения на три нормальных компоненты. Для большинства ТФ применение этого алгоритма приводило к идентификации в явном виде соответствующих типов РСТФ. Рисунок 5.3.2 демонстрирует разложение эмпирических распределений мета-скоров для FOXA1 и JUND на три группы надежности.

Также новый алгоритм IMETARA был использован для идентификации мета-кластеров, предназначенных для описания профилей открытого хроматина, основанных на данных DNase-seq экспериментов. В данном случае исходные наборы данных для построения мета-кластеров группировались на уровне клеточных типов.

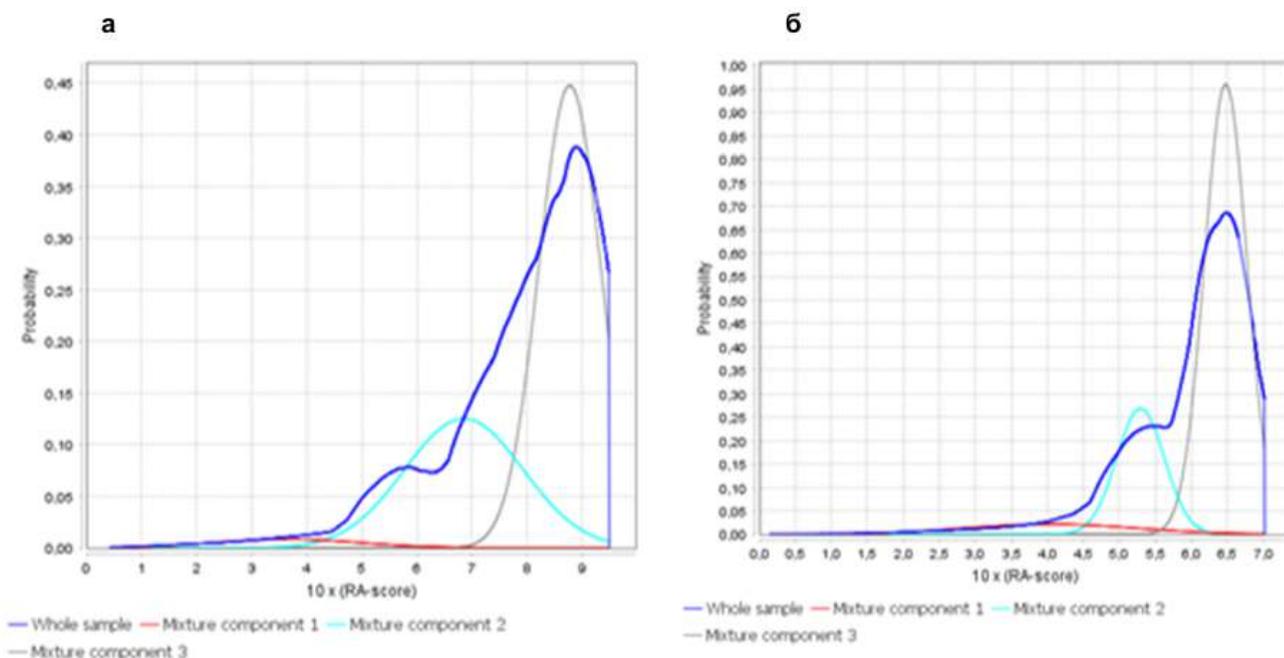


Рисунок 5.3.2 – Разложение эмпирического распределения мета-скоров для (а) FOXA1 и (б) JUND на три нормальных компоненты: красный – надежные, голубой – умеренно надежные; серый – ненадежные, синий – вся выборка

5.3.2 Ресурсы, построенные на основе данных GTRD

На основе БД GTRD коллективом член-корреспондента, д.б.н., проф. Макеева В.Ю. и д.б.н. Кулаковского И.В. совместно были созданы ресурсы:

- HOCOMOCO (Kulakovskiy et al., 2018) – HOMO sapiens COMprehensive MODEL Collection – коллекция мотивов для ССТФ человека и мыши (<https://hocomoco11.autosome.ru/>);
- ADASTRA (Abramov et al., 2021) – Allelic Dosage-corrected Allele-Specific human Transcription factor binding sites (<https://adastra.autosome.ru/>) – коллекция данных по аллель-специфичному связыванию факторов транскрипции в геноме человека;
- ANANASTRA (Boytsov et al., 2022) – ANnotation and enrichment ANALysis of SNP sets with allele-specific TRANscription factor binding (<https://ananastra.autosome.ru/>) – коллекция данных по влиянию SNP на аллель-специфичное связывание факторов транскрипции в геноме человека;
- версия цистрома для человека и мыши (Vorontsov et al., 2018).

Также информация из БД GTRD была использована при создании известных международных БД:

- BaMM motif (Siebert et al., 2016) – Bayesian Markov Models (BaMMs) – для распознавания ССТФ (<https://bammotif.soedinglab.org/>);
- mSigDB (Liberzon et al., 2015) – Molecular Signatures DataBase, C3: regulatory target gene sets (<https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp#C3>).

5.3.3 Предсказание уровня экспрессии генов

Для построения моделей регуляции транскрипции в отдельных клеточных линиях был разработан новый метод, в котором используется регрессионный анализ ChIP-Seq данных из БД GTRD (мета-кластеров) и активности сайтов инициации транскрипции, экспериментально измеренных в рамках международного проекта FANTOM5. Для этой цели мы изначально генерировали большое количество бинарных переменных (features), а затем применяли пошаговую регрессию с целью выбора наиболее важных переменных (feature selection).

Исходные бинарные переменные описывали наличие или отсутствие РСТФ в различных промоторных районах ([-5000, -1001], [-1000, -501], [-500, -201], [-200, -101], [-100, 0], [1, 100], [101, 500], [501, 1000] относительно TSS). Используя данный метод были идентифицированы наборы наиболее важных ТФ, которые играют ключевую роль в экспрессии генов в 3 клеточных линиях человека как HepG2, K562 и HEK293, которые наиболее полно исследованы с помощью ChIP-Seq экспериментов. Важно отметить, что разработанный метод позволяет не только идентифицировать наборы наиболее важных ТФ, но и классифицировать их как активаторы или

репрессоры. Данный метод и полученные результаты были опубликованы в статье Sharipov et al., 2020.

Точность построенных моделей регуляции транскрипции оценивалась на основе коэффициента корреляции R между наблюдаемыми и предсказанными (с помощью регрессии) значениями активностей СТС. Для HepG2, K562 и HEK293 корреляция R достигала 0.743, 0.733 и 0.732, соответственно.

5.4 Обсуждение

5.4.1 Сравнение с другими БД по ChIP-seq экспериментам

Высокие темпы обновления информации с постоянным контролем качества данных позволяют БД GTRD оставаться одной из самых больших в мире БД по количеству единообразно аннотированных и обработанных ChIP-seq экспериментов (Yevshin et al., 2017; Kolmykov et al., 2019, 2021). Сравнение с другими БД по ChIP-seq экспериментам приведено в приложении 5.3. Основным конкурентом является БД ChIP-Atlas (Zou et al., 2021; <https://chip-atlas.org>), сравнение с которым приведено в отдельной таблице в приложении 5.3. Основным отличием БД ChIP-Atlas от БД GTRD является способ аннотирования NGS экспериментов по регуляции транскрипции: в БД GTRD используется ручная аннотация мета-данных, в то время как в БД ChIP-Atlas используется автоматический интеллектуальный анализ данных (Zou et al., 2021). Мы полагаем, что хотя ручная аннотация является более медленной, она обеспечивает более высокое качество.

Также стоит отметить тот факт, что в GTRD при аннотации данных ChIP-seq происходит объединение повторяющихся реплик благодаря этапу ручной аннотации, если индивидуальный контроль представлен не для каждой из них. Тогда как в ChIP-Atlas каждая реплика считается отдельным экспериментом.

5.4.2 Полнота покрытия ТФ

Интересен вопрос, для какого процента известных ТФ проведены соответствующие ChIP-seq эксперименты. Для ответа на этот вопрос в качестве "золотого стандарта" по известным ТФ была использована БД CIS-BP (Weirauch et al., 2014). ТФ, представленные в CIS-BP были сопоставлены по UNIPROT ID, который используется и в CIS-BP, и в GTRD. Результаты сопоставления представлены в таблице 5.4.3 Как видно, наибольшее покрытие достигается для *Homo sapiens*, *Saccharomyces cerevisiae* и *Drosophila melanogaster*, для которых в GTRD представлены ChIP-seq и ChIP-ехо эксперименты для 77%, 73.6% и 65.5% из известных ТФ.

Таблица 5.4.1 – Полнота покрытия множества ТФ ChIP-seq экспериментами в GTRD

Вид	CIS-BP #ТФ	GTRD v23.07 #ТФ и кофакторы	CIS-BP – GTRD пересечение по ТФ	% от CIS-BP
<i>Homo sapiens</i>	1639	2123	1264	77.1
<i>Mus musculus</i>	1513	1098	660	43.6
<i>Rattus norvegicus</i>	1362	198	163	11.8
<i>Danio rerio</i>	2350	28	18	0.7
<i>Caenorhabditis elegans</i>	766	364	234	30.5
<i>Drosophila melanogaster</i>	719	683	471	65.5
<i>Saccharomyces cerevisiae</i>	239	828	176	73.6
<i>Schizosaccharomyces pombe</i>	115	79	12	10.4
<i>Arabidopsis thaliana</i>	1749	218	132	7.5

5.4.3 Полнота покрытия ССТФ

Следующий интересный вопрос – насколько полным является множество мета-кластеров в БД GTRD, которые, в первом приближении, соответствуют ССТФ. Для этого из БД TRANSFAC, которая является “золотым стандартом” по экспериментально подтвержденным ССТФ, были извлечены ССТФ, идентифицированные в индивидуальных экспериментах, таких как ChIP-PCR, EMSA, анализ мутаций и т.п. Эти ССТФ сравнивались с мета-кластерами РСТФ в БД GTRD и определялись случаи, когда ССТФ TRANSFAC не пересекался с мета-кластером. Эту информацию можно использовать как оценку полноты покрытия всех сайтов связывания для заданного ТФ соответствующими мета-кластерами GTRD. Результаты представлены в таблице 5.4.4 Как видно, такое покрытие варьирует от 21% до 86%.

Поэтому на следующем этапе был проведен ручной анализ причин отсутствия мета-кластеров для нескольких ТФ, с малой долей покрытия: SP1 (только для 1-ой хромосомы) и ESR1 (все сайты). Стоит отметить, что эта очень трудоемкая работа, которая требует ручного анализа каждого ССТФ, была проделана Евшиным И.С.

Таблица 5.4.2 – Полнота покрытия ССТФ, представленных в БД TRANSFAC, соответствующими мета-кластерами в GTRD для человека

ТФ (Uniprot ID)	Количество экспериментов в GTRD	Количество мета-кластеров	Количество сайтов в TRANSFAC	Полнота покрытия %
Sp1 (P08047)	24	711 089	1430	43
Sp3 (Q02447)	2	24 167	446	82
NFKB3 (Q04206)	100	3 335 438	424	39
Jun (P05412)	81	1 371 244	309	45
NFKB1 (P19838)	8	2 882 858	277	57
ESR1 (P03372)	389	8 492 377	124	86
FOXA1 (P55317)	248	4 998 092	47	21

Строки таблицы упорядочены в порядке убывания количества сайтов в TRANSFAC.

Причины несовпадений были классифицированы следующим образом:

- различие клеточных линий или экспериментальных условий – многие ССТФ являются специфичными для клеточных линий/типов или работают в определенных условиях;
- расположение в геномных повторах – методика ChIP-seq не позволяет однозначно идентифицировать РСТФ в геномных повторах, поскольку короткие прочтения (36-70 п.о.) не могут быть однозначно выровнены на геном;
- Chip-PCR/EMSA/мутационный анализ противоречит ChIP-seq. В этом случае методом ChIP-seq не определяется ССТФ, хотя определяется одним или несколькими индивидуальными экспериментами. При этом такие случаи не попадают ни в одну из других категорий. Истинной причиной таких несовпадений могут быть неясные различия в экспериментальных условиях, не описанные в исходных публикациях, различие в маркированных одинаково клеточных линиях, стохастическая природа ССТФ, а также ложноположительные ответы в индивидуальных экспериментах. Иногда ложно-положительные ответы являются неотъемлемой частью экспериментального метода. Так EMSA может показать ССТФ in vitro, в то время как связывание in vivo не происходит (например, вследствие закрытой структуры хроматина в этом районе in vivo). Мутационный анализ может показать влияние мутации возле предполагаемого ССТФ на экспрессию гена, хотя это влияние может реализоваться через сайты связывания другого ТФ;

- ошибки в исходном исследовании, плохой дизайн эксперимента – в этом случае подробный анализ публикации выявил ошибки или недостатки в исходном исследовании, что позволяет отбросить эти ССТФ как ошибочные;
- неточность определения границ РСТФ или кластеризации – используемые в БД GTRD методы анализа могут не точно определять границы РСТФ. В основном, это касается случаев, когда несколько ССТФ расположены на небольшом расстоянии друг от друга;
- неточность аннотации в БД TRANSFAC;
- отсутствие экспериментальных доказательств;
- причина несовпадения не найдена.

Был подсчитан вклад отдельных категорий ошибок в общую картину (таблица 5.4.5). Ожидаемо, наиболее частой причиной несовпадений оказалось различие клеточных линий и экспериментальных условий (68,9% случаев для SP1 и 37,9% для ESR1).

Также большой вклад вносят ССТФ, находящиеся в геномных повторах, картирование которых короткими прочтениями ChIP-seq затруднено (10,8% случаев для SP1 и 10,3% для ESR1).

В случае ESR1, 20,7% несовпадений можно объяснить несовершенством алгоритмов определения пиков и кластеризации, в то время как для SP1 в эту категорию попадает лишь 5,4% ошибок.

Оставшиеся несовпадения объясняются ошибками/неточностями в исходных исследованиях (2,7% для SP1, 12,6% для ESR1), неверной интерпретацией исследований при занесении в базу TRANSFAC и т.п (4,1% для SP1, 2,3% для ESR1).

Лишь для 5,4% несовпадения SP1 и 11,5% несовпадений ESR1 не было найдено причин их возникновения (т.е. или ChIP-PCR, или EMSA, или анализ мутаций противоречит данным ChIP-seq).

Была проанализирована зависимость процента вновь сгенерированных мета-кластеров РСТФ при увеличении количества ChIP-Seq экспериментов (рисунок 5.4.1). В результате этого анализа было обнаружено, что процент отсутствующих мета-кластеров падает при увеличении количества обработанных ChIP-seq экспериментов, используемых для построения мета-кластеров.

Таким образом, для построения достаточно полного цистрома необходимо:

- 1) большее количество ChIP-seq экспериментов с большим покрытием клеточных линий и типов;
- 2) дальнейшее совершенствование алгоритмов определения РСТФ и их кластеризации;
- 3) разработка алгоритмов для локализации РСТФ в геномных повторах.

Таблица 5.4.3 – Причины неполного покрытия ССТФ в БД TRANSFAC соответствующими мета-кластерами в БД GTRD для Sp1 и ESR1 человека

Причина	Sp1, %	ESR1, %
Различие клеточных линий или экспериментальных условий	68,9	37,9
Расположение в геномных повторах	10,8	10,3
ChIP-PCR/EMSA/мутационный анализ противоречит ChIP-seq	5,4	11,5
Ошибки в исходном исследовании, плохой дизайн эксперимента	2,7	12,6
Неточность определения пиков или кластеризации	5,4	20,7
Неточность аннотации в БД TRANSFAC	4,1	2,3
Отсутствие экспериментальных доказательств	2,7	4,6

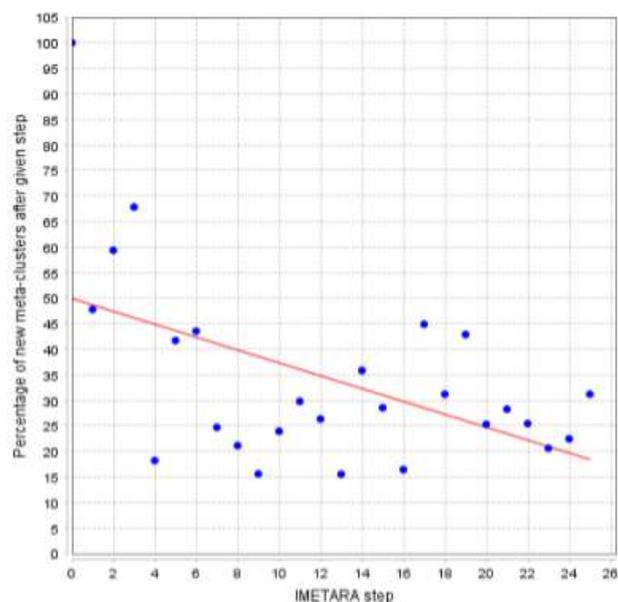
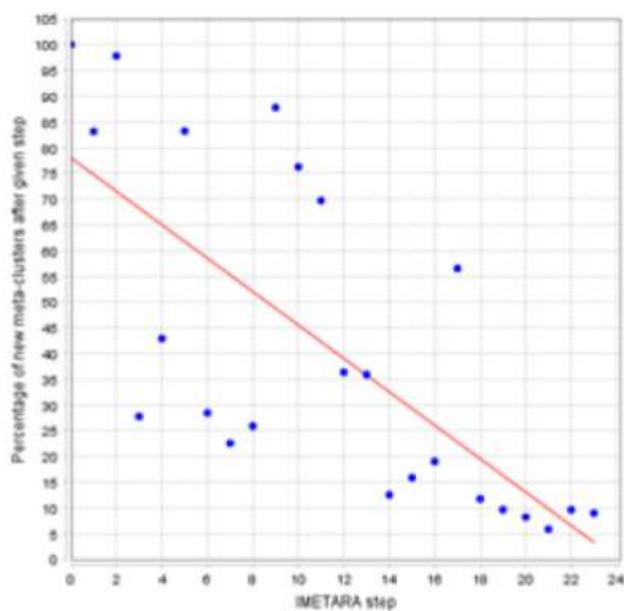


Рисунок 5.4.1 – Зависимость процента вновь сгенерированных мета-кластеров при увеличении количества ChIP-Seq экспериментов для транскрипционных факторов ESR1 и FOXA1

ГЛАВА 6. МОДЕЛИРОВАНИЕ СЛОЖНЫХ БИОЛОГИЧЕСКИХ СИСТЕМ

В главе 6 продемонстрированы возможности разработанной технологии и ПК BioUML для создания и использования сложных моделей биологических систем.

6.1 Модульная модель апоптоза

Рассмотрим основные этапы предложенной технологии на примере создания модульной модели апоптоза и используя модифицированную для этого технологию ХР.

Апоптоз – многоэтапный процесс, который может быть разделен на три основные фазы: сигнальную (индукторную), эффекторную и деградиционную (Kroemer et al., 1995).

Сигнальная фаза (рисунок 6.1.1) – апоптоз запускается внутренними (intrinsic) или внешними (extrinsic) сигналами, источниками которых являются различные формы клеточного стресса или так называемые “рецепторы смерти”, расположенные на поверхности клеток (Adams, 2003).

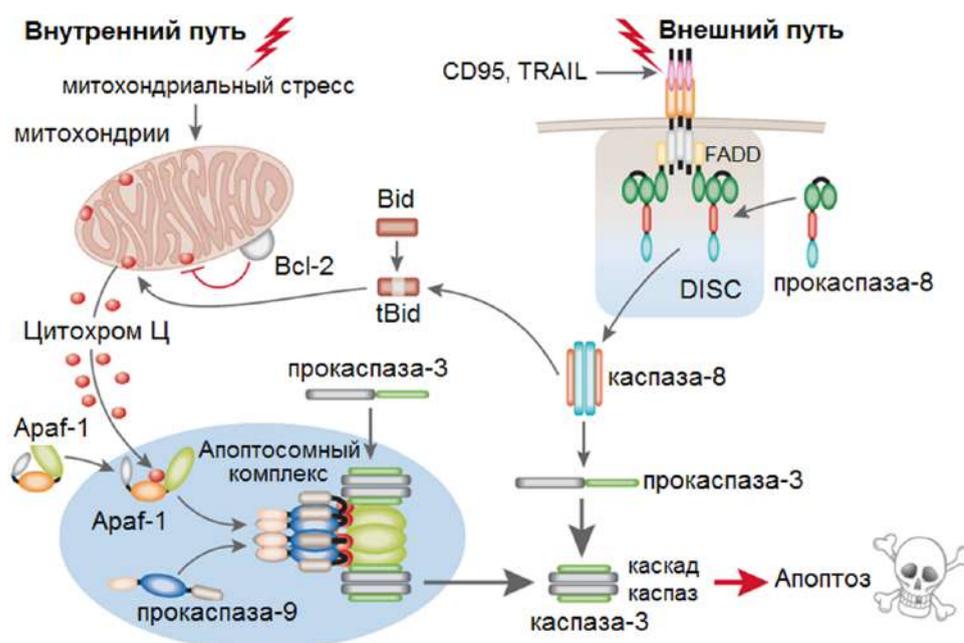


Рисунок 6.1.1 – Два основных пути апоптоза: один активируется через воздействие рецепторов смерти (внешний путь), а другой – посредством стресс-индуцирующих стимулов (внутренний путь). Инициация рецепторов смерти, расположенных на поверхности клеток (CD95, TRAIL), приводит к активации каспазы-8 посредством ее взаимодействия с комплексом DISC через адаптерную молекулу FADD. При внутреннем пути стресс-индуцированный апоптоз приводит к высвобождению цитохрома С из митохондриального межмембранного пространства, при этом антиапоптотические (Bcl-2/Bcl-XL/Mcl1) и проапоптотические (Bax, Bak и tBid) белки соответственно ингибируют или стимулируют высвобождение. Далее цитохром С связывается с фактором Apaf-1, что приводит к активации каспазы-9, а также последующей активации эффекторных каспаз-3, -6 и -7, которые отвечают за расщепление важных клеточных субстратов (MacFarlane and Williams, 2004)

Эффекторная (execution) фаза – характеризуется последовательностью протеолитических событий, целью которых является апоптотический “демонтаж” клетки. При этом мишенями для действия эффекторных каспаз являются белки, отвечающие за ее жизненные функции. В частности, каспазы-3 и -7 запускают расщепление поли(АДФ-рибоза)-полимеразы (PARP), участвующей в репарации повреждений ДНК, после чего процесс апоптоза становится необратимым (Chaitanya et al., 2010).

Деградиционная фаза – происходит фрагментации клетки на отдельные апоптотические тельца, ограниченные плазматической мембраной. Фрагменты погибшей клетки поглощаются макрофагами либо соседними клетками, минуя развитие воспалительной реакции.

В создаваемой модели мы будем рассматривать только сигнальную и эффекторную фазу (рисунок 6.1.1), которые представляют наибольший интерес для моделирования.

6.1.1 Эволюционное создание моделей

Сначала проведем анализ литературы с целью нахождения наиболее подходящих моделей («предков», от которых будет идти дальнейшая эволюция), разработанных в последнее время (здесь необходимо учесть, что период разработки описываемой модели – 2009-2011 гг.).

При создании обобщенной модели апоптоза рассматривали следующее ранее опубликованные модели (Кутумова, 2012):

- **Bentele et al., 2004** – модель CD95-индуцируемого апоптоза подробно описывает процесс активации каспазы-8 внутри комплекса DISC-CD95 с дальнейшим расщеплением прокаспазы-3, характерным для клеток типа I, и активацией митохондриального пути. Итогом индукции этих сигнальных путей является расщепление белка PARP, после чего процесс апоптоза становится необратимым.
- **Albeck et al., 2008** – модель апоптоза, индуцируемого рецептором TRAIL-R, изучает роль прокаспазы-8 и ее субстратов в задержке проницаемости внешней мембраны митохондрий, а также исследует про- и антиапоптотические свойства белков семейства Bcl-2.
- **Rangamani and Sirovich, 2007** – рассматривает сигнальные пути выживания и гибели клеток под действием TNF- α . Выживание возможно за счет активации транскрипционного фактора NF- κ B, который стимулирует продуцирование ингибитора апоптоза cIAP.
- **Bagci et al., 2006** – описывает митохондриально-зависимый путь апоптоза и характеризует роль белков Bax и Bcl-2 (с учетом скорости их синтеза и дегградации), а также числа митохондриальных пор в этом процессе.
- **Legewie et al., 2006** – изучает процесс ингибирования каспаз-3 и -9 посредством IAP.

- **Schoeberl et al., 2002** – модель сигнальных путей эпидермального фактора роста (EGF), который связываясь со своим рецептором (EGFR) на поверхности клеток, приводит к передаче сигнала через ряд адаптерных (Grb-2, Shc) и эффекторных (Sos) белков к активации киназного каскада Ras/Raf/MEK/ERK.
- **Borisov et al., 2009** – модель стимуляции клеточной линии HEK293 инсулином и фактором роста EGF, в результате чего происходит активация киназных каскадов Ras/ERK и PI3K/Akt.
- **Hoffmann et al., 2002** – изучает сигнальные пути транскрипционного фактора NF-κB, локализация которого в ядре клеток млекопитающих контролируется тремя изоформами ингибитора: IκB-α, -β и -ε.
- **Hamada et al., 2009** – описывает внутренний путь индукции апоптоза посредством активации процесса синтеза опухолевого супрессора p53 и влияние поврежденной ДНК на фазу клеточного цикла G2/M.

Перечисленные модели представляют одни и те же сигнальные пути апоптоза, используя разные цепочки реакций и значения кинетических параметров.

Наша задача при предложенном подходе:

- выбрать стартовые модели для каждой итерации;
- воспроизвести эти модели;
- внести необходимые изменения:
 - для возможности представления в виде модулей и стыковки с другими модулями;
 - для соответствия тестам, отслеживающим воспроизведение экспериментальных данных;
 - для учета новых данных и знаний.

6.1.2 Планирование

Для примера, сначала сформулируем следующие требования к модели в виде набора историй пользователя (user stories) и соответствующих приемочных тестов (unit tests):

- Требование 1: эффекторная фаза – если нет внешних сигналов, то апоптоз не происходит;
 - основа модуля – модель Bentele et al., 2004;
 - тест – концентрация PARP равна нулю.
- Требование 2: сигнальная фаза, запуск внешнего пути апоптоза через рецепторы CD95.
 - основа модуля – модель Bentele et al., 2004;
 - тест – динамика каспазы-8 и прокаспазы-8 должна соответствовать экспериментальным данным Bentele et al., 2004 (рисунок 6.1.2).
- Требование 3: сигнальная фаза, запуск внешнего пути апоптоза через рецепторы TNF-α;

- основа модуля – модель Rangamani and Sirovich, 2007;
- тест – динамика каспазы-8 и прокаспазы-8 должна соответствовать экспериментальным данным Janes et al., 2006 (рисунок 6.1.2).

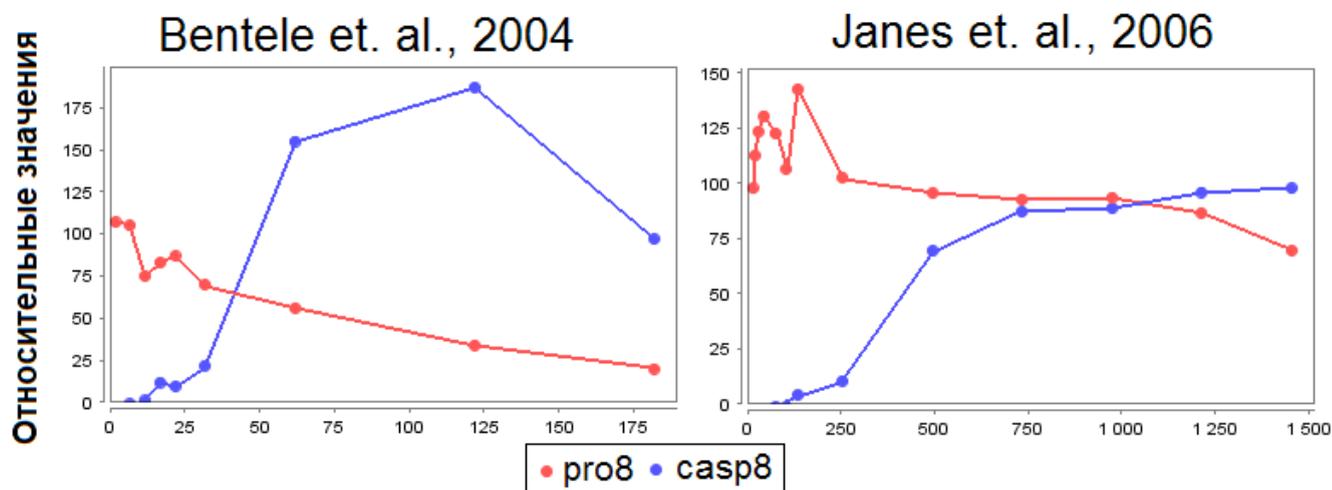


Рисунок 6.1.2 – Экспериментальные данные по динамике каспазы-8 (casp8) и прокаспазы-8 (pro8) при индукции апоптоза

Следующим шагом распределим требования по итерациям. В данном случае, для примера, пусть будет:

- итерация 1 – требование 1;
- итерация 2 – требование 2;
- итерация 3 – требование 3.

6.1.3 Приемочные тесты

В соответствии с предложенным подходом сначала создадим в ПК BioUML набор тестов для каждой итерации (рисунок 6.1.3), как было описано в разделе 3.10 (рисунок 3.5):

Unit test				
	Name	State	Info	Time limit
1	Steady-state case			10.0
2			Steady state: cleaved PARP	10.0
3	CD95L signaling	CD95L signaling		50.0
4		CD95L signaling	Experiment: Bentele et. al., procaspase-8	20.0
5		CD95L signaling	Experiment: Bentele et. al., caspase-8	30.0
6	TNFa signaling	TNFa signaling		40.0
7		TNFa signaling	Experiment: Janes et. al., procaspase-8	10.0
8		TNFa signaling	Experiment: Janes et. al., caspase-8	30.0

Diagram showing iteration requirements:

- Итерация 1 (red box) points to row 1.
- Итерация 2 (blue box) points to row 3.
- Итерация 3 (black box) points to row 6.

Рисунок 6.1.3 – Пример набора приемочных тестов для первых трех итераций по созданию модульной модели апоптоза

6.1.4 Кодирование

В соответствии с таблицей 3.10.1 (раздел кодирование) код должен быть написан с учетом согласованных стандартов, а именно: SBGN, SBML и Antimony.

При визуальном создании моделей ПК BioUML автоматически выполняет эти требования. Для этого предназначены специальные типы диаграмм:

- SBML model in SBGN notation – для немодульной модели;
- SBML 'comp' model in SBGN notation – для модульной модели или отдельного модуля.

Соответственно эти диаграммы хранятся в формате SBML, используя тег <annotation> для хранения информации для графического представления в расширенной графической нотации SBGN Process Diagram. Также по такой диаграмме автоматически генерируется ее текстовое представление в формате Antimony, и происходит синхронизация между графическим и текстовым представлениями модели.

ПК BioUML обеспечивает 2 варианта коллективной собственности и совместной работы над моделью:

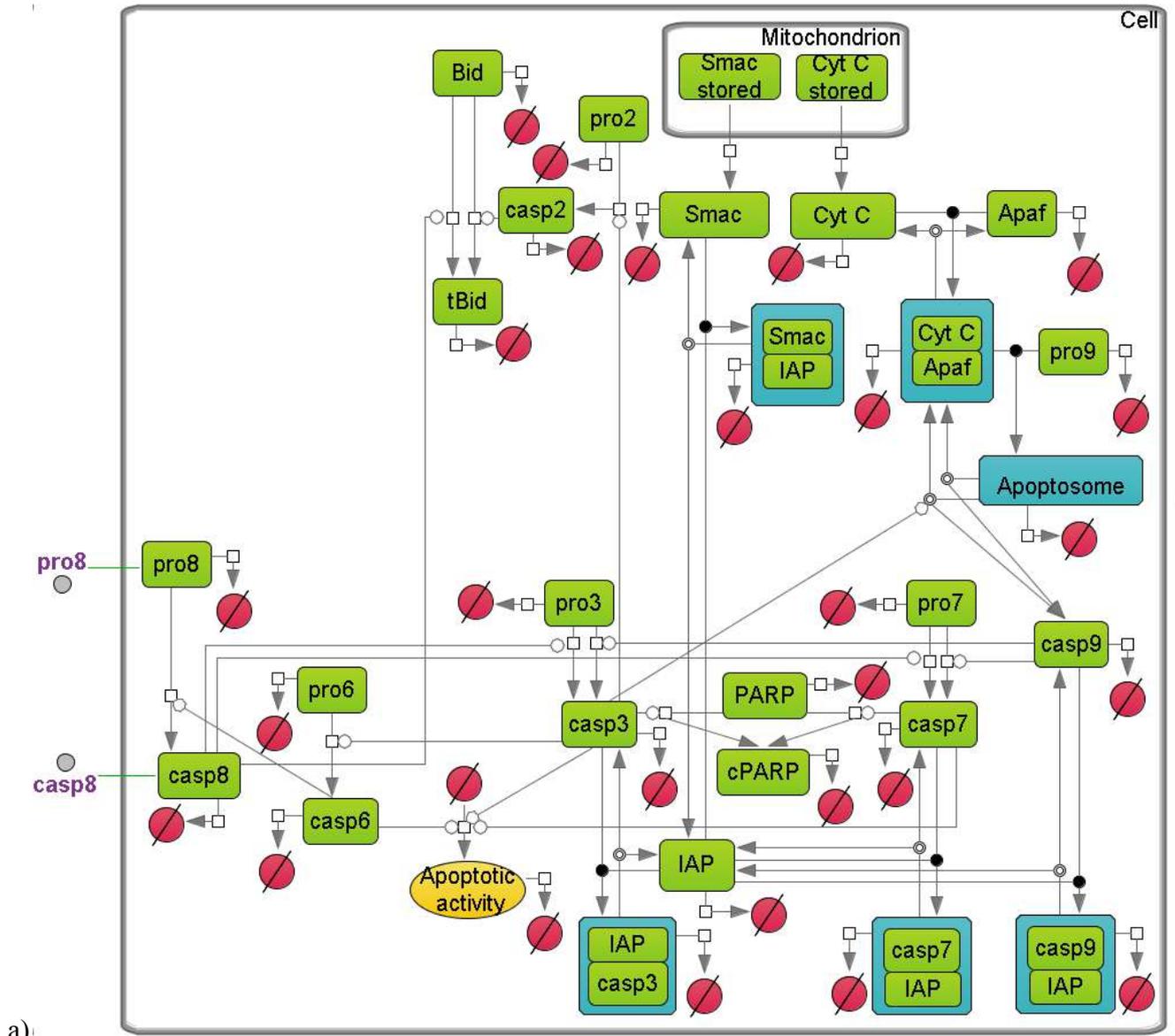
- использование git репозитория для коллективного владения кодом;
- использование "облачной" версии ПК BioUML, которая обеспечивает возможность одновременной работы нескольких пользователей над моделью через веб-интерфейс, подобно Google documents.

6.1.5 Итерация 1

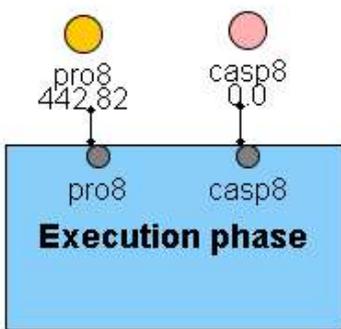
На рисунке 6.1.4.а представлена детальная структура подмодели (модуля) для эффекторной фазы апоптоза в графической нотации SBGN. Для дальнейшей интеграции этого модуля в общую модель апоптоза вынесем 2 порта типа "контакт" для каспазы- и прокаспазы-8. Напомним, порт типа "контакт" обозначает, что значение соответствующей переменной модели может изменяться как этим, так и другими модулями, т.е. вычисляется по значению на предыдущем шаге с учетом всех требуемых изменений.

На рисунке 6.1.5 показаны результаты приемочного теста для первой итерации. Как видно, тест прошел успешно. При нажатии на кнопку "View" пользователь может просмотреть график изменения концентрации расщепленного белка PARP (сPARP) и увидеть, что динамика отсутствует и концентрация остается равной нулю.

Таким образом, первая итерация успешно выполнена и можно переходить ко второй итерации.



a)



б)

Рисунок 6.1.4 – Модуль для эффекторной фазы апоптоза, построенный на основе модели Bentele et al., 2004; а) детальная структура модуля в графической нотации SBGN; б) представление модуля с портами типа "контакт" для каспазы и прокаспазы-8

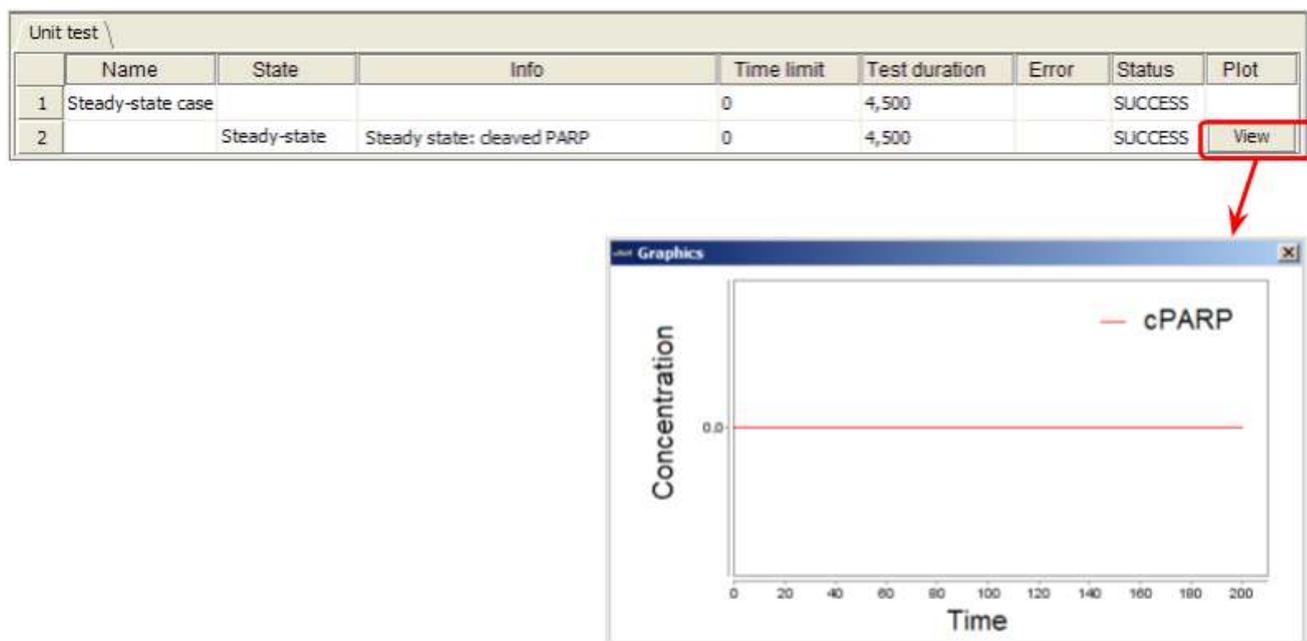


Рисунок 6.1.5 – Результаты приемочного теста для первой итерации

6.1.6 Итерация 2

В ходе второй итерации на основе модели Bentele et al., 2004, разработан модуль лиганда CD95L (рисунок 6.1.6.а) для запуска внешнего пути апоптоза через рецепторы CD95. Используя порты типа "контакт" для каспазы- и прокаспазы-8, мы можем объединить эти 2 модуля в единую модель (рисунок 6.1.6.б). Чтобы проверить правильность полученной модели, запустим соответствующие тесты для второй итерации.

На рисунке 6.1.7 показаны результаты приемочного теста для второй итерации. Как видно, тест прошел успешно. В данном случае при нажатии на кнопку "View" пользователь может посмотреть, что результаты моделирования динамики каспазы- и прокаспазы-8 хорошо соответствуют экспериментальным данным. Таким образом, вторая итерация успешно выполнена и можно переходить к третьей итерации.

6.1.7 Итерация 3

В ходе третьей итерации на основе модели Rangamani and Sirovich, 2007 разработан модуль TNF- α (рисунок 6.1.8.а) для запуска внешнего пути апоптоза через рецепторы TNF- α . Используя порты типа "контакт" для каспазы- и прокаспазы-8, мы можем объединить уже 3 модуля в единую модель (рисунок 6.1.8.б). Чтобы проверить правильность полученной модели, запустим соответствующие тесты для третьей итерации.

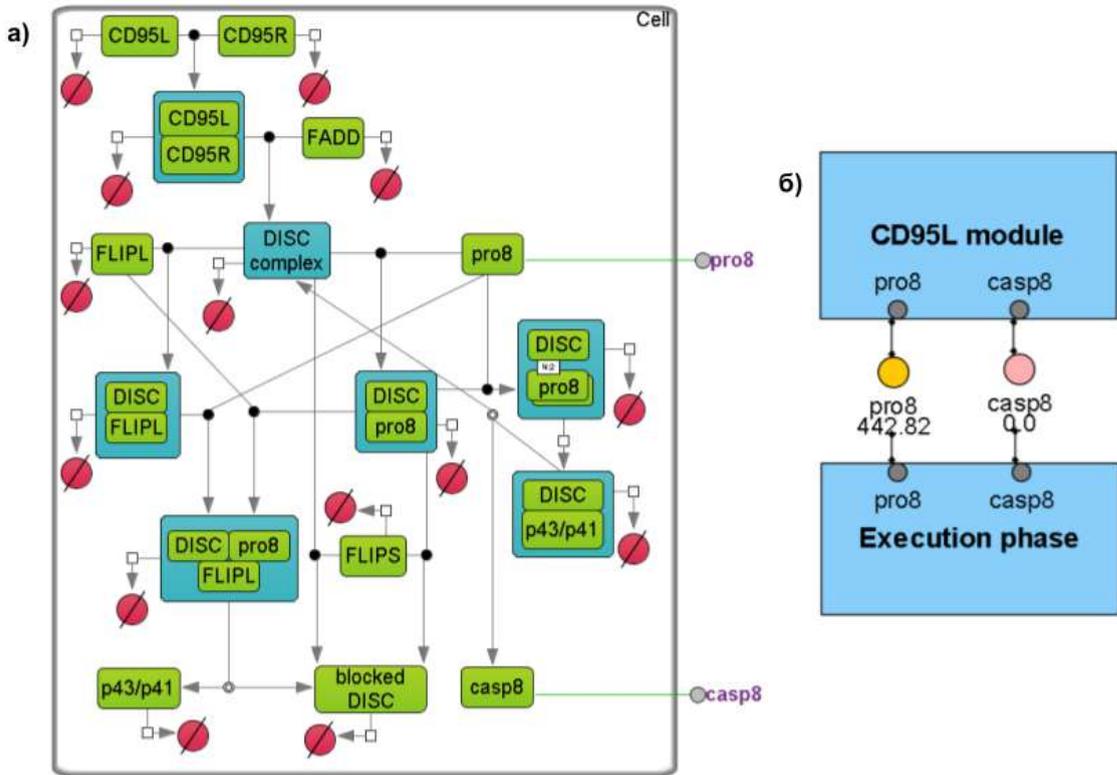


Рисунок 6.1.6 – Модуль CD95L запуска внешнего пути апоптоза через рецепторы CD95, построенный на основе модели Bentele et al., 2004; а) детальная структура модуля в графической нотации SBGN; б) объединение модулей CD95L с модулем эффекторной фазы апоптоза через порты типа "контакт" для каспазы- и прокаспазы-8

Unit test								
	Name	State	Info	Time limit	Test duration	Error	Status	Plot
1	Steady-state case			0	4,500		SUCCESS	
2			Steady state: cleaved PARP	0	4,500		SUCCESS	View
3	CD95L signaling	CD95L signaling		0	2,657		SUCCESS	
4		CD95L signaling	Experiment: Bentele et. al., procaspase-8	0	2,657		SUCCESS	View
5		CD95L signaling	Experiment: Bentele et. al., caspase-8	0	2,657		SUCCESS	View

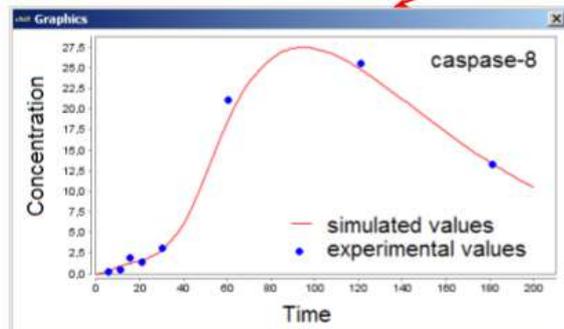
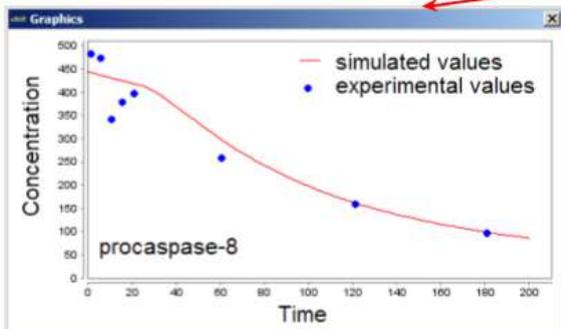


Рисунок 6.1.7 – Результаты приемочного теста для второй итерации

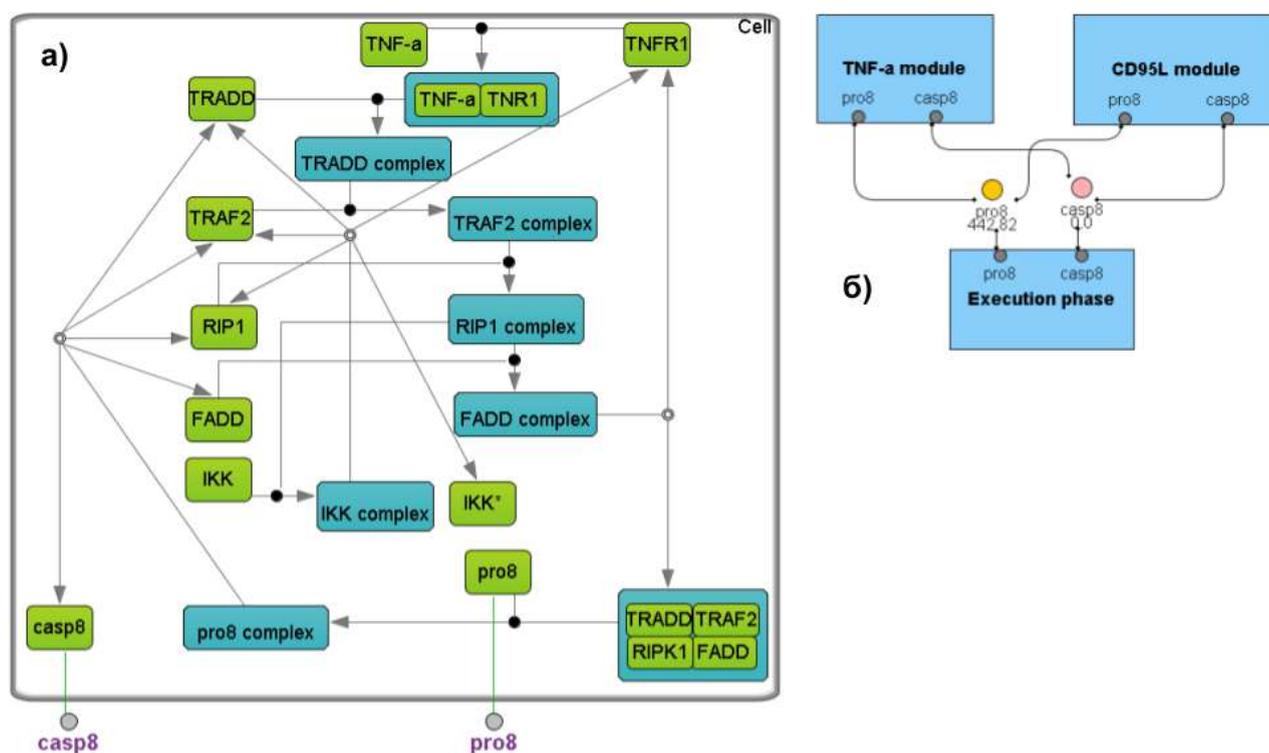


Рисунок 6.1.8 – Модуль TNF- α запуска внешнего пути апоптоза, построенный на основе модели Rangamani and Sirovich, 2007; а) детальная структура модуля в графической нотации SBGN; б) объединение модулей TNF- α , CD95L и эффекторной фазы апоптоза через порты типа "контакт" для каспазы-8 и прокаспазы-8

На рисунке 6.1.9 показаны результаты приемочного теста для третьей итерации. На этот раз тест завершился с ошибкой. Пользователь может просмотреть как подробное сообщение об ошибке, так и результаты моделирования для тестов, которые не прошли. В данном случае мы видим, что результаты моделирования динамики прокаспазы-8 не соответствуют экспериментальным данным Janes et al., 2006.

Этот пример хорошо показывает, что если мы попробуем механистически объединить фрагменты из двух разных моделей, в данном случае Bentele et al., 2004 и Rangamani and Sirovich, 2007, то получившийся результат скорее всего не будет соответствовать экспериментальным данным. Тогда необходим более тщательный анализ объединенной модели, ее модификация и оценка параметров, чтобы результаты численного моделирования соответствовали экспериментальным данным.

Данный процесс подробно описан в работе (Кутумова, 2012).

Unit test								
	Name	State	Info	Time limit	Test duration	Error	Status	Plot
1	Steady-state case			0	4,500		SUCCESS	
2			Steady state: deaved PARP	0	4,500		SUCCESS	View
3	CD95L signaling	CD95L signaling		0	2,657		SUCCESS	
4		CD95L signaling	Experiment: Bentele et. al., procaspase-8	0	2,657		SUCCESS	View
5		CD95L signaling	Experiment: Bentele et. al., caspase-8	0	2,657		SUCCESS	View
6	TNFa signaling	TNFa signaling		0	3,844		ERROR	
7		TNFa signaling	Experiment: Janes et. al., procaspase-8	0	3,844	The deviation...	ERROR	View
8		TNFa signaling	Experiment: Janes et. al., caspase-8	0	3,844	The deviation...	ERROR	View

The deviation between the table data and the simulation result is greater than the maximum allowable value.

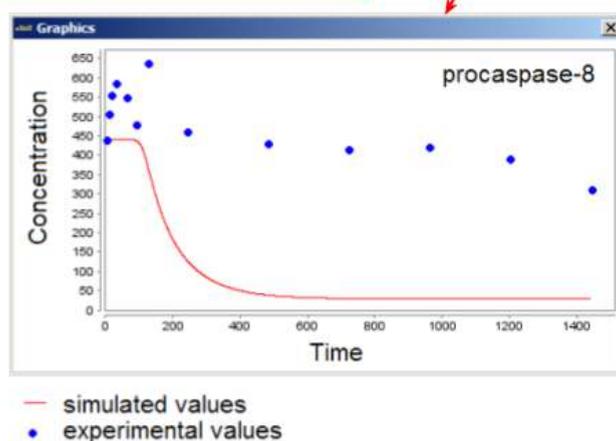


Рисунок 6.1.9 – Результаты приемочного теста для третьей итерации. Последний тест завершился с ошибкой и выдал сообщение о том, что отклонение между табличными данными и результатом моделирования превосходит максимально допустимое значение

6.1.8 Итоговая модель

На рисунке 6.1.10 представлена итоговая модульная модель апоптоза. Она включает: 13 модулей, 280 белков и их комплексов, 372 реакции и 459 параметров. Краткое описание модулей, статистика по компонентам, входящим в их состав, а также использованные для их построения модели приведены в таблице 6.1.1.

При создании данной модели были разработаны новые подмодели для сигнальных путей CD95 и NF-κB путем разбиения существующих моделей на отдельные функциональные модули, их упрощения и объединения в комплексную модульную модель с сохранением воспроизведения экспериментальных данных исходных моделей. Более подробное описание дано в работах Kutumova et al., 2012; Kutumova et al., 2013b; Кутумова и др., 2012.

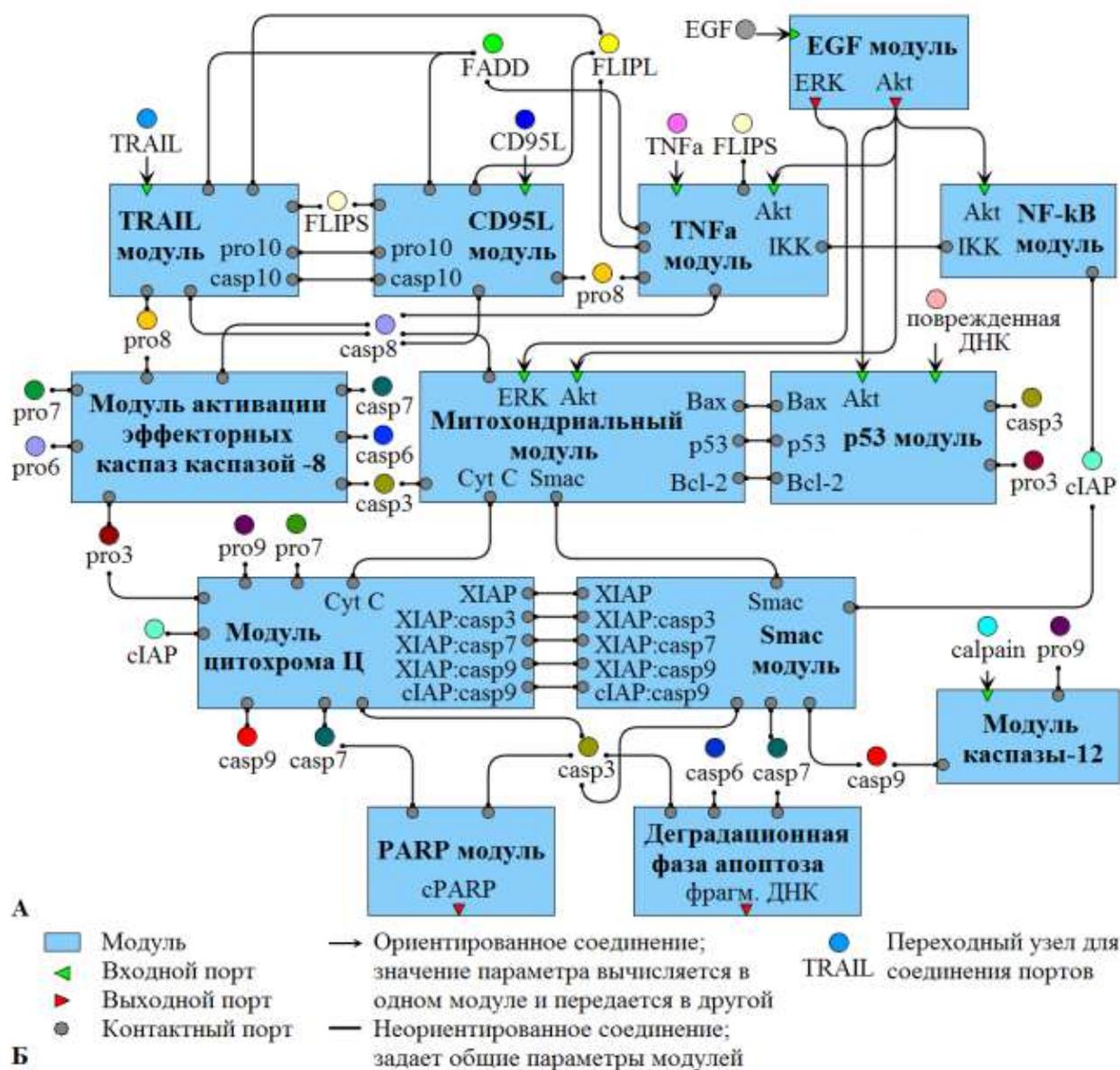


Рисунок 6.1.10 – Модульная модель апоптоза (Kutumova et al., 2012). А. Модель, состоящая из 13 функциональных модулей. Б. Б. Графическая нотация представления модульной модели в ПК BioUML

Таблица 6.1.1 – Краткое описание модулей обобщенной модели апоптоза (Кутумова, 2012)

Модули	Порты			Статистика ¹	Исходные модели и БД ²
	Контактные	Входные	Выходные		
TRAIL	FADD, FLIPS, FLIPL, pro-8, -10, casp-8, -10	TRAIL	–	(18, 17, 26)	<i>Albeck et al., 2008;</i> <i>Toivonen et al., 2011;</i> <i>Bentele et al., 2004;</i> Reactome
CD95L	FADD, FLIPS, FLIPL, pro-8, -10, casp-8, -10	CD95L	–	(19, 17, 26)	<i>Bentele et al., 2004;</i> <i>Toivonen et al., 2011;</i> Reactome
TNF- α	FADD, IKK, FLIPS, FLIPL, pro-8, casp-8	TNF- α , Akt-PP	–	(24, 16, 26)	<i>Rangamani and Sirovich, 2007;</i> <i>Bentele et al., 2004;</i> <i>Toivonen et al., 2011;</i> Reactome
EGF	–	EGF	ERK-PP, Akt-PP	(102, 134, 128)	<i>Schoeberl et al., 2002;</i> <i>Borisov et al., 2009;</i> Reactome
NF- κ B	IKK, cIAP1	Akt-PP	–	(26, 53, 53)	<i>Hoffmann et al., 2002;</i> <i>Rangamani and Sirovich, 2007</i>
Митохондриальный модуль	Bax, Bcl-2, p53, Cyt C, Smac, casp-3, -8	ERK-PP, Akt-PP	–	(52, 48, 71)	<i>Albeck et al., 2008;</i> <i>Hamada et al., 2009;</i> <i>Bentele et al., 2004;</i> <i>Bagci et al., 2006;</i> Reactome
p53	Bax, Bcl-2, p53, pro-3, casp-3	damaged DNA, Akt-PP	–	(25, 34, 51)	<i>Hamada et al., 2009;</i> Reactome
Активация эффекторных каспаз каспазой-8	pro-3, -6, -7, -8, casp-3, -6, -7, -8	–	–	(8, 5, 10)	<i>Bentele et al., 2004</i>
Цитохром C	Cyt C, XIAP, cIAP1, pro-3, -7, -9, casp-3, -7, -9, XIAP:casp-3,-7,-9, cIAP1:casp-9	–	–	(22, 29, 43)	<i>Bentele et al., 2004;</i> <i>Legewie et al., 2006;</i> Reactome
Smac	Smac, XIAP, cIAP1, casp-3, -7, -9, XIAP:casp-3,-7,-9, cIAP1:casp-9	–	–	(17, 12, 14)	<i>Bentele et al., 2004;</i> <i>Albeck et al., 2008;</i> Reactome
Каспаза-12	pro-9, casp-9	Calpain	–	(5, 2, 4)	<i>Fan et al., 2005</i>
PARP	casp-3, -7	–	cPARP	(4, 2, 4)	<i>Bentele et al., 2004</i>
Деградиционная фаза апоптоза	casp-3, -6, -7	–	fragmented DNA	(8, 3, 3)	Reactome

¹ В скобках слева направо указано число веществ, реакций и параметров модулей.² Курсивом выделены курируемые модели, сохраненные в БД BioModels.

6.2 Модульная модель регуляции генной экспрессии при физической нагрузке в скелетных мышцах

Данная работа выполнялась в рамках гранта РФФИ 17-00-00296 "Построение модульной математической модели, связывающей метаболические процессы с регуляцией генной экспрессии в скелетной мышце при функциональных сдвигах, связанных с изменением уровня двигательной активности" под руководством автора.

Используя предложенную выше технологию работа также была разделена на 3 итерации:

- построение модульной модели энергетического метаболизма в клетках скелетной мышцы (Киселев и др., 2019а);
- математическая модель, связывающая Ca^{2+} -зависимый сигнальный путь с регуляцией экспрессии генов в клетках скелетной мышцы человека (Акбердин и др., 2020). При построении этой модели использовались экспериментально подтвержденные районы связывания ТФ CREB, FOS и JUN семейств в промоторных областях генов *NR4A2*, *NR4A3* и *PPARGC1A* из БД GTRD;
- интеграция описанных выше двух моделей в единую модульную модель, что впервые позволило построить математическую модель, описывающую изменение экспрессии отдельных генов в ответ на физическую нагрузку (Akberdin et al., 2020, 2021).

6.2.1 Модель энергетического метаболизма в клетках скелетной мышцы

В качестве исходной была использована комплексная модель энергетического метаболизма клетки скелетной мышцы, разработанная Li Y. et al., 2009. На первом этапе данная модель была воспроизведена и разделена на модули (рисунок 6.2.1), при этом модули могут состоять из вложенных подмодулей (рисунок 6.2.2). Ниже перечислены и кратко описаны основные модули полученной модели (Киселев и др., 2019а):

1. Артерии – представлены как компартмент фиксированного объема, содержащий основные рассматриваемые в модели метаболиты (в скобках даны соответствующие сокращения на рисунках ниже): H^+ (H), O_2 (O_2), CO_2 (CO_2), лактат (Lac), глюкоза (Glc), пируват (Pyr), аланин (Ala), глицерин (Glr), свободные жирные кислоты (FFA).
2. Вены – также представлены как компартмент фиксированного объема, содержащий основные рассматриваемые в модели метаболиты.
3. Поток крови – модуль, описывающий поток метаболитов из артерий через кровеносные капилляры в межклеточное пространство мышечных волокон и, затем, в вены. В представленной модели предполагается, что для каждого метаболита его концентрация в капиллярах, межклеточном пространстве мышечных волокон и венах одинакова. Кроме того, для каждого метаболита задана реакция, которая описывает перенос метаболитов из артерий на венозный конец капилляра. Транспорт – модуль, описывающий реакции транспортного обмена основных метаболитов между капиллярами/межклеточным

пространством и мышечными волокнами. Для каждого метаболита задана соответствующая реакция переноса.

4. Транспорт – модуль, описывающий транспорт основных метаболитов между цитозолем и митохондрией.
5. Мышечная ткань – модуль, иерархически описывающий структуру мышечного волокна (рисунок 6.2.2). Он состоит из следующих модулей:
 - Цитозоль – модуль основных метаболических процессов, протекающих в цитоплазме скелетных мышц: гликолиз, гликогенолиз, метаболизм липидов (рисунок 6.2.3).
 - Митохондрия – модуль основных метаболических процессов, протекающих в митохондриях скелетных мышц: окисление пирувата (цикл Кребса), окисление жирных кислот, окислительное фосфорилирование.

Адаптация математической модели к опубликованным данным (Li Y. et al., 2009) показала, что модель, реализованная на основе модульного подхода в ПК BioUML, достаточно точно воспроизводит ранее опубликованную динамику (Киселев и др., 2019а). Это относится к изменению концентрации ключевых метаболитов гликолиза и окислительного фосфорилирования (рисунок 6.2.4), а также скоростей потребления кислорода и синтеза АТФ в различных метаболических путях при имитации ишемического состояния (Киселев и др., 2019а). При этом стоит подчеркнуть, что экспериментально наблюдаемая динамика в метаболической системе клеток скелетной мышцы человека воспроизводится с помощью модели в BioUML только при изменении значений кинетических параметров в реакциях гликолиза по сравнению с опубликованными значениями модели Li Y. et al., 2009.

Для моделирования метаболических изменений в клетках скелетных мышц здоровых молодых мужчин в ответ на максимальную аэробную физическую нагрузку использовали среднее значение максимальной мощности (W), достигнутой в тесте с непрерывно возрастающей нагрузкой на велоэргометре (общепринятый подход для тестирования аэробной выносливости человека), которое составило 234 Вт. Согласно экспериментальным данным продолжительность работы с такой интенсивностью составляет 2–5 минут. Однако расчеты модели при физической нагрузке проводились более длительном интервале времени (5–15 минут) с целью обнаружения нового устойчивого стационарного решения модели по сравнению с функционированием скелетной мышцы в покое. В результате расчетов модульной модели (Киселев и др., 2019а) при симуляции удалось показать, что уровни АТФ и АДФ при физической нагрузке изменяются незначительно (рисунок 6.2.5), в отличие от уровней концентрации креатина и фосфокреатина, что соответствует расчетам модели Li (Li Y. et al., 2009, рисунок 8).

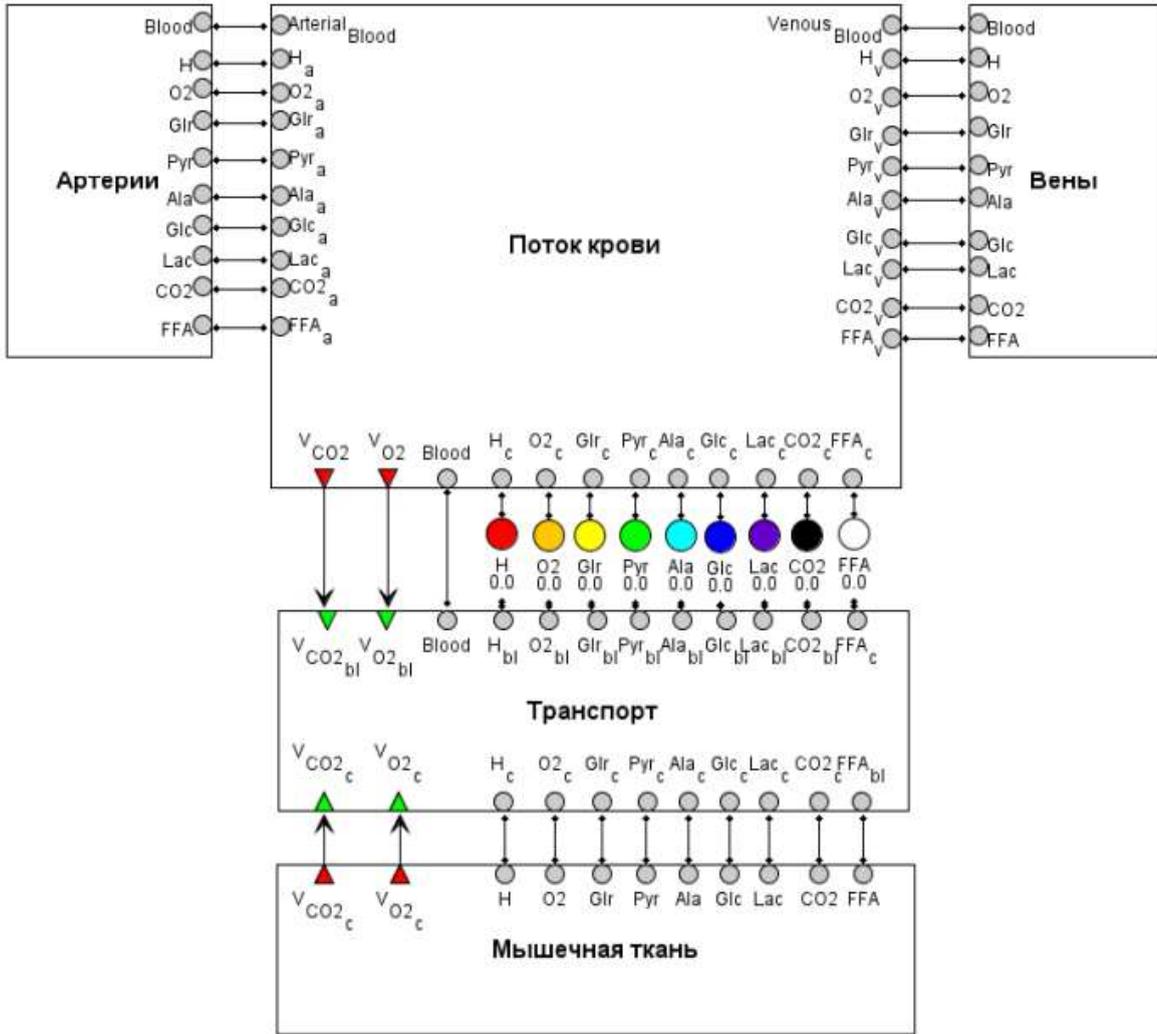


Рисунок 6.2.1 – Модульная модель метаболических процессов в скелетных мышцах (Киселев и др., 2019а)

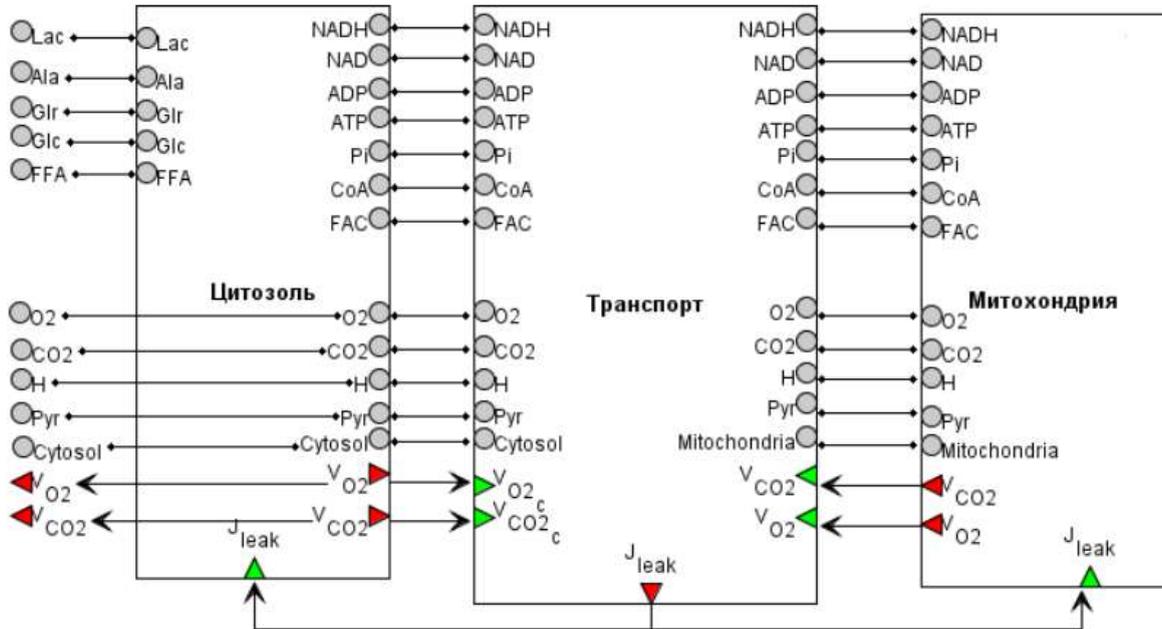


Рисунок 6.2.2 – Структура модуля "Мышечная ткань" (Киселев и др., 2019а)

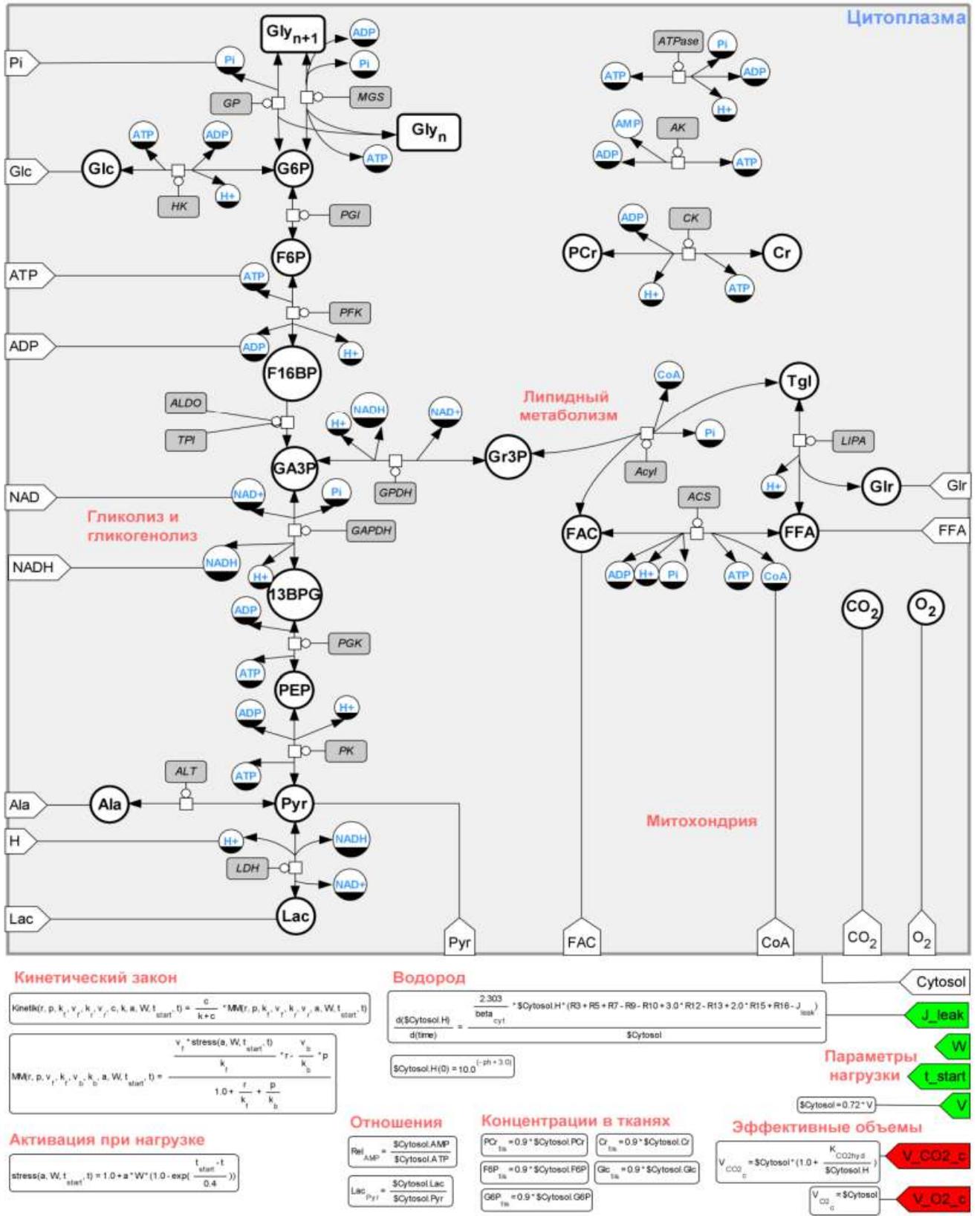


Рисунок 6.2.3 – Модуль «Цитозоль» в расширенной графической нотации SBGN (Киселев и др., 2019a)

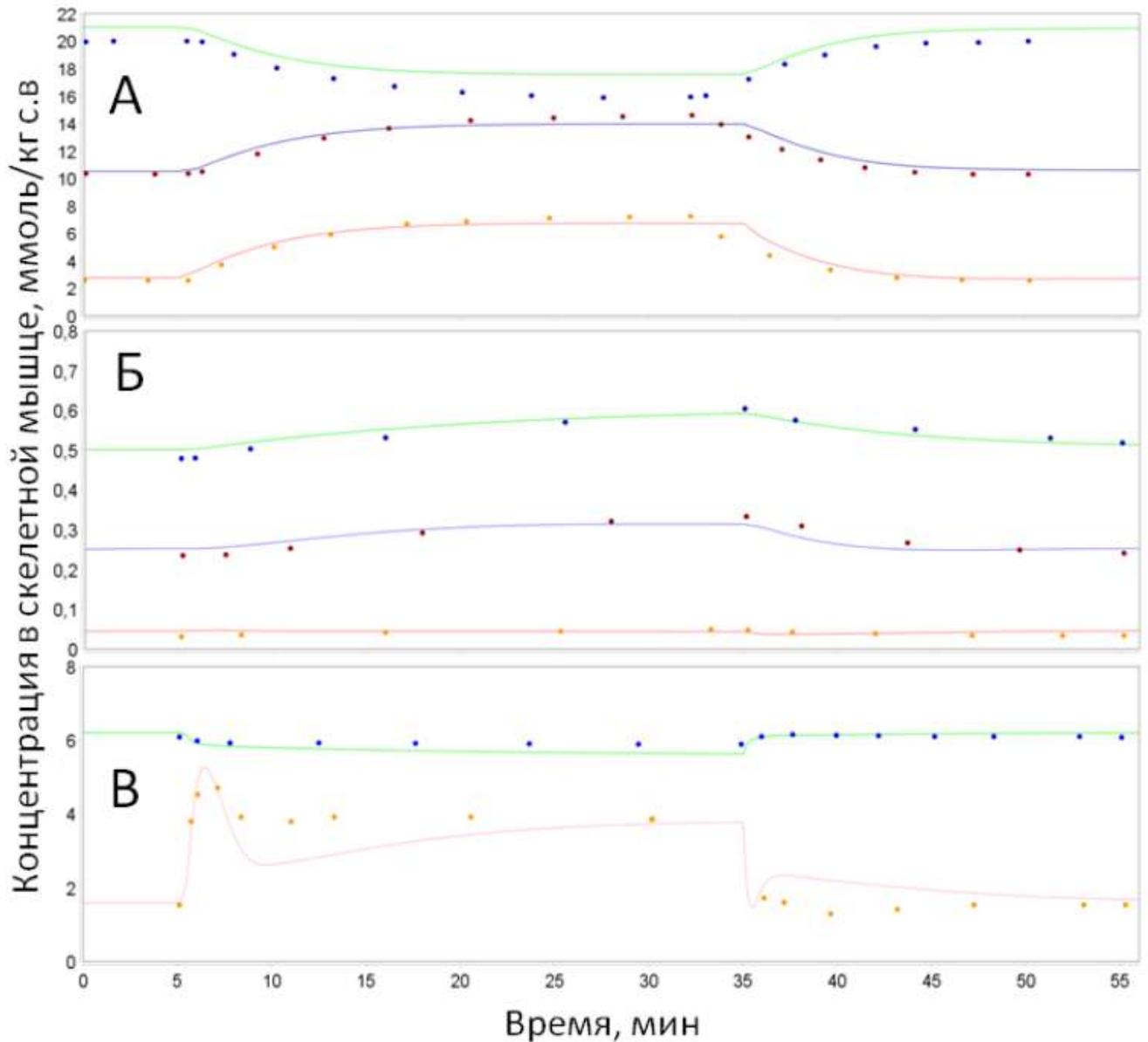


Рисунок 6.2.4 – Изменение концентрации креатинфосфата (зеленая кривая, ЗК), креатина (синяя кривая, СК) и фосфата (красная кривая, КК) (А), Glc (ЗК), G6P (СК) и F6P (КК) (Б), АТФ (ЗК) и NADH/NAD в митохондриях (КК) (В) в покое, при ишемии (время инициации соответствует 5 минутам по оси X) и восстановлении после ишемии (время начала соответствует 35 минутам). Ишемическое состояние моделировалось на основе изменения значения параметра Q_0 (скорость потока крови) на $0.18 \text{ л} \cdot \text{мин}^{-1}$, тогда как процесс восстановления симулировали за счет возврата к изначальному значению этого параметра, равного $0.9 \text{ л} \cdot \text{мин}^{-1}$. Точками на графиках указаны значения соответствующего расчета модели из работы Li Y. et al., 2009, кривая – результаты расчета модели в ПК BioUML (Киселев и др., 2019а)

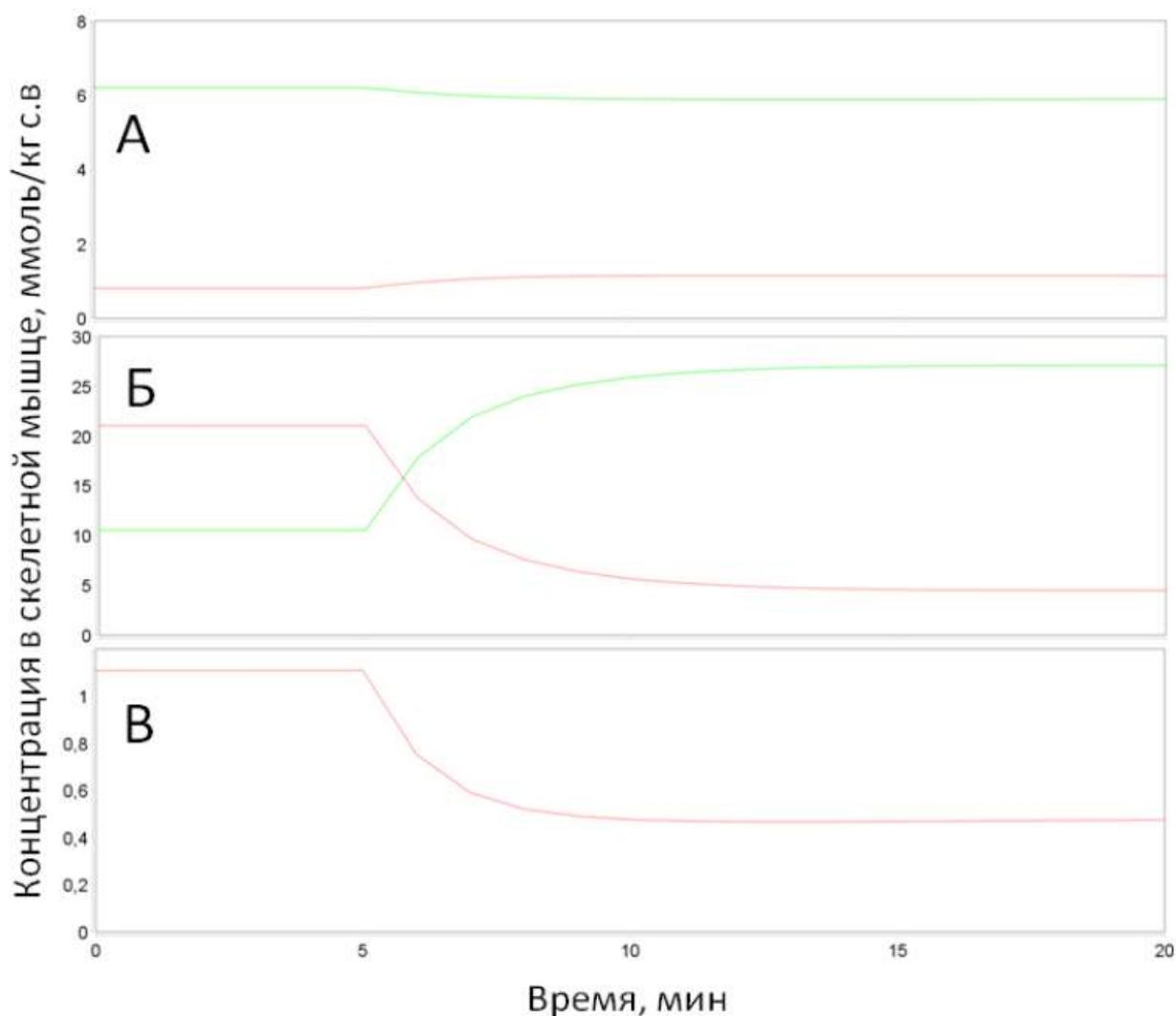


Рисунок 6.2.5 – Изменение концентрации метаболитов в скелетной мышце. ADP (красная кривая, КК) и АТФ (зеленая кривая, ЗК) (А) креатинфосфата (КК) и креатина (ЗК) (Б), АТФ\АДР (В) в скелетных мышцах в состоянии покоя и при физической нагрузке (время инициации соответствует 5 минутам). Физическая нагрузка моделировалось на основе изменения значения параметра W (мощность физической нагрузки) на 234 Вт. Кривые – результаты расчета модели в ПК BioUML (Киселев и др., 2019а).

6.2.2 Регуляция экспрессии генов через Ca^{2+} -зависимый сигнальный путь

Во время аэробной физической нагрузки в работающей мышце происходят существенные метаболические изменения, вызывающие активацию множества сигнальных молекул. Один из наиболее важных сигнальных путей, активируемых при сократительной активности, это Ca^{2+} -зависимый сигнальный путь (Hawley et al., 2014). Увеличение содержания ионов кальция (Ca^{2+}) в миоплазме, вызванное мышечными сокращениями, значимо влияет на активацию сигнальных белков (Ca^{2+} /кальмодулин-зависимые киназы, кальциневрин и др.) и может быть потенциально вовлечено в регуляцию экспрессии множества генов, ответственных за адаптацию мышечных клеток к аэробным упражнениям (Koullmann and Bigard, 2006).

В исследовании Porov et al., 2019, в ходе сопоставления транскриптомного ответа в работавшей и не работавшей (контрольной) мышце удалось выделить специфический для сократительной активности транскриптомный ответ для 1-го и 4-го часов восстановления после аэробного упражнения (рисунок 6.2.6). Было показано, что доминирующий биологический процесс, ассоциированный с этим ответом – это регуляция транскрипции, т.е. увеличение экспрессии генов, кодирующих различные транскрипционные факторы и коактиваторы. Через час после окончания нагрузки экспрессировались гены, кодирующие транскрипционные факторы, играющие важную роль в регуляции мышечного метаболизма, такие как гены семейств *NR4A*, *AP-1* и *EGR*. Тогда как на четвёртом часе восстановления экспрессировался другой набор генов-регуляторов транскрипции (таких как *PPARGC1A*, *ESRRG* и *VGLL2* гены), также играющих важную роль в регуляции метаболизма мышц (Акбердин и др., 2020).

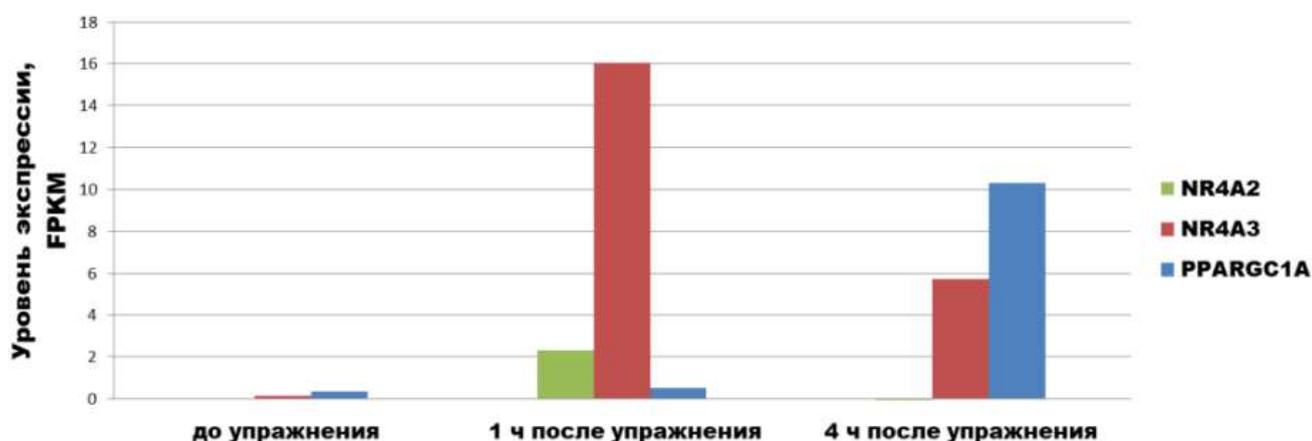


Рисунок 6.2.6 – Уровни экспрессии генов раннего (*NR4A2*, *NR4A3*) и отложенного ответов (*PPARGC1A*), специфические для сократительной активности (разница экспрессий в работающей и неработающей ногах), до, через 1 и 4 часа после физического упражнения на основе анализа транскриптомных данных (Porov et al., 2019)

Для исследования механизмов регуляции двух наблюдаемых паттернов экспрессии генов в ответ на аэробную нагрузку была разработана математическая модель (рисунок 6.2.7; Акбердин и др., 2020), которая описывает последовательную активацию Ca^{2+} -зависимого сигнального пути и экспрессии генов раннего и отложенного ответов в клетке скелетных мышц человека. В качестве генов, моделирующих ранний и отложенный ответы, мы рассмотрели хорошо изученные гены *NR4A2*, *NR4A3* и *PPARGC1A*, соответственно (рисунок 6.2.6). Известно, что белковые продукты этих генов являются важными регуляторами митохондриального биогенеза и углеводно-жирового обмена (Scarpulla, 2008; Olesen et al., 2010; Pearen et al., 2012, 2013), а Ca^{2+} -зависимый сигнальный путь за счет активации CREB1, CREB-связанных транскрипционных факторов и коактиваторов играет ключевую роль в регуляции экспрессии генов *NR4A2*, *NR4A3* и *PPARGC1A* (Yoshioka et al., 2009; Bruno et al., 2014; Goode et al., 2016; Berdeaux and Hutchins, 2019).

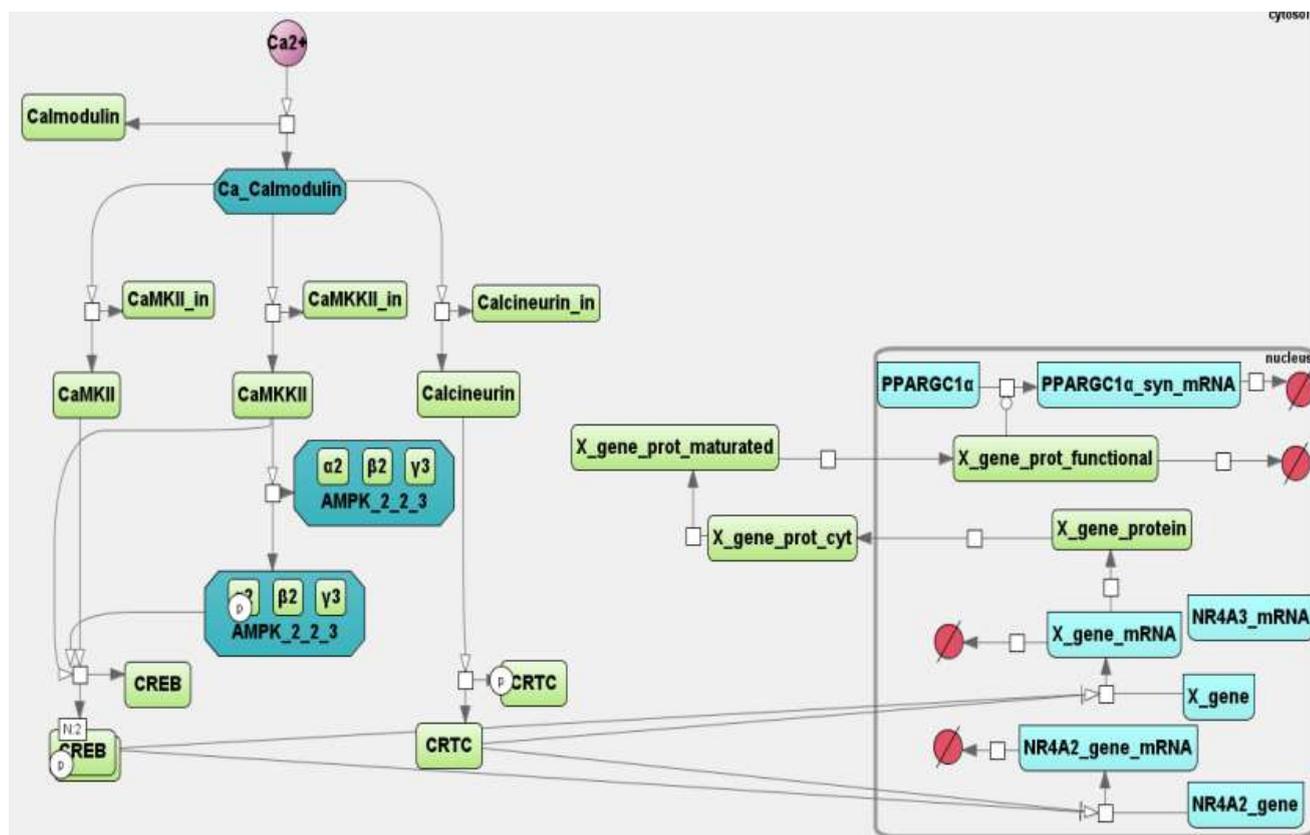


Рисунок 6.2.7 – Модель Ca^{2+} -зависимого сигнального пути, который приводит к активации экспрессии генов раннего (гены семейства *NR4A*) и отложенного (ген *PPARGC1A*) ответов в процессе адаптации мышечных клеток к физической нагрузке. События сигнального каскада происходят в цитоплазме, а процессы генетической регуляции экспрессии генов описаны в ядре мышечной клетки; X_gene – ген раннего ответа, кодирующий промежуточный фактор X, необходимый для активации транскрипции *PPARGC1A* (Акбердин и др., 2020)

Реконструированная сеть Ca^{2+} -зависимого сигнального пути приведена на (рисунок 6.2.7). Резкое изменение миоплазматической концентрации кальция при сократительной активности приводит к связыванию ионов Ca^{2+} с кальмодулином. Формируемые комплексы Ca^{2+} -кальмодулин стимулируют образование активных форм Ca^{2+} -кальмодулин-зависимых протеинкиназ II (CaMKII), CaMK киназ II (CaMKKII) и кальциневрина (кальций-зависимой серин/треонин-фосфатазы). Активация CaMKKII приводит к увеличению уровня фосфорилирования АМФ-зависимой протеинкиназы (АМПК), которая играет одну из ключевых ролей в регуляции внутриклеточного метаболизма скелетных мышц. В дополнение к этому, активность CaMKII, индуцируемая при сократительной активности, приводит к фосфорилированию транскрипционного фактора CREB1, тогда как CREB-регулируемый транскрипционный коактиватор 1, CRTC дефосфорилируется и активируется функционально активным кальциневрином (Акбердин и др., 2020).

Известно, что CREB1 (и, по-видимому, другие CREB-подобные транскрипционные факторы) и его коактиватор CRTC играют важную роль в регуляции экспрессии гена *PPARGC1A* с канонического (проксимального) и с индуцибельного (дистального) промотора

гена (Popov, 2018). Хорошо известно, что в скелетных мышцах человека в ответ на однократную нагрузку экспрессия гена *PPARGC1A* происходит с задержкой (рисунок 6.2.6): максимум транскрипционной активности достигается лишь на ~3–5 часах восстановления (Popov et al., 2019). Поэтому, мы предположили существование промежуточного фактора, регулирующего транскрипцию *PPARGC1A* при активации Ca^{2+} -зависимого сигнального пути, за счет увеличения экспрессии одного из генов раннего ответа.

Анализ транскриптомных данных по специфическому для сократительной активности геному ответу (Popov et al., 2019) показал, что в качестве потенциального промежуточного активатора генов раннего ответа могут выступать CREB-подобные белки, а также белки семейств FOS и JUN. Более того, CREB1 может связываться с ДНК и регулировать экспрессию целевых генов в виде гомо- и гетеродимерного комплекса с белками этих семейств (Hai and Curran, 1991; Newman and Keating, 2003).

Был проведен анализ районов связывания CREB, FOS и JUN семейств в промоторных областях генов *NR4A2*, *NR4A3* и *PPARGC1A* на основе информации из БД GTRD. Его результаты представлены на рисунке 6.2.8 и приложении 6.2. Видно, что ТФ из семейств FOS, JUN и CREB имеют общие сайты связывания в промоторах исследуемых генов раннего и отложенного ответов.

Построенная модель включает 25 биологических молекул (мРНК, белки и низкомолекулярные соединения), представляющих вершины графа взаимодействий, а математическая модель состоит из 14 ОДУ, в основе которых были использованы механизмы моно- и бимолекулярных реакций для пути передачи сигнала, а также обобщенные функции Хилла – для процессов регуляции транскрипции генов (Likhoshvai and Ratushny, 2007). Соответствующие уравнения и начальные значения переменных и параметров приведены в работе Акбердин и др., 2020.

Численный анализ модели показал, что для адекватного воспроизведения экспериментально наблюдаемых изменений экспрессии генов раннего и отложенного ответов, необходимо учитывать синтез и созревание промежуточных транскрипционных факторов, регулирующих транскрипцию генов отложенного ответа, тогда как биоинформатический анализ оригинальных транскриптомных данных позволил предположить, что таковыми транскрипционными факторами могут являться CREB-подобные белки из семейств FOS и JUN, образующие гетеродимерные комплексы с белком CREB1.

Построенная модель качественно и количественно воспроизводит динамику транскрипционной активности генов раннего и отложенного ответа на аэробную физическую нагрузку (рисунок 6.2.9) в соответствии с экспериментальными данными (Popov et al., 2019).

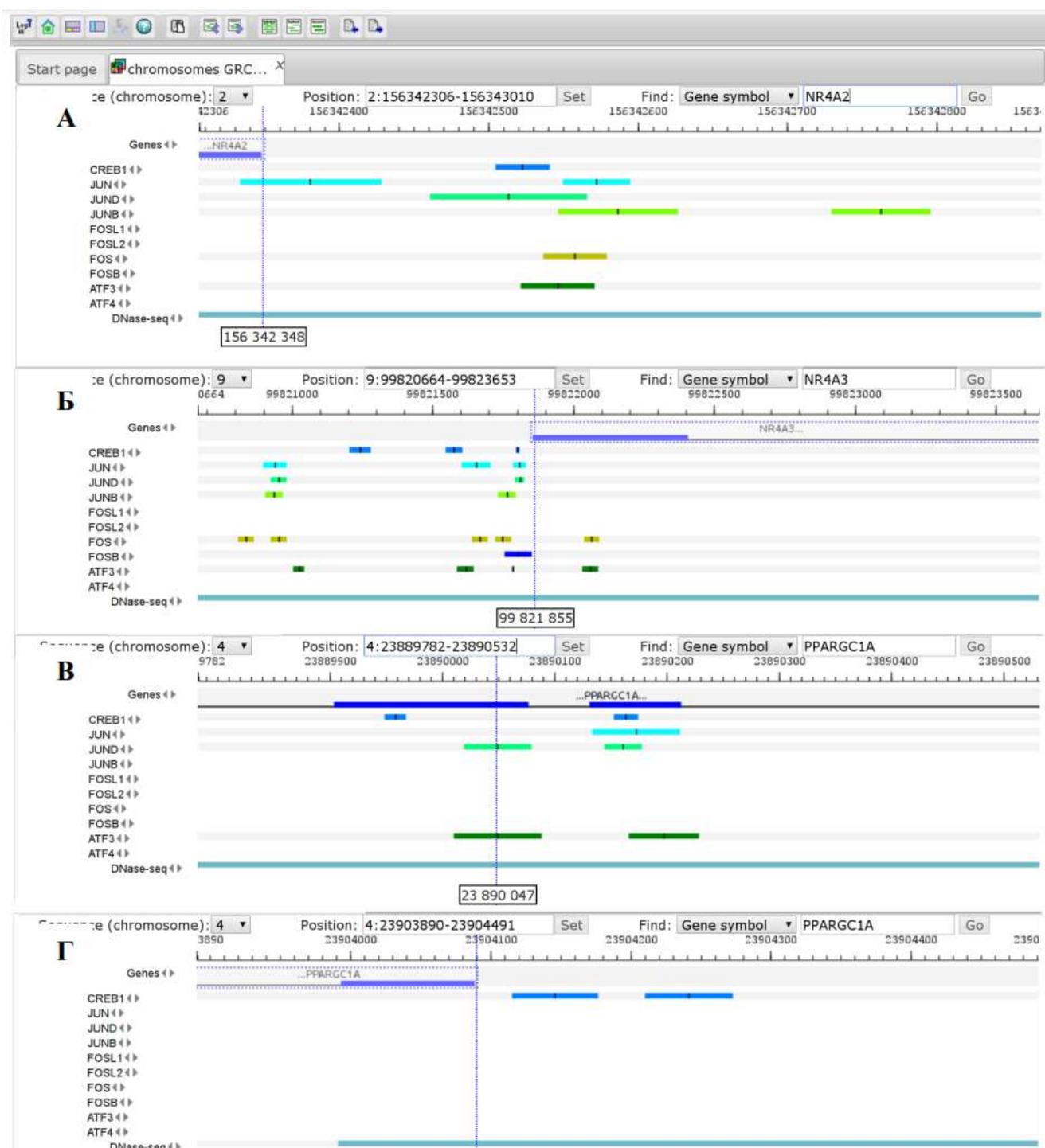


Рисунок 6.2.8 – Экспериментально подтвержденные районы связывания для ТФ семейств CREB, FOS и JUN в промоторных областях генов *NR4A2*, *NR4A3* и *PPARGC1A* по информации БД GTRD в геномном браузере ПК BioUML. Вертикальным пунктиром отмечены старты транскрипции, координата каждого старта транскрипции дана внизу в прямоугольнике: а) промоторная область гена *NR4A2* (ENSG00000153234); б) промоторная область гена *NR4A3* (ENSG00000119508); в) область канонического промотора гена *PPARGC1A* (транскрипт NM_013261, координата старта транскрипта дана по аннотации NCBI версии 109.20190905); г) область альтернативного промотора гена *PPARGC1A* (транскрипт XM_005248132, координата дана по аннотации NCBI версии 109.20190905) (Акбердин и др., 2020)

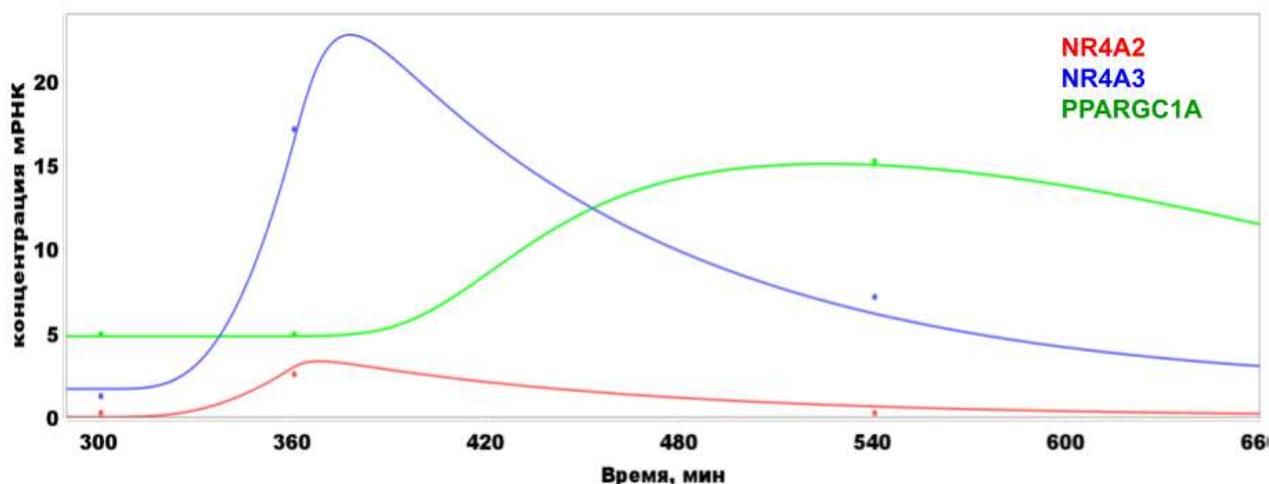


Рисунок 6.2.9 – Численный анализ изменения транскрипционной активности генов раннего и отложенного ответа на аэробную нагрузку. Экспериментальные данные (уровень мРНК в скелетной мышце до, через 1 и 4 часа после физической нагрузки), отмеченные на рисунке точками. По оси X – время расчета (минуты), по оси Y – концентрация мРНК: для экспериментальных данных – в ФРКМ, для теоретических расчетов – в μM . Момент времени, соответствующий 300 и 360 минуте расчета – начало и окончание физической нагрузки, соответственно

6.2.3 Итоговая модульная модель

Итоговая модульная модель состоит из нескольких иерархических уровней (рисунок 6.2.10):

- 1) физиологический уровень – на этом уровне моделируется кровоток и транспорт метаболитов в/из мышцы. Для более реалистичного описания механизмов функционирования мышечной ткани при разных типах нагрузок учтено функционирование мышечных волокон двух типов (быстрые и медленные), имеющих различные паттерны активации в ответ на физическую нагрузку в зависимости от ее интенсивности.
- 2) клеточный уровень – на этом уровне моделируется работа одной мышечной клетки, при этом данный уровень разделен на подуровни: метаболизм, передача сигнала в клетке и регуляция генной экспрессии.

Каждый уровень организации представляет собой отдельный модуль, внутри которого могут быть представлены свои подсистемы-подмодули (например, для метаболического уровня – метаболические пути в модулях “Цитозоль”, “Митохондрии”, “Транспортный обмен” между этими компартментами и между кровью и мышечной клеткой).

Данная модель (все уровни) включает 25 модулей, 238 типов молекул, 185 реакций, 171 ОДУ и 647 параметров. Она подробно описана в приложении к статье Akberdin et al., 2021.

Стоит отметить, что механизм активации, повышающий энергетический метаболизм через увеличение объема мышц, кровотока, скоростей транспорта и метаболических реакций во время упражнений, задавался функцией активации вида:

$$Function(W) = 1 + \alpha_i * W * (1 - e^{-\frac{t_{start}-t}{\tau_i}}) \quad (6.3.1)$$

где W – мощность, α_i – коэффициент активации, τ_i – временная константа, характеризующая изменение скорости транспорта/реакций, t_{start} – время начала упражнений. Во время выполнения упражнения (т. е. при $W > 0$) скорости реакций, транспорта, кровотока и объем мышц умножались на вышеприведенную функцию активации.

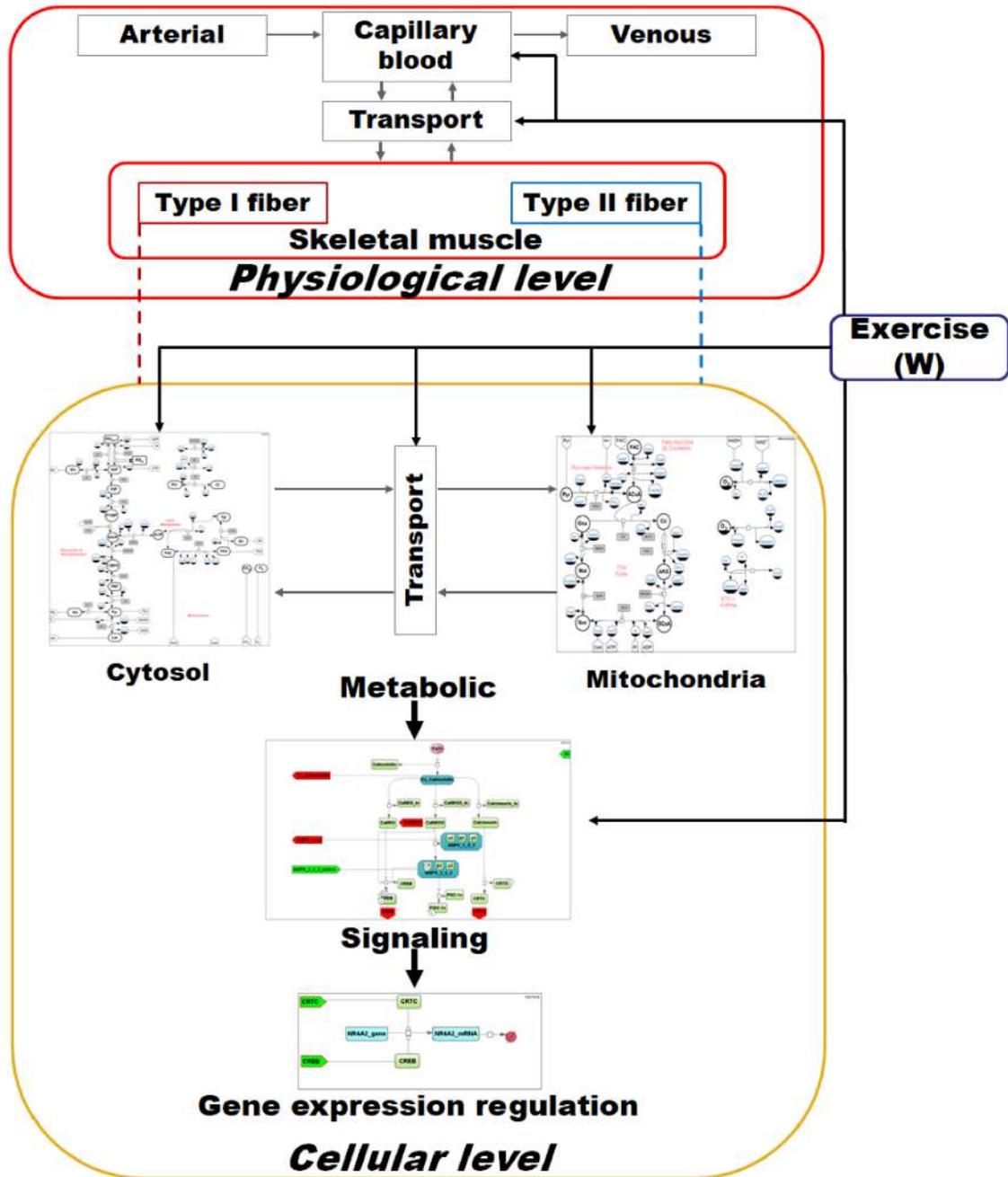


Рисунок 6.2.10 – Иерархическая структура модульной модели для регуляции генной экспрессии при физической нагрузке в скелетных мышцах (Akberdin et al., 2021).

Для валидации численных расчетов модели было исследовано динамическое поведение системы в ответ на различные типы нагрузок: тест с повышающейся нагрузкой (incremental ramp test – постепенное повышение W в модели) и различные по интенсивности непрерывные аэробные упражнения (от низкоинтенсивных до высокоинтенсивных: 50-100 % от VO_{2max} , где VO_{2max} – максимальная скорость потребления кислорода организмом, регистрируемая во время теста с возрастающей физической нагрузкой на велоэргометре до отказа). Численные результаты расчетов модели достаточно хорошо соответствуют экспериментальным измерениям (Roussel et al., 2003; Krstrup et al., 2004; Greiner et al., 2007; Barker et al., 2008; Cannon et al., 2014; Bartlett et al., 2020) в соответствующих дизайнах физических нагрузок (рисунок 6.2.11).

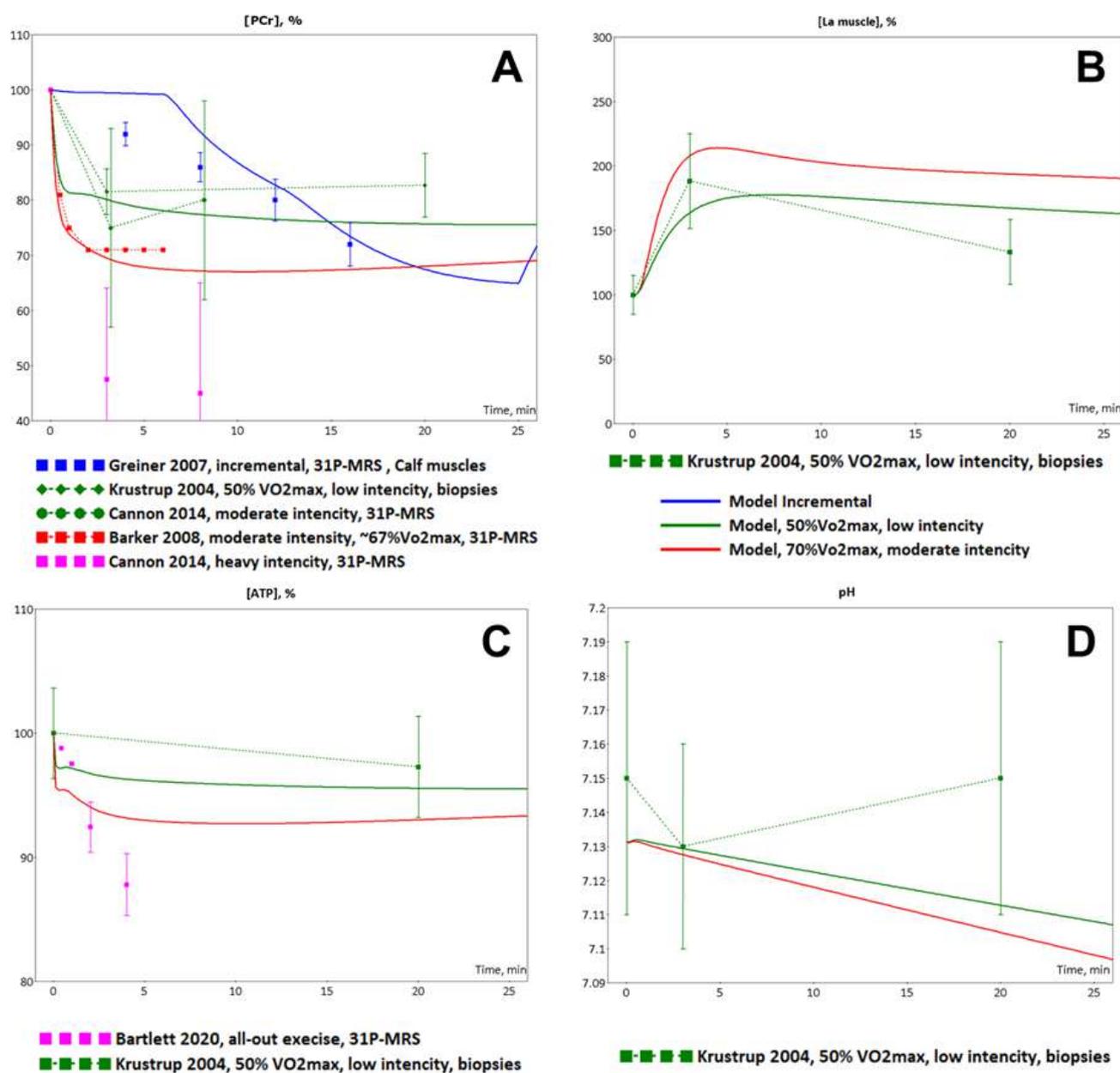


Рисунок 6.2.11 – Валидация модели. Сравнение результатов симуляции модели с экспериментальными данными, отражающими различные типы физических нагрузок: теста с

повышающейся нагрузкой и непрерывные упражнения с низкой, умеренной и высокой интенсивностями (А) Динамическое изменения концентрации фосфокреатина, PCr в мышечных волокнах, предсказанное в модели в процессе incremental ramp теста (синяя линия), низкоинтенсивного упражнения (зеленая линия) и упражнения с умеренной аэробной нагрузкой (красная линия) (ось X – время выполнения упражнения в минутах) и экспериментальные оценки его концентрации в течение incremental ramp теста (синие точки, Greiner et al., 2007), низкоинтенсивного упражнения и упражнения умеренной интенсивности (темно-красные точки, Krstrup et al., 2004; зеленые точки, Barker et al., 2008; розовые точки, Cannon et al., 2014) и упражнения высокой интенсивности (темно-розовые точки, Cannon et al., 2014); (В) Динамическое изменения концентрации лактата, La в мышечных волокнах, предсказанное в модели в процессе incremental ramp теста (синяя линия), низкоинтенсивного упражнения (зеленая линия) и упражнения с умеренной аэробной нагрузкой (красная линия) (ось X – время выполнения упражнения в минутах) и экспериментальные оценки его концентрации в течение низкоинтенсивного упражнения (темно-красные точки, Krstrup et al., 2004); (С) Динамическое изменения концентрации АТФ, АТР в мышечных волокнах, предсказанное в модели в процессе incremental ramp теста (синяя линия), низкоинтенсивного упражнения (зеленая линия) и упражнения с умеренной аэробной нагрузкой (красная линия) (ось X – время выполнения упражнения в минутах) и экспериментальные оценки её концентрации в течение низкоинтенсивного упражнения (темно-красные точки, Krstrup et al., 2004) и упражнения на полную мощность (all-out intensity; красные точки, Bartlett et al., 2020); (D) Динамическое изменения рН в мышечных волокнах, предсказанное в модели в процессе incremental ramp теста (синяя линия), низкоинтенсивного упражнения (зеленая линия) и упражнения с умеренной аэробной нагрузкой (красная линия) (ось X – время выполнения упражнения в минутах) и экспериментальные его значения в течение низкоинтенсивного упражнения (темно-красные точки (значение \pm ст.отклонение), Krstrup et al., 2004) (Akberdin et al., 2021)

Однако, стоит отметить, что текущая версия модели имеет ряд ограничений, в частности:

- не учитывается эффект утомления мышечных волокон в течение теста с повышающейся нагрузкой;
- не учитываются механизмы рекрутирования мышечных волокон I и II типов при физических нагрузках;
- не учитываются обратные связи между рассматриваемыми иерархическими уровнями организации мышечной клетки: метаболическим, сигнальным и уровнем регуляции экспрессии генов. Например, сигнальные киназы влияют на активность ферментов, катализирующих реакции гликолиза и цикла трикарбоновых кислот. Другой пример – экспрессия генов, кодирующих ферменты гликолиза или малат-аспартатного обмена в мышечных клетках может регулироваться PPARGC1A.

Предлагаемый эволюционный подход создания биологических моделей позволит в дальнейшем осуществить развитие модели для преодоления описанных выше ограничений.

6.3 Модели COVID-19

Данная работа выполнялась при поддержке гранта РФФИ 20-04-60355 "Разработка мультимасштабной иммуно-эпидемиологической математической модели COVID-19 с учетом воздействия на экономику региона и сценариев действия органов власти" под руководством автора диссертационной работы.

Итерационная разработка мультимасштабной модели COVID-19 шла по трем направлениям:

- эпидемиологическая модель на основе систем дифференциальных уравнений;
- агентная модель эпидемиологического процесса;
- модель инфекционного процесса в организме человека.

6.3.1 Эпидемиологические модели

В рамках предложенной технологии были разработаны три версии (последовательно улучшающие друг друга) эпидемиологической модели распространения коронавируса SARS-CoV-2, использующие обыкновенные дифференциальные уравнения и дифференциальные уравнения с запаздывающим аргументом. Модели были адаптированы к Новосибирской области, г. Москве, а также трем европейским странам: Германии, Франции и Италии. Помимо этого была реализована агентная модель распространения коронавируса в Новосибирской области.

6.3.1.1. Первая версия модели

В качестве исходной модели (рисунок 6.3.1) для последующего ее эволюционного развития была выбрана модель профессора Вестерхоффа (Westerhoff and Kolodkin, 2020). Она была доработана для описания эпидемий в Москве и Новосибирской области. Основные отличия от исходной модели:

- Добавлена подгруппа не чувствительных к заболеванию людей. Они не заражаются вирусом и не могут заразить других людей благодаря уже существующему иммунитету (Ng et al., 2020; Pinto et al., 2020; Shrock et al., 2020). Доля таких людей в популяции предполагается 25-50% (Doshi, 2020; Mateus et al., 2020).
- Добавлена группа бессимптомных носителей вируса. Носители могут заражать других, но сами не болеют и не проявляют симптомов. Предполагается, что эта группа среди всех чувствительных может составлять до 50 % (Pollock and Lancaster, 2020; Yanes-Lane et al., 2020; Stadler et al., 2021).
- добавлена группа с тяжелым течением болезни – такие больные должны быть помещены в реанимацию, чтобы избежать гибели;

- добавлен расчет ежедневных новых случаев заражения и смерти;
- добавлена вакцинация – общее количество ежедневно вакцинированных делится пропорционально между подходящими группами (чувствительные, чувствительные протестированные, нечувствительные, нечувствительные протестированные, выздоровевшие, выздоровевшие протестированные).

Диаграмма этой модели в графической нотации SBGN приведена на рисунке 6.3.2.

6.3.1.2. Вторая версия модели

Вторая версия модели (рисунок 6.3.3) была построена на основе предыдущей итерации для преодоления двух основных недостатков:

- 1) значение коэффициента социальной дистанции, соответствующего тем или иным противоэпидемическим мерам, подбиралось для соответствия текущим графикам, а не исходя из реального содержания принятых мер;
- 2) при адаптации модели к статистическим данным эпидемии в двух регионах, мы получили два различных набора значений параметров. Однако, по крайней мере часть параметров модели, касающаяся свойств вируса, не должна сильно варьировать от одного региона к другому (при условии совпадения штаммов).

Для решения первой проблемы мы использовали Stringency Index (Hale et al., 2021), разработанный в Blavatnik School of Government (University of Oxford), учитывающий 10 различных противоэпидемических мер, включая закрытие школ, перевод на удаленную работу, отмена публичных мероприятий, запрет на массовые собрания, ограничение внутреннего и международного перемещения, запрет на покидание домов. Значение индекса варьирует между 0 (нет ограничений) и 100 (максимально возможные ограничения). В новой версии модели параметр социальной дистанции не подбирается на основе статистических данных по новым выявленным случаям, а рассчитывается, исходя из текущего значения Stringency Index для данного региона:

$$S = 100 / (100 - SI_Comply * SI (t - SI_Delay)), \text{ где} \quad (6.3.1)$$

- S – параметр социальной дистанции,
- SI – величина Stringency Index,
- SI_Comply – степень соблюдения населением принимаемых ограничений,
- SI_Delay – задержка между принятием противоэпидемических мер и их эффектом.

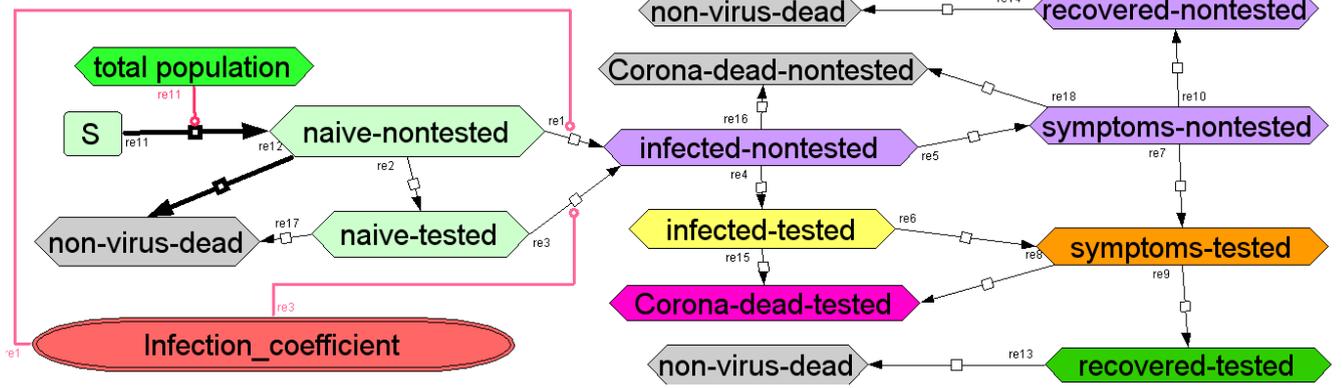


Рисунок 6.3.1 – Модель эпидемиологии COVID-19 (Westerhoff and Kolodkin, 2020)

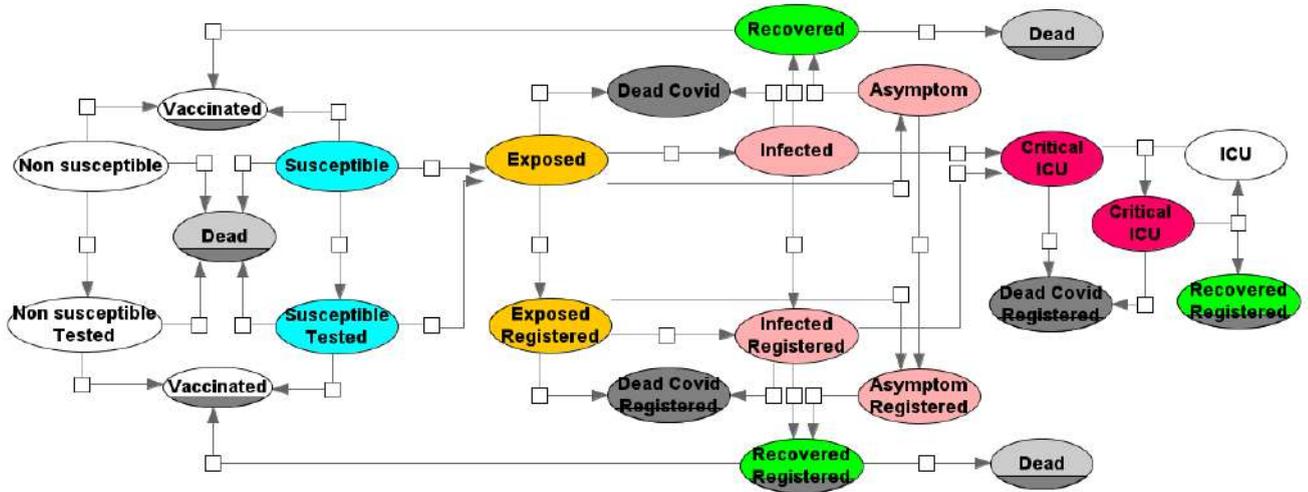


Рисунок 6.3.2 – Первая версия модели эпидемиологии COVID-19

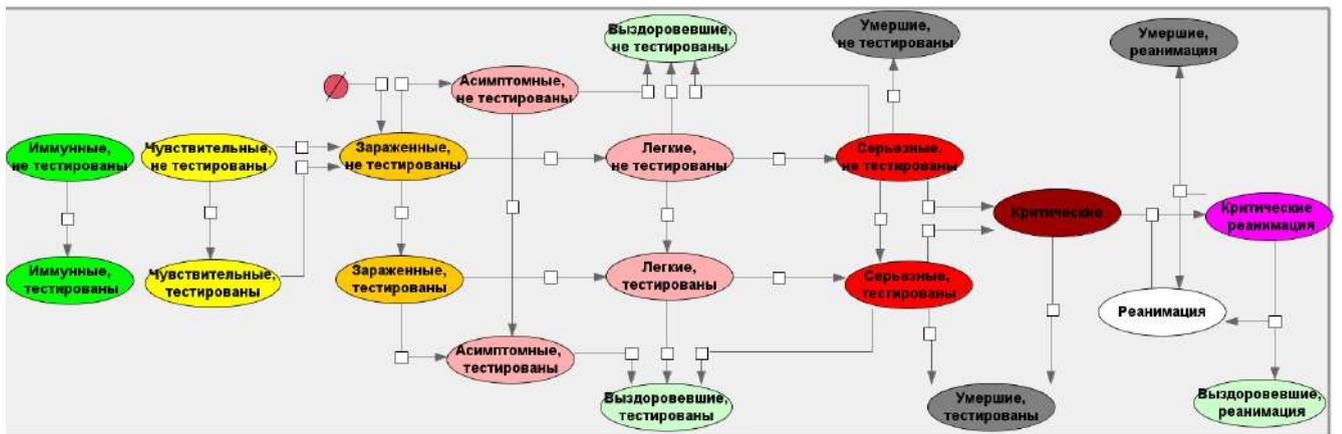


Рисунок 6.3.3 – Вторая версия модели эпидемиологии COVID-19

Также по сравнению с первой итерацией в модели были сделаны следующие изменения:

- 1) Группа инфицированных разделена на имеющих легкие и серьезные симптомы. Предполагается, что серьезные симптомы развиваются со временем у части инфицированных, имеющих легкие симптомы. Для таких пациентов требуется госпитализация.
- 2) Добавлен учет контактного тестирования (т.е. тестирование людей, не демонстрирующих симптомы, проведенное в результате отслеживания контактов уже выявленных зараженных).
- 3) Игнорируются смерти от причин, отличных от COVID-19. Кроме того, зараженные не могут умереть, находясь в инкубационном периоде.

Модель была адаптирована для Новосибирской области, г. Москвы и трех стран Европы (Германия, Франция и Италия). Результаты численных расчетов и их сопоставление со статистическими данными приведены в приложении 6.2. Как видно из приведенных в приложении рисунков, результаты численного моделирования хорошо соответствуют статистическим данным.

6.3.1.3. Третья версия модели

Модели, подобные тем, что представлены на двух предыдущих итерациях, используют законы действующих масс при описании процессов перехода между состояниями (например, из инкубационного периода в симптоматическое) и имеют существенный недостаток, заключающийся в том что модель не всегда может адекватно воспроизвести динамику перехода между состояниями с учетом его временной задержки. Для преодоления этих проблем мы использовали дифференциальные уравнения с запаздывающим аргументом и задание пропорций распределений переходов пациентов из одной группы в другую (Kiselev et al., 2021). Рассмотрим на двух примерах эти подходы более детально.

1. **Инкубационный период.** По имеющимся данным, например для Уханьского штамма, медианное время инкубационного периода SARS-CoV-2 составляет 5 дней и имеются данные по процентилям (таблица 6.3.1).

Модель А – в случае использования закона действующих масс, выход из инкубационного периода описывается дифференциальным уравнением с единственным параметром k :

$$dE/dt = -k * E \quad (6.3.2)$$

Для того чтобы получить медианное время выхода из инкубационного периода, равное определенному значению x , обычно используют значение $k = \ln(2)/x$. Таким образом, при фиксировании медианы мы не имеем возможности получить требуемые процентиля данного

процесса. При использовании $k = \ln(2)/5,1$ мы получим время инкубационного периода, которое соответствует медианному, но имеет другие процентиля (таблица 6.3.1).

Модель В – предлагаемым решением данной проблемы является использование вместо закона действующих масс взвешенной суммы, включающей количество людей в инкубационном периоде с различным временным запаздыванием:

$$dE/dt = -k_1 * E(t - d_1) - \dots - k_n * E(t - d_n) \quad (6.3.3)$$

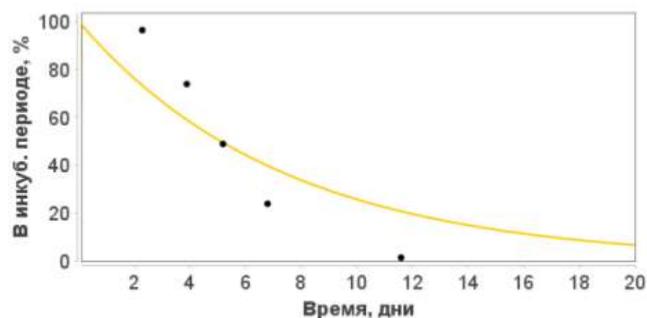
Параметры $k_1, \dots, k_n, d_1, \dots, d_n$ идентифицируются, исходя из известных процентилей моделируемого процесса, отдельно от всей остальной модели. Как правило, достаточно 1-2 слагаемых. Уравнение

$$0.14 * E(t - 2) + 0.42 * E(t - 3) \quad (6.3.4)$$

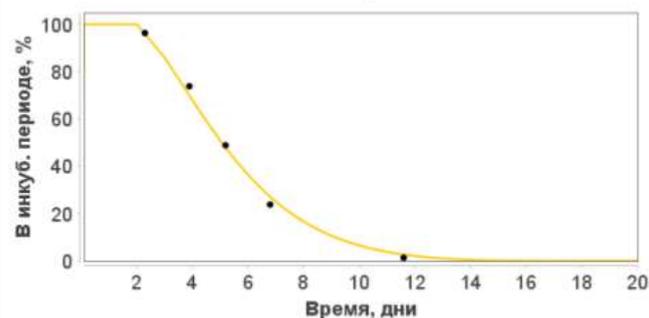
дает процентиля длительности инкубационного периода, совпадающие с данными Lauer et al. 2020 (Таблица 6.3.1, Рисунок 6.3.4).

Таблица 6.3.1 – Время инкубационного периода SARS-CoV-2

Данные или результаты моделирования	Процентиль, дней				
	2.5%	25%	50%	75%	97.5%
Lauer et al., 2020	2.2	3.8	5.1	6.7	11.5
Модель А	0.18	2.1	5.1	10.2	27.1
Модель В	2.2	3.8	5.1	6.7	11.5



Модель А: $dE/dt = -\ln(2)/5.1 * E(t)$



Модель В: $dE/dt = -0.14 * E(t-2) - 0.04 * E(t-3)$

Рисунок 6.3.4 – Сравнение распределения инкубационного периода (Lauer S.A., et al, 2020), смоделированного с помощью закона действующих масс (модель А) и уравнения с запаздыванием (модель В).

Стоит отметить, что параметры инкубационного процесса определяются независимо от подгонки параметров всей модели к наблюдаемым эпидемиологическим данным. Это позволяет с одной стороны более точно описать динамику инфицирования, а с другой – уменьшить количество параметров, идентифицируемых в ходе решения обратной задачи, что делает оставшиеся параметры более идентифицируемыми.

2. Конкурирующие процессы выздоровления и ухудшения симптомов. Другая проблема возникает, когда из некоторого состояния (например, появление легких симптомов) есть два возможных перехода: выздоровление либо ухудшение симптомов. При этом помимо корректных значений длительности обоих процессах необходимо учитывать долю пациентов в обоих исходах. Очевидно что при использовании закона действующих масс процесс с большей константой скорости реакции не только будет происходить быстрее, но и задействует большее количество пациентов. Например, не удастся смоделировать ситуацию, когда медианное время выздоровления пациентов 14 дней, а медианное время до ухудшения симптомов – 5 дней, при этом выздоравливает без ухудшения симптомов около 80% больных, что соответствует данным ВОЗ о COVID-19 на момент выполнения работы (6 марта 2020 г., WHO Situation Report 6 March) и данным по длительностям госпитализации (Boëlle et al., 2020).

В качестве решения Киселевым И.Н. было предложено разбить группу симптомных на три подгруппы: 1) “Новые симптомные”, 2) “Симптомные выздоравливающие” и “Симптомные с ухудшением”. Пациенты из первой группы мгновенно распределяются между двумя остальными в соответствии с заданным процентным соотношением выздоровевших и пациентов с тяжелыми симптомами. Процессы выздоровления и ухудшения могут моделироваться как с помощью закона действующих масс, так и с применением запаздывающего аргумента, как описано в предыдущем пункте. Таким образом, возможно добиться как правильных пропорций, так и длительности конкурирующих процессов (рисунок 6.3.5).

Сочетая описанные выше два подхода в третьей итерации нами была создана новая модель распространения коронавируса (рисунок 6.3.6). Большая часть переходных процессов была идентифицирована по имеющимся статистическим данным (Boelle et al., 2020) отдельно от остальной модели:

- инкубационный период;
- выздоровление при легких симптомах или их отсутствии;
- появление серьезных симптомов, требующих госпитализации;
- выздоровление при серьезных симптомах;
- появление критических симптомов, требующих реанимации;
- выздоровление при критических симптомах.

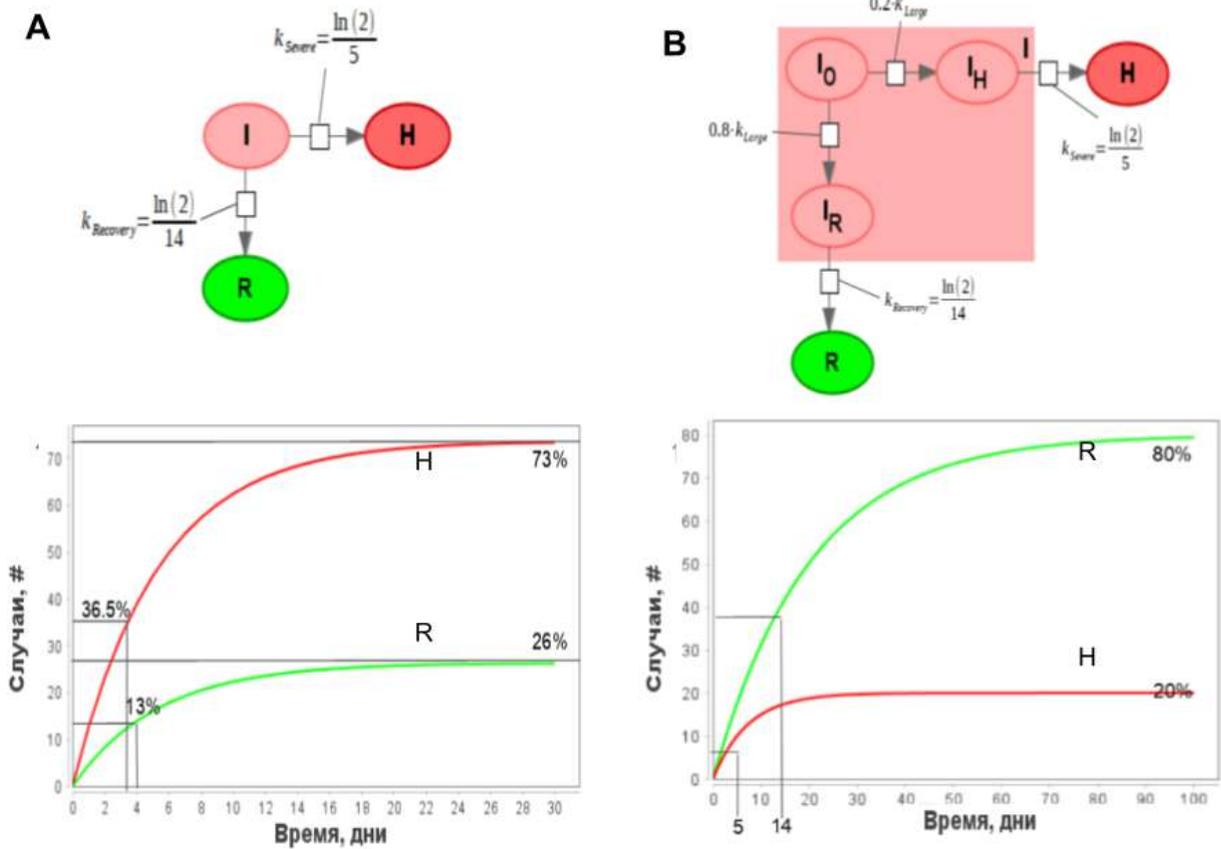


Рисунок 6.3.5 – Модель выздоровления/ухудшения симптомов с использованием закона действующих масс: стандартная (А) и дополненная мгновенным разделением на подгруппы (В)

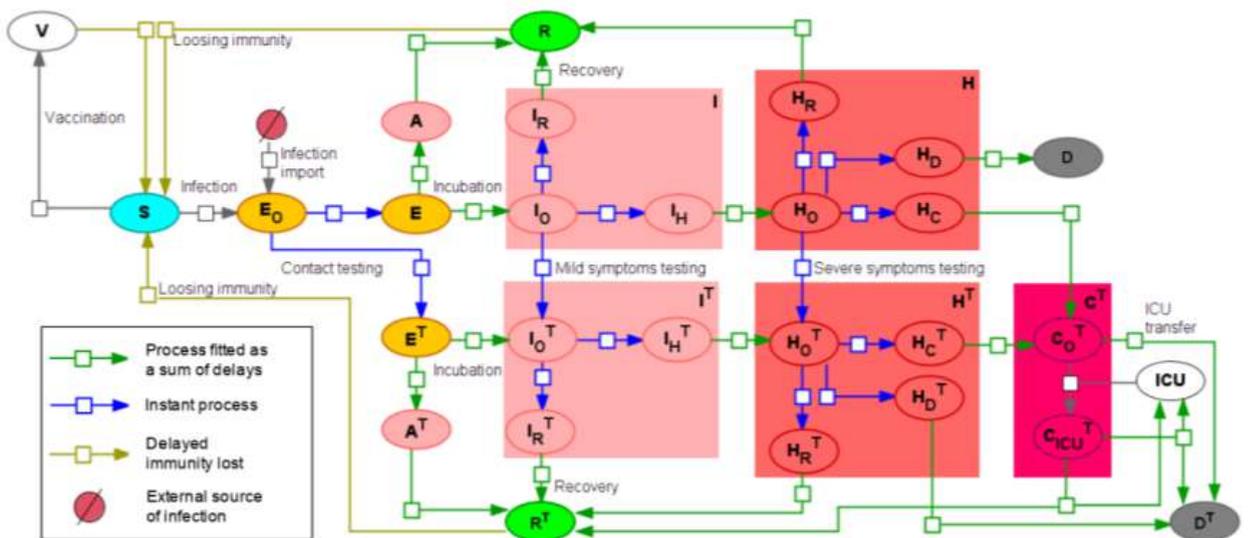


Рисунок 6.3.6 – Третья версия модели эпидемиологии COVID-19 с параметром с запаздыванием (Kiselev et al., 2021).

Также в модель были внесены следующие изменения и дополнения:

- добавлена потеря приобретенного иммунитета со временем. В модели считается, что переболевшие либо вакцинированные теряют приобретенный иммунитет спустя фиксированное время – 180 дней.
- количество вакцинированных в день берется из известных данных с сайта ourworldindata.org. В модели предполагается, что вакцинируются только выздоровевшие пациенты либо вакцинированные после того, как у них исчез иммунитет.
- новые волны пандемии ассоциированы с появлением новых штаммов коронавируса, а именно: европейского, альфа, дельта, омикрон. При появлении нового штамма в модели скачкообразно меняются характеристики вируса, такие как заразность, доля тяжелых больных, летальность.

Данная версия модели адаптирована к данным по Германии и Франции (приложение 6.3). Адаптация проводилась с целью минимизировать количество параметров, значения которых отличаются между данными странами.

Основным преимуществом данной модели является то, что большая часть ее параметров (не считая констант идентифицированных процессов) имеют явный эпидемиологический смысл: доля бессимптомных, доля тяжелобольных, процент смертности в зависимости от тяжести симптомов, вместо абстрактных констант скоростей реакций переходных процессов. Данная модель подробно описана в статье Kiselev et al., 2021, а исходный код модели и Jupyter ноутбуки доступны в репозитории по адресу:

<https://gitlab.sirius-web.org/covid-19/dde-epidemiology-model>

6.3.2 Агентная модель

На основе существовавшей ранее агентной модели распространения коронавируса REINA (Tuomisto et al., 2020) на языке Java была реализована агентная модель пандемии для ПК BioUML. Структура модели представлена на рисунке 6.3.7, а ее модульная диаграмма в ПК BioUML, адаптированная к Новосибирской области показана на рисунке 6.3.8.

Для уменьшения вычислительной сложности, агентами в модели являются только зараженный коронавирусом индивиды. Новый агент добавляется в модель в момент времени, когда происходит заражение здорового человека. Далее для агента рассчитывается его траектория болезни в зависимости от его возраста и характеристик вируса. Модель учитывает симптомы разной степени тяжести и бессимптомных носителей. Шаг времени – день. В течение одного дня каждый агент встречается с определенным количеством людей и с некоторой вероятностью заражает их, создавая новых агентов. Дополнительным агентом в модели является система здравоохранения, которая осуществляет тестирование на коронавирус в

популяции, госпитализацию больных с серьезными симптомами и перемещение в реанимацию критически больных.

По сравнению с исходной моделью в нее были внесены следующие изменения:

- добавлена длительность эффективности иммунитета;
- добавлен процесс вакцинации, который также осуществляется системой здравоохранения. При этом число вакцинированных людей каждый день задается, исходя из известных данных;
- добавлена возможность появления новых штаммов вируса. У каждого агента имеется метка того, каким штаммом он заражен. При заражении других людей, они помечаются тем же штаммом. У штаммов отличаются характеристики такие как заразность (вероятность заразить восприимчивого человека при встрече), вероятность появления тяжелых симптомов и летальность.

Модель была адаптирована к распространению нового коронавируса в Новосибирской области в 2020 году. Для учета количества людей, нарушающих противокоронавирусные ограничения, использовался индекс самоизоляции Яндекса. Результаты численного моделирования приведены в приложении 6.4.

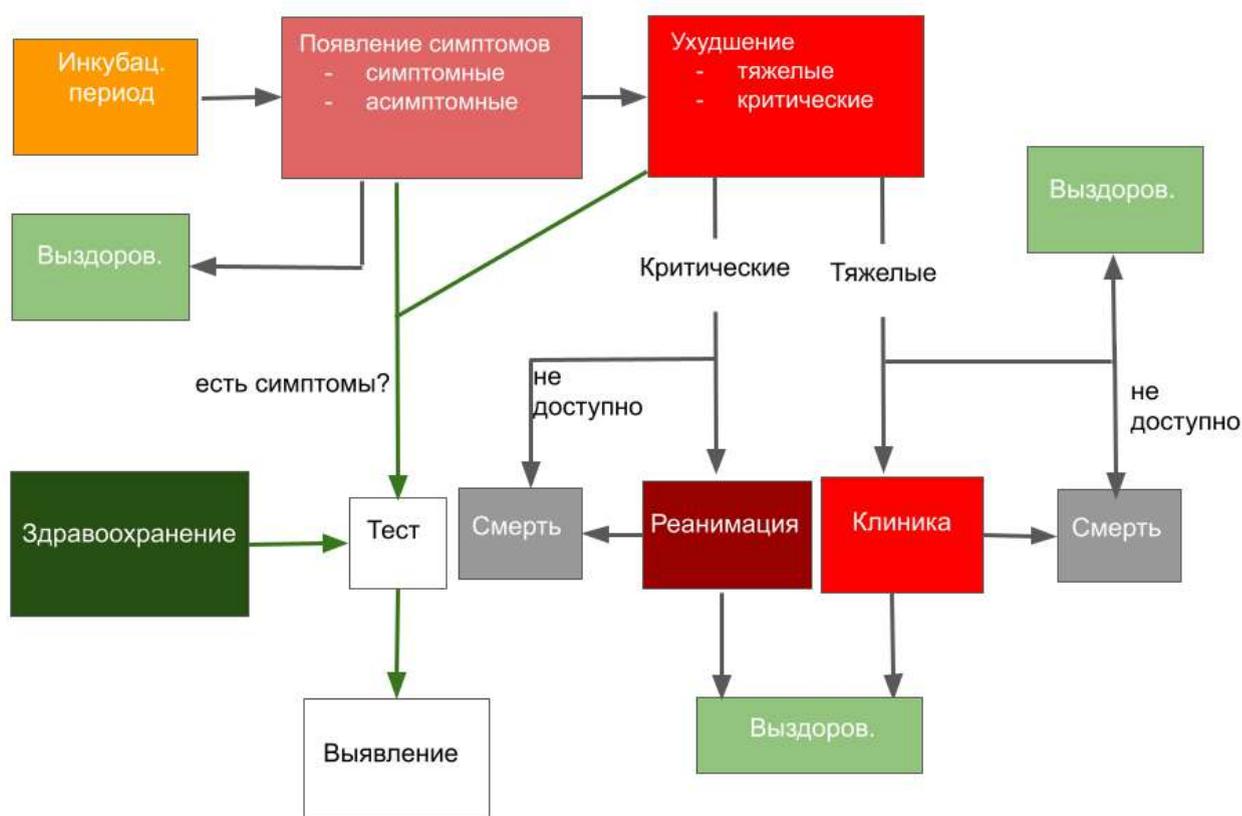


Рисунок 6.3.7 – Схема агентной модели модели распространения коронавируса

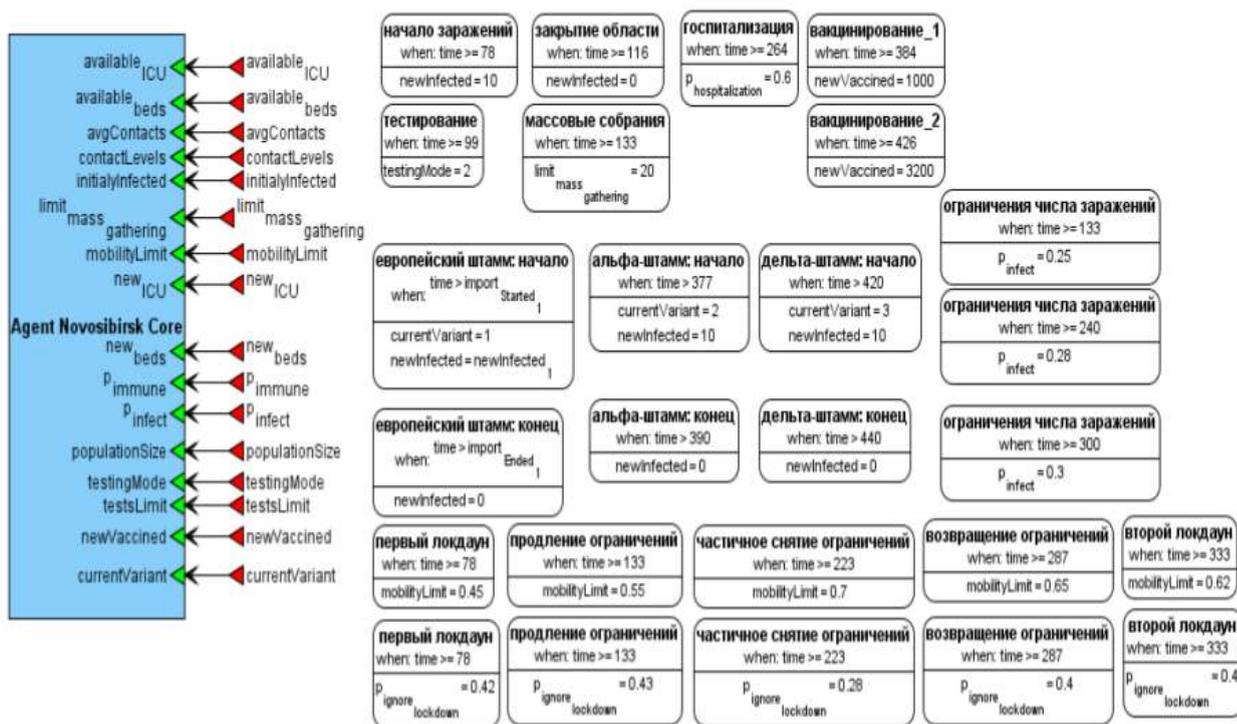


Рисунок 6.3.8 – Агентная модели модель распространения коронавируса, адаптированная к Новосибирской области.

6.3.3 Мультимасштабная эпидемиологическая модель

Мультимасштабная модель распространения коронавируса (рисунок 6.3.9) и включает в себя следующие модули:

1. Модуль эпидемиологической модели (Template_model на рисунке 6.3.9) – может содержать любую из моделей, описывающих распространение нового коронавируса, включая SEIR-модель, модель Вестерхоффа-Колодкина, одну из двух итераций расширения этой модели, модель с параметром с задержкой или агентную модель.
2. Модуль, описывающий противоэпидемические государственные меры (NPI) – рассчитывает следующие величины:
 - $Limit_{Asymptom}$ – ограничение передвижения бессимптомных носителей;
 - $Limit_{Exposed}$ – ограничение передвижения зараженных в инкубационном периоде;
 - $Limit_{Mild}$ – ограничение передвижения больных с легкими симптомами;
 - $Limit_{Severe}$ – ограничение передвижения больных с серьезными симптомами;
 - $Stringency_{Delayed}$ – общий уровень противоэпидемиологических мер, которой может быть рассчитан, например на основе Stringency Index или индекса самоизоляции Яндекса.
3. Модуль, описывающий свойства популяции (Population):
 - $Initial_{Size}$ – начальный размер популяции в моделируемом регионе;

- N – количество контактов в популяции в день;
- $Strigency_{Comply}$ – уровень соблюдения противоэпидемиологических мер;
- N – количество контактов в популяции в день;
- $Susceptible_{Fraction}$ – долю чувствительных к заражению в популяции.

4. Модуль тестирования (Testing):

- $Test_{Delay}$ – задержку между сдачей теста и получением результатов;
- $Test_{Exposed}$ – уровень контактного тестирования;
- $Test_{Mild}$ – уровень тестирования больных с легкими симптомами;
- $Test_{Severe}$ – уровень тестирования больных с серьезными симптомами.

5. Модуль вакцинирования (Vaccination):

- $Vaccinated_{Day}$ – количество проведенных тестов в день.

6. Модуль, описывающий свойства инфекционного заболевания (Virus):

- $Critical_{Fraction}$ – долю зараженных с критическими симптомами;
- $Severe_{Fraction}$ – долю зараженных с серьезными симптомами;
- $Severe_{CFR}$ – уровень летальности больных с серьезными симптомами в госпитале;
- ICU_{CFR} – уровень летальности в реанимации;
- I_{Scale} – степень заразности текущего штамма вируса по сравнению с изначальным штаммом;
- $I_{Asymptom}$ – вероятность заразиться при контакте с бессимптомным носителем;
- $I_{Exposed}$ – вероятность заразиться при контакте с зараженным в инкубационном периоде;
- I_{Mild} – вероятность заразиться при контакте с больным с легкими симптомами;
- I_{Severe} – вероятность заразиться при контакте с больным с серьезными симптомами.

В данной модели любой из модулей может быть модифицирован или заменен на другой модуль, описывающий соответствующую подсистему. Таким образом, заменяя модули, пользователь может собрать модель, описывающий конкретное инфекционное заболевание в нужном регионе при заданном уровне противоэпидемических мер.

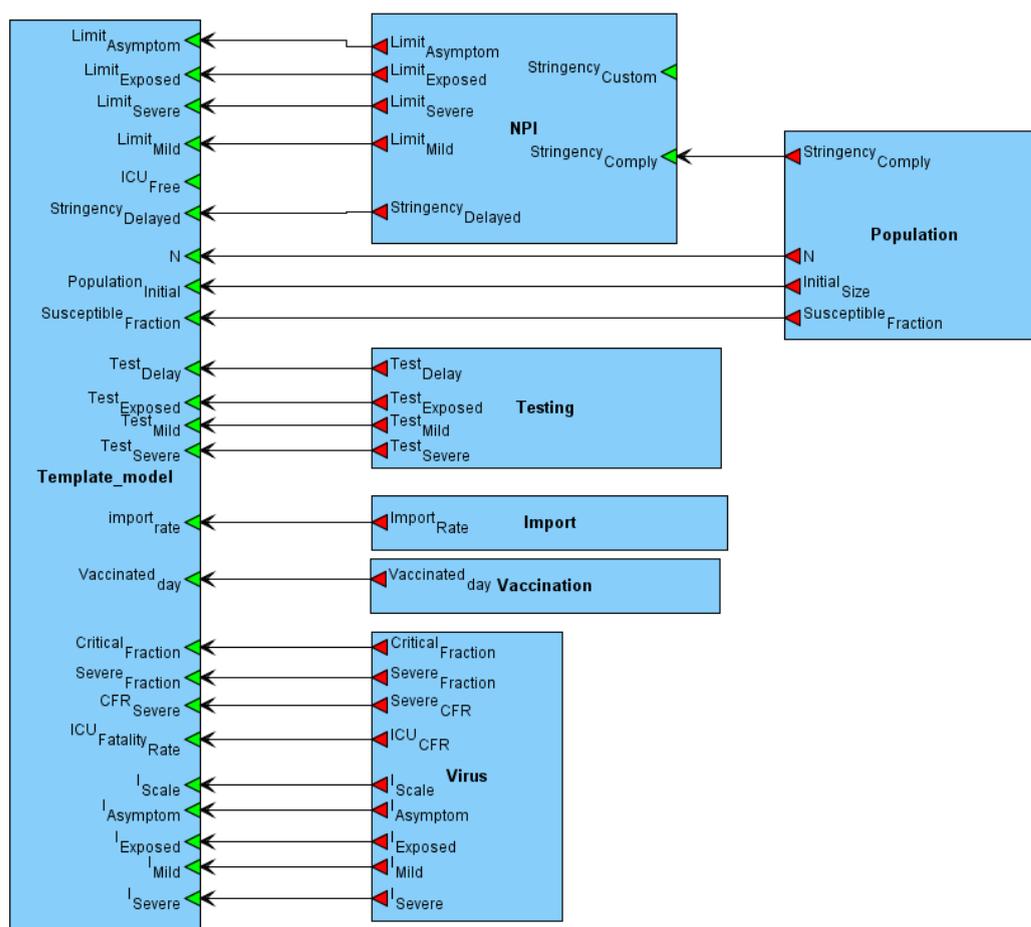


Рисунок 6.3.9 – Модульная мультимасштабная модель распространения коронавируса (см. описание модулей в тексте выше)

6.3.4 Взаимодействие вируса и организма хозяина

Также, используя ПК BioUML, были воспроизведены модели, описывающие взаимодействие вируса, организма хозяина и его иммунной системы:

- модульные модели иммунного ответа на заражение вирусом гриппа А и *M. tuberculosis* (Lee J. et al., 2009; Marino and Kirschner, 2004);
- динамика иммунного ответа на заражение SARS-CoV-2 (Grebennikov et al., 2022).

Также была построена новая модель распространения вируса SARS-CoV-2 по органам и тканям человека (Afonyushkin et al., 2022).

6.4 Модели регуляции артериального давления у человека

Принципы модульного подхода, реализованные в рамках ПК BioUML и описанные в разделе 3.4, хорошо согласуются с иерархической структурой физиологических систем и позволяют создавать модели различных элементов и различных уровней этой структуры и объединять всё в рамках одной модели.

Важно заметить, что многие принципы и законы, действующие в физиологических системах, не имеют на сегодняшний день полного теоретического обоснования и представляют собой аппроксимацию огромного количества экспериментальных данных, выраженную уравнениями или графиками зависимости одних параметров от других. Поэтому для визуального моделирования таких физиологических систем в ПК BioUML используется специальный тип диаграмм, описанный в разделе 3.3.5.

Отметим также, что наряду со специфическими физиологическими принципами при моделировании процессов, протекающих в живых организмах, имеют место и фундаментальные физические законы, такие как закон Гука, гидродинамические законы, закон фильтрации через пористую среду и т. д. (Бибердорф и др., 2015). Подобные законы использованы в пяти моделях, кратко рассмотренных в данной главе:

1. Модель Гайтона (Guyton et al., 1972), которая была создана в модульном виде, используя ПК BioUML.
2. Резервуарная модель, предложенная профессором Ю.В. Солодянниковым в 1994 году (Солодянников, 1994), в которой большой круг кровообращения человека представлен системой из трех резервуаров с буфером капилляров. Данная система может находиться в одном из двух состояний – систола или диастола. Каждое состояние описывается отдельной системой дифференциальных уравнений, полученной на основе фундаментальных физиологических и гидродинамических законов.
3. Модель долговременной почечной регуляции, предложенная Караасланом и соавторами (Karaaslan et al., 2005), использующая идеи своих предшественников – моделей Гайтона (Guyton et al., 1972), Уттамсинга (Uttamsingh et al., 1985) и Колмана-Холла (Coleman and Hall, 1992) – детально описывает механизмы почечной регуляции кровотока.
4. Гидродинамическая модель (ГДМ) артериальной системы, полученная из уравнений Навье-Стокса посредством интегрирования их по поперечному сечению сосуда (Lamroni, 2004). Данная модель описывает кровоток в артериальном дереве, состоящем из 55 наиболее крупных артерий человека, и позволяет получить значения давления, потока крови и других параметров в каждой точке каждой артерии. Для поиска численных решений ГДМ предложен оригинальный алгоритм упаковки артериального дерева и использован метод ортогональной прогонки (Бибердорф и др., 2015).
5. Комплексная модель регуляции артериального давления человека (Kutumova et al., 2021), которая является объединением и развитием моделей 1, 2 и 3.

6.4.1 Модель Гайтона

Модель, предложенная профессором Гайтоном (Guyton et al., 1972), представляет собой первое глобальное описание процессов в человеческом организме, связанных с кровообращением, и демонстрирует долговременные эффекты регуляции артериального давления у человека со значительной ролью почечной регуляции. При этом модель оперирует усредненными по времени величинами. Краткосрочные эффекты, такие как биение сердца, в модели не учитываются.

Модель была изначально написана на языке программирования Fortran, позже была переписана на язык C и длительное время дорабатывалась и расширялась ее авторами. Исходная модель снабжена структурной схемой, представляющей ее в виде блочной структуры, состоящей из более чем 350 элементарных блоков. Блоки соответствуют числовым константам, переменным и операциям, таким как сложение, умножение, интегрирование (рисунок 1.2.3.б). Как уже рассматривалось в главе 1, такая графическая нотация плохо подходит для представления биологических систем.

Поэтому модель Гайтона была реализована в ПК BioUML в виде модульной диаграммы (рисунок 6.4.1; Киселев, 2016). За основу построения модульной модели был взят вариант модели Гайтона, который Kofranek и Rusz реконструировали в 2010 году в среде MATLAB SIMULINK (Kofranek and Rusz, 2010). При этом набор блоков алгебраических операций (суммирование, умножение, интегрирование и т.д.), из которых состоит модель в SIMULINK, был преобразован в систему ОДУ.

Реализованная модель (Киселев, 2016) состоит из 18 модулей, суммарно включающих 229 параметров, 133 операции присваивания, 39 обыкновенных дифференциальных уравнений.

Модуль “Гемодинамика”, описывающий кровоток по резервуарам большого и малого кругов кровообращения, реализован в виде модульной модели (рисунок 6.4.2), включающей 5 модулей (в порядке следования крови): “Левый желудочек”, “Артерии большого круга”, “Вены большого круга”, “Правый желудочек”, “Артерии и вены малого круга”. В качестве примера на рисунке 6.4.3 приведен модуль “Артерии большого круга”. Обозначения переменных модели сохранено.

Данный пример наглядно демонстрирует, насколько графическая нотация для создания модульных моделей физиологических систем в ПК BioUML лучше подходит для моделирования сложных физиологических систем.

6.4.2 Резервуарная модель

В данном разделе кратко описывается реализация в ПК BioUML резервуарной модели, предложенной профессором Ю. В. Солодяниковым в 1994 году (Солодяников, 1994). В

дальнейшем эта модель была расширена (Прошин и Солодянников, 2006), чтобы учесть большой и малый круг кровообращения.

Исходная модель описывает процессы, протекающие в кровеносной системе человека, и представляет большой круг кровообращения, содержащий три резервуара – резервуар артерий, вен и сердца и систему капилляров, которая играет роль некоторого «буфера» между артериями и венами (рисунок 6.4.4). Резервуар сердца представляет модель левого желудочка сердца человека и отделяется от венозного резервуара митральным клапаном М, между артериальной системой и сердцем находится аортальный клапан А. Все величины, характеризующие артериальный и венозный кровоток данной модели, усреднены по объему соответствующих резервуаров и являются только функциями времени.

Система кровообращения модели Солодянникова 1994 года может находиться в двух качественно различных состояниях (фазах):

- 1) сокращение желудочка и изгнание крови из сердца (систола);
- 2) расслабление желудочка сердца и заполнение его кровью (диастола).

Вместе с тем наличие митрального и аортального клапанов М и А означает наличие в системе пороговых нелинейностей: при переходе от одной фазы к другой некоторые параметры изменяются скачком. Математической идеализацией такого объекта является динамическая система класса дифференциальных уравнений, когда в различных фазах объект описывается различными системами уравнений.

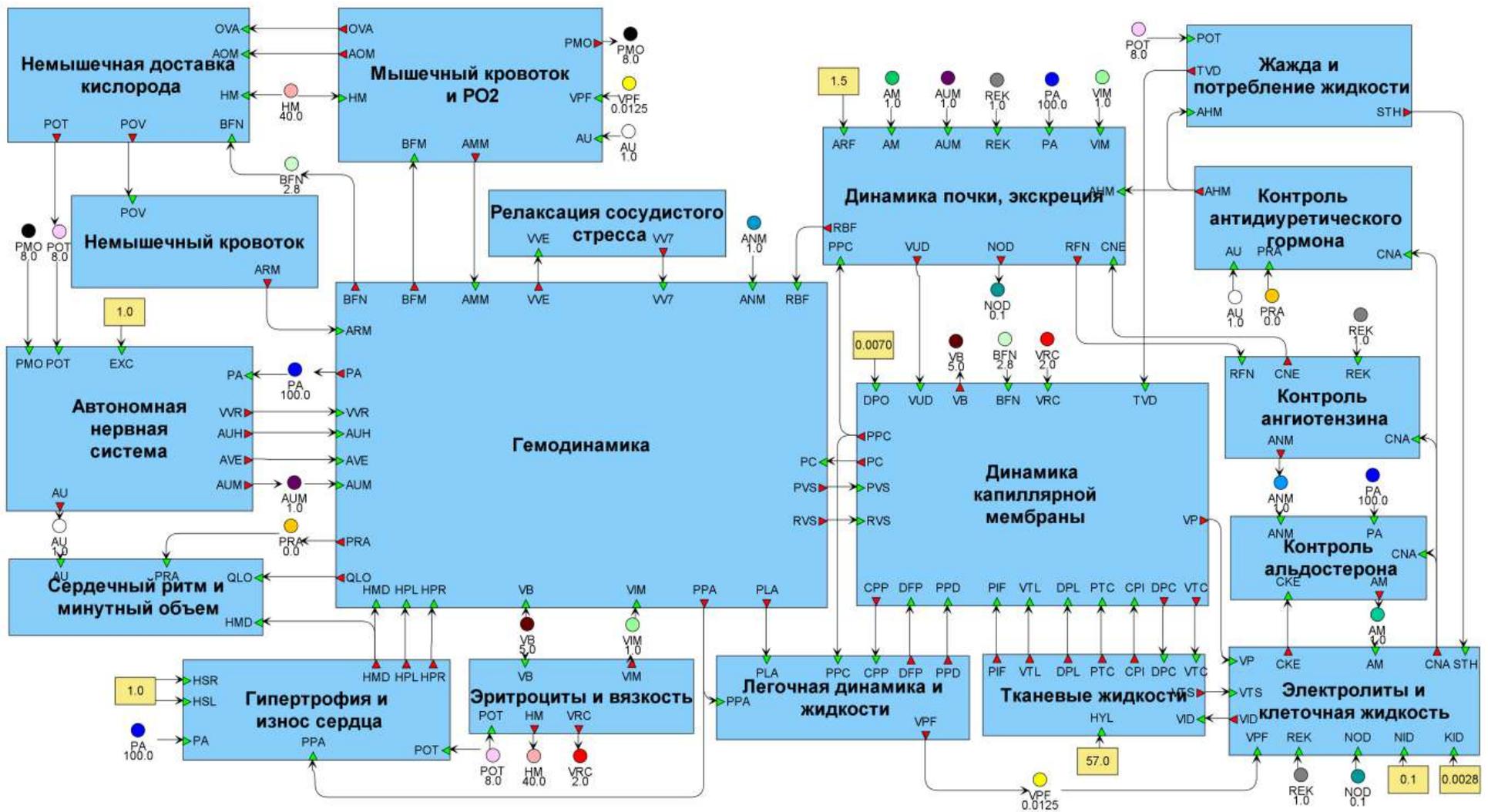


Рисунок 6.4.1 – Общая модель кровообращения (Guyton et al., 1972), реализованная в BioUML в виде модульной диаграммы (Киселев, 2016)

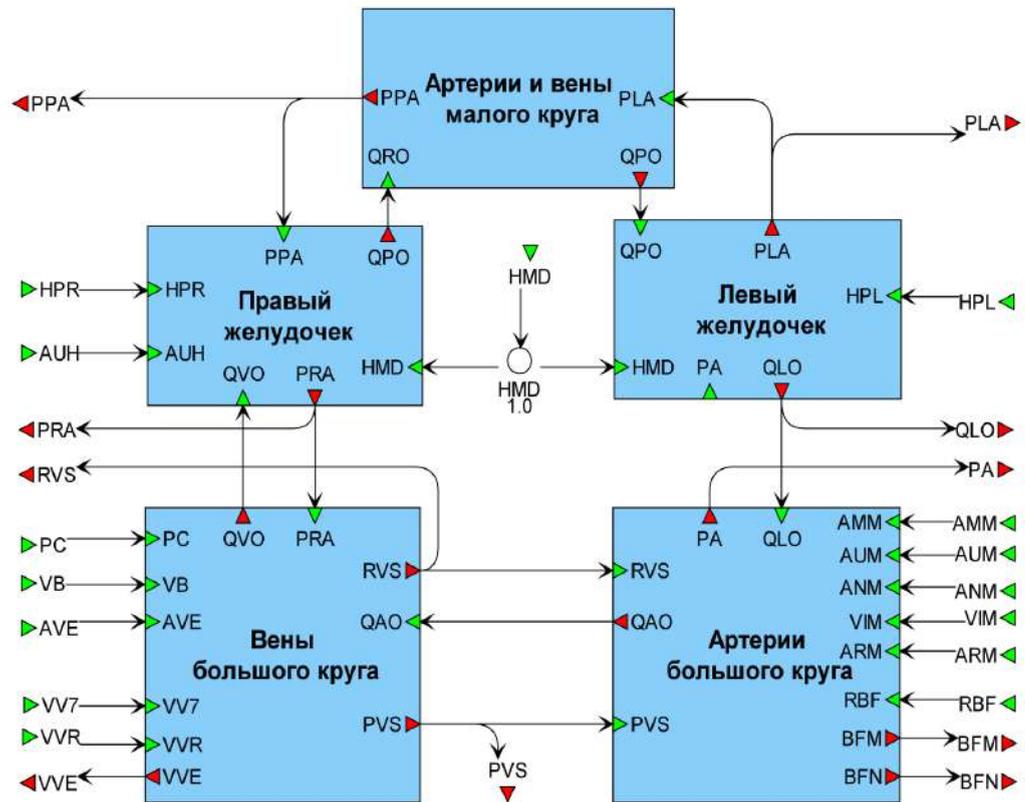


Рисунок 6.4.2 – Модуль “Гемодинамика” модульной модели общей регуляции, реализованный в BioUML (Киселев, 2016)

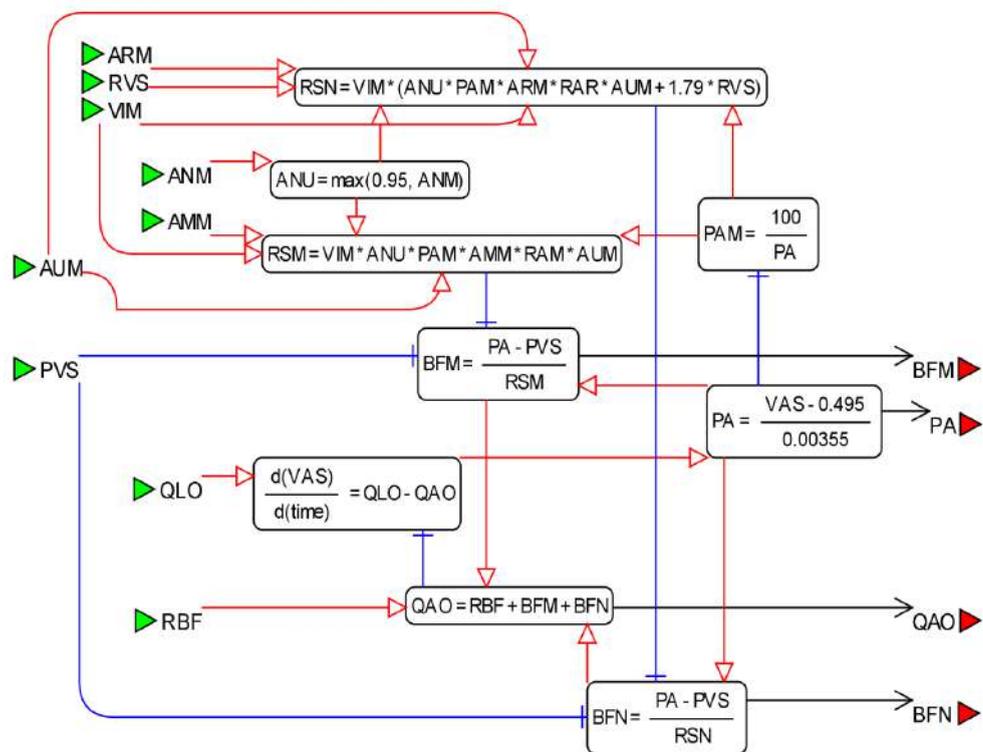


Рисунок 6.4.3 – Модуль “Артерии большого круга” в виде диаграммы “Модель физиологических процессов” (Киселев, 2016)



Рисунок 6.4.4 – Резервуары модели Солодяникова (Солодяников, 1994; Бибердорф и др., 2015).
М – митральный клапан, А – аортальный клапан

Для каждого резервуара модели Солодяникова 1994 года имеется набор уравнений, определяющих его состояние. Кровообращение между резервуарами артерий, вен и сердца регулируется следующим набором законов:

$$\frac{dV_i}{dt} = Q_{ki} - Q_{ij}, \quad P_i = G_i(V_i - W_i), \quad Q_{ij} = \frac{P_i - P_j}{R_{ij}}, \quad (6.4.1)$$

где V_i – объем крови в i -м резервуаре; Q_{ij} – поток крови из i -го в j -й резервуар; P_i – давление крови в i -м резервуаре; R_{ij} – сопротивляемость j -го резервуара току крови из i -го; G_i – объемная упругость i -го резервуара; W_i – ненапряженный объем i -го резервуара. Далее вместо индексов i, j, k мы будем использовать буквы «А», «В», «Н» обозначающие резервуары артерий, вен и сердца соответственно.

Объем крови венозного резервуара рассчитывается исходя из предположения постоянства общего объема крови в системе (A_{30} – общий объем крови в системе) :

$$V_V = A_{30} - V_H - V_A. \quad (6.4.2)$$

Уравнения 6.4.1 представляют фундаментальные физические законы. Так, первое из них следует из закона сохранения масс и определяет динамику объема несжимаемого вещества (в данном случае объема крови V_i) через входящий и исходящий потоки Q_{ki} и Q_{ij} ; второе есть некоторый аналог закона Гука и выражает линейную зависимость давления P_i от объемной деформации ($V_i - W_i$); третье представляет некоторое обобщение гидродинамических законов, таких как закон Пуазейля или закон фильтрации жидкости через пористую среду,

утверждающих, что поток жидкости через среду Q_{ij} пропорционален градиенту давления $P_i - P_j$ в среде и обратно пропорционален сопротивляемости среды R_{ij} .

Согласование всех параметров этих уравнений 6.4.1 реализовано с помощью эмпирических данных и фундаментальных физиологических законов – контуров управления модели Солодянникова. В основу модели положены четыре контура.

1. Контур нейрогуморального управления обеспечивает коррекцию свойств резервуаров системы и поддержание частоты сердечных сокращений на необходимом уровне (рисунок 6.4.4). Длительность сердечного цикла T определяется в данном контуре через величину нейрогуморального фактора γ , представляющего воздействие нервной и гуморальной систем организма на кровеносную. Уравнения, задающие значения γ и T , имеют вид

$$\frac{d\gamma}{dt} = A_{12} \left(A_{13} \frac{Q_{AV}}{V_A} - A_{14}(P_A - A_{28}) - \gamma \right), \quad T = \frac{1}{\gamma}. \quad (6.4.3)$$

Здесь Q_{AV} — мгновенный поток через капилляры; V_A — объём артерий; P_A — артериальное давление; $A_{12}, A_{13}, A_{14}, A_{28}$ — параметры, характеризующие конкретный организм (Солодянников, 1994). Важно отметить, что уравнение, определяющее динамику фактора γ , может быть представлено в виде

$$\frac{d\gamma}{dt} = \frac{\gamma_0(t) - \gamma(t)}{T_\gamma}, \quad (6.4.4)$$

где $\gamma_0(t) = A_{13} \frac{Q_{AV}}{V_A} - A_{14}(P_A - A_{28})$ — характеризует стационарное значение γ , а $T_\gamma = \frac{1}{A_{12}}$ — скорость выхода на стационарный режим.

Таким образом, уравнения (6.4.3) отражают свойство динамического равновесия рассматриваемой биологической системы. При завершении сердечного цикла осуществляется переход от системы уравнений, описывающей систолическую фазу, к системе, описывающей фазу диастолы. При этом происходит скачкообразное изменение систолического давления в желудочке P_S , производных объёма желудочка и объёма артерий V_A по времени.

2. Данный скачок контролируется контуром управления длительностью систолы, который обеспечивает постоянное соотношение между временем заполнения и временем сокращения желудочка путём изменения систолического давления P_S .

3. Момент перехода от систолы к диастоле определяется в рамках контура управления величиной ударного выброса в соответствии с законом Франка-Старлинга («чем больше мышца сердца растягивается в диастолу, тем сильнее ее сокращение в период систолы»), выражающего объём крови, выбрасываемый желудочком за одно сокращение v (ударный объём) от объёма желудочка V_{ED} в момент начала сжатия сердца (конечно-диастолического объёма), k_1 и k_2 — инотропные коэффициенты:

$$\nu = k_1 * V_{ED} - k_2. \quad (6.4.5)$$

4. Контур управления тканевым метаболизмом необходим для поддержания требуемого значения кровотока через капилляры Q_{AV} и вены Q_{VH} . Управляющий сигнал в нём — кислородный долг DO_2

$$\begin{cases} \frac{dDO_2}{dt} = A_2(A_{15} - gO_2), \\ gO_2 = Q_{AV}(A_{22} - BO_2), \\ \frac{dBO_2}{dt} = (A_1gO_2 - A_{15}). \end{cases} \quad (6.4.6)$$

Здесь gO_2 – мгновенная доставка кислорода тканям; BO_2 – содержание кислорода в венозной крови; A_{15} – мгновенная потребность организма в кислороде; A_{22} – содержание кислорода в артериальной крови; A_1 – чувствительность вен к DO_2 ; A_2 – скорость утилизации O_2 .

Данная модель была реализована в КП BioUML в виде модульной диаграммы (рисунок 6.4.5), состоящей из 6 модулей: “Артериальная система”, “Левый желудочек”, “Венозная система”, “Капилляры”, “Нейрогуморальный контроль”, “Тканевый метаболизм” и 22 связи между ними (Киселев, 2016).

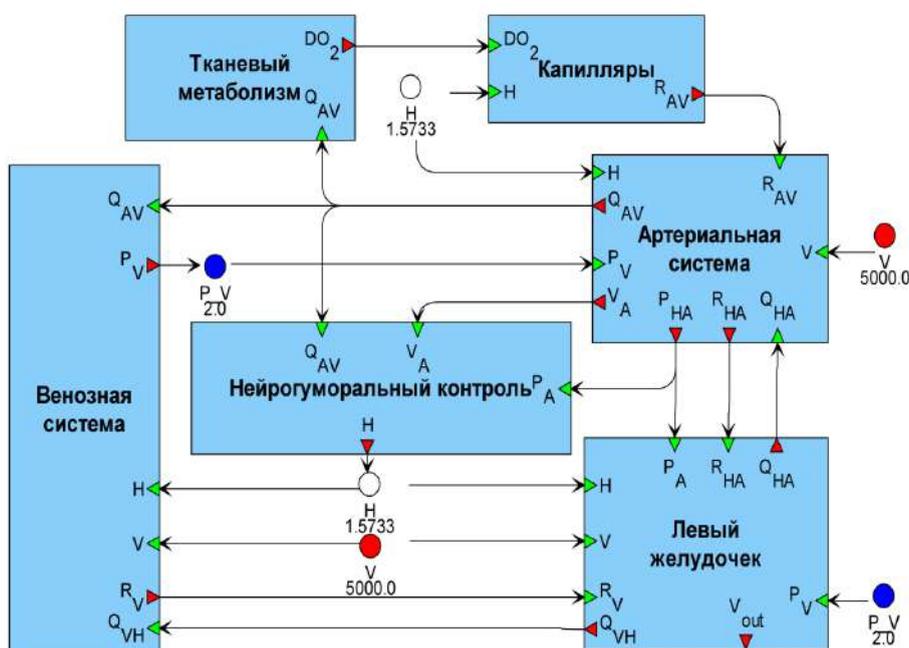


Рисунок 6.4.5 – Модель краткосрочной регуляции сердечно-сосудистой системы человека с пульсирующим сердцем (Солодяников, 1994), реализованная в виде блочной диаграммы в BioUML (Киселев, 2016)

Модель включает в себя 58 переменных, 7 дифференциальных уравнений, 14 операций присваивания и 2 исполняемых события для мгновенного перехода между состояниями.

Алгоритм для численного решения, встроенный в BioUML, на каждом шаге численных расчетов проверяет условие перехода и рассчитывает точное время перехода модели из одного состояния в другое.

При реализации модели в BioUML были сделаны модификации дающие возможность варьирования общего объема крови в системе. Изменение объема крови в резервуаре определяется уравнением:

$$\frac{dV_i}{dt} = Q_{ki} - Q_{ij} + c_i \frac{dV}{dt}, \quad (6.4.7)$$

где c_i – коэффициенты, задающие распределение изменения объема крови между резервуарами. Данные коэффициентов были взяты из Thomas et al., 2008. В разрешенной относительно производных форме этот закон задается в следующем виде:

$$\begin{cases} \frac{d\tilde{V}_i}{dt} = Q_{ki} - Q_{ij}, \\ V_i = \tilde{V}_i + c_i V. \end{cases} \quad (6.4.8)$$

Уравнение для объема крови венозного резервуара (6.4.2) заменено дифференциальным уравнением согласно (6.4.1). Также была изменена формула расчета периферической проводимости C_{AV} , а именно взята формула из более поздней модели того же автора (<http://www.samara-dialog.ru/help/rus/help.htm>):

$$C_{AV} = C_{AV_0} + DO_2 - C_H H, \quad \text{вместо} \quad C_{AV} = C_{AV_0} + DO_2 + C_H H. \quad (6.4.9)$$

Новая формула отражает вазоконстрикторный эффект при увеличении нейрогуморальных воздействий. Данное изменение также позволило более точно смоделировать физическую нагрузку (Киселев, 2016).

Модель позволяет имитировать состояние физической нагрузки, различные патологии, связанные с нарушением сердечного цикла, такие как аритмия, кардиогенный шок и т.д. Далее в тексте модель упоминается как “модель сердечных сокращений”.

6.4.3 Модель долговременной почечной регуляции

Почка является важным компонентом регуляции артериального давления, регулирующим объем и состав жидкости в организме и, соответственно, артериальное давление, поэтому без нее комплексная модель была бы неполной. Одной из наиболее проработанных моделей, подробно описывающих работу почки, является модель, разработанная Ф. Караасланом с соавторами (Karaaslan et al., 2005). Данная модель была создана на основе предыдущих работ:

Guyton et al., 1972; Uttamsingh et al., 1985; Coleman and Hall, 1992 – здесь мы как раз видим реализацию принципа эволюционного развития модели.

Наиболее важной частью модели является блок, описывающий регуляцию объема жидкости в организме почкой. Данный блок значительно более детализирован, чем в моделях-предшественниках. Важной частью модели является регуляция уровня различных гормонов и веществ в организме – вазопрессина, ангиотензина, альдостерона, натрийуретического пептида, пищевой соли и др. в организме человека.

Реализованная в ПК BioUML версия модели состоит из 6 модулей: “Сердце”, “Артерии”, “Почка”, “Нервная система”, “Гормональная система”, “Контроль объема жидкости” (рисунок 2), содержит 85 переменных, 61 операцию присваивания и 11 обыкновенных дифференциальных уравнений.

6.4.4 Гидродинамическая модель артериальной системы

Одномерная модель артериального древа человека, созданная на основе модели (Lamponi, 2004), описывает течение крови в артериальном сосудистом русле человека. Модель включает в себя 55 основных артерий, составляющих бинарное дерево. Каждая артерия описана набором параметров: длина, площадь поперечного сечения, объемная эластичность и два «дочерних» сосуда, на которые она разветвляется (кроме терминальных сосудов).

Математическая модель была получена из уравнений Навье-Стокса посредством интегрирования их по произвольному осевому сечению сосуда, т.е. в некотором смысле посредством усреднения этих уравнений по “поперечному” направлению. При этом были использованы следующие предположения:

1. Осевая симметрия сосуда. Все величины не зависят от угловой компоненты координаты φ . Как следствие, каждое осевое сечение S остается круговым в течение всего времени движения стенок сосуда. Радиус трубки R является функцией времени t и осевой координаты z .
2. Стенка сосуда является упругой и перемещается только вдоль радиального направления.
3. Ось цилиндра зафиксирована и не перемещается со временем.
4. Давление p постоянно на каждом осевом сечении.
5. Массовые силы отсутствуют.
6. Преобладает осевая компонента скорости u_z . По сравнению с ней компоненты скорости, ортогональные оси z , являются пренебрежимо малыми.

Для поиска приближённых решений краевой задачи была осуществлена дискретизация по времени. На каждом временном слое решение ищется методом ортогональной прогонки. Алгоритм решения уравнений модели был реализован на языке Java в виде модуля для пакета программ BioUML (Киселев и др., 2012; Бибердорф и др., 2015). Кроме того, в BioUML был

реализован пользовательский интерфейс для графического редактирования структуры артериального дерева и задания всех необходимых параметров.

6.4.5 Комплексная модель регуляции артериального давления

В результате применения агентного подхода нами была создана комплексная модель сердечно-сосудистой системы человека, которая затем была исследована. Модель была создана путем интеграции описанных выше математических моделей, которые были реализованы в ПК BioUML в виде модульных диаграмм, после чего необходимым образом доработаны. Кроме того, были определены и согласованы между собой входные и выходные переменные каждой подмодели.

Была выполнена калибровка модели, чтобы найти состояние равновесия в пределах нормальных диапазонов показателей жизнедеятельности для здорового взрослого человека. Мы проверили способность модели воспроизводить равновесные состояния с аномальными физиологическими значениями, связанными с различными сочетаниями сердечно-сосудистых заболеваний (таких как системная гипертензия, хроническая сердечная недостаточность, легочная гипертензия и др.). Для создания и проверки модели были использованы результаты более 200 научных исследований, охватывающих известные модели сердечно-сосудистой и почечной функций человека, платформы биомоделирования и клинические измерения физиологических величин в норме и при патологии.

Последняя версия этой модели (рисунок 6.4.1) подробно описана в работе Kutumova et al., 2021. На рисунке 6.4.6 приведен отзыв одного из рецензентов на эту работу: «Никто со времен Гайтона и его команды в Миссисипи не сделал такого качественного скачка в моделировании регуляции и контроля кровяного давления на уровне всего организма».

Q 5 Please provide your detailed review report to the editor and authors (including any comments on the Q4 Check List):

Reviewer 2 | 12 Aug 2021 | 20:57

#1

This brilliant paper develops a mathematical model that can be applied to human hypertension. It constructs a set of basic modules that can then be used to assemble a model for a given patient. They then develop a treatment simulation for a virtual population that can identify virtual groups with a similar reaction to particular classes of anti-hypertensive drugs. Next, they determine which features of the patient can contribute to the efficacy of the therapy. Their model takes into account the interconnection between the heart and the kidneys. No-one since Guyton and his team in Mississippi in the 1960-70s has made such a quantum leap in the field of modelling whole body cardiovascular regulation of blood pressure and its control.

The authors provide an adequate description of limitations to their model.

They also describe future directions, such as influences of aging and genetics.

There are no Comments to add related to the Q4 checklist.

QUALITY ASSESSMENT

Q 6	Originality	
Q 7	Rigor	
Q 8	Significance to the field	
Q 9	Interest to a general audience	
Q 10	Quality of the writing	
Q 11	Overall quality of the study	

Рисунок 6.4.6 – Рецензия на статью Kutumova et al., 2021

ГЛАВА 7. ЦИФРОВОЙ ДВОЙНИК ПАЦИЕНТА

"... истинная персонализация лекарственной терапии будет основываться на «виртуальном пациенте» – модели, основанной на данных индивидуального пациента, полученных при помощи молекулярных и сенсорных технологий."²

Hans Lehrach, 2016

Основная часть данной работы была выполнена по гранту РФФИ №16-01-00779 «Виртуальный пациент: разработка технологии, тестирование ее полезности на примере модульной математической модели сердечно-сосудистой системы человека и предсказание эффективности лечения артериальной гипертензии» под руководством автора диссертационной работы. Полученные результаты представлены в статьях (Киселев и др., 2019b, 2019c; Колпаков и др., 2019b, 2019c; Kutumova et al., 2021, 2022a).

7.1 Концепция цифрового двойника пациента

Концепция цифровых двойников была рассмотрена в разделе 1.2.5. В применении к цифровому двойнику пациента она выглядит следующим образом:

- физический продукт в реальном пространстве – реальный пациент;
- виртуальный продукт в виртуальном пространстве – виртуальный пациент, который представлен математической моделью физиологии и биохимии пациента;
- данные и информация, которые объединяют виртуальный и физический продукт – персонализация математической модели на основе данных электронной медицинской карты, генетических данных, опросников, носимых мобильных устройств и т.п.

Таким образом, "виртуальный пациент" – это часть цифрового двойника реального пациента, который представлен математической моделью физиологии и биохимии человека, причем параметры этой модели подобраны так, чтобы максимально соответствовать наблюдаемым клиническим данным для заданного реального пациента (Lehrach, 2016).

Основной целью построения цифрового двойника пациента является персонализация лекарственной терапии, которая в свою очередь, является частью персонализированной медицины. Цель персонализированной медицины состоит в том, чтобы найти подходящий лекарственный препарат для конкретного больного, и в некоторых случаях даже разработать схему лечения пациента в соответствии с его индивидуальными данными (приказ Министерства

² *"... true personalization of drug therapies will rely on "virtual patient" models based on a detailed characterization of the individual patient by molecular, imaging, and sensor techniques."*

здравоохранения РФ № 186). Необходимость этого обусловлена тем, что традиционные, создаваемые для лечения конкретного заболевания лекарственные препараты, оказываются неэффективными для 30–60% пациентов наряду с высокой частотой возникновения побочных эффектов (Кропачева и др., 2015; Spear et al., 2001).

В зависимости от заболевания, его механизмов и терапии виртуальные модели пациентов могут быть реализованы как на довольно высоком уровне абстракции, так и включая модели, представляющими релевантные клетки, типы клеток или органы, которые могут взаимодействовать не только друг с другом, но и с окружающей средой. Однако, пока невозможно построить единую математическую модель для всех заболеваний. Для каждой группы болезней необходима своя математическая модель для цифрового двойника пациента.

АГ является одним из наиболее важных социально-значимых заболеваний и главным фактором риска развития сердечно-сосудистых катастроф, таких как инфаркт миокарда, инсульт, ишемическая болезнь сердца, хроническая сердечная недостаточность. Поскольку АГ – это сложное мультифакторное заболевание, то для ее лечение используются разные классы лекарственных препаратов, и часто врач пробует несколько схем лечения, чтобы найти оптимальный вариант для конкретного пациента. Поэтому использование цифрового двойника пациента для оптимизации лекарственной терапии является крайне актуальной задачей.

В данной работе нами предложена технология построения цифрового двойника пациента и продемонстрирована ее работоспособность на примере оптимизации выбора лекарственной терапии для лечения АГ. Предложенная технология состоит из 4 этапов:

- 1) **Построение модульной математической модели биохимии и физиологии человека с достаточным уровнем детализации для заданной болезни.** На данный момент не реально построить "виртуального пациента" на все случаи жизни. Поэтому предложен следующий подход – создать набор основных блоков, а из них собирать модель под конкретного пациента и болезнь (как из блоков конструктора Лего). Созданная для этого математическая модель регуляции артериального давления у человека была описана в предыдущей главе.
- 2) **Моделирование фармакодинамики лекарственных препаратов** – для основных классов АГП были определены их точки воздействия на построенную модель регуляции артериального давления, после чего разрабатывались соответствующие модели фармакодинамики. Для валидации полученной модели были использованы данные клинических исследований, найденные в литературе. Для этого мы сгенерировали популяцию виртуальных пациентов с повышенным артериальным давлением и смоделировали их лечение АГП.

3) **Персонализация модели** – определение параметров модели для заданного реального пациента – самая сложная задача. Часть параметров модели может быть идентифицирована следующим образом:

- взята из истории болезни пациента;
- рассчитана на основе известных физиологических зависимостей от роста, веса, пола и возраста пациента;
- найдена путем решения обратной задачи на основе клинических измерений, из истории болезни наблюдаемого пациента. Однако, при решении обратной задачи только меньшая часть параметров является идентифицируемой.

Таким образом, большая часть параметров модели не может быть однозначно идентифицирована для заданного пациента. Чтобы решить эту проблему был предложен следующий подход:

- отобрать физиологические параметры, значимо влияющие на артериальное давление и другие характеристики сердечно-сосудистой системы;
- сгенерировать популяцию виртуальных пациентов (200-500), в которой значения этих параметров существенно варьируются в заданных физиологических границах (нормальных или патологических в зависимости от клинического состояния пациента), а наблюдаемые параметры соответствуют данным из истории болезни.

4) **“Лечение” созданной популяции виртуальных пациентов**, т.е. моделирование воздействия АГП. При этом каждый виртуальный пациент реагирует на “лечение” по-своему, не для всех оно будет эффективно. Это позволяет выделить группы виртуальных пациентов со схожей реакцией на лекарственный препарат и определить какие именно параметры определяют разделение на эти группы. В частности, такой анализ дает возможность установить, какие еще лабораторные исследования нужно провести для реального пациента, чтобы понять к какой группе виртуальных пациентов он относится и какое лекарственное лечение будет для него наиболее эффективно.

7.2 Генерация виртуального пациента

В данной работе виртуальный пациент – это математическая модель регуляции артериального давления человека, описанная в разделе 6.4.5 и определяемая системой алгебро-дифференциальных уравнений, в совокупности с множеством значений всех ее параметров и начальных значений переменных.

Как и в исследовании Cheng et al., 2017 мы определяем виртуального пациента как единичную равновесную параметризацию модели. Под равновесной параметризацией мы понимаем начальное состояние модели, при котором значения переменных либо не изменяются (например, систолическое/диастолическое давление), либо устойчиво осциллируют с амплитудой, равной длине сердечного цикла (например, артериальное давление). Эта параметризация соответствует конкретному состоянию пациента, например, «больной» или «здоровый». Реакция пациента на внешние раздражители (например, прием АГП) симулируется путем внесения возмущения в начальные значения модели и последующего решения модифицированной системы для поиска нового равновесия. Параметры (постоянные значения) модели делятся на две группы:

- персональные параметры варьируются в пределах заданных физиологических диапазонов или принимают фиксированное значение в соответствии с клиническим измерением у реального пациента;
- общие параметры принимают одинаковые значения для всех виртуальных пациентов.

Среди переменных модели мы выделяем наблюдаемые (т.е. клинически измеримые) характеристики. Такие переменные можно использовать для проверки соответствия виртуального пациента реальному человеку.

Список оптимизируемых параметров с учетом границ пространства поиска для генерации виртуальных пациентов с АГ, а также коэффициентов чувствительности среднего артериального давления (МАР) приведен в приложении 7.1.

Задача генерации виртуального пациента состоит в том, чтобы найти равновесное состояние модели в заданных физиологических диапазонах. Допустим, у нас есть история болезни с результатами лабораторной диагностики заданного пациента (мониторинг артериального давления, электрокардиография, эхокардиография, анализы крови и т. д.). Мы можем напрямую подставить в модель те измеренные значения, которые являются в ней постоянными (это масса тела и показатели крови). Клинические значения, рассчитываемые в модели (такие как артериальное давление, частота сердечных сокращений, скорость клубочковой фильтрации и т.д.), могут быть получены путем оптимизации параметров модели. Если анамнез пациента подтверждает наличие (или отсутствие) каких-либо сердечно-сосудистых и/или почечных заболеваний, мы также можем учитывать ограничения, налагаемые на ряд параметров и переменных модели, которые не были измерены в явном виде. В случае параметров модели такие ограничения задают границы пространства поиска (нижний и верхний пределы подгоночных параметров). В случае переменных они определяют допустимую область, которая используется для вычисления штрафной функции в задаче

оптимизации параметров. Полный список диапазонов параметров и переменных приводится в работе (Kutumova et al., 2021).

Заметим, агентная модель сходится к состоянию равновесия гораздо медленнее, чем каждый из агентов независимо. Поэтому оптимизация всей модели занимает больше времени. Однако значения среднего артериального давления (*MAP*), сердечного выброса (*CO*) и гематокрита (*Hct*), передаваемые подмоделью сердечно-сосудистой системы почечному агенту (рисунок 7.2.1), поддаются клиническому измерению и могут быть найдены в анамнезе пациента (если нет, то мы можем случайным образом выбрать их из соответствующих физиологических диапазонов). Зафиксировав эти значения, мы можем оценить параметры только почечной системы, передать необходимые значения равновесия сердечно-сосудистому агенту, а затем оптимизировать его параметры, минимизируя расстояния между смоделированными и фиксированными равновесиями *MAP* и *CO* при постоянном значении *Hct*. Рисунок 7.2.1 демонстрирует описанный алгоритм.

7.3 Генерация виртуальной популяции

Виртуальная популяция — это набор уникальных виртуальных пациентов. Для создания популяции с различными значениями систолического артериального давления (САД), диастолического артериального давления (ДАД), частоты сердечных сокращений (ЧСС), массы тела и индекса массы тела (ИМТ) генерировались случайные числа с требуемыми средним значением (*mean*) и стандартным отклонением (*SD*).

Мы проанализировали исходные характеристики пациентов с АГ, включенных в различные клинические испытания алискирена, амлодипина, бисопролола, эналаприла, гидрохлоротиазида и лозартана (Abate et al., 1998; Agabiti Rosei et al., 2005; Brown et al., 2011; Corea et al., 1996; Cushman et al., 2012; Derosa et al., 2014; Fagher et al., 1990; Fermé et al., 1990; Gradman et al., 2005; Koh et al., 2004; Kwakernaak et al., 2017; Lee et al., 2012; Lithell et al., 1987; Mallion, 2007; Martina et al., 2000; Narkiewicz, 2007; Nedogoda et al., 2013; Oparil et al., 1996; Oparil et al., 2008; Pareek et al., 2016; Paterna et al., 2007; Pool et al., 2007; Porthan et al., 2009; Puig et al., 2007; Sun et al., 2016; Tham et al., 1993; Weber et al., 1995; Wing et al., 2003). Основываясь на этих характеристиках, мы выбрали подходящие средние значения и стандартные отклонения для необходимых величин. В таблице 7.3.1 представлены параметры САД, ДАД, ЧСС, ИМТ, массы тела, роста и пола, которые мы использовали для генерации популяции виртуальных пациентов с АГ. Отметим, что для других параметров и переменных модели мы также учитывали физиологические диапазоны, характерные для пациента с неосложненной АГ (Kutumova et al., 2021).

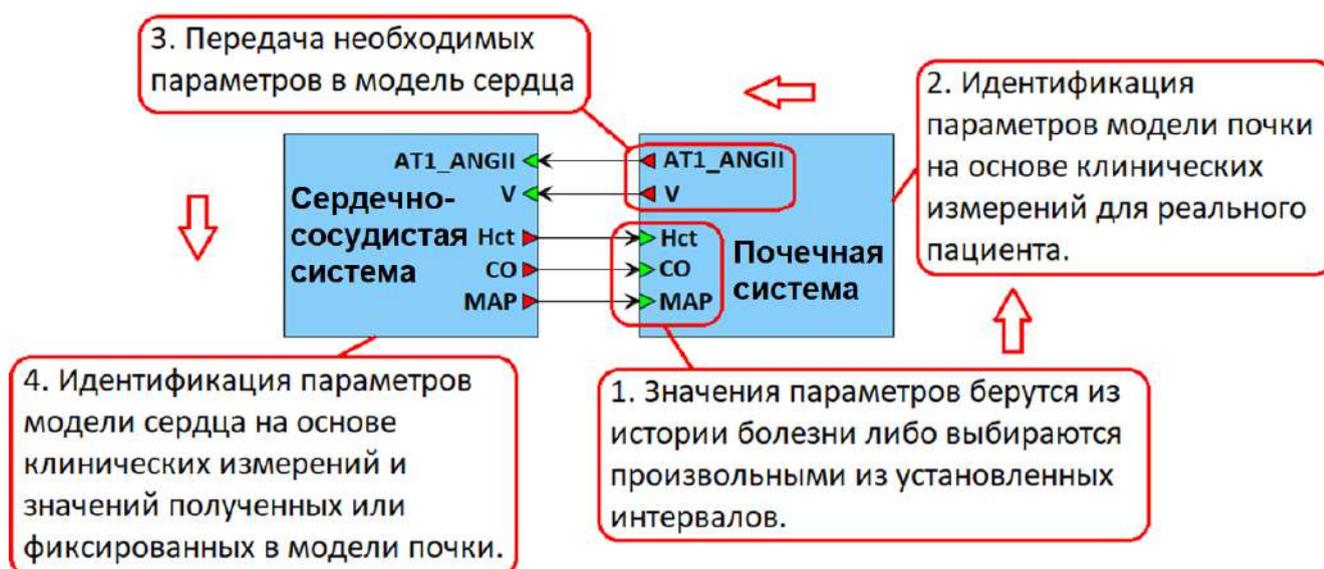


Рисунок 7.2.1 – Алгоритм создания виртуального пациента. Среднее артериальное давление (*MAP*), сердечный выброс (*CO*) и гематокрит (*Hct*) берутся из анамнеза реального пациента или выбираются случайным образом из физиологических диапазонов в соответствии с его заболеваниями. Когда оценка параметров почечной подмодели завершена, равновесные значения общего объема крови (*V*) и концентрации ангиотензина II, связанного с AT1-рецепторами (*AT1_ANGII*), передаются в сердечно-сосудистую подмодель, которая калибруется таким образом, чтобы симулируемые равновесные значения *MAP* и *CO*, а также постоянное значение *Hct* совпадали со значениями, установленными в подмодели почек. Когда калибровка обеих подмоделей завершена, равновесные значения объединяются в единую параметризацию – виртуального пациента.

В качестве критериев включения мы рассматривали гипертоническую болезнь (САД ≥ 140 мм рт. ст. и/или ДАД ≥ 90 мм рт. ст.; САД > 130 мм рт. ст.; ДАД > 80 мм рт. ст.), ИМТ > 22 кг/м² и средний рост (160–180 см). Критериями исключения были тяжелая АГ (САД > 179.5 мм рт. ст. или ДАД > 109.5 мм рт. ст.), сердечно-сосудистые и/или почечные осложнения, аномальная частота сердечных сокращений (ЧСС < 60 или ЧСС > 90 уд/мин), тяжелое ожирение (ИМТ > 36 кг/м²).

Выбрав произвольным образом значения САД, ДАД, ЧСС, ИМТ и массы тела в соответствие с базовыми настройками из таблицы 7.3.1, мы устанавливали значение массы тела в модель и использовали рост (рассчитываемый как $100 \cdot \sqrt{\text{вес}/\text{ИМТ}}$), вес и пол, чтобы оценить границы поиска для начального значения общего количества воды в организме и ограничений для общего объема крови с использованием формул Надлера и Мура (Nadler et al., 1962; Moore, 1967). При этом значение среднего артериального давления рассчитывалось по формуле:

$$MAP = (САД + 2 \cdot ДАД) / 3 \quad (7.3.1)$$

После этого виртуальный пациент генерировался в соответствии с алгоритмом, описанном в пункте 7.2.

Таблица 7.3.1 – Базовые настройки для генерации виртуальной популяции с АГ

Физиологические характеристики	Использование в модели	Базовые настройки (mean ± SD)	Критерий исключения (логический оператор)
САД, мм рт.ст.	переменная	160 ± 10	(САД < 140 и ДАД < 90) или ДАД > 109.5 или САД > 179.5 или САД < 130 или ДАД < 80
ДАД, мм рт.ст.	переменная	100 ± 10	
ЧСС, уд./мин	переменная	75 ± 10	ЧСС < 60 или ЧСС > 90
ИМТ, кг/м ²	–	29 ± 5	ИМТ < 22 или ИМТ > 36
Вес, кг	параметр	80 ± 20	–
Рост, см	Оценка общего объема крови по формуле Надлера (Nadler et al., 1962)	–	рост < 160 или рост > 180
Пол, М:Ж		50% : 50%	–

7.4 Моделирование действия антигипертензивных препаратов

Всем пациентам с АГ или повышенным артериальным давлением рекомендуется изменение образа жизни (отказ от курения, повышение физической активности, нормализация массы тела, ограничение соли в рационе и т.д.). Время начала лекарственной терапии (одномоментно с немедикаментозными вмешательствами или после них) определяется уровнем клинического артериального давления, уровнем сердечно-сосудистого риска, наличием поражения органов-мишеней или сопутствующих сердечно-сосудистых заболеваний.

Ренин-ангиотензин-альдостероновая система (РААС) и ее гиперактивация является одним из ключевых факторов в патогенезе артериальной гипертензии. Основным звеном патогенеза является ангиотензин II, который реализует свое действие через рецепторы ангиотензина I типа и обладает выраженным сосудосуживающим действием на гладкомышечную мускулатуру, а также усиливает сократимость миокарда, стимулирует продукцию альдостерона, выброс катехоламинов, оказывает профибротические и провоспалительные эффекты, тем самым способствуя ремоделированию миокарда, прогрессированию эндотелиальной дисфункции и повреждению почек (рисунок 7.4.1).

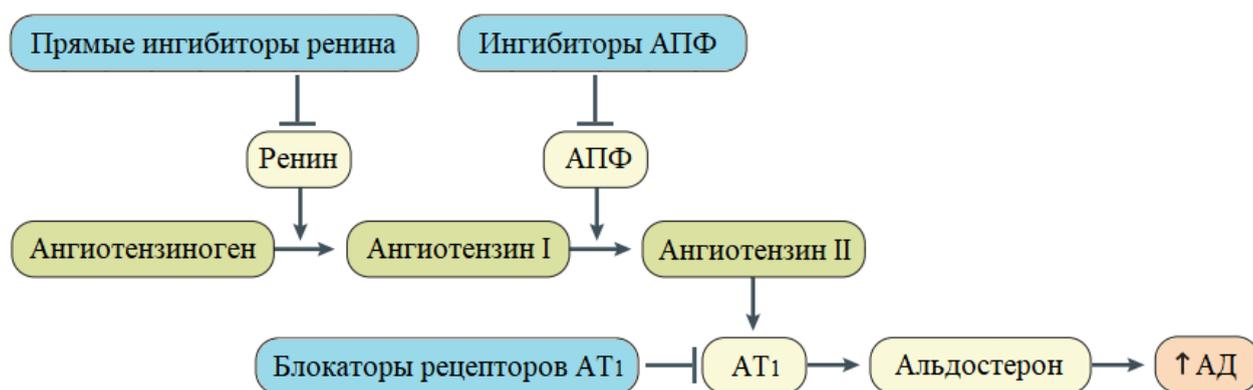


Рисунок 7.4.1 – Роль РААС в регуляции артериального давления (АД), а также влияние АПФ на эту систему. Выброс ренина происходит в ответ на различные сигналы, например, стимуляцию адреналином, уменьшение абсорбции хлорида натрия, а также уменьшение кровяного давления в клубочках почки (Sethi et al., 2003). АТ₁-рецепторы ангиотензина II способствуют секреции альдостерона, усилению реабсорбции натрия в почечных канальцах, сужению просвета кровеносных сосудов, повышению их общего периферического сопротивления и, как следствие, повышению артериального давления (Opavil et al., 2018). Ингибиторы ренина и АПФ, а также блокаторы рецепторов АТ₁ снижают активность РААС, стимулируя снижение артериального давления

В настоящее время для лечения АГ рекомендованы пять основных классов АГП (Чазова и др., 2015; Opavil et al., 2018):

- 1) ингибиторы ангиотензинпревращающего фермента (АПФ) – снижение артериального давления через воздействие на РААС (рисунок 7.4.1);
- 2) блокаторы АТ₁-рецепторов ангиотензина II – снижение артериального давления через воздействие на РААС (рисунок 7.4.1);
- 3) антагонисты кальция – выраженное сосудорасширяющее действие;
- 4) β-адреноблокаторы – снижение сердечного выброса, ЧСС, угнетение активности симпатической нервной системы;
- 5) диуретики – уменьшение объема жидкости, циркулирующей в крови, ускорение выведения натрия.

В качестве дополнительных классов АГП для комбинированной терапии могут использоваться α-адреноблокаторы, агонисты имидазолиновых рецепторов и прямые ингибиторы ренина (Чазова и др., 2015). Последние совместно с ингибиторами АПФ и блокаторами АТ₁-рецепторов способствуют понижению артериального давления посредством воздействия на компоненты РААС (рисунок 7.4.1).

Тактика лечения (моно- или комбинированная терапия, препараты быстрого всасывания или с пролонгированным действием) подбирается индивидуально для каждого пациента. В силу разнообразия патогенетических механизмов выбор лекарственной терапии остается сложной задачей (Sethi et al., 2003).

В качестве представителей различных классов АГП мы рассмотрели:

- алискирен – селективный ингибитор ренина (Gradman et al., 2005; Колпакова, 2014);
- лозартан – блокатор AT_1 -рецепторов ангиотензина II (Sica et al., 2005);
- бисопролол – селективный β_1 -адреноблокатор (Steinmann et al., 1986);
- эналаприл – ингибитор АПФ (Ferguson, 1982);
- амлодипин – блокатор кальциевых каналов (Faulkner, 1986);
- гидрохлортиазид – тиазидный диуретик (Carter et al., 2004).

Для каждого АГП были определены его точки воздействия на модель регуляции артериального давления (рисунок 7.4.2, таблица 7.4.1). Хотя в данной работе мы моделировали только фармакодинамику АГП, при дальнейшем развитии этого подхода могут также использоваться фармакокинетические модели. Поэтому на примере алискирена и лозартана показано использование ПК BioUML для создания соответствующих фармакокинетических моделей.

7.4.1 Алискирен

Алискирен – низкомолекулярный ингибитор ренина непептидной структуры (Gradman et al., 2005; Колпакова 2014). В качестве основы для моделирования его действия на сердечно-сосудистую систему человека человека мы взяли ФК-модель, описанную в работе (Hong et al., 2008), и ФД-модель из работы (Steinmann et al., 1986), после чего дополнили эти модели дискретным событием, имитирующим прием препарата. Данное событие (рисунок 7.4.3.A) характеризует мгновенное увеличение суммарной дозы алискирена A_p на дозу однократного приема $Dose$ и происходит в том случае, когда время t , прошедшее с момента предыдущего приема, достигает определенного значения *Interval* при условии, что количество приемов не превысило заданного порога *Injections*.

ФК-модель алискирена (Hong et al., 2008) – классическая двухкомпарментная модель, которая делит организм в зависимости от скорости распределения вещества на центральный и периферический компартменты (Bonate, 2011). К центральному компартменту относят плазму крови и хорошо перфузируемые органы – сердце, легкие, печень и почки; к периферическому – кожу, жир и мышцы (Гельфанд и Салтанов, 2011).

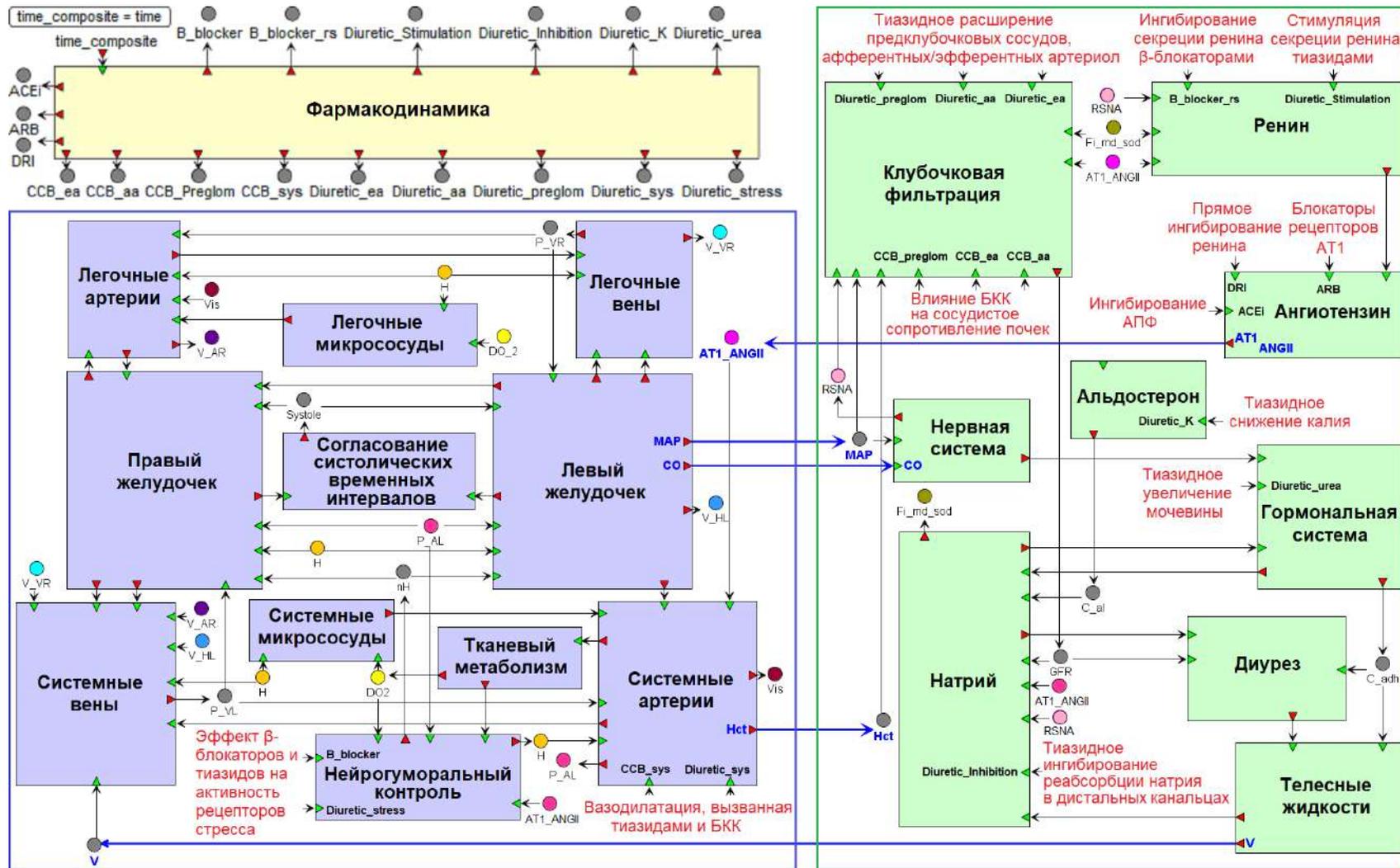


Рисунок 7.4.2 – Комплексная модель регуляции артериального давления человека, реализованная в платформе BioUML. Зеленым отмечены модули, взятые из расширенной модели Гайтона – Карааслана, синим – из модели Солодяникова. Красными стрелками отмечены точки воздействия АГП (Kutumova et al., 2022a)

Таблица 7.4.1 – Точки воздействия АГП на модель регуляции артериального давления человека

АГП	Модуль	Точки воздействия	Коэффициенты воздействия
Алискирен, 150 или 300 мг/день	Ангиотензин	Активность ренина плазмы	$\frac{0.99 \cdot D}{D + 20}$
Амлодипин, 5 мг/день	Клубочковая филтрация	Сопротивление афферентных артериол	0.413
		Сопротивление эфферентных артериол	0.107
		Сопротивление междольковых, дугowych, междольковых артерий	0.413
	Системные артерии	Сопротивление системных микрососудов	0.107
Бисопролол, 5 мг/день	Нейрогумораль- ный контроль	Активность рецепторов стресса	0.371
	Ренин	Скорость секреции ренина	0.933
Эналаприл, 20 мг/день	Ангиотензин	Скорость превращения ангиотензина I в ангиотензин II под действием АПФ	0.996
Гидрохлоротиа зид, 12.5 мг/день	Натрий	Фракционная дистальная скорость реабсорбции натрия	0.304
	Ренин	Скорость секреции ренина	1.113
	Клубочковая филтрация	Сопротивление афферентных артериол	0.469
		Сопротивление эфферентных артериол	0.302
		Сопротивление междольковых, дугowych, междольковых артерий	0.581
	Системные артерии	Сопротивление системных микрососудов	$\min(\frac{time}{1950000} \cdot 0.074, 0.0$
	Нейрогумораль- ный контроль	Активность стрессовых рецепторов	0.389
	Альдостерон	Уровень калия в крови	0.030
Гормональная система	Уровень мочевины в крови	0.100	
Лозартан, 50 или 100 мг/день	Ангиотензин	Скорость связывания ангиотензина II с рецепторами AT1	0.886 (для 50 мг) 0.954 (для 100 мг)

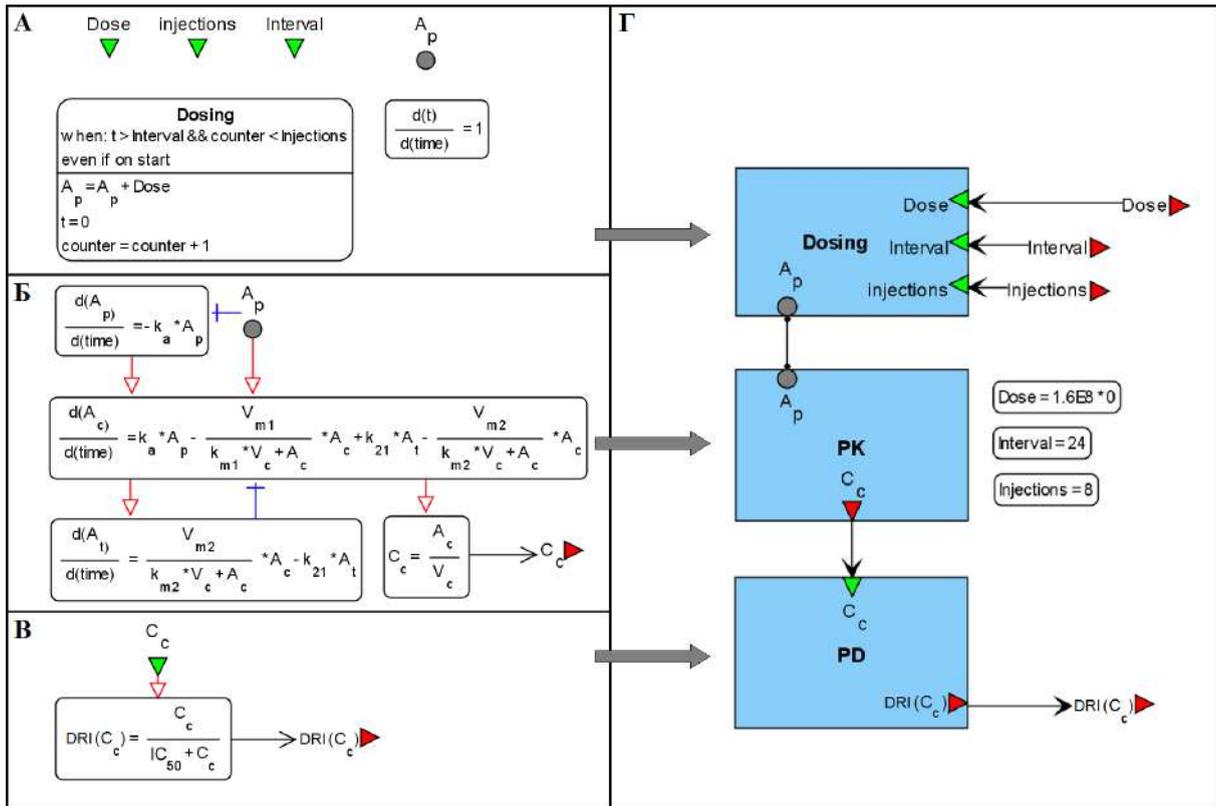


Рисунок 7.4.3 – Модульная модель фармакокинетики и фармакодинамики алискирена. **А.** Прием препарата задается при помощи дискретного события и зависит от трех параметров: лекарственной дозы (*Dose*), числа приемов (*Injections*) и интервала между ними (*Interval*). **Б.** ФК-модель алискирена. A_p – суммарная доза принятого препарата; A_c и A_t – количество алискирена в центральном и периферическом компартменте соответственно; C_c – концентрация лекарства, оказывающая терапевтический эффект. Детальное описание модели и численные значения параметров взяты из работы (Hong et al., 2008). **В.** ФД-модель алискирена – терапевтический эффект, оказываемый препаратом. IC_{50} – концентрация алискирена в центральном компартменте, которая обеспечивает эффект препарата, равный половине от максимально возможного. **Г.** Модульное представление модели алискирена

Алискирен абсорбируется с линейной скоростью $k_a A_p$ в центральный компартмент, где одна его часть элиминируется, а другая часть транспортируется в периферический компартмент со скоростями, определяемыми по закону Михаэлиса – Ментен (рисунок 7.4.3.Б).

Концентрация алискирена C_c , оказывающая терапевтический эффект, определяется как количество лекарства в центральном компартменте A_c , отнесенное к его объему V_c :

$$C_c = \frac{A_c}{V_c} \quad (7.4.1)$$

Данный эффект (*DRI* – direct renin inhibition, прямое ингибирование ренина) выражается как функция от дозы препарата (Hallow et al., 2014):

$$DRI(dose) = \frac{E_{\max} \cdot dose}{dose + ED_{50}}, \quad (7.4.2)$$

где $E_{\max} = 0.99$ – максимальный эффект, который он может дать (т.е. активность ренина снижается на 99 %), а $ED_{50} = 20$ мг – доза, эффект от которой составляет половину E_{\max} .

Имея в наличии ФК-модель алискирена (Hong et al., 2008), мы модифицировали формулу (7.4.2) с учетом функции изменения его концентрации в плазме крови (7.4.1):

$$DRI(C_c) = \frac{C_c}{IC_{50} + C_c}. \quad (7.4.3)$$

Здесь IC_{50} обозначает концентрацию лекарства в центральном компартменте, которая обеспечит эффект $0.5 \cdot E_{\max}$. Численные расчеты, которые мы произвели для данной модели при входном значении $A_p = 20$ мг, дали на выходе среднее значение $C_c^{cp} = 0.58$ нг/мл. Его мы и взяли в качестве значения для параметра IC_{50} .

Приняв формулу (7.4.3) в качестве ФД-модели алискирена (рисунок 7.4.3.В), мы построили для него модульную модель (рисунок 7.4.3.Г), входными параметрами которой являются доза алискирена для однократного приема, число приемов и интервал времени между ними. Выходной параметр – антигипертензивный эффект алискирена, на основе которого происходит снижение активности ренина в модуле «Ангиотензин». Результаты сравнения численной динамики модели и экспериментальных данных приведены на рисунке 7.4.4.

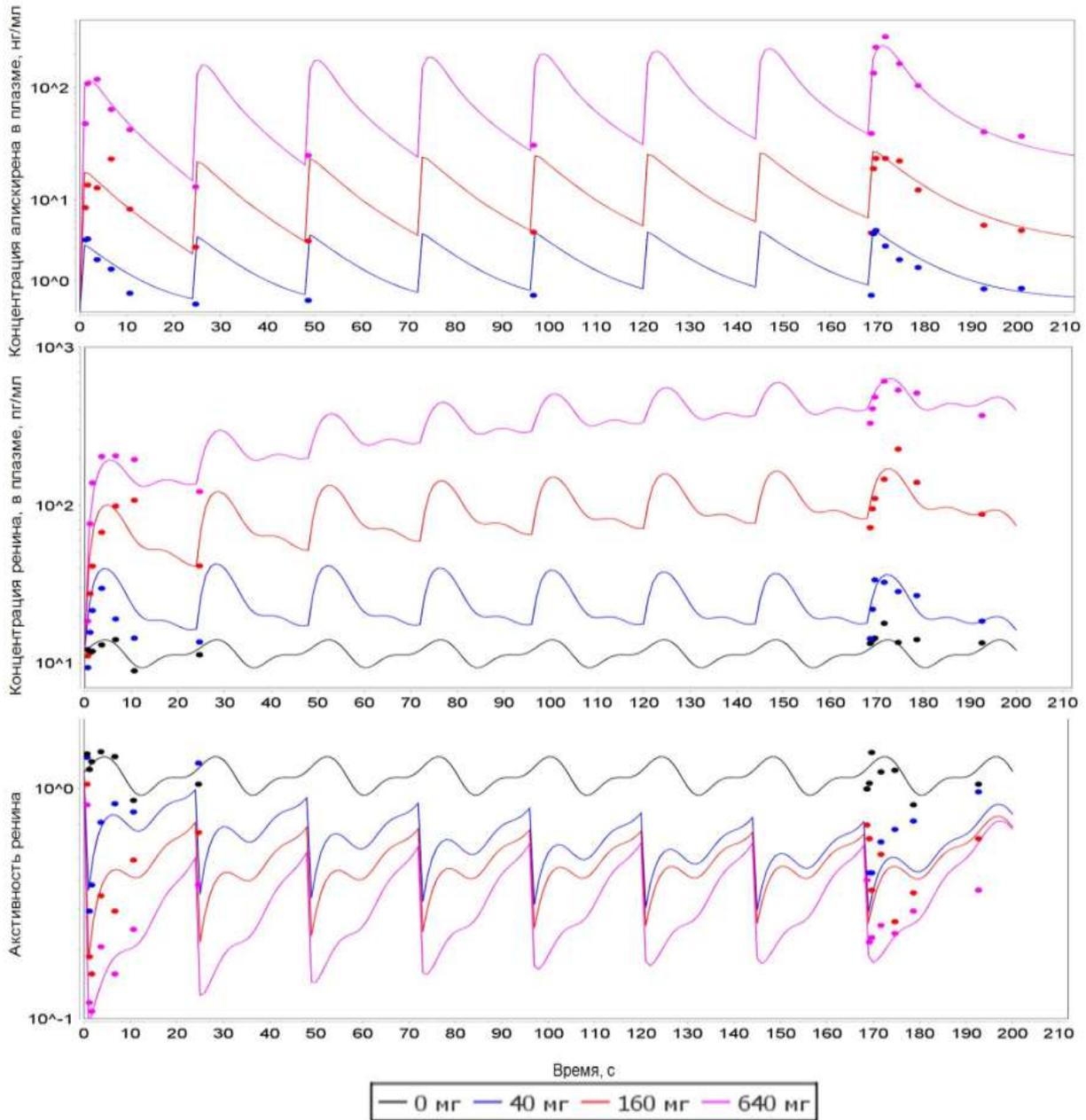


Рисунок 7.4.4 – Результаты расчетов модульной модели ФК/ФД алискирена в сравнении с экспериментальными данными (Nussberger et al., 2002) для 4 различных дозировок (0 мг, 40 мг, 160 мг, и 640 мг)

7.4.2 Лозартан

Лозартан – специфический блокатор рецепторов ангиотензина II (ANG II) типа AT_1 (Sica et al., 2005). Активация этих рецепторов происходит в модуле «Ангиотензин» посредством их связывания с ANG II и формирования комплекса $AT_1:ANGII$ со скоростью, вычисляемой по формуле (Hallow et al., 2014):

$$v_{AT_1:ANG II}(t) = c_{AT_1} \cdot [ANG II] \quad (7.4.4)$$

Здесь квадратные скобки обозначают концентрацию соответствующего вещества, а c_{AT_1} – константа. Фармакодинамический эффект лозартана *ARB* (angiotensin receptor-blocking, блокирование рецепторов ангиотензина) в работе (Hallow et al., 2014) моделируется путем замены этой константы на функцию от дозы препарата:

$$c_{AT_1}(dose) = c'_{AT_1} \cdot (1 - ARB(dose)), \quad (7.4.5)$$

где c'_{AT_1} – константа скорости связывания AT_1 и ANG II при отсутствии терапии, а значение $ARB(dose)$ за неимением достаточного количества экспериментальных данных для идентификации параметров E_{max} и ED_{50} в формуле (7.4.2) полагается равным 0.954 (оценка на основе экспериментальных данных Nedogoda et al., 2013) и характеризует терапевтический эффект для дозы лозартана, равной 100 мг. Значение $ARB = 0.886$ для дозировки 50 мг было рассчитано на основе данных клинических испытаний, приведенных в (Porthan et al., 2009).

Также нами была построена фармакокинетическая модель лозартана (Nguyen et al., 2017), которая учитывает превращение лозартана в его карбокси-кислотный метаболит (EXP 3174) под действием цитохрома CYP2C9 в клетках печени (рисунок 7.4.5). Стоит отметить, что именно EXP 3174 (carboxylosartan) оказывает основное терапевтическое действие. Результаты сравнения численной динамики модели и экспериментальных данных приведены на рисунке 7.4.6.

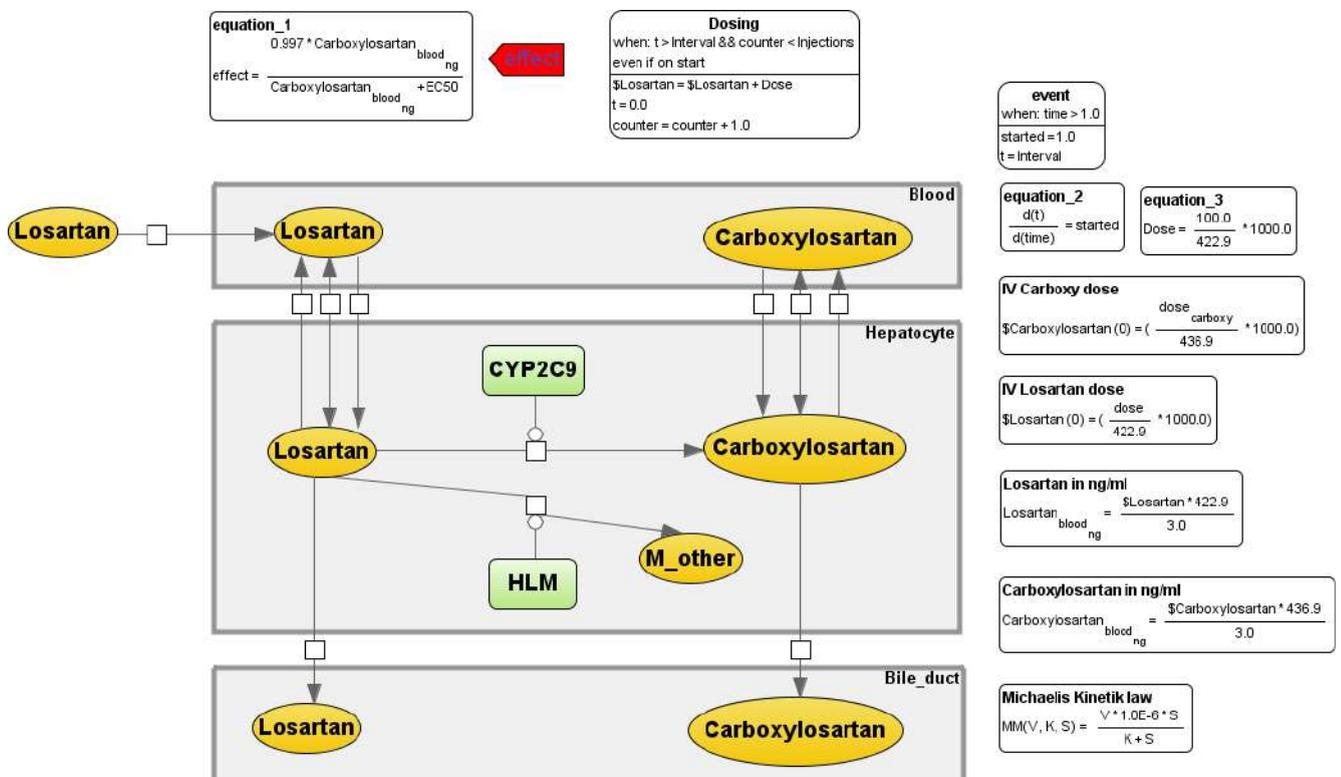


Рисунок 7.4.5 – Модель фармакокинетики лозартана (Nguyen et al., 2017)

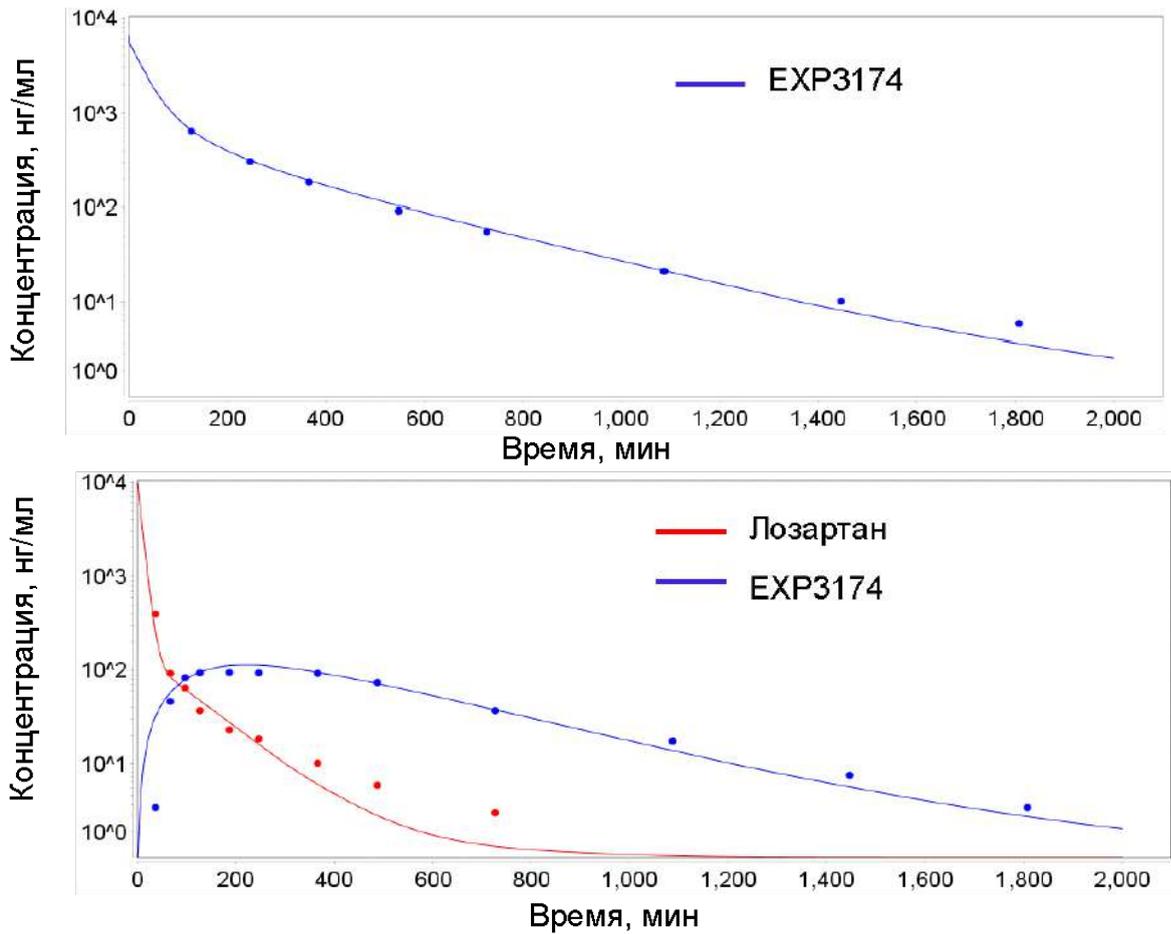


Рисунок 7.4.6 – Сравнение моделирования фармакокинетики лозартана с экспериментальными данными

7.4.3 Бисопролол

Бисопролол – ингибитор β_1 -адренорецепторов (Steinmann et al., 1986), фармакодинамический эффект оказывает в модуле «Нейрогуморальный контроль». В этом модуле сердечный центр представлен как система управления, которая задает выходную величину нервно-гуморального фактора H на основе суммарного сигнала рецепторов r_{sum} (Прошин и Солодянников, 2006):

$$\frac{dH}{dt} = a \cdot \left(H_{\text{max}} \cdot \frac{1 - e^{-2r_{\text{sum}}}}{1 + 18e^{-2r_{\text{sum}}}} - H \right), \quad (7.4.6)$$

где a и H_{max} – константы. Величину r_{sum} формируют значения барорецептора rB , а также рецепторов дыхательного центра rV , фактора усталости rD и стресса rS с учетом соответствующих констант h_{baro} , h_{VO_2} , h_{oxygen} и h_{stress} , а также константы h_{base} :

$$r_{\text{sum}} = h_{\text{baro}} \cdot rB + h_{VO_2} \cdot rV + h_{\text{oxygen}} \cdot rD + h_{\text{stress}} \cdot rS + h_{\text{base}}. \quad (7.4.7)$$

Значения рецепторов вычисляются по формулам:

$$rB = r^-(P_{AL}), \quad rV = r^-(VO_2), \quad (7.4.8)$$

$$rD = r^+(DO_2), \quad rS = r^+(S), \quad (7.4.9)$$

где P_{AL} – давление в артериях большого круга (определяемое в модуле «Артериальное дерево»), VO_2 и DO_2 – венозная концентрация кислорода и кислородный долг (модуль «Тканевый метоболизм»), $S \in [0,1]$ – фактор стресса, отражающий уровень адреналина в крови, а $r^+(\varphi)$ и $r^-(\varphi)$ – сигмоидные функции активности:

$$r^+(\varphi) = \frac{1 - e^{-\alpha(\varphi - \varphi_0)}}{1 + \beta e^{-\alpha(\varphi - \varphi_0)}}, \quad (7.4.10)$$

$$r^-(\varphi) = 1 - \frac{1 - e^{-\alpha(\varphi - \varphi_0)}}{1 + \beta e^{-\alpha(\varphi - \varphi_0)}}. \quad (7.4.11)$$

Значения констант α , β и φ_0 для всех рецепторов взяты из работы (Прошин и Солодяников, 2006).

Действие бисопролола мы моделируем путем уменьшения рецептора стресса:

$$rS = r^+(S) \cdot (1 - B_{\text{block}}). \quad (7.4.12)$$

Помимо этого мы учли, что β -блокаторы подавляют секрецию ренина в почках, снижают его уровень и активность (Laragh and Sealey, 2011), в связи с чем добавили множитель $(1 - B_{\text{block_rs}})$ для переменной, определяющей скорость секреции ренина в модуле «Ренин».

Для идентификации значения параметров $B_{\text{block}} = 0.371$ и $B_{\text{block_rs}} = 0.933$, характеризующих прием 5 мг препарата, мы использовали клинические данные (Porthan et al., 2009).

7.4.4 Эналаприл

Эналаприл – ингибитор АПФ (Ferguson et al., 1982), катализирующий расщепление ангиотензина I (ANG I) до ангиотензина II (ANG II) в модуле «Ангиотензин» со скоростью (Hallow et al., 2014):

$$v_{\text{ANG II}}(t) = c_{\text{ACE}} \cdot [\text{ANG I}] \quad (7.4.13)$$

Фармакодинамический эффект $ACEi$ данного препарата можно задать через уменьшение константы c_{ACE} :

$$c_{ACE} = c'_{ACE} \cdot (1 - ACEi), \quad (7.4.14)$$

где c'_{ACE} – скорость преобразования ANG I в ANG II при отсутствии лекарственной дозы, а $ACEi = 0.996$ для дозы в 20 мг.

7.4.5 Амлодипин

Амлодипин блокирует кальциевые каналы L- и N-типа, вызывает расширение (и, как следствие, снижение сопротивления) афферентных (приносящих) и эфферентных (выносящих) артериол (Hayashi et al., 2007), а также предклубочковых артерий почки (таких как междольевые и дугообразные артерии, корковые лучевые артерии и т.д.) (Hallow et al., 2014).

Оказывает воздействие в модуле «Клубочковая фильтрация». Согласно Hallow et al., 2014, почка определена как сеть параллельных нефронов. Полагая, что сопротивление артериол нефункционирующих нефронов равно бесконечности, получаем, что общее сосудистое сопротивление почек вычисляется по формуле:

$$R = \frac{R_{aa} + R_{ea}}{n} + R_{preglom}, \quad (7.4.15)$$

где n – число функционирующих нефронов, R_{aa} и R_{ea} – сопротивления афферентных и эфферентных артериол соответственно, а $R_{preglom}$ – общее предклубочковое сопротивление. Значения R_{aa} , R_{ea} и $R_{preglom}$ зависят от многих параметров модели. На них, в частности, влияет концентрация комплекса $AT_1:ANGII$ (вычисляемая в модуле «Ангиотензин»), почечная симпатическая нервная активность (модуль «Нервная система») и скорость тока натрия через Macula densa (модуль «Натрий»). Опустив детали, которые подробно описаны в работе (Hallow et al., 2014), обозначим через R'_{aa} , R'_{ea} и $R'_{preglom}$ значения параметров R_{aa} , R_{ea} и $R_{preglom}$ при отсутствии антигипертензивной терапии. Тогда фармакодинамический эффект амлодипина можно смоделировать по формулам:

$$R_{aa} = R'_{aa} \cdot (1 - CCB_{aa}), \quad (7.4.16)$$

$$R_{ea} = R'_{ea} \cdot (1 - CCB_{ea}), \quad (7.4.17)$$

$$R_{preglom} = R'_{preglom} \cdot (1 - CCB_{preglom}), \quad (7.4.18)$$

где значения $CCB_{aa} = CCB_{preglom} = 0.413$ и $CCB_{ea} = 0.107$ для дозы 5 мг препарата характеризуют более сильное влияние амлодипина на афферентные артериолы и предклубочковые артерии и

более слабое влияние на эфферентные артериолы (оценка параметров производилась на основе данных Nedogoda et al., 2013).

Отметим, что амлодипин действует не только на параметры общего сосудистого сопротивления почек, но и на общее сопротивление артериальной системы R_a (Hallow et al., 2014). В нашей модели данное сопротивление участвует опосредованно через параметр проводимости капилляров большого круга кровообращения Y_{ALVL} , обратной величиной которого он является. Значение Y_{ALVL} используется в модуле «Системные артерии» для расчета потока крови из артерий большого круга (с давлением P_{AL}) в вены большого круга (с давлением P_{VL}) по формуле (Karaaslan et al., 2005):

$$F_{ALVL} = Y_{ALVL} \cdot (P_{AL} - P_{VL}) \quad (7.4.19)$$

Чтобы учесть влияние амлодипина CCB_{sys} , мы изменили формулу (8.19) по аналогии с формулами 8.16-8.18:

$$F_{ALVL} = \frac{Y_{ALVL}}{1 - CCB_{sys}} \cdot (P_{AL} - P_{VL}), \quad (7.4.20)$$

положив при этом, как и в модели (Hallow et al., 2014), $CCB_{sys} = CCB_{ea}$.

7.4.6 Гидрохлоротиазид

Гидрохлоротиазид относится к классу бензотиадиазинов, называемых просто тиазидными диуретиками (Carter et al., 2004). Тиазид-индуцированное снижение артериального давления включает дифференцировку на острую и хроническую фазы (Rapoport and Soleimani, 2019). Резкое снижение коррелирует с диурезом и уменьшением объема плазмы, связанным с ингибированием реабсорбции натрия в дистальных канальцах (Duarte and Cooper-DeHoff, 2010). Этот феномен описывается действием препарата ($1 - Diuretic_{Inhibition}$) на нормальное значение фракционной реабсорбции натрия в дистальных канальцах (Hallow et al., 2014) в модуле «Натрий».

Возможный механизм хронического снижения давления включает расширение сосудов (Rapoport and Soleimani, 2019; Duarte and Cooper-DeHoff, 2010). Поэтому по аналогии с Hallow et al., 2014, мы ввели в модуль «Клубочечная фильтрация» влияния ($1 - Diuretic_{aa}$), ($1 - Diuretic_{ea}$) и ($1 - Diuretic_{preglom}$) на резистентность афферентных артериол, эфферентных артериол и междольковые/аркуатные/междольковые артерии соответственно. Рассматривая сходное влияние на системные микрососуды, мы также учитывали следующие факты:

- внеклеточная жидкость (как и объем плазмы) сначала уменьшается под действием тиазидов, а затем почти полностью восстанавливается в течение 4–6 недель непрерывного лечения (Duarte and Cooper-DeHoff, 2010; Rapoport and Soleimani, 2019; Leth, 1970);

- сердечный выброс уменьшается с потерей объема плазмы, а затем возвращается к исходному уровню при длительном лечении тиазидами (Duarte and Cooper-DeHoff, 2010; van Brummelen et al., 1980).

Мы обнаружили, что восстановление этих показателей можно воспроизвести в модуле «Системные артерии» при нарастающем влиянии $1/(1 - Diuretic_{sys})$ на проводимость системных микрососудов (обратное их сопротивлению):

$$Diuretic_{sys}(time) = \min\left(E_{max} \cdot \frac{time}{duration}, E_{max}\right), \quad (7.4.21)$$

где E_{max} обозначает максимальное влияние, а $duration$ — время достижения E_{max} .

В качестве других мишеней гидрохлоротиазида в модели мы рассматривали постоянные уровни мочевины и калия в плазме крови. Поскольку гидрохлоротиазид повышает уровень мочевины и снижает содержание калия (Scaglione et al., 1992; Scaglione et al., 1995; Devineni et al., 2014), мы добавили к этим параметрам множители $(1 + Diuretic_{urea})$ и $(1 - Diuretic_{potassium})$ в модулях «Гормональная система» и «Альдостерон» соответственно.

Кроме того, клиническое исследование Villamil et al., 2007, показывает, что монотерапия гидрохлоротиазидом повышает активность ренина в плазме. Поэтому, следуя модели Hallow et al., 2014, мы учитывали прямое влияние $(1 + Diuretic_{stimulation})$ на секрецию ренина (модуль «Ренин»).

Другая терапевтическая мишень препарата основана на том факте, что частота сердечных сокращений существенно не изменяется при длительном лечении (van Brummelen et al., 1980; Shah et al., 1978; Scaglione et al., 1992; Scaglione et al., 1995). Различные исследования показывают, что прессорный эффект норадреналина снижается или остается неизменной при применении тиазидных диуретиков (Rapoport and Soleimani, 2019). Поэтому мы предположили влияние $(1 - Diuretic_{stress})$ на активность рецепторов стресса в модуле «Нейрогуморальный контроль».

Для оценки фармакодинамических параметров гидрохлоротиазида в дозировке 12.5 мг, значения которых приведены в таблице 7.4.1, мы использовали клинические данные MacKay et al., 1996.

7.5 Валидация модели на основе клинических данных

Использование известных экспериментально полученных данных фармакологического действия АГП для валидации реализованной модели было сопряжено с рядом трудностей, поскольку в клинических исследованиях:

- приведены обобщенные статистические данные по исследованным группам пациентов (как правило, среднее и стандартное отклонение);

- приведено очень ограниченное количество физиологических и биохимических показателей, по которым можно определить только часть параметров модели;
- пациенты наряду с АГ могут иметь другие заболевания (например, диабет, ишемическая болезнь сердца, ожирение), которые пока не могут быть адекватно отражены в созданной модели (поэтому такие клинические данные не были использованы);
- пациенты могут быть подобраны или разбиты по группам, признаки которых мы не можем учесть в модели (например, беременность, курение и т.д.). Такие клинические данные так же не были использованы.

С учетом вышесказанного для идентификации фармакодинамических параметров модели применялся следующий подход:

- мы сгенерировали популяцию виртуальных пациентов с неосложненной артериальной гипертонией на основе алгоритма, изложенного в разделе 7.3 ;
- смоделировали лечение этой популяции описанными выше АГП;
- подобрали результаты клинических исследований, в которых выборка пациентов была похожа на сгенерированную виртуальную популяцию;
- сравнили полученные в результате моделирования результаты и соответствующие клинические данные.

Численные расчеты показали, что снижение систолического и диастолического давления при приеме перечисленных АГП находится в хорошем согласии с результатами клинических испытаний, которые мы нашли в литературе. Ниже мы рассмотрим эти шаги более подробно.

7.5.1 Генерация популяции виртуальных пациентов

Одна из методик генерации виртуальных пациентов описана в работе (Allen et al., 2016). Авторы предлагают создать большое количество (сотни тысяч) «правдоподобных пациентов», определяемых как модельные параметризации в пределах биологически правдоподобных диапазонов, а затем выбрать из них виртуальную популяцию, соответствующую желаемому эмпирическому распределению. Чтобы найти правдоподобного пациента, они берут входные данные модели в пределах предопределенных правдоподобных границ и оптимизируют этот выбор до тех пор, пока требуемые выходные данные также не попадут в правдоподобные диапазоны. Однако авторы отмечают, что их подход неприменим к моделям, которые рассчитываются медленно (например, с динамикой в нескольких временных масштабах, как в случае с нашей моделью) из-за высоких вычислительных затрат для получения большой правдоподобной совокупности. Более простой способ создания виртуальной популяции обсуждается Hallow et al. (2014; 2020), где входные данные отбираются из предопределенных диапазонов и считаются действительными виртуальными пациентами только в том случае, если они дают выходные данные с физиологически приемлемыми значениями. Такой подход может

быть актуален при относительно небольшом количестве входов и выходов (авторы рассматривают 11 входов/13 выходов в одном исследовании, Hallow et al., 2014, и 14 входов/3 выхода в другом исследовании, Hallow et al., 2020). Однако при большом количестве входов/выходов (в нашем случае 57 входов и 54 выхода) это приводит к перебору огромного количества решений с высокой степенью отбраковки. Поэтому мы применяем собственный метод генерации виртуальной популяции. Как и Аллен с соавт. (Allen et al., 2016), мы используем методы оптимизации с ограничениями для создания виртуальных пациентов, что гарантирует попадание выходных данных в физиологически приемлемые диапазоны. Кроме того, для каждого пациента мы минимизируем целевую функцию расстояний между смоделированными и нормально распределенными случайными значениями САД, ДАД и ЧСС, что непосредственно дает нам популяцию с заранее заданными эмпирическими распределениями этих величин (подробнее см. пункт 7.3). Для тестирования модели мы создали популяцию из 186 виртуальных пациентов с АГ. Исходные характеристики этой популяции показаны в таблице 7.5.1.

Таблица 7.5.1 – Базовые характеристики виртуальной популяции для тестирования модели

Переменные и параметры модели	Значения в виртуальной популяции ($n = 186$), среднее \pm стандартное отклонение
САД, мм рт.ст.	154.26 \pm 7.00
ДАД, мм рт.ст.	101.34 \pm 5.64
ЧСС, уд./мин	75.81 \pm 7.38
ИМТ, кг/м ²	28.48 \pm 3.04
Вес, кг	81.54 \pm 9.57
Рост, см	169.20 \pm 5.72
Пол, М/Ж	100/86

7.5.2 Моделирование АГП терапии

В качестве стандартных дозировок для изучаемых лекарственных препаратов мы рассмотрели 150 и 300 мг в день (мг/д) для алискирена, 50 и 100 мг/д для лозартана, 20 мг/д для эналаприла, 5 мг/д для амлодипина, 5 мг/д для бисопролола и 12.5 мг/д для гидрохлоротиазида. В качестве модельного времени мы использовали интервал, равный 4 неделям.

В таблице 7.5.2 даны результаты численного моделирования артериального давления для популяции виртуальных пациентов до и после антигипертензивной терапии.

7.5.3 Сравнение с результатами клинических исследований

На рисунке 7.5.1 представлено сравнение результатов численного моделирования с данными клинических исследований, перечисленными в таблице 7.5.1. Как видно из

приведенных на рисунке 7.5.1 данных, результаты численного моделирования лечения АГ основными классами АГП хорошо соответствуют клиническим данным.

Таблица 7.5.2 – Результаты численного моделирования изменения систолического и диастолического давления в популяции 186 виртуальных пациентов под действием АГП (среднее значение \pm стандартное отклонение)

Препарат	Доза (мг/день)	Систолическое давление		Диастолическое давление	
		До лечения	После лечения	До лечения	После лечения
Алискирен	150	154.3 \pm 7.0	145.6 \pm 7.3	101.3 \pm 5.6	92.8 \pm 5.5
	300		142.8 \pm 7.4		91.1 \pm 5.5
Лозартан	50		145.7 \pm 7.3		92.8 \pm 5.5
	100		141.2 \pm 7.5		90.2 \pm 5.5
Бисопролол	5		147.4 \pm 8.3		88.3 \pm 5.7
Эналаприл	20		138.6 \pm 7.7		88.9 \pm 5.5
Амлодипин	5		144.4 \pm 6.7		93.3 \pm 5.2
Гидрохлоротиазид	12.5		140.6 \pm 8.6		91.9 \pm 5.3

7.6 Персонализация параметров модели виртуального пациента

Самый сложный этап построения виртуального пациента – это персонализация модели, т.е. задания параметров модели для заданного пациента на основе его клинических данных и истории болезни.

Как было описано в разделе 7.2 мы делим все параметры (константы и переменные модели) на 3 группы:

- 1) измеряемые клинические параметры – берутся из истории болезни реальных людей. Используются для создания виртуального пациента, соответствующего некоторому реальному человеку;
- 2) скрытые персональные параметры – физиологические параметры, которые напрямую связаны с артериальным давлением и другими показателями сердечно-сосудистой системы, но прямое измерение которых является затруднительным (например, из-за сложности или высокой стоимости эксперимента), т.е. значения этих параметров недоступны из истории болезни пациентов;
- 3) остальные (“технические”) параметры – их значения считаются общими для всех виртуальных пациентов.

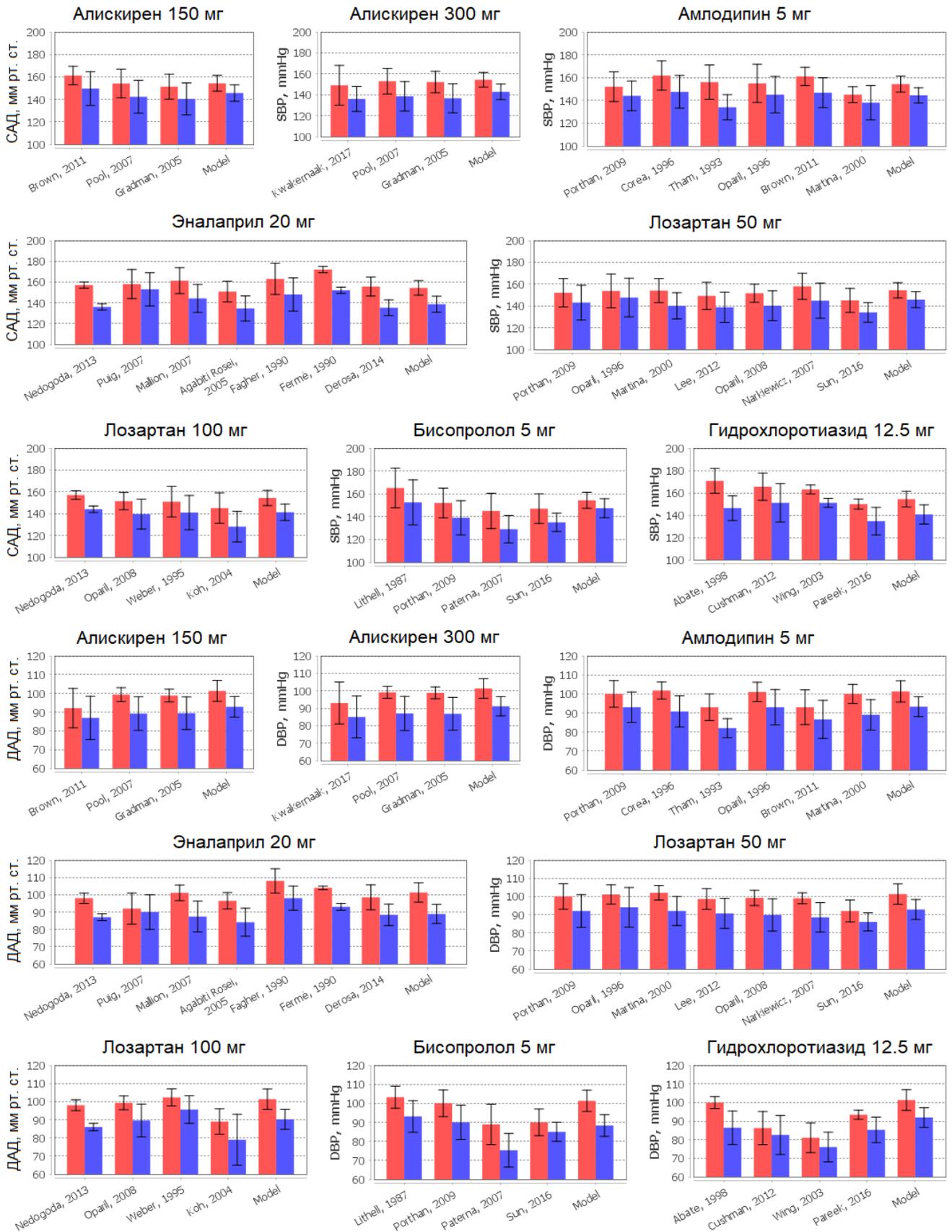


Рисунок 7.5.1 – Средние значения систолического и диастолического давления, полученные в ходе клинических испытаний и численного моделирования в BioUML (значения с подписью “Model”) до лечения АГП (красный) и после лечения (синий)

С целью апробации модели для анализа прогнозируемой эффективности антигипертензивного лечения в реальных условиях были проанализированы истории болезни шести анонимных пациентов с АГ, предоставленные Центром новых медицинских технологий (подразделение Института химической биологии и фундаментальной медицины СО РАН). Анамнез пациентов включал следующие показатели:

- 1) суточный мониторинг артериального давления (АД):
 - среднее дневное и ночное АД,
 - максимальное и минимальное АД за сутки,
 - индексы времени и вариабельность САД и ДАД на протяжении суток (повышены или нет),
 - степень ночного снижения АД («диппер»);
- 2) суточный мониторинг ЭКГ по Холтеру:
 - тип ритма,
 - средняя, минимальная и максимальная частота ритма,
 - наличие отклонений (тахикардия, брадикардия, паузы ритма),
 - урежение ритма в ночное время,
 - желудочковые и наджелудочковые экстрасистолы;
- 3) общий анализ крови (в частности, гематокрит и гемоглобин);
- 4) биохимический анализ крови (в частности, калий и натрий);
- 5) биохимический анализ мочи;
- 6) иммуноферментные и иммунохимические анализы крови (для некоторых пациентов: глобулин-связывающий половой гормон, пролактин, тестостерон, эстрадиол, гомоцистеин, витамин В12, фолиевая кислота, трийодтиронин свободный, тиреотропный гормон, свободный тироксин, иммунореактивный инсулин, антитела к тиреопероксидазе);
- 7) система гемостаза (фибриноген, D-димер, РФМК и т.д.);
- 8) эхокардиография с ЦДК:
 - показатели аорты;
 - показатели левого предсердия;
 - показатели правого предсердия;
 - показатели левого желудочка (размер в состоянии покоя, конечный диастолический размер, конечный систолический размер, фракция удара, фракция выброса, ударный объем);
 - показатели правого желудочка;
 - толщина межжелудочковой перегородки в диастолу;
 - толщина задней стенки левого желудочка в диастолу;

- наличие нарушения локальной сократимости;
 - наличие добавочной хорды;
 - показатели митрального, аортального, трехстворчатого и легочного клапанов сердца (уплотнены или нет створки, максимальный градиент давления, пиковая скорость, степень регургитации).
- 9) УЗИ сосудов шеи (дуплексное сканирование брахиоцефальных артерий);
- 10) дуплексное сканирование с ЦДК артерий нижних конечностей;
- 11) дуплексное сканирование с ЦДК вен нижних конечностей.

Однако при использовании перечисленных клинических данных для персонализации параметров математической модели существуют две проблемы.

Первая проблема – не все из этих данных могут быть использованы, поскольку для многих клинических показателей в модели не определены соответствующие аналоги. Данная проблема может быть решена последовательным расширением и детализацией модели для учета большего количества параметров. На данный момент для персонализации модели мы используем 18 клинических показателей, представленных на рисунке 7.6.1. Часть из них (масса тела, а также такие показатели крови, как гематокрит, гемоглобин, общий белок, глюкоза, мочевины и калий) используются как константы в модели. Другая часть (возраст, рост, пол, раса, креатинин) применяется опосредованно для оценки модельных значений скорости клубочковой фильтрации и общего объема воды организма. Другие показатели (систолическое и диастолическое давление, частота сердечных сокращений, ударный объем, фракция выброса и концентрация натрия в крови) используются для решения обратной задачи, чтобы определить 50 персональных параметров, представленных в таблице 7.1 приложения (заметим, в эту таблицу также включены такие параметры, как гематокрит, гемоглобин, мочевины, глюкоза, общий белок и концентрация калия в крови, значения которых также подвергаются оценке в случае отсутствия информации о них в истории болезни пациента).

Вторая проблема связана с идентифицируемостью скрытых персональных параметров на основе клинических данных. Как было сказано выше, мы производим идентификацию 50 параметров на основе 18 экспериментальных значений. Как следствие такого существенного несоответствия между числом искомых и известных величин мы получим переобучение (overfitting) модели. То есть с математической точки зрения, задача однозначного определения параметров модели виртуального пациента на основе доступных клинических данных является неразрешимой.

Чтобы решить изложенную проблему, нами был предложен новый подход – создавать множество виртуальных пациентов с вариативностью скрытых персональных параметров в заданных физиологических границах, диапазоны которых могут определяться в зависимости от

заболеваний и анамнеза пациента. Данная работа была выполнена Кутумовой Е.О. и представлена в статье (Kutumova et al., 2022a).

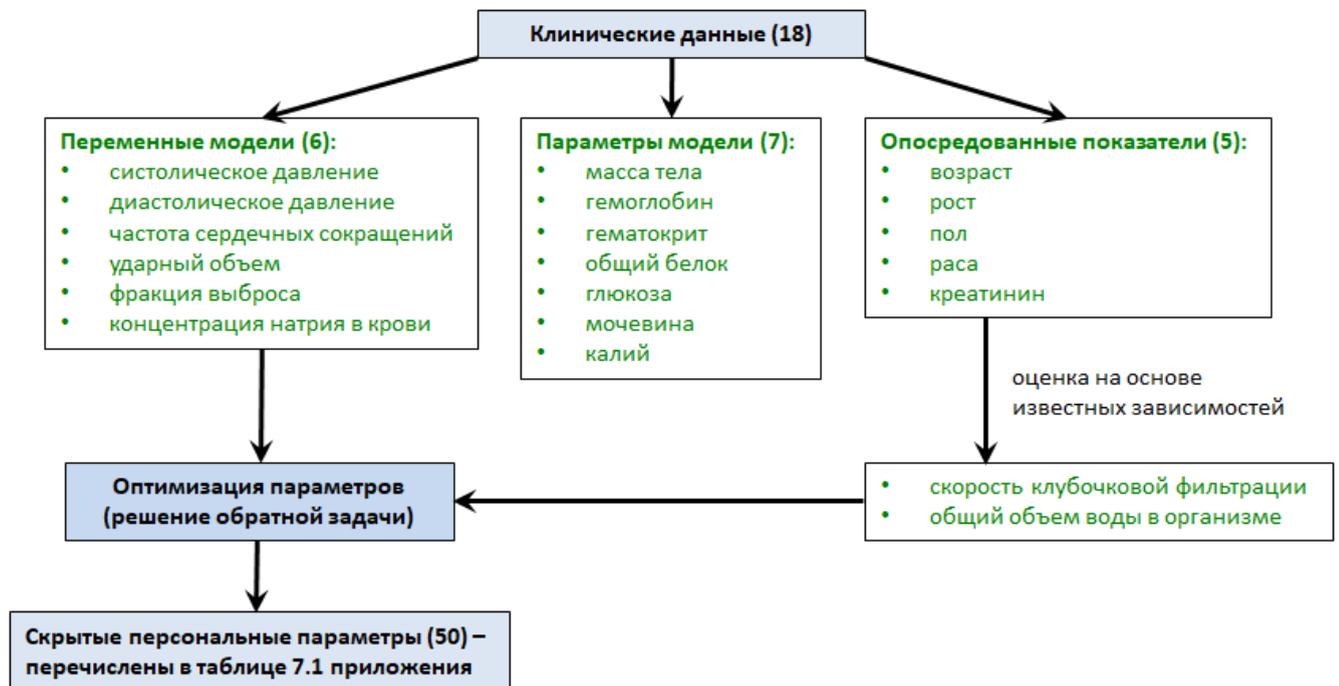


Рисунок 7.6.1 – Используемые клинические данные для персонализации модели виртуального пациента

7.7 Моделирование лечения пациента

Для моделирования лечения пациента в ПК BioUML был разработан ряд методов:

- Population generation – генерирует популяцию виртуальных пациентов для реального пациента с заданными клиническими параметрами (рисунок 7.7.1). Помимо результатов лабораторных исследований и общих данных о пациенте, вводится информация о его заболеваниях с возможностью выбора одной из 3 опций: есть, нет, неизвестно. Эта информация о заболеваниях используется, чтобы задать ограничения на возможные диапазоны значений для многих параметров модели, как это описано в работе (Kutumova et al., 2021).
- Population treatment – моделирует лечение популяции виртуальных пациентов, сгенерированных предыдущим методом, при помощи заданного набора АГП. Результаты прогноза лечения представляются в виде отчета (рисунок 7.7.2).

На основе этой информации может быть сделан вывод о большей эффективности того или иного препарата, а также (после дополнительного статистического анализа) о том какие еще показатели пациента нужно исследовать, чтобы сделать более однозначный вывод об

эффективности препаратов. Например, мы можем разделить популяцию на 2 группы: тех пациентов, кто хорошо отвечает на лечение, и тех, кто не отвечает. Далее можно провести статистический анализ параметров модели для каждой группы и определить какие из них способствуют разделению на эти группы. В частности, такой анализ позволяет оценить, какие еще лабораторные исследования нужно провести для данного пациента, чтобы понять к какой виртуальной группе он относится и какое лечение будет для него наиболее эффективно.

Population generation	
Property	Value
Физиология пациента	
Общая информация	
Возраст, лет	40
Рост, см	170
Вес, кг	70
Пол	мужчина
Раса	европеоидная
Мониторинг АД	
Систолическое давление, мм рт. ст.	140
Диастолическое артериальное давление, мм рт. ст.	90
Электрокардиография	
Частота сердечных сокращений, уд/мин	70
УЗИ сердца	
Ударный Объем, мл	75
Фракция Выброса, %	60
Заболевания	
Артериальная гипертензия	есть
Хроническая сердечная недостаточность	есть
- Функциональный класс (NYHA)	класс II
Легочная гипертензия	нет
Диастолическая дисфункция ЛЖ	есть
Гипертрофия ЛЖ	есть
Хроническая почечная недостаточность	неизвестно
Атеросклероз сонных и/или коронарных артерий	неизвестно
Митральная регургитация	есть
- Стадия митральной регургитации	легкая
Трикуспидальная регургитация	нет
Аортальная регургитация	есть
- Стадия аортальной регургитации	легкая
Легочная регургитация	нет

Run analysis Generate script Expert mode

0%

Рисунок 7.7.1 – Интерфейс в ПК BioUML для генерации популяции для конкретного пациента

Отчет о моделировании лечения антигипертензивными препаратами

Данные пациента

Характеристики	Значения, среднее ± SD (мин - макс)
Возраст, лет	47,0
Рост, см	160,0
Вес, кг	70,0
Пол	женщина
Раса	европеидная
Систолическое давление, мм рт. ст.	152,0
Диастолическое артериальное давление, мм рт. ст.	104,0
Частота сердечных сокращений, уд/мин	70,0
Ударный Объем, мл	41,0
Фракция Выброса, %	70,0
Гемоглобин, г/л	160,0
Гематокрит, %	49,4
Белок общий, г/л	72,6
Глюкоза, ммоль/л	4,62
Мочевина, ммоль/л	4,8
Креатинин, мкмоль/л	97,0
Калий, ммоль/л	4,5
Натрий, ммоль/л	143,0
Артериальная гипертензия	есть
Хроническая сердечная недостаточность	есть
- Функциональный класс (NYHA)	класс I
Легочная гипертензия	нет
Диастолическая дисфункция ЛЖ	нет
Гипертрофия ЛЖ	нет
Хроническая почечная недостаточность	нет
Атеросклероз сонных и/или коронарных артерий	нет
Митральная регургитация	есть
- Стадия митральной регургитации	легкая
Трикуспидальная регургитация	есть
- Стадия трикуспидальной регургитации	легкая
Аортальная регургитация	нет
Легочная регургитация	есть
- Стадия легочной регургитации	легкая

Характеристики популяции

Характеристики	Значения, среднее ± SD (мин - макс)
Вес, кг	70,0
Систолическое давление, мм рт. ст.	152,0 ± 0,2 (151,7 – 153,0)
Диастолическое артериальное давление, мм рт. ст.	104,0 ± 0,1 (103,5 – 104,2)
Среднее артериальное давление, мм рт. ст.	120,0 ± 0,0 (120,0 – 120,0)
Частота сердечных сокращений, уд/мин	69,9 ± 0,2 (68,4 – 70,1)
Ударный Объем, мл	41,3 ± 0,5 (40,9 – 42,9)
Сердечный выброс, л/мин	2,9 ± 0,0 (2,9 – 3,0)
Фракция Выброса, %	70,0 ± 0,1 (69,9 – 70,5)
Конечный диастолический объем левого желудочка, мл	58,9 ± 0,6 (58,5 – 61,3)
Конечный систолический объем левого желудочка, мл	17,7 ± 0,2 (17,6 – 18,4)
Гемоглобин, г/л	160,0
Гематокрит, %	49,4
Белок общий, г/л	72,6
Глюкоза, ммоль/л	4,62
Мочевина, ммоль/л	4,8
Калий, ммоль/л	4,5
Натрий, ммоль/л	143,0 ± 0,0 (143,0 – 143,0)
Скорость клубочковой фильтрации, л/мин (CKD-EPI)	0,060 ± 0,000 (0,060 – 0,060)

Размер популяции: 175

Результаты лечения

Схемы лечения	САД, мм рт. ст.	Снижение САД, мм рт. ст.	ДАД, мм рт. ст.	Снижение ДАД, мм рт. ст.
Алискирен 150 мг	145,9 ± 2,5 (137,5 – 150,5)	6,1 ± 2,5 (1,5 – 14,5)	94,6 ± 1,2 (90,1 – 96,9)	9,3 ± 1,2 (7,1 – 13,9)
Амлодипин 5 мг	141,1 ± 1,5 (137,6 – 144,1)	10,9 ± 1,5 (7,9 – 14,4)	96,4 ± 0,6 (95,2 – 98,5)	7,5 ± 0,6 (5,5 – 8,7)
Бисопролол 5 мг	148,8 ± 4,5 (138,5 – 159,1)	3,2 ± 4,5 (-7,1 – 13,4)	89,8 ± 2,4 (81,4 – 96,2)	14,2 ± 2,4 (7,8 – 22,6)
Гидрохлоротиазид 12,5 мг	133,5 ± 8,3 (110,3 – 143,9)	18,5 ± 8,3 (8,1 – 41,7)	94,8 ± 2,2 (91,0 – 105,8)	9,2 ± 2,2 (-1,8 – 12,8)
Лозартан 50 мг	146,0 ± 2,5 (137,7 – 150,6)	6,0 ± 2,5 (1,4 – 14,2)	94,7 ± 1,2 (90,2 – 96,9)	9,3 ± 1,2 (7,1 – 13,8)
Эналаприл 20 мг	139,4 ± 3,8 (115,7 – 144,5)	12,6 ± 3,7 (7,6 – 36,2)	90,9 ± 1,5 (85,9 – 99,2)	13,1 ± 1,4 (4,8 – 18,0)

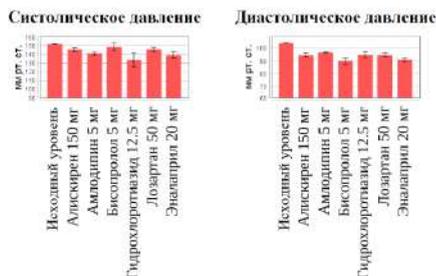


Рисунок 7.7.2 – Пример отчета по прогнозу лечения АГ основными классами АГП в ПК BioUML

7.8 Валидация прогноза результата лечения артериальной гипертензии

С помощью предложенного подхода было проведено моделирование лечения АГ для 6 реальных пациентов клиники ЦНМТ. В качестве примера приведены результаты для одного из них.

Данные пациента

- Рост 160 см, вес 70 кг
- Возраст 47 лет
- Пол женский
- Среднее дневное артериальное давление 152 / 100 мм. рт. ст.
- Средняя частота сердечных сокращений 70 уд/мин.
- Ударный объем 41 мл
- Фракция выброса 70%
- Глюкоза 4.62 ммоль/л (в норме)
- Гемоглобин 160 г/л
- Гематокрит 49.4 %
- Калий 4.5 ммоль/л
- Натрий 143 ммоль/л
- Общий белок 72.6 г/л
- Мочевина 4.8 ммоль/л
- Креатинин 97 мкмоль/л

Заболевания: гипертоническая болезнь 1 стадии, артериальная гипертензия 1 степени, риск 2 (умеренный). Дислипотеидемия. Хроническая сердечная недостаточность 0-1, функциональный класс 1 (NYHA).

Среднее дневное артериальное давление после лечения препаратом Лозап (Лозартан) 50 мг: 132/94.

Результаты предсказания результатов лечения приведены в отчете на рисунке 7.7.2 Как видно, в реальности лозартан более существенно снижал систолическое давление пациента, чем дает среднестатистическая оценка, показанная моделью в соответствии с клиническими исследованиями препарата. Результаты реального лечения были сравнимы с наилучшими результатами, показанными моделью в популяции виртуальных пациентов. Это говорит о том, что именно эти виртуальные пациенты могут быть использованы для тестирования и анализа действия различных препаратов и схем лечения в отношении исследуемого пациента.

По среднестатистическим результатам моделирования можно заключить, что из шести препаратов лучший результат по снижению давления дает тиазидный диуретик гидрохлоротиазид. Это может быть связано с тем, что сердечная недостаточность, наблюдаемая у пациента, как правило сопровождается отеками и, как следствие, повышенным объемом воды в организме, с которым справляются диуретики, тем самым провоцируя снижение артериального давления. По предсказаниям модели наилучший результат лечения показал

бисопролол. В то время как у лозартана и алискирена (оба препарата ингибируют ренин-ангиотензин-альдостероновую систему) результаты оказались сопоставимы.

7.9 Обсуждение

Разработанная технология построения цифрового двойника пациента и оптимизации выбора лекарственной терапии на примере лечения АГ является прототипом, доказывающим возможность решения такой задачи вычислительным путем и работоспособность предложенного подхода, а не готовым решением для внедрения в медицинскую практику. Для этого нужно пройти еще большой путь, включая:

- существенное расширение модели регуляции артериального давления у человека в норме и при АГ с учетом:
 - физиологических параметров, связанных с сосудистым и общим старением человека;
 - молекулярных механизмов, позволяющих учесть данные генотипирования, связанные с риском возникновения АГ;
 - молекулярных механизмов, связанных с фармакогенетикой (в первую очередь, системы цитохромов);
 - основных особенностей образа жизни человека, связанных с риском возникновения АГ (в первую очередь, индекса массы тела, курения).
- расширение сценариев лечения, включая тестирование комплексной терапии несколькими АГП;
- представление существующих методических рекомендаций по лечению гипертонии в виде цифровой регламентной модели, включающей в себя в том числе и набор правил для первичного ограничения возможных вариантов лекарственной терапии;
- усовершенствование алгоритма персонализации параметров модели по клиническим, биохимическим, патофизиологическим и генетическим данным пациента. Алгоритм персонализации должен уметь работать в ситуациях неполного набора данных о пациенте;
- исследование полноты и репрезентативности генерируемой популяции виртуальных пациентов;
- оптимизацию скорости генерации популяции виртуальных пациентов для заданного пациента и/или генерацию большой базовой популяции (“на все случаи жизни”), а затем выбор из нее тех виртуальных пациентов, характеристики которых соответствуют клиническим данным этого пациента;
- интеграцию методов персонализированного моделирования, логико-вероятностных методов, методов машинного обучения и регламентного моделирования, состоящего из

наборов правил, в единую информационную систему гибридного искусственного интеллекта;

- сертификацию и валидацию созданной системы на основе большого количества клинических данных.

Также важно подчеркнуть, что создание цифрового двойника пациента не является разовым действием. В идеале, цифровой двойник пациента должен создаваться с момента нахождения его в утробе матери на основе его генетических данных и истории беременности и сопровождать его в течение всей жизни, учитывая образ жизни пациента, его историю болезни и взаимодействия с системой здравоохранения. Такой подход требует развития соответствующей технологии и программного комплекса, который бы модифицировал цифрового двойника пациента в соответствии с жизненными событиями, развитием и старением реального пациента. В качестве первого приближения на примере АГ при построении цифрового двойника пациента могут использоваться данные генетического тестирования, а для ведения ("жизни") цифрового двойника – как ретроспективные данные из истории болезни с учетом образа жизни, так и данные, полученные в ходе клинических исследований.

В настоящее время на методы искусственного интеллекта в медицине возлагаются большие надежды. По сравнению с персонализированным моделированием и экспертными системами, основанными на правилах, основным преимуществом методов машинного обучения является гораздо более быстрое построение предсказательных моделей, которое не требует от их создателей экспертного знания предметной области. В ряде медицинских областей, например, в области распознавания изображений (в частности, рентгенография) методы ИИ уже превзошли людей.

Активно развиваются системы помощи принятия врачебного решения на основе искусственного интеллекта и в России. Наиболее известные в этом направлении компании (проекты):

- WEBIOMED (<https://webiomed.ai/>) – платформа предиктивной аналитики и управления рисками в здравоохранении на основе машинного обучения, которая позволяет выявлять факторы риска заболеваний и прогнозирует вероятность осложнений или даже смерти пациента, а также формирует индивидуальные рекомендации по предотвращению заболевания. Платформа также обеспечивает интеграцию со многими МИС (Медицинскими Информационными Системами) и автоматический анализ медицинских данных, включая, извлечение информации из неструктурированных врачебных записей с помощью NLP (Natural Language Processing) технологий. В список поддерживаемых заболеваний входят и сердечно-сосудистые заболевания.

- Botkin.AI (<https://botkin.ai/>) – платформа для анализа и обработки медицинских изображений с технологиями искусственного интеллекта. Платформа позволяет:
 - по КТ – определение злокачественных новообразований в легких;
 - распознавание рентгенограмм органов грудной клетки;
 - диагностика рака молочной железы по данным маммографии;
 - диагностика туберкулеза по данным рентгенографии.
- SberMedAI (ранее CoBrain-analytics, <https://cobrain.ai/ru>, <https://sbermed.ai/>) – единая витрина для медицинских приложений на базе искусственного интеллекта.

Однако уже сейчас ряд надежд, связанных с ИИ, не оправдался. Наиболее крупным таким разочарованием может служить проект компании IBM Watson Health (Frownfelter, 2021). Ниже переведенный фрагмент из (Vespoli, 2021):

“ В 2015 году, всего через два года и 62 миллиона долларов затрат после начала их партнерства, онкологический центр доктора медицины Андерсона прекратил сотрудничество с IBM Watson. Проект, над которым они работали, «Советник по онкологии», был задуман как источник рекомендаций по лечению. Организации рекламировали способность платформы использовать обработку естественного языка для чтения медицинских карт пациентов и поиска в БД исследований рака.

К 2018 году более десятка партнеров и клиентов IBM остановили или свернули свои онкологические проекты с помощью Watson. Ранее в этом году The Wall Street Journal сообщила, что IBM изучает возможность продажи Watson Health, которая, несмотря на то что принесла доход почти в 1 миллиард долларов, была нерентабельной”.

Текущий взлет ИИ – это не первый взлет в его истории, что можно сказать и про периоды спада, для которых даже существует термин – “зима ИИ” (https://ru.wikipedia.org/wiki/Зима_искусственного_интеллекта).

В качестве метафоры мы можем рассматривать персонализированное моделирование как “белый ящик”, где мы знаем, что и как работает и почему система принимает то или иное решение. Методы искусственного интеллекта, наоборот, часто работают как “черный ящик”, где непонятно, какие зависимости и почему используются для принятия решения.

На наш взгляд, наиболее адекватным подходом, совмещающим преимущества трех подходов: персонализированного моделирования, наборов правил (экспертных систем) и методов ИИ – являются гибридные интеллектуальные системы. Данное направление представляется перспективным для дальнейшего развития предложенного подхода.

ГЛАВА 8. ДРУГИЕ РАЗРАБОТКИ НА ОСНОВЕ ПК BioUML

Модульная архитектура ПК BioUML позволила создать на его основе ряд различных программ для моделирования, анализа и графического представления данных, не только в области биологии и медицины, но и в других предметных областях (например, физика высоких энергий). В результате на основе ПК BioUML сейчас разрабатывается универсальная научная платформа u-science.

8.1 Биоинформатическая платформа для одномолекулярного секвенирования ДНК

Одномолекулярное секвенирование в реальном времени (single molecule real time sequencing – SMRT) – метод секвенирования ДНК нового поколения, разработанный компанией Pacific Biosciences. Идея метода состоит в определении последовательности ДНК путем наблюдения за работой единичной молекулы ДНК-полимеразы в реальном времени. При этом ДНК-полимераза достраивает вторую цепь исследуемой молекулы ДНК, используя нуклеотиды, меченные различными флуоресцентными метками; регистрируя данные метки, можно понять, какой нуклеотид ДНК-полимераза встраивает в настоящий момент (Wenger et al., 2019). Длина прочтения при таком секвенировании достигает 20 000 пар нуклеотидов и больше.

В рамках проекта «Разработка опытного образца одномолекулярного секвенатора ДНК, набора реагентов и расходных материалов для него» (Соглашение № 075-15-2021-1057 от 28 сентября 2021 г., заключенное между Минобрнауки России и Институтом аналитического приборостроения РАН в рамках реализации отдельных мероприятий Федеральной научно-технической программы развития генетических технологий на 2019 – 2027 годы) разрабатывается отечественный секвенатор для одномолекулярного секвенирования ДНК «Нанофор-ОМС». Для обеспечения решения «под ключ» к нему нами разрабатывается биоинформатическая платформа для хранения, анализа и графического представления полученных с него данных на основе отечественного ПК BioUML, что особенно важно для обеспечения технологического суверенитета Российской Федерации в этой области.

Именно в ходе этого проекта были существенно расширены возможности ПК BioUML:

- выполнение сценариев анализа данных на языке WDL (см. подраздел 4.2.1)
 - загрузка/импорт сценариев на языке WDL;
 - просмотр и редактирование сценариев на языке WDL, включая подсветку синтаксиса, как документа Платформы;
 - просмотр и генерация графа сценария анализа данных, используя иерархический алгоритм для расположения вершин графа;

- автоматическая генерация форм для ввода параметров для методов и сценариев анализа данных одномолекулярного секвенирования;
- использование системы управления сценариями Cromwell (<https://cromwell.readthedocs.io/en/stable/>) как отдельного сервиса;
- возможность использования отдельных программ и сценариев, описанных на языке WDL, внутри сценариев платформы BioUML.
- выполнение сценариев анализа данных на языке/платформе Nextflow (см. подраздел 4.2.2);
- интеграция с новой версии Galaxy и Galaxy Toolshed (см. подраздел 4.2.3);
- разработка новой очереди задач внутри платформы. Оказалось, что для составления сложных сценариев анализа данных одномолекулярного секвенирования в одном сценарии могут использоваться инструменты из разных платформ, например, одни программы используются из платформы Galaxy, другие для WDL/Cromwell. При выполнении такого гетерогенного сценария очередь задач ПК BioUML должна уметь распределять выполнение задач между соответствующими сервисами;
- улучшение интеграции с Jupyter notebook/hub (см. подраздел 2.3.9):
 - поддержка доступа к репозиторию и методам анализа данных ПК BioUML из скриптов Python 3, выполняемых в Jupyter блокнотах;
 - возможность использования нескольких языков в одном скрипте;
- разработка "облачной" версии платформы, используя среду Kubernetes;
- нагрузочное тестирование – тестирование производительности платформы при большом числе пользователей одновременно. Для этого были написаны несколько сценариев имитации нагрузки пользователей, используя JMeter (<https://jmeter.apache.org/>).
- разработка комплекта документации.

В ходе работы были рассмотрены основные сценарии анализа геномных данных на платформе SMRT Link полученных при одномолекулярном секвенировании ДНК при помощи Sequel II System компании PacBio, который является ближайшим зарубежным аналогом для разрабатываемого отечественного секвенатора. Как результат, были реализованы следующие сценарии анализа данных одномолекулярного секвенирования:

- сборка геномов de-novo (см. приложение 8.1 как пример такого сценария);
- определение полиморфизмов, в том числе структурных;
- анализ результатов секвенирования РНК;
- сценарий для анализа результатов таргетного секвенирования;
- анализ результатов эпигенетическое секвенирование микробов;
- анализ результатов эпигенетическое секвенирование эукариот.

The screenshot displays a web-based bioinformatics platform interface. On the left, a file tree under 'data' shows a folder 'Sequencer examples' containing 'De Novo Genome Assembly', which is further divided into 'Doc', 'Input', 'Results', 'wdfs', and 'De_novo_genome_assembly_workflow.wdl'. The main content area features a title 'Анализ данных мономолекулярного секвенирования' and a list of six scenarios. At the bottom left, a workflow diagram shows a sequence of steps: 'reads_1' (green box) leads to 'indexes' (yellow box), which leads to a circular node 'i', and finally to a green box labeled 'r'. The bottom right area contains a search bar and a description field that is currently empty.

Start page
← [Рабочие окружения](#)

Анализ данных мономолекулярного секвенирования

- [Сценарий "Сборка геномов de-novo"](#)
- [Сценарий "Секвенирование целевых локусов"](#)
Реконструкция и анализ целевых локусов на основе данных таргетного мономолекулярного секвенирования
- [Сценарий "Определение однонуклеотидных полиморфизмов и структурных вариаций"](#)
Идентификация однонуклеотидных полиморфизмов и структурных вариаций (≥ 20 п. н.) на основе данных мономолекулярного секвенирования
- [Сценарий "Эпигеномное секвенирование микробов"](#)
Определение и характеристика мотивов ДНК, распознаваемых метилтрансферазой при модификациях нуклеотидов (бактериальная эпигенетика)
- [Сценарий "Секвенирование РНК"](#)
Реконструкция последовательностей РНК и функциональная характеристика полноразмерных изоформ транскриптов на основе данных мономолекулярного секвенирования РНК
- [Сценарий "Эпигеномное секвенирование эукариот"](#)
Идентификация гипер- и гипометилированных CpG-островков для исследования генной экспрессии и регуляции на основе данных мономолекулярного секвенирования

Search Info Default

ID: De_novo_genome_assembly_workflow.wdl

User description is not available

My description Graph search Script Clipboard Tasks

Рисунок 8.1.1 – Стартовая страница биоинформатической платформы для анализа данных одномолекулярного секвенирования.

8.2 Подсистемы для прототипа «Национальной базы генетической информации»

На основе ПК BioUML в ходе реализации работ по контракту № 43-03/21/44/152 от 13.12.2021 "Исследование, обоснование и выбор программных решений для визуализации генетических данных и обеспечения инструментов для работы с генетической информацией в «Национальной базе генетической информации» нами были успешно созданы и приняты комиссией заказчика в лице Федерального государственного бюджетного учреждения «Национальный исследовательский центр «Курчатовский институт» две подсистемы: "Геномный браузер" и "Рабочие окружения". Основные фрагменты технического задания на разработку этих подсистем приведены в приложении 8.2. Полный текст конкурсной документации доступен на официальном сайте Единой информационной системы в сфере закупок:

<https://zakupki.gov.ru/epz/order/notice/ea44/view/documents.html?regNumber=0373100001021000793>

В соответствии с требованиями технического задания подсистема "Геномный браузер" обеспечивает визуализацию больших геномов растений. Например, на рисунке п.8.2.1 приведена визуализация самого большого секвенированного генома из БД NCBI – *Protopterus annectens* и его аннотации (40+ млрд. пар нуклеотидов). Также разработанная подсистема поддерживает визуализацию трехмерных структур молекул (рисунок п.8.2.2), используя библиотеку JsMol.

8.3 Платформа geneXplain

Платформа geneXplain (<https://genexplain.com/genexplain-platform>) – это коммерческий вариант ПК BioUML, разрабатываемый компаниями geneXplain GmbH (Вольфенбютель, Германия) и ООО "Биософт.Ру" для анализа омиксных данных (Kel et al., 2011). В ее основе лежат 3 идеи:

- 1) Создать большой набор готовых сценариев для анализа различных типов омиксных данных, которые позволят пользователю, не являющемуся экспертом в области биоинформатики, проанализировать его данные. Другими словами, платформа geneXplain подобна книге рецептов, где пользователь просто должен выбрать подходящий ему «рецепт», указать входные данные и на выходе получит готовый отчет и набор файлов с результатами анализа.
- 2) Использовать коммерчески-доступные БД для глубокого анализа омиксных данных. Для этого используются следующие БД:
 - TRANSFAC® (<https://genexplain.com/transfac>) – БД ТФ и их сайтов связывания, также содержит большую коллекцию весовых матриц для определения ССТФ (Wingender, 2008, Wingender et al., 2018).

- TRANSPATH® (<https://genexplain.com/transpath>) – БД метаболических путей и путей передачи сигнала в клетке у млекопитающих (Krull et al., 2006).

3) Разработать собственные методы глубокого анализа омических данных с использованием БД TRANSFAC® и TRANSPATH®.

Краткое описание веб-интерфейса и ключевых собственных методов глубокого анализа омических данных платформы платформы geneXplain приведено в приложении 8.3.

8.4 Платформа Genome Enhancer

Платформа Genome Enhancer (<https://genexplain.com/genexplain-platform>) – это специализированный вариант ПК BioUML для анализа омических данных в автоматическом режиме (Kel et al., 2011), изначально разработанный ООО "Биософт.Ру" с участием автора диссертационной работы при поддержке гранта Фонда "Сколково". Сейчас она поддерживается и распространяется компанией geneXplain GmbH (Вольфенбютель, Германия).

Основные идеи Genome Enhancer, предложенные А.Э. Келем:

- от генома к мишени – по загруженным пользователем омическими данным платформа автоматически определяет:
 - гены, изменившие экспрессию;
 - соответствующие им мастер-регуляторы (см. выше);
 - возможные лекарственные препараты или химические вещества, которые могли бы активировать или ингибировать мастер-регуляторы, чтобы вернуть экспрессию генов к нормальному состоянию;
- максимально упростить интерфейс пользователя для анализа данных;
- максимально автоматизировать анализ омических данных – в зависимости от загруженных пользователем данных платформа автоматически составляет нужный сценарий анализа этих данных;
- представить результаты анализа данных в виде отчета, сходного по содержанию и оформлению с готовой научной статьей.

Общий сценарий автоматического анализа мультиомических данных в платформе Genome Enhancer представлен на рисунке 8.4.1. Он состоит из 4-х основных этапов:

- 1) Построение и анализ списка дифференциально экспрессирующихся генов для соответствующего патологического процесса. Для полученного списка генов строятся тепловые карты, а также проводится анализ для выявления категорий GO, наиболее значимых для этого списка генов. В зависимости от набора представленных омических данных могут выполняться различные шаги.

- 2) Анализ регуляторных районов генов (промоторы и энхансеры) для выявления транскрипционных факторов, чьи сайты связывания наиболее часто встречаются в дифференциально экспрессируемых генах. Для этого используется информация из БД TRANSFAC® и алгоритмы Match and Composite Module Analyst, описанные в разделе приложения 8.3.
- 3) Поиск мастер-регуляторов – для этого используется информация из БД TRANSPATH® и алгоритм поиска мастер-регуляторов, описанный в приложении 8.3.
- 4) Поиск потенциальных лекарственных препаратов или химических веществ, которые могли бы активировать или ингибировать мастер-регуляторы, чтобы вернуть экспрессию генов к нормальному состоянию. Для этого используется информация из БД HumanPSD™ и результаты предсказания компьютерной программы PASS (Стельмашенко и др., 2019).

По результатам работы каждого этапа анализа формируется соответствующий раздел отчета и входящие в него таблицы и графики. Более подробно алгоритм работы описан на веб-странице платформы Genome Enhancer: <http://my-genome-enhancer.com/analysisDescription.php>.

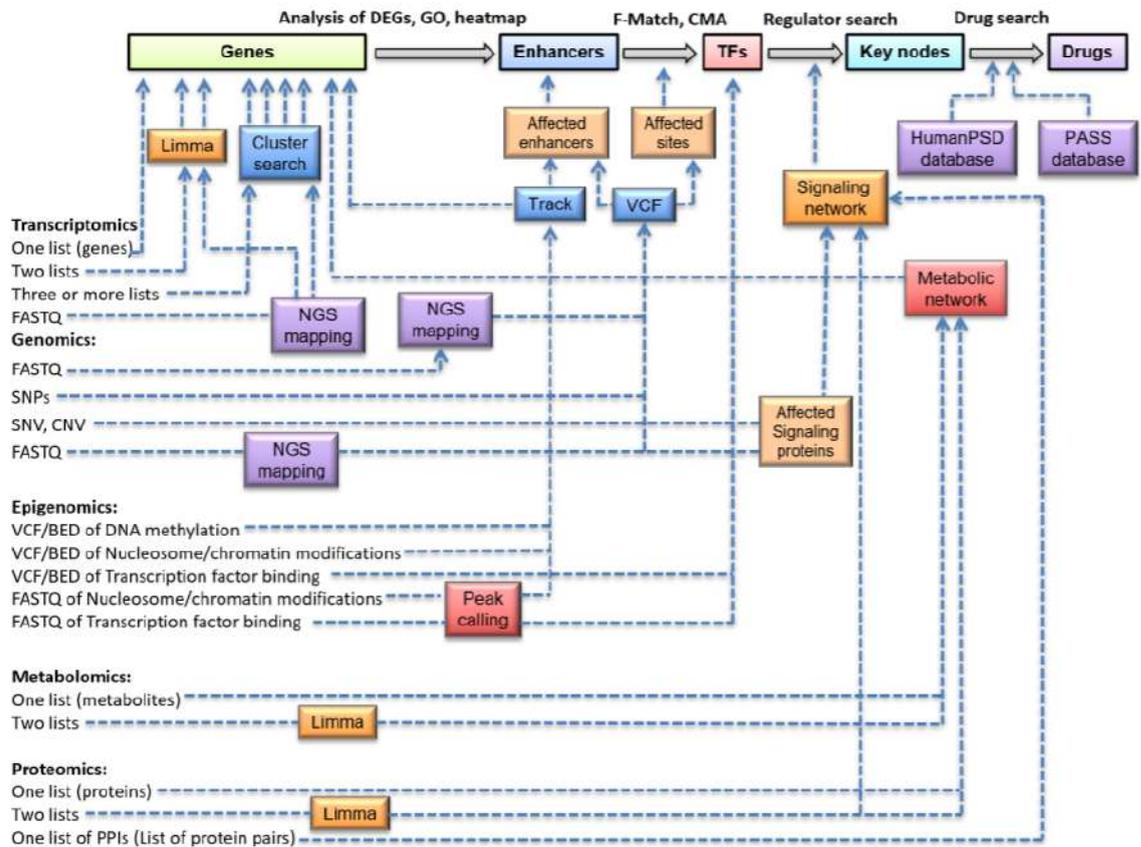


Рисунок 8.4.1 – Общий сценарий автоматического анализа мультиомиксных данных в платформе Genome Enhancer

8.5 Платформа Sirius-web

*Что пользы, если Моцарт будет жив
И новой высоты ещё достигнет?
Подымет ли он тем искусство? Нет;
Оно падёт опять, как он исчезнет.*

А.С. Пушкин, "Моцарт и Сальери"

Sirius-web – это информационная платформа для проектов, связанных с анализом данных и моделированием для образовательных и научных проектов. Постановка задачи, идея и архитектура платформы были сформулированы автором диссертационной работы. Сейчас эта платформа активно используется коллективом под руководством автора диссертационной работы в научно-технологическом университете "Сириус" для:

- ведения курсов "Программирование на Python", "Системная биология";
- выполнения студентами дипломных проектов;
- выполнения научных исследований, связанных с моделированием сложных биологических систем.

Однако, задачи, поставленные при ее разработке, делают ее полезной для использования в других научных и научно-образовательных организациях, связанных с информационными технологиями.

8.5.1 Постановка задачи

Научные образовательные организации (университеты) по определению являются проточными системами – каждый год одни магистранты и аспиранты приходят, а другие уходят.

Вопрос – что остается в университете после того как они уйдут? Опыт современной университетской науки показывает, что, как правило, это бумажные документы: написанные статьи, дипломные и кандидатские работы, а также патенты и свидетельства на регистрацию программ для ЭВМ и баз данных.

Если это были программы, алгоритмы, анализ данных, то часто новое поколение исследователей начинает работу заново, потому что без автора это все не работает. Каждое поколение магистрантов и аспирантов изобретает свой "велосипед". Более того, в соседних лабораториях или кафедрах часто разрабатывают вариации одного и того же "велосипеда".

Опыт успешных компаний и открытых проектов показывает, что можно так организовать проточную систему (люди приходят и уходят), что каждое поколение оставляет свой вклад, а новое поколение начинает с того места, где закончило предыдущее. В основе этого лежат:

- некоторая информационная система для хранения, поиска, документирования и изменения артефактов (техническая часть);
- регламент, определяющий жизненный цикл артефактов (бюрократическая часть).

Интеллектуальный капитал делится обычно на три категории:

- личный капитал — это то, что находится в голове сотрудника;
- капитал клиентов – относится к отношениям с клиентами;
- капитал структуры — это то знание, которое остается внутри организации и может передаваться новым работникам.

Эта самая третья категория и является ключом к управлению данными. По словам Томаса А. Стюарта, много написавшего по этому вопросу: «Капитал структуры — это знание, которое вечером не возвращается домой». Он включает в себя все — процессы, системы и установки, представляющие накопление опыта организации за все время ее существования.

Управление данными прежде всего касается попыток превратить остальные два вида интеллектуального капитала в капитал структуры. Идея состоит в том, что преимущество перед конкурентами можно черпать из ноу-хау, которое может быть получено, каталогизировано и сделано доступным для каждого сотрудника организации. Управление данными знаний является одной из основ успеха компании Microsoft (Дерлав, 2003).

8.5.2 Накапливаемые данные

Рассмотрим, что можно накапливать в такой информационной системе.

1. Учебные курсы, связанные с информационными технологиями и анализом данных.

Многие университеты имеют свои информационные системы, где хранятся учебные курсы, представленные, как правило, презентациями и документами. В данном подходе, с учетом того, что это курсы по информационным технологиям и анализу данных, мы предлагаем хранить данные курсы в виде набора Jupyter ноутбуков. Это позволяет использовать такие ноутбуки как для интерактивных занятий, так и для домашних заданий студентам. Более того, результаты выполнения домашних заданий могут быть сохранены в виде отдельного репозитория для каждого студента. Это облегчает как проверку домашних заданий, так и позволяет сохранить студенту результаты работы, которые могут им быть использованы в дальнейшем (например, быстро вспомнить пройденную тему).

2. Дипломные проекты, связанные с информационными технологиями и анализом данных. Результаты дипломного проекта должны быть представлены в виде репозитория с разработанными алгоритмами, программами анализа данных, полученными результатами, документацией и т.п. Для этих целей необходимо разработать правила (методические рекомендации), чтобы результаты работы можно было воспроизвести, проверить и использовать в дальнейшем.

3. Научные проекты по анализу данных, осуществляемые в ходе научно-исследовательской работы организации. Сейчас в науке остро стоит проблема воспроизведения результатов научных исследований. Во многих статьях приводится ссылка на

репозиторий на GitHub, где хранятся разработанные программы, методы и т.п. Для этих целей также необходимо некоторые правила (методические рекомендации), чтобы результаты научной работы можно было воспроизвести, проверить и использовать в дальнейшем.

4. Сценарии анализа данных – формализованные сценарии анализа данных (CWL, Galaxy, NextFlow и т.п.) вместе с их описанием и примерами. Например, опыт работы биоинформатиков может быть суммирован для биологов в сценариях обработки их данных, что позволит формализовать и упростить для экспериментаторов, получаемых ими данных.

5. Математические модели – для каждой публикации может создаваться свой репозиторий, где содержится код модели, исходные настройки и данные для их работы, результаты работы моделей и визуализация результатов в виде графиков. Документация по моделям. Модели могут быть разные, а для работы с ними могут использоваться Jupyter notebooks.

6. Экспериментальные данные – например, экспериментальные биологические данные (генетические данные, аграрные, сельхоз. и т.п.) и результаты их анализа. Важно с самого начала разработать регламенты для хранения и аннотации таких экспериментальных данных, в зависимости от их типов. При этом нужно отталкиваться от существующих стандартов для каждого типа данных.

8.5.3 Архитектура

Архитектура платформы Sirius-web представлена на рисунке 8.5.1 Она состоит из следующих основных компонентов:

- 1) платформа GitLab – используется для:
 - накопления и структурирования научно-образовательных результатов в виде git репозитория;
 - управления ходом исследований, следуя современным Agile технологиям. Для этого платформа GitLab представляет возможность разбиения проекта на этапы. В каждом этапе формируется набор задач, чей статус может быть отражен на специальной канбан доске;
 - документирования этих результатов, используя встроенную систему wiki.
- 2) ПК BioUML – используется для построения моделей, их анализа, численного моделирования их динамики, а также для графического представления и анализа биомедицинских данных, как было описано выше.

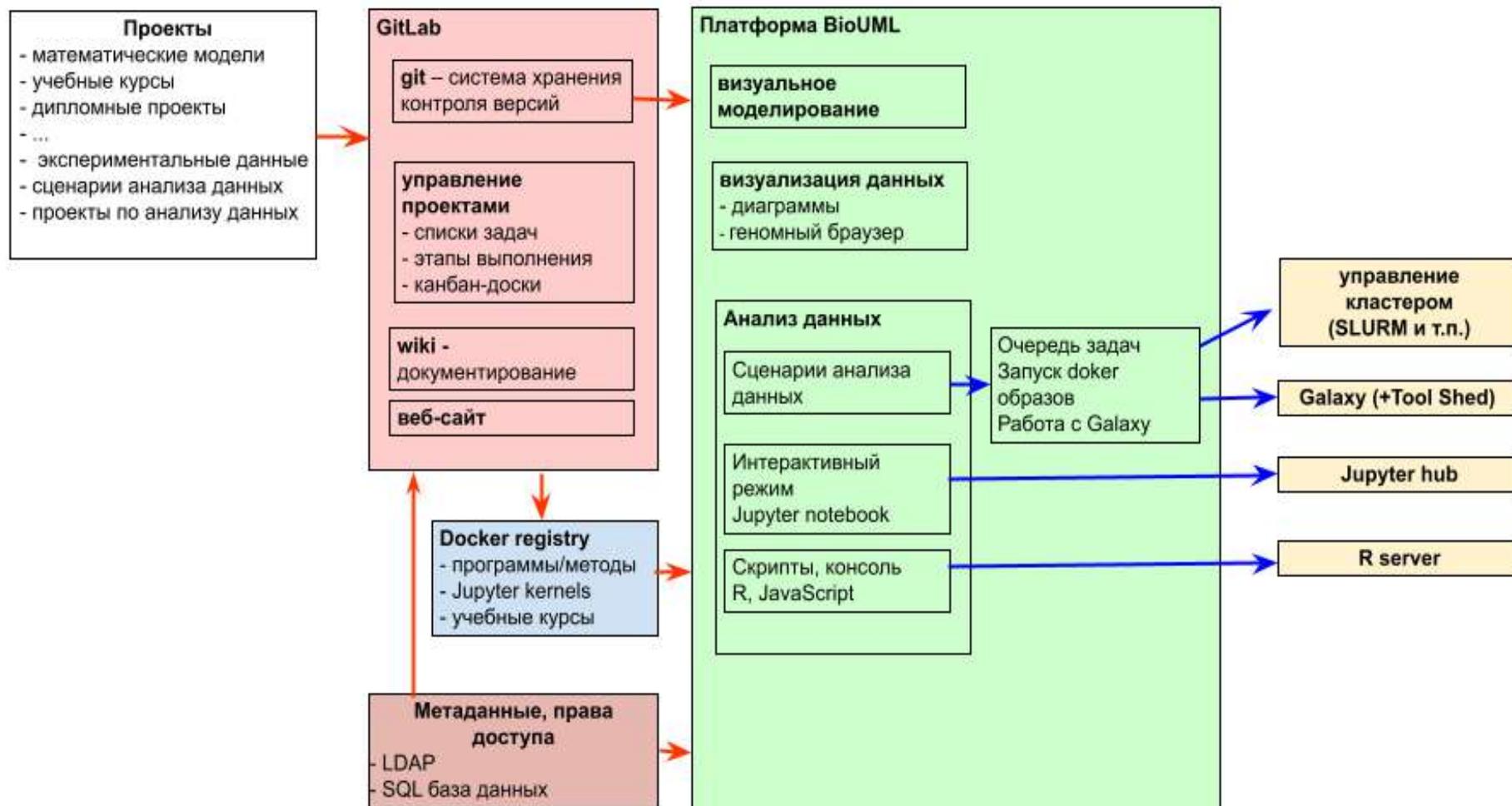


Рисунок 8.5.1 – Архитектура платформы Sirius-web

- 3) Мета-информация – для удобства поиска и структурирования данных, мета-информация о проектах хранится в специальной реляционной БД. С помощью платформы BeanExplorer Enterprise Edition разработан веб-интерфейс для поиска и просмотра информации в этой БД.
- 4) Реестр docker образов – в нем хранятся docker образы для необходимых программ, а также ядра (kernels) для Jupyter ноутбуков. Такой механизм позволяет легко добавлять новые функции в платформу Sirius web.

8.5.4 Использование

Примеры практического использования платформы Sirius-web приведены в приложении 8.5. Она также используется для хранения, редактирования, анализа и документирования математических моделей.

Кроме того мы перенесли в платформу Sirius-web основные модели (модульные модели апоптоза, работы скелетных мышц, регуляции артериального давления, эпидемиологии COVID-19), описанные в диссертации.

При этом использовались следующие правила:

- для каждой публикации создавался отдельный Git репозиторий;
- для воспроизведения всех графиков и результатов, приведенных в статье, в репозитории содержатся соответствующие Jupyter ноутбуки;
- на стартовой странице (README) приведена ссылка на соответствующую публикацию, что служит дополнительной документацией по модели;
- на стартовой странице приведена ссылка для открытия соответствующего проекта в платформе BioUML, а также могут быть приведены дополнительные ссылки, чтобы открыть диаграммы моделей и Jupyter ноутбуки;
- для модификации и развития модели для следующей публикации создается новый клон исходного репозитория, в котором уже идет дальнейшая работа.

8.6 Платформа u-science

Несмотря на существенную разницу в данных, программах и методах для их анализа и визуализации в разных научных областях (биология, ядерная физика, обработка спутниковых данных и т.п.) возможно создание универсальной научной платформы, которая обеспечит:

- следование принципам FAIR – Findable, Accessible, Interoperable, and Re-usable (Wilkinson et al., 2016);

- воспроизводимость результатов исследований (reproducible research);
- совместную работу над проектами (collaborative research);
- удобный пользовательский веб-интерфейс для поиска и анализа данных, а также их графического представления.

В основе такого подхода лежат 3 идеи:

- 1) мы не собираемся заменять или создавать заново отличные инструменты, которые уже созданы в каждой предметной области, но хотим поместить их в единый контекст, где бы они могли быть легко найдены и использованы;
- 2) формализовать, единообразно описать и структурировать существующие в каждой предметной области данные, программы и методы для их анализа и визуализации;
- 3) существует ряд рутинных IT задач, которые не специфичны для предметной области, но требуют много усилий для их решения:
 - управление правами доступа к данным и программам;
 - управление доступом к IT ресурсам, управление очередью задач;
 - обеспечение совместной работы над анализом данных;
 - веб-интерфейс для доступа к данным и программам;
 - система контроля версий данных и программ.

Эти задачи могут быть эффективно решены в рамках универсальной научной платформы u-science, которая построена на основе ПК BioUML.

На рисунке 8.6.1 представлена архитектура платформы u-science. Как видно, она фактически совпадает с последней версией архитектуры ПК BioUML (Kolpakov et al., 2022). Основное отличие платформы u-science – из нее удалены модули, специфичные для биологических задач – моделированием биологических систем и методы анализа биологических данных.

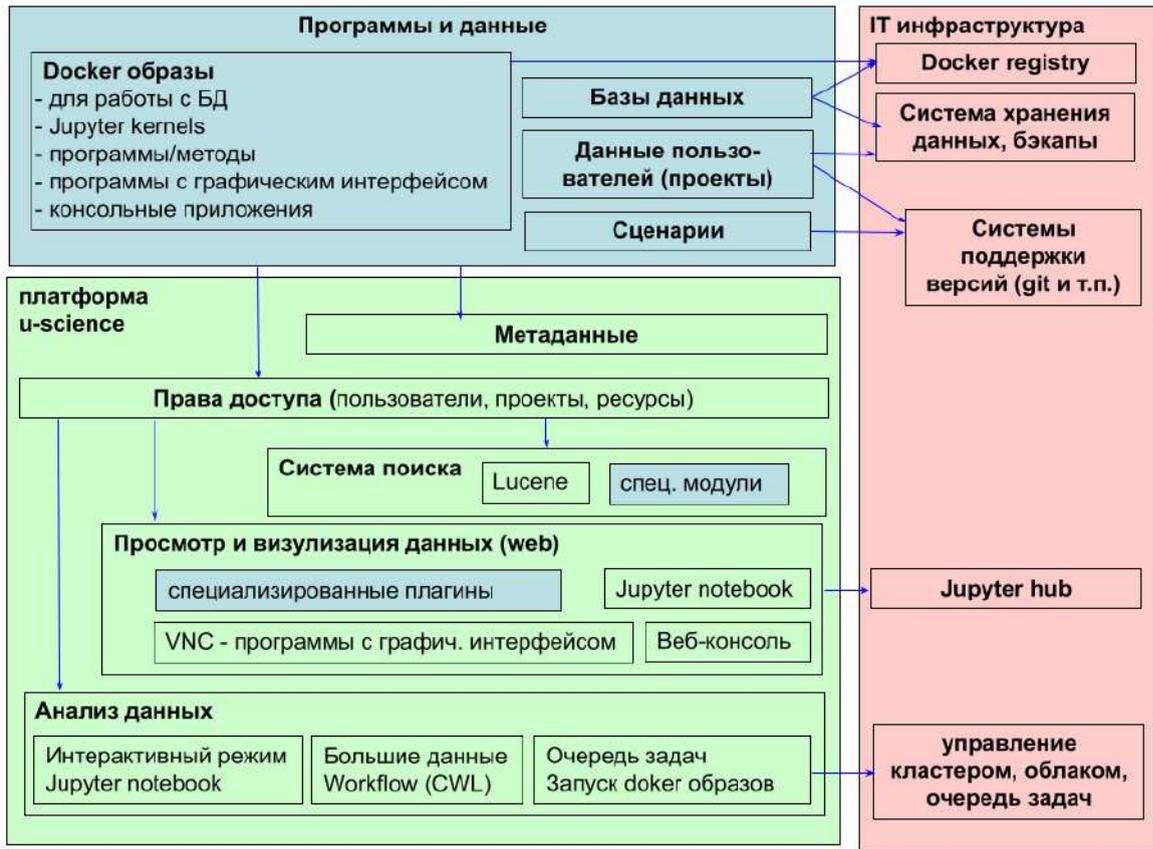


Рисунок 8.6.1 – Архитектура платформы u-science

В приложении 8.6 описана апробация платформы u-science для решения задач ядерной физики в ходе гранта РФФИ 19-72-20114 "Разработка системы моделирования, обработки и хранения данных установки класса мегасайенс "Супер С-тау фабрика" (руководитель – д.ф.м.н. Логашенко И.Б., исполнители от ФИЦ ИВТ: Колпаков Ф.А., Киселев И.Н.) (Колпаков и др., 2019а).

ЗАКЛЮЧЕНИЕ

Разработан ПК BioUML, реализующий весь инструментарий, необходимый для успешного использования предложенной технологии. Разработанная технология и ПК BioUML существенно ускоряет и упрощает процесс создания, отладки и валидации сложных моделей биологических систем, содержащих сотни и тысячи переменных и уравнений.

По данным независимого сравнения (Maggioli et al., 2019) BioUML – самый качественный и быстрый SBML симулятор в мире (рисунок 9.1а, б).

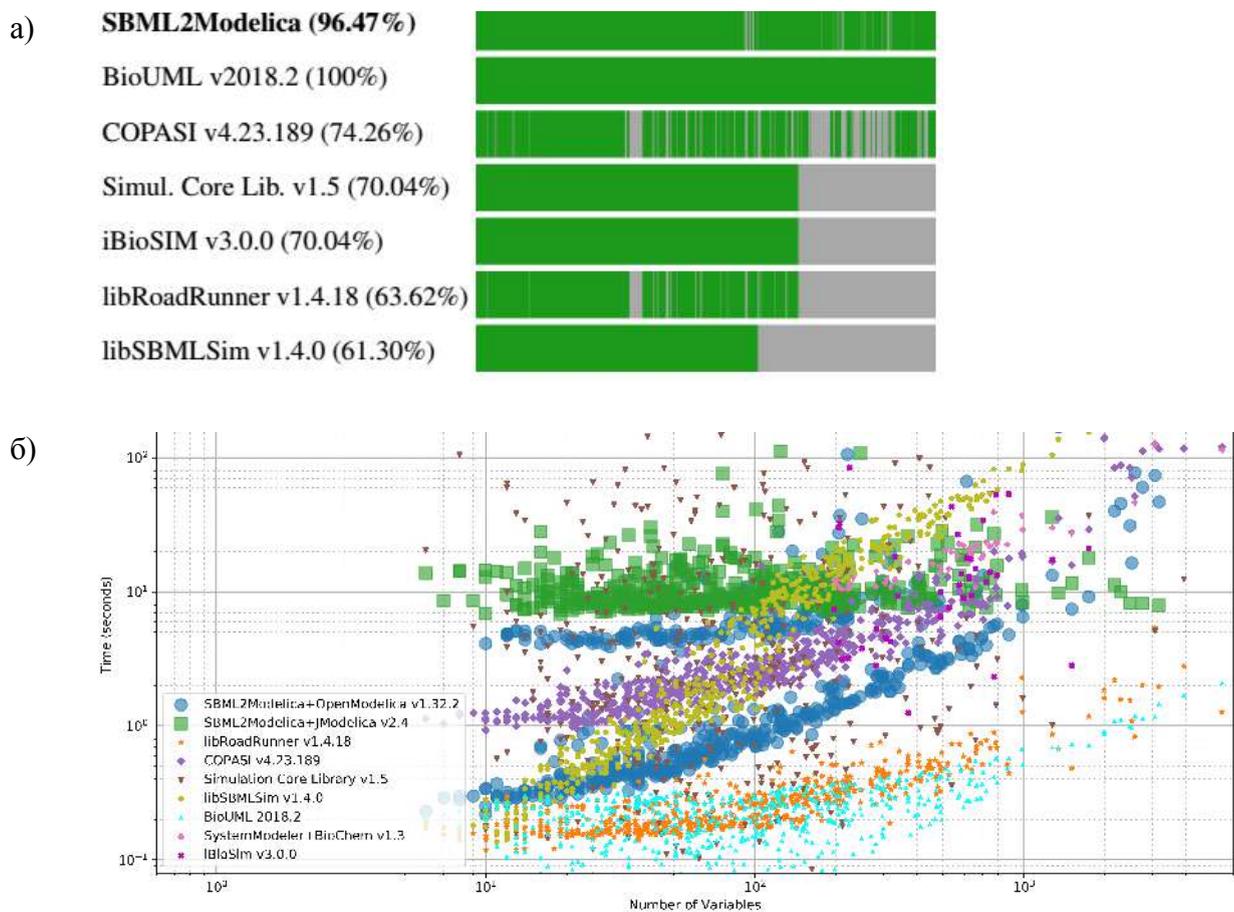


Рисунок 9.1 – Сравнение наиболее часто используемых симуляторов биологических моделей (взято из Maggioli et al., 2019):

- а) по полноте поддержки стандарта SBML на основе набора тестов SBML Test Suite Core v3.3.0
 б) по скорости работы по набору из 613 курируемых моделей из БД BioModels

Платформа Sirius-web, которая сочетает в себе возможности платформы BioUML для моделирования и анализа данных и платформы GitLab для хранения, управления и

документирования проектов, создают такую среду, на которой модели могут развиваться подобно живым существам:

- размножение – репозиторий GitLab, из которого можно легко скопировать математические модели. Подробная документация о структуре и параметрах моделей также представлена в репозитории и важна для дальнейшего развития моделей;
- мутации – редактор (как визуальный, так и текстовый) для редактирования модели;
- рекомбинация – поддержка модульного подхода для составления сложных моделей из готовых модулей;
- естественный отбор – методы подгонка параметров модели под экспериментальные данные и методы тестирования для проверки соответствия модели определенным условиям и экспериментальным данным.

Предложенная технология и разработанный ПК BioUML позволили создать наиболее сложные в мире модульные модели процессов апоптоза, регуляции генной экспрессии при работе скелетных мышц и регуляции артериального давления.

Отдельно выделим модель регуляции генной экспрессии при работе скелетных мышц, поскольку при ее создании были использованы все основные разработки, представленные в данной диссертации:

- разработанная технология сложных модульных моделей;
- ПК BioUML, реализующий данную технологию;
- ПК BioUML для анализа геномных данных: 1) анализ CAGE-seq данных для выделения генов, чья экспрессия значительно меняется при физической нагрузке; 2) анализ ChIP-seq данных для выделения подтвержденные районов связывания ТФ CREB, FOS и JUN;
- БД GTRD, в которой представлены результаты единообразного анализа данных ChIP-seq;
- платформа Sirius-web для хранения созданных моделей, воспроизводимых вычислительных экспериментов над созданными моделями в виде Jupyter ноутбуков.

Выводы

1. Разработана эффективная технология для итерационного создания, тестирования и повторного использования модульных моделей сложных биологических систем, имеющих иерархическую структуру.
2. Разработан ПК BioUML реализующий весь инструментарий, необходимый для успешного использования предложенной технологии.
3. В составе разработанного ПК BioUML включены методы, программы, сценарии для анализа и графического представления биомедицинских данных.
4. Создана новая база данных GTRD, которая обеспечивает единообразную аннотацию, контроль качества, обработку и интеграцию большого объема экспериментальных NGS данных по регуляции транскрипции для 10 видов эукариот: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Gallus gallus*, *Danio rerio*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* и *Arabidopsis thaliana*. На основе методов машинного обучения разработаны модели для предсказания уровня транскрипции для наиболее экспериментально исследованных клеточных линий на основе собранных и интегрированных NGS данных.
5. Продемонстрированы возможности разработанной технологии и ПК BioUML для создания и использования сложных моделей биологических систем, включая: процессы, происходящие на клеточном уровне (метаболизм, транскрипция, отдельные пути передачи сигнала, апоптоз); регуляцию артериального давления у человека; лекарственную терапию артериальной гипертензии; эпидемиологию COVID-19.
6. Разработана технология построения цифрового двойника пациента. Показана ее применимость на примере оптимизации выбора лекарственной терапии для лечения артериальной гипертензии.

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

АГ – артериальная гипертензия

АГП – антигипертензивные препараты

АД – артериальное давление

АПФ – ангиотензинпревращающий фермент

АТФ – АденозинТриФосфат, нуклеозидтрифосфат, играющий основную роль в обмене энергии в клетках живых организмов

БД – база данных

ГДМ – гидродинамическая модель

ДАД – диастолическое артериальное давление

ИИ – искусственный интеллект

ИМТ – индекс массы тела

КТ – компьютерная томография – метод диагностики, основанный на послойном исследовании структуры внутренних органов и систем

МИС – медицинская информационная система

МЭГ – магнитоэнцефалография – технология, позволяющая измерять и визуализировать магнитные поля, возникающие вследствие электрической активности мозга

ОДУ (ODE) – обыкновенные дифференциальные уравнения (Ordinary Differential Equations)

ОС – операционная система

ПК – программный комплекс

ПЦР – полимеразная Цепная Реакция, экспериментальный метод молекулярной биологии для значительного увеличения малых концентраций определённых фрагментов ДНК в биологической пробе

РААС – ренин-ангиотензин-альдостероновая система

РСТФ – районы связывания транскрипционных факторов

РФМК – растворимые фибрин-мономерные комплексы

САД – систолическое артериальное давление

ССТФ – сайты связывания транскрипционных факторов

СУБД – система управления БД, комплекс программ, позволяющих создать БД и манипулировать данными (вставлять, обновлять, удалять и выбирать)

ТФ – транскрипционные факторы

УЗИ – ультразвуковое исследование

УЧП – уравнения с частными производными

ФД-модель – фармакодинамическая модель

ФК-модель – фармакокинетическая модель

фМРТ – функциональная магнитно-резонансная томография, разновидность магнитно-резонансной томографии, которая проводится с целью измерения гемодинамических реакций (изменений в токе крови), вызванных нейронной активностью головного или спинного мозга

ЦДК – цветное доплеровское картирование, прямая визуализация потоков крови в сердце и в относительно крупных сосудах

ЧСС – частота сердечных сокращений

ЭКГ – электрокардиограмма, регистрация и исследование электрических полей, образующихся при работе сердца

ЭЭГ – электроэнцефалография, неинвазивный метод исследования функционального состояния головного мозга путем регистрации его биоэлектрической активности

ADAstra – Allelic Dosage-corrected Allele-Specific human Transcription factor binding sites, коллекция данных по аллель-специфичному связыванию факторов транскрипции в геноме человека

ADP – adenosine diphosphate, аденозин дифосфат

ANANASTRA – ANnotation and enrichment ANALysis of SNP sets with allele-specific TRANscription factor binding, коллекция данных по влиянию SNP на аллель-специфичное связывание факторов транскрипции в геноме человека

ANG I – ангиотензин I

ANG II – ангиотензин II

API – Application Programming Interface, программный интерфейс приложения

ARB – Angiotensin Receptor Blockers, блокаторы рецепторов ангиотензина

AST – Abstract Syntax Tree, абстрактное синтаксическое дерево, конечное помеченное ориентированное дерево, в котором внутренние вершины сопоставлены с операторами языка программирования, а листья — с соответствующими операндами

AT1 – рецепторы ангиотензина II, тип 1

ATP – adenosine triphosphate, аденозин трифосфат

ВAM (bam) – Binary Alignment Map, формат файлов, который представляет собой бинарный эквивалент SAM

ВaMM – Bayesian Markov Model, байесовская марковская модель

ВED (bed) – Browser Extensible Data, формат текстового файла, используемый для хранения геномных областей в виде координат и связанных с ними аннотаций

bigBed – формат, аналогичный BED, но предназначенный для хранения очень больших (более 50 МБ) наборов данных

bigWig – формат для отображения плотных непрерывных данных; файлы bigWig изначально создаются из файлов типа WIG, имеют индексированный двоичный формат и для больших наборов данных работают значительно быстрее, чем WIG-файлы

BioPAX – Biological Pathway Exchange, основанный на RDF/OWL стандартный язык для представления биологических путей на молекулярном и клеточном уровне

BioUML – Biological Universal Modelling Language, универсальный биологический язык моделирования, который представляет собой интегрированную Java-платформу для биомедицинских исследований

cAMP – cyclic adenosine monophosphate, циклический аденозинмонофосфат – органическое соединение, производное АТФ, выполняющее в организме роль вторичного посредника, используемого для внутриклеточного распространения сигналов некоторых гормонов (например, глюкагона или адреналина), которые не могут проходить через клеточную мембрану

ССВ – Calcium Channel Blockers, блокаторы кальциевых каналов

CellML – Cell Markup Language, язык разметки на основе XML для описания математических моделей

ChIP-exo – модификация протокола ChIP-seq, улучшающая разрешение сайтов связывания с сотен пар оснований до почти одной пары оснований; использует лямбда-эксонуклеазу (exonuclease) для переработки обработанного ультразвуком хроматина до точки сшивания индуцированного формальдегидом белка и ДНК

ChIP-seq – Chromatin Immunoprecipitation-sequencing, метод полногеномных исследований регуляции генов, который сочетает в себе иммунопреципитацию хроматина с массовым параллельным секвенированием ДНК

CLI – Command Line Interface, интерфейс командной строки, способ взаимодействия между человеком и компьютером путём отправки компьютеру команд, представляющих собой последовательность символов

CO – cardiac output, сердечный выброс

COMBINE – COmputational Modeling in BIOlogy' NEtwork, организация, координирующая разработку стандартов в системной и синтетической биологии

COVID-19 – инфекционное заболевание, вызываемое коронавирусом SARS-CoV-2

CRAN – The Comprehensive R Archive Network, сеть ftp и веб-серверов по всему миру, на которых хранятся идентичные актуальные версии кода и документации для R

CREB – cAMP Response Element-Binding protein, транскрипционный фактор, способный связывать CRE-последовательности ДНК и регулировать экспрессию генов

CSS – Cascading Style Sheets, формальный язык декорирования и описания внешнего вида документа (веб-страницы), написанного с использованием языка разметки (например, HTML)

CWL – Common Workflow Language, общий язык рабочих процессов, стандарт для описания рабочих процессов вычислительного анализа данных

Cy3 – cyanine-3, флуоресцентный краситель

Cy5 – cyanine-5, флуоресцентный краситель

DevOps – технология активного взаимодействия специалистов по разработке со специалистами по информационно-технологическому обслуживанию и взаимная интеграция их рабочих процессов друг в друга для обеспечения качества продукта

DML – Diagram Markup Language, XML формат, используемый ПК BioUML для хранения диаграмм

DRI – Direct Renin Inhibitors, прямые ингибиторы ренина

DSL – Domain Specific Language, доменный язык

FAIR-данные – данные, соответствующие принципам Findability, Accessibility, Interoperability, Reusability — находимости, доступности, совместимости и переиспользования

FNCM – False Negative Control Metric, метрика для контроля ложно-невьявленных ССТФ

FPCM – False Positive Control Metric, метрика для контроля ложно-выявленных ССТФ

GFF (gff) – General Feature Format, простой текстовый формат с разделителями табуляцией для описания геномных особенностей

GPU – Graphics Processing Unit, графический процессор, один из видов микропроцессоров

GTRD – Gene Transcription Regulation Database, БД по регуляции транскрипции

Hct – hematocrit, гематокрит, процентное содержание форменных элементов крови в ее общем объеме

НОСОМОСО – HOmo sapiens COmprehensive MOdel COllection, коллекция мотивов для ССТФ человека и мыши

JIT-компиляция – компиляция Just-in-Time (точно в нужное время), технология увеличения производительности программных систем, использующих байт-код, путём компиляции байт-кода в машинный код или в другой формат непосредственно во время работы программы

JSON – JavaScript Object Notation, текстовый формат обмена данными, основанный на JavaScript

KEGG – Kyoto Encyclopedia of Genes and Genomes, веб-ресурс, предоставляющий доступ к ряду биологических БД и инструментам для анализа биологических и медицинских данных

KiSAO – the Kinetic Simulation Algorithm Ontology, классификация алгоритмов симуляционной биологии, их параметров и результатов

LLVM – Low Level Virtual Machine, проект программной инфраструктуры для создания

компиляторов и сопутствующих им утилит

MAP – Medium Arterial Pressure, среднее артериальное давление

MathML – Mathematical Markup Language, язык разметки на основе XML для представления математических символов и формул в веб-документах

MIAME – Minimum Information About a Microarray Experiment, стандарт для мета-информации для микрочиповых экспериментов

MIASE – Minimum Information About A Simulation Experiment, стандарт мета-информации для выполнения и воспроизведения численного имитационного эксперимента

MIRIAM – Minimum information requested in the annotation of biochemical models, стандарт мета-информации для аннотирования и курирования количественных моделей биологических систем

mSigDB – Molecular Signatures DataBase, БД аннотированных наборов генов для человека и мыши

NCBI – National Center for Biotechnology Information, Национальный центр биотехнологической информации США

NGS – Next Generation Sequencing, технология массового параллельного секвенирования, которая обеспечивает сверхвысокую пропускную способность, масштабируемость и скорость

NLP – Natural Language Processing, область компьютерных наук, связанная с предоставлением компьютерам возможности понимать текст и произносимые слова почти так же, как люди

NYHA classification – New York Heart Association classification, классификация Нью-Йоркской кардиологической ассоциации

OSGi – Open Services Gateway Initiative, спецификация динамической модульной системы и сервисной платформы для Java-приложений

SAM – Sequence Alignment Map, текстовый формат, первоначально предназначенный для хранения биологических последовательностей, выровненных по эталонной последовательности

SARS-CoV-2 – Severe Acute Respiratory Syndrome-related CoronaVirus 2, оболочечный вирус с одноцепочечной РНК позитивной полярности, относящийся к семейству Coronaviridae

SBGN – Systems Biology Graphical Notation, язык графического представления элементов биологических систем и взаимодействий между ними

SBML – Systems Biology Markup Language, XML формат для представления биологических моделей

SD – standard deviation, стандартное отклонение

SED-ML – Simulation Experiment Description Markup Language, XML формат для воспроизведения результатов численных экспериментов с математическими моделями биологических систем

SELDI – Surface-enhanced laser desorption/ionization, лазерная десорбция/ионизация с усилением поверхности, метод мягкой ионизации в масс-спектрометрии, используемый для анализа белковых смесей

SIR модель – математическая модель распространения инфекционных заболеваний на основе популяций восприимчивых (S), инфицированных (I) и выздоровевших (R) индивидов

SISSRs – Site Identification from Short Sequence Reads, идентификация сайта по чтению коротких последовательностей

SMRT – Single Molecule Real Time Sequencing – одномолекулярное секвенирование в реальном времени – метод секвенирования ДНК нового поколения, разработанный компанией Pacific Biosciences

SNP – Single Nucleotide Polymorphism, однонуклеотидный полиморфизм

SNV – single nucleotide variants, замена одного нуклеотида в геноме

SOAP – Simple Object Access Protocol, протокол обмена структурированными сообщениями в распределенной вычислительной среде

SQL – Structured Query Language, декларативный язык программирования, применяемый для создания, модификации и управления данными в реляционной БД, управляемой соответствующей системой управления БД

TAP-MS – Tandem Affinity Purification – Mass Spectrometry, тандемная аффинная очистка на основе иммунопреципитации для изучения белок-белковых взаимодействий методом масспектрометрии

TSS – Transcription Start Sites, сайт начала транскрипции – место, где первый нуклеотид ДНК транскрибируется в РНК

URL – Uniform Resource Locator, унифицированный указатель ресурса, система унифицированных адресов электронных ресурсов, или единообразный определитель местонахождения ресурса

VCF (vcf) – Variant Call Format, текстовый формат для представления однонуклеотидных вариантов, вставок, вариантов числа копий

WDL – Workflow Definition Language, язык описания рабочих процессов обработки данных с помощью удобочитаемого и записываемого синтаксиса

WIG (wig) – wiggle format, текстовый формат для отображения количественных данных в браузере генома

XML – eXtensible Markup Language, расширяемый язык разметки

XP – eXtreme Programming, экстремальное программирование – технология разработки программного обеспечения, предназначенная для повышения его качества и реагирования на меняющиеся требования клиентов

Y2H – Yeast Two-Hybrid system, двугибридный анализ, молекулярно-биологический метод, позволяющий с высокой точностью детектировать белок-белковые и ДНК-белковые взаимодействия в условиях *in vivo*

YAML – Ain't Markup Language, формат сериализации данных, концептуально близкий к языкам разметки, но ориентированный на удобство ввода-вывода типичных структур данных многих языков программирования

СПИСОК ЛИТЕРАТУРЫ

1. Об утверждении Концепции предиктивной, превентивной и персонализированной медицины: приказ Министерства здравоохранения РФ N 186 от 24 апреля 2018 г.). – Доступ из справ.-правовой системы КонсультантПлюс. – Текст: электронный.
2. Акбердин, И. Р. Математическая модель, связывающая Ca^{2+} -зависимый сигнальный путь с регуляцией экспрессии генов в клетках скелетной мышцы человека / Акбердин И.Р., Вертышев А.Ю., Пинтус С.С. [и др.] // Математическая биология и биоинформатика. – 2020. – Т. 15. – №. 1. – С. 20-39.
3. Бенькович, Е. Практическое моделирование динамических систем / Бенькович Е., Колесов Ю., Сениченков Ю. – СПб: ВHV, 2002. – 464 с. – ISBN: 5-94157-099-6.
4. Бибердорф, Э. А. Система кровообращения человека – математическое и компьютерное моделирование: учебное пособие / Бибердорф Э. А., Блохин Ф. М., Киселев И. Н. – РИЦ НГУ, Новосиб. гос. ун.-т., Новосибирск, 2015. 189 с.
5. Гольдштейн, Б.Н. Кинетические графы в энзимологии. // М.: Наука, 1989. – 166 с. – ISBN ТАГ 60-04042018-15.
6. Деарлав Д. Бизнес путь: Билл Гейтс. 10 секретов самого богатого в мире бизнес-лидера / Деарлав Д. – СПб.: Изд-во «Крылов», 2003. – 208 с. – ISBN 5-94371-402-2.
7. Интенсивная терапия. Национальное руководство. Т. 2. / Под редакцией Гельфанда Б. Р., Салтанова А. И. – Москва: ГЭОТАР-Медиа, 2011. – 784 с. – ISBN: 978-5-9704-1789-8.
8. Канторович, Л. В. О методе Ньютона / Канторович Л. В. // Труды Математического института имени В.А. Стеклова. – 1949. – Т. 28. – С. 104-144.
9. Карпов Ю. Имитационное моделирование систем. Введение в моделирование с AnyLogic 5 / Карпов Ю. – СПб: БХВ-Петербург, 2006. – 400 с. – ISBN 5-94157-148-8.
10. Киселев, И. Н. Модульное моделирование биологических систем на примере сердечно-сосудистой системы человека : специальность 03.01.09 «Математическая биология, биоинформатика» : диссертация на соискание ученой степени кандидата физико-математических наук // Киселев Илья Николаевич ; ФГБУН Институт вычислительных технологий СО РАН. – Новосибирск, 2016. – 177 с.
11. Киселев, И. Н. Модульное моделирование сердечно-сосудистой системы человека / Киселев И. Н., Семисалов Б. В., Бибердорф Э. А. [и др.] // Математическая биология и биоинформатика. – 2012. – Т. 7. – №. 2. – С. 703-736.
12. Киселев, И. Н. Модульная графическая модель энергетического метаболизма в клетках скелетной мышцы / Киселев И.Н., Акбердин И.Р., Вертышев А.Ю. [и др.] // Математическая биология и биоинформатика. – 2019а. – Т. 14. – №. 2. – С. 373-392.
13. Киселев, И. Н. Математическое моделирование действия антигипертензивных препаратов /

- Киселев И.Н., Кутумова Е.О., Колпакова А.Ф. [и др.] // Математическая биология и биоинформатика. – 2019b. – Т. 14. – №. 1. – С. 233-256.
14. Колпаков, Ф. А. Формальное описание и визуальное моделирование биологических систем : специальность 03.01.09 «Математическая биология, биоинформатика» : диссертация на соискание ученой степени кандидата биологических наук // Колпаков Федор Анатольевич ; ФГБУН Институт биомедицинской химии имени В. Н. Ореховича РАМН. – Москва, 2011. – 150 с. – Библиогр.: с. 138–147. – Текст : непосредственный
 15. Колпаков, Ф.А. Компьютерная система MGL – инструмент для составления выборок, графического представления и анализа регуляторных геномных последовательностей / Колпаков Ф.А., Бабенко В.Н. // Молекулярная биология. – 1997. – Т. 31. – №.4 – С. 647-655.
 16. Колпаков, Ф.А. Универсальная платформа для моделирования и анализа научных данных / Колпаков Ф.А., Киселев И.Н., Юрченко А.В. [и др.] // Международная научно-практическая конференция «Распределенные информационно-вычислительные ресурсы: Цифровые двойники и большие данные» (DICR 2019): Новосибирск, 3-6 декабря 2019а г.). – URL: <http://conf.ict.nsc.ru/dicr2019/>
 17. Колпаков, Ф. А. Виртуальный пациент: разработка технологии на примере модульной математической модели сердечно-сосудистой системы человека и предсказание эффективности лечения артериальной гипертензии / Колпаков Ф.А., Киселев И.Н., Кутумова Е.О. [и др.] // Биотехнология-медицине будущего. – 2019b. – С. 74.
 18. Колпаков, Ф. А. Виртуальный пациент / Колпаков Ф. А., Киселев И. Н. // Коммерсантъ Наука. – 2019с. – №. 4 (1). – С. 28.
 19. Колпаков, Ф. А. Sirius-web-среда для эволюционного создания математических моделей сложных биологических систем / Колпаков Ф.А., Кутумова Е.О., Акбердин И.Р. [и др.] // Международная конференция Прикладная математика иммунологии и вирусологии, Сочи, Россия, 14-18 ноября 2022 года. Программа. – Сочи: Университет «Сириус». – 2022. – С. 53-56. – URL: https://siriusmathcenter.ru/pr_img/1918100371/20221112/62063230/Program_027w.pdf.
 20. Колпакова, А. Ф. О возможностях применения прямого ингибитора ренина алискирена для лечения социально значимых заболеваний / Колпакова А. Ф. // Сибирский журнал клинической и экспериментальной медицины. – 2014. – Т. 29. – №. 2. – С. 14-21.
 21. Кропачева Е. С. Быстрые темпы насыщения варфарином-предиктор развития чрезмерной гипокоагуляции. Модернизация алгоритма подбора дозы варфарина / Кропачева Е.С., Боровков Н.Н., Вавилова Т.В. [и др.] // Атеротромбоз. – 2015. – №. 1. – С. 74-86.
 22. Кутумова, Е. О. Модульная модель процесса программируемой гибели клеток: построение, верификация и упрощение : специальность 03.01.02 «Биофизика» : диссертация на

- соискание ученой степени кандидата физико-математических наук // Кутумова Елена Олеговна ; ФГБУН Конструкторско-технологический институт вычислительной техники СО РАН. – Красноярск, 2012. – 131 с. – Библиогр.: с. 114–128. – Текст : непосредственный
23. Кутумова Е. О. Применение методов редукции для построения комплексной модели апоптотических путей / Кутумова Е.О., Зиновьев А.Ю., Шарипов Р.Н. [и др.] // Математическая биология и биоинформатика. – 2012. – V. 7. – №. 2. – P. 572-588.
 24. Прошин, А. П. Математическое моделирование системы кровообращения и его практические применения / Прошин А. П., Солодянников Ю. В. // Автоматика и телемеханика. – 2006. – №. 2. – С.174-188.
 25. Ратнер, В. А. Блочно-модульный принцип организации и эволюции молекулярно-генетических систем управления (МГСУ) / Ратнер, В.А. // Генетика. – 1992. – Т. 28. – №. 2. – С. 5-23.
 26. Ратнер, В. А. Проблемы теории молекулярной эволюции / Ратнер В. А., Жарких А.А., Колчанов Н.А. [и др.] – Новосибирск: Наука, 1985. – 259 с.
 27. Солодянников, Ю. В. Элементы математического моделирования и идентификация системы кровообращения / Солодянников Ю.В. – Самара: Самар. ун-т, 1994. – 315 с. – ISBN: 5-230-06001-8.
 28. Стельмашенко, Д.Е. Персонализированная медицина в один клик: автоматическая идентификация терапевтической мишени по ДНК и РНК пациента / Стельмашенко Д.Е., Колпаков Ф.А., Евшин И.С. [и др.] // Лаборатория и производство. – 2019. – Т. 3. – №.7 – С. 76-86.
 29. Тахтеев, И. А. Обзор систем контейнеризации Docker и Singularity в рамках кластеров суперкомпьютеров / Тахтеев, И. А. // Молодой ученый. – 2022. – №. 21. – С. 217-220.
 30. Терентьев, А. А. Динамическая протеомика в моделировании живой клетки. Белок-белковые взаимодействия / Терентьев А. А., Молдогазиева Н. Т., Шайтан К. В. // Успехи биологической химии. – 2009. – Т. 49. – С. 429-480.
 31. Чазова, И. Е. Клинические рекомендации диагностика и лечение артериальной гипертензии / Чазова И.Е., Ощепкова Е.В., Жернакова Ю.В. // Кардиологический вестник. – 2015. – Т. 10. – №. 1. – С.3-30.
 32. Чураев, Р. Н. Моделирование реальных эукариотических управляющих генных подсетей на основе метода обобщенных пороговых моделей / Чураев Р.Н., Галимзянов А.В. // Молекулярная биология. – 2001. – Т. 35. – №. 6. – С. 1088-1094.
 33. Abate, G. Pinacidil and hydrochlorothiazide alone or in combination in the treatment of hypertension in the elderly / G. Abate, M. Zito, P. U. Carbonin [et al.] // Current therapeutic research. – 1998. – V. 59. – №. 1. – P. 62-71.

34. Abramov, S. Landscape of allele-specific transcription factor binding in the human genome / Abramov S., Boytsov A., Bykova D. [et al.] // *Nature Communications*. – 2021. – V.12. – Art. 2751.
35. Adams, J. M. Ways of dying: multiple pathways to apoptosis / J. M. Adams // *Genes & development*. – 2003. – V. 17. – №. 20. – P. 2481-2495.
36. Afonyushkin, V. Multicompartmental Mathematical Model of SARS-CoV-2 Distribution in Human Organs and Their Treatment / Afonyushkin V., Akberdin I.R., Kozlova Y. [et al.] // *Mathematics*. – 2022. – V. 10. – №. 11. – Art. 1925.
37. Agabiti Rosei, E. Effects of nifedipine GITS 20 mg or enalapril 20 mg on blood pressure and inflammatory markers in patients with mild–moderate hypertension / E. Agabiti Rosei, P. Morelli, D. Rizzoni // *Blood pressure*. – 2005. – V. 14. – №. sup1. – P. 14-22.
38. Akberdin, I. R. A modular mathematical model of exercise-induced changes in metabolism, signaling, and gene expression in human skeletal muscle / Akberdin I. R., Kiselev I. N., Pintus S. S. // *International Journal of Molecular Sciences*. – 2021. – V. 22. – №. 19. – Art. 10353.
39. Akberdin I. R. Integrated modular model linking metabolism, signaling transduction and gene expression regulation in human skeletal muscle1 / Akberdin I. R., Kiselev I. N., Pintus S. S. [et al.] // *CEUR Workshop Proceedings*. – 2019. – V. 2569. – P. 1-5.
40. Albeck, J. G. Modeling a snap-action, variable-delay switch controlling extrinsic cell death / J. G. Albeck, J. M. Burke, S. L. Spencer, D. A. Lauffenburger [et al.] // *PLoS biology*. – 2008. – V. 6. – №. 12. – P. e299.
41. Allen, R. J. Efficient generation and selection of virtual populations in quantitative systems pharmacology models / R. J. Allen, T. R. Rieger, C. J. Musante // *CPT Pharmacomet. Syst. Pharmacol.* – 2016. – V. 5. – N 3. – P. 140–146.
42. Bagci, E. Z. Bistability in apoptosis: roles of bax, bcl-2, and mitochondrial permeability transition pores / E. Z. Bagci, Y. Vodovotz, T. R. Billiar, G. B. Ermentrout, I. Bahar // *Biophysical journal*. – 2006. – V. 90. – №. 5. – P. 1546-1559.
43. Bairoch, A. The Cellosaurus, a cell-line knowledge resource / A. Bairoch // *Journal of biomolecular techniques: JBT*. – 2018. – V. 29. – №. 2. – P. 25.
44. Bajracharya, K. Comparison of three agent-based platforms on the basis of a simple epidemiological model (WIP) / K. Bajracharya, R. Duboz // *Proceedings of the Symposium on Theory of Modeling & Simulation-DEVS Integrative M&S Symposium*. – 2013. – P. 1-6.
45. Barker, A. R. Muscle phosphocreatine kinetics in children and adults at the onset and offset of moderate-intensity exercise / A. R. Barker, J. R. Welsman, J. Fulford, D. Welford, N. Armstrong // *Journal of Applied Physiology*. – 2008. – V. 105. – №. 2. – P. 446-456.
46. Barrand, G. GAUDI—A software architecture and framework for building HEP data processing

- applications / Barrant, G., Belyaev, I., Binko, P. [et al.] // *Computer Physics Communications*. – 2001. – V. 140. – №. 1-2. – P. 45-55.
47. Bartlett, M. F. Oxidative ATP synthesis in human quadriceps declines during 4 minutes of maximal contractions / M. F. Bartlett, L. F. Fitzgerald, R. Nagarajan, Y. Hiroi, J. A. Kent // *The Journal of Physiology*. – 2020. – V. 598. – №. 10. – P. 1847-1863.
48. Beard, D. A. Multiscale modeling and data integration in the virtual physiological rat project / Beard, D. A., Neal, M. L., Tabesh-Saleki, N. [et al.] // *Annals of Biomedical Engineering*. – 2012. – V. 40. – P. 2365-2378.
49. Bellifemine, F. L. JADE-A White Paper. Exp in Search of Innovation / F. L. Bellifemine, G. Caire, A. Poggi, G. Rimassa // *EXP in search of innovation*. – Telecom Italia, 2003. – V. 3. – P. 6-19.
50. Bentele, M. Mathematical modeling reveals threshold mechanism in CD95-induced apoptosis / M. Bentele, I. Lavrik, M. Ulrich, S. Stösser [et al.] // *The Journal of Cell Biology*. – 2004. – V. 166. – №. 6. – P. 839-851.
51. Berdeaux, R. Anabolic and pro-metabolic functions of CREB-CRTC in skeletal muscle: advantages and obstacles for type 2 diabetes and cancer cachexia / R. Berdeaux, C. Hutchins // *Frontiers in Endocrinology*. – 2019. – V. 10. – P. 535.
52. Bergmann, F. T. COPASI and its applications in biotechnology / F. T. Bergmann, S. Hoops, B. Klahn, U. Kummer [et al.] // *Journal of Biotechnology*. – 2017. – V. 261. – P. 215-220.
53. Biasini, M. pv: v1. 8.1 / Biasini M. – Текст: электронный // Zenodo: интернет-портал. 2015. – URL: <https://zenodo.org/record/20980> (дата обращения: 04.07.2023)
54. Birmingham, C. L. Methods to monitor autophagy of Salmonella enterica serovar Typhimurium / Birmingham C. L., Brumell J. H. // *Methods in Enzymology*. – 2009. – V. 452. – P. 325-343.
55. Björkman, M. Global optimization using the DIRECT algorithm in Matlab . Björkman M., Holmström K. // *Advanced Modeling and Optimization*. – 1999. – V. 1. – №. 2. – P. 17-37.
56. Blinov, M. L. Compartmental and spatial rule-based modeling with virtual cell / Blinov, M. L., Schaff, J. C., Vasilescu, D. [et al.] // *Biophysical Journal*. – 2017. – V. 113. – №. 7. – P. 1365-1372.
57. Bliss, R. D. A. G. Role of feedback inhibition in stabilizing the classical operon / R. D. Bliss, P. R. Painter, A. G. Marr // *Journal of Theoretical Biology*. – 1982. – V. 97. – №. 2. – P. 177-193.
58. Bock, C. Computational epigenetics / Bock C., Lengauer T. // *Bioinformatics*. – 2008. – V. 24. – №. 1. – P. 1-10.
59. Boëlle, P. Y. Trajectories of hospitalization in COVID-19 patients: an observational study in France / Boëlle, P. Y., Delory, T., Maynadier, X. [et al.] // *Journal of Clinical Medicine*. – 2020. – V. 9. – №. 10. – Art. 3148.
60. Boëlle, P. Y. Excess cases of influenza-like illnesses synchronous with coronavirus disease (COVID-19) epidemic, France, March 2020 / Boëlle, P. Y., Souty, C., Launay, T. [et al.] //

- Eurosurveillance. – 2020. – V. 25. – №. 14. – Art. 2000326.
61. Bonabeau, E. Agent-based modeling: Methods and techniques for simulating human systems / E. Bonabeau // Proceedings of the national academy of sciences. – 2002. – V. 99. – №. suppl_3. – P. 7280-7287.
 62. Bonate, P. L. Pharmacokinetic-pharmacodynamic modeling and simulation / P. L. Bonate. – New-York: Springer, 2011. – 618 c.
 63. Borisov, N. Systems-level interactions between insulin-EGF networks amplify mitogenic signaling / N. Borisov, E. Aksamitiene, A. Kiyatkin, S. Legewie S. [et al.] // Molecular Systems Biology. – 2009. – V. 5. – №. 1. – P. 256.
 64. Boyarskikh, U. Computational master-regulator search reveals mTOR and PI3K pathways responsible for low sensitivity of NCI-H292 and A427 lung cancer cell lines to cytotoxic action of p53 activator Nutlin-3 / U. Boyarskikh, S. Pintus, N. Mandrik, D. Stelmashenko [et al.] // BMC Medical Genomics. – 2018. – V. 11. – P. 43-59.
 65. Boytsov, A. ANANASTRA: annotation and enrichment analysis of allele-specific transcription factor binding at SNPs / Boytsov A., Abramov S., Aiusheeva A. [et al.] // Nucleic Acids Research. – 2022. – V. 50. – №. W1. – P. W51-W56.
 66. Brown, M. J. Aliskiren and the calcium channel blocker amlodipine combination as an initial treatment strategy for hypertension control (ACCELERATE): a randomised, parallel-group trial / M. J. Brown, G. T. McInnes, C. C. Papst, J. Zhang [et al.] // The Lancet. – 2011. – V. 377. – №. 9762. – P. 312-320.
 67. Bruno, N. E. Creb coactivators direct anabolic responses and enhance performance of skeletal muscle / N. E. Bruno, K. A. Kelly, R. Hawkins, M. Bramah-Lawani [et al.] // The EMBO Journal. – 2014. – V. 33. – №. 9. – P. 1027-1043.
 68. Cannon, D. T. Skeletal muscle ATP turnover by ³¹P magnetic resonance spectroscopy during moderate and heavy bilateral knee extension / D. T. Cannon, W. E. Bimson, S. A. Hampson, T.S. Bowen [et al.] // The Journal of Physiology. – 2014. – V. 592. – №. 23. – P. 5287-5300.
 69. Carrier, T. A. Investigating autocatalytic gene expression systems through mechanistic modeling / T. A. Carrier, J. D. Keasling // Journal of Theoretical Biology. – 1999. – V. 201. – №. 1. – P. 25-36.
 70. Carter, B. L. Hydrochlorothiazide versus chlorthalidone: evidence supporting their interchangeability / B. L. Carter, M. E. Ernst, J. D. Cohen // Hypertension. – 2004. – V. 43. – №. 1. – P. 4-9.
 71. Chaitanya, G. V. PARP-1 cleavage fragments: signatures of cell-death proteases in neurodegeneration / G. V. Chaitanya, J. S. Alexander, P. P. Babu // Cell Communication and Signaling. – 2010. – V. 8. – P. 1-11.
 72. Chan, T. Artificial markets and intelligent agents: дис. – Massachusetts Institute of Technology,

2001. – 178 pp.
73. Chandran, D. Hierarchical modeling for synthetic biology / D. Chandran, H. M. Sauro // *ACS Synthetic Biology*. – 2012. – V. 1. – №. 8. – P. 353-364.
 74. Chandran, D. TinkerCell: modular CAD tool for synthetic biology / D. Chandran, F. T. Bergmann, H. M. Sauro // *Journal of Biological Engineering*. – 2009. – V. 3. – №. 1. – P. 1-17.
 75. Chao, A. Estimating the population size for capture-recapture data with unequal catchability / A. Chao // *Biometrics*. – 1987. – P. 783-791.
 76. Chen K.C. Kinetic analysis of a molecular model of the budding yeast cell cycle / K.C. Chen, A. Csikasz-Nagy, B. Gyorffy [et al.] // *Molecular Biology of The Cell*. – 2000. – V. 11. – №. 1. – P. 369-391.
 77. Chen, K. C. Integrative analysis of cell cycle control in budding yeast / K. C. Chen, L. Calzone, A. Csikasz-Nagy [et al.] // *Molecular Biology of The Cell*. – 2004. – V. 15. – №. 8. – P. 3841-3862.
 78. Cheng, Y. QSP toolbox: computational implementation of integrated workflow components for deploying multi-scale mechanistic models / Y. Cheng, C. J. Thalhauser, S. Smithline [et al.] // *The AAPS journal*. – 2017. – V. 19. – P. 1002-1016.
 79. Chernykh, I. Using Simulation System AGNES for Modeling Execution of Parallel Algorithms on Supercomputers / I. Chernykh, B. Glinskiy, I. Kulikov // *Computers, Automatic Control, Signal Processing and Systems Science*. – 2014. – P. 66-70.
 80. Clough, E. The gene expression omnibus database / E. Clough, T. Barrett // *Statistical Genomics: Methods and Protocols*. – 2016. – V. 1418. – P. 93-110.
 81. Cohen, S. D. CVODE, a stiff/nonstiff ODE solver in C / Cohen S. D., Hindmarsh A. C., Dubois P. F. // *Computers in Physics*. – 1996. – V. 10. – №. 2. – P. 138-143.
 82. Coleman, T. G. A mathematical model of renal hemodynamics and excretory function / T. G. Coleman, J. E. Hall. *Structuring Biological Systems: A Computer Modelling Approach* ed. by S. S. Iyengar. – CRC Press, 1992. P. 89-124.
 83. Cooper, L. The plant ontology as a tool for comparative plant anatomy and genomic analyses / Cooper L, Walls RL, Elser J [et al.] // *Plant and Cell Physiology*. – 2013. – V. 54. – №. 2. – Art. e1.
 84. Copeland, T. E. *Generating parsers with JavaCC* / T. E. Copeland. – Centennial Books, 2007. – 258 pp. – ISBN 0-9762214-3-8.
 85. Corea, L. Valsartan, a new angiotensin II antagonist for the treatment of essential hypertension: a comparative study of the efficacy and safety against amlodipine / Corea L, Cardoni O, Fogari R. [et al.] // *Clinical Pharmacology & Therapeutics*. – 1996. – V. 60. – №. 3. – P. 341-346.
 86. Courtot, M. Controlled vocabularies and semantics in systems biology / Courtot M., Juty N., Knüpfer C. [et al.] // *Molecular Systems Biology*. – 2011. – V. 7. – №. 1. – Art. 543.

87. Covert, M. W. Constraints-based models: regulation of gene expression reduces the steady-state solution space / Covert M. W., Palsson B. O. // *Journal of Theoretical Biology*. – 2003. – V. 221. – №. 3. – P. 309-325.
88. Covert, M. W. Integrating high-throughput and computational data elucidates bacterial networks / Covert M.W, Knight E.M, Reed J.L. [et al.] // *Nature*. – 2004. – V. 429. – №. 6987. – P. 92-96.
89. Crusoe, M. R. Methods included: standardizing computational reuse and portability with the common workflow language / Crusoe M.R., Abeln S., Iosup A. [et al] // *Communications of The ACM*. – 2022. – V. 65. – №. 6. – P. 54-63.
90. Cushman, W. C. Home and clinic blood pressure responses in elderly individuals with systolic hypertension / Cushman WC, Duprez DA, Weintraub HS [et al.] // *Journal of the American Society of Hypertension*. – 2012. – V. 6. – №. 3. – P. 210-218.
91. Da Veiga Leprevost, F. BioContainers: an open-source and community-driven framework for software standardization / da Veiga Leprevost F, Grüning BA, Alves Aflitos S [et al.] // *Bioinformatics*. – 2017. – V. 33. – №. 16. – P. 2580-2582.
92. De Jong H. Modeling and simulation of genetic regulatory systems: a literature review / De Jong H. // *Journal of Computational Biology*. – 2002. – V. 9. – №. 1. – P. 67-103.
93. DeLano, W.L. The PyMOL Molecular Graphics System / DeLano W.L. – San Carlos: DeLano Scientific, CA, USA, 2002.
94. Derosa, G. Effects of enalapril/lercanidipine combination on some emerging biomarkers in cardiovascular risk stratification in hypertensive patients / Derosa G, Bonaventura A, Romano D. [et al.] // *Journal of Clinical Pharmacy and Therapeutics*. – 2014. – V. 39. – №. 3. – P. 277-285.
95. Devineni, D. Effects of hydrochlorothiazide on the pharmacokinetics, pharmacodynamics, and tolerability of canagliflozin, a sodium glucose co-transporter 2 inhibitor, in healthy participants / Devineni D, Vaccaro N, Polidori D. [et al.] // *Clinical therapeutics*. – 2014. – V. 36. – №. 5. – P. 698-710.
96. Di Tommaso, P. Nextflow enables reproducible computational workflows / Di Tommaso P, Chatzou M, Floden EW. [et al.] // *Nature Biotechnology*. – 2017. – V. 35. – №. 4. – P. 316-319.
97. Diehl, A. D. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability / A. D. Diehl, T. F. Meehan, Y. M. Bradford [et al.] // *Journal of Biomedical Semantics*. – 2016. – V. 7. – P. 1-10.
98. Doi, A. Constructing biological pathway models with hybrid functional Petri nets / Doi A., Fujita S, Matsuno H. [et al.] // *In Silico Biology*. – 2004. – V. 4. – №. 3. – P. 271-291.
99. Dormand, J. R. A family of embedded Runge-Kutta formulae / Dormand J. R., Prince P. J. // *Journal of Computational and Applied Mathematics*. – 1980. – V. 6. – №. 1. – P. 19-26.
100. Doshi, P. Covid-19: Do many people have pre-existing immunity? / Doshi, P. // *The BMJ*. – 2020.

- T. 370. – Art. m3563.
101. Drasdo, D. The virtual liver: state of the art and future perspectives / Drasdo, D., Bode, J., Dahmen, U. [et al.] // *Archives of Toxicology*. – 2014. – V. 88. – №. 12. – P. 2071–2075.
102. Duarte J. D. Mechanisms for blood pressure lowering and metabolic effects of thiazide and thiazide-like diuretics / Duarte JD, Cooper-DeHoff RM // *Expert Review of Cardiovascular Therapy*. – 2010. – V. 8. – №. 6. – P. 793-802.
103. Endy, D. Intracellular kinetics of a growing virus: a genetically structured simulation for bacteriophage T7 / Endy D., Kong D., Yin J. // *Biotechnology and Bioengineering*. – 1997. – V. 55. – №. 2. – P. 375-389.
104. Erhard, F. FERN—a Java framework for stochastic simulation and evaluation of reaction networks / Erhard F., Friedel C. C., Zimmer R. // *BMC Bioinformatics*. – 2008. – V. 9. – №. 1. – P. 1-12.
105. Evans, T. W. The SBML discrete stochastic models test suite / Evans T. W., Gillespie C. S., Wilkinson D. // *Bioinformatics*. – 2008. – V. 24. – №. 2. – P. 285-286.
106. Fabregat, A. Reactome diagram viewer: data structures and strategies to boost performance / Fabregat A, Sidiropoulos K, Viteri G. [et al.] // *Bioinformatics*. – 2018. – V. 34. – №. 7. – P. 1208-1214.
107. Fagher, B. Antihypertensive and renal effects of enalapril and slow-release verapamil in essential hypertension: A double-blind, randomized study / Fagher B, Henningsen N, Hulthén L. [et al.] // *European Journal of Clinical Pharmacology*. – 1990. – V. 39. – P. S41-S43.
108. Fan, T. J. Caspase family proteases and apoptosis / Fan T.J., Han L.H., Cong R.S., Liang J. // *Acta Biochimica et Biophysica Sinica*. – 2005. – V. 37. – №. 11. – P. 719-727.
109. Faulkner, J. K. The pharmacokinetics of amlodipine in healthy volunteers after single intravenous and oral doses and after 14 repeated oral doses given once daily / Faulkner J.K., McGibney D., Chasseaud L.F. [et al.] // *British Journal of Clinical Pharmacology*. – 1986. – V. 22. – №. 1. – P. 21-25.
110. Ferguson, R. K. Effects of enalapril, a new converting enzyme inhibitor, in hypertension / Ferguson R.K., Vlasses P.H., Swanson B.N. // *Clinical Pharmacology & Therapeutics*. – 1982. – V. 32. – №. 1. – P. 48-53.
111. Ferme, I. Comparative study on monotherapy with sustained-release diltiazem 300 mg and enalapril 20 mg in mild to moderate arterial hypertension / Ferme I, Djian J, Tcherdakoff P. // *Journal of Cardiovascular Pharmacology*. – 1990. – V. 16. – P. S46-50.
112. Frownfelter, J. Why IBM Watson Health could never live up to the promises / Frownfelter J. – Текст: электронный // MedCityNews: интернет-портал. 2021. – URL: <https://medcitynews.com/2021/04/why-ibm-watson-health-could-never-live-up-to-the-promises/> (дата обращения: 04.07.2023).

113. Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update / Afgan E., Nekrutenko A., Grüning B. A. [et al.] // *Nucleic Acids Research*. – 2022. – V. 50. – №. W1. – P. W345-W351.
114. Gibson, M. A. Efficient exact stochastic simulation of chemical systems with many species and many channels / Gibson M. A., Bruck J. // *The Journal of Physical Chemistry A*. – 2000. – V. 104. – №. 9. – P. 1876-1889.
115. Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions / Gillespie D. T. // *Journal of Computational Physics*. – 1976. – V. 22. – №. 4. – P. 403-434.
116. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions / Gillespie D. T. // *The Journal of Physical Chemistry*. – 1977. – V. 81. – №. 25. – P. 2340-2361.
117. Goldbeter, A. A model for circadian oscillations in the *Drosophila* period protein (PER) / Goldbeter A. // *Proceedings of the Royal Society of London. Series B: Biological Sciences*. – 1995. – V. 261. – №. 1362. – P. 319-324.
118. Goode, J. M. The nuclear receptor, Nor-1, induces the physiological responses associated with exercise / Goode J.M., Pearen M.A., Tuong Z.K. [et al.] // *Molecular Endocrinology*. – 2016. – V. 30. – №. 6. – P. 660-676.
119. Goss, P. J. E. Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets / Goss P. J. E., Peccoud J. // *Proceedings of the National Academy of Sciences*. – 1998. – V. 95. – №. 12. – P. 6750-6755.
120. Gradman, A. H. Aliskiren, a novel orally effective renin inhibitor, provides dose-dependent antihypertensive efficacy and placebo-like tolerability in hypertensive patients / Gradman AH, Schmieder RE, Lins RL. [et al.] // *Circulation*. – 2005. – V. 111. – №. 8. – P. 1012-1018.
121. Grebennikov, D.. Predicting the kinetic coordination of immune response dynamics in SARS-CoV-2 infection: implications for disease pathogenesis / Grebennikov, D., Karsonova, A., Loguinova, M. [et al.] // *Mathematics*. – 2022. – V. 10. – №. 17. – Art. 3154.
122. Greiner, A. High-energy phosphate metabolism in the calf muscle of healthy humans during incremental calf exercise with and without moderate cuff stenosis / Greiner A, Esterhammer R, Bammer D. [et al.] // *European Journal of Applied Physiology*. – 2007. – V. 99. – P. 519-531.
123. Gremse, M. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources / Gremse M, Chang A, Schomburg I. [et al.] // *Nucleic Acids Research*. – 2010. – V. 39. – №. suppl_1. – P. D507-D513.
124. Grieves, M. Digital twin: manufacturing excellence through virtual factory replication / Grieves M. // *White paper*. – 2014. – V. 1. – №. 2014. – P. 1-7.
125. Grieves, M. Digital twin: mitigating unpredictable, undesirable emergent behavior in complex

- systems / Grieves M., Vickers J. // *Transdisciplinary perspectives on complex systems: New findings and approaches*. – 2017. – P. 85-113.
- 126.Grüning, B. Bioconda: sustainable and comprehensive software distribution for the life sciences / Grüning B., Dale R., Sjödin A. [et al.] // *Nature Methods*. – 2018. – V. 15. – №. 7. – P. 475-476.
- 127.Guo, Y. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints / Guo Y., Mahony S., Gifford D. K. // *PLoS Computational Biology*. – 2012. – V. 8. – №. 8. – Art. e1002638.
- 128.Guyton, A. C. Circulation: overall regulation / Guyton A. C., Coleman T. G., Granger H. J. // *Annual Review of Physiology*. – 1972. – V. 34. – №. 1. – P. 13-44.
- 129.Hai, T. Cross-family dimerization of transcription factors Fos/Jun and ATF/CREB alters DNA binding specificity / Hai T., Curran T. // *Proceedings of the National Academy of Sciences*. – 1991. – V. 88. – №. 9. – P. 3720-3724.
- 130.Hairer, E. Solving ordinary differential equations II / Hairer E., Wanner G. – New York : Springer Berlin Heidelberg, 1996. – V. 375.
- 131.Hale, T. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker) / Hale, T., Angrist, N., Goldszmidt, R. [et al.] // *Nature Human Behaviour*. – 2021. – V. 5. – №. 4. – P. 529-538.
- 132.Hallow, K. M. A model-based approach to investigating the pathophysiological mechanisms of hypertension and response to antihypertensive therapies: extending the Guyton model / Hallow K.M., Lo A., Beh J. [et al.] // *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*. – 2014. – V. 306. – №. 9. – P. R647-R662.
- 133.Hallow, K. M. Renal effects of dapagliflozin in people with and without diabetes with moderate or severe renal dysfunction: prospective modeling of an ongoing clinical trial / Hallow KM, Boulton DW, Penland RC. [et al.] // *Journal of Pharmacology and Experimental Therapeutics*. – 2020. – V. 375. – №. 1. – P. 76-91.
- 134.Hamada, H. Sophisticated framework between cell cycle arrest and apoptosis induction based on p53 dynamics / Hamada H., Tashima Y., Kisaka Y. [et al.] // *PloS One*. – 2009. – V. 4. – №. 3. – Art. e4795.
- 135.Hammal, F. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments / Hammal, F., de Langen, P., Bergon, A. [et al.] // *Nucleic Acids Research*. – 2022. – V. 50. – №. D1. – P. D316-D325.
- 136.Hanai, T. Application of knowledge information processing methods to biochemical engineering, biomedical and bioinformatics fields / Hanai T., Honda H. // *Advances in Biochemical Engineering/Biotechnology*. – 2004. – V. 91. – P. 51-73.

137. Hanson, R. M. Jmol—a paradigm shift in crystallographic visualization / Hanson R. M. // *Journal of Applied Crystallography*. – 2010. – V. 43. – №. 5. – P. 1250-1260.
138. Harris, L. A. BioNetGen 2.2: advances in rule-based modeling / Harris, L. A., Hogg, J. S., Tapia, J. J. [et al.] // *Bioinformatics*. – 2016. – V. 32. – №. 21. – P. 3366-3368.
139. Hawley, J. A. Integrative biology of exercise / Hawley J.A., Hargreaves M., Joyne, M.J., Zierath J.R. // *Cell*. – 2014. – V. 159. – №. 4. – P. 738-749.
140. Hayashi, K. Ca²⁺ channel subtypes and pharmacology in the kidney / Hayashi K., Wakino S., Sugano N. // *Circulation Research*. – 2007. – V. 100. – №. 3. – P. 342-353.
141. Heinemeyer, T. Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL / Heinemeyer T., Wingender E., Reuter I. [et al.] // *Nucleic Acids Research*. – 1998. – V. 26. – №. 1. – P. 362-367.
142. Hernández, A. I. A multiformalism and multiresolution modelling environment: application to the cardiovascular system and its regulation / Hernández A. I., Le Rolle V., Defontaine A., Carrault G. // *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. – 2009. – V. 367. – №. 1908. – P. 4923-4940.
143. Hestenes, M. R. Methods of conjugate gradients for solving linear systems / Hestenes, M. R., Stiefel, E. // *Journal of Research of the National Bureau of Standards*. – 1952. – V. 49. – №. 6. – P. 409-436.
144. Hindmarsh, A. C. SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers / Hindmarsh A. C., Brown P. N., Grant K. E. [et al.] // *ACM Transactions on Mathematical Software (TOMS)*. – 2005. – V. 31. – №. 3. – P. 363-396.
145. Hofestädt, R. Interactive modelling and simulation of biochemical networks / Hofestädt R., Meineke F. // *Computers in Biology and Medicine*. – 1995. – V. 25. – №. 3. – P. 321-334.
146. Hoffmann, A. The I κ B-NF- κ B signaling module: temporal control and selective gene activation / Hoffmann A., Levchenko A., Scott M.L., Baltimore D. // *Science*. – 2002. – V. 298. – №. 5596. – P. 1241-1245.
147. Hoffman, B. G. Genome-wide identification of DNA-protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing / Hoffman B. G., Jones S. J. M. // *Journal of Endocrinology*. – 2009. – V. 201. – №. 1. – P. 1-13.
148. Hong, Y. Pharmacokinetic/pharmacodynamic modeling of renin biomarkers in subjects treated with the renin inhibitor aliskiren / Hong Y., Dingemans J., Mager D.E. // *Clinical Pharmacology & Therapeutics*. – 2008. – V. 84. – №. 1. – P. 136-143.
149. Hsin, J. Using VMD: an introductory tutorial / Hsin, J., Arkhipov, A., Yin, Y. [et al.] // *Current Protocols in Bioinformatics*. – 2008. – V. 24. – №. 1. – P. 5.7.1-5.7.48.
150. Huang, J. ChIPBase v3.0: the encyclopedia of transcriptional regulations of non-coding RNAs and

- protein-coding genes / Huang, J., Zheng, W., Zhang, P. [et al.] // *Nucleic Acids Research*. – 2023. – V. 51. – №. D1. – P. D46-D56.
151. Huang, S. Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery // *Journal of Molecular Medicine*. – 1999. – V. 77. – №. 6. – P. 469-480.
152. Huber, W. Orchestrating high-throughput genomic analysis with Bioconductor / Huber W., Carey V., Gentleman R. [et al.] // *Nature Methods*. – 2015. – V. 12. – №. 2. – P. 115-121.
153. Hucka, M. The ERATO systems biology workbench: an integrated environment for multiscale and multitheoretic simulations in systems biology / Hucka M, Sauro HM, Finney A. [et al.] // *Foundations of Systems Biology*. – Cambridge, MA: MIT Press, 2001. – P. 125–143.
154. Hucka, M. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models / Hucka M, Finney A, Sauro HM. [et al.] // *Bioinformatics*. – 2003. – V. 19. – №. 4. – P. 524-531.
155. Hucka, M. The systems biology markup language (SBML): language specification for level 3 version 2 core release 2 / Hucka M, Bergmann FT, Chaouiya C. [et al.] // *Journal of Integrative Bioinformatics*. – 2019. – V. 16. – №. 2. – Art. 20190021.
156. Hucka, M. sbmlteam/sbml-test-suite: v3.4.0 / Hucka M., Keating S., Smith L. [et al.] – Текст: электронный // Zenodo: интернет-портал. 2022. – URL: <https://zenodo.org/record/5834571> (дата обращения: 04.07.2023)
157. Huerta-Cepas, J. ETE: a python Environment for Tree Exploration / Huerta-Cepas J., Dopazo J., Gabaldón T. // *BMC Bioinformatics*. – 2010. – V. 11. – P. 1-7.
158. Humphrey, W. VMD: visual molecular dynamics / Humphrey W., Dalke A., Schulten K. // *Journal of Molecular Graphics*. – 1996. – V. 14. – №. 1. – P. 33-38.
159. Inchiosa, M. E. Overcoming design and development challenges in agent-based modeling using ASCAPE / Inchiosa M. E., Parker M. T. // *Proceedings of the National Academy of Sciences*. – 2002. – V. 99. – №. suppl_3. – P. 7304-7308.
160. Ingber, L. Adaptive simulated annealing (ASA): Lessons learned / Ingber L. // *Control and Cybernetics*. – 1996. – V. 25. – №. 1. – P. 33-54.
161. Ison, J. The bio. tools registry of software tools and data resources for the life sciences / Ison J, Ienasescu H, Chmura P. [et al.] // *Genome Biology*. – 2019. – V. 20. – №. 1. – P. 1-4.
162. Janes, K. A. The response of human epithelial cells to TNF involves an inducible autocrine cascade / Janes K.A., Gaudet S., Albeck J.G. [et al.] // *Cell*. – 2006. – V. 124. – №. 6. – P. 1225-1239.
163. JavaBeans API specification / – Текст: электронный // Oracle: интернет-портал. – URL: <https://edelivery.oracle.com/otn-pub/jcp/7224-javabeans-1.01-fr-spec-oth-JSpec/beans.101.pdf>

(дата обращения: 04.07.2023)

164. Jones, T. C. JigCell Model Connector: building large molecular network models from components / Jones T. C., Jr; Hoops, S., Watson, L. T. [et al.] // *Simulation*. – 2018. – V. 94. – №. 11. – P. 993-1008.
165. Kanehisa, M. KEGG: integrating viruses and cellular organisms / Kanehisa M, Furumichi M, Sato Y. [et al.] // *Nucleic Acids Research*. – 2021. – V. 49. – №. D1. – P. D545-D551.
166. Karaaslan, F. Long-term mathematical model involving renal sympathetic nerve activity, arterial pressure, and sodium excretion / Karaaslan F., Denizhan Y., Kayserilioglu A., Gulcur H.O. // *Annals of Biomedical Engineering*. – 2005. – V. 33. – P. 1607-1630.
167. Karr, J. R. A whole-cell computational model predicts phenotype from genotype // Karr, J. R., Sanghvi, J. C., Macklin, D. N. [et al.] // *Cell*. – 2012. – V. 150. – №. 2. – P. 389–401.
168. Karr J. R. WholeCellSimDB: a hybrid relational/HDF database for whole-cell model predictions // Karr J. R., Phillips N. C., Covert M. W. // *Database*. – 2014. – V. 2014. – Art. bau095.
169. Kauffman, S. A. The origins of order: Self-organization and selection in evolution / Kauffman, S. A. – New York: Oxford University Press, USA, 1993. – 709 pp.
170. Kawaji, H. The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types / Kawaji H, Kasukawa T, Forrest A. [et al.] // *Scientific Data*. – 2017. – V. 4. – №. 1. – P. 1-3.
171. Keating, S. M. SBML Level 3: an extensible format for the exchange and reuse of biological models / Keating SM, Waltemath D, König M. [et al.] // *Molecular Systems Biology*. – 2020. – V. 16. – №. 8. – Art. e9110.
172. Kel, A. E. Computer tool FUNSITE for analysis of eukaryotic regulatory genomic sequences / Kel A. E., Kondrakhin Y. V., Kolpakov Ph. A. [et al.] // *Proceedings. International Conference on Intelligent Systems for Molecular Biology*. – 1995. – V. 3. – P. 197-205.
173. Kel, A. Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations / Kel A., Konovalova T., Waleev T. [et al.] // *Bioinformatics*. – 2006. – V. 22. – №. 10. – P. 1190-1197.
174. Kel, A. GeneXplain—identification of causal biomarkers and drug targets in personalized cancer pathways / Kel A., Kolpakov F., Poroikov V., Selivanova G. // *Journal of Biomolecular Techniques: JBT*. – 2011. – V. 22. – №. Suppl. – Art. S16.
175. Kel, A. E. Multi-omics “upstream analysis” of regulatory genomic regions helps identifying targets against methotrexate resistance of colon cancer / Kel AE, Stegmaier P, Valeev T. [et al.] // *EuPA open proteomics*. – 2016. – V. 13. – P. 1-13.
176. Kel, A. Walking pathways with positive feedback loops reveal DNA methylation biomarkers of colorectal cancer / Kel A., Boyarskikh U., Stegmaier P. [et al.] // *BMC Bioinformatics*. – 2019. –

V. 20. – №. 4. – P. 1-20.

- 177.Kholodenko B.N. Mechanistic and modular approaches to modeling and inference of cellular regulatory networks / Kholodenko B.N., Bruggeman F.J., Sauro H.M. In: Systems Biology: Definitions and Perspectives; eds. Alberghina L. and Westerhoff H.V. – Heidelberg, Germany: Springer-Verlag, 2005. – 408 pp.
- 178.Kierzek, A. M. The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression / Kierzek AM, Zaim J, Zielenkiewicz P // Journal of Biological Chemistry. – 2001. – V. 276. – №. 11. – P. 8165-8171.
- 179.Kim, S. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data / Kim S., Imoto S., Miyano S. // Biosystems. – 2004. – V. 75. – №. 1-3. – P. 57-65.
- 180.King, E. L. A schematic method of deriving the rate laws for enzyme-catalyzed reactions / King E. L., Altman C. // The Journal of Physical Chemistry. – 1956. – V. 60. – №. 10. – P. 1375-1378.
- 181.Kiselev, I. N. BioUML: modular modeling of complex biological / Kiselev I. N., Kolpakov F. A. // Proceedings of the 8th International Conference on Bioinformatics of Genome Regulation and Structure\Systems Biology (BGRS\SB'2012), June 25-29, 2012. – Novosibirsk, 2012. – P. 146.
- 182.Kiselev, I. N. Modular approach to modeling complex biological systems / Kiselev I. N., Semisalov B. V., Sharipov R. N., Kolpakov F. // Circulatory System and Arterial Hypertension: Experimental Investigation, Mathematical and Computer Simulation; eds. Ivanova L. N., Markel A. L., Blokhin A. M., Mishchenko E. V. – Nova Science Publishers, Inc., 2012a. – P. 143–154
- 183.Kiselev, I. N. Modular modeling of the human cardiovascular system / Kiselev I. N., Semisalov B. V., Sharipov R. N., Kolpakov F. A. // Circulatory System and Arterial Hypertension: Experimental Investigation, Mathematical and Computer Simulation; eds. L. N. Ivanova, A. L. Markel, A. M. Blokhin, E. V. Mishchenko. – Nova Science Publishers, Inc., 2012b. – P. 155–206.
- 184.Kiselev, I. Modular modeling of biological systems / Kiselev I. N., Kolpakov F. A. // Virtual Biology. – 2013. – V. 1. – №. 1. – P. 30-44.
- 185.Kiselev, I. BioUML: plugin for population-based modeling / Kiselev I., Kolpakov F. // Virtual Biology. – 2014. – V. 1. – №. 2. – P. 7-15.
- 186.Kiselev, I. N. Improved SBGN (ML) support in BioUML / Kiselev I. N., Kinsht S., Kolpakov F. A. // Proceedings of the Conference “Mathematical Modeling And High Performance Computing In Bioinformatics, Biomedicine And Biotechnology” (MM-HPC-BBB-2016), Novosibirsk, August 29 – September 2. – Novosibirsk, 2016. – P. 57.
- 187.Kiselev, I. N. A delay differential equation approach to model the COVID-19 pandemic / Kiselev I. N., Akberdin I. R., Kolpakov F. A. // medRxiv. – 2021. – Art. 2021.09.01.21263002.
- 188.Kitano, H. Computational systems biology / Kitano, H. // Nature. – 2002a. – V. 420. – №.6912. –

- P. 206-210.
189. Kitano, H. Systems biology: a brief overview / Kitano H. // *Science*. – 2002b. – V. 295. – №. 5560. – P. 1662-1664.
 190. Kitano, H. Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology / Kitano H. // *Current Genetics*. – 2002c. – V. 41. – №. 1. – P. 1-10.
 191. Kofránek, J. Restoration of Guyton's diagram for regulation of the circulation as a basis for quantitative physiological model development / Kofránek J., Rusz J. // *Physiological Research*. – 2010. – V. 59. – №. 6. – P. 897-908.
 192. Koh, B. T., Tan R. B. H., Yap M. G. S. Genetically structured mathematical modeling of trp attenuator mechanism / Koh B. T., Tan R. B. H., Yap M. G. S. // *Biotechnology and Bioengineering*. – 1998. – V. 58. – №. 5. – P. 502-509.
 193. Koh, K. K. Additive beneficial effects of losartan combined with simvastatin in the treatment of hypercholesterolemic, hypertensive patients / Koh KK, Quon MJ, Han SH. [et al.] // *Circulation*. – 2004. – V. 110. – №. 24. – P. 3687-3692.
 194. Kolchanov, N. A. Transcription regulatory regions database (TRRD): its status in 1999 / Kolchanov N.A., Ananko E.A., Podkolodnaya O.A. [et al.] // *Nucleic Acids Research*. – 1999. – V. 27. – №. 1. – P. 303-306.
 195. Kolmykov, S. K. Population size estimation for quality control of ChIP-Seq datasets / Kolmykov, S. K., Kondrakhin, Y. V., Yevshin, I. S. [et al.] // *PLoS One*. – 2019. – V. 14. – №. 8. – Art. e0221760.
 196. Kolmykov, S. K. Analysis of NGS Data on the Transcriptional Regulation / Kolmykov, S. K., Evshin, I. S., Kolpakov, F. A. // *CEUR Workshop Proceedings*. – 2020a. – P. 19-22.
 197. Kolmykov, S. K. Meta-analysis of ChIP-seq datasets through the rank aggregation approach / Kolmykov S. K., Kondrakhin Y. V., Sharipov R. N. [et al.] // *2020 Cognitive Sciences, Genomics and Bioinformatics (CSGB). IEEE*. – 2020b. – P. 180-184.
 198. Kolmykov, S. GTRD: an integrated view of transcription regulation / Kolmykov, S. K., Yevshin, I. S., Kulyashov M. K. [et al.] // *Nucleic Acids Research*. – 2021. – V. 49. – №. D1. – P. D104-D111.
 199. Kolmykov, S. Transcription factor binding sites: data integration, stable identifiers and incremental builds / Kolmykov, S., Kondrakhin, Y., Sharipov, R. [et al.] // *Proceedings. Bioinformatics of genome regulation and structure/systems biology (BGRS/SB-2022)*. – 2022. – C. 77.
 200. Kolpakov, F. A. BioUML – Framework for visual modeling and simulation of biological systems / Kolpakov F. A. // *Proceedings of the 3rd International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002)*. – Novosibirsk, 2002. – V.2 – P. 128-131.
 201. Kolpakov, F.A. BioUML – open source extensible workbench for systems biology / Kolpakov

- F.A. // Proceedings of the 4th International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2004), July 25-30, 2004. – Novosibirsk, 2004. – P. 77-80.
202. Kolpakov, F. A. GeneNet: a gene network database and its automated visualization / Kolpakov F.A., Ananko E.A., Kolesov G.B. [et al.] // Bioinformatics (Oxford, England). – 1998. – V. 14. – №. 6. – P. 529-537.
203. Kolpakov, F. A. Interactive data input into the GeneNet database / Kolpakov F. A., Ananko E. A. // Bioinformatics (Oxford, England). – 1999. – V. 15. – №. 7. – P. 713-714.
204. Kolpakov, F. BioUML: visual modeling, automated code generation and simulation of biological systems / Kolpakov F., Puzanov M., Koshukov A. // Proceedings of the 5th International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2006), July 16-22, 2006. – Novosibirsk, 2006. – P. 281-284.
205. Kolpakov, F. A.. BioUML – integrated platform for building virtual cell and virtual physiological human / Kolpakov, F. A., Tolstyh, N., Kutumova, E. O. [et al.] // Proceedings of the 7th International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2010), June 20-27, 2010. – Novosibirsk, 2010. – P. 144.
206. Kolpakov, F. A. BioUML Platform - From Systems Biology to Systems Pharmacology / Kolpakov F. A., Kiselev I. N., Valeev T. F. [et al.] // 7th Noordwijkerhout Symposium on Pharmacokinetics, Pharmacodynamics and Systems Pharmacology. Systems Pharmacology in Drug Discovery & Development: Noordwijkerhout, The Netherlands, 23-25 April 2014. – Noordwijkerhout: Leiden Academic Centre for Drug Research. – 2014. – P. 156-157.
207. Kolpakov, F. BioUML: an integrated environment for systems biology and collaborative analysis of biomedical data / Kolpakov F., Akberdin I., Kashapov T. [et al.] // Nucleic Acids Research. – 2019. – V. 47. – №. W1. – P. W225-W233.
208. Kolpakov, F. BioUML—Towards a universal research platform / Kolpakov F., Akberdin I.R., Kiselev I.N. [et al.] // Nucleic Acids Research. – 2022. – V. 50. – №. W1. – P. W124-W131.
209. Kononowicz A. A. Virtual patients-what are we talking about? A framework to classify the meanings of the term in healthcare education / Kononowicz AA, Zary N, Edelbring S. [et al.] // BMC Medical Education. – 2015. – V. 15. – №. 1. – P. 1-7.
210. Koschmann, J. “Upstream analysis”: an integrated promoter-pathway analysis approach to causal interpretation of microarray data / Koschmann J., Bhar A., Stegmaier P. [et al.] // Microarrays. – 2015. – V. 4. – №. 2. – P. 270-286.
211. Koulmann, N. Interaction between signalling pathways involved in skeletal muscle responses to endurance exercise / Koulmann N., Bigard A. X. // Pflügers Archiv. – 2006. – V. 452. – P. 125-139.
212. Kroemer, G. The biochemistry of programmed cell death / Kroemer G, Petit P, Zamzami N. [et al.]

// The FASEB Journal. – 1995. – V. 9. – №. 13. – P. 1277-1287.

213. Krull, M. TRANSPATH®: an information resource for storing and visualizing signaling pathways and their pathological aberrations / Krull, M., Pistor, S., Voss, N. [et al.] // *Nucleic Acids Research*. – 2006. – V. 34. – №. suppl_1. – P. D546-D551.
214. Krustup, P. Slow-Twitch Fiber Glycogen Depletion Elevates Moderate-Exercise Fast-Twitch Fiber Activity and O₂ Uptake / Krustup P, Söderlund K, Mohr M. [et al.] // *Medicine and Science in Sports and Exercise*. – 2004. – V. 36. – №. 6. – P. 973-982.
215. Kulakovskiy, I. V. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis / Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S. [et al.] // *Nucleic Acids Research*. – 2018. – V. 46. – №. D1. – P. D252-D259.
216. Kulyashov, M. A. Description, characteristic and algorithm for creation of a dictionary of cell types and tissues in the GTRD database / Kulyashov, M. A., Kolmykov, S. K., Yevshin, I. S., Kolpakov, F. A. // *CEUR Workshop Proceedings*. – 2020a. – V. 2569. – P. 13-18.
217. Kulyashov, M. A. Advanced data curation in GTRD database: hierarchical dictionaries of cell types and experimental factors / Kulyashov, M. A., Kolmykov, S. K., Yevshin, I. S., Kolpakov, F. A. // *2020 Cognitive Sciences, Genomics and Bioinformatics (CSGB)*. – IEEE, 2020b. – P. 23-27.
218. Kutumova, E. O. A modular model of the apoptosis machinery / Kutumova E.O., Kiselev I. N., Sharipov R. N. [et al.] // *Advances in Systems Biology*. – 2012. – V. 736. – P. 235-245.
219. Kutumova, E. BioUML plug-in for nonlinear parameter estimation using multiple experimental data / Kutumova E., Ryabova A., Valeev T., Kolpakov F. // *Virtual Biology*. – 2013a. – V. 1. – №. 1. – P. 47-58.
220. Kutumova, E. Model composition through model reduction: a combined model of CD95 and NF- κ B signaling pathways / Kutumova E., Sharipov R., Kolpakov F., Zinovyev A. // *BMC Systems Biology*. – 2013b. – V. 7. – P. 1-21.
221. Kutumova, E. Thoroughly calibrated modular agent-based model of the human cardiovascular and renal systems for blood pressure regulation in health and disease / Kutumova E, Kiselev I, Sharipov R. [et al.] // *Frontiers in Physiology*. – 2021. – V. 12. – Art. 746300.
222. Kutumova, E. Mathematical modeling of antihypertensive therapy / Kutumova E., Kiselev I., Sharipov R. [et al.] // *Frontiers in Physiology*. – 2022a. – V.13. – Art.1070115.
223. Kutumova, E. Modular representation of physiologically based pharmacokinetic models: Nanoparticle delivery to solid tumors in mice as an example / Kutumova E, Akberdin I, Kiselev I. [et al.] // *Mathematics*. – 2022b. – V. 10. – №. 7. – Art. 1176.
224. Kwakernaak, A. J. Effects of direct renin blockade on renal & systemic hemodynamics and on RAAS activity, in weight excess and hypertension: a randomized clinical trial / Kwakernaak AJ, Roksnoer LC, Lambers Heerspink HJ. [et al.] // *PLOS One*. – 2017. – V. 12. – №. 1. – Art.

- e0169258.
- 225.Lamponi, D. One dimensional and multiscale models for blood flow circulation. – EPFL, Lausanne, 2004. – №. THESIS № 3006.
- 226.Landahl, H. D. Some conditions for sustained oscillations in biochemical chains with feedback inhibition / Landahl H. D. // *The Bulletin of Mathematical Biophysics*. – 1969. – V. 31. – P. 775-787.
- 227.Landt, S. G. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia / Landt, S. G., Marinov, G. K., Kundaje, A. [et al.] // *Genome Research*. – 2012. – V. 22. – №. 9. – P. 1813-1831.
- 228.Langmead, B. Fast gapped-read alignment with Bowtie 2 / Langmead B., Salzberg S. L. // *Nature Methods*. – 2012. – V. 9. – №. 4. – P. 357-359.
- 229.Lanumteang, K., Böhning D. An extension of Chao's estimator of population size based on the first three capture frequency counts / Lanumteang K., Böhning D. // *Computational Statistics & Data Analysis*. – 2011. – V. 55. – №. 7. – P. 2302-2311.
- 230.Laragh, J. H. The plasma renin test reveals the contribution of body sodium-volume content (V) and renin-angiotensin (R) vasoconstriction to long-term blood pressure / Laragh J. H., Sealey J. E. // *American Journal of Hypertension*. – 2011. – V. 24. – №. 11. – P. 1164-1180.
- 231.Lauer S. A. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application / Lauer, S. A., Grantz, K. H., Bi, Q. [et al.] // *Annals of Internal Medicine*. – 2020. – V. 172. – №. 9. – P. 577-582.
- 232.Le Novere, N. Systems biology graphical notation: process diagram level 1 / Le Novere, N., Moodie, S., Sorokin, A. [et al.] // *Nature Precedings*. – 2008. – P. 1-1.
- 233.Lee, C. T. Non-local concepts and models in biology / Lee C.T., Hoopes M.F., Diehl J. [et al.] // *Journal of Theoretical Biology*. – 2001. – V. 210. – №. 2. – P. 201-219.
- 234.Lee, J. H. Isolation and genetic characterization of H5N2 influenza viruses from pigs in Korea / Lee, J. H., Pascua, P. N. Q., Song, M. S. [et al.] // *Journal of Virology*. – 2009. – V. 83. – №. 9. – P. 4205-4215.
- 235.Lee, S. E. Efficacy and tolerability of fimasartan, a new angiotensin receptor blocker, compared with losartan (50/100 mg): a 12-week, phase III, multicenter, prospective, randomized, double-blind, parallel-group, dose escalation clinical trial with an optional 12-week extension phase in adult Korean patients with mild-to-moderate hypertension / Lee SE, Kim YJ, Lee HY. [et al.] // *Clinical Therapeutics*. – 2012. – V. 34. – №. 3. – P. 552-568. e9.
- 236.Legewie, S. Mathematical modeling identifies inhibitors of apoptosis as mediators of positive feedback and bistability / Legewie S., Bluthgen N., Herzel H. // *PLOS Computational Biology*. – 2006. – V. 2. – №. 9. – Art. e120.

237. Lehrach H. Omics approaches to individual variation: modeling networks and the virtual patient / Lehrach H. // *Dialogues in Clinical Neuroscience*. – 2016. – V. 18. – №. 3. – P. 253-265.
238. Leinonen, R. The sequence read archive / Leinonen R, Sugawara H, Shumway M. // *Nucleic Acids Research*. – 2010. – V. 39. – №. suppl_1. – P. D19-D21.
239. Leth, A. Changes in plasma and extracellular fluid volumes in patients with essential hypertension during long-term treatment with hydrochlorothiazide / Leth A. // *Circulation*. – 1970. – V. 42. – №. 3. – P. 479-485.
240. Levenberg, K. A method for the solution of certain non-linear problems in least squares / Levenberg K. // *Quarterly of Applied Mathematics*. – 1944. – V. 2. – №. 2. – P. 164-168.
241. Li, H. Algorithms and software for stochastic simulation of biochemical reacting systems / Li H., Cao Y., Petzold L. [et al.] // *Biotechnology Progress*. – 2008. – V. 24. – №. 1. – P. 56-61.
242. Li, H. 1000 genome project data processing subgroup. The sequence alignment/map (SAM) format and SAMtools / Li H., Handsaker B., Wysoker A. [et al.] // *Bioinformatics*. – 2009. – V. 25. – №. 16. – P. 2078-2079.
243. Li, Y. Role of NADH/NAD⁺ transport activity and glycogen store on skeletal muscle energy metabolism during exercise: in silico studies / Li Y., Dash R. K., Kim J. [et al.] // *American Journal of Physiology-Cell Physiology*. – 2009. – V. 296. – №. 1. – P. 25-46.
244. Liberzon, A. The molecular signatures database hallmark gene set collection / Liberzon A, Birger C, Thorvaldsdóttir H. // *Cell Systems*. – 2015. – V. 1. – №. 6. – P. 417-425.
245. Likhoshvai, V. Generalized Hill function method for modeling molecular processes / Likhoshvai V., Ratushny A. // *Journal of Bioinformatics and Computational Biology*. – 2007. – V. 5. – №. 02b. – P. 521-531.
246. Lithell, H. Efficacy and safety of bisoprolol and atenolol in patients with mild to moderate hypertension: a double-blind, parallel group international multicentre study / Lithell H, Selinus I, Hosie J. [et al.] // *European Heart Journal*. – 1987. – V. 8. – №. suppl_M. – P. 55-64.
247. Liu, T. Cistrome: an integrative platform for transcriptional regulation studies / Liu, T., Ortiz, J. A., Taing, L. [et al.] // *Genome Biology*. – 2011. – V. 12. – №. 8. – P. 1-10.
248. Luke, S. Mason: A multiagent simulation environment / Luke S, Cioffi-Revilla C, Panait L. [et al.] // *Simulation*. – 2005. – V. 81. – №. 7. – P. 517-527.
249. Luo, Y. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal / Luo, Y., Hitz, B. C., Gabdank, I. [et al.] // *Nucleic Acids Research*. – 2020. – V. 48. – №. D1. – P. D882-D889.
250. Macal, C. M. Agent-based modeling and simulation / Macal C. M., North M. J. // *Proceedings of the 2009 Winter Simulation Conference (WSC)*. – IEEE, 2009. – P. 86-98.
251. MacDonald, N. Biological delay systems: linear stability theory / MacDonald N. – Cambridge

- University Press, 2008. – 248 pp.
252. MacFarlane, M. Apoptosis and disease: a life or death decision / MacFarlane M., Williams AC // *EMBO Reports*. – 2004. – V. 5. – №.7 – P. 674–678.
253. MacKay, J. H. Losartan and low-dose hydrochlorothiazide in patients with essential hypertension: a double-blind, placebo-controlled trial of concomitant administration compared with individual components / MacKay JH, Arcuri KE, Goldberg AI. [et al.] // *Archives of Internal Medicine*. – 1996. – V. 156. – №. 3. – P. 278-285.
254. Macklin, D. N. Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation / Macklin, D. N., Ahn-Horst, T. A., Choi, H. [et al.] // *Science*. – 2020. – V. 369. – №. 6502. – Art. eaav3751.
255. Maggioli, F. SBML2Modelica: integrating biochemical models within open-standard simulation ecosystems / Maggioli F., Mancini T., Tronci E. // *Bioinformatics*. – 2020. – V. 36. – №. 7. – P. 2165-2172.
256. Mahaffy, J. M. Cellular control models with linked positive and negative feedback and delays. I. The models / Mahaffy J. M. // *Journal of Theoretical Biology*. – 1984. – V. 106. – №. 2. – P. 89-102.
257. Mallion, J. M. An evaluation of the initial and long-term antihypertensive efficacy of zofenopril compared with enalapril in mild to moderate hypertension / Mallion J. M. // *Blood Pressure*. – 2007. – V. 16. – №. sup2. – P. 13-18.
258. Malone, J. Modeling sample variables with an Experimental Factor Ontology / Malone J, Holloway E, Adamusiak T. [et al.] // *Bioinformatics*. – 2010. – V. 26. – №. 8. – P. 1112-1118.
259. Mandrik N. Rule-based modeling in BioUML / Mandrik N., Kutumova E. O., Kolpakov F. A. // *Proceedings of The Second International Conference “Mathematical Modeling and High Performance Computing in Bioinformatics, Biomedicine and Biotechnology”* (within BGRS\SB-2016): Novosibirsk, Russia. 29 August – 2 September 2016. – Novosibirsk, 2016. – P. 72.
260. Marino S. The human immune response to Mycobacterium tuberculosis in lung and lymph node / Marino S., Kirschner D. E. // *Journal of Theoretical Biology*. – 2004. – V. 227. – №. 4. – P. 463-486.
261. Maritan M. Building structural models of a whole mycoplasma cell / Maritan, M., Autin, L., Karr, J. [et al.] // *Journal of Molecular Biology*. – 2022. – V. 434. – №. 2. – Art. 167351.
262. Marquardt, D. W. An algorithm for least-squares estimation of nonlinear parameters / Marquardt D. W. // *Journal of the Society for Industrial and Applied Mathematics*. – 1963. – V. 11. – №. 2. – P. 431-441.
263. Martina, B. Effects of Losartan Titrated to Losartan/Hydrochlorothiazide and Amlodipine on Left

- Ventricular Mass in Patients with Mild-to-Moderate Hypertension A Double-Blind Randomized Controlled Study / Martina B, Dieterle T, Weinbacher M, Battagay E. // *Cardiology*. – 2000. – V. 92. – №. 2. – P. 110-114.
264. Maruyama, G. Continuous Markov processes and stochastic equations / Maruyama G. // *Rendiconti del Circolo Matematico di Palermo*. – 1955. – V. 4. – P. 48-90.
265. Mateus, J. Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans / Mateus, J., Grifoni, A., Tarke, A. [et al.] // *Science*. – 2020. – V. 370. – №. 6512. – P. 89-94.
266. Matsuno, H. Hybrid Petri net representation of gene regulatory network / Matsuno H. // *Biocomputing 2000*. – 1999. – P. 341-352.
267. Matsuoka, Y. Modeling and simulation using CellDesigner / Matsuoka, Y., Funahashi, A., Ghosh, S. [et al.] // *Transcription factor regulatory networks: Methods and protocols*. – 2014. – P. 121-145.
268. Matys, V. TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes / Matys V, Kel-Margoulis OV, Fricke E. [et al.] // *Nucleic Acids Research*. – 2006. – V. 34. – №. suppl_1. – P. D108-D110.
269. Matzko, R. O. Novel Ground-Up 3D Multicellular Simulators for Synthetic Biology CAD Integrating Stochastic Gillespie Simulations Benchmarked with Topologically Variable SBML Models / Matzko R. O., Mierla L., Konur S. // *Genes*. – 2023. – V. 14. – №. 1. – P. 154.
270. McCrea, R. S. Analysis of capture-recapture data / McCrea R. S., Morgan B. J. T. – CRC Press, 2014. – 292 pp.
271. Mendoza, L. Genetic control of flower morphogenesis in *Arabidopsis thaliana*: a logical analysis / Mendoza L., Thieffry D., Alvarez-Buylla E. R. // *Bioinformatics (Oxford, England)*. – 1999. – V. 15. – №. 7. – P. 593-606.
272. Mendoza-Parra, M. A. A quality control system for profiles obtained by ChIP sequencing / Mendoza-Parra, M. A., Van Gool, W., Mohamed Saleem, M. A. [et al.] // *Nucleic Acids Research*. – 2013. – V. 41. – №. 21. – Art. e196.
273. Mesarović, M. D. Systems theory and biology – view of a theoretician / Mesarović M. D. // *Systems Theory and Biology: Proceedings of the III Systems Symposium at Case Institute of Technology*. – Berlin, Heidelberg : Springer Berlin Heidelberg, 1968. – P. 59-87.
274. Mestl, T. A mathematical framework for describing and analysing gene regulatory networks / Mestl T., Plahte E., Omholt S. // *Journal of Theoretical Biology*. – 1995a. – V. 176. – №. 2. – P. 291-300.
275. Mestl, T. Periodic solutions in systems of piecewise-linear differential equations / Mestl T., Plahte E., Omholt S. W. // *Dynamics and Stability of Systems*. – 1995b. – V. 10. – №. 2. – P. 179-193.
276. Minar, N. The swarm simulation system: A toolkit for building multi-agent simulations / Minar N,

- Burkhart R, Langton C. [et al.] – Santa Fe; Santa Fe University, 1996. – 11pp. – URL: – <https://sfi-edu.s3.amazonaws.com/sfi-edu/production/uploads/sfi-com/dev/uploads/filer/8a/2a/8a2ae001-9ad5-43e6-b7e3-4d951223e9e8/96-06-042.pdf> (дата обращения: 04.07.2023). – Текст: электронный.
- 277.Mirschel, S. PROMOT: modular modeling for systems biology / Mirschel S, Steinmetz K, Rempel M. [et al.] // *Bioinformatics*. – 2009. – V. 25. – №. 5. – P. 687-689.
- 278.Moore, F. D. Body composition and its measurement in vivo / Moore F. D. // *British Journal of Surgery*. – 1967. – V. 54. – №.13. – P. 431–435.
- 279.Moore, J. H. Petri net modeling of high-order genetic systems using grammatical evolution / Moore J. H., Hahn L. W. // *BioSystems*. – 2003. – V. 72. – №. 1-2. – P. 177-186.
- 280.Mullighan, C. G. Genome-wide profiling of genetic alterations in acute lymphoblastic leukemia: recent insights and future directions / Mullighan C. G., Downing J. R. // *Leukemia*. – 2009. – V. 23. – №. 7. – P. 1209-1218.
- 281.Mungall, C. J. Uberon, an integrative multi-species anatomy ontology / Mungall, C.J., Torniai, C., Gkoutos, G.V. [et al.] // *Genome Biology*. – 2012. – V. 13. – №.1 – P. 1-12.
- 282.Myers, C. J. iBioSim: a tool for the analysis and design of genetic circuits / Myers CJ, Barker N, Jones K. [et al.] // *Bioinformatics*. – 2009. – V. 25. – №. 21. – P. 2848-2849.
- 283.Nadler, S. B. Prediction of blood volume in normal human adults / Nadler SB, Hidalgo JU, Bloch T. // *Surgery*. – 1962. – V. 51. – P. 224–232.
- 284.Narkiewicz, K. Comparison of home and office blood pressure in hypertensive patients treated with zofenopril or losartan / Narkiewicz K. // *Blood Pressure*. – 2007. – V. 16. – №. suppl_2. – P. 7–12.
- 285.Narlikar, L. ChIP-Seq data analysis: identification of Protein–DNA binding sites with SISRr peak-finder / Narlikar L., Jothi R. // *Next Generation Microarray Bioinformatics: Methods and Protocols*. – 2012. – V. 802. – P. 305-322.
- 286.Nebro, A. J. Mocell: A cellular genetic algorithm for multiobjective optimization / Nebro A.J., Durillo J.J., Luna F. [et al.] // *International Journal of Intelligent Systems*. – 2009. – V. 24. – №. 7. – P. 726-746.
- 287.Nedogoda, S. V. Randomized trial of perindopril, enalapril, losartan and telmisartan in overweight or obese patients with hypertension / Nedogoda SV, Ledyeva AA, Chumachok EV. [et al.] // *Clinical Drug Investigation*. – 2013. – V. 33. – P. 553-561.
- 288.Nelder, J. A. A simplex method for function minimization / Nelder J. A., Mead R. // *The Computer Journal*. – 1965. – V. 7. – №. 4. – P. 308-313.
- 289.Newman, J. R. S. Comprehensive identification of human bZIP interactions with coiled-coil arrays / Newman J. R. S., Keating A. E. // *Science*. – 2003. – V. 300. – №. 5628. – P. 2097-2101.

290. Ng, K. W. Preexisting and de novo humoral immunity to SARS-CoV-2 in humans / Ng, K. W., Faulkner, N., Cornish, G. H. [et al.] // *Science*. – 2020. – V. 370. – №. 6522. – P. 1339-1343.
291. Nguyen, H. Q. Prediction of losartan-active carboxylic acid metabolite exposure following losartan administration using static and physiologically based pharmacokinetic models / Nguyen H.Q., Lin J., Kimoto E. [et al.] // *Journal of Pharmaceutical Sciences*. – 2017. – V. 106. – №. 9. – P. 2758-2770.
292. Niarakis, A. Addressing barriers in comprehensiveness, accessibility, reusability, interoperability and reproducibility of computational models in systems biology / Niarakis A., Waltemath D., Glazier J. // *Briefings in Bioinformatics*. – 2022. – V. 23. – №. 4. – Art. bbac212.
293. North, M. J. Complex adaptive systems modeling with Repast Symphony / North, M.J., Collier, N.T., Ozik, J. [et al.] // *Complex Adaptive Systems Modeling*. – 2013. – V. 1. – P. 1-26.
294. Novak, B. Mathematical model of the cell division cycle of fission yeast / Novak B, Pataki Z, Ciliberto A, Tyson JJ. // *Chaos: An Interdisciplinary Journal of Nonlinear Science*. – 2001. – V. 11. – №. 1. – P. 277-286.
295. Novotny, M. V. Biochemical individuality reflected in chromatographic, electrophoretic and mass-spectrometric profiles / Novotny M. V., Soini H. A., Mechref Y. // *Journal of Chromatography B*. – 2008. – V. 866. – №. 1-2. – P. 26-47.
296. Nussberger, J. Angiotensin II Suppression in humans by the orally active renin inhibitor aliskiren (SPP100) comparison with enalapril / Nussberger, J., Wuerzner, G., Jensen, C. [et al.] // *Hypertension*. – 2002. – V. 39. – №. 1. – P. e1-e8.
297. Okonechnikov, K. Unipro UGENE: a unified bioinformatics toolkit / Okonechnikov K, Golosova O, Fursov M. [et al.] // *Bioinformatics*. – 2012. – V. 28. – №. 8. – P. 1166-1167.
298. Olejniczak, S. H. Distinct molecular mechanisms responsible for bortezomib-induced death of therapy-resistant versus-sensitive B-NHL cells / Olejniczak S.H, Blickwedehl J., Belicha-Villanueva A. [et al.] // *Blood, The Journal of the American Society of Hematology*. – 2010. – V. 116. – №. 25. – P. 5605-5614.
299. Olesen, J. PGC-1 α -mediated adaptations in skeletal muscle / Olesen J., Kiilerich K., Pilegaard H. // *Pflügers Archiv-European Journal of Physiology*. – 2010. – V. 460. – P. 153-162.
300. Olivier, B. G. SBML level 3 package: flux balance constraints version 2 / Olivier B. G., Bergmann F. T. // *Journal of Integrative Bioinformatics*. – 2018. – V. 15. – №. 1. – Art. 20170082.
301. Oparil, S. Efficacy, tolerability, and effects on quality of life of losartan, alone or with hydrochlorothiazide, versus amlodipine, alone or with hydrochlorothiazide, in patients with essential hypertension / Oparil S, Barr E, Elkins M. [et al.] // *Clinical Therapeutics*. – 1996. – V. 18. – №. 4. – P. 608-625.
302. Oparil, S. A double-blind, randomized study evaluating losartan potassium monotherapy or in

- combination with hydrochlorothiazide versus placebo in obese patients with hypertension / Oparil S, Abate N, Chen E. [et al.] // *Current Medical Research and Opinion*. – 2008. – V. 24. – №. 4. – P. 1101-1114.
303. Oparil, S. Hypertension / Oparil S., Acelajado M.C., Bakris G.L. [et al.] // *Nature Reviews Disease Primers*. – 2018. – V. 4. – Art. 18014.
304. Orth, J. D. What is flux balance analysis? / J. D. Orth, I. Thiele, B. Ø. Palsson // *Nature Biotechnology*. – 2010. – V. 28. – №. 3. – P. 245-248.
305. Oster, G. F. Mechanochemistry and morphogenesis / Oster G. F. // *Biological structures and coupled flows*. – 1983. – P. 417-443.
306. Panchiwala, H. The systems biology simulation core library / Panchiwala, H., Shah, S., Planatscher, H. [et al.] // *Bioinformatics*. – 2022. – V. 38. – №. 3. – P. 864-865.
307. Pareek, A. K. Efficacy of low-dose chlorthalidone and hydrochlorothiazide as assessed by 24-h ambulatory blood pressure monitoring / Pareek AK, Messerli FH, Chandurkar NB. [et al.] // *Journal of the American College of Cardiology*. – 2016. – V. 67. – №. 4. – P. 379-389.
308. Parker, M. T. Ascape: Abstracting complexity / Parker M. T. // *Natural Resources and Environmental Issues*. – 2001. – V. 8. – №. 1. – P. 5.
309. Paterna, S. Medium-term effects of bisoprolol administration on renal hemodynamics and function in mild to moderate essential hypertension / Paterna S, Parrinello G, Di Pasquale P. [et al.] // *Advances in Therapy*. – 2007. – V. 24. – №. 6. – P. 1260-1270.
310. Patterson M. D. Implementing Runge-Kutta solvers in Java : дисс. – Acadia University, 2003.
311. Pearen, M. A. The nuclear receptor, Nor-1, markedly increases type II oxidative muscle fibers and resistance to fatigue / Pearen M.A., Eriksson N.A., Fitzsimmons R.L. [et al.] // *Molecular Endocrinology*. – 2012. – V. 26. – №. 3. – P. 372-384.
312. Pearen, M. A. Transgenic muscle-specific Nor-1 expression regulates multiple pathways that effect adiposity, metabolism, and endurance / Pearen M.A., Goode J.M., Fitzsimmons R.L. [et al.] // *Molecular Endocrinology*. – 2013. – V. 27. – №. 11. – P. 1897-1917.
313. Peleg, M. Modelling biological processes using workflow and petri net models / Peleg M., Yeh I., Altman R. B. // *Bioinformatics*. – 2002. – V. 18. – №. 6. – P. 825-837.
314. Perez, L. An agent-based approach for modeling dynamics of contagious disease spread / Perez L, Dragicevic S. // *International Journal of Health Geographics*. – 2009. – V. 8. – №. 1. – P. 1-17.
315. Pettersen, E. F. UCSF Chimera—a visualization system for exploratory research and analysis / Pettersen E, Goddard T, Huang C. // *Journal of Computational Chemistry*. – 2004. – V. 25. – №. 13. – P. 1605-1612.
316. Petuhoff. Ударим ТАУ по пандемии COVID-19. Численное моделирования распространения инфекции / Petuhoff – Текст: электронный // Хабр: интернет-портал, 2020. – URL:

<https://habr.com/ru/post/506556/> (дата обращения: 04.07.2023)

317. Pinto, D. Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody / Pinto, D., Park, Y.J., Beltramello, M. [et al.] // *Nature*. – 2020. – V. 583. – №. 7815. – P. 290-295.
318. Pintus, S. Biostore: a cloud-compatible hub for bioinformatics related tools and platforms / Pintus S., Valeev T. F., Yevshin I. S. [et al.] // *Proceedings of The Tenth International Conference on Genome Regulation and Structure\Systems Biology (BGRS\SB-2016)*, Novosibirsk, August 29 – September 2, 2016. – Novosibirsk, 2016. – P. 237.
319. Pintus, S. S. Genome-wide atlas of promoter expression reveals contribution of transcribed regulatory elements to genetic control of disuse-mediated atrophy of skeletal muscle / S. S. Pintus, I. R. Akberdin, I. Yevshin. [et al.] // *Biology*. – 2021. – V. 10. – №. 6. – Art. 557.
320. Pollock, A. M. Asymptomatic transmission of covid-19 / Pollock A. M., Lancaster J. // *The BMJ*. – 2020. – T. 371. – Art. m4851.
321. Pool, J. L. Aliskiren, an orally effective renin inhibitor, provides antihypertensive efficacy alone and in combination with valsartan / Pool JL, Schmieder RE, Azizi M. [et al.] // *American Journal of Hypertension*. – 2007. – V. 20. – №. 1. – P. 11-20.
322. Popov, D. V. Adaptation of skeletal muscle to contractile activity of various duration and intensity: role of PGC-1 α / Popov D.V. // *Biochemistry (Moscow)*. – 2018. – V. 83. – №. 6. – P. 613-628.
323. Popov, D. V. Contractile activity-specific transcriptome response to acute endurance exercise and training in human skeletal muscle / Popov D.V., Makhnovskii P.A., Shagimardanova E.I. [et al.] // *American Journal of Physiology-Endocrinology and Metabolism*. – 2019. – V. 316. – №. 4. – P. E605-E614.
324. Poroikov, V. PASS: identification of probable targets and mechanisms of toxicity / Poroikov, V., Filimonov, D., Lagunin, A. [et al.] // *SAR and QSAR in Environmental Research*. – 2007. – V. 18. – №. 1-2. – P. 101-110.
325. Porthan, K. Short-term electrophysiological effects of losartan, bisoprolol, amlodipine, and hydrochlorothiazide in hypertensive men / Porthan K, Viitasalo M, Hiltunen TP. [et al.] // *Annals of Medicine*. – 2009. – V. 41. – №. 1. – P. 29-37.
326. Porubsky, V. L. Best practices for making reproducible biochemical models / Porubsky VL, Goldberg AP, Rampadarath AK. [et al.] // *Cell Systems*. – 2020. – V. 11. – №. 2. – P. 109-120.
327. Powers, R. NMR metabolomics and drug discovery / Powers, R. // *Magnetic Resonance in Chemistry*. – 2009. – V. 47. – №. S1. – P. S2-S11.
328. Pratt, H. E. Factorbook: an updated catalog of transcription factor motifs and candidate regulatory motif sites / Pratt, H. E., Andrews, G. R., Phalke, N. [et al.] // *Nucleic Acids Research*. – 2022. – V. 50. – №. D1. – P. D141-D149.
329. Price, N. D. Genome-scale microbial in silico models: the constraints-based approach / Price ND,

- Papin JA, Schilling CH. [et al.] // *Trends in Biotechnology*. – 2003. – V. 21. – №. 4. – P. 162-169.
- 330.Prokudina, E. I., Valeev R. Y., Tchuraev R. N. A new method for the analysis of the dynamics of the molecular genetic control systems. II. Application of the method of generalized threshold models in the investigation of concrete genetic systems / Prokudina E. I., Valeev R. Y., Tchuraev R. N. // *Journal of Theoretical Biology*. – 1991. – V. 151. – №. 1. – P. 89-110.
- 331.Puchałka, J. Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks / Puchałka J., Kierzek A. M. // *Biophysical Journal*. – 2004. – V. 86. – №. 3. – P. 1357-1372.
- 332.Puig, J. G. Lercanidipine, enalapril and their combination in the treatment of elderly hypertensive patients: placebo-controlled, randomized, crossover study with four ABPM / Puig JG, Calvo C, Luurila O. [et al.] // *Journal of Human Hypertension*. – 2007. – V. 21. – №. 12. – P. 917-924.
- 333.Quinlan, A. R., Hall I. M. BEDTools: a flexible suite of utilities for comparing genomic features / Quinlan A. R., Hall I. M. // *Bioinformatics*. – 2010. – V. 26. – №. 6. – P. 841-842.
- 334.Rangamani, P. Survival and apoptotic pathways initiated by TNF- α : Modeling and predictions / Rangamani P., Sirovich L. // *Biotechnology and Bioengineering*. – 2007. – V. 97. – №. 5. – P. 1216-1229.
- 335.Rapoport, R. M. Mechanism of thiazide diuretic arterial pressure reduction: the search continues / Rapoport R. M., Soleimani M. // *Frontiers in Pharmacology*. – 2019. – V. 10. – P. 815.
- 336.Raue A. Identifiability and observability analysis for experimental design in nonlinear dynamical models / Raue A, Becker V, Klingmüller U. [et al.] // *Chaos: An Interdisciplinary Journal of Nonlinear Science*. – 2010. – V. 20. – №. 4. – Art. 045105.
- 337.Reddy, V. N. Petri net representations in metabolic pathways / Reddy VN, Mavrovouniotis ML, Liebman MN. // *Proceedings. International Conference on Intelligent Systems for Molecular Biology*. – 1993. – V. 93. – P. 328-336.
- 338.Reguant, R. AlignmentViewer: sequence analysis of large protein families / Reguant R, Antipin Y, Sheridan R. [et al.] // *F1000Research*. – 2020. – V. 9. – Art. ISCB Comm J-213.
- 339.Reimel, B. A. Proteomics on fixed tissue specimens—a review / Reimel B.A., Pan S., May D.H. [et al.] // *Current proteomics*. – 2009. – V. 6. – №. 1. – P. 63-69.
- 340.Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes / Roadmap Epigenomics Consortium // *Nature*. – 2015. – V. 518. – №. 7539. – P. 317-330.
- 341.Rohn, H. VANTED v2: a framework for systems biology applications / Rohn H, Junker A, Hartmann A. [et al.] // *BMC Systems Biology*. – 2012. – V. 6. – P. 1-13.
- 342.Roussel, M. Metabolic determinants of the onset of acidosis in exercising human muscle: a ³¹P-MRS study / Roussel M, Mattei JP, Le Fur Y. [et al.] // *Journal of Applied Physiology*. – 2003. – V. 94. – №. 3. – P. 1145-1152.

343. Runarsson, T. P. Stochastic ranking for constrained evolutionary optimization / Runarsson T. P., Yao X. // *IEEE Transactions on evolutionary computation*. – 2000. – V. 4. – №. 3. – P. 284-294.
344. Sánchez, L. Establishment of the Dorso-ventral Pattern During Embryonic Development of *Drosophila melanogaster*: a Logical Analysis / Sánchez L., van Helden J., Thieffry D. // *Journal of Theoretical Biology*. – 1997. – V. 189. – №. 4. – P. 377-389.
345. Sanz Leon, P. The Virtual Brain: a simulator of primate brain network dynamics / Sanz Leon, P., Knock, S. A., Woodman, M. M. // *Frontiers in Neuroinformatics*. – 2013. – V. 7. – P. 10.
346. Savageau, M. A. Reconstructionist molecular biology / Savageau, M. A. // *The New Biologist*. – 1991. – V. 3. – №. 2. – P. 190-197.
347. Scaglione, R. Antihypertensive efficacy and effects of nitrendipine on cardiac and renal hemodynamics in mild to moderate hypertensive patients: randomized controlled trial versus hydrochlorothiazide / Scaglione R, Indovina A, Parrinello G. // *Cardiovascular Drugs and Therapy*. – 1992. – V. 6. – P. 141-146.
348. Scaglione, R. Effects of cilazapril on renal haemodynamics and function in hypertensive patients: a randomised controlled study versus hydrochlorothiazide / Scaglione R, Ganguzza A, Corrao S. [et al.] // *Blood Pressure*. – 1995. – V. 4. – №. 6. – P. 363-368.
349. Scalbert, A. Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research / Scalbert A., Brennan L., Fiehn O. [et al.] // *Metabolomics*. – 2009. – V. 5. – P. 435-458.
350. Scarpulla, R. C. Transcriptional paradigms in mammalian mitochondrial biogenesis and function / Scarpulla R. C. // *Physiological Reviews*. – 2008. – V. 88. – №. 2. – P. 611-638.
351. Schoeberl, B. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors / Schoeberl B., Eichler-Jonsson C., Gilles E.D. [et al.] // *Nature Biotechnology*. – 2002. – V. 20. – №. 4. – P. 370-375.
352. Schreiber, F. Specifications of standards in systems and synthetic biology: status and developments in 2021 / Schreiber F, Gleeson P, Golebiewski M. [et al.] // *Journal of Integrative Bioinformatics*. – 2021. – V. 18. – №. 3. – Art. 20210026.
353. Schwanhäusser, B. Global quantification of mammalian gene expression control / Schwanhäusser, B., Busse, D., Li, N. [et al.] // *Nature*. – 2011. – V. 473. – №. 7347. – P. 337–342.
354. Semisalov, B. V. BioUML: plugin for stochastic modeling of biological systems / Semisalov, B. V., Kiselev, I. N., Sharipov, R. N. [et al.] // *Proceedings of the 8th International Conference on Bioinformatics of Genome Regulation and Structure / Systems Biology (BGRS/SB'2012)*, June 25-29, 2012. – Novosibirsk, 2012. – P. 283.
355. Seppänen-Laakso, T. How to study lipidomes / Seppänen-Laakso T., Orešič M. // *Journal of Molecular Endocrinology*. – 2009. – V. 42. – №. 3. – P. 185-190.

356. Sethi, A. A. Angiotensinogen gene polymorphism, plasma angiotensinogen, and risk of hypertension and ischemic heart disease: a meta-analysis / Sethi A.A., Nordestgaard B.G., Tybjaerg-Hansen A. // *Arteriosclerosis, Thrombosis, and Vascular biology*. – 2003. – V. 23. – №. 7. – P. 1269-1275.
357. Shah, S. Mechanism of antihypertensive effect of thiazide diuretics / Shah S., Khatri I., Freis E. D. // *American Heart Journal*. – 1978. – V. 95. – №. 5. – P. 611-618.
358. Shaikh, B. BioSimulators: a central registry of simulation engines and services for recommending specific tools / Shaikh, B., Smith, L. P., Vasilescu, D. [et al.] // *Nucleic Acids Research*. – 2022. – V. 50. – №. W1. – P. W108-W114.
359. Shank, S. D., Weaver S., Kosakovsky Pond S. L. phylotree. js-a JavaScript library for application development and interactive data visualization in phylogenetics / Shank S. D., Weaver S., Kosakovsky Pond S. L. // *BMC bioinformatics*. – 2018. – V. 19. – P. 1-5.
360. Sharipov, R. N. Assessment of transcriptional importance of cell line-specific features based on GTRD and FANTOM5 data / Sharipov R.N., Kondrakhin Y.V., Ryabova A.S. [et al.] // *PLoS One*. – 2020. – V. 15. – №. 12. – P. e0243332.
361. Shmulevich, I. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks / Shmulevich I., Dougherty E.R., Kim S. [et al.] // *Bioinformatics*. – 2002a. – V. 18. – №. 2. – P. 261-274.
362. Shmulevich, I. Gene perturbation and intervention in probabilistic Boolean networks / Shmulevich I., Dougherty E. R., Zhang W. // *Bioinformatics*. – 2002b. – V. 18. – №. 10. – P. 1319-1331.
363. Shmulevich, I. Activities and sensitivities in Boolean network models / Shmulevich I., Kauffman S. A. // *Physical Review Letters*. – 2004. – V. 93. – №. 4. – P. 048701.
364. Shrock, E. Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity / Shrock, E., Fujimura, E., Kula, T. [et al.] // *Science*. – 2020. – V. 370. – №. 6520. – Art. eabd4250.
365. Sica, D. A. Clinical pharmacokinetics of losartan / Sica D. A., Gehr T. W. B., Ghosh S. // *Clinical Pharmacokinetics*. – 2005. – V. 44. – P. 797-814.
366. Siebert, M. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences / Siebert M., Söding J. // *Nucleic Acids Research*. – 2016. – V. 44. – №. 13. – P. 6055-6069.
367. Sierra, M. R. Improving PSO-based multi-objective optimization using crowding, mutation and ϵ -dominance / Sierra M. R., Coello Coello C. A. // *International conference on evolutionary multi-criterion optimization*. – Berlin, Heidelberg : Springer Berlin Heidelberg, 2005. – P. 505-519.
368. Singh O. V. Transcriptomics, proteomics and interactomics: unique approaches to track the

- insights of bioremediation / Singh O. V., Nagaraj N. S. // *Briefings in Functional Genomics*. – 2006. – V. 4. – №. 4. – P. 355-362.
369. Smith, L. P. Antimony: a modular model definition language / Smith, L. P., Bergmann, F. T., Chandran, D. [et al.] // *Bioinformatics*. – 2009. – V. 25. – №. 18. – P. 2452-2454.
370. Smith, L. P. Hierarchical Model Composition, Version 1 Release 2 / Smith L. P., Hucka M., Hoops S. [et al.] – Текст: электронный // COMBINE: интернет-портал, 2013. – URL: <https://identifiers.org/combine.specifications:sbml.level-3.version-1.comp.version-1.release-32013> (дата обращения: 04.07.2023)
371. Smith, L. P. SBML level 3 package: hierarchical model composition, version 1 release 3 / Smith LP, Hucka M, Hoops S. [et al.] // *Journal of Integrative Bioinformatics*. – 2015. – V. 12. – №. 2. – P. 603-659.
372. Smolen, P. Modeling transcriptional control in gene networks—methods, recent results, and future directions / Smolen P., Baxter D. A., Byrne J. H. // *Bulletin of Mathematical Biology*. – 2000. – V. 62. – №. 2. – P. 247-292.
373. Somogyi, R. Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation / Somogyi R., Sniegoski C. A. // *Complexity*. – 1996. – V. 1. – №. 6. – P. 45-63.
374. Spear, B. B. Clinical application of pharmacogenetics / Spear B. B., Heath-Chiozzi M., Huff J. // *Trends in Molecular Medicine*. – 2001. – V. 7. – №. 5. – P. 201-204.
375. Stadler, R. N. Systematic screening on admission for SARS-CoV-2 to detect asymptomatic infections / Stadler, R. N., Maurer, L., Aguilar-Bultet, L. [et al.] // *Antimicrobial Resistance & Infection Control*. – 2021. – V. 10. – №. 1. – Art. 44.
376. Steinbeck, C. Recent developments of the chemistry development kit (CDK)-an open-source java library for chemo-and bioinformatics / Steinbeck C, Hoppe C, Kuhn S. [et al.] // *Current Pharmaceutical Design*. – 2006. – V. 12. – №. 17. – P. 2111-2120.
377. Steinmann, E. Acute hemodynamic effects of bisoprolol, a new beta 1 selective adrenoreceptor blocking agent, in patients with coronary artery disease / Steinmann E., Pfisterer M., Burkart F. // *Journal of Cardiovascular Pharmacology*. – 1986. – V. 8. – №. 5. – P. 1044-1050.
378. Stevens, J. T. Dynamic modeling of cellular populations within iBioSim / Stevens J. T., Myers C. J. // *ACS Synthetic Biology*. – 2013. – V. 2. – №. 5. – P. 223-229.
379. Sun, F. Efficacy of losartan for improving insulin resistance and vascular remodeling in hemodialysis patients / Sun F, Song Y, Liu J. [et al.] // *Hemodialysis International*. – 2016. – V. 20. – №. 1. – P. 22-30.
380. Sütterlin, T. Bridging the scales: semantic integration of quantitative SBML in graphical multi-cellular models and simulations with EPISIM and COPASI / Sütterlin T, Kolb C, Dickhaus

- H. [et al.] // *Bioinformatics*. – 2013. – V. 29. – №. 2. – P. 223-229.
381. Taylor, J. G. On the circuit complexity of sigmoid feedforward neural networks / Taylor J.G., Beiu V. // *Neural Networks*. – 1996. – V. 9. – №. 7. – P. 1155-1171.
382. Tchuraev, R. N. A new method for the analysis of the dynamics of the molecular genetic control systems. I. Description of the method of generalized threshold models / Tchuraev R.N. // *Journal of Theoretical Biology*. – 1991. – V. 151. – №. 1. – P. 71-87.
383. Tchuraev, R. N. Parametric stability evaluation in computer experiments on the mathematical model of *Drosophila* control gene subnetwork / Tchuraev R. N., Galimzyanov A. V. // *In Silico Biology*. – 2003. – V. 3. – №. 1-2. – P. 101-115.
384. Tham, T. C. Haemodynamic comparison of amlodipine and atenolol in essential hypertension using the quantascope / Tham T.C., Herity N., Guy S. [et al.] // *British Journal of Clinical Pharmacology*. – 1993. – V. 36. – №. 6. – P. 555.
385. Thieffry, D. Dynamical behaviour of biological regulatory networks—II. Immunity control in bacteriophage lambda / Thieffry D., Thomas R. // *Bulletin of Mathematical Biology*. – 1995. – V. 57. – P. 277-297.
386. Thomas, R. *Biological feedback* / Thomas R., d'Ari R. – CRC Press, Boca Raton, FL, 1990. – 328 pp. – ISBN 9780849367663.
387. Thomas, R. Dynamical behaviour of biological regulatory networks—I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state / Thomas R., Thieffry D., Kaufman M. // *Bulletin of Mathematical Biology*. – 1995. – V. 57. – P. 247-276.
388. Thomas, R. Features that define the best ChIP-seq peak calling algorithms / Thomas, R., Thomas, S., Holloway, A. K. [et al.] // *Briefings in Bioinformatics*. – 2017. – V. 18. – №. 3. – P. 441-450.
389. Thomas, S. R. SAPHIR: a physiome core model of body fluid homeostasis and blood pressure regulation / Thomas S.R., Baconnier P., Fontecave J. [et al.] // *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. – 2008. – V. 366. – №. 1878. – P. 3175-3197.
390. Tiwari, K. Reproducibility in systems biology modelling / Tiwari K, Kananathan S, Roberts MG. [et al.] // *Molecular Systems Biology*. – 2021. – V. 17. – №. 2. – Art. e9982.
391. Toivonen, H. T. Modeling reveals that dynamic regulation of c-FLIP levels determines cell-to-cell distribution of CD95-mediated apoptosis / Toivonen HT, Meinander A, Asaoka T. [et al.] // *Journal of Biological Chemistry*. – 2011. – V. 286. – №. 21. – P. 18375-18382.
392. Tucker, T. Massively parallel sequencing: the next big thing in genetic medicine / Tucker T., Marra M., Friedman J. M. // *The American Journal of Human Genetics*. – 2009. – V. 85. – №. 2. – P. 142-154.
393. Tuomisto J. T. An agent-based epidemic model REINA for COVID-19 to identify destructive

- policies / Tuomisto, J. T., Yrjölä, J., Kolehmainen, M. [et al.] // MedRxiv. – 2020. – Art. 2020.04.09.20047498.
394. Uttamsingh, R. J. Mathematical model of the human renal system / Uttamsingh R. J., Leaning M. S., Bushman J. A. [et al.] // Medical and Biological Engineering and Computing. – 1985. – V. 23. – P. 525-535.
395. Valeev, T. BioUML genome browser / Valeev T., Yevshin I., Kolpakov F. // Virtual Biology. – 2013. – V. 1. – №. 1. – P. 15-26.
396. van Brummelen, P. Hemodynamic changes during long-term thiazide treatment of essential hypertension in responders and nonresponders / van Brummelen P., Man In't Veld A. J., Schalekamp M. A. D. H. // Clinical Pharmacology & Therapeutics. – 1980. – V. 27. – №. 3. – P. 328-336.
397. Vangheluwe, H. L. M. DEVS as a common denominator for multi-formalism hybrid systems modelling / Vangheluwe H. L. M. // CACSD. Conference Proceedings. IEEE international symposium on computer-aided control system design (cat. no. 00th8537). – IEEE, 2000. – P. 129-134.
398. Vespoli, L. Where Watson went wrong / Vespoli L. – Текст: электронный // Medical marketing and media: интернет-портал, 2021. URL: <https://www.mmm-online.com/home/channel/features/where-watson-went-wrong/> (дата обращения: 04.07.2023).
399. Vieira, L. S. Computational models in systems biology: standards, dissemination, and best practices / Vieira L. S., Laubenbacher R. C. // Current Opinion in Biotechnology. – 2022. – V. 75. – Art. 102702.
400. Villamil, A. Renin inhibition with aliskiren provides additive antihypertensive efficacy when used in combination with hydrochlorothiazide / Villamil A, Chrysant SG, Calhoun D. [et al.] // Journal of Hypertension. – 2007. – V. 25. – №. 1. – P. 217-226.
401. Volkenstein, M. V. Allosteric enzyme models and their analysis by the theory of graphs / Volkenstein M. V., Goldstein B. N. // Biochimica et Biophysica Acta (BBA)-General Subjects. – 1966. – V. 115. – №. 2. – P. 478-485.
402. Von Bertalanffy, L. The theory of open systems in physics and biology / Von Bertalanffy L. // Science. – 1950. – V. 111. – №. 2872. – P. 23-29.
403. Von Bertalanffy, L. General System Theory - A critical review / Von Bertalanffy L. // General Systems. – 1962. – V. 7. – P. 1-20.
404. Von Bertalanffy, L. General System Theory: A new approach to unity of science / Von Bertalanffy, L., Hempel, C. G., Bass, R. E. [et al.] // Human Biology. – 1951. – V. 23. – №. 4. – P. 302-361.
405. Vorontsov, I. E. Genome-wide map of human and mouse transcription factor binding sites

- aggregated from ChIP-Seq data / Vorontsov I.E., Fedorova A.D., Yevshin I.S. [et al.] // *BMC Research Notes*. – 2018. – V. 11. – Art.756.
406. Waleev, T. Composite Module Analyst: identification of transcription factor binding site combinations using genetic algorithm / Waleev T, Shtokalo D, Konovalova T. [et al.] // *Nucleic Acids Research*. – 2006. – V. 34. – №. suppl_2. – P. W541-W545.
407. Waltemath, D. Minimum information about a simulation experiment (MIASE) / Waltemath D, Adams R, Beard DA. [et al.] // *PLoS Computational Biology*. – 2011a. – V. 7. – №. 4. – Art. e1001122.
408. Waltemath, D. Reproducible computational biology experiments with SED-ML-the simulation experiment description markup language / Waltemath D, Adams R, Bergmann FT. [et al.] // *BMC Systems Biology*. – 2011b. – V. 5. – №. 1. – P. 1-10.
409. Waltemath, D. The first 10 years of the international coordination network for standards in systems and synthetic biology (COMBINE) / Waltemath, D, Golebiewski, M, Blinov, ML. [et al.] // *Journal of Integrative Bioinformatics*. – 2020. – V. 17. – №. 2-3. – Art. 20200005.
410. Watanabe, L. iBioSim 3: a tool for model-based genetic circuit design / Watanabe, L., Nguyen, T., Zhang, M. [et al.] // *ACS Synthetic Biology*. – 2018. – V. 8. – №. 7. – P. 1560-1563.
411. Weber, M. A. Blood pressure effects of the angiotensin II receptor blocker, losartan / Weber MA, Byyny RL, Pratt JH. [et al.] // *Archives of Internal Medicine*. – 1995. – V. 155. – №. 4. – P. 405-411.
412. Weins, D. Development of simulation model of HPC system for Super Charm-Tau factory / D. Weins, V. Vorobyev, I. Chernykh [et al.] // *Journal of Physics: Conference Series*. – IOP Publishing, 2019. – V. 1336. – №. 1. – Art. 012025.
413. Weirauch, M. T. Determination and inference of eukaryotic transcription factor sequence specificity / Weirauch, M. T., Yang, A., Albu, M. [et al.] // *Cell*. – 2014. – V. 158. – №. 6. – P. 1431-1443.
414. Weisbuch, G. Networks of automata and biological organization / Weisbuch G. // *Journal of Theoretical Biology*. – 1986. – V. 121. – №. 3. – P. 255-267.
415. Welsh, C. libRoadRunner 2.0: a high performance SBML simulation and analysis library / Welsh C., Xu J., Smith L. [et al.] // *Bioinformatics*. – 2023. – V. 39. – №. 1. – Art. btac770.
416. Wenger, A. M. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome / Wenger, A. M., Peluso, P., Rowell, W. J. [et al.] // *Nature Biotechnology*. – 2019. – V. 37. – №. 10. – P. 1155-1162.
417. Westerhoff H. V. Advice from a systems-biology model of the Corona epidemics / Westerhoff H. V., Kolodkin A. N. // *npj Systems Biology and Applications*. – 2020. – V. 6. – Art. 18.
418. WHO Situation Report 6 March / – Текст: электронный // Reliefweb : интернет-портал, 2020.

URL:

<https://reliefweb.int/report/china/coronavirus-disease-2019-covid-19-situation-report-46-6-march-2020> (дата обращения: 24.07.2023).

419. Wiechert, W. The thermodynamic meaning of metabolic exchange fluxes / Wiechert W. // *Biophysical Journal*. – 2007. – V. 93. – №. 6. – P. 2255-2264.
420. Wilensky, U. An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo / Wilensky U., Rand W. – MIT Press, 2015. – 504 pp. – ISBN 0262731894.
421. Wilkinson, M. D. The FAIR Guiding Principles for scientific data management and stewardship / Wilkinson M.D., Dumontier M., Aalbersberg I.J. [et al.] // *Scientific Data*. – 2016. – V. 3. – №. 1. – P. 1-9.
422. Wing, L. Candesartan and hydrochlorothiazide in isolated systolic hypertension / Wing LMH, Arnolda LF, Upton J. [et al.] // *Blood Pressure*. – 2003. – V. 12. – №. 4. – P. 246-254.
423. Wingender, E. TFClass: expanding the classification of human transcription factors to their mammalian orthologs / Wingender E, Schoeps T, Haubrock M. [et al.] // *Nucleic Acids Research*. – 2018. – V. 46. – №. D1. – P. D343-D347.
424. Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation / Wingender E. // *Briefings in Bioinformatics*. – 2008. – V. 9. – №. 4. – P. 326-332.
425. Wooldridge, M. An introduction to multiagent systems / M. Woolridge – John Wiley & Sons, 2009. – 488 pp. – ISBN 978-0-470-51946-2.
426. Yanes-Lane, M. Proportion of asymptomatic infection among COVID-19 positive persons and their transmission potential: A systematic review and meta-analysis / Yanes-Lane, M., Winters, N., Fregonese, F. [et al.] // *PLoS One*. – 2020. – V. 15. – №. 11. – Art. e0241536.
427. Yevshin, I. S. BioUML – software platform for analysis of Next Generation Sequencing data using collaborative and reproducible research / Yevshin I.S., Valeev T.F., Sharipov R.N. [et al.] // *Proceedings. International conference “High-throughput sequencing in genomes – 2013”*: Novosibirsk, July 21-25, 2013. – Novosibirsk: Institute of Chemical Biology and Fundamental Medicine, 2013. – P. 66.
428. Yevshin, I. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments / Yevshin I., Sharipov R., Valeev T. [et al.] // *Nucleic Acids Research* – 2017. – V. 45. – №. D1. – P. D61-D67.
429. Yevshin, I. GTRD: a database on gene transcription regulation—2019 update / Yevshin I., Sharipov R., Kolmykov S. [et al.] // *Nucleic Acids Research*. – 2019. – V. 47. – №. D1. – P. D100-D105.

430. Yoshioka, T. Identification and characterization of an alternative promoter of the human PGC-1 α gene / Yoshioka T., Inagaki K., Noguchi T. [et al.] // Biochemical and Biophysical Research Communications. – 2009. – V. 381. – №. 4. – P. 537-543.
431. Yuen, D. The Dockstore: enhancing a community platform for sharing reproducible and accessible computational protocols / Yuen D, Cabansay L, Duncan A. [et al.] // Nucleic Acids Research. – 2021. – V. 49. – №. W1. – P. W624-W632.
432. Zaia, J. Mass spectrometry and the emerging field of glycomics / Zaia J. // Chemistry & Biology. – 2008. – V. 15. – №. 9. – P. 881-892.
433. Zelnerman, D. Robust estimation in truncated discrete distributions with application to capture-recapture experiments / Zelnerman D. // Journal of Statistical Planning and Inference. – 1988. – V. 18. – №. 2. – P. 225-237.
434. Zhadan, A. AURORA software package for parametric simulation and feasibility study / Zhadan A., Suharev A., Zhadan D. [et al.] – Текст: электронный // Workshop on future Super c -tau factories 2021: Novosibirsk, November 15-17, 2013. – URL: <https://indico.inp.nsk.su/event/62/contributions/2280/> (дата обращения: 04.07.2023).
435. Zhang, X. PICS: probabilistic inference for ChIP-seq / Zhang X., Robertson G., Krzywinski M. [et al.] // Biometrics. – 2011. – V. 67. – №. 1. – P. 151-163.
436. Zhang, Y. Model-based analysis of ChIP-Seq (MACS) / Zhang Y, Liu T, Meyer CA. [et al.] // Genome Biology. – 2008. – V. 9. – №. 9. – Art. R137.
437. Zheng, R. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis / Zheng, R., Wan, C., Mei, S. [et al.] // Nucleic Acids Research. – 2019. – V. 47. – №. D1. – P. D729-D735.
438. Zou, Z. ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data / Zou, Z., Ohta, T., Miura, F. [et al.] // Nucleic Acids Research. – 2022. – V. 50. – №. W1. – P. W175-W182.

СВИДЕТЕЛЬСТВА О РЕГИСТРАЦИИ ПРОГРАММ ДЛЯ ЭВМ

1. **Колпаков Ф.А.**, Ананько Е.А., Колчанов Н.А., Колесов Г.Б., Подколотная О.А., Игнатьева Е.В. БД GeneNet. РОСПАТЕНТ, св-во № 990006 от 15.02.1999 г.
2. **Колпаков Ф.А.** BioUML workbench (BioUML). Свидетельство о государственной регистрации программы для ЭВМ № 2008611606 от 27.03.2008 г.
3. **Колпаков Ф.А.** BioUML Enterprise Edition. Свидетельство о государственной регистрации программы для ЭВМ № 2008612975 от 19.06.2008 г.
4. **Колпаков Ф.А.** Модульная математическая модель регуляции липидного метаболизма на уровне организма человека. Свидетельство о государственной регистрации программы для ЭВМ № 2015618214 от 03.08.2015 г.
5. Евшин И.С., **Колпаков Ф.А.**, Шарипов Р.Н. GTRD – gene transcription regulation database. Свидетельство о государственной регистрации БД № 2017620872 от 10.08.2017 г.
6. Кель А.Э., **Колпаков Ф.А.**, Шарипов Р.Н., Киселев И.Н., Стельмашенко Д.Е., Мандрик Н.В., Евшин И.С., Пинтус С.С. Программа «От генома к мишени» // Свидетельство о государственной регистрации программы для ЭВМ № 2017616241 от 02.06.2017 г.
7. Кель А.Э., Киселев И.Н., **Колпаков Ф.А.**, Пинтус С.С. Модуль графической аннотации экспериментальных данных. Свидетельство о государственной регистрации программы для ЭВМ № 2017615715 от 15.06.2017.

СПИСОК ОСНОВНЫХ ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

Журналы из списка Web of Science и/или Scopus, Q1-Q2

1. **Kolpakov F.**, Akberdin I.R., Kiselev I.N., Kolmykov S.K., Kondrakhin Y., Kulyashov M., Kutumova E.O., Pintus S.S., Ryabova A., Sharipov R.N., Yevshin I.S., Zhatchenko S., Kel A. BioUML - towards a universal research platform // *Nucleic Acids Research* – 2022. – V.50(W1). – P.W124-W131.
2. Kutumova E., Kiselev I., Sharipov R., Lifshits G., **Kolpakov F.** Mathematical modeling of antihypertensive therapy // *Frontiers in Physiology* – 2022. – V.13. – Art.1070115.
3. Boytsov A., Abramov S., Aiusheeva A., Kasianova A., Baulin E., Kuznetsov I., Aulchenko Y., Kolmykov S.K., Yevshin I.S., **Kolpakov F.**, Vorontsov I.E., Makeev V.J., Kulakovskiy I.V. ANANASTRA: annotation and enrichment analysis of allele-specific transcription factor binding at SNPs // *Nucleic Acids Research* – 2022. – V.50(W1). –P.W51-W56.
4. Kutumova E.O., Akberdin I.R., Kiselev I.N., Sharipov R.N., Egorova V.S., Syrocheva A.O., Parodi A., Zamyatnin A.A., **Kolpakov F.A.** Physiologically Based Pharmacokinetic Modeling of Nanoparticle Biodistribution: A Review of Existing Models, Simulation Software, and Data Analysis Tools // *International Journal of Molecular Sciences* – 2022. – V.23(20). –Art.12560.
5. Afonyushkin V., Akberdin I.R., Kozlova Y., Schukin I., Mironova T., Bobikova A., Cherepushkina V., Donchenko N., Poletaeva Y., **Kolpakov F.A.** Multicompartmental Mathematical Model of SARS-CoV-2 Distribution in Human Organs and Their Treatment // *Mathematics* – 2022. – V.10(11). – Art.1925.
6. Kutumova E.O., Akberdin I.R., Kiselev I.N., Sharipov R.N., **Kolpakov F.A.** Modular Representation of Physiologically Based Pharmacokinetic Models: Nanoparticle Delivery to Solid Tumors in Mice as an Example // *Mathematics* – 2022. – V.10(7). – Art.1176.
7. Makhnovskii P.A., Gusev O.A., Bokov R.O., Gazizova G.R., Vepkhvadze T., Lysenko E.A., Vinogradova O.L., **Kolpakov F.A.**, Popov D.V. Alternative transcription start sites contribute to acute-stress-induced transcriptome response in human skeletal muscle // *Human Genomics* – 2022. – V.16(1). –Art.24.
8. Abramov S., Boytsov A., Bykova D., Penzar D.D., Yevshin I., Kolmykov S.K., Fridman M.V., Favorov A.V., Vorontsov I.E., Baulin E., **Kolpakov F.A.**, Makeev V.J., Kulakovskiy I.V. Landscape of allele-specific transcription factor binding in the human genome // *Nature Communications* – 2021. – V.12 – Art.2751.
9. Akberdin I.R., Kiselev I.N., Pintus S.S., Sharipov R.N., Vertyshev A.Y., Vinogradova O.L., Popov D.V., **Kolpakov F.A.** A modular mathematical model of exercise-induced changes in metabolism, signaling, and gene expression in human skeletal muscle // *International Journal of Molecular Sciences* – 2021. – V.22(19). – Art.10353.

10. Kolmykov S., Yevshin I.S., Kulyashov M., Sharipov R.N., Kondrakhin Y., Makeev V.J., Kulakovskiy I.V., Kel A., **Kolpakov F.** GTRD: an integrated view of transcription regulation // *Nucleic Acids Research* – 2021. – V.49(D1). – P.D104-D111.
11. Kutumova E., Kiselev I., Sharipov R., Lifshits G., **Kolpakov F.** Thoroughly Calibrated Modular Agent-Based Model of the Human Cardiovascular and Renal Systems for Blood Pressure Regulation in Health and Disease // *Frontiers in Physiology* – 2021. – V.12 – Art.746300.
12. Makhnovskii P.A., Bokov R.O., **Kolpakov F.A.**, Popov D.V. Transcriptomic Signatures and Upstream Regulation in Human Skeletal Muscle Adapted to Disuse and Aerobic Exercise // *International Journal of Molecular Sciences* – 2021. – V.22(3). – Art.1208.
13. Pintus S.S., Akberdin I.R., Yevshin I.S., Makhnovskii P.A., Tyapkina O., Nigmatzyanov I., Nurullin L., Devyatyyarov R., Shagimardanova E., Popov D., **Kolpakov F.A.**, Gusev O., Gazizova G.R. Genome-Wide Atlas of Promoter Expression Reveals Contribution of Transcribed Regulatory Elements to Genetic Control of Disuse-Mediated Atrophy of Skeletal Muscle // *Biology (Basel)* – 2021. – V.10(6). – Art.557.
14. Makhnovskii P.A., Zgoda V.G., Bokov R.O., Shagimardanova E.I., Gazizova G.R., Gusev O.A., Lysenko E.A., **Kolpakov F.A.**, Vinogradova O.L., Popov D.V. Regulation of Proteins in Human Skeletal Muscle: The Role of Transcription // *Scientific Reports* – 2020. – V.10(1). – Art.3514.
15. Sharipov R.N., Kondrakhin Y.V., Ryabova A.S., Yevshin I.S., **Kolpakov F.A.** Assessment of transcriptional importance of cell line-specific features based on GTRD and FANTOM5 data // *PloS ONE* – 2020. – V.15(12). – Art.e0243332.
16. Keating S.M., Waltemath D., et al. SBML Level 3: an extensible format for the exchange and reuse of biological models // *Molecular Systems Biology* – 2020. – V.16. – Art.e9110.
17. Kel A., Boyarskikh U., Stegmaier P., Leskov L.S., Sokolov A.V., Yevshin I.S., Mandrik N., Stelmashenko D., Koschmann J., Kel-Margoulis O., Krull M., Martínez-Cardús A., Moran S., Esteller M., **Kolpakov F.**, Filipenko M., Wingender E. Walking pathways with positive feedback loops reveal DNA methylation biomarkers of colorectal cancer // *BMC Bioinformatics* – 2019. – V.20. – Art.119.
18. Kolmykov S.K., Kondrakhin Y.V., Yevshin I.S., Sharipov R.N., Ryabova A.S., **Kolpakov F.A.** Population size estimation for quality control of ChIP-Seq datasets // *PloS ONE* – 2019. – V.14(8). – Art.e0221760.
19. **Kolpakov F.**, Akberdin I., Kashapov T., Kiselev I., Kolmykov S., Kondrakhin Y., Kutumova E., Mandrik N., Pintus S., Ryabova A., Sharipov R.N., Yevshin I., Kel A. BioUML: an integrated environment for systems biology and collaborative analysis of biomedical data // *Nucleic Acids Research* – 2019. – V.47(W1). – P.W225-W233.
20. Yevshin I., Sharipov R., Kolmykov S., Kondrakhin Y., **Kolpakov F.** GTRD: a database on gene

- transcription regulation-2019 update // *Nucleic Acids Research* – 2018. – V. 47(D1). – P. D100-D105.
21. Kel A., Boyarskikh U., Stegmaier P., Leskov L.S., Sokolov A.V., Yevshin I., Mandrik N., Stelmashenko D., Koschmann J., Kel-Margoulis O., Krull M., Martínez-Cardús A., Moran S., Esteller M., **Kolpakov F.**, Filipenko M., Wingender E. Walking pathways with positive feedback loops reveal DNA methylation biomarkers of colorectal cancer // *BMC Bioinformatics* – 2019. –V. 20(Suppl 4). – Art.119.
 22. Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., **Kolpakov, F.A.**, Makeev, V.J. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis // *Nucleic Acids Research* – 2018. – V.46(D1). – P.D252-D259.
 23. Boyarskikh U., Pintus S., Mandrik N., Stelmashenko D., Kiselev I., Evshin I., Sharipov R., Stegmaier P., **Kolpakov F.**, Filipenko M., Kel A. Computational master-regulator search reveals mTOR and PI3K pathways responsible for low sensitivity of NCI-H292 and A427 lung cancer cell lines to cytotoxic action of p53 activator Nutlin-3 // *BMC Med Genomics* – 2018. – V.11(Suppl. 1). – Art.12.
 24. Yevshin I., Sharipov R., Valeev T., Kel A., Kolpakov F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments // *Nucleic Acids Research* – 2017. – V.45(D1). – P.D61-D67.
 25. Kulakovskiy I.V., Makeev V.J., Vorontsov I.E., Kasianov A.S., Medvedeva Y.A., Yevshin I.S., **Kolpakov F.A.**, Soboleva A.V., Ashoor H., Ba-Alawi W., Bajic V.B. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models // *Nucleic Acids Research* – 2016. – V.44(D1). – P.D116-D125.
 26. Kutumova E., Sharipov R., **Kolpakov F.**, Zinovyev A. Model composition through model reduction: a combined model of CD95 and NF- κ B signaling pathways // *BMC Systems Biology* – 2013. – V.7 – Art.13.
 27. Kel A., **Kolpakov F.**, Poroikov V., and Selivanova G. geneXplain -identification of causal biomarkers and drug targets in personalized cancer pathways // *Journal of Biomolecular Techniques* – 2011. – V. 22(Supplement). – Art.S16.
 28. Waltemath D., Adams R., Bergmann F.T., Hucka M., **Kolpakov F.**, Miller A.K., Nickerson D., Moraru I.I., Sahle S., Snoep J.L., Le Novère N. Reproducible computational biology experiments with SED-ML - the simulation experiment description markup language // *BMC Systems Biology* – 2011. – V.5 – Art.198.
 29. Le Novère N., Hucka M., Mi H., Moodie S., Schreiber F., Sorokin A., Demir E., Wegner K.,

- Aladjem M.I., Wimalaratne S.M., Bergman F.T., Gauges R., Ghazal P., Kawaji H., Li L., Matsuoka Y., Villéger A., Boyd S.E., Calzone L., Courtot M., Dogrusoz U., Freeman T.C., Funahashi A., Ghosh S., Jouraku A., Kim S., **Kolpakov F.**, Luna A., Sahle S., Schmidt E., Watterson S., Wu G., Goryanin I., Kell D.B., Sander C., Sauro H., Snoep J.L., Kohn K., Kitano H. The Systems Biology Graphical Notation // *Nature Biotechnology* – 2009. – V.27(8). – P.735-741.
30. **Kolpakov F.**, Poroikov V, Sharipov R, Kondrakhin Y, Zakharov A, Lagunin A, Milanesi L, Kel A. CYCLONET - an integrated database on cell cycle regulation and carcinogenesis // *Nucleic Acids Research* – 2007. – V 35(Database issue). – D550-D556.
31. **Kolpakov F.A.**, Ananko E.A. Interactive data input into the GeneNet database // *Bioinformatics* – 1999. – V.15(7-8). – P.713-714.
32. Kolchanov N.A., Ananko E.A., Podkolodnaya O.A., Ignatieva E.V., Stepanenko I.L., Kel-Margoulis O.V., Kel A.E., Merkulova T.I., Goryachkovskaya T.N., Busygina T.V., **Kolpakov F.A.**, Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. Transcription Regulatory Regions Database (TRRD):its status in 1999 // *Nucleic Acids Research* – 1999. – V.27(1). – P.303-306.
33. **Kolpakov F.A.**, Ananko E.A., Kolesov G.B., Kolchanov N.A. GeneNet: a gene network database and its automated visualization // *Bioinformatics* – 1998. – V.14(6). – P. 529-537.
34. Heinemeyer T., Wingender E., Reuter I., Hermjakob H., Kel A.E., Kel O.V., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., **Kolpakov F.A.**, Podkolodny N.L., Kolchanov N.A. Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL // *Nucleic Acids Research* – 1998. – V.26(1). – P.362-367.

Журналы из списка Web of Science и/или Scopus, Q3-Q4

35. Акбердин И.Р., Вертышев А.Ю., Пинтус С.С., Попов Д.В., **Колпаков Ф.А.** Математическая модель, связывающая Ca²⁺-зависимый сигнальный путь с регуляцией экспрессии генов в клетках скелетной мышцы человека // *Математическая биология и биоинформатика* – 2020. – Т.15(1). – С.20-39.
36. Киселев И.Н., Акбердин И.Р., Вертышев А.Ю., Попов Д.В., **Колпаков Ф.А.** Модульная графическая модель энергетического метаболизма в клетках скелетной мышцы // *Математическая биология и биоинформатика* – 2019. – Т.14(2). – С.373-392.
37. Киселев И.Н., Кутумова Е.О., Колпакова А.Ф., Лифшиц Г.И., **Колпаков Ф.А.** Математическое моделирование действия антигипертензивных препаратов // *Математическая биология и биоинформатика* – 2019. – Т.14(1). – С.233-256.
38. Vorontsov I.E., Fedorova A.D., Yevshin I.S., Sharipov R.N., **Kolpakov F.A.**, Makeev V.J., Kulakovskiy I.V. Genome-wide map of human and mouse transcription factor binding sites aggregated from ChIP-Seq data 06 Biological Sciences 0604 Genetics // *BMC Research Notes* –

2018. – V.11(1). – Art.756.

39. Kondrakhin Y., Valeev T., Sharipov R., Yevshin I., **Kolpakov F.**, Kel A. Prediction of protein-DNA interactions of transcription factors linking proteomics and transcriptomics data // *EuPA Open Proteomics* – 2016. – V.13. – P.14-23.
40. Киселев И. Н., Бибердорф Э. А., Баранов В. И., Комлягина Т. Г., Мельников В. Н., Суворова И. Ю., Кривошеков С. Г., **Колпаков Ф. А.** Персонализация параметров и валидация модели сердечно-сосудистой системы человека // *Математическая биология и биоинформатика* – 2015. – Т.10(2). – С.526-547.
41. Kutumova E. O., Kiselev I. N., Sharipov R. N., Lavrik I. N., **Kolpakov F. A.** A modular model of the apoptosis machinery // *Advances in Experimental Medicine and Biology* – 2012. – V.736. – P.235-245.
42. Кутумова Е.О., Зиновьев А.Ю., Шарипов Р.Н., **Колпаков Ф.А.** Применение методов редукции для построения комплексной модели апоптотических путей // *Математическая биология и биоинформатика* – 2012. – Т.7(2). – С.572-588.
43. Киселев И. Н, Семисалов Б. В, Бибердорф Э. А., Шарипов Р. Н., Блохин А. М., **Колпаков Ф.А.** Модульное моделирование сердечно-сосудистой системы человек // *Математическая биология и биоинформатика* – 2012. – Т.7(2). – С.703-736.
44. Колчанов Н.А., Ананько Е.А., **Колпаков Ф.А.**, Подколотная О.А., Игнатъева Е.В., Горячковская Т.Н., Степаненко И.Л. Генные сети // *Молекулярная биология* – 2000. – Т.34(4). – С.533-544.
45. Ананько Е.А., **Колпаков Ф.А.**, Колесов Г.Б., Колчанов Н.А. Автоматическая генерация схем генных сетей на основе их формализованного описания в БД GeneNet // *Биофизика* – 1999. – Т.44(4). – С. 628-631.
46. **Колпаков Ф.А.**, Бабенко В.Н. Компьютерная система MGL – инструмент для составления выборок, графического представления и анализа регуляторных геномных последовательностей // *Молекулярная биология* – 1997. – Т.31(4). – С.647-655.

Журналы из списка Web of Science и/или Scopus, без Q

47. Akberdin I.R., Kiselev I.N., Pintus S.S., Vertyshev A., Makhnovskii P.A., Popov D.V., **Kolpakov F.A.** Integrated modular model linking metabolism, signaling transduction and gene expression regulation in human skeletal muscle // *CEUR Workshop Proceedings* – 2020. – V.2569. – P.1-5.
48. Kolmykov S.K., Evshin I.S., **Kolpakov F.A.** Analysis of NGS Data on the Transcriptional Regulation // *CEUR Workshop Proceedings* – 2020. – V.2569. – P.19-22.
49. Kulyashov M.A., Kolmykov S.K., Evshin I.S., **Kolpakov F.A.** Description, Characteristic And Algorithm For Creation Of A Dictionary Of Cell Types And Tissues In The Gtrd Database // *CEUR Workshop Proceedings* – 2020. – V.2569. – P.13-18.

Рецензируемые журналы, рекомендованные ВАК

50. Миронова Т.Е., Афонюшкин В.Н., Козлова Ю.Н., Бобикова А.С., Коптев В.Ю., Черепушкина В.С., Сигарева Н.А., **Колпаков Ф.А.** Изучение защитных эффектов вирицидных препаратов на модели коронавирусной пневмонии // Ветеринария и кормление – 2020. - № 7. – С.35-38.
51. Баранов В.И., Бибердорф Э.А., Киселев И.Н., **Колпаков Ф.А.**, Комлягина Т.Г., Кривошеков С.Г., Мельников В.Н., Суворова И.Ю. Персонализированная одномерная модель сердечно-сосудистой системы человека // Сибирский научный медицинский журнал – 2016. – Т.36(1). – С.70-79.

Прочие журналы

52. Стельмашенко Д.Е., Колпаков Ф.А., Евшин И.С., Мандрик Н.В., Рябова А.С., Пинтус С.С., Киселев И.Н., Шарипов Р.Н., Штегмайер Ф., Кошманн Д., Кель-Маргулис О.В., Крул М., Тандон Н., Вингендер Э., Кель А.Э. Персонализированная медицина в один клик: автоматическая идентификация терапевтической мишени по ДНК и РНК пациента // Лаборатория и производство. – 2019. – №3 (7). – С. 76-86.
53. Boyarskikh U.A., Pintus S.S., Mandrik N.V., Stelmashenko D.E, Kiselev I.N., Sharipov R.N., Kolpakova A.F., Stegmaier P., Filipenko M.L., Kolpakov F.A., Kel A.E.. Master-regulators driving resistance of non-small cell lung cancer cells to p53 reactivator Nutlin-3 // Virtual Biology – 2017. – V.4 – P.1–31.
54. Kondrakhin Yu., Valeev T., Sharipov R., Yevshin I., **Kolpakov F.** Toolkit for ChIP-seq based comparative analysis of the PWM methods for prediction of transcription factor binding sites // Virtual Biology – 2014. – V.2 – P.19-33.
55. Kiselev I., **Kolpakov F.** BioUML: plugin for population-based modeling // Virtual Biology – 2014. – V.2 – P.7-15.
56. Kiselev I. N., **Kolpakov F. A.** Modular Modeling of Biological Systems // Virtual Biology – 2013. – V.1 – P.30-44.
57. Valeev T., Yevshin I., **Kolpakov F.** BioUML genome browser // Virtual Biology – 2013. – V.1 – P.15-26.
58. Kutumova E., Ryabova A., Valeev T., **Kolpakov F.** BioUML plug-in for nonlinear parameter estimation using multiple experimental data // Virtual Biology – 2013. – V.1 – P.47-58.

Главы в книгах

59. Kiselev I.N., Semisalov B.V., Sharipov R.N., **Kolpakov F.A.** Modular approach to modeling complex biological systems. Circulatory System and Arterial Hypertension: Experimental Investigation, Mathematical and Computer Simulation / Eds. Ivanova L. N., Markel A. L., Blokhin

- A. M., Mishchenko E. V. Nova Science Publishers, Inc. 2012. – 258 pp. – ISBN-10: 1622572750.
60. Kiselev I.N., Semisalov B.V., Sharipov R.N., **Kolpakov F.A.** Modular modeling of the human cardiovascular system. Circulatory System and Arterial Hypertension: Experimental Investigation, Mathematical and Computer Simulation / Eds. L. N. Ivanova, A. L. Markel, A. M. Blokhin, E. V. Mishchenko. Nova Science Publishers, Inc. 2012. – 258 pp. – ISBN-10: 1622572750.
61. **Колпаков Ф.А.**, Шарипов Р.Н., Евшин И.С., Леонова Т.И., Семисалов Б.В. Компьютерное моделирование системы кровообращения. Система кровообращения и артериальная гипертония: биофизические и генетико-физиологические механизмы, математическое и компьютерное моделирование / под ред. Л.Н. Ивановой, А.М.Блохина, А.Л. Маркеля. Новосибирск, Россия. Издательство СО РАН 2008. – 252 с. ISBN: 978-5-7692-1021-1.

Учебные пособия

62. Бибердорф Э.А., Блохин Ф.М., Киселев И.Н. Семисалов Б.В. Колпаков Ф.А. Система кровообращения человека – математическое и компьютерное моделирование: учебное пособие. Новосибирск: РИЦ НГУ 2015. – 187 с.

Труды конференций (избранное)

63. Киселев И.Н., Колпакова А.Ф., **Колпаков Ф.А.** Математическое моделирование фармакокинетики и фармакодинамики алискирена // XVI Российская конференция «Распределенные информационно - вычислительные ресурсы. Наука – цифровой экономике» (DICR-2017): Труды XVI Всероссийской конференции (4-7 декабря 2017 г.). Новосибирск / Под ред. О.Л. Жижимова, А.М. Федотова. – Новосибирск: ИВТ СО РАН. – 2017. – С.218-223.
64. Киселев И.Н., Шарипов Р.Н., Мандрик Н.В., **Колпаков Ф.А.** Популяционное моделирование сердечно-сосудистой системы человека в платформе BioUML // XVI Российская конференция «Распределенные информационно - вычислительные ресурсы. Наука – цифровой экономике» (DICR-2017): Труды XVI Всероссийской конференции (4-7 декабря 2017 г.). Новосибирск / Под ред. О.Л. Жижимова, А.М. Федотова. – Новосибирск: ИВТ СО РАН. – 2017. – С.212-217.
65. Pintus S., Valeev T.F., Yevshin I.S., **Kolpakov F.A.** Biostore: a cloud-compatible hub for bioinformatics related tools and platforms // Tenth International Conference on Genome Regulation and Structure\Systems Biology (BGRS\SB-2016): Novosibirsk, 29 August – 2 September, 2016. Abstracts. – Novosibirsk. – 2016. – P.237.
66. Mandrik N., Kutumova E.O., Kolpakov F.A. Rule-based modeling in BioUML // Proceedings of The Second International Conference “Mathematical Modeling and High Performance Computing in Bioinformatics, Biomedicine and Biotechnology” (within BGRS\SB-2016): Novosibirsk, Russia. 29 August – 2 September 2016. – Novosibirsk. – 2016. – P.72.

67. Kiselev I.N., **Kolpakov F.A.**, Biberdorf E.A., Baranov V.I., Komlyagina T.G., Suvorova I.Y., Melnikov V.N., Krivoschekov S.G. Patient-specific 1d model of the human cardiovascular system // Proceedings - 2015 International Conference on Biomedical Engineering and Computational Technologies, SIBIRCON 2015. Institute of Electrical and Electronics Engineers Inc.– 2015. – P.76-81.
68. **Kolpakov F.A.**, Kiselev I.N., Valeev T.F., Bukharov A., Mandrik N., Kondrakhin Y.V., Yevshin I.S., Kolpakova A. BioUML Platform - From Systems Biology to Systems Pharmacology // 7th Noordwijkerhout Symposium on Pharmacokinetics, Pharmacodynamics and Systems Pharmacology. Systems Pharmacology in Drug Discovery & Development: Noordwijkerhout, The Netherlands, 23-25 April 2014. – Noordwijkerhout: Leiden Academic Centre for Drug Research. – 2014. – P.156-157.
69. Yevshin I.S., Valeev T.F., Sharipov R.N., **Kolpakov F.A.** BioUML – software platform for analysis of Next generation sequencing data using collaborative and reproducible research // International conference “High-throughput sequencing in genomes - 2013”: Novosibirsk, July 21-25, 2013. – Novosibirsk: Institute of Chemical Biology and Fundamental Medicine. – 2013. – P.66.
70. Semisalov B.V., Kiselev I.N., Sharipov R.N., **Kolpakov F.A.** BioUML: plugin for stochastic modeling of biological systems // The 8th International Conference on Bioinformatics of Genome Regulation and Structure/Systems Biology: Novosibirsk, June 25-29 2012. – Novosibirsk. – 2012. – P.283.
71. **Kolpakov F.**, Puzanov M., Koshukov A. BioUML: visual modeling, automated code generation and simulation of biological systems // Proceedings of the 5th International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2006), July 16-22, 2006. – Novosibirsk. –2006. – P.281-284.
72. **Kolpakov F.A.** BioUML – open source extensible workbench for systems biology // Proceedings of the 4th International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2004), July 25-30, 2004. – Novosibirsk. –2004. – P.77-80.
73. **Kolpakov F.A.** Comprehensive model for formalized description, visualization and simulation of biological systems // Proceedings of German Conference on Bioinformatics. October 12-14, 2003. Proceedings of the 3rd International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002). – Novosibirsk. – 2002. – V.2 P.128-131.
74. Kel, A. E. Computer tool FUNSITE for analysis of eukaryotic regulatory genomic sequences / Kel A. E., Kondrakhin Y. V., **Kolpakov Ph. A.** [et al.] // Proceedings. International Conference on – Neuherberg/Garching, Germany. – 2003. – V.II. – P.155-157.
75. **Kolpakov F.A.** BioUML – framework for visual modeling and simulation of biological systems. // Intelligent Systems for Molecular Biology. – 1995. – V. 3. – P. 197-205.

БЛАГОДАРНОСТИ

Автор благодарен академику РАН, профессору, д.б.н. Н.А. Колчанову, в лаборатории которого начался трудовой путь автора и был сформулирован ряд идей, которые получили дальнейшее развитие в течение последующих лет работы над диссертацией; научному консультанту член-корреспонденту РАН, профессору, д.б.н, В.В. Поройкову за помощь в подготовке диссертации; академику РАН, профессору, д.ф-м.н. Ю.И. Шокину за стимулирование автора к более быстрому написанию диссертации.

Автор признателен директору ООО "Биософт.Ру" и ООО "Девелопмент груп" С.А. Жатченко за всемерную поддержку работы и идей автора; директору geneXplain GmbH А.Э. Келю за поддержку и совместную работу над проектами. Автор благодарен Шарипову Р.Н., к.ф-м.н. Кутумовой Е.О. и другим коллегам за помощь по оформлению диссертационной работы.

Автор благодарен коллегам, которые внесли свой большой вклад в развитие ПК BioUML, созданные при его помощи математические модели сложных биологических систем, в создание и развитие БД GTRD.

Разработка ПК BioUML	Разработка моделей	Разработка БД GTRD
в настоящее время (по алфавиту)		
Евшин И.С. Жатченко С.А. к.ф-м.н. Киселев И.Н. Колмыков С.К. к.б.н. Кондрахин Ю.В. Куляшов М.А. к.ф-м.н. Кутумова Е.О. к.б.н. Пинтус С.С. Рябова А.С. Сон А.	к.б.н. Акбердин И.Р. к.б.н. Афонюшкин В.Н. Вертышев А.Ю. Евшин И.С. к.ф-м.н. Киселев И.Н. к.ф-м.н. Кутумова Е.О. д.м.н. Лифшиц Г.И. д.б.н. Попов Д.В. Шарипов Р.Н.	Евшин И.С. к.б.н. Кондрахин Ю.В. Куляшов М.А. к.б.н. Пинтус С.С. Спасельникова А. Шарипов Р.Н. Эрнст Т.
раньше (в историческом порядке)		
Валеев Т. Мандрик Н. Толстых Н. Киншт С. Бухаров А. Семисалов Б. Леонова Т. Шадрин А. Пузанов М.	Бухаров А. Семисалов Б. Леонова Т. Басманова Е. к.б.н. Баранов В.И. к.ф-м.н. Бибердорф Э.А. д.ф-м.н. Блохин А.М.	Замятин В.

Рюмин Д. Магдысюк А. Кошуков А. Худяков В. Шайдуков А. Тяжев И. Жвалеев В. Онегов О.		
---	--	--

Приложение 1.1 Высокопроизводительные технологии для сбора численных данных

Область знаний Краткое описание технологий	Основные характеристики	Литература
<p>Геномика высокопроизводительные методы секвенирования ДНК, включая изучение вариабельности в различных клетках одного организма</p>	<p>На примере технологии Solid4 (Applied Biosystems, USA):</p> <ul style="list-style-type: none"> - загрузка до 96 капилляров одновременно (24 образца), - секвенирование до 60 млрд пар оснований за один прогон (фрагменты по 50-70 п.н. до 10 дней) - точность более 99,94% - возможность идентификации SNP и аллельных вариантов 	<p>The SOLiD™ System: Next-Generation Sequencing.</p> <p>Tucker et al., 2009</p>
<p>Эпигеномика, эпигенетика изучение факторов транскрипции, не кодируемых в ДНК (метилирование ДНК, и т. д.)</p>	<p>Наиболее популярной в силу ряда преимуществ сейчас является ChIP-seq технология:</p> <ul style="list-style-type: none"> - чтение более 700 млн фрагментов за прогон, - длина ридов до 250 оснований, - точность определения сайта связывания +50 п.н. (против 500-1000 п.н. у ChIP-chip), - низкий фоновый шум (в сравнении с ChIP-chip), - отсутствие перекрестной гибридизации (в сравнении с ChIP-chip), - малое количество ДНК (менее 10 нг против 4 мкг у ChIP-chip) и значит малое число реакций иммунопреципитации, - высокая чувствительность и надежность в распознавании аллельных вариантов, SNP и т.д. <p>ChIP-chip и ChIP-PET являются менее распространенными в силу меньшей эффективности данных технологий. Использование дрожжевой двугибридной системы (Y2H) также используется для выявления белок-ДНК взаимодействий и позволяет добиться хороших результатов.</p>	<p>Chromatin Immunoprecipitation Sequencing (ChIP-Seq) Illumina genome analyzer scalability.</p> <p>Hoffman et al., 2009. Bock and Lengauer, 2008.</p>
<p>Транскриптомика измерение экспрессии генов, используя ДНК-микрочипы и другие методы</p>	<p>Наиболее популярен анализ экспрессии с использованием микрочипов – мультиплексной технологии, основанной на гибридизации меченых специально подобранных олигонуклеотидов, иммобилизованных на подложке, с мечеными кДНК – копиями иРНК, выделенных из клетки. На основе данных гибридизации делается вывод о повышенной или пониженной по сравнению с контролем представленности продуктов тех или иных генов в клетке, выявляются SNP и генотируются мутантные геномы. Существует большое разнообразие типов микрочипов, отличающихся</p>	<p>Birmingham et al., 2009. Mullighan and Downing, 2009.</p>

	<p>принципом изготовления, точностью, эффективностью и стоимостью. Наибольшее распространение получили двухканальные микрочипы (Cy3/Cy5) и одноканальные (производства Affymetrix, Agilent и др.).</p>	
<p>Интерферомика исследование взаимодействия транскрибируемых РНК</p>	<p>Технология, основанная на принципе комплементарности нуклеиновых кислот, позволяющая, внося в клетки антисмысловые малые РНК, комплементарные иРНК определенных генов, блокировать их трансляцию. Это позволяет выключать определенные гены и исследовать их функции. Подобное явление наблюдается <i>in vivo</i>, и может быть продемонстрировано на примере регуляции экспрессии ряда генов, например, эмбриональных. Данная технология считается перспективной и уже применяется для лечения рака у человека (блокирование трансляции иРНК онкогенов).</p>	<p>Olejniczak et al., 2010.</p>
<p>Протеомика измерение уровня белков или пептидов методами гель-электрофореза, масс-спектрометрии или многомерных методик измерения белков.</p>	<p>Протеомные методы достаточно разнообразны и включают в себя задержку в геле, иммуногистохимический и иммуноферментный анализ, масс-спектрометрию (MALDI и ESI), микроскопические методы (например, флуоресцентную, FRET), и различные компьютерные методы (программы симуляции докинга, создания белковых сетей и т.п.).</p>	<p>Терентьев и др., 2009. /Terentiev et al., 2009. Reimel et al., 2009.</p>
<p>Метаболомика измерение концентраций малых молекул, метаболитов</p>	<p>Использование высокопроизводительных методов масс-спектрометрии и ЯМР позволяет получать количественные данные о метаболитах с диагностическими и научными целями и выявлять паттерны, характерные для различных нормальных и патологических состояний организма, ткани, клетки и т.д. Масс-спектрометрия определяет массу молекул путем их ионизации и распределения полученных ионов в электрическом поле в зависимости от отношения массы к заряду. Чаще всего используются:</p> <ul style="list-style-type: none"> - метод мягкой ионизации – матрично-активированная лазерная десорбция/ионизация (MALDI), при котором образец, содержащий пептиды, смешивается с молекулами специфической матрицы, которая облучается ионизирующим лазером; - активированная поверхностью лазерная десорбция/ионизация (SELDI). В этом случае образец наносится на специальный чип, а не смешивается с матрицей, и ионизация происходит в вакуумной камере. Этот метод не позволяет непосредственно определить белок, приходится использовать дополнительно ионообменную хроматографию и 	<p>Scalbert et al., 2009. Powers, 2009.</p>

	полиакриламидный гель-электрофорез.	
Липидомика измерение уровня липидов	Для анализа в основном применяется сочетание методов высокоэффективной жидкостной/ газовой хроматографии и масс-спектрометрии. Существует также метод «струйной» липидомики (shotgun lipidomics), появившийся в результате прогресса методов масс-спектрометрии с ионизацией электрораспылением, и практическому применению разделения компонентов внутри источника ионов в качестве стратегии применения двумерной масс-спектрометрии для изучения состава липидов биологических образцов. Полученные количественные данные используются для создания липидного профиля с помощью компьютерных методов анализа.	Seppänen-Laakso and Oresic, 2009.
Гликомика измерение уровня углеводов	Методы исследования представленности и количества углеводов базируются на методах масс-спектрометрии, капиллярной хроматографии и 1D/2D гель-электрофореза.	Zaia, 2008. Novotny et al., 2008.
Интерактомика исследование взаимодействий между молекулами (например, измерение белок-белковых взаимодействий: PPI)	Методы интерактомики позволяют выявить парные и более комплексные взаимодействия белковых молекул, составить карты таких взаимодействий и верифицировать или предсказать возможные функции белков в клетке и организме (в том числе на основе гомологии с уже изученными белками других организмов). В настоящее время широкое применение получили двугибридная дрожжевая система (Y2H), фаговый дисплей, аффинная очистка совместно с масс-спектрометрией (TAP-MS).	Терентьев и др., 2009. /Terentiev et al., 2009. Singh and Nagaraj, 2006.
Флаксомика измерение динамики потоков и концентраций во времени (как правило, метаболитов).	Методы флаксомики представлены пульсовым мечением ^{13}C с последующим анализом методами масс-спектрометрии распределения метки по потокам, а также использование флуоресцентно меченых белков-наносенсоров, изменяющих свою конформацию в результате взаимодействия с лигандом. Все эти методы дают хорошее разрешение по времени. Последний метод позволяет работать только с отдельными веществами в отличие от метода масс-спектрометрии, тем не менее, он позволяет работать на клеточном и внутриклеточном уровне с минимальным вмешательством в ее процессы.	Wiechert et al., 2007.

ПРИЛОЖЕНИЕ 1.2 КРАТКОЕ ОПИСАНИЕ ФОРМАТА SED-ML

Формат SED-ML включает в себя следующие основные элементы (см. рисунок ниже):

- **DataDescription** – позволяет ссылаться на внешние данные (например, таблица с экспериментальными значениями концентраций веществ в заданных временных точках) и содержит описание того, как получить доступ к данным, в каком они формате и какое подмножество данных нужно извлекать. Такие данные могут использоваться, например, для определения параметров модели или для построения графиков численного моделирования вместе с соответствующими экспериментальными значениями;
- **Model** – определяет модели, используемые в численном эксперименте. Класс `Change` позволяет описать, как модели должны быть предварительно изменены (например, изменение значения параметра модели, удаление или добавление веществ и реакций, и т.п.).
- **Simulation** – является абстрактным классом и служит контейнером для различных типов экспериментов по моделированию. SED-ML Level 1 версии 4 предоставляет предопределенные классы моделирования:
 - `UniformTimeCourse` – численное моделирование динамики модели с заданным временным шагом;
 - `OneStep` – делает один шаг по расчету динамики модели;
 - `SteadyState` – поиск стационарного состояния;
 - `Analysis` – любой вид анализа или симуляции, который далее описывается в элементе `Algorithm`.
- **AbstractTask** – определяет какую задачу моделирования, для какой модели и с какими параметрами необходимо запустить.
- **DataGenerator** – позволяет описать процесс обработки результатом моделирования для последующего их вывода на график или таблицу. Например, можно нормализовать результаты вычислений или вычислить среднее значение нескольких переменных. Для описания таких вычислений используется формат `MathML`.
- **Output** – позволяет задать вывод результатов численного моделирования. Это может быть:
 - `Plot` – двух- или трех- мерный график;
 - `Figure` – несколько графиков могут быть объединены в одну фигуру;
 - `Report` – одна или несколько таблиц;
 - `ParameterEstimationReport` – результаты определения параметров модели по экспериментальным данным.
- **Style** – определяет графические стили для графиков и фигур.

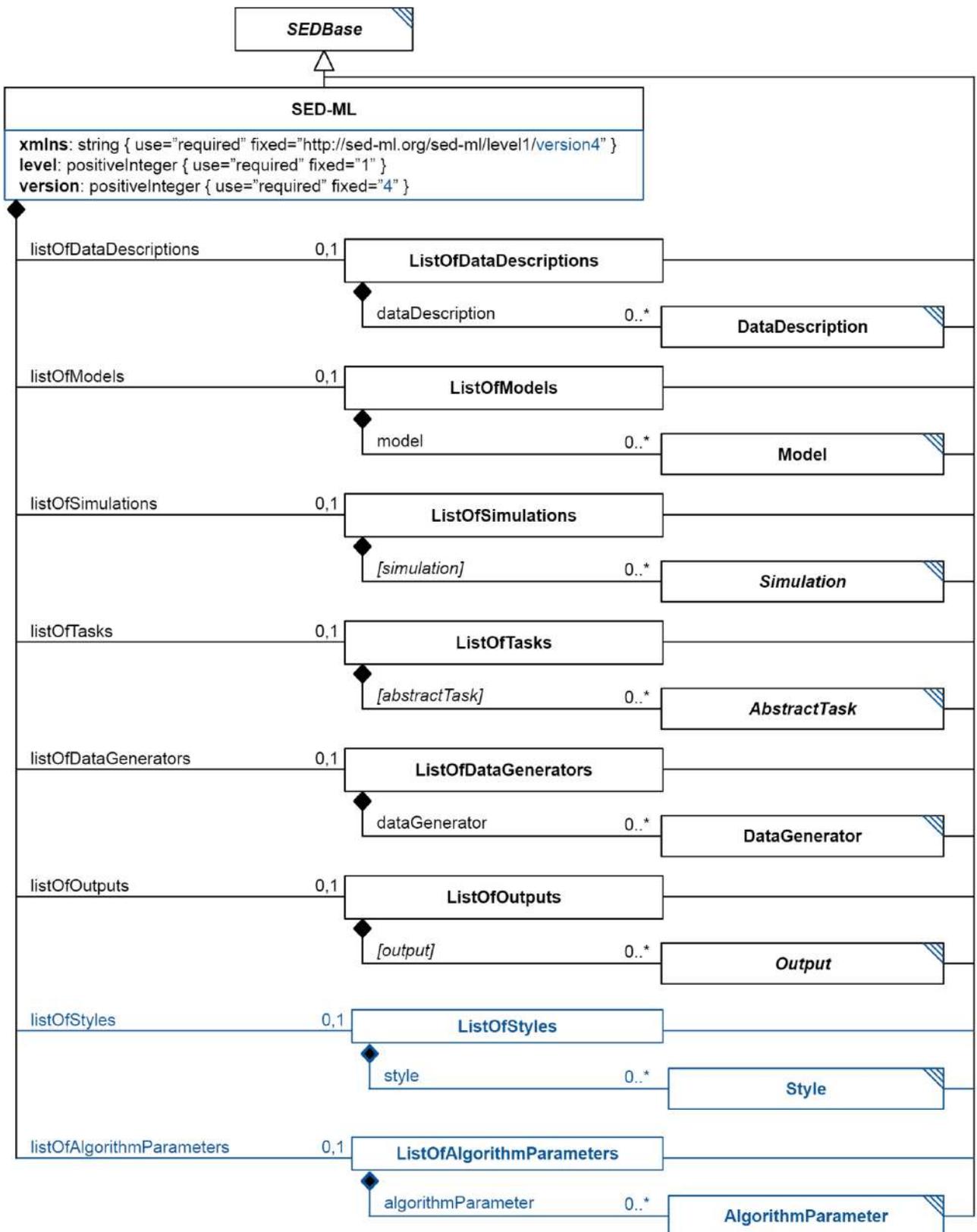


Рисунок П.1.2 – Структура основных компонентов (классов) формата SED-ML level 1 version 4 (<https://sed-ml.org/documents/sed-ml-L1V4.pdf>)

ПРИЛОЖЕНИЕ 2.1 ОСНОВНЫЕ ЭТАПЫ РАЗВИТИЯ ПК BioUML

Год, конференция (журнал), ссылка	Название публикации, функционал	Основные авторы
2002 BGRS'2002, (Kolpakov, 2002)	<p><i>BioUML – framework for visual modeling and simulation of biological systems.</i></p> <p>Первая версия BioUML. Разработано ядро BioUML: - мета-модель для визуального моделирования биологических систем; - автоматическая генерация кода для численного моделирования в MATLAB.</p>	Колпаков Ф.А.
2004 BGRS'2004, Kolpakov 2004	<p><i>BioUML – open source extensible workbench for systems biology.</i></p> <p>Описание мета-модели, графической нотации, модульной архитектуры на основе. Поддержка форматов SBML и CellML и БД KEGG/Pathways, TRANSPATH, GeneNet.</p>	Колпаков Ф.А.
2006 BGRS'2006, Kolpakov et al., 2006	<p><i>BioUML: visual modeling, automated code generation and simulation of biological systems.</i></p> <p>Генерация Java кода для численного решения систем ОДУ. Первые ОДУ-солверы. ПК BioUML впервые прошел SBML Semantic test suite на 100%.</p>	Колпаков Ф., Пузанов М., Кошуков А.
2009 Nature Biotechnology, Le Novere et al., 2009	<p><i>The Systems Biology Graphical Notation</i></p> <p>Поддержка SBGN нотации в BioUML.</p>	Колпаков Ф.А., Толстых Н.И.
2010 BGRS/SB 2010, Kolpakov et al., 2010	<p><i>BioUML — Integrated Platform for Building Virtual Cell and Virtual Physiological Human</i></p> <p>Первая веб-версия платформы BioUML. Интеграция с R. Поддержка R и JavaScript для</p>	Колпаков Ф.А. Толстых Н.И. Валеев Т.Ф., Рябова А.С., Кутумова Е.О.,

	написания скриптов.	Киселев И.Н.
2012 BGRS/SB 2012, Kiselev et al., 2012	<i>BioUML: modular modeling of complex biological system</i> Модульное визуальное моделирование сложных систем. Примеры: модульная модель апоптоза (Kutumova et al., 2012), модульная модель сердечно-сосудистой системы человека (Kiselev et al., 2012a,b).	Киселев И.Н., Колпаков Ф.А.
2012 BGRS/SB 2012, Semisalov et al., 2012	<i>BioUML: Plugin for stochastic modeling of biological systems</i> Реализованы методы для стохастического моделирования биологических систем: Гиллеспи, Гибсона-Брука и тау-прыжка (tau-leaping).	Семисалов Б.В., Киселев И.Н., Шарипов Р.Н., Колпаков Ф.А.
2013 International conference “High-throughput sequencing in genomes - 2013” Yevshin et al., 2013	<i>BioUML – software platform for analysis of Next generation sequencing data using collaborative and reproducible research</i> Начата реализация методов для анализа данных секвенирования, полученных с помощью технологий Next-generation sequencing.	Евшин И.С., Валеев Т.Ф., Шарипов Р.Н., Колпаков Ф.А.
2013 Virtual biology Valeev et al, 2013	<i>BioUML genome browser</i> Реализован геномный браузер	Валеев Т., Евшин И., Колпаков Ф.
2014 Virtual biology Kiselev and Kolpakov, 2014	<i>BioUML: plugin for population-based modeling</i> Реализован модули для популяционного моделирования, используя нелинейные модели со смешанными эффектами.	Киселев И.Н., Колпаков Ф.А.
2014 7th Noordwijkerhout Symposium on Pharmakokinetics, Pharmacodynamics and Systems	<i>BioUML Platform - From Systems Biology to Systems Pharmacology</i> Реализованы модули для: - хемоинформатики (интеграция пакета CDK) - предсказание спектров биологической	Колпаков Ф.А., Киселев И.Н., Валеев Т.Ф., Бухаров А., Мандрик Н., Кондрахин Ю.В., Евшин И.С.,

Pharmacology. Systems Pharmacology in Drug Discovery & Development Kolpakov et al., 2014	активности химических веществ, используя программу PASS (Poroikov et al., 2007).	Колпакова А.Ф.
2016 BGRS\SB-2016 Mandrick at al., 2016	<i>Rule-based modeling in BioUML</i> Поддержка моделирования на основе задания шаблонов реакций: специальный тип диаграмм и генерация полного набора реакций на основе заданных шаблонов (правил).	Мандрик Н.В., Кутумова Е.О., Колпаков Ф.А.
2016 BGRS\SB-2016 Pintus et al., 2016	<i>BioStore: a cloud-compatible hub for bioinformatics related tools and platforms</i> Использование BioStore как центрального хаба для управления правами доступа для BioUML серверов.	Пинтус С.С., Валеев Т.Ф., Евшин И.С., Колпаков Ф.А.
2016 ММ-НРС-BBB-20 16 Kiselev et al., 2016	<i>Improved SBGN (ML) support in BioUML</i> Улучшена поддержка стандарта SBGN.	Киселев И.Н., Киншт С., Колпаков Ф.А.
2021 Nucleic Acids Research, Kolpakov et al., 2021	<i>BioUML—towards a universal research platform</i> Реализованы: - интеграция с Jupyter notebook/hub; - интеграция с docker; - интеграция с Galaxy Tool Shed; - поддержка CWL и nextFlow Предложена архитектура универсальной научной платформы – u-science	Колпаков Ф., Киселев И., Колмыков С., Куляшов М., Кутумова Е., Рябова А., Евшин И. Жатченко С.

История всех версий ПК BioUML и реализованного в них функционала доступна на веб-сайте: http://wiki.biouml.org/index.php/BioUML_development_history

ПРИЛОЖЕНИЕ 2.2 ОСНОВНЫЕ МОДУЛИ (PLUG-INS) ПК BioUML

Модуль	Краткое описание
<i>ядро системы</i>	
org.eclipse.core.contenttype_3.6.0 org.eclipse.core.jobs_3.9.0 org.eclipse.core.runtime_3.13.0 org.eclipse.equinox.app_1.3.400 org.eclipse.equinox.common_3.9.0 org.eclipse.equinox.launcher_1.4.0 org.eclipse.equinox.preferences_3.7.0 org.eclipse.equinox.registry_3.7.0 org.eclipse.osgi_3.12.0 org.eclipse.update.configurator 3.3.100	подсистема управления модулями (Eclipse run-time)
ru.biosoft.workbench.editors_0.9.10 ru.biosoft.workbench_0.9.10	основные компоненты интерфейса пользователя (главное окно, меню, панель инструментов, документ и его части, управление панелями и т.п.); библиотека графических объектов
biouml.workbench_0.9.10	компоненты меню, диаграмма как документ; мета-модель и объектно-ориентированное представление предметной области; графическая нотация
ru.biosoft.access_0.9.10	доступ к БД, репозиторий, панель репозитория
ru.biosoft.table_0.9.10 ru.biosoft.treetable 0.9.10	работа с табличными данными
ru.biosoft.access.search_0.9.10 org.apache.lucene_2.3.2 biouml.plugins.lucene 0.9.10	подсистема поиска по базам данных; интерфейс пользователя
com.developmentontheedge.beans-3.0.4 com.beanexplorer.logging 0.0.1	BeanExplorer, инспектор свойств
ru.biosoft.jobcontrol-2.0.0	контроль процесса выполнения длительных операций
ru.biosoft.exception-1.0.0	определение основных типов исключений
ru.biosoft.templates 0.9.10	работа с шаблонами Velocity
<i>визуальное моделирование</i>	
biouml.plugins.sbgm 0.9.10	поддержка SBGN – Systems Biology Graphic Notation
biouml.plugins.gne 0.9.10	визуальный редактор графических нотаций
ru.biosoft.math_0.9.10	разбор и генерация математических выражений на основе абстрактного синтаксического дерева
biouml.plugins.simulation_0.9.10	подсистема моделирования: автоматическая генерация кода, численное моделирование (системы алгебро-дифференциальных уравнений, гибридные системы)
biouml.plugins.simulation test 0.9.10	система тестов
biouml.plugins.antimony 0.9.10	поддержка Antimony
<i>методы моделирования</i>	

biouml.plugins.agentmodeling_0.9.10 jade	мультиагентное моделирование фреймворк JADE
biouml.plugins.bionetgen_0.9.10	генерация моделей на основе правил (rule based modeling), используя формат BioNetGen
biouml.plugins.fbc.server_0.9.10 biouml.plugins.fbc_0.9.10 glpk gurobi	анализ метаболических потоков (fbc) методы линейного программирования для анализа метаболических потоков
biouml.plugins.hemodynamics_0.9.10	моделирование гемодинамики
biouml.plugins.pharm_0.9.10	популяционное моделирование на основе моделей со смешанными эффектами (nlme)
biouml.plugins.stochastic_0.9.10	стохастическое моделирование
<i>Методы оптимизации и анализа моделей</i>	
biouml.plugins.modelreduction_0.9.10	методы упрощения моделей
ru.biosoft.analysis.optimization_0.9.10	определение общего интерфейса для оптимизационных задач; ряд алгоритмов для их решения
biouml.plugins.optimization_0.9.10	методы подгонки параметров модели на основе экспериментальных данных
<i>Доступ к биологическим БД</i>	
biouml.plugins.chebi_0.9.10	ChEBI
biouml.plugins.ensembl_0.9.10 org.ensembl.ensj_39.2	Ensembl
biouml.plugins.go_0.9.10	GeneOntology
biouml.plugins.kegg_0.9.10	KEGG/Ligands, Pathways
biouml.plugins.reactome_0.9.10	Reactome (MySQL)
biouml.plugins.miriam_0.9.10	MIRIAM – каталог БД
biouml.plugins.sabiork_0.9.10	SABIO-RK
biouml.plugins.uniprot_0.9.10	UNIPROT
<i>Чтение/запись биологических данных в разных форматах</i>	
biouml.plugins.biopax_0.9.10 org.semanticweb.owl_2.1.0	BioPAX
biouml.plugins.cellml_0.9.10	CellML
biouml.plugins.obo_0.9.10	OBO (Open Biology Ontology)
biouml.plugins.psimi_0.9.10	PSI-MI
biouml.plugins.sbml_0.9.10	SBML
org.jetbrains.bio.big_0.9.1	bigBed, bigWig
<i>Импорт/экспорт диаграмм (графов) в разных форматах</i>	
biouml.plugins.gxl_0.9.10	GXL – Graph eXchange Format
biouml.plugins.svg_0.9.10 org.apache.batik.1.7	SVG – Scalable Vector Graphics
<i>Хемоинформатика</i>	
org.openscience.cdk_1.3.5	Chemical Development Kit – библиотека для чтения/записи и визуализации химических структур
org.openscience.jchempaint_3.1.2	редактор структурных формул
biouml.plugins.cheminformatics_0.9.10	интеграция CDK, редактор структурных формул как документ BioUML

biouml.plugins.pass_0.9.10	интеграция с программой PASS
org.openscience.jmol_12.2.4	JMol – программа для просмотра 3-х мерных структур макромолекул
<i>визуализация и анализ биомедицинских данных</i>	
ru.biosoft.bsa_0.9.10	объектно-ориентированное представление биологических данных, связанных с нуклеотидными и аминокислотными последовательностями
ru.biosoft.bsastats_0.9.10	сбор ряда статистических показателей по набору коротких прочтений
ru.biosoft.analysiscore_0.9.10	определение общего интерфейса для методов анализа
ru.biosoft.analysis_0.9.10	интерфейс пользователя; ряд методов анализа
ru.biosoft.graph_0.9.10 ru.biosoft.plugins.graph_0.9.10 biouml.workbench.graph_0.9.10	библиотеки для работы с графами
ru.biosoft.graphics_0.9.10	библиотека графических объектов
affymetrix.fusion_1.1.1 biouml.plugins.affymetrix_0.9.10 biouml.plugins.agilent_0.9.10 biouml.plugins.microarray_0.9.10.jar	анализ микрочиповых данных
biouml.plugins.bindingregions_0.9.10	различные методы для поиска и анализа регуляторных элементов
biouml.plugins.bowtie_0.9.10	обертка для программы Bowtie
biouml.plugins.chipmunk_0.9.10	обертка для программы ChIPMunk для поиска мотивов в ChIP-seq данных
biouml.plugins.cma_0.9.10	поиск композиционных модулей (CMA – Composite Module Analyst)
biouml.plugins.enrichment_0.9.10	анализ обогащения групп генов по разным критериям (GeneOntology, метаболические пути и т.п.)
biouml.plugins.genemodels_0.9.10	моделирование экспрессии генов при помощи разного типа моделей
biouml.plugins.gtrd_0.9.10	методы анализа для БД GTRD
biouml.plugins.machinelearning_0.9.10	различные методы машинного обучения
biouml.plugins.riboseq_0.9.10	методы анализа ribo-seq данных
biouml.plugins.research_0.9.10	журнал исследований – хранит историю всех анализов в каждом проекте пользователя
<i>Интеграция с R, JavaScript, docker, Jupyter</i>	
ru.biosoft.plugins.jri_0.9.10 org.rosuda.JRI_0.6 org.rosuda.engine_0.6.0	интеграция с R, используя библиотеку JRI
ru.biosoft.plugins.javascript_0.9.10 org.mozilla.javascript_1.7.0 com.Ostermiller.syntax_1.0.0	поддержка JavaScript (консоль, редактор и дебагер для выполнения JavaScript)
ru.biosoft.plugins.docker_0.9.10	интеграция с docker
ru.biosoft.plugins.jupyter_0.9.10	интеграция с Jupyter hub
<i>сценарии анализа данных</i>	
biouml.plugins.cwl-1.0.0	поддержка CWL
biouml.plugins.galaxy2_0.9.10 ru.biosoft.galaxy_0.9.10	интеграция с Galaxy

biouml.plugins.nextflow 0.9.10	интеграция с Nextflow
biouml.plugins.sedml_0.9.10 org.jlibsedml 2.2.1	поддержка SED-ML
biouml.plugins.wdl_0.9.10 cromwell_83	поддержка WDL, интеграция с Cromwell
<i>веб-сервер</i>	
biouml.plugins.server_0.9.10 com.developmentontheedge.server_0.9.10 ru.biosoft.bsa.server_0.9.10 ru.biosoft.server_0.9.10 ru.biosoft.services_0.9.10 ru.biosoft.server.servlets_0.9.10 ru.biosoft.server.tomcat_0.9.10	различные функции веб-сервера ПК BioUML
biouml.plugins.node_0.9.10	запускает ПК BioUML как отдельный узел в кластере для выполнения расчетных задач
<i>разные библиотеки</i>	
org.apache.axis_1.1.0 org.apache.bcel_6.0.0 org.apache.commons.cli_1.0.0 org.apache.commons.codec_1.6.0 org.apache.commons.collections_3.2.1 org.apache.commons.collections4_4.1 org.apache.commons.compress_1.6 org.apache.commons.fileupload_1.2.1 org.apache.commons.io_1.4 org.apache.commons.lang_2.4.0 org.apache.commons.logging_1.1.1 org.apache.commons.math3_3.3 org.apache.http_4.5 org.apache.log4j_1.2.12 org.apache.mina.core_2.0.9 org.apache.poi_3.17 org.apache.velocity_1.6.2	библиотеки проекта Apache
com.incors.plaf.alloy_1.4.4.jar com.incors.plaf.kunststoff_2.0.2.jar com.jgoodies.looks_2.1.4.jar net.java.plaf.windows_0.5.0 plaf.metal_1.5.0.jar	настройка внешнего вида приложения (look & feel)
org.yaml.snakeyaml_1.11.jar	чтение-запись данных в YAML формате

Приложение 2.3 JAVA БИБЛИОТЕКА ДЛЯ ДОСТУПА И ПОИСКА ИНФОРМАЦИИ В БД

Для доступа и поиска информации в биологических БД используется разработанная автором библиотека Java классов (Колпаков, 2011), обеспечивающая унифицированный интерфейс для работы с гетерогенными БД, представление информации из них в объектно-ориентированном виде, а также универсальную полнотекстовую поисковую систему.

Каждый объект (запись) в БД представляется в виде Java класса, реализующего интерфейс `DataElement`, содержащий 2 метода:

- `String getName()` – возвращает уникальный идентификатор (первичный ключ) объекта;
- `DataCollection getOrigin()` – возвращает коллекцию данных, которой он принадлежит.

Набор однотипных объектов в БД представляется в виде Java объекта, реализующего интерфейс `DataCollection`, описывающий методы для доступа, редактирования, удаления и добавления объектов типа `DataElement`. БД может быть представлена несколькими объектами типа `DataCollection`, например, каждой таблице БД может соответствовать свой экземпляр `DataCollection`. Поскольку интерфейс `DataCollection` является расширением интерфейса `DataElement`, то возможно построение древовидной структуры, соответствующей сложно организованному биологическим данным.

Набор подклассов, реализующих интерфейс `DataCollection`, обеспечивает доступ к разным типам данных (текстовые файлы, реляционные БД, XML данные и другие);

Концепция `DerivedDataCollection` обеспечивает возможность построения цепочки обработки данных, состоящей из нескольких `DataCollection`. Информация из БД представляется в виде объектов `DataElement` первичной `DataCollection`.

Основные подклассы `DerivedDataCollection`:

- `FilteredDataCollection` – осуществляет подвыборку элементов из первичной `DataCollection`. Java объект с интерфейсом `Filter` определяет правила отбора элементов. Отбор элементов может осуществляться как перебором всего множества, так и путем использования индексов (хэш-индексы или B-деревья).
- `TransformedDataCollection` – при помощи объекта типа `Transformer` осуществляет преобразование типов данных между первичной и вторичной `DataCollection`.

Информация о `DataCollection` (или цепочке) для работы с биологическими БД задается при помощи специальных файлов конфигураций, где указывается тип `DataCollection`, трансформер, путь к файлам, строка соединения для реляционных БД, используемые индексы, и другие параметры. Специальный класс `CollectionFactory` используется для создания экземпляров `DataCollection` на основе файлов конфигураций.

Например (рисунок П.2.3.1), информация из текстового файла БД UNIPROT в первичной `DataCollection` представляется в виде объектов типа `EntryDataElement`, содержащих имя записи и ее текст. `UniprotEntryTransformer` осуществляет разбор данного текста и соответствующая информация представляется в `TransformedDataCollection` в виде

объекта Protein.

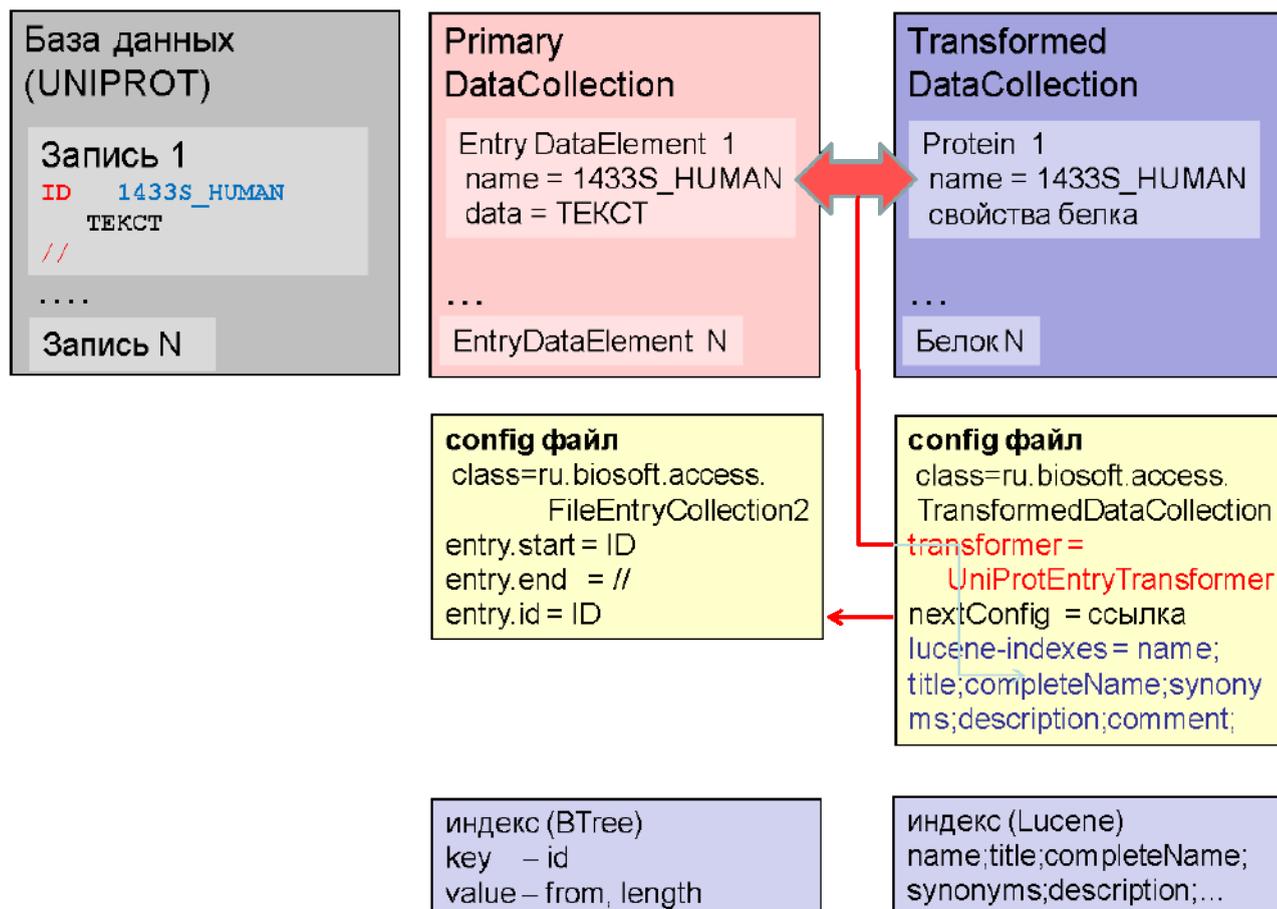


Рисунок П.2.3.1 – Пример цепочки DataCollection для объектно-ориентированного представления данных из базы UNIPROT (см. пояснения в тексте).

Таким образом при помощи цепочек из DataCollection нам удастся четко разделить работу с БД на 2 этапа:

- поиск, чтение, запись и индексация данных – осуществляется первичной DataCollection;
- семантический разбор записи и представление информации в виде Java объекта из онтологии предметной области – осуществляется специальными трансформерами.

На данный момент библиотека ru.biosoft.access и система BioUML содержат более 100 трансформеров для разбора различных биологических данных из более чем 20 БД и форматов.

Класс LocalRepository позволяет организовать репозиторий данных, состоящий из гетерогенных БД. В репозитории хранятся файлы конфигураций для работы с БД, сами данные могут храниться как в репозитории, так и удаленно (в этом случае доступ к данным может осуществляться по протоколам JDBC, SOAP, HTTP и другим).

Объектно-ориентированное представление информации из БД в виде компонентов JavaBeans позволили создать универсальную полнотекстовую поисковую систему на основе технологии Apache Lucene (lucene.apache.org). Для этого свойства объектов отображаются в поля документа Lucene и индексируются. Предопределенные типы данных и их свойства

позволяет единообразным образом строить запросы к различным информационным источникам. При необходимости учесть их специфику индексируются динамические свойства.

Более подробно данная библиотека описана в работе автора (Колпаков, 2011). Базовые классы данной библиотеки свободно доступны в Git репозитории по адресу: <https://github.com/Biosoft-ru/access>

ПРИЛОЖЕНИЕ 2.4 JAVA БИБЛИОТЕКА ГРАФИЧЕСКИХ ОБЪЕКТОВ

Библиотека графических объектов, разработанная автором, обеспечивает графические примитивы, функции и компоненты для визуального представления биологических данных и обеспечивает следующие функции:

1. рисование двумерных графических примитивов, таких как линии, прямоугольники, овалы, текст и т.д.;
2. создание комплексных графических объектов, т.е. объектов, частями которого являются другие графические объекты;
3. механизм управления элементами комплексных объектов;
4. ассоциация графических элементов с произвольными объектами приложения, поддержка событий.

Ниже кратко описаны основные классы этой библиотеки. Более подробно она описана в работе автора (Колпаков, 2011). Данная библиотека свободно доступна в Git репозитории по адресу: <https://github.com/Biosoft-ru/graphics>

В основе библиотеки лежат 2 класса `View` и `CompositeView`.

Класс `View` – это базовый класс для всех графических элементов. Класс содержит описания параметров и методов, свойственных всем без исключения графическим элементам.

Содержит следующие методы:

- `getModel/setModel` – возвращает/устанавливает объект приложения, ассоциированный с данным графическим элементом. Этот объект можно интерпретировать как «модель» в стандартном шаблоне «модель – вид – контроллер». В качестве модели может выступать произвольный Java объект;
- `getBounds` – возвращает прямоугольник минимального размера, содержащий данный графический элемент;
- `getShape` – возвращает геометрическую форму, описывающую данный элемент;
- `intersects` – проверка пересечения элемента с прямоугольной областью. Может использоваться, например, при поиске элементов, на которые попадает указатель мыши;
- `getSelectionPriority` – возвращает приоритет, по которому определяется текущий выбранный объект, если таковых имеется несколько;
- `setLocation` – устанавливает положение данного элемента;
- `move` – сдвиг объекта на заданное расстояние относительно предыдущего положения;
- `scale` – коэффициент масштабирования элемента;
- `isActive/setActive` – проверка/установка флага доступности элемента для выделения и добавления в него других элементов;
- `isVisible/setVisible` – проверка/установка флага, отвечающего за необходимость прорисовки данного элемента;
- `paint` – осуществляет прорисовку данного элемента.

Класс CompositeView – базовый класс для комплексных (составных) графических элементов. Комплексный элемент представляет собой набор других графических элементов (View). Компонентами такого элемента могут быть как элементарные элементы (линии, прямоугольники и т.д.), так и другие комплексные элементы. Графический элемент для диаграммы, таким образом, также представляется комплексным элементом. Класс CompositeView наследуется от класса View, расширяя его и добавляя следующие функции:

- `getChildren` – возвращает объект перечислимого типа для доступа к дочерним элементам;
- `elementAt` – возвращает дочерний графический элемент по номеру;
- `insert` – добавляет новый дочерний элемент в указанную позицию;
- `add` – добавляет новый дочерний элемент с указанием взаимного расположения относительно родительского элемента;
- `remove` – удаление элемента из комплекса.

При добавлении элемента в комплекс может быть указано взаимное расположение этого элемента относительно всего комплекса с помощью параметра `mode`. Набор значений для данного параметра определен в виде статических констант класса CompositeView. Имена этих параметров интерпретируются следующим образом:

```

X_RL
|  ||
| |--- граница нового элемента
| ---- граница комплекса
----- координата x или y

```

X – координата x:
L – левая граница объекта
C – центр объекта
R – правая граница объекта

Y – координата y:
T – верхняя граница объекта
C – центр объекта
B – нижняя граница объекта

Типы графических элементов

В библиотеке графических объектов реализованы все основные типы элементов, используемых для рисования диаграмм. Ниже приведена таблица с их описаниями:

Таблица п.2.2.1 – Основные классы для представления графических примитивов

Имя класса	Тип	Описание
LineView	базовый	Линия. Задается объектом <code>Pen</code> и двумя крайними точками
ImageView	базовый	Позволяет представить картинку (объект <code>java.awt.Image</code>) в виде графического элемента (View)
HtmlView	базовый	Позволяет представить фрагмент html страницы в виде

		графического элемента
BoxView	базовый	Прямоугольник. Задается размерами, положением, а также объектами Pen и Brush
EllipseView	базовый	Эллипс. Задается размерами, положением, а также объектами Pen и Brush
PoligonView	базовый	Ломаная линия. Задается объектом Pen и координатами точек.
TextView	базовый	Текстовый элемент. Задается текстом, шрифтом, размером и выравниванием.
Complex TextView	комплексный	Текстовый элемент с поддержкой переносов по словам, некоторых тегов html, надстрочного и подстрочного текста.
ArrowView	комплексный	Стрелка. Представляет собой объект PoligonView с добавлением с одного или обоих концов указателя (треугольник, ромбик, овал и т.д.)
Ruler	комплексный	«Линейка». Линия нанесенными делениями и указаниями значений через определенные интервалы.

Отрисовка и редактирование графических объектов

Данная библиотека используется для отрисовки и редактирования диаграмм как в настольной, так и в серверной версии ПК BioUML.

В настольной версии, когда ПК BioUML запускается как Java приложение на компьютере пользователя, библиотека классов представляет специальный компонент (ViewPane), который наследуется от `javax.swing.JPanel` и, следовательно, может быть интегрировано в любое Swing приложение.

ViewPane получает на вход CompositeView и отрисовывает его на экране пользователя, используя графические методы Graphics2D библиотеки Java Swing. ViewPane также поддерживает обработку событий для элементов элементов CompositeView, например, приложение может получить элемент, ассоциированный с графическим элементом по нажатию мыши на нем.

Основным расширением класса ViewPane является класс ViewEditorPane, который предоставляет дополнительную функциональность, необходимую для редактирования графических элементов, а именно:

- добавление новых элементов;
- перемещение элементов (с помощью мыши или клавиатуры);
- изменение размеров элементов;
- удаление элементов.

При всех этих действиях может возникнуть необходимость изменения модели. Например, перемещение элемента может повлечь смену родительского элемента для него. Работа с моделью осуществляется через специальный интерфейс ViewEditorHelper.

Для отрисовки и редактирования диаграмм в веб-браузере (серверная версия) для каждого Java класса, наследуемого от View разработаны соответствующие JavaScript классы-партнеры. Они позволяют нарисовать изображение в браузере, используя технологию canvas, введенную в стандарт HTML 5.0. Таким образом, в серверной версии ПК BioUML

построение и редактирование диаграмм происходит следующим образом:

1. Java код на сервере строит `CompositeView` для соответствующей диаграммы;
2. `CompositeView` сериализуется в формате json и передается в веб-браузер клиента;
3. на стороне клиента из переданного json текста строятся соответствующие JavaScript объекты;
4. JavaScript объекты отрисовывают себя в веб-браузере на холсте (canvas);
5. при редактировании диаграммы в веб-браузере специальный JavaScript компонент передает соответствующее событие серверу;
6. Java код на стороне сервера обрабатывает полученное событие, изменяет диаграмму и переходит к шагу 1.

Таким образом, разработанная архитектура позволяет использовать один и тот же Java код для отрисовки и редактирования графических объектов, как в Java Swing так и в клиент-серверных приложениях.

ПРИЛОЖЕНИЕ 3.1 ОСНОВНЫЕ JAVA ИНТЕРФЕЙСЫ ДЛЯ ЗАДАНИЯ ГРАФИЧЕСКОЙ НОТАЦИИ

На рисунке П.3.1.1 представлены Java классы и интерфейсы (пакета biouml.model) для формального определения графической нотации в программе BioUML.

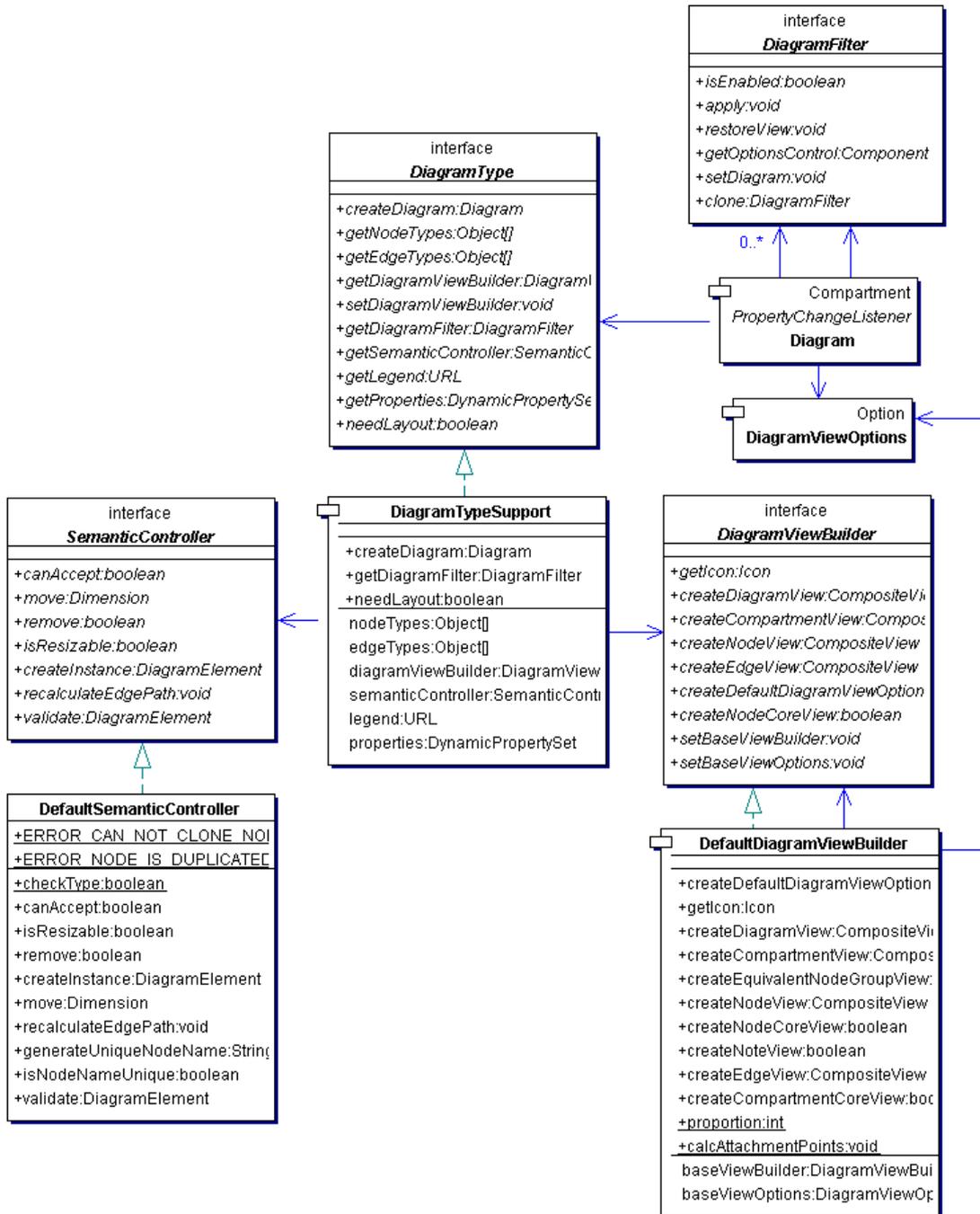


Рисунок П.3.1.1 – Интерфейсы Java для формального определения графической нотации

Интерфейс **DiagramType** формально определяет графическую нотацию, используемую при рисовании и редактировании диаграммы. Для этого в нем определены следующие методы:

- `Object[] getNodeTypes()` – возвращает типы вершин, которые могут быть добавлены и использованы на диаграмме данного типа. Типы вершин представлены объектами типа `Class`, например, `Substance.class`;
- `Object[] getEdgeTypes()` – возвращает типы ребер, которые могут быть добавлены и использованы на диаграмме данного типа;
- `DiagramViewBuilder getDiagramViewBuilder()` – возвращает правила отрисовки вершин и ребер графа на диаграмме. Данные правила формализованы в виде интерфейса `DiagramViewBuilder` (см. ниже);
- `SemanticController getSemanticController()` – возвращает семантические правила и ограничения, используемые при редактировании диаграммы. Эти правила и ограничения формализованы в виде интерфейса `SemanticController`;
- `URL getLegend()` – возвращает URL для страницы, где описана легенда для данного типа диаграммы. Как правило, легенда представляется в виде некоторого рисунка или таблицы с комментариями;
- `Diagram createDiagram(DataCollection origin, String name)` – создает новый экземпляр диаграммы соответствующего типа с заданным именем и помещает его в указанную `DataCollection`.
- `public boolean needLayout(Compartment node)` – указывает, нужно ли применять автоматическое упорядочивание вершин для данного компартмента или рассматривать его как простую вершину.

Интерфейс `DiagramViewBuilder` формализует правила отрисовки вершин и ребер графа на диаграмме в виде набора методов, которые создают соответствующий графический элемент/представление:

- `CompositeView createDiagramView(Diagram diagram, Graphics g)`
создает графическое представление для всей диаграммы;
- `CompositeView createCompartmentView(Compartment compartment, DiagramViewOptions options, Graphics g)` – создает графическое представление для заданного компартмента и вложенных в него вершин и ребер;
- `CompositeView createNodeView(Node node, DiagramViewOptions options, Graphics g)` – создает графическое представление для заданной вершины;
- `boolean createNodeCoreView(CompositeView container, Node node, DiagramViewOptions options, Graphics g)` – создает графическое представление для заданной вершины без заголовка. Возвращает `true` если к вершине нужно отдельно добавить соответствующий заголовок;
- `CompositeView createEdgeView(Edge edge, DiagramViewOptions options, Graphics g)` – создает графическое представление для заданного ребра.

Данный интерфейс использует паттерн проектирования `Builder`, когда построение сложного объекта (графического представления всей диаграммы) разбивается на ряд более простых шагов: построение графических представлений для компартментов, вершин и ребер. Процесс построения графического представления для диаграммы можно представить в виде дерева:

- `createDiagramView`
 - `createCompartmentView` (для всех компарментов)
 - `createNodeView` (для всех вершин компармента)
 - `createNodeCoreView`
 - если нужно, добавить заголовок к представлению вершины
 - `createEdgeView` (для всех вершин компармента)

В интерфейсе `DiagramViewBuider` определено несколько вспомогательных методов:

- `javax.swing.Icon Icon getIcon(Object type)`
возвращает иконку для панели инструментов для вставки на диаграмму объекта соответствующего типа.
- `DiagramViewOptions createDefaultDiagramViewOptions()`
возвращает графические настройки, используемые при рисовании диаграммы данного типа.

Интерфейс `SemanticController` формализует семантические правила и ограничения, используемые при редактировании диаграммы. При каждом событии (вставка, удаление, перемещение объекта на диаграмме или изменение его размеров), редактор обращается к объекту `SemanticController` чтобы проверить допустимость этого действия и/или внести в диаграмму дополнительные изменения, чтобы восстановить ее семантическую целостность и корректность. Если `SemanticController` обнаруживает, что какое-то действие нарушает его правила, он выбрасывает исключение, в котором описана причина ошибки. В этом случае, редактор выводит на экран сообщение, переданное в исключении и отменяет действие пользователя. Для этого в интерфейсе `SemanticController` определены следующие методы:

- `boolean isResizable(DiagramElement diagramElement)`
указывает, может ли размер данной вершины или компармента быть изменен.
- `public boolean canAccept(Compartment compartment, DiagramElement de)`
указывает, можно ли данный элемент диаграммы поместить в указанный компармент. При помощи этого метода задаются семантические ограничения на вложенность биологических и математических объектов друг в друга.
- `Dimension move(DiagramElement de, Compartment parent, Dimension offset, Rectangle oldBounds) throws Exception`
передвигает вершину (`de`) в заданный компармент (`parent`). Параметр `offset` указывает расстояние, на которое вершина должна быть сдвинута относительно исходного положения. Возвращает расстояние, на которое вершина была передвинута. Если вершина передвигается в другой компармент, то вначале вызывается метод `canAccept`, чтобы проверить допустимость этого действия.
- `boolean remove(DiagramElement de) throws Exception`
удаляет заданный элемент, осуществляя необходимые изменения в диаграмме, чтобы сохранить семантическую целостность. Например, при удалении вещества с диаграммы

удаляются все реакции, в которых оно участвует.

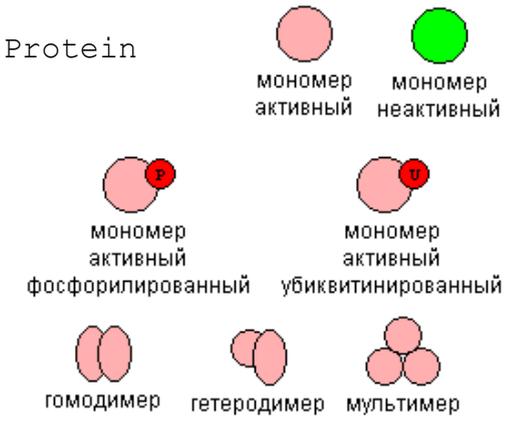
- `DiagramElement createInstance (Compartment compartment, Object type, Point point, ViewEditorPane editor)`
создает новый элемент заданного типа (`type`) в заданной позиции заданного компартамента.
- `DiagramElement validate(Compartment compartment, DiagramElement de) throws Exception`
этот метод вызывается редактором диаграмм после обработки любого действия пользователя, чтобы проверить и при необходимости восстановить семантическую целостность диаграмм. Как правило, этот метод используется, когда главный `SemanticController` вызывает вложенный `SemanticController` при реализации сложных типов диаграмм, например на диаграмму типа SBML накладывается графическая нотация SBGN.

Приложение 3.2 Исходная графическая нотация ПК BioUML

В таблице п.3.2.1 представлены компоненты (изображения вершин и ребер) исходной графической нотации ПК BioUML.

Предложенная графическая нотация обеспечивает процесс реконструкции и моделирования биологических систем на молекулярно-клеточном уровне и приближена к естественной работе пользователя-биолога (рисунок П.3.2.1), который начинает работу с полуформального представления исследуемой биологической системы в виде семантической сети, и заканчивает – математической моделью системы, которая качественно и/или количественно может воспроизвести требуемые экспериментальные наблюдения.

Таблица П.3.2.1 – Исходная графическая нотация ПК BioUML

Java класс*	Изображение	Комментарий	Тип диаграмм				
			1	2	3	4	5
<i>Биологические понятия и объекты</i>							
Concept	 процесс функция	биологическое понятие (концепция, состояние, процесс и функция)	+			+	
Substance	 вещество	лиганд или метаболит небелковой природы	+	+	+	+	+
Gene	 ген	ген	+	+	+	+	+
RNA	 РНК	РНК	+	+	+	+	+
Protein	 мономер активный мономер неактивный мономер активный фосфорилированный мономер активный убиквитинированный гомодимер гетеродимер мультимер	функциональное состояние - отображается цветом: - розовый: активная форма - зеленый – неактивная форма. модификации – отображаются в виде красного круга и буквы, показывающей модификацию: - P – фосфорилирование - U – убиквитинирование, и т.п. структура – отображается формой белка или комплекса	+	+	+	+	+
Compartment	 компармент	компармент клетки или организма	+	+	+	+	+
Cell	 клетка или клеточная культура	клетка или клеточная культура	+			+	
<i>Связи между биологическими объектами</i>							

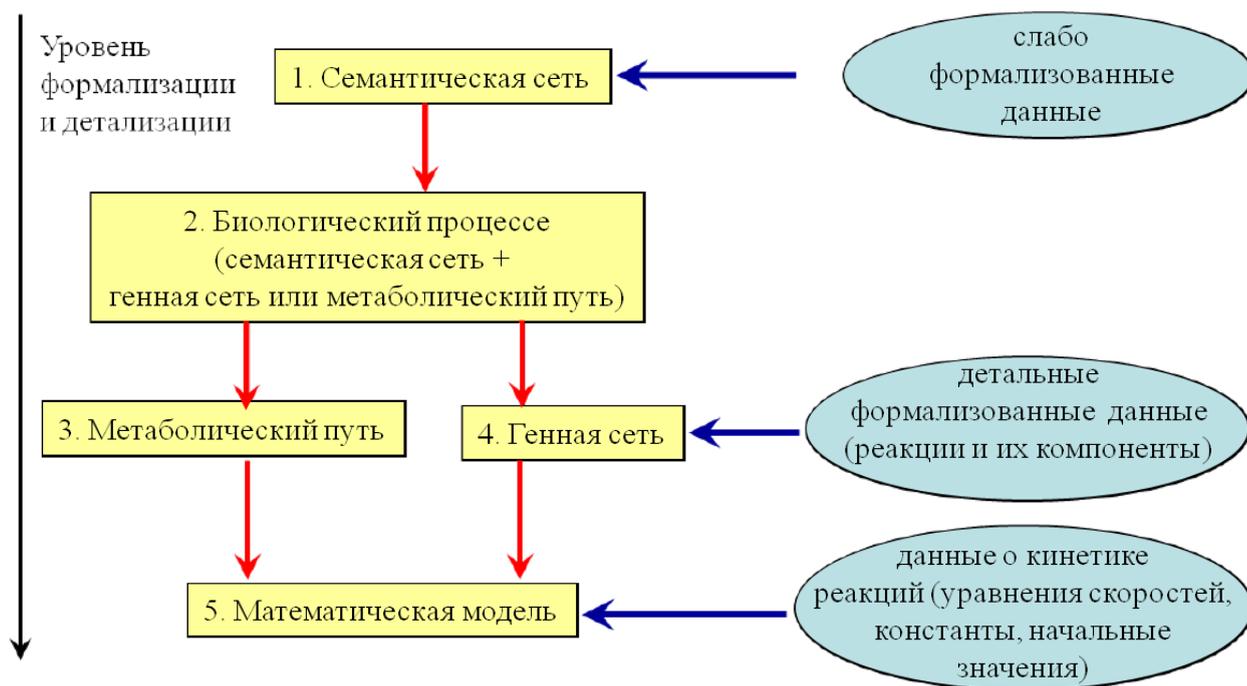


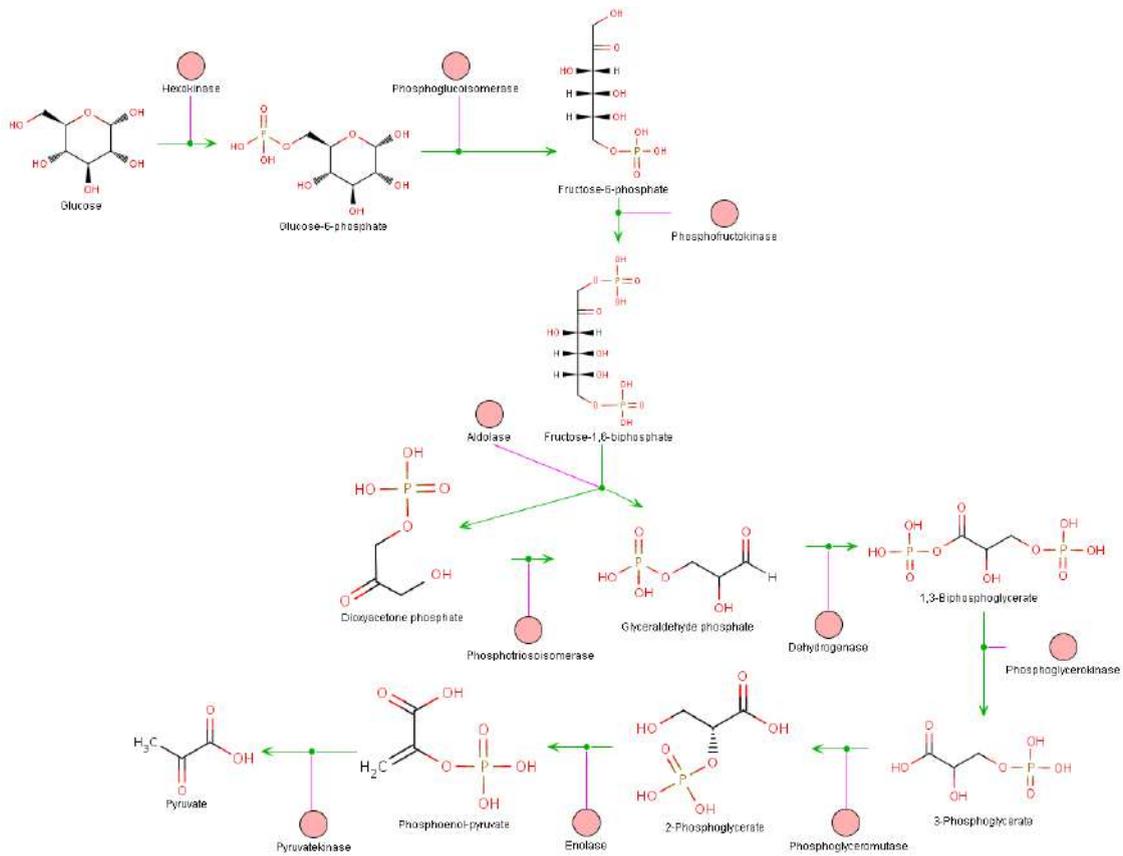
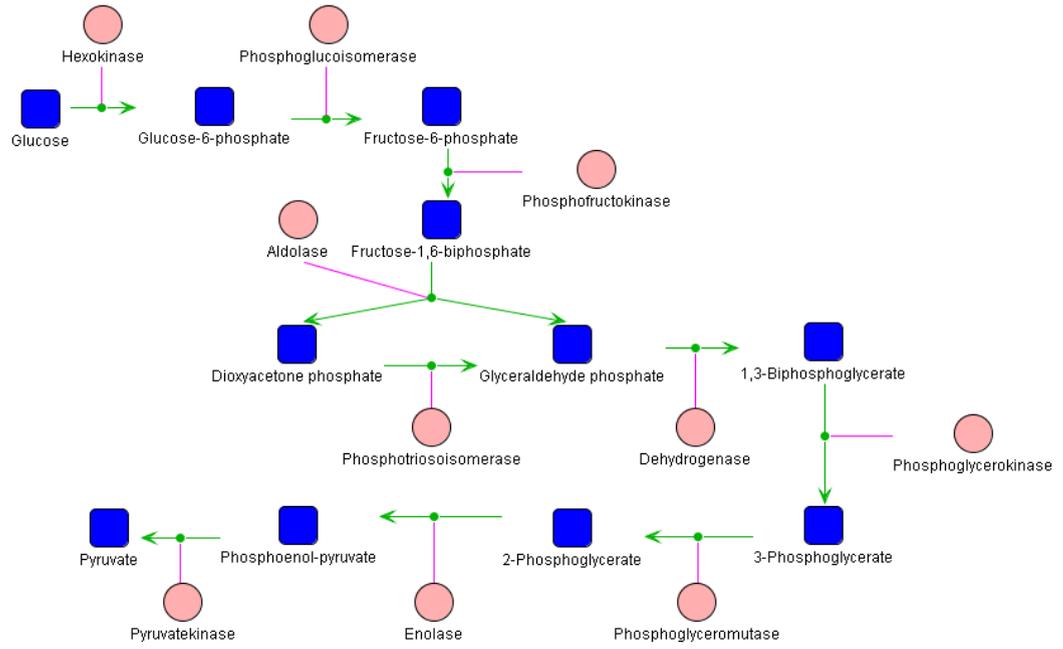
Рисунок П.3.2.1 – Использование различных типов диаграмм для реконструкции, формального описания и моделирования биологических систем

В работе автора (Колпаков, 2011, глава 6) рассмотрены примеры использования разработанной графической нотации для формального описания и визуального моделирования широкого круга биологических систем. Здесь же мы приведем только несколько примеров.

Метаболические пути

Представление и анализ кинетических схем при помощи графов был предложен во второй половине 60-х годов и развит в работах М.В. Волькенштейна, Б.Н. Гольдштейна, Е. Кинга, К.А. Альтмана и других (King and Altman, 1956; Volkenstein and Goldstein, 1966; Гольдштейн Б.Н., 1989). На данный момент представление метаболических путей в виде графов является общепринятым подходом.

На рисунке п.3.2.2.а приведена упрощенная схема гликолиза, представленная в графической нотации BioUML. На рисунке п.3.2.2.б представлена та же схема, но низкомолекулярные вещества представлены на ней не в виде пиктограмм, а в виде структурных формул, что входит в разработанную графическую нотацию и поддерживается ПК BioUML – структурные формулы задаются в MOL формате и отображаются при помощи библиотеку CDK.



б)

Рисунок П.3.2.2 – Упрощенная схема гликолиза в графической нотации BioUML.
а) с использованием пиктограмм; б) с использованием структурных формул

Генные сети

Графическая нотация BioUML для генных сетей является развитием графической нотации системы GeneNet, которая была разработана для представления структуры генных сетей в виде компартиментализованных графов (Kolpakov et al., 1998).

Тип диаграммы «генная сеть» в BioUML соответствует графической нотации GeneNet за следующими исключениями:

- биологические понятия разделены на концепции, состояния, процессы и функции; для каждого из них предусмотрено свое графическое отображение (см. таблицу п.3.2.1);
- расширен список отображаемых модификаций белков (см. таблицу п.3.2.1);
- для соответствия формату SBML, в графической BioUML нотации не используется понятие «регуляторное событие», такие взаимодействия между объектами теперь описывается как модификатор реакции;
- добавлена возможность размещать примечания, что позволяет отобразить на диаграмме недостаточно формализованную информацию и специальные заметки.

Новый, по сравнению с GeneNet, тип диаграммы «генная сеть + семантическая сеть» позволяет графически представить недостаточно формализованные данные, а так же опосредованные взаимодействия между биологическими объектами (без детализации таких взаимодействий). Пример такой диаграммы представлен на рисунке п.3.2.3, показывающем пересечение сигнальных путей, вовлекающих активацию транскрипционных факторов AP-1 и CRE в регуляцию гена тирозин гидроксилазы.

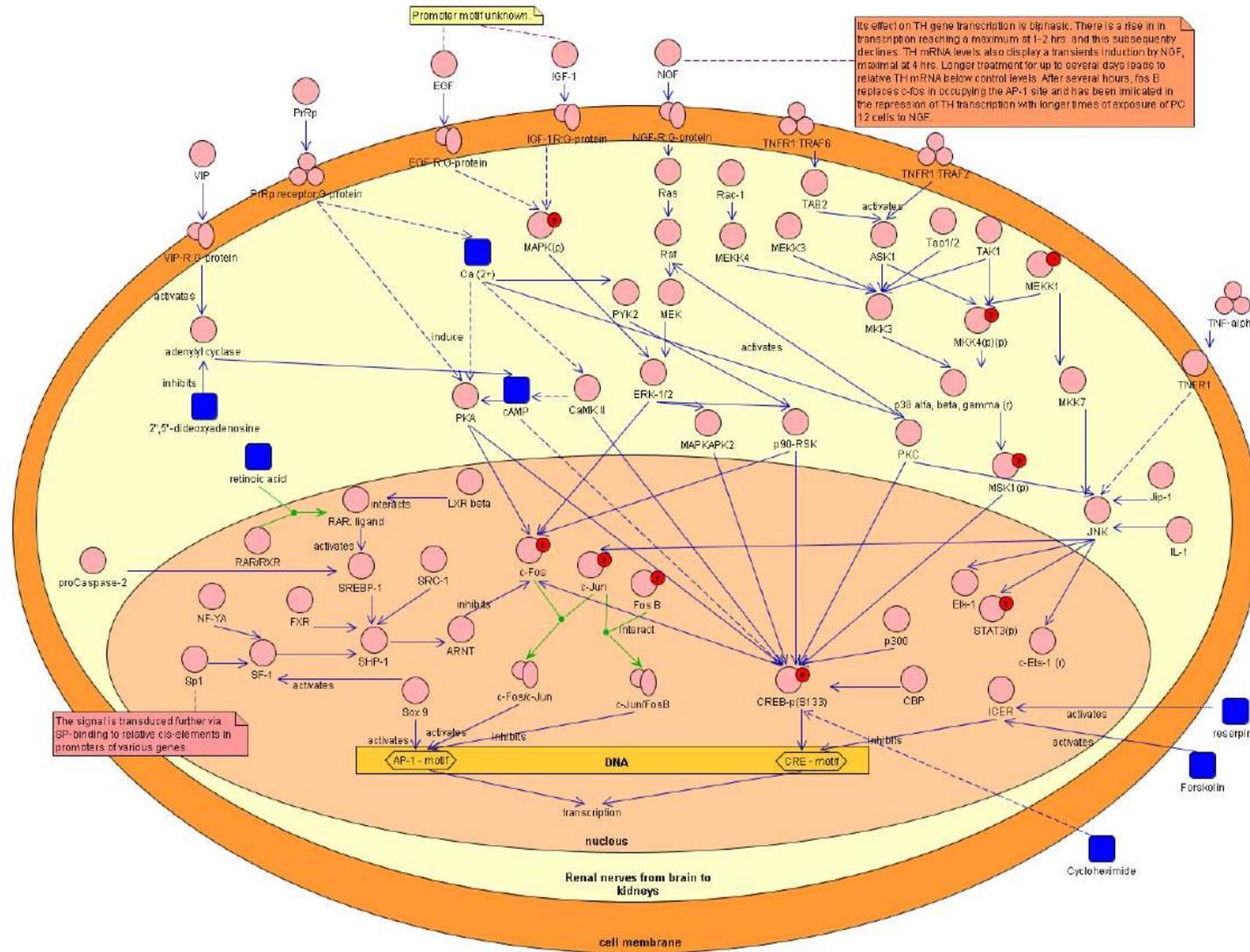


Рисунок П.3.2.3 – Пересечение сигнальных путей, вовлекающих активацию транскрипционных факторов AP-1 и CRE в регуляции гена тирозин гидроксилазы

ПРИЛОЖЕНИЕ 3.3 РАСШИРЕНИЕ И синхронизация SBGN PROCESS DIAGRAM И ANTIMONY

Спецификация аннотаций

Аннотация позволяет расширять язык Antimony, подобно тэгу <annotation> в формате SBML.

Аннотация имеет формат:

```
@annotation_name object_id.property = value
```

```
@annotation_name object_id = {property1 = value1,property2 = value2}
```

Каждая аннотация объявляется (импортируется) в Antimony файле в любом месте, но до первого использования соответствующей аннотации, используя следующий формат:

```
@import annotation_name=URL
```

Пример:

```
@import sbgn=https://biouml.org/antimony/sbgn.yaml
```

По соответствующему URL должен находиться YAML файл, описывающий все свойства и их формат, используемые в соответствующей аннотации. Соответственно, Antimony парсер ПК BioUML использует эту информацию для проверки корректности аннотаций.

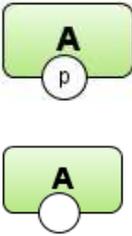
В YAML файле аннотации описываются свойства и типы данных:

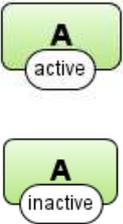
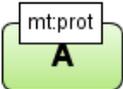
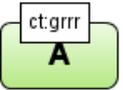
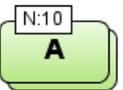
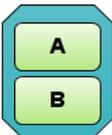
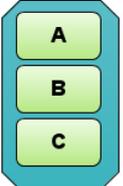
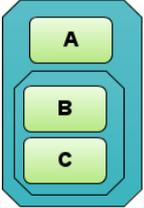
```
name: annotation_name
description: annotation_description
properties:
  - name:
    type: int|float|string|enum|array|map|any
    enum:
      - value_1: description
      - value_2: description
    description:
    properties:      если map
      - nested_property_1:
        type:
        description:
```

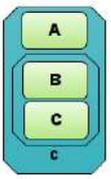
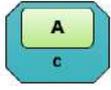
На данный момент для ПК BioUML предложены следующие аннотации:

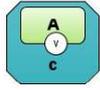
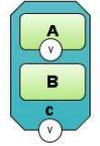
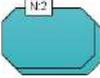
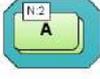
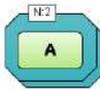
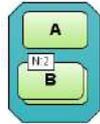
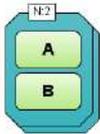
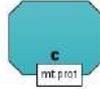
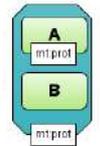
- @sbgn – свойства для SBGN нотации
- @glycans – представление структуры гликанов
- @smiles – представление структуры химических веществ, используя нотацию SMILES
- @rdf – аннотация (ссылки на БД, статьи и т.п.)
- @layout – расположение вершин и ребер на диаграмме

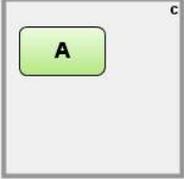
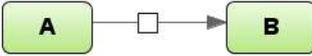
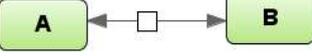
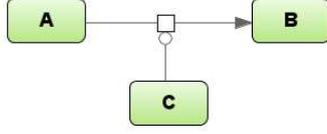
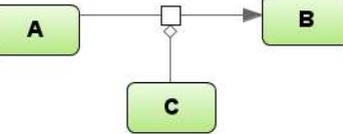
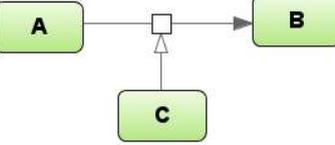
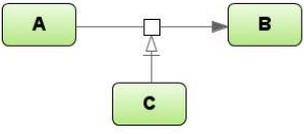
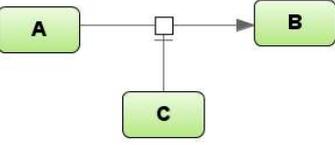
Таблица п.3.3.1 – Соответствие расширений SBGN Process Diagram и аннотаций Antimony

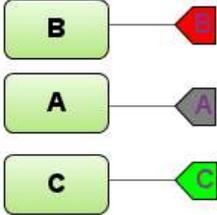
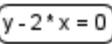
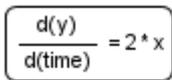
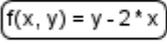
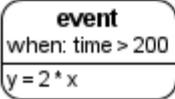
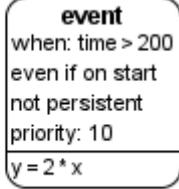
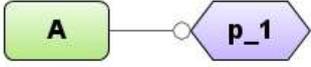
1. Entities		
Description	SBGN glyph	Antimony code
Macromolecule		species A
Simple chemical		species A @sbgn A.type = "simple chemical"
Unspecified entity		species A @sbgn A.type = "unspecified"
Perturbing agent		species A @sbgn A.type = "perturbing agent"
Nucleic Acid Feature		gene A; @sbgn A.type = "nucleic acid feature";
2. State variable (modification in Transpath terms)		
Entity A has state variable with name var and value val		species A; @sbgn A.structure = "A{val@var}";
Entity A has state variable with name Y701 and value P (phosphate group at certain location)		species A; @sbgn A.structure = "A{P@Y701}";
Entity has unnamed state variable with no name and value p (phosphorylated) as opposed to A with state variable with empty name (not phosphorylated)		species A1, A2; A1 is "A"; A2 is "A"; @sbgn A1.structure = "A{p}"; @sbgn A2.structure = "A{}";

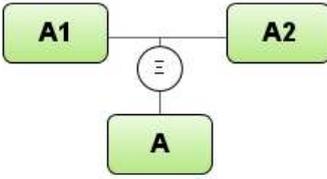
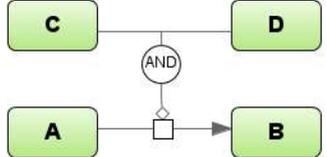
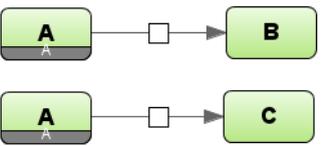
		species A1, A2; A1 is "A"; A2 is "A"; @sbgn A1.structure ="A {active}"; @sbgn A2.structure ="A {inactive}";
3. Unit of information (UOI)		
Material type - protein		species A; @sbgn A.structure ="A[mt:prot]";
Conceptual type - Gene regulatory region		species A; @sbgn A.structure ="A[ct:grr]";
4. Multimer		
		species A; @sbgn A.structure ="(A)10";
5. Complexes		
Description	SBGN image	Structure
Complex with two components		species c; @sbgn c.type ="complex"; @sbgn c.structure ="(A:B)"; ИЛИ @sbgn c = {type="complex", structure="(A:B)"}
Complex with three components		species c; @sbgn c.type ="complex"; @sbgn c.structure ="(A:B:C)";
Complex with nested complex		species c; @sbgn c.type ="complex"; @sbgn c.structure ="(A:(B:C))";

Complex with one known entity and unknown (or omitted) other components		species c; @sbgn c.type = "complex"; @sbgn c.structure = "(A:)"
Complex with unknown or omitted components.		species c; @sbgn c.type = "complex"; @sbgn c.structure = "(:)";
		species c; @sbgn c.type = "complex"; @sbgn c.structure = "c(c_2(:))";
6. Named complexes		
In some cases we would like to specify name of the complex		
Description	SBGN image	Structure
Named complexes		species c; @sbgn c.type = "complex"; @sbgn c.structure = "c(A:B:C)"
		species c; @sbgn c.type = "complex"; @sbgn c.structure = "c(A:(B:C))";
		species c; @sbgn c.type = "complex"; @sbgn c.structure = "c(A:)"
		species c; @sbgn c.type = "complex"; @sbgn c.structure = "c(:)";
7. Complexes with annotation		
Complex with state variable		species c; @sbgn c.type = "complex"; @sbgn c.structure = "c(:){v}";

		species c; @sbgn c.type = "complex"; @sbgn c.structure = "c(A {v} :)";
		species c; @sbgn c.type = "complex"; @sbgn c.structure = "c(A {v} :B) {v}";
Complex with multimers		species c; @sbgn c.type = "complex"; @sbgn c.structure = "(:)2";
		species c; @sbgn c.type = "complex"; @sbgn c.structure = "(A)2";
		species c; @sbgn c.type = "complex"; @sbgn c.structure = "(A :)2";
		species c; @sbgn c.type = "complex"; @sbgn c.structure = "(A :B)2";
		species c; @sbgn c.type = "complex"; @sbgn c.structure = "(A :B)2";
		
Complex with Unit Of Information		species c; @sbgn c.type = "complex"; @sbgn c.structure = "(A[mt:prot])";
		species c; @sbgn c.type = "complex"; @sbgn c.structure = "(A[mt:prot]:B)[mt:prot]";

Compartment		compartment c; species A in c;
Reaction		species A,B; R: A=>B; k*A;
Reversible reaction		species A,B; R: A->B; k*A;
Reaction catalysis		species A,B,C; R: A->B; k*A; R_mod: C-(R;
Reaction modulation		species A,B,C; R: A->B; k*A; R_mod: C-<diamond>R; species A,B,C; R: A->B; k*A; R_mod: C-(R; @sbgn R_mod.type = "modulation";
Reaction stimulation		species A,B,C; R: A->B; k*A; R_mod: C-(R;
Reaction necessary stimulation		species A,B,C; R: A->B; k*A; R_mod: C- >R; @R_mod.sbgn.type="necessary stimulation"
Reaction inhibition		species A,B,C; R: A->B; k*A; R_mod: C- R; @sbgn R_mod.type="inhibition"

Connection port (red -output, green -input, gray -contact)		<pre>model m(A, <B, >C) species A,B,C; end</pre>
Elements not in SBGN		
Description	BioUML Glyph	Antimony
Scalar rule		$y := 2 * x$
Algebraic rule		$0 = y - 2 * x$
Rate rule		$y' = 2 * x$
Function		<pre>function f(x,y) y-2*x end</pre>
Event		<pre>event: at time>200, t0 = true, persistent = true : y = 2*x;</pre>
Event		<pre>event: at time>200,priority = 10, fromTrigger = true : y = 2*x;</pre>
Constraint		<pre>const_1: assert y>=2*x else "Y must be < 2*x"</pre>
Additional		
Phenotype		<pre>p_1: =>; p_1_mod: A -(p_1; @sbgm p_1.type = "phenotype";</pre>

<p>Equivalence</p>		<p>A1 subtype A; A2 subtype A:</p>
<p>Logical operator</p>		<p>species A,B,C,D; R: A->B; k*A;</p> <p>R_mod: (C & D)-(R;</p> <p>R_mod1: C-(R; R_mod2: D-(R;</p> <p>R_mod.C_edge.path R_mod.AND_edge.path R_mod.path</p> <p>@R_mod.path=[]</p>
<p>Clone (two nodes corresponds to the same entity)</p>		<p>species A, B, C;</p> <p>R: A=>B; R_1: A=>C;</p> <p>@sbgn A.clone = {reactions=[R, R_1],title="A"}</p>

ПРИЛОЖЕНИЕ 5.1 ПЕРСПЕКТИВА GTRD для ПК BioUML

Для работы пользователей с БД GTRD используется отдельная инсталляция ПК BioUML со специально разработанной перспективой. Для этого на стартовой странице вынесены и сгруппированы соответствующие методы (рисунок П.5.1.1):

- **Searching, browsing** – позволяет найти эксперименты в БД по заданному типу данных (ChIP-seq, ChIP-exo, DNase-seq и т.п.) и/или ТФ, антителу (для ChIP-seq/exo экспериментов), клеточной линии, воздействию, идентификатору в GEO или GTRD. Результаты поиска представлены в виде таблицы (рисунок П.5.1.2). При выборе ряда из этой таблицы в панели слева, снизу отображается подробная информация об эксперименте, включая метрики качества. В панели слева, снизу пользователь может нажать гиперссылку, чтобы посмотреть найденные пики (районы связывания) в виде таблицы (рисунок 5.2.9) или в геномном браузере.
- **Genome browser**
 - **Display tracks** – позволяет находить и открывать в геномном браузере треки для заданного вида организма по типам данных (мета-кластеры; кластеры; пики для отдельных пик-коллеров; открытый хроматин; предсказанные ССТФ, используя весовые матрицы из БД HOCOMOСO; районы повторов, на которые не могут быть выровнены уникально прочтения – unmappable regions; РСТФ для аллель специфичных вариантов из БД ADASTRA) и соответствующим дополнительным полям.
 - **Display per TF workflow results** – отображает все треки, связанные с построением мета-кластеров для заданного ТФ (рисунок П.5.1.3).
 - **Advanced search**
 - **Binding sites near the specified gene** – поиск РСТФ вблизи заданного гена (рисунок 5.2.11).
 - **Genes regulated by the specified transcription factor** – поиск генов, потенциально регулируемых заданным ТФ. Для этого задается размер промоторного района или окрестности вокруг гена, где должны находиться экспериментально подтвержденные РСТФ.
 - **Filter cell type experiments** – отбирает аннотированные экспериментальные данные по клеточным типам и заданному типу экспериментов.

Для обучения пользователя разработанным интерфейсом на стартовой веб-странице (<https://gtrd.biouml.org>) представлены в виде набора последовательных слайдов ("карусель") основные сценарии использования (use cases) – рисунок П.5.1.4.

databases

- databases
 - EnsemblArabidopsisThaliana91
 - EnsemblChicken104
 - EnsemblFruitfly91
 - EnsemblHuman85_38
 - EnsemblMouse81_38
 - EnsemblNematoda91
 - EnsemblRat91
 - EnsemblSaccharomycesCerevisiae91
 - EnsemblSchizosaccharomycesPombe91
 - EnsemblZebrafish92
 - GTRD
 - Data
 - ATAC-seq experiments
 - ChIP-exo experiments
 - ChIP-seq HM experiments
 - DNase experiments
 - FAIRE-seq experiments
 - MNase-seq experiments
 - bigBed
 - clusters
 - experiments
 - generic
 - peaks
 - Dictionarys
 - HOCOMOCO v11
 - Utils

Start page

Searching, browsing

Experimental design: ChIP-seq Transcription Factors Search Browse as table

Enter transcription factor, antibody, cell line, treatment or GTRD/GEO ID

Genome browser

Display tracks

Organism	Human (Homo sapiens)
Data type	meta clusters

Show

Display per TF workflow results

Organism	Human (Homo sapiens)
Transcription factor	(not selected)

Show

Advanced search

Binding sites near the specified gene

Organism	Homo sapiens
Gene symbol or ID	Any
Transcription factor	Any
Data set	meta clusters
Max gene distance	5000
Output type	Open in genome browser

Run

Filter cell type experiments

All experiments for cell types

Organism	Homo sapiens
Cluster	
Cell type	
Choose available experiment types	(no selection)

Show

Genes regulated by the specified transcription factor

Method	Analysis of TF binding
TF binding site location	promoter[-100,+10]
Organism	Human (Homo sapiens)
Transcription factor	(not selected)

Run

Regulatory mutations
check your SNVs for regulatory potential

Import VCF file Upload file

Рисунок П.5.1.1 – Стартовая страница для перспективы GTRD

Start page

Gene Transcription Regulation Database

- [Documentation, help \(wiki pages\)](#)
- [Transcription factor classification tree](#)

Searching, browsing

Experimental design: ChIP-seq Transcription Factors Browse [as table](#)

Search
Enter transcription factor, antibody, cell line, treatment or GTRD/GEO ID

Search Info GTRD Experiment

EXP030850

Transcription factor class: 6.3.1.0.1
Transcription factor: TP53
Antibody: mouse monoclonal anti-human p53 (BD Pharmingen, cat# 554294)
Cell: 566
Specie: Human (Homo sapiens)
Treatment: Doxo
Quality metrics:
 ● FNCM(GEM) = 0.0347952405868333
 ● FNCM(MACS) = NaN
 ● FNCM(PICS) = 0.319832996324322
 ● FNCM(SISSRS) = 0.197969961154192
 ● FPCM = 1.52384072944295
 ▲ NRF = NaN

Search result

Showing 1 to 50 of 229 entries

ID	Path	Name	Antibody	TF class	TF title	Cell	Treatment	Specie	External References	Is control experiment	Control	Articles
0	EXP030850	EXP030850	mouse mo...	6.3.1.0.1	TP53	lymphoblastoid cells	Doxo	Human (Homo sapiens)	GSM1142696, GSE46991, 24120139	false	EXP030846	[Wang X, Zeron-Me...
1	EXP030851	EXP030851	mouse mo...	6.3.1.0.1	TP53	lymphoblastoid cells	vehicle	Human (Homo sapiens)	GSM1142697, GSE46991, 24120139	false	EXP030849	[Wang X, Zeron-Me...
2	EXP030853	EXP030853	mouse mo...	6.3.1.0.1	TP53	lymphoblastoid cells	nutilin-3	Human (Homo sapiens)	GSM1142700, GSE46992, 24120139	false	EXP030852	[Wang X, Zeron-Me...
3	EXP030855	EXP030855	mouse mo...	6.3.1.0.1	TP53	lymphoblastoid cells	ionizing radiation	Human (Homo sapiens)	GSM1142702, GSE46993, 24120139	false	EXP030854	[Wang X, Zeron-Me...
			anti-p53 mouse					Human				

Рисунок П.5.1.2 – Результаты поиска ChIP-seq экспериментов для ТФ tp53 в БД GTRD

Start page | PEAKS034633 X | chromosomes GRCh38 X

First Previous Page 1 of 1741 Next Last Showing 1 to 50 of 87004 entries Show 50 entries

ID	Sequence (chromosome) name	From	To	Length	Strand	Type	Property: -log10(pvalue)	Property: -log10(qvalue)	Property: abs_summit	Property: fold_enrichment	Property: name	Property: pileup
1	1	775214	775340	127	?	unsure	5.73815	3.01216	775272	4.19661	MACS2_TP53	4
2	1	778627	778827	201	?	unsure	13.7565	10.3736	778707	8.57871	MACS2_TP53	10
3	1	778909	779014	106	?	unsure	4.31839	1.84073	778912	3.35728	MACS2_TP53	3
4	1	826927	827205	279	?	unsure	5.73815	3.01216	827087	4.19661	MACS2_TP53	4
5	1	869810	870009	200	?	unsure	9.10931	6.02395	869916	6.23906	MACS2_TP53	7
6	1	875394	875888	495	?	unsure	45.3529	41.2646	875677	21.8223	MACS2_TP53	25
7	1	883008	883302	295	?	unsure	21.0695	17.4064	883186	11.7505	MACS2_TP53	13
8	1	898457	898631	175	?	unsure	15.575	12.1035	898525	9.23253	MACS2_TP53	10

Search Info Filters Search result

GTRD Experiment

First Previous 1 2 3 4 5 Next Last Showing 1 to 50 of 229 entries Show 50 entries

ID	Path	Name	Antibody	TF class	TF title	Cell	Treatment	Specie	External References	Is control experiment	Control	Articles
0	EXP030850	EXP030850	mouse mo...	6.3.1.0.1	TP53	lymphoblastoid cells	Doxo	Human (Homo sapiens)	GSM1142696, GSE46991, 24120139	false	EXP030848	[Wang X, Zeron-Me...
1	EXP030851	EXP030851	mouse mo...	6.3.1.0.1	TP53	lymphoblastoid cells	vehicle	Human (Homo sapiens)	GSM1142697, GSE46991, 24120139	false	EXP030849	[Wang X, Zeron-Me...
2	EXP030853	EXP030853	mouse mo...	6.3.1.0.1	TP53	lymphoblastoid cells	nutlin-3	Human (Homo sapiens)	GSM1142700, GSE46992, 24120139	false	EXP030852	[Wang X, Zeron-Me...
3	EXP030855	EXP030855	mouse mo...	6.3.1.0.1	TP53	lymphoblastoid cells	ionizing radiation	Human (Homo sapiens)	GSM1142702, GSE46993, 24120139	false	EXP030854	[Wang X, Zeron-Me...
4	EXP030857	EXP030857	anti-p53 mouse monoclonal	6.3.1.0.1	TP53	IMR90 (lung)	none	Human (Homo sapiens)	GSM1142704, GSE46994, 24120139	false	EXP030856	[Lyons SK,

▲ NSC = NaN
 ▲ RSC = NaN
 ▲ FRiP_GEM = NaN
 ▲ FRiP_MACS = NaN
 ▲ FRiP_SISSRS = NaN
 ▲ FRiP_PICS = NaN
Control: [EXP030848](#)
Reads:
 READS031316.fastq.gz (fetched from URL)
Alignments: ALIGNS035255.bam (not available)
Peaks: PEAKS034633 (open as [table](#) or in [genome browser](#) or [export](#))
External references:
 • GEO : [GSM1142696](#)
 • GEO : [GSE46991](#)
 • PUBMED : 24120139

Рисунок П.5.1.3 – Пики (районы связывания) для ТФ tp53, определенные в эксперименте EXP030850

Use cases

- ▶ Search ChIP-seq experiments by transcription factor
- ▶ Browse ChIP-seq peaks in genome browser
- ▶ Browse all peaks and clusters for given TF in genome browser
- ▼ Find transcription factor binding sites on a gene
- ▼ Find transcription factor binding sites on gene

Step 1. Search for Jun binding sites near the TP53 gene

Advanced search

Binding sites near the specified gene

Organism	Human (Homo sapiens)
Gene symbol or ID	TP53
Transcription factor	JUN P05412
Data set	macs2 peaks
Cell line	Any
Treatment	Any
Max gene distance	5000
Output type	Open in genome browser

Run

1. Enter TP53 into 'Gene symbol' field and select Jun transcription factor

2. Select Dataset

3. You also can restrict search to specific cell line and treatment

- ▶ Find genes regulated by transcription factor
- ▶ Filter experiments for chose parameters
- ▶ View predicted sites

Рисунок П.5.1.4 – Демонстрация основных сценариев использования BioUML интерфейса к БД GTRD

Приложение 5.2 Веб-интерфейс БД GTRD на платформе BeanExplorer EE

Для интеграции результатов из различных типов NGS экспериментов в контексте заданного клеточного типа и условий, был разработан новый веб-интерфейс с использованием технологии BeanExplorer (<https://github.com/DevelopmentOnTheEdge/beanexplorer>). Данная технология позволяет задать мета-информацию для реляционной БД, по которой далее автоматически генерируется веб-интерфейс.

Например, выбрав пункт меню "Cells -> with Experiments" (рисунок П.5.2.1) пользователь может увидеть список всех клеточных типов в БД GTRD и сводную статистику: сколько и каких экспериментов проаннотировано и проанализировано для соответствующего клеточного типа. Операция "Filter" позволяет выбрать клеточные типы для заданного вида организма, по его названию или идентификатору в связанных онтологиях.

Нажав соответствующую гиперссылку, пользователь может просмотреть список экспериментов для соответствующей клеточной линии (рисунок П.5.2.2). Для этого все эксперименты сгруппированы по типам, каждый из которых показывается в отдельной вкладке. Каждый тип эксперимента имеет соответствующий тип полей. Например, для ChIP-seq экспериментов специфичными являются поля:

- Antibody – антитело, использованное в ChIP-seq эксперименте;
- TF Class – класс соответствующего ТФ, согласно классификации Wingender et al., 2018;
- Uniprot – идентификатор ТФ в БД Uniprot;
- Gene – краткое название гена, кодирующего ТФ.

Далее, по гиперссылке может просмотреть подробный отчет по выбранному эксперименту (рисунок П.5.2.3), который включает ссылки на результаты обработки данного эксперимента. Например, для ChIP-seq эксперимента это будет набор пиков, полученных при помощи различных алгоритмов. Полученные результаты можно как скачать в формате bigBed, так и просмотреть в геномном браузере. Также приводятся все посчитанные метрики качества (Kolnykov et al., 2019).

Также BeanExplorer веб-интерфейс содержит следующие пункты меню:

- Experiments – просмотр списка экспериментов. Фильтр позволяет отобразить эксперименты по виду организма, типу эксперимента, типу клеток, названию ТФ, ссылке на GEO. Нажав на соответствующую гиперссылку, пользователь может просмотреть подробный отчет по выбранному эксперименту (такой же, как на рисунке п.5.2.3).
- TF & cofactors – просмотр списка ТФ и кофакторов, объединенного по БД GTRD, CIS-BP, TFClass и AnimalTFDB 3.0. Фильтр позволяет отобразить эксперименты по виду организма,

типу белка (TF, кофактор, не классифицирован), названию гена или ID в CIS-BP, TFClass или AnimalTFDB 3.0. В поле Classification для каждого вида организма можно посмотреть, насколько полно покрыта коллекция TF из БД CIS-BP, TFClass или AnimalTFDB 3.0 соответствующими экспериментами в БД GTRD:

- AnimalTFDB, Cis-BP, TFClass, GTRD – показывает полный список TF и кофакторов в соответствующей БД;
- GTRD with AnimalTFDB, GTRD with Cis-BP, GTRD with TFClass – пересечение (т.е. общие) TF и кофакторы в GTRD и соответствующей БД;
- GTRD missing in AnimalTFDB, GTRD missing in Cis-BP, GTRD missing in TFClass – TF и кофакторы, которых нет в GTRD, но есть в соответствующей БД. Такие списки позволяют понять, насколько полно покрыты TF и кофакторы для выбранного вида организма данными ChIP-seq (ChIP-exo, ChIP-nexus) экспериментами.
- GTRD unclassified in AnimalTFDB, GTRD unclassified in Cis-BP, GTRD unclassified in TFClass – список белков, чье взаимодействие с ДНК было изучено при помощи экспериментов ChIP-seq (ChIP-exo, ChIP-nexus), но которых нет в соответствующей БД. Это могут быть как потенциальные TF, белки, участвующие в ремоделировании хроматина, так и другие белки, как-то взаимодействующие с ДНК
- Downloads – список файлов с результатами анализа в bigBed или bigWig формате, которые можно скачать с сайта GTRD. Фильтр обеспечивает поиск по виду организма, типу экспериментов, типу данных (пики, кластеры, мета-кластеры), клеточному типу, TF или кофактору, идентификатору в базе GEO.
- TSS – атлас сайтов инициации транскрипции, полученных на основе анализа и объединения данных CAGE-seq экспериментов для человека, мыши, крысы и курицы, а также их сопоставление с данными FANTOM5.
- Enhancers – списки потенциальных энхансеров, выявленные в описанных выше сайтах инициации транскрипции;
- Statistics – on-line статистика по БД GTRD с возможностью получить список результатов заданного типа для заданного организма. Для каждой группы треков приведена гиперссылка, при переходе по которой пользователь получает список соответствующих треков и может скачать, интересующие его.

Разработан стабильный API для ссылок на объекты из БД GTRD. Он основан на соглашениях API, используемых в технологии BeanExplorer. Общий вид ссылки имеет следующий вид:

<https://gtrd.biouml.org/#!/table/схема.таблица/представление/фильтр>

- схема – схема БД (как правило, gtrd_current – текущая версия GTRD)
- таблица – таблица БД, например, experiments, cells и т.п.
- представление – название SQL запроса, используемого для представления данных
- фильтр – фильтр для отбора записей. Общий вид: поле=значение.

Например, описание эксперимента с заданным идентификатором:

https://gtrd.biouml.org/#!/table/gtrd_current.experiments/Details/ID=AEXP000001.

Main Experiments **Cells** TF & cofactors Downloads TSS Enhancers Statistics

Species **with experiments** with external links

Cells: with experiments

Filter

Other columns:

#	Title	References	Source	Experiments
1	IB10 (embryonic stem cells)		cell line	ChIP-seq: 3 (+4 control), TF-2
2	mESCs (mouse embryonic stem cells)	EFO:0004038	stem cell	ChIP-seq: 1020 (+456 control), TF-199 ChIP-exo: 2 (+0 control), TF-2 Histone marks: 722 (+106 control) DNase-seq: 8 ATAC-seq: 349 MNase-seq: 29 RNA-seq: 12
3	epididymal white adipose tissue	UBERON:0001347	adipose tissue	ChIP-seq: 16 (+2 control), TF-3 Histone marks: 29 (+12 control)
4	inguinal white adipose tissue	UBERON:0001347	adipose tissue	ChIP-seq: 6 (+4 control), TF-3 ATAC-seq: 6
5	liver	UBERON:0002107	liver	ChIP-seq: 636 (+279 control), TF-75 ChIP-exo: 8 (+0 control), TF-3 Histone marks: 318 (+91 control) DNase-seq: 85 ATAC-seq: 150 MNase-seq: 68 FAIRE-seq: 44 CAGE-seq: 9

tt

Previous **1** 2 3 4 5 ... 1268 Next 5 entries Showing 1 to 5 of 6,340 entries

Рисунок П.5.2.1 – Обзор и статистика по всем типам экспериментов для всех клеточных линий

Cell report: mESCs (mouse embryonic stem cells)

[Description](#)
[ChIP-seq](#)
[ChIP-exo](#)
[Chromatin](#)
[Histone marks](#)
[RNA-seq](#)
[CAGE-seq](#)

Other columns:

#	ID	Treatment	Antibody	TF class	Uniprot	Gene	Protein
1	EXP000018		sc-8628	3.1.10.5.2	P20263	Pou5f1	POU domain, class 5, transcription factor 1 (NF-A3) (Octamer-binding protein 3) (Oct-3) (Octamer-binding protein 4) (Oct-4) (Octamer-binding transcription factor 3) (OTF-3)
2	EXP000020		AF2018		P48432	Sox2	Transcription factor SOX-2
3	EXP000021		bl1662		Q80Z64	Nanog	Homeobox protein NANOG (ES cell-associated protein 4) (Early embryo specific expression NK-type homeobox protein) (Homeobox transcription factor Nanog)
4	EXP000022		sc-8635		Q9Z1J1	Tcf7l1	Transcription factor 7-like 1 (HMG box transcription factor 3) (TCF-3) (mTCF-3)
5	EXP000179		anti-HA	2.1.5.0.2	P45448	Nr5a2	Nuclear receptor subfamily 5 group A member 2 (Liver receptor homolog 1) (LRH-1)

[Previous](#)
[1](#)
[2](#)
[3](#)
[4](#)
[5](#)
[...](#)
[204](#)
[Next](#)

Showing 1 to 5 of 1,020 entries

Рисунок П.5.2.2 – Детальный отчет по всем экспериментам в GTRD для заданной клеточной линии

Cell report: mESCs (mouse embryonic stem cells)

Description **ChIP-seq** [ChIP-exo](#) [Chromatin](#) [Histone marks](#) [RNA-seq](#) [CAGE-seq](#)

Experiment EXP000018, ChIP-seq

Species: [Mus musculus](#)
 Cell type: [mESCs \(mouse embryonic stem cells\)](#)
 Control: [EXP000019](#)
 Design: [unspecified](#)

Antibody

Antibody: [sc-8628](#)
 TF Class: [3.1.10.5.2](#)
 Gene: [Pou5f1](#)
 Uniprot: [P20263](#)
 Protein synonyms: [POU domain, class 5, transcription factor 1 \(NF-A3\) \(Octamer-binding protein 3\) \(Oct-3\) \(Octamer-binding protein 4\) \(Oct-4\) \(Octamer-binding transcription factor 3\) \(OTF-3\)](#)

References

GEO series: [GSE11724](#)
 GEO samples: [GSM307137](#)
 Article: [Jaenisch R, Bartel DP, Sharp PA, Hannett N, Love J, Gupta S, Volkert TL, Dennis LM, Calabrese JM, Newman J, Wernig M, Johnston WK, Guenther MG, Johnstone S, Brambrink T, Frampton GM, Cole MF, Marson A, Young RA, Levine SS. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell 2008 Aug 8;134\(3\):521-33.](#)

Files

Accession	Data type	Peak caller	Assembly	# elements	File size	Format	Download	Genome browser
PEAKS035242	Peaks	SISSRS	mm10	20984	433KB	bigBed	PEAKS035242_Pou5f1_P20263_SISSRS_2.bb	UCSC
PEAKS035242	Peaks	GEM	mm10	16394	934KB	bigBed	PEAKS035242_Pou5f1_P20263_GEM_2.bb	UCSC
PEAKS035242	Peaks	MACS2	mm10	29054	1.44MB	bigBed	PEAKS035242_Pou5f1_P20263_MACS2_2.bb	UCSC
PEAKS035242	Peaks	PICS	mm10	36497	917KB	bigBed	PEAKS035242_Pou5f1_P20263_PICS_2.bb	UCSC

Quality metrics:

● FNCM(GEM) = 0.550334352498763
 ● FNCM(MACS) = undefined
 ● FNCM(PICS) = 0.947270911341261
 ● FNCM(SISSRS) = 0.713590789177902
 ● FPCM = 3.47782209623542
 ▲ NRF = 0.946009
 ▲ PBC1 = 0.951222
 ▲ PBC2 = 21.091186
 ▲ NSC = 2.350677
 ▲ RSC = 1.526317
 ▲ FRiP_GEM = undefined
 ▲ FRiP_MACS = undefined
 ▲ FRiP_PICS = undefined
 ▲ FRiP_SISSRS = undefined

Рисунок П.5.2.3 – Подробное описание ChIP-seq эксперимента, вместе с метриками качества

ПРИЛОЖЕНИЕ 5.3 СРАВНЕНИЕ GTRD С ДРУГИМИ БД ПО ChIP-SEQ ЭКСПЕРИМЕНТАМ

БД	Кол-во ChIP-seq экспериментов	Кол-во ТФ	Методы для пиков ChIP-seq	Виды
GTRD v23.07 (Kolmykov et al., 2021)	всего: 62 947 <i>H. sapiens</i> : 31 818	всего: 5619 <i>H. sapiens</i> : 2 123	MACS2, MACS, GEM, PICS, SISRrs	<i>H. sapiens</i> , <i>M. musculus</i> , <i>R. norvegicus</i> , <i>D. melanogaster</i> , <i>C. elegans</i> , <i>D. rerio</i> , <i>S. pombe</i> , <i>S. cerevisiae</i> , <i>A. thaliana</i> , <i>Gallus gallus</i>
ChIP-Atlas (Zou et al., 2022)	всего: 90 717 <i>H. sapiens</i> : 45 477	всего: <1715 <i>H. sapiens</i> : <1 029	MACS2	<i>H. sapiens</i> , <i>M. musculus</i> , <i>R. norvegicus</i> , <i>D. melanogaster</i> , <i>C. elegans</i> , <i>S. cerevisiae</i>
CistromeDB (Zheng et al., 2019)	всего: 24 065 <i>H. sapiens</i> : 13 976	всего: ? <i>H. sapiens</i> : ?	MACS2	<i>H. sapiens</i> , <i>M. musculus</i>
NGS-QC (Mendoza-Parra et al., 2013)	всего: 22 398 <i>H. sapiens</i> : 11 597	всего: ? <i>H. sapiens</i> : ?	-	<i>H. sapiens</i> , <i>M. musculus</i> , <i>R. norvegicus</i> , <i>D. rerio</i> , <i>C. elegans</i> , <i>D. melanogaster</i> , <i>S. cerevisiae</i> , <i>A. thaliana</i> , <i>G. gallus</i> , <i>P. troglodytes</i>
ENCODE (Luo et al., 2020)	всего: 3 609 <i>H. sapiens</i> : 2 407	всего: ? <i>H. sapiens</i> : ?	SPP	<i>H. sapiens</i> , <i>M. musculus</i> , <i>D. melanogaster</i> , <i>C. elegans</i>
ChIPBase v3.0 (Huang et al. 2023)	всего: 55 420 <i>H. sapiens</i> : 16 309	всего: <870 <i>H. sapiens</i> : ?	нет единообразного анализа	<i>H. sapiens</i> , <i>M. musculus</i> , <i>R. norvegicus</i> , <i>D. rerio</i> , <i>X. tropicalis</i> , <i>C. elegans</i> , <i>D. melanogaster</i> , <i>S. cerevisiae</i> , <i>A. thaliana</i> , <i>G. gallus</i>
ReMap2022 (Hammal et al., 2022)	всего: 19 983 <i>H. sapiens</i> : 8 103	всего: 2 831 <i>H. sapiens</i> : 1210	MACS2	<i>H. sapiens</i> , <i>M. musculus</i> , <i>D. melanogaster</i> , <i>A. thaliana</i>
Factorbook (Pratt et al., 2022)	всего: 2 863 <i>H. sapiens</i> : 2 688	всего: 930 <i>H. sapiens</i> : 881	None	<i>H. sapiens</i> , <i>M. musculus</i>

**Сравнение БД GTRD с БД ChIP-Atlas по ChIP-seq экспериментам
без учета контрольных экспериментов**

Вид	GTRD v23.07			ChIP-Atlas (состояние в 2021 г.)		
	Транскрипционные факторы и кофакторы		Гистоновые модификации и ChIP-seq	Транскрипционные факторы и кофакторы		Гистоновые модификации ChIP-seq
	ChIP-seq	ChIP-exo		ChIP-seq	ChIP-exo	
<i>Mus musculus</i>	14 695	134	16 284	20 318	53	38 000
<i>Homo sapiens</i>	20 677	2 021	22 977	29 451	48	32 537
<i>Rattus norvegicus</i>	295	2	680	324	1	688
<i>Drosophila melanogaster</i>	2 171	51	2 056	2 470	0	2 375
<i>Caenorhabditis elegans</i>	1 058	0	868	1 286	0	865
<i>Arabidopsis thaliana</i>	699	0	2 162	0	0	0
<i>Saccharomyces cerevisiae</i>	2 167	1 120	1862	3 619	231	2 843
<i>Danio rerio</i>	73	0	597	0	0	0
<i>Schizosaccharomy ces pombe</i>	461	13	919	0	0	0
Всего	42 296	3 341	48 405	57 468	333	77 308
		45 637			57 8011	

Жирным шрифтом отмечены колонки, где больше соответствующих экспериментов.

ПРИЛОЖЕНИЕ 6.1 ЭКСПЕРИМЕНТАЛЬНО ПОДТВЕРЖДЕННЫЕ РСТФ В ПРОМОТОРНЫХ ОБЛАСТЯХ ГЕНОВ NR4A2, NR4A3 И PPARGC1A

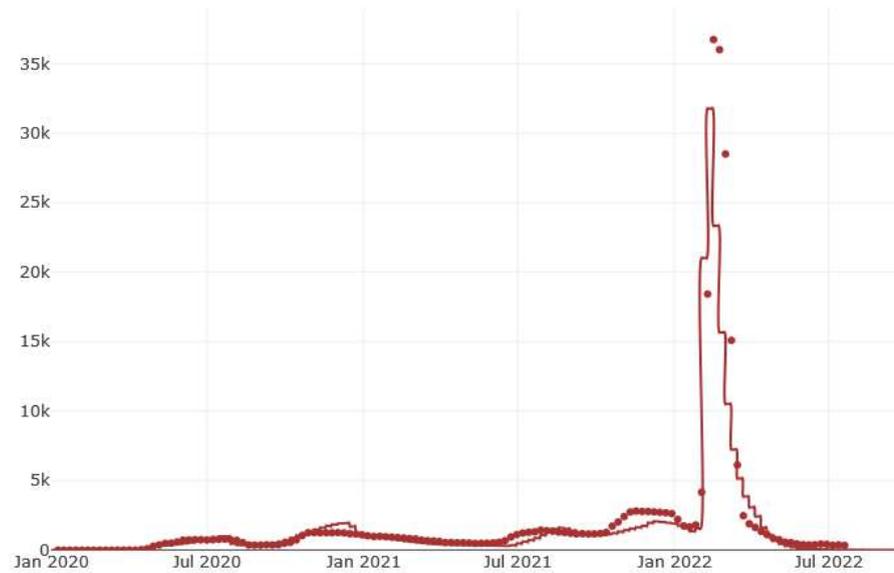
Таблица п.6.1 – Экспериментально подтвержденные (БД GTRD) районы связывания ТФ семейств факторов CREB, FOS и JUN семейств в промоторных областях генов NR4A2, NR4A3 и PPARGC1A (Акбердин и др., 2020)

ТФ	Промотор гена	Старт транскрипции (TSS)	Позиция сайта ТФ относительно TSS
pCRE B1	<i>NR4A2 (ENSG00000153234)</i>	chr2:156342348	-175
	<i>NR4A3 (ENSG00000119508)</i>	chr9:99821855	612
			-279
	<i>PPARGC1A* (XM_005248132)</i>	chr4:23904089	-53
			58
<i>PPARGC1A# (NM_013261)</i>	chr4:23890047	-153	
		89	
JUN	<i>NR4A2</i>	chr2:156342348	-224
			-33
	<i>NR4A3</i>	chr9:99821855	915
			-200
	<i>PPARGC1A*</i>	chr4:23904089	-46
<i>PPARGC1A#</i>	chr4:23890047	-126	
JUND	<i>NR4A2</i>	chr2:156342348	-166
	<i>NR4A3</i>	chr9:99821855	-901
			-45
	<i>PPARGC1A*</i>	chr4:23904089	-3
<i>PPARGC1A#</i>	chr4:23890047	-114	
JUNB	<i>NR4A2</i>	chr2:156342348	-239
			-415
	<i>NR4A3</i>	chr9:99821855	-917
			-90
	<i>PPARGC1A*</i>	chr4:23904089	
<i>PPARGC1A#</i>	chr4:23890047		
FOSL 1	<i>NR4A2</i>	chr2:156342348	
	<i>NR4A3</i>	chr9:99821855	
	<i>PPARGC1A*</i>	chr4:23904089	

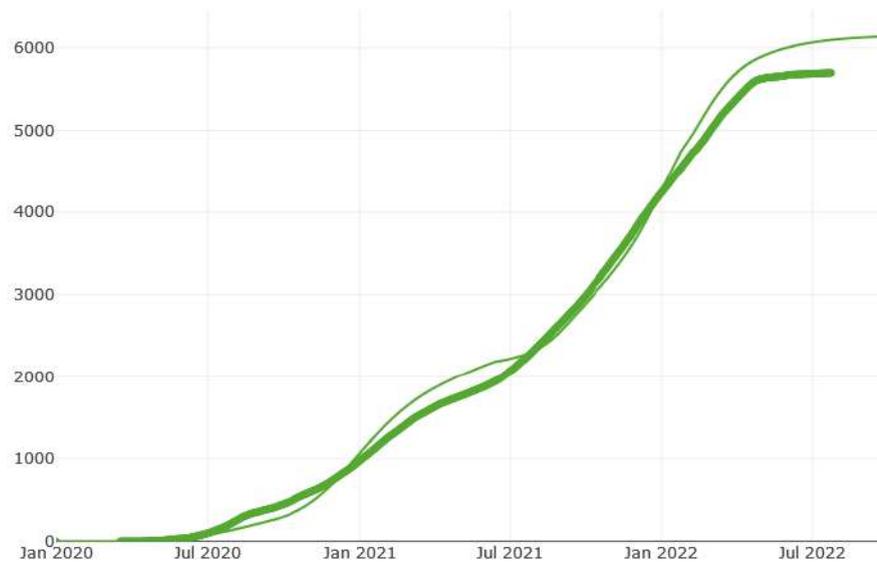
	<i>PPARGC1A</i> [#]	chr4:23890047	
FOSL 2	<i>NR4A2</i>	chr2:156342348	
	<i>NR4A3</i>	chr9:99821855	
	<i>PPARGC1A</i> [*]	chr4:23904089	
	<i>PPARGC1A</i> [#]	chr4:23890047	
FOS	<i>NR4A2</i>	chr2:156342348	-210
	<i>NR4A3</i>	chr9:99821855	-1017
			-901
			-188
			-105
			-210
<i>PPARGC1A</i> [*]	chr4:23904089		
<i>PPARGC1A</i> [#]	chr4:23890047		
FOSB	<i>NR4A2</i>	chr2:156342348	
	<i>NR4A3</i>	chr9:99821855	50
	<i>PPARGC1A</i> [*]	chr4:23904089	
	<i>PPARGC1A</i> [#]	chr4:23890047	
ATF3	<i>NR4A2</i>	chr2:156342348	-198
	<i>NR4A3</i>	chr9:99821855	-828
			-237
			-69
			-205
	<i>PPARGC1A</i> [*]	chr4:23904089	
<i>PPARGC1A</i> [#]	chr2:156342348	-3	
			-151
ATF4	<i>NR4A2</i>	chr2:156342348	
	<i>NR4A3</i>	chr9:99821855	
	<i>PPARGC1A</i> [*]	chr4:23904089	
	<i>PPARGC1A</i> [#]	chr4:23890047	

* – альтернативный промотор *PPARGC1A*; # – канонический промотор *PPARGC1A*.

ПРИЛОЖЕНИЕ 6.2 РЕЗУЛЬТАТЫ ЧИСЛЕННЫХ РАСЧЕТОВ ВТОРОЙ ВЕРСИИ МОДЕЛИ ДЛЯ ЭПИДЕМИОЛОГИИ COVID-19



Количество новых случаев в неделю



Общее количество смертей

Рисунок п.6.2.1 – Результаты численных расчетов второй версии модели эпидемиологии COVID-19 для Новосибирской области. Статистические данные (отмечены толстой линией) были взяты с официального канала “Коронавирус. Оперативный штаб Новосибирской области”.

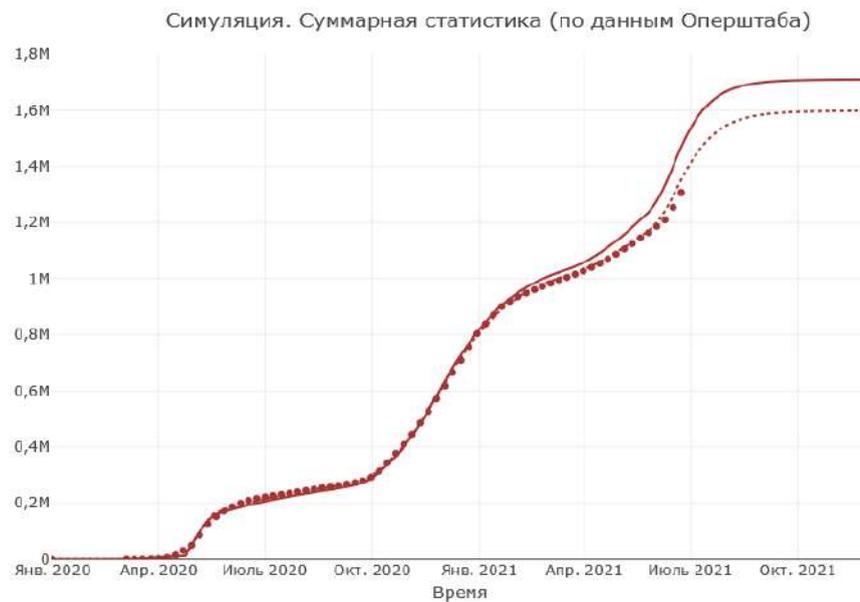
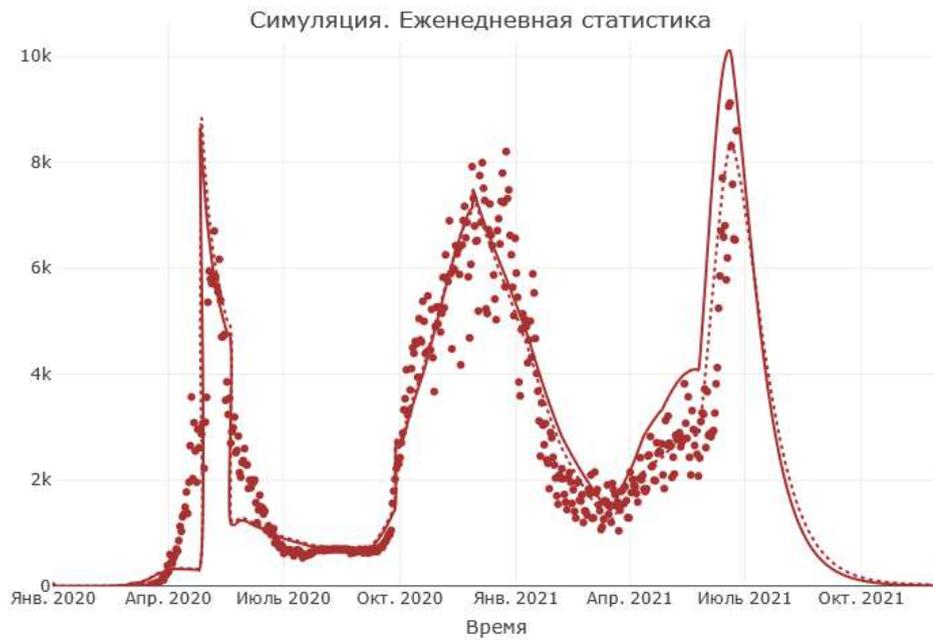
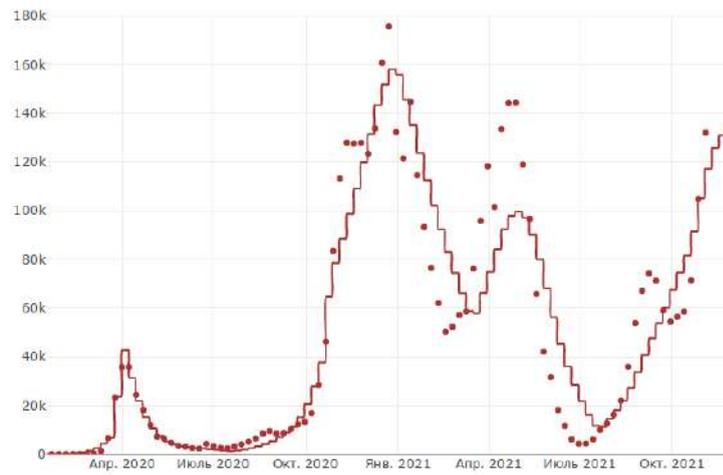
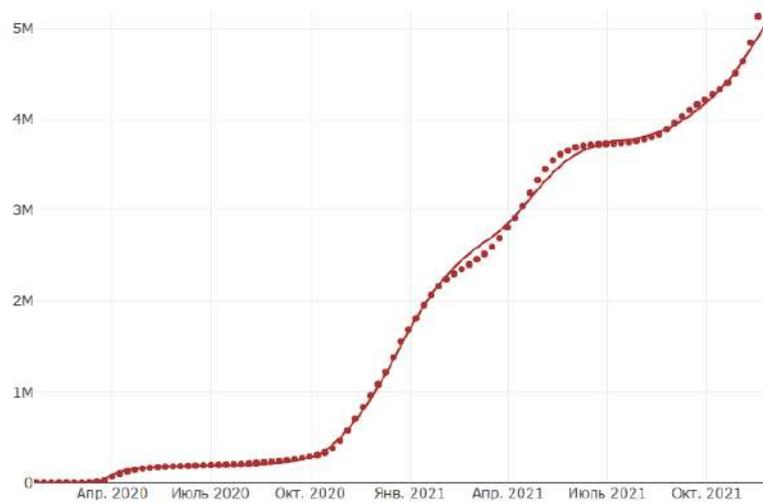


Рисунок п.6.2.2 – Результаты численных расчетов второй версии модели эпидемиологии COVID-19 для г. Москвы области. Статистические данные (отмечены толстой линией) при были взяты с официального канала “Коронавирус. Оперативный штаб г. Москвы”.



Количество новых случаев в неделю

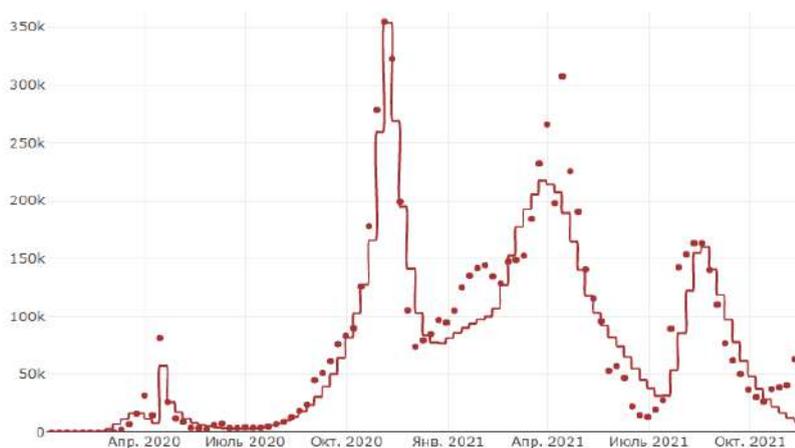


Общее количество новых случаев

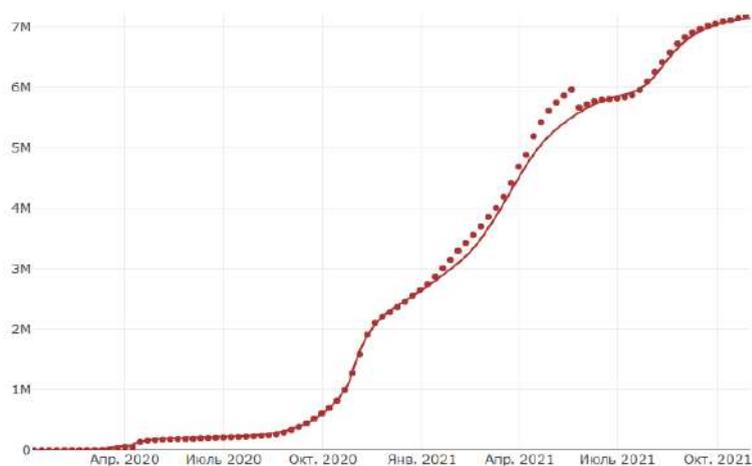


Общее количество смертей

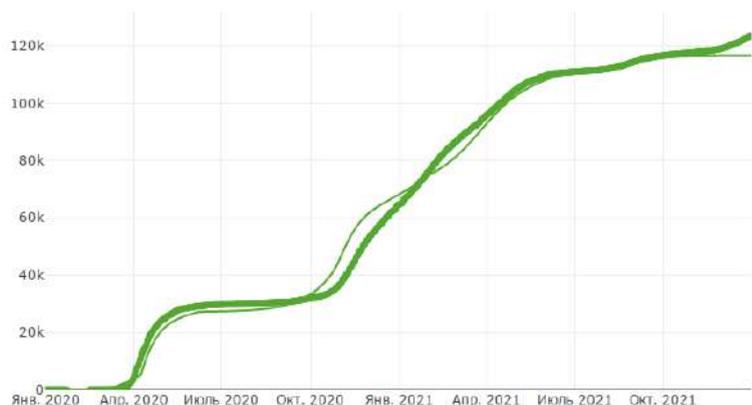
Рисунок п.6.2.3 – Результаты численных расчетов второй версии модели эпидемиологии COVID-19 для Германии. Статистические данные приведены по <https://ourworldindata.org/>.



Количество новых случаев в неделю

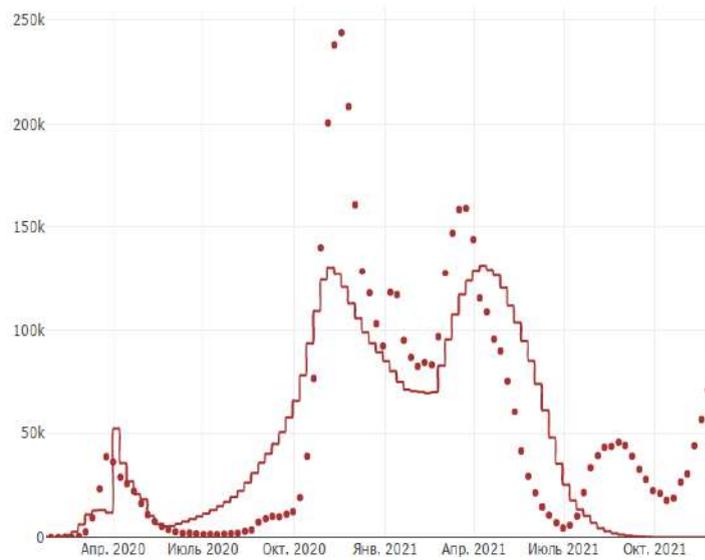


Общее количество новых случаев

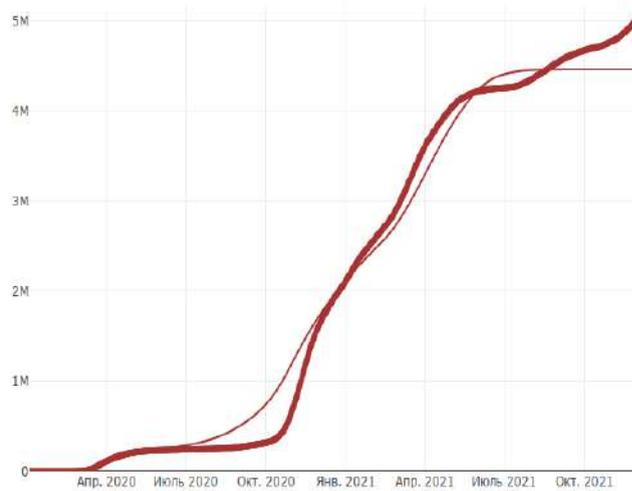


Общее количество смертей

Рисунок п.6.2.4 – Результаты численных расчетов второй версии модели эпидемиологии COVID-19 для Франции. Статистические данные приведены по <https://ourworldindata.org/>. 20 мая 2020 г общее количество выявленных случаев было пересчитано Французским правительством в сторону уменьшения на 364 261 человек, так как некоторые выявленные были учтены дважды. Этим объясняется разрыв в статистических данных на графике «Общее количество новых случаев».



Количество новых случаев в неделю



Общее количество новых случаев



Общее количество смертей

Рисунок п.6.2.5 – Результаты численных расчетов второй версии модели эпидемиологии COVID-19 для Италии. Статистические данные приведены по <https://ourworldindata.org/>.

**ПРИЛОЖЕНИЕ 6.3 РЕЗУЛЬТАТЫ ЧИСЛЕННЫХ РАСЧЕТОВ ТРЕТЬЕЙ ВЕРСИИ МОДЕЛИ ДЛЯ
ЭПИДЕМИОЛОГИИ COVID-19**

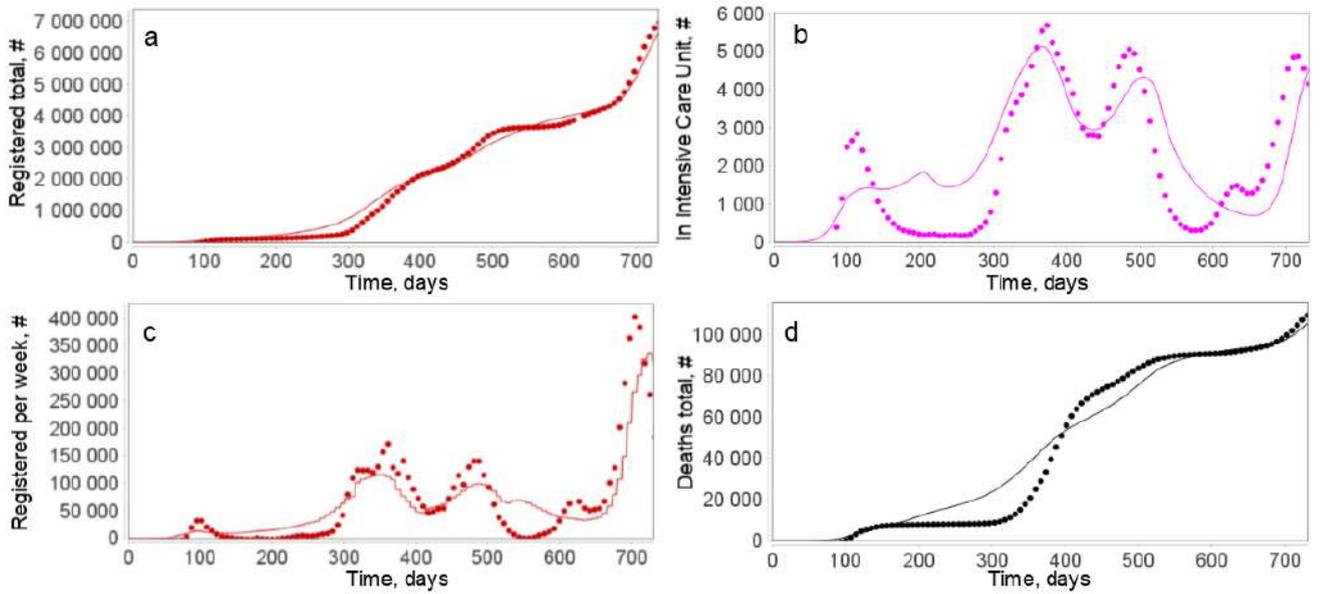


Рисунок п.6.3.1 – Результаты численных расчетов модели с запаздыванием для Германии

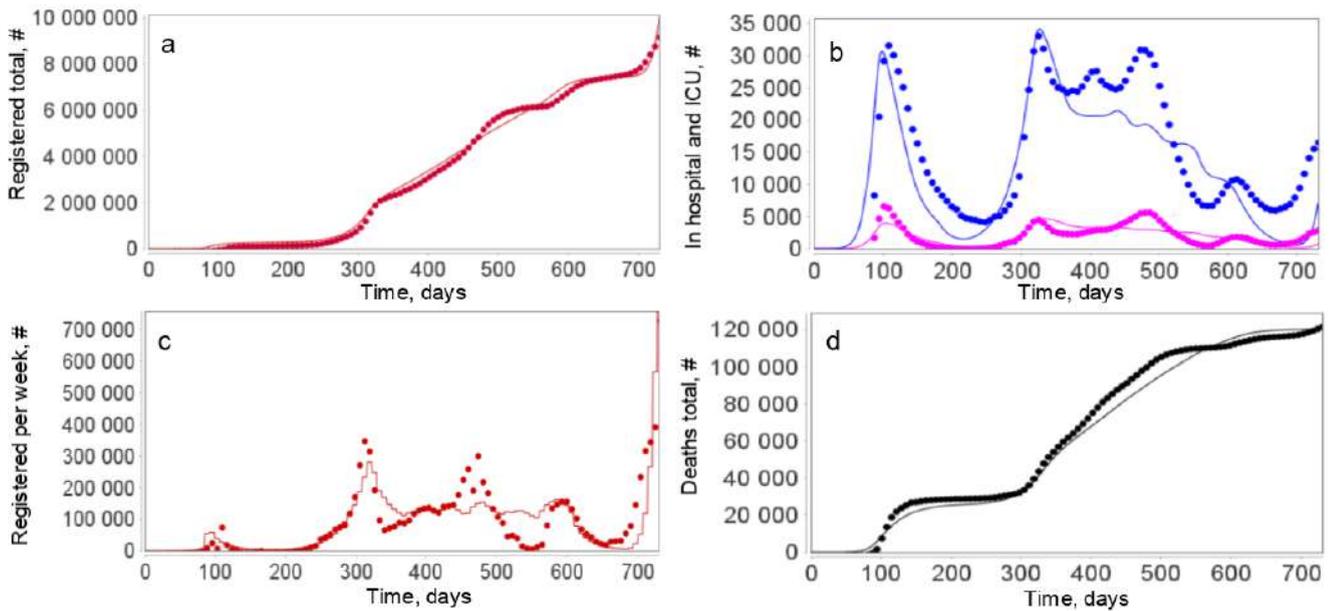


Рисунок п.6.3.2 – Результаты численных расчетов модели с запаздыванием для Франции

ПРИЛОЖЕНИЕ 6.4 РЕЗУЛЬТАТЫ ЧИСЛЕННЫХ РАСЧЕТОВ АГЕНТНОЙ МОДЕЛИ ДЛЯ ЭПИДЕМИОЛОГИИ COVID-19

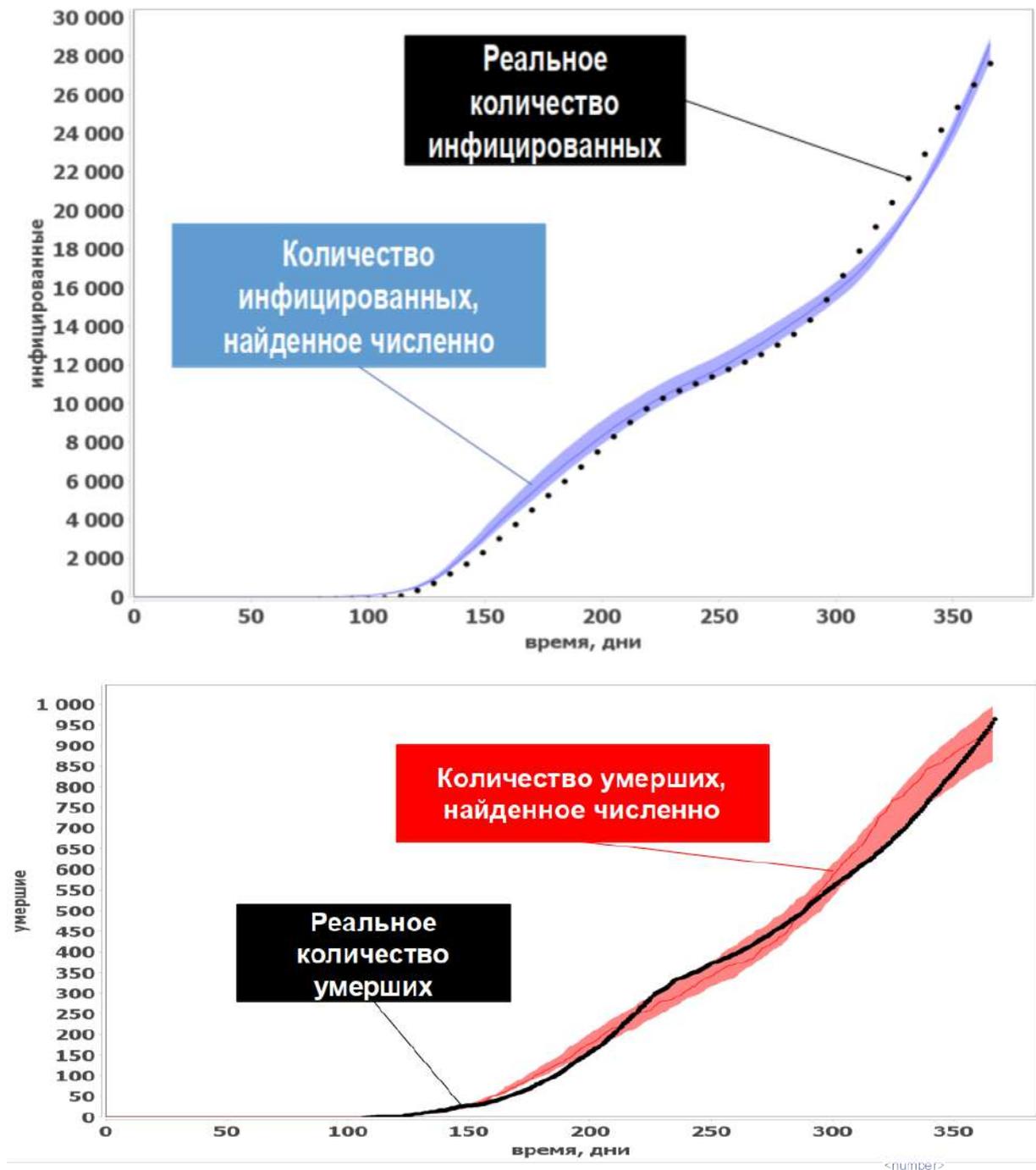


Рисунок п.6.4.1 – Результаты численных расчетов агентной модели для Новосибирска

**Приложение 7.1 Список оптимизируемых параметров для генерации
виртуальных пациентов с АГ**

Таблица П.7.1 – Список оптимизируемых параметров с учетом границ пространства поиска для генерации виртуальных пациентов с артериальной гипертензией, а также коэффициентов чувствительности среднего артериального давления (*MAP*)

№	Параметры	Ограничения ³	Ед. изм.	Коэффициенты чувствительности <i>MAP</i>
01	Общая интенсивность метаболизма, A_1	0.00032 – 0.00128	мл ⁻¹	334312.37
02	Нормальное значение потребления воды, Φ_{win_norm}	0.00096 – 0.00312	л·мин ⁻¹	45476.97
03	Базовая эластичность легочных вен, G_{VRO}	0.01 – 0.05	мм рт. ст.·мл ⁻¹	24578.73
04	Базовая эластичность системных вен, G_{VLO}	0.01 – 0.05	мм рт. ст.·мл ⁻¹	24552.53
05	Нормальный коэффициент фильтрации, K_{FG_0}	0.0039 – 0.0162	л·мин ⁻¹ ·мм рт. ст. ⁻¹	-9518.17
06	Потребление натрия, Φ_{sodin}	0.0280 – 0.2088	мэкв·мин ⁻¹	4634.81
07	Тонус системных артерий, A_9	0.07 – 0.09	мм рт. ст.·с·мл ⁻¹	3736.51
08	Симпатическая чувствительность системных микрососудов, A_3	0.1 – 0.6	мл·мм рт. ст. ⁻¹	2807.93
09	Базовая активность сердечного центра, $Heart_{Base}$	0.01 – 1.00	–	2064.45
10	Жесткость стенок левого желудочка, G_{HL}	0.02 – 0.72	мм рт. ст.·мл ⁻¹	1126.78
11	Нормальная фракционная реабсорбция натрия в дистальных канальцах, n_{e_dt}	0.3 – 0.7	–	552.41
12	Равновесное отношение активности ренина в плазме к его концентрации, X_{PRC_PRA}	0.61 – 1.42	фмоль·мин ⁻¹ ·пг ⁻¹	279.42
13	Базовая эластичность легочных артерий, G_{ARO}	0.08 – 0.26	мм рт. ст.·мл ⁻¹	269.34
14	Жесткость стенок правого желудочка, G_{HR}	0.02 – 0.72	мм рт. ст.·мл ⁻¹	-207.91
15	Нормальная фракционная проксимальная реабсорбция натрия, n_{η_pl}	0.67 – 0.97	–	158.12
16	Нормальная фракционная реабсорбция натрия в собирательных трубочках, n_{η_cd}	0.6 – 1.0	–	137.35
17	Базовый инотропный коэффициент левого желудочка, K_{LO}	0.5 – 0.8	–	107.16
18	Базовая проводимость системных микрососудов, Y_{ALVLO}	0.5 – 2.0	мл·с ⁻¹ ·мм рт.ст. ⁻¹	73.63

³ М – мужчина; Ж – женщина.

19	Базовая эластичность системных артерий, G_{ALO}	0.33 – 1.67	мм рт. ст.·мл ⁻¹	-72.27
20	Концентрация калия в крови, C_K	3.5 – 5.5	мЭКВ·л ⁻¹	60.52
21	Концентрация мочевины в плазме крови, <i>urea</i>	1.8 – 7.1	ммоль·л ⁻¹	52.49
22	Базовый инотропный коэффициент правого желудочка, K_{R0}	0.5 – 0.8	–	42.85
23	Номинальное значение скорости потока натрия через область <i>macula densa</i> , $\Phi_{md_sod_0}$	1.0 – 4.0	мЭКВ·мин ⁻¹	27.84
24	Номинальная потребность организма в кислороде, RO_{20}	2.52 – 5.88	мл·с ⁻¹	26.07
25	Сатурация артериальной крови, SpO_2	0.92 – 0.99	–	25.16
26	Концентрация глюкозы в плазме, <i>glucose</i>	3.9 – 6.1	ммоль·л ⁻¹	18.48
27	Почечное венозное сопротивление, R_V	11.3 – 20.1	мм рт. ст.·мин·л ⁻¹	18.43
28	Кислородная емкость гемоглобина, C_H	1.32 – 1.39	мл·г ⁻¹	17.63
29	Почечное венозное давление, P_V	2.0 – 6.0	мм рт. ст.	14.04
30	Диаметр афферентной артериолы, d_{aa}	8.7 – 23.9	мкм	11.35
31	Нормальный уровень вазопрессина, C_{adh_norm}	1.0 – 13.3	пг·мл ⁻¹	-10.64
32	Гидростатическое давление в пространстве Боумена-Шумлянского, P_B	10.0 – 15.0	мм рт. ст.	10.02
33	Номинальное сопротивление междольковых, дуговых и междольковых артерий, $R_{preglom_0}$	7.0 – 28.0	мм рт. ст.·мин·л ⁻¹	9.75
34	Диаметр эфферентной артериолы, d_{ea}	12.2 – 20.1	мкм	5.89
35	Порог действия закона Франка-Старлинга, $FS_{threshold0}$	0.0 – 40.0	мл	4.82
36	Нормальный уровень альдостерона, C_{al_norm}	70 – 300	пг·мл ⁻¹	3.22
37	Базовая проводимость легочных микрососудов, Y_{ARVRO}	9.0 – 21.0	мл·с ⁻¹ ·мм рт.ст. ⁻¹	-2.63
38	Нормальный уровень атриопептина, C_{anp_norm}	7.4 – 152.0	нг·л ⁻¹	2.40
39	Нормальная концентрация ренина в плазме, PRC_{nom}	3.42 – 69.4	пг·мл ⁻¹	-2.24
40	Гематокрит, Hct	М: 40 – 54 Ж: 36 – 48	%	2.16
41	Нормальное значение клубочкового гидростатического давления, P_{gh_norm}	48.0 – 63.0	мм рт. ст.	1.64
42	Общий белок, TP	60.0 – 86.0	г·л ⁻¹	1.45
43	Длина афферентной артериолы, L_{aa}	101 – 123	мкм	0.89
44	Длина эфферентной артериолы, L_{ea}	124 – 152	мкм	-0.40
45	Гемоглобин, He	М: 140 – 180 Ж: 120 – 160	г·л ⁻¹	0.17
46	Количество нефронов в почках, $N_{nephrons}$	0.80E6 – 2.75E6	–	9.48E-5

47	Доля ненапряженного объема в системных артериях, k_{AL}	0.7 – 1.0	–	
48	Доля ненапряженного объема в легочных артериях, k_{AR}	0.7 – 1.0	–	
49	Доля ненапряженного объема в системных венах, k_{VL}	0.7 – 1.0	–	
50	Доля ненапряженного объема в легочных венах, k_{VR}	0.7 – 1.0	–	
51	Доля ненапряженного объема в левом желудочке, k_{HL}	0.0 – 0.3	–	
52	Доля ненапряженного объема в правом желудочке, k_{HR}	0.0 – 0.3	–	
53	Начальное значение общего обменного натрия, M_{sod}	2040 – 3950	мЭКв	
54	Начальное значение клубочкового капиллярного онкотического давления, P_{go}	23.6 – 34.0	мм рт. ст.	
55	Начальное значение общего количества воды в организме, TBW	$f(0.9 \cdot N_{BV}) - f(1.1 \cdot N_{BV})^4$	л	
56	Начальное значение венозного содержания кислорода, VO_2	0.0855 – 0.1848	–	

⁴ Для расчета TBW мы использовали формулу $f(N_{BV}) = (N_{BV} - 650)/111.5$ (Moore, 1967), где N_{BV} – оценка, определяемая уравнением Надлера (Nadler et al., 1962). В качестве нормального диапазона общего объема крови мы рассматривали $N_{BV} \pm 10\%$.

ПРИЛОЖЕНИЕ 8.1 ПРИМЕР СЦЕНАРИЯ ДЛЯ БИОИНФОРМАТИЧЕСКОЙ ПЛАТФОРМЫ ДЛЯ АНАЛИЗА ДАННЫХ ОДНОМОЛЕКУЛЯРНОГО СЕКВЕНИРОВАНИЯ

Все сценарии для биоинформатической платформы для анализа данных одномолекулярного секвенирования реализованы в виде отдельных git репозиториях. Каждый репозиторий имеет следующую структуру:

- папка Doc – содержит описание сценария в виде файла README.md;
- папка Input – входные тестовые данные для демонстрации работы сценария (если позволяет размер соответствующих файлов);
- папка Results – примеры результатов анализа, в том числе, отчётов по результатам выполнения сценария в формате HTML (если позволяет размер соответствующих файлов);
- папка wdl – файлы, описывающие запуск отдельных программ, используемых в сценарии, в формате wdl. Данные программы могут быть запущены как самостоятельные wdl-скрипты.
- сценарий анализа данных одномолекулярного секвенирования в формате WDL;
- inputs.json – текстовый файл в формате JSON, содержащий информацию о входных параметрах. Представляет собой альтернативный способ задания входных параметров выполнения wdl-скрипта, доступный для выбора в интерфейсе запуска сценария.

Сценарий “Сборка геномов *de-novo*”

Сценарий "Сборка геномов *de novo*" предназначен для *de novo* сборки и аннотации геномов на основе данных мономолекулярного секвенирования (рисунок п.8.1.1). Данный сценарий основывается на сценарии для сборки геномов *de novo*, реализованном в программном пакете SMRT Link для анализа данных одномолекулярного секвенирования PacBio. Из сценария были исключены этапы: оценка структурных вариаций, а также определение нуклеотидов, подвергнувшихся различным модификациям, поскольку данные этапы представляют собой отдельную задачу и представлены в виде отдельных сценариев.

Рассматриваемый сценарий состоит из 10 основных шагов, краткое описание используемых программ приведено в таблице п.8.1.1. На рисунке п.8.1.2 приведена форма для запуска сценария, а в таблице п.8.1.2 – описание входных и выходных параметров этой формы.

Ссылка на git-репозиторий: <https://github.com/Biosoft-ru/De-Novo-Genome-Assembly>

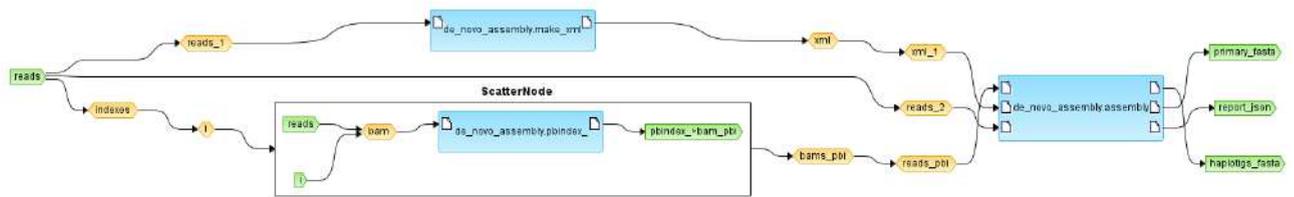


Рисунок п.8.1.1 – Схема сценария “Сборка геномов *de novo*” на платформе BioUML

Таблица п.8.1.1 – Список шагов сценария “Сборка генома *de novo*”

№	Описание шага	Используемые программы
1	Фильтрация исходных данных по качеству, отсечение слишком коротких прочтений	python3 pbcorettools.tasks.dataset_filter
2	Конвертирование отфильтрованных прочтений в базу данных	pancake seqdb
3	Нахождение пересечений в прочтениях	falconc ipa-shard-ovl-async, pancake ovl-hifi
4	Сортировка и фильтрация локальных выравниваний	ipa2 ovl-async-merge, sort
5	Фазирование прочтений	nighthawk split, nighthawk phase
6	Фильтрация пересечений	falconc m4filt
7	Сборка генома <i>de novo</i>	pblayout
8	Полировка собранного генома	falconc ipa-polish-paf-split, pancake seqfetch, racon
9	Удаление гаплотигов и перекрытий в сборке генома на основе глубины прочтений	pancake, pbmm2
10	Генерация отчета по сборке генома	ipa2 report_assembly2

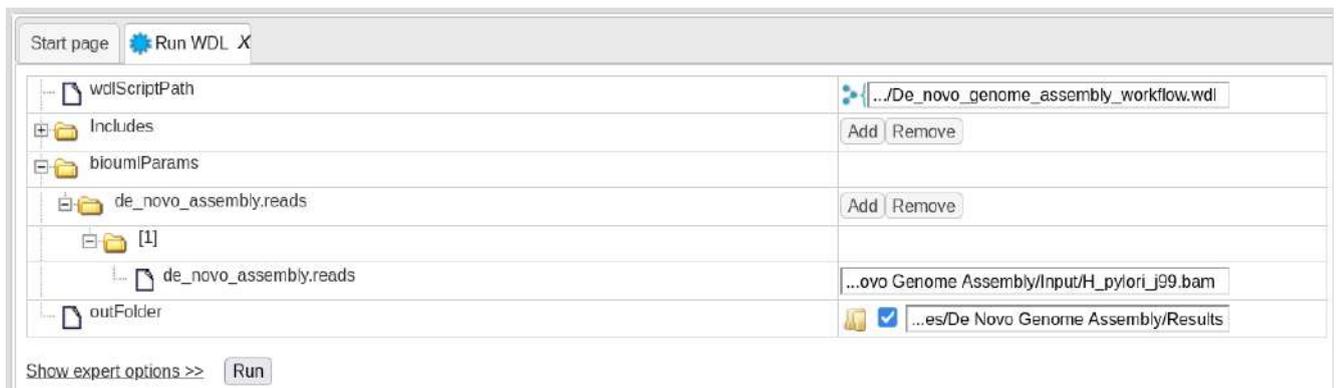


Рисунок п.8.1.2 – Интерфейс запуска сценария “Сборка геномов *de-novo*”

Таблица п.8.1.2 – Параметры сценария “Сборка геномов *de-novo*”

Параметр	Формат	Краткое описание
<i>Входные данные и параметры</i>		
de_novo_assembly.reads	BAM/FASTQ	Данные одномолекулярного секвенирования: консенсусные прочтения
outFolder	Папка	Путь до директории с результатами
<i>Выходные данные</i>		
	FASTA	Окончательно отполированная сборка гаплотипов
	FASTA	Окончательно отполированная сборка первичных контигов
	HTML	Отчёт о результатах сборки генома

Приложение 8.2 Подсистемы для прототипа «Национальной базы генетической информации»

Ниже приведены основные фрагменты технического задания на выполнение научно-исследовательской работы по теме: «Исследование, обоснование и выбор программных решений для визуализации генетических данных и обеспечения инструментов для работы с генетической информацией в «Национальной базе генетической информации».

Полный текст конкурсной документации доступен на официальном сайте Единой информационной системы в сфере закупок:

<https://zakupki.gov.ru/epz/order/notice/ea44/view/documents.html?regNumber=0373100001021000793>

Далее приведены фрагменты отчетной документации, демонстрирующие интерфейс подсистемы для прототипа «Национальной базы генетической информации».

2. Цели и задачи выполнения работ

2.1. Цель работ

Целью работ является исследование функциональных возможностей существующих программных решений в части визуализации генетических данных и инструментов для работы с генетической информацией, а также формирование детальных требований к подсистемам «Геномный браузер» и «Рабочие окружения» для «Национальной базы генетической информации» (далее – ИАС НБГИ).

2.2. Задачи работы

Задачами работы являются:

- формирование функциональных возможностей подсистем «Геномный браузер» и «Рабочие окружения»;
- разработка эскизов экранных форм;
- создание программно-технических действующих прототипов на основе технологий, предлагаемых для использования, в целях подтверждения принципиальной работоспособности, выбранных решений;
- определение принципов и механизмов интеграции подсистем с программными сервисами Заказчика.

3. Требования к составу и содержанию работ

3.1. В рамках НИР Исполнитель должен провести исследование, обоснование и выбор программных решений в части визуализации генетических данных и инструментов для работы с генетической информацией, а также сформировать детальные требования к подсистемам

«Геномный браузер» и «Рабочие окружения» для «Национальной базы генетической информации».

3.2. Выбор проектных решений должен быть выполнен с учетом актуальности предлагаемых и используемых технологий (с предоставлением документов, подтверждающих правообладание). Предпочтение при выборе технологий, протоколов и готовых программно-технических решений должно быть отдано программному обеспечению с открытым исходным кодом.

...

3.15. Подсистема «Геномный браузер» должна обеспечивать:

- а) представление генетической информации и связанных треков метаинформации (структурная аннотация, функциональная аннотация и др.) в графическом интерактивном режиме;
- б) представление данных о картировании прочтений на геномную последовательность;
- в) визуализацию генетической информации и соответствующих метаданных для геномов размером вплоть до 50 млрд пар нуклеотидов, включающих непрерывные последовательности нуклеотидов длиной до 1 млрд пар нуклеотидов;
- г) масштабируемость представления, регулируемую пользователем;
- д) переход к произвольному, заданному пользователем региону последовательности, включенной в визуализируемый геном;
- е) изменение типа отображаемой информации в зависимости от масштаба: отображение прочтений (ридов); треков; отдельных нуклеотидов;
- ж) применение цветового кодирования при отображении информации;
- з) отображение контактирующих участков хромосом.
- и) отображение графика относительного состава нуклеотидов;
- к) интерактивное объединение и пересечение треков с возможностью комбинации их отображения (объединение, только общие области, различия);
- л) отображение совпадающих или отличающихся фрагментов при сравнении ГД;
- м) поддержку следующих форматов данных:
 - BED (Browser Extensible Data);
 - WIG (Wiggle Track Format);
 - bigWig (bigWig Track Format);
 - bigBed (bigBed Track Format);
 - VCF (Variant Call Format);
 - GFF (General Feature Format);
- н) возможность выбора нескольких источников данных для отображения;
- о) отображение метаданных о треках и самих треков;
- п) возможность поиска треков по подключенным источникам данных;

- р) табличное представление с отображением координат и аннотации для отображения полного перечня треков загруженных ГД;
- с) вызов доступных в системе обработчиков для выбранного в области просмотра участка;
- т) визуализацию трёхмерной структуры молекул и данных аннотации;
- у) поддержку возможности адресации к заданному участку ГД через гиперссылку с указанием таких параметров как: источник геномных данных, геномная сборка; район – хромосома, позиции; набор отображаемых ГД (треков);
- ф) интеграцию с Jupyter Notebook.

3.16. Подсистема «Рабочие окружения» должна обеспечивать:

а) решение следующих задач:

- сборка и аннотирования первичных последовательностей;
- поиск и выравнивание последовательностей;
- анализ ДНК и РНК;
- анализ транскриптомных данных;
- анализ экспрессии генов;
- функциональный анализ;
- филогенетический анализ;
- поиск и анализ вариаций геномов.

б) определение области применения рабочего окружения, включая указание:

- типов ГД, с которыми может работать пользователь в рамках окружения;
- конфигурационных параметров инструментов анализа;
- поддерживаемых категорий био-образцов;
- используемых референсных БД;

в) предоставление промежуточных и итоговых файлов с результатами анализа.

...

4.9. Отчет о НИР должен включать следующие основные темы:

- сравнительный анализ существующих программных решений для подсистемы «Геномный браузер»;
- архитектурные решения для реализации подсистемы «Геномный браузер», детальное описание функциональных возможностей и интерфейса пользователя; требования к системному окружению;
- архитектурные решения для реализации подсистемы «Рабочие окружения», детальное описание функциональных возможностей и интерфейса пользователя; требования к системному окружению.
- результаты тестирования рассмотренных программных решений на технологическом стенде.

Search Info SQL track

Site ID: 551
 Type: mRNA
 Sequence name: NC_056725.1
 Position: 23174059 - 24429368 (1255310)
 Properties:

- Dbxref: GeneID:122808636,Genbank:XM_044079637.1
- ID: rna-XM_044079637.1
- Name: XM_044079637.1
- Parent: gene-DPF1
- experiment: COORDINATES: polyA evidence [ECO:0006239]

Track: GCF_019279795.1_PAN1.0_genomic

First Previous Page 1 of 18 Next Last Showing 1 to 50 of 888 entries Show 50 entries

ID	Sequence (chromosome name)	From	To	Length	Strand	Type	Property: Dbxref	Property: Gap
1	NC_056725.1	1	2000000000	2000000000	+	region	taxon:7888	
62	NC_056725.1	3247212	4991190	1743979	-	gene	GeneID:122807035	

Рисунок п.8.2.1 – Проверка производительности подсистемы "Геномный браузер" – визуализация самого большого секвенированного генома из БД NCBI – *Protopterus annectens* и его аннотации (40+ млрд. пар нуклеотидов).

3ins.pdb - BioUML 2020.4 Геном

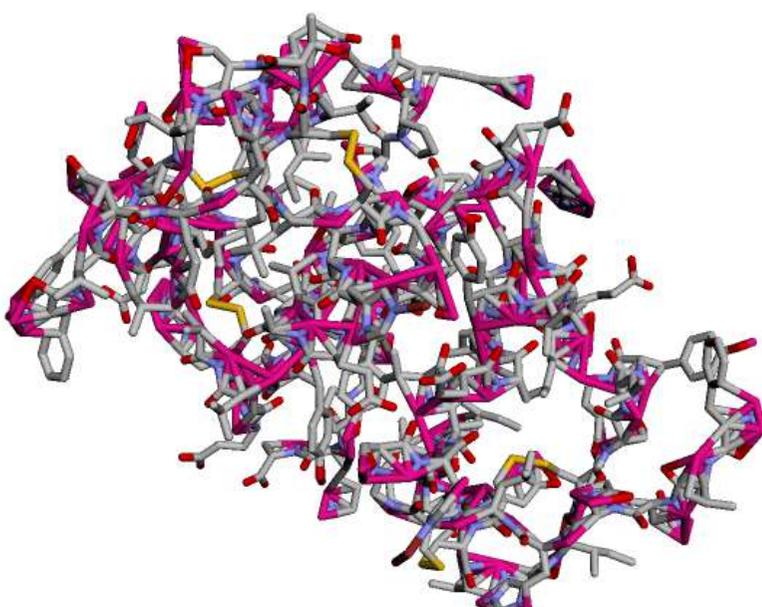
Не защищено | nbgi.computing.kiae.ru/bioumlweb/#de=data/Collaboration/demo_nbgi/Data/PDB/3ins.pdb

Research: Demo | Default

Start page | 3ins 3D X

STRUCTURE OF INSULIN. RESULTS OF JOINT NEUTRON AND X-RAY REFINEMENT

Stick



Search | Info | Default | My description | Graph search | Script | Clipboard | Tasks

ID: [3ins.pdb](#)
Created: Thu Dec 09 05:28:52 UTC 2021
Modified: Thu Dec 09 05:28:52 UTC 2021
Size on disk: 197.5kb (202,257 bytes)
Complete name: data/Collaboration/demo_nbgi/Data/PDB/3ins.pdb

User description is not available

Рисунок п.8.2.2 – Визуализация трехмерной структуры структуры молекулы инсулина в подсистеме “Геномный браузер”

ПРИЛОЖЕНИЕ 8.3 ПЛАТФОРМА geneXplain

Интерфейс платформы geneXplain

Интерфейс пользователя платформы geneXplain построен путем настройки ПК BioUML. Для этого используется свой стиль – цвета компонентов интерфейса, иконки и стартовая страница. На рисунке п.8.3.1 (правая часть) представлена стартовая страница, в ней методы и сценарии сгруппированы по типу омиксных данных и решаемых задач.

На этой стартовой странице пользователь выбирает соответствующий раздел и попадает на страницу для выбора конкретного метода анализа или сценария. Например, если на рисунке п.8.3.1 пользователь выберет раздел RNA-seq, то далее он попадает на страницу, представленную на рисунке п.8.3.2 для выбора метода или сценария анализа RNA-seq данных. После выбора соответствующего метода или сценария откроется форма для ввода входных и выходных данных а также параметров анализа, как это было описано в главе 4.

Собственные методы анализа омиксных данных

Помимо методов анализа омиксных данных, входящих в ПК BioUML, платформа geneXplain содержит ряд дополнительных методов глубокого анализа омиксных данных с использованием БД TRANSFAC® и TRANSPATH®. Эти методы реализованы как модули для ПК BioUML. Ниже кратко описаны наиболее важные из них.

Поиск композиционных модулей (CMA – Composite Module Analyst) – данный метод позволяет найти комбинации ССТФ – композиционные модули, которые максимально разделяет два набора генов (предполагается, что эти наборы генов регулируются сходным образом на уровне транскрипции). Для поиска потенциальных ССТФ в промоторах этих генов используется библиотека весовых матриц из БД TRANSFAC®. Композиционный модуль представляется как функция от набора сайтов связывания отобранных ТФ, локализованных в определенной ориентации и на определенном расстоянии друг от друга. Для построения таких композиционных модулей и оптимизации параметров разделяющей функции используется генетический алгоритм (Kel et al., 2006; Waleev et al., 2006).

Поиск "мастер-регуляторов" – анализ графов регуляторных путей для поиска ключевых молекул, связанных с изменениями экспрессии генов в заданных условиях (заболевание, воздействие и т.п.) (Koschmann et al., 2015; Kel et al., 2016; Boyarskikh et al., 2018). Для этого проводится анализ промоторов генов, изменивших экспрессию: с использованием БД TRANSFAC® выявляются ТФ, которыми они обогащены. Для найденных ТФ по БД TRANSPATH® строится граф ("восходящий анализ", upstream analyses), который связывает эти ТФ с воздействующими на них компонентами путей передачи сигнала в клетке, что позволяет

найти общие регуляторы ("мастер-регуляторы", рисунок п.8.3.3). Такие мастер-регуляторы служат перспективными кандидатами терапевтических мишеней, поскольку их ингибирование приводит к блокировке сигналов, приводящих к активации дифференциально экспрессирующихся при патологии генов (Стельмашенко и др., 2019).

"Шагающие пути" (walking pathways) – метод поиска возможных причинно-следственных связей между эпигенетическими изменениями (метилированием ДНК) в регуляторных областях генов, которые влияют на ССТФ, и изменениями экспрессии генов. Этот метод рассматривает топологию задействованных путей передачи сигнала в клетке и ищет петли положительной обратной связи, которые могут вызывать канцерогенные aberrации в экспрессии генов. Динамических изменения статуса метилирования ДНК важных регуляторных областей генов рассматриваются как отдельные "шаги" («эпигеномная ходьба»), приводящие к изменению структуры регуляторных путей в клетке при раке, что и дало такое название методу (Kel et al., 2019).

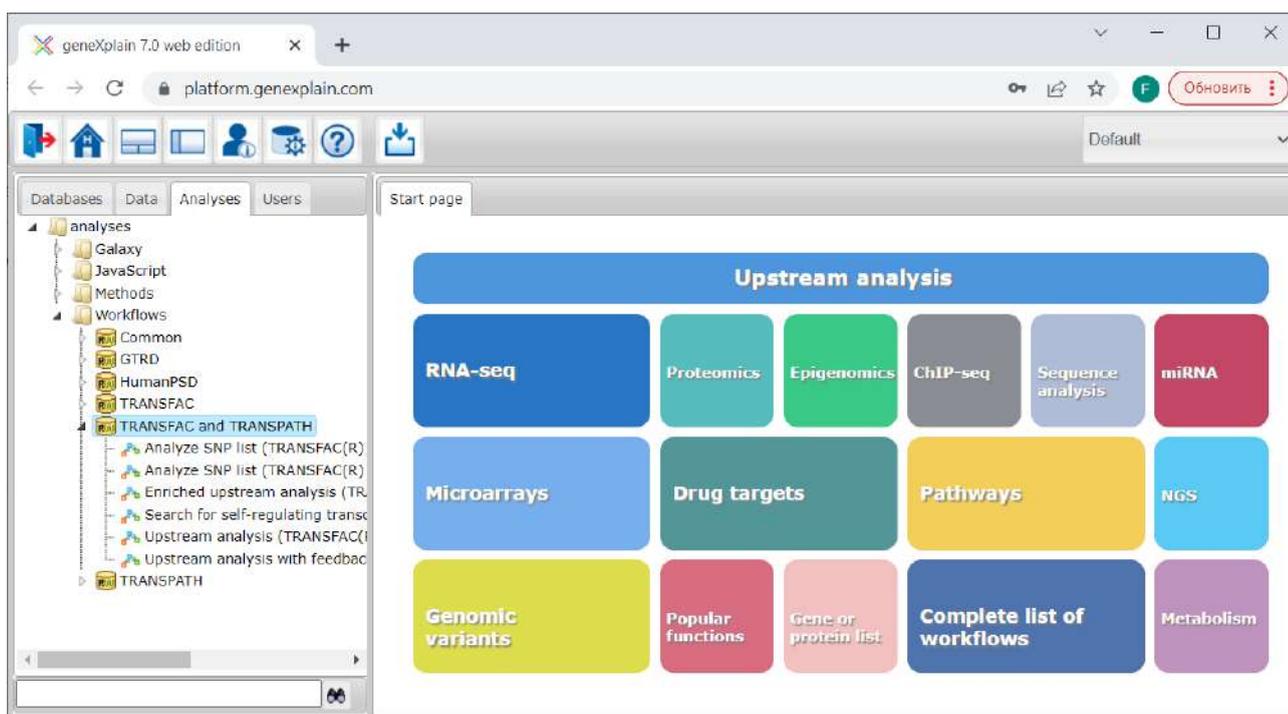


Рисунок 8.3.1 – Стартовая страница платформы geneXplain. Левая часть – вкладка с деревом методов анализов и сценариями. Правая часть – группы сценариев и анализа

The screenshot displays the geneXplain 7.0 web edition interface. The browser address bar shows the URL `platform.genexplain.com`. The left sidebar contains a navigation menu with categories: Databases, Data, Analyses, and Users. Under 'Analyses', a tree view shows folders for Galaxy, JavaScript, Methods, and Workflows. The 'Workflows' folder is expanded, listing sub-folders: Common, GTRD, HumanPSD, TRANSFAC, TRANSFAC and TRANSPATH, and TRANSPATH.

The main content area is titled 'RNA-seq' and features a blue arrow pointing left. It lists several analysis methods and scenarios, each with a folder icon and a corresponding icon on the right:

- Load RNA-seq data** (Icon: Download)
- RNA-seq preprocessing** (Icon: Document with list)
 - [Conversion of RNAseq data](#)
 - [Alignment of RNAseq data](#)
 - [Counting reads to genomic features](#)
 - [From raw data to gene/transcript counts](#)
 - [Discover genome variants and indels](#)
- Detect differentially expressed genes (DEGs)** (Icon: Line graph)
 - [Full RNAseq data analysis](#)
 - [Statistics for RNAseq data](#)
- Discover functional enrichment of DEGs** (Icon: Line graph)
 - [Gene set enrichment analyses \(GSEA\)](#)
 - [Functional classification](#)
- Analyze networks of DEGs** (Icon: Network diagram)
 - [Find master regulators](#)
 - [Find common effectors](#)
 - [Identify functional gene cluster](#)
- Analyze regulatory regions of DEGs** (Icon: Document with list)
- Find drug targets of DEGs** (Icon: Molecular structure)

Рисунок п.8.3.2 – Страница для выбора метода или сценария анализа RNA-seq данных

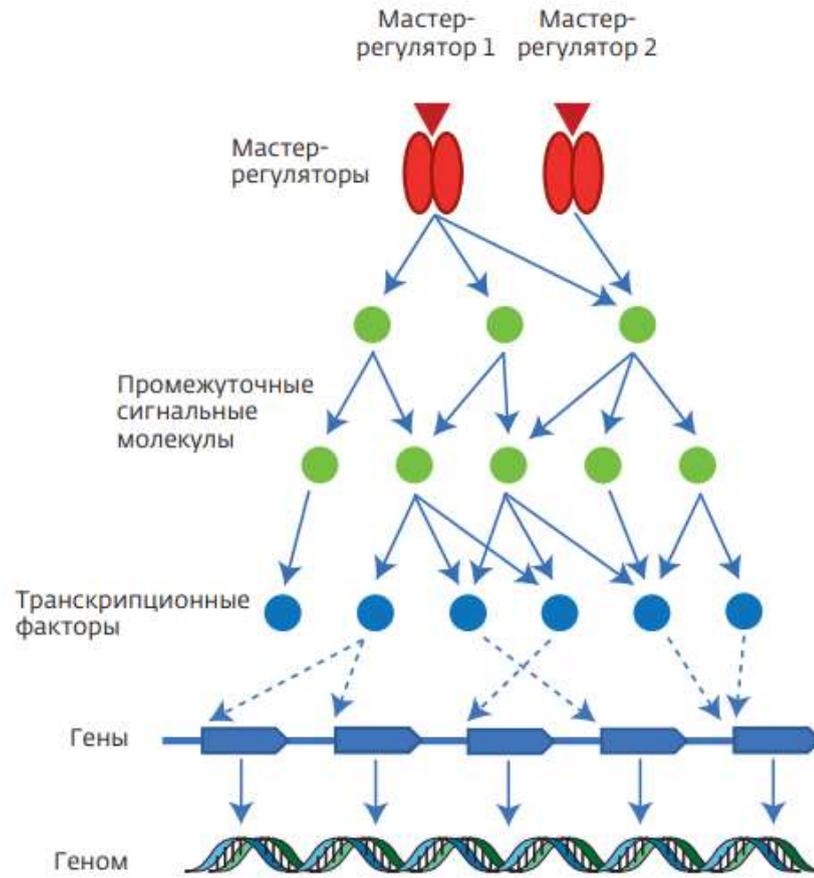


Рисунок п.8.3.3 – Схема поиска мастер-регуляторов (Стельмашенко и др., 2019)

ПРИЛОЖЕНИЕ 8.4 ПЛАТФОРМА GENOME ENHANCER

Веб интерфейс платформы Genome Enhancer представлен на рисунке п.8.4.1 Основным компонентом является мастер (wizard) по проведению анализа, состоящий из 5 шагов:

- **Select project** – на этом шаге пользователь создает новый проект или выбирает уже существующий для продолжения анализа или просмотра результатов;
- **Upload my data** – здесь пользователь загружает свои данные для анализа. Список типов данных и типов файлов приведен на рисунке п.8.4.2.
- **Prepare my data** – на этом шаге пользователь из загруженных файлов, отображаемых в репозитории проекта (рисунок п.8.4.3, левая часть) выбирает файлы для анализа, путем перетаскивания мышкой файлов из проекта на диаграмму для аннотации данных для анализа (рисунок п.8.4.3, правая часть).
- **Describe my data** – на данном шаге происходит графическая аннотация данных путем их разбиения на группы, например контроль и эксперимент (рисунок п.8.4.4). Идея графической аннотации была предложена Колпаковым Ф.А. (Свидетельство о государственной регистрации программы для ЭВМ № 2017615715 от 15.06.2017). Для этого для каждой группы данных создается свой компартимент, отображаемый на диаграмме в виде прямоугольника, куда пользователь при помощи drag & drop может поместить названия соответствующих файлов с данными.
- **Start analysis** – на этом шаге пользователь задает дополнительную информацию для анализа данных (рисунок п.8.4.5), в частности:
 - какой болезни соответствуют загруженные данные;
 - какая ткань или орган была использована для получения омиксных данных;
 - определенные на шаге 3 группы данных для сравнения друг с другом.
- **See report** – результаты анализа представляются в виде 2 отчетов:
 - краткий отчет – основные результаты анализа: список перспективных лекарственных мишеней и препаратов для лечения исследуемого заболевания, выявленных на основе анализа дифференциально экспрессируемых генов;
 - полный отчет – подробные результаты анализа данных в виде отчета, сходного по содержанию и оформлению с готовой научной статьей.

Пример первой страницы автоматически генерируемого отчета в форме статьи приведен на рисунке п.8.4.6. На сайте платформы Genome Enhancer выложены примеры отчетов с результатами анализа омиксных данных при разных патологиях:

<http://my-genome-enhancer.com/demoReports.php>.

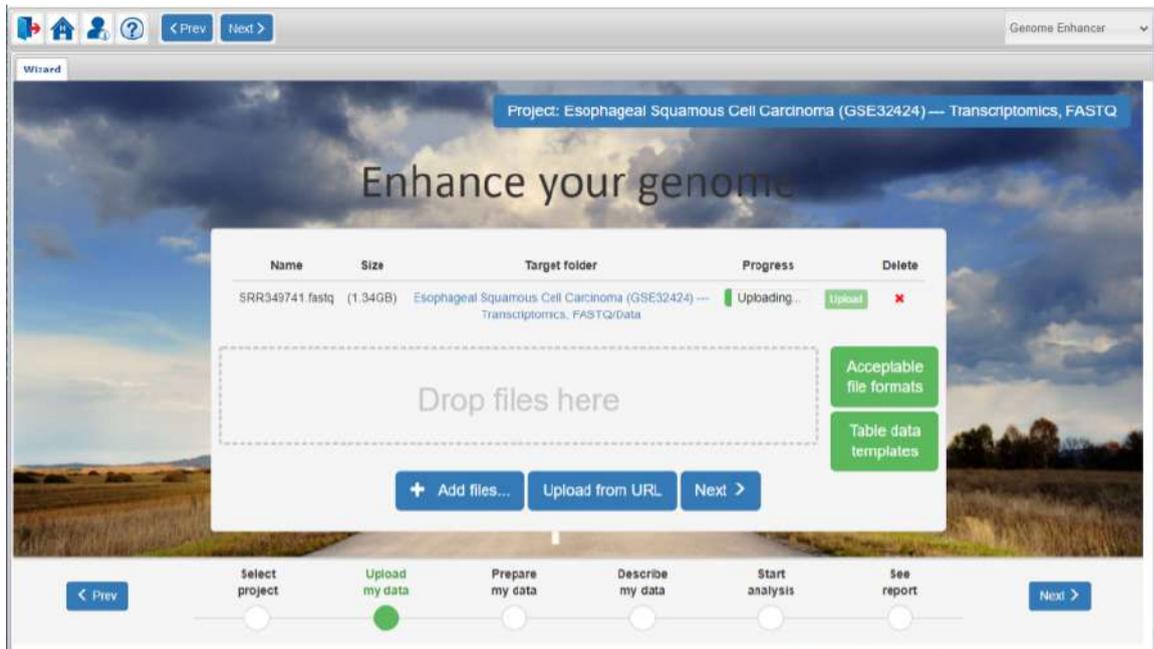


Рисунок п.8.4.1 – Интерфейс платформы Genome Enhancer, мастер (wizard) по проведению анализа, шаг по загрузке данных

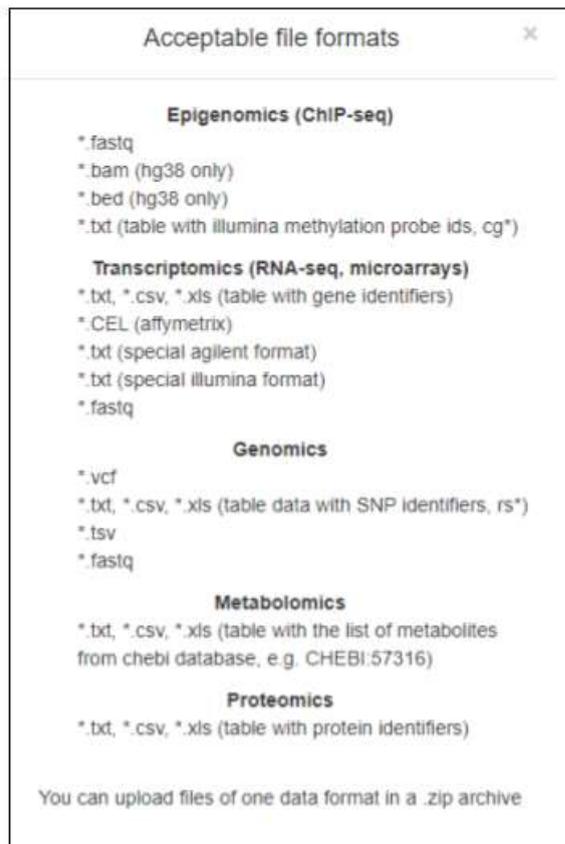


Рисунок п.8.4.2 – Список типов данных и типов файлов, которые можно загрузить для анализа в Genome Enhancer

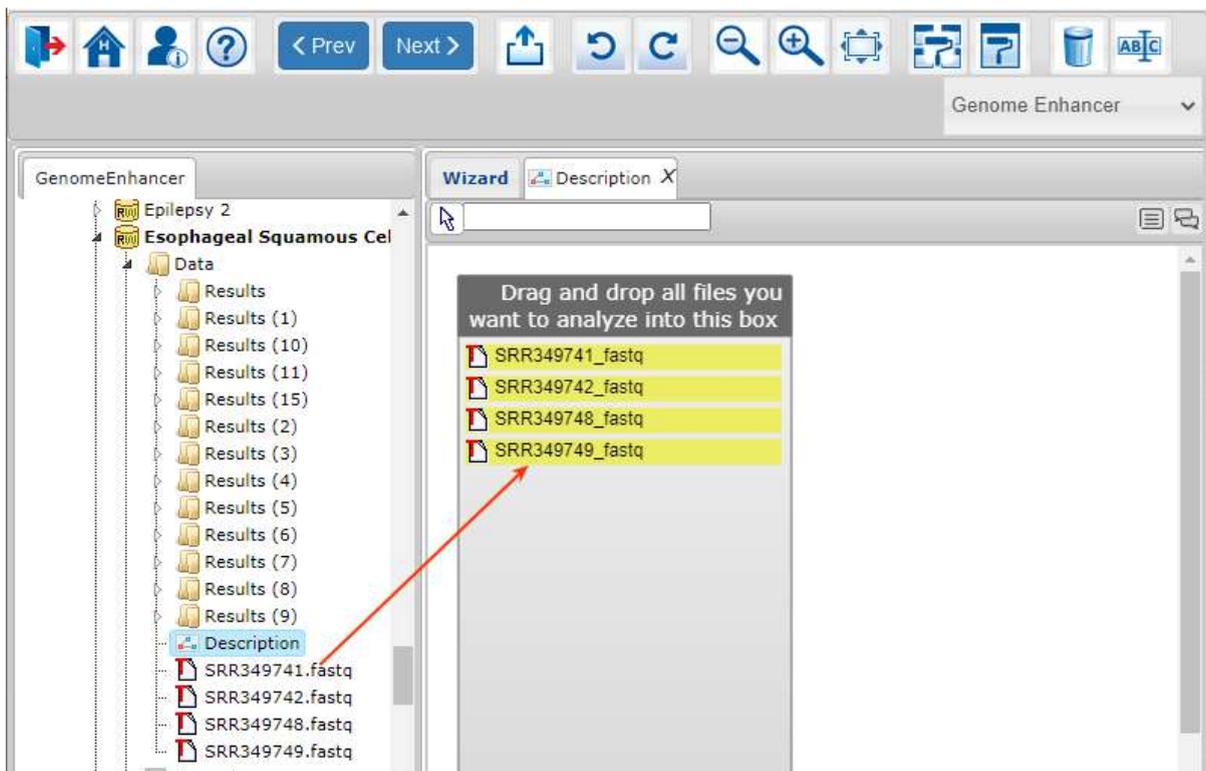


Рисунок п.8.4.3 – Выбор файлов для анализа в платформе Genome Enhancer

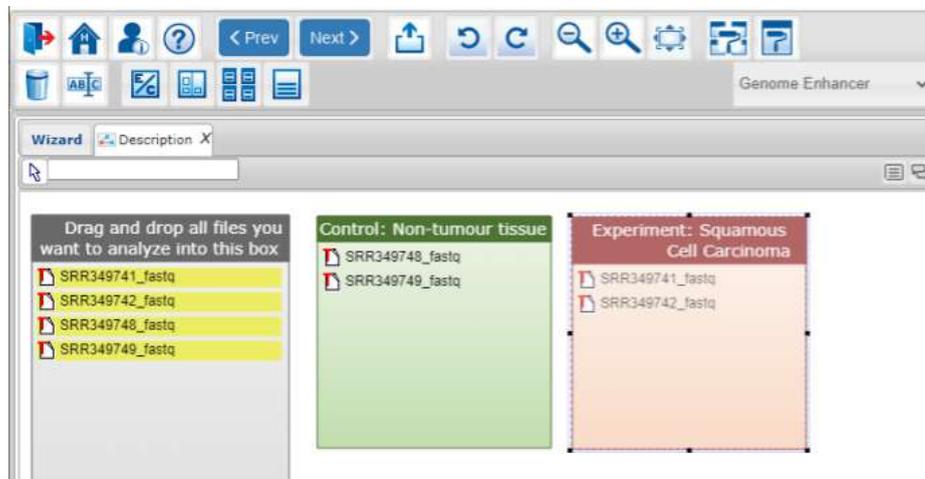


Рисунок п.8.4.4 – Графическая аннотация данных для анализа в платформе Genome Enhancer
(см. пояснение в тексте)

Genome Enhancer

Wizard

Project: Esophageal Squamous Cell Carcinoma (GSE32424) — Transcriptomics, FASTQ

Enhance your genome

Disease:

Tissue:

Select one or more diseases you are studying.

Choose your data for analysis:

Select the conditions you want to compare in your analysis. If only one condition will be analyzed, remove the unnecessary Condition 2 by clicking on *Less conditions* button. If two or more conditions will be analyzed, specify the condition which refers to the baseline (control/background set).

Baseline

Condition 1:

Condition 2:

Select project Upload my data Prepare my data Describe my data **Start analysis** See report

Рисунок п.8.4.5 – Настройка параметров анализа данных в Genome Enhancer

PSMA7 and PSMC5 are promising druggable targets for treating Squamous Cell Carcinoma that control activity of TP53, JUN and SMAD3 transcription factors on promoters of differentially expressed genes

Demo User

geneXplain GmbH

info@genexplain.com

Data received on 13/08/2019 ; Run on 28/01/2021 ; Report generated on 28/01/2021

Genome Enhancer release 2.3 (TRANSFAC®, TRANSPATH® and HumanPSD™ release 2021.1)



Abstract

In the present study we applied the software package "Genome Enhancer" to a data set that contains *transcriptomics* data. The study is done in the context of *Squamous Cell Carcinoma*. The goal of this pipeline is to identify potential drug targets in the molecular network that governs the studied pathological process. In the first step of analysis pipeline discovers transcription factors (TFs) that regulate genes activities in the pathological state. The activities of these TFs are controlled by so-called master regulators, which are identified in the second step of analysis. After a subsequent druggability checkup, the most promising master regulators are chosen as potential drug targets for the analyzed pathology. At the end the pipeline comes up with (a) a list of known drugs and (b) investigational active chemical compounds with the potential to interact with selected drug targets.

From the data set analyzed in this study, we found the following TFs to be potentially involved in the regulation of the differentially expressed genes: TP53, JUN, NFATC2, SMAD3, SMAD2 and FOS. The subsequent network analysis suggested

- IL-1beta-p17:IL-1RI:IL-1RacP:MyD88:tollip:IRAK-1{pS376}{pT387}:IRAK-4:IRAK-2
- PP2A
- 26S proteasome

as the most promising molecular targets for further research, drug development and drug repurposing initiatives on the basis of identified molecular mechanism of the studied pathology. Having checked the actual druggability potential of the full list of identified targets, both, via information available in medical literature and via cheminformatics analysis of drug compounds, we have identified the following drugs as the most promising treatment candidates for the studied pathology: Dasatinib, Bosutinib and {(2Z)-4-AMINO-2-[(4-METHOXYPHENYL)IMINO]-2,3-DIHYDRO-1,3-THIAZOL-5-YL}(4-METHOXYPHENYL)METHANONE.

1. Introduction

Recording "-omics" data to measure gene activities, protein expression or metabolic events is becoming a standard approach to characterize the pathological state of an affected organism or tissue. Increasingly, several of these methods are applied in a combined approach leading to large "multiomics" datasets. Still the challenge remains how to reveal the underlying molecular mechanisms that render a given pathological state different from the norm. The disease-causing mechanism can be described by a re-wiring of the cellular regulatory network, for instance as a result of a genetic or epigenetic alterations influencing the activity of relevant genes. Reconstruction of the disease-specific regulatory networks can help identify potential master regulators of the respective pathological process. Knowledge about these master regulators can point to ways how to block a pathological regulatory cascade. Suppression of certain molecular targets as components of these cascades may stop the pathological process and cure the disease.

Рисунок п.8.4.6 – Пример отчета (фрагмент 1-ой страниц), сформированного платформой Genome Enhancer

ПРИЛОЖЕНИЕ 8.5 ПРИМЕРЫ ИСПОЛЬЗОВАНИЯ ПЛАТФОРМЫ SIRIUS-WEB

Учебные курсы

Учебные курсы объединены в группу <https://gitlab.sirius-web.org/courses/> (рисунок п.8.5.1).

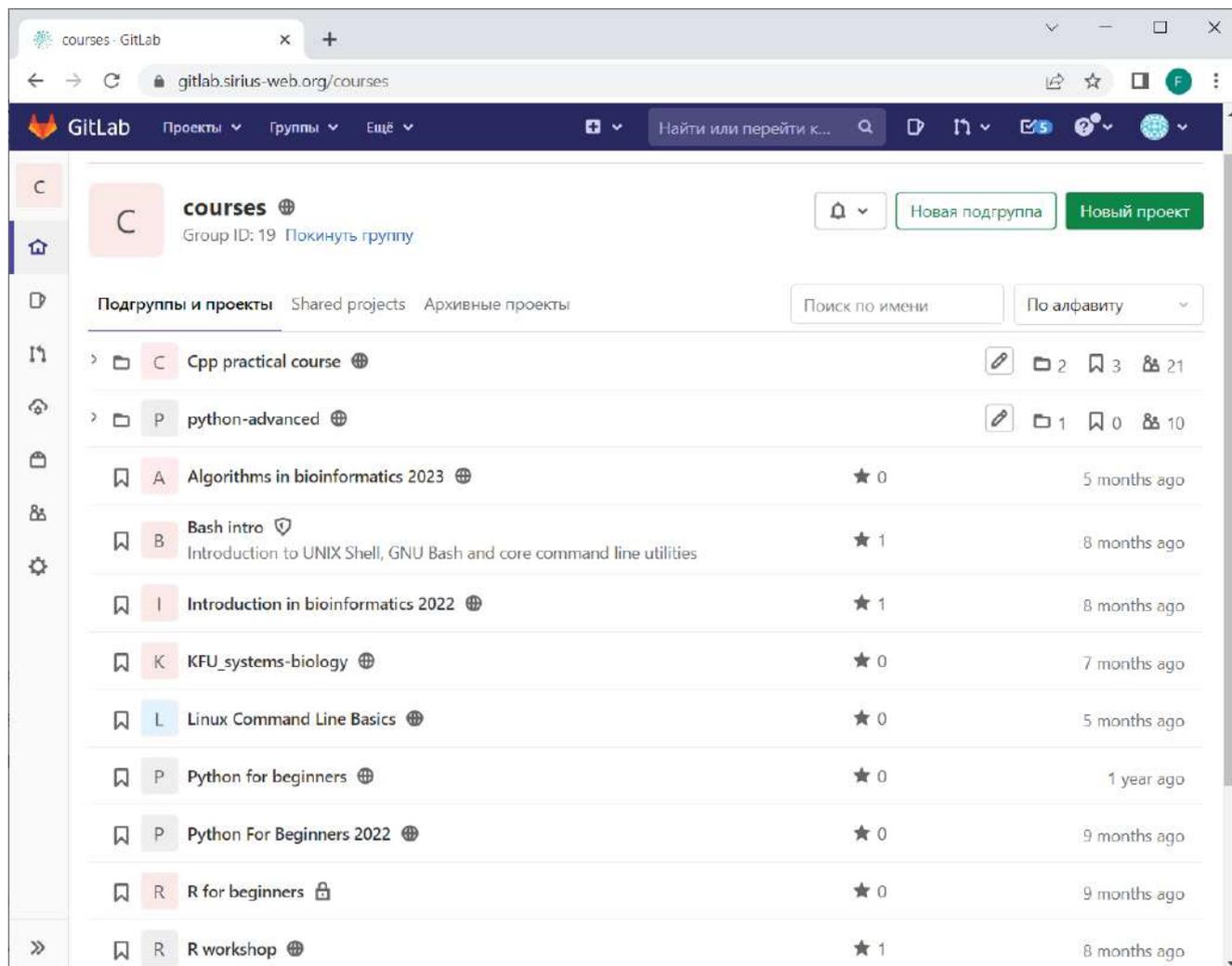


Рисунок п.8.5.1 – Стартовая страница группы «courses» для учебных курсов.

Основные идеи/правила, как организована информация по курсам в репозитории:

- каждый курс представлен в виде отдельного git репозитория (рисунок п.8.5.2);
- если возможно, каждое занятие представлено в виде отдельного Jupyter ноутбука;
- домашние и контрольные задания также представлены в виде Jupyter ноутбуков;
- каждый студент клонирует себе исходный репозиторий и в нем выполняет домашние и контрольные задания;
- преподаватель заходит в репозиторий студента и проверяет задания;

- по завершению курса студент может оставить себе свой репозиторий, в котором пройденные занятия могут служить ему конспектом и справочником по пройденным темам. Кроме того, в ходе обучения студент может в своем репозитории расширять и дополнять конспекты пройденных тем.

На рисунке п.8.5.3 приведен пример выполнения домашнего задания в виде Jupyter ноутбука, открытого в платформе BioUML, в составе платформы Sirius-web.

The screenshot shows the GitLab interface for a repository. At the top, there's a navigation bar with 'GitLab', 'Projects', 'Groups', 'Snippets', and 'Help'. A search bar and 'Sign in / Register' button are also present. Below the navigation bar, there's a sidebar with icons for home, repository, discussions, merge requests, issues, and other features. The main content area shows a table of repository files and folders:

Name	Last commit	Last update
exam	Update exam/Bonpocы_2022.pdf	1 month ago
hands-on trainings	Update hands-on trainings/HT_9-10_Modeli...	1 month ago
homeworks	Update homeworks/HW_8_Modeling_Immu...	1 month ago
lections	Update lections/lection_10_Modeling_Epide...	1 month ago
README.md	Update README.md	1 month ago

Below the table, the 'README.md' file is expanded, showing the following content:

Курс "Системная биология"

Курс по системной биологии используется в программе магистратуры «Биоинформатика и математическая биология» (направление подготовки 01.04.02. Прикладная математика и информатика) в качестве обязательной дисциплины.

Цель курса:

Освоение магистрантами современных методов и подходов математического и компьютерного моделирования биологических систем, позволяющих проводить теоретический анализ механизмов функционирования живых систем и динамических особенностей их поведения на молекулярном, клеточном, организменном и популяционных уровнях их иерархической организации.

Задачи курса:

Формирование у студентов теоретических знаний и практических навыков, связанных с разработкой и анализом математических моделей биологических систем на основе уже имеющейся информации и экспериментальных данных с помощью компьютерной системы BioUML для приобретения новых знаний о механизмах и динамике функционирования изучаемой системы.

Объем курса

Объем дисциплины составляет 3 зачетных единицы, 108 академических часов (ак.ч.), из которых 76 ак. ч. составляет контактная работа обучающихся с преподавателем (26 ак. ч. занятий лекционного типа, 50 ак. ч. лабораторных, практических занятий, семинаров), 32 ак. ч. составляет самостоятельная работа обучающихся.

Рисунок п.8.5.2 – Стартовая страница и структура репозитория для курса "Системная биология"

Research: Covid19 | Default

Start page | hw0.ipynb X | pract.ipynb X

Edit View Insert Cell Kernel Navigate Help | Not Trusted | JS_BioUML

1) Оптимизация модели CD95-NFκB

```
In [1]: M var diagram = data.get("data/Collaboration/sysbio22_kondrakhin/Data/Homeworks/Homework_0/CD95-NFκB_simple_model");
var expData = data.get("data/Collaboration/sysbio22_kondrakhin/Data/Homeworks/Homework_0/exp_data_1");
var span = se.createSpan(0, 1000, 1.0);
var result = se.simulate(diagram, span);

plot = pm.createPlot("CD95-NFκB optimized model", "time, min", "Concentration, mole/l");
plot.addCurve("p43p41", result, "$cell.p43p41", "time", pm.createPen(1, "red", "-"));
plot.addExperiment("p43p41 exp", expData, "p43p41", "time", pm.createPen(1, "red", "-"));
plot.addCurve("casp8", result, "$cell.C8star", "time", pm.createPen(1, "blue", "-"));
plot.addExperiment("casp8 exp", expData, "casp8", "time", pm.createPen(1, "blue", "-"));

var htmlTable = html.createTable();
htmlTable.addRow(plot.getImage());
htmlTable.out();
```

Out[1]:

Больше всего на динамику изменения концентрации p43p41 и casp8 влияют параметры **k2**, **k5**, **k8** и **k11**.
k2 - константа скорость синтеза соединения, участвующего в синтезе p43p41. **k5** - константа скорости синтеза p43p41. **k8**, **k11** - константы

Рисунок п.8.5.3 – Пример выполнения домашнего задания в виде Jupyter ноутбука, открытого в платформе BioUML, в составе платформы Sirius-web

Дипломные проекты

Платформа Sirius-web была использована для выполнения ряда дипломных проектов, связанных с моделированием сложных биологических систем и анализом биомедицинских данных. Стоит отметить, что во многих случаях платформа обеспечивала весь функционал, необходимый для написания дипломной работы.

Как было подчеркнуто выше, такой подход позволяет:

- сохранить в виде git репозитория все результаты работы студента;
- легко воспроизвести результаты, представленные в дипломе – для этого создается соответствующий Jupyter ноутбук;
- описать и задокументировать результаты. С одной стороны, для этого служит сам текст диплома, с другой стороны для этого используется документирование при помощи страниц wiki.

Ниже приведено несколько примеров дипломных работ, выполненных под руководством автора, и соответствующие им репозитории в платформе Sirius web:

- «Построение многоуровневой математической модели эпилептических припадков», Кондрахин П.Ю., диплом магистра, 2023 г., направление «Вычислительная биология», научно-технологический университет «Сириус»
<https://gitlab.sirius-web.org/diploms/2021/kondrakhin.py>
- «Предсказания эффекта SNV на регуляцию транскрипции», Ежов Д.О., диплом магистра, 2023 г., направление «Вычислительная биология», научно-технологический университет «Сириус»
<https://gitlab.sirius-web.org/diploms/2021/ezhov.do>
- «Использование нейронных сетей и методов машинного обучения для предсказания уровня экспрессии генов и их функциональности», Суханов Н.А., диплом магистра, 2023 г., направление «Вычислительная биология», научно-технологический университет «Сириус»
<https://gitlab.sirius-web.org/diploms/2021/sukhanov.na>
- «Математическое моделирование распространения SARS-CoV-2 по организму и его взаимодействия с иммунной системой человека», Шукин И.А., диплом бакалавра, 2021 г., Механико-математический факультет, Новосибирский Государственный Университет
<https://gitlab.sirius-web.org/diploms/covid19-compartments>

Научные проекты

Платформа Sirius web позволяет не только накапливать и структурировать научно-образовательных результатов в виде git репозитория, но также управлять ходом исследований, следуя современным Agile технологиям, используя возможности платформы GitLab.

Данный подход был апробирован нами на примере проекта по гранту РФФИ № 21-75-30020 "Исследование и моделирование метаболических и гомеостатических переменных при онкотранспорте наночастиц и доставке лекарственных средств", руководитель – д.б.н. Замятнин А.А. В данном случае, для проекта была создана отдельная группа:

<https://gitlab.sirius-web.org/nanoparticles>

Она включает 3 Git репозитория:

- *<https://gitlab.sirius-web.org/nanoparticles/main>*
репозиторий для для экспериментальных данных;
- *<https://gitlab.sirius-web.org/nanoparticles/models>*
подпроект по математическому моделированию доставки наночастиц в раковые опухоли (Kutumova et al., 2022b). В этом репозитории хранятся математические модели, разработанные в платформе BioUML, Jupyter ноутбуки с графиками и кодом для воспроизведения результатов моделирования, а также документация по модели. Данный репозиторий подключен к платформе BioUML, поэтому перейдя по соответствующей ссылке (рисунок п.8.5.4) пользователь сразу может открыть соответствующие модели и ноутбуки в платформе BioUML.
- *<https://gitlab.sirius-web.org/nanoparticles/nanoparticles.pages.sirius-web.org>*
платформа GitLab позволяет создавать простые веб-сайты. В данном репозитории хранится код для сайта-визитки по этому проекту. Сам сайт доступен по следующей ссылке – *<https://nanoparticles.pages.sirius-web.org>*.

Наименование	Последний коммит	Последнее обновление
Diagrams	Latest	6 months ago
Images	Upload New File	6 months ago
.gitignore	Sync	6 months ago
.project	Config files	1 year ago
Dockerfile	latest	3 months ago
Nanoparticles.format.config	Config files	1 year ago
README.md	Update README.md	6 months ago
default.config	Config files	1 year ago
default.primary.config	Config files	1 year ago
description.html	Config files	1 year ago

README.md

PBPK model of nanoparticle delivery in mice

[launch binder](#)



Following the divide-and-conquer strategy, modular modeling is an effective way to simulate complex biological processes. Physiologically based pharmacokinetic (PBPK) models represent one of the classes of such complex models allowing to take into account both the physiological characteristics of the organism (the size of organs, their composition and blood flow), and the physicochemical properties of the drug. The consideration of the organism as a set of compartments representing individual organs or their groups used in this approach, sets the natural division of the model into modules. One part of the modules coincides in structure but differs in parameters, while the other part varies both in structure and in parameters. Herein, we demonstrate how the PBPK model of nanoparticle delivery to solid tumors, as an example, can be represented in the modular graphical view in the BioUML platform. The main advantage of the proposed approach is its visibility and ease of perception. Also, the modular representation simplifies the work with the model and reduces the risk of technical errors during its reuse and extension by the accompanying biochemical processes.

Рисунок п.8.5.4 – Стартовая страница Git репозитория подпроекта по математическому моделированию доставки наночастиц в раковые опухоли. Используя ссылку (иконка BioUML) пользователь может открыть этот проект в платформе BioUML

Для управления проектами платформа GitLab представляет возможность разбить проект на этапы (рисунок п.8.3.5, верхняя часть). В данном случае в подпроекте по математическому моделированию доставки наночастиц в раковые опухоли мы выделили этапы в соответствии с основными группами экспериментов, проводимых в рамках этого гранта.

В каждом этапе формируется набор задач, чей статус может быть отражен на специальной канбан доске (рисунок п.8.3.5, нижняя часть). В правой части доски ("Closed") располагается список уже выполненных задач, а в левой части доски ("Открыто") – задачи, которые в ходе выполнения или нужно еще выполнить.

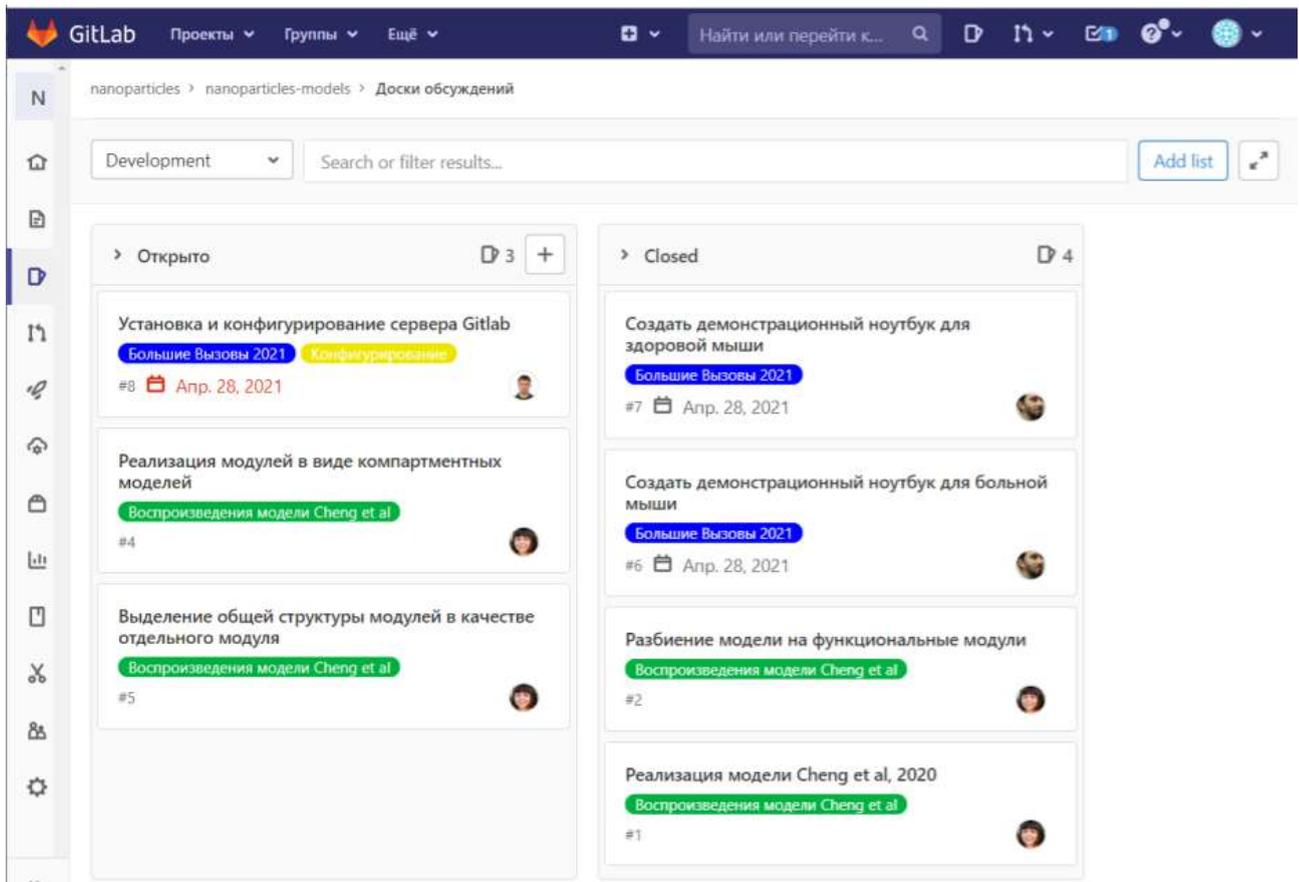
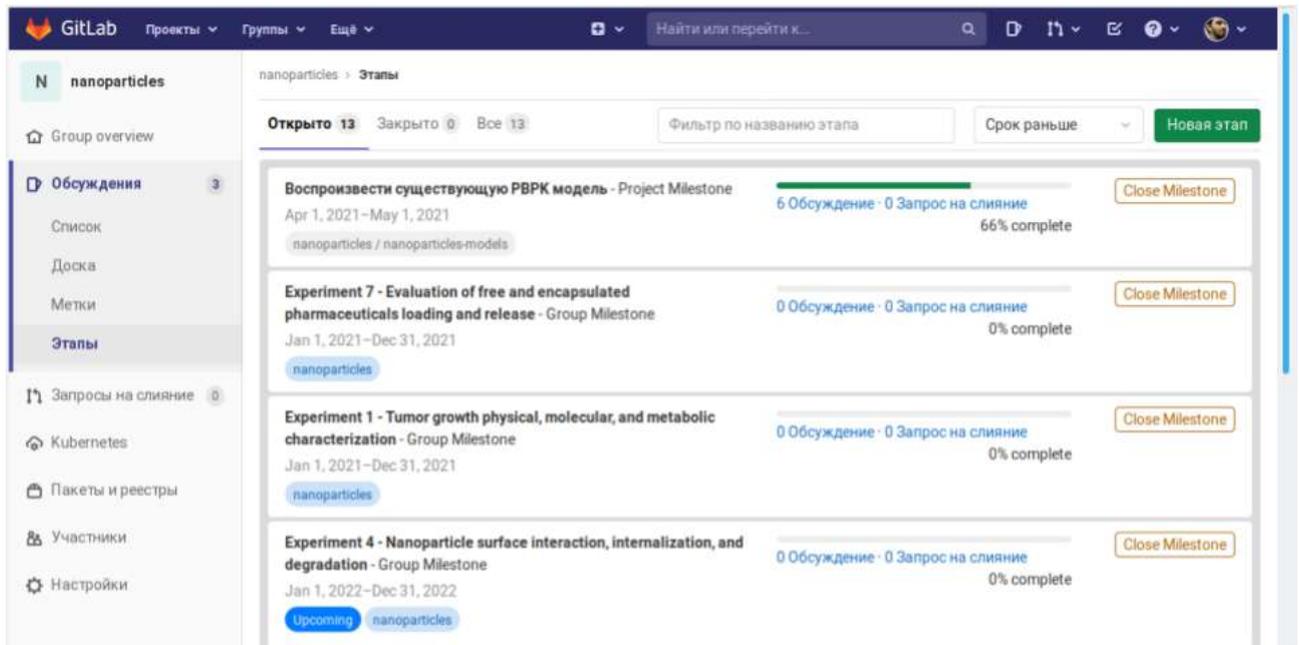


Рисунок п.8.3.5 – Пример управления проектом по моделированию доставки наночастиц в раковые опухоли в платформе GitLab, сверху – разбиение проекта на этапе, снизу – канбан доска с задачами 1-го этапа

Подсистема wiki в платформе GitLab позволяет создавать подробную документацию по модели в виде набора редактируемых страниц (рисунок п. 8.3.6).

The screenshot shows a web browser window displaying a GitLab wiki page. The browser tabs include 'nanoparticles - GitLab' and 'Physiologically based pharmacokinetic (PBPK) modelling'. The address bar shows the URL 'gitlab.sirius-web.org/nanoparticles/models/-/wikis/Physiologically-base...'. The page title is 'Physiologically based pharmacokinetic (PBPK) modelling', last edited by 'Elena Kutumova 1 год назад'. The main content features a diagram of a human torso with labels for 'Brain', 'Blood vessels', 'Heart', 'Lungs', 'Liver', 'Stomach', and 'Intestines'. To the right is a schematic diagram of a compartmental model with 'Venous blood' and 'Arterial blood' compartments, and internal compartments for 'Lung', 'Adipose', 'Bone', 'Brain', 'Heart', 'Muscle', 'Skin', 'Liver', and 'Kidney'. Below the diagram is a paragraph explaining PBPK modeling as a technique for predicting ADME. The right sidebar contains options to 'Клонировать репозиторий' and 'Edit sidebar', and a search bar with the text 'Nanoparticle delivery to tumors using a PBPK modeling'.

Рисунок п.8.3.6 – Пример документации по модели в подсистеме wiki платформы GitLab

ПРИЛОЖЕНИЕ 8.6 АПРОБАЦИЯ ПЛАТФОРМЫ U-SCIENCE В ПРОЕКТЕ "СУПЕР С-ТАУ ФАБРИКА"

Супер С-тау фабрика – это симметричный электрон-позитронный коллайдер ультравысокой светимости с энергией пучков в системе центра масс от 2 до 6 ГэВ, ее проект разрабатывается в Институте Ядерной Физики СО РАН им. Г.И. Будкера. Цель этого проекта – проведение фундаментальных экспериментов для прецизионная проверка существующей теории (Стандартной Модели, СМ) и поиск феноменов, не описываемых в рамках СМ.

Проект включает в себя уникальный ускорительно-накопительный комплекс со светимостью $10^{35} \text{ см}^{-2}\text{с}^{-1}$ и универсальный детектор элементарных частиц. Платформа u-science обеспечивает интеграцию и веб-интерфейс для программного обеспечения для моделирования и обработки данных с детектора, которое будет использоваться для оптимизации конструкции детектора Супер С-тау фабрики, для разработки алгоритмов анализа данных и решения других задач. Для этого было создан ряд специальных docker образов и специализированных модулей, которые описаны ниже.

Фреймворк Aurora

Модульная среда (фреймворк) Aurora (Zhadan et al., 2021) разрабатывается в Институте Ядерной Физики СО РАН им. Г.И. Будкера на основе универсальной модельной среды GAUDI (Barrand et al., 2001). Для запуска моделирования и анализа данных Супер с-тау фабрики, среда Aurora была интегрирована в платформу u-science с помощью специального docker образа, содержащего:

- базовую операционную систему Scientific Linux 7;
- подключаемый по NFS фреймворк Aurora;
- средства его настройки и запуска сценариев.

Используя этот образ, пользователь может в среде u-science:

- запустить сценарии для генерации событий, их моделирования, оцифровки и реконструкции в детекторе (рисунок п.8.6.1);
- просмотреть полученные результаты (рисунок п.8.6.2).

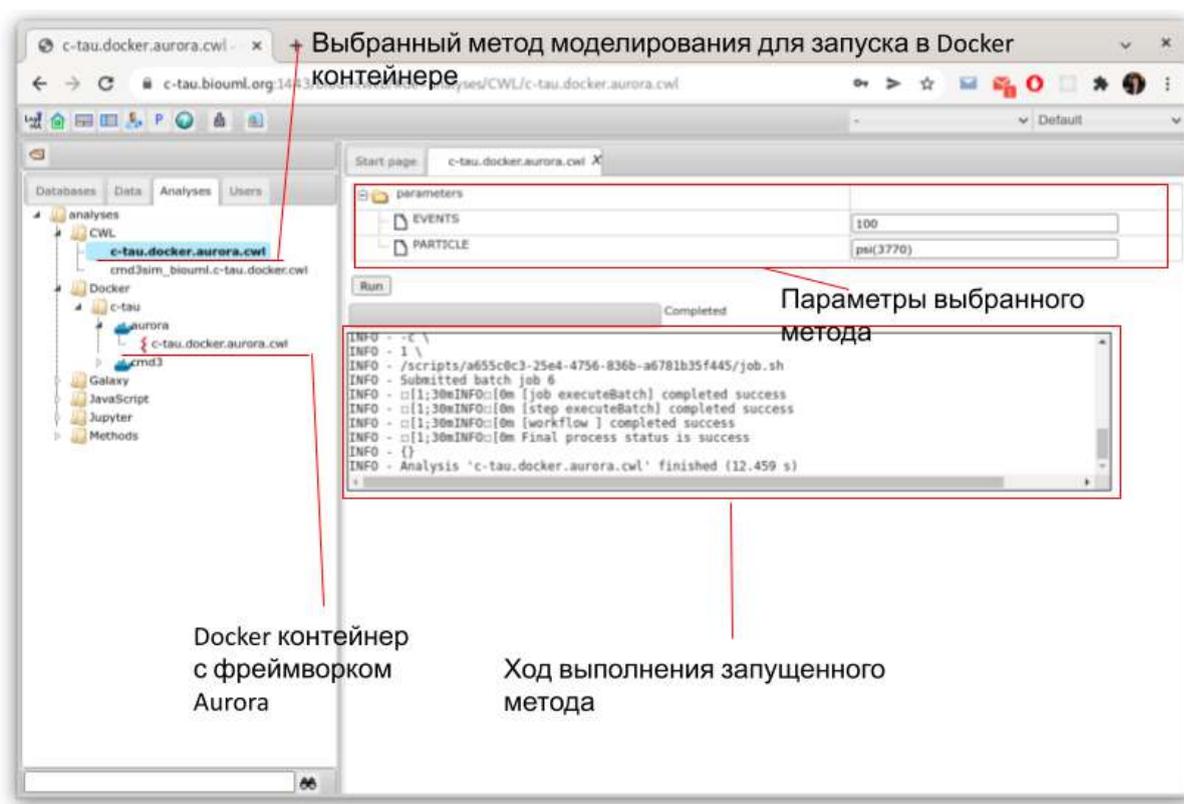


Рисунок п.8.6.1 – Интеграция среды Ауорога в платформу u-science. Запуск анализа загруженного в Docker контейнер

Платформа Root

Root (<http://root.cern.ch>) – платформа для анализа данных с открытым исходным кодом, которая широко используется в физике высоких энергий. Она также включает собственный платформо-независимый формат файлов данных. Для интеграции Root в платформу u-science были разработаны:

- docker образ, который включает в себя ядро для Jupyter hub, которое позволяет выполнить команды root в Jupyter notebook;
- модуль, который содержит библиотеку jsRoot (<https://root.cern.ch/js/>), которая позволяет интерактивно просматривать и визуализировать содержимое root файлов (рисунок п. 8.6.2).

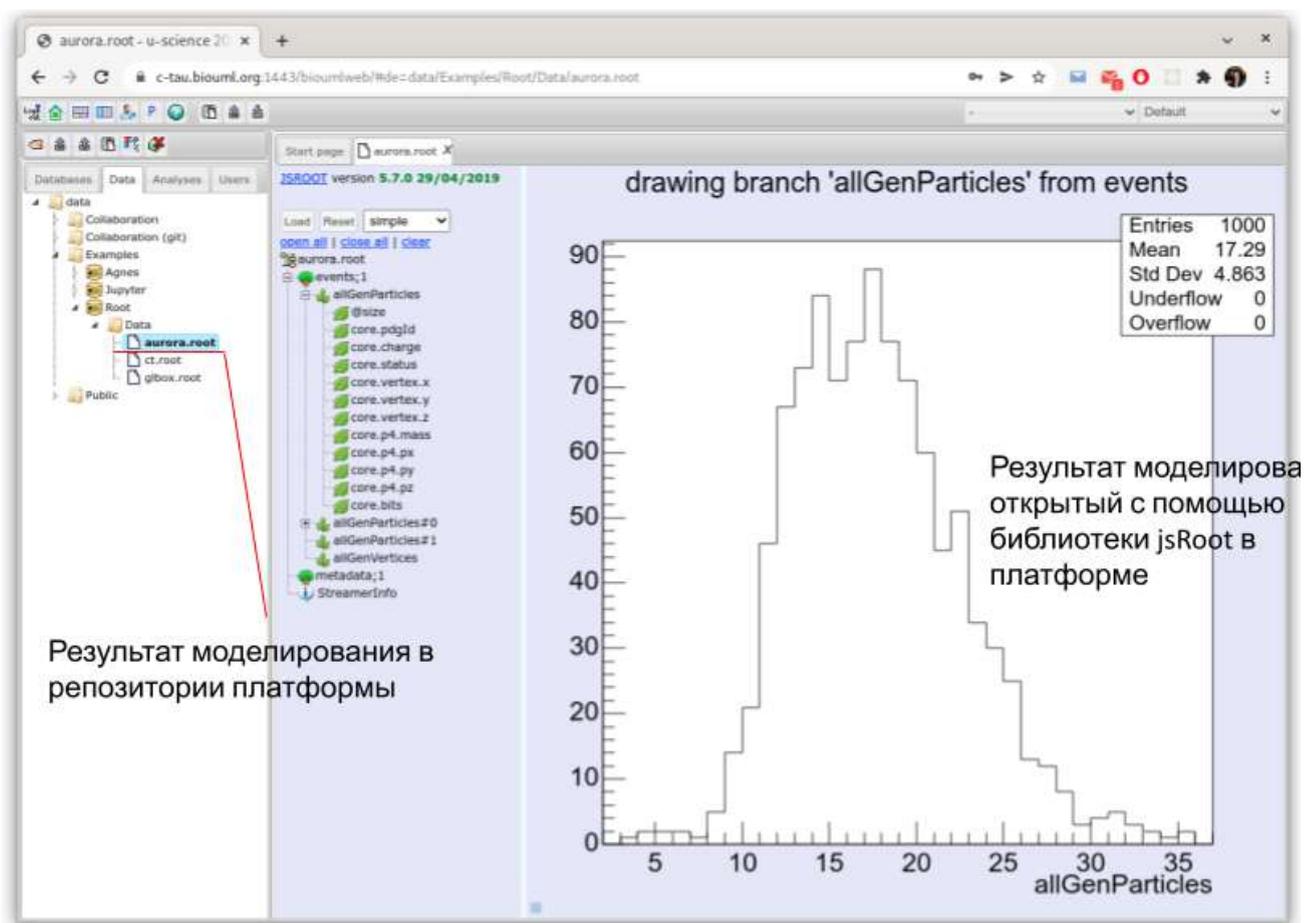


Рисунок п.8.6.2 – Просмотр результатов моделирования в физике высоких энергий, используя библиотеку jsRoot

Система AGNES

В ходе проведения экспериментов с детектора элементарных частиц супер С-тау фабрики будет получено порядка 100 петабайт «сырых» данных, поэтому один из ее важнейших элементов – это система обработки и хранения данных, в задачи которой входит первичная обработка данных, передача данных на хранение в долговременную систему хранения данных и обработка с использованием высокопроизводительных вычислительных систем.

В рамках проекта предлагается решить проблему проектирования вычислительной инфраструктуры и многоуровневых систем хранения данных экспериментов путем имитационного моделирования.

Для такого моделирования используется система AGNES, разработанная в ИВМиМГ СО РАН для моделирования телекоммуникационных и информационных сетей (Chernykh et al., 2014; Weins et al., 2019). Основной областью применения AGNES является имитационное моделирование распределённых сложных неоднородных систем с дискретными событиями, в частности вычислительных систем любой сложности, таких как локальные вычислительные

сети, сенсорные сети, суперкомпьютерные центры коллективного пользования и т.д. AGNES построена на основе платформы мультиагентного программирования Jade (<https://jade.tilab.com/>).

Для интеграции системы AGNES в платформу u-science был разработан модуль, обеспечивающий графическое представление и редактирование в виде диаграмм вычислительной инфраструктуры Супер С-тау Фабрики и его сопряжение с имитационной моделью. Модуль использует библиотеки Jade и AGNES, а также Java кода для графического представления моделей AGNES, генерации соответствующего xml кода и запуска моделей.

Каждому типу агентов соответствует свое графическое обозначение (рисунок п.8.6.3). Каждый объект на диаграмме соответствует одной группе агентов с одинаковыми свойствами. В скобках указываются количества для агентов каждой группы. Стрелка между объектами обозначают связи между агентами. При запуске расчетов по визуальному представлению генерируется xml файл с настройками AGNES и запускается симуляция с помощью Jade.

Например, для решения задач по хранению и обработке данных поступающих с Супер С-тау Фабрики в имитационной модели предусмотрена имитация работы параллельной системы хранения данных (Weins et al., 2019). В модель введены интеллектуальные программные агенты коллекторы данных. Каждый агент коллектор данных содержит в себе счетчик записанных данных, информацию о состоянии и характеристиках всех полок с носителями информации, находящимися под его управлением. При обращении к нему агентов контроллеров работающих доменов он передает им информацию о времени, необходимом для чтения и записи искомого объема данных. Разработанная имитационная модель предусматривает все аспекты функционирования вычислительной инфраструктуры от параллельной системы хранения данных до организации параллельного запуска пользовательских заданий (Weins et al., 2019).

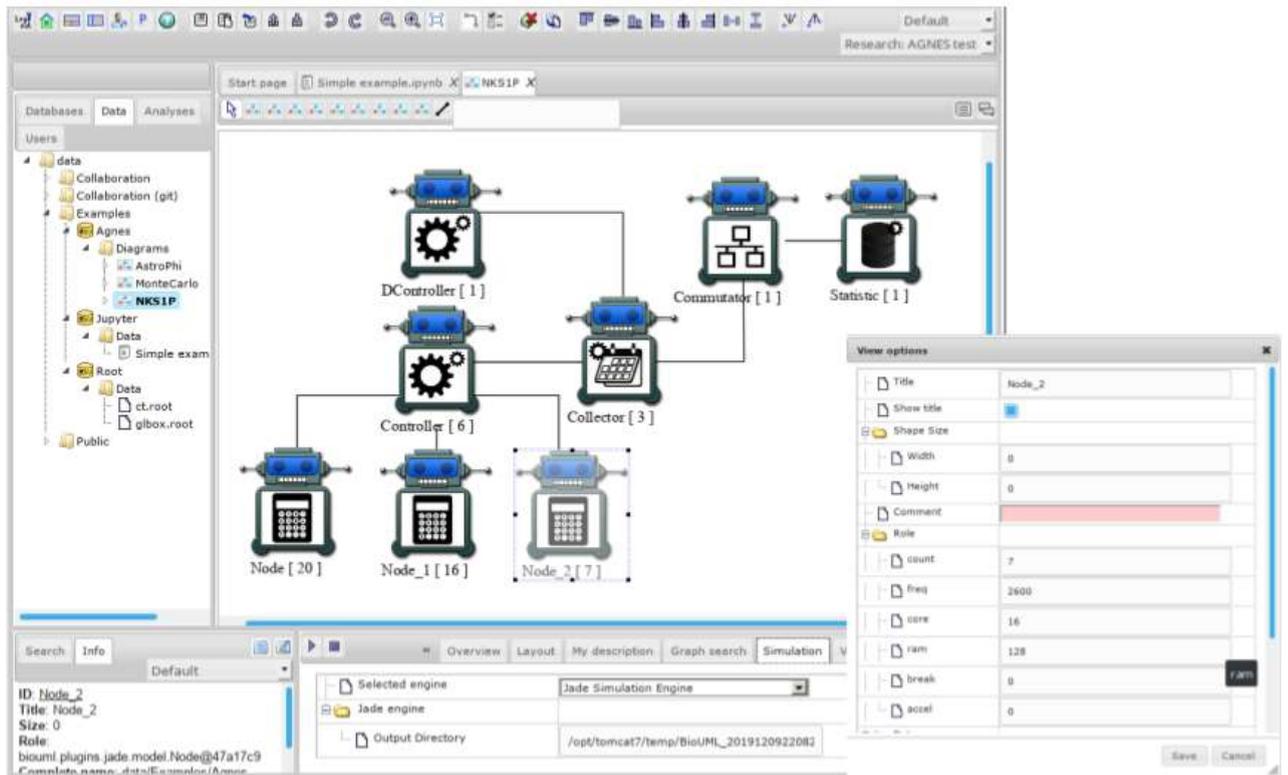


Рисунок п.8.6.3 – Графическое представление имитационной модели AGNES в платформе u-science (см. описание выше)

В результате такого подхода можно будет заранее оценить структуру и объемы данных, производимых физической установкой, а также просчитать эффективные характеристики вычислительной системы и системы хранения данных, что позволит сформировать проект на создание системы хранения данных и высокопроизводительной вычислительной системы для обработки данных, который будет рассчитан на структуру и объемы экспериментальных данных Супер С-тау Фабрики.

**Приложение 8.7 Акты внедрения результатов диссертационной работы: ПК
ВioUML и разработанных на его основе компьютерных систем**



АКТ

о внедрении результатов работы Колпакова Ф.А. в проектах, выполняемых ИАП РАН

Научно-техническая комиссия в составе:

- руководителя научного направления, д-ра техн. наук Курочкина В.Е.,
- ведущего инженера, к.т.н. Петрова А.И.,
- старшего научного сотрудника, к.ф.-м.н. Бородинова А.Г.

составила настоящий акт в том, что в Институте аналитического приборостроения РАН при выполнении работ по теме Исследовательской программы «Разработка опытного образца одномолекулярного секвенатора ДНК, набора реагентов и расходных материалов для него» в рамках Соглашения о предоставлении из федерального бюджета грантов в форме субсидий в соответствии с пунктом 4 статьи 78.1 Бюджетного кодекса Российской Федерации от 28.09.2021 г. № 075-15-2021-1057, а именно, при разработке биоинформатической платформы для хранения, анализа и графического представления данных, полученных при одномолекулярном секвенировании ДНК, был использован программный комплекс (ПК) BioUML и результаты диссертационной работы Колпакова Ф.А. на тему «Компьютерное моделирование биологических систем и анализ биомедицинских данных».

ПК BioUML является полнофункциональной отечественной биоинформатической платформой, обеспечивающей весь функционал для глубокого анализа и графического представления генетических данных. Это особенно важно для обеспечения технологического суверенитета Российской Федерации в этой области. Использование ПК BioUML в комплексе с разрабатываемым одномолекулярным секвенатором ДНК, набором реагентов и расходных материалов для него позволяет обеспечить готовое решение "под ключ" для проведения комплексных генетических исследований.

Руководитель научн. направления, д-р техн. наук  Курочкин В.Е.,

Ведущий инженер, к.т.н.  Петров А.И.,

Старший научный сотрудник, к.ф.-м.н.  Бородинов А.Г.

МИНОБНАУКИ РОССИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ НАУЧНОЕ УЧРЕЖДЕНИЕ

**«ФЕДЕРАЛЬНЫЙ
ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР
ИНФОРМАЦИОННЫХ И
ВЫЧИСЛИТЕЛЬНЫХ
ТЕХНОЛОГИЙ»
(ФИЦ ИВТ)**

Пр-т Ак. Лаврентьева, 6, г. Новосибирск, 630090
Тел.: +7 (383) 330-6150; факс: +7 (383) 330-1824
e-mail: ict@ict.nsc.ru; http://www.ict.nsc.ru

19.06.2023 № 15312-01/5114/163

на № _____ от _____



УТВЕРЖДАЮ

И.о. директора ФИЦ ИВТ

Д.Ф.-М.Н.  Медведев С.Б.« 19 » июня 2023 г.

АКТ

о внедрении результатов диссертационной работы

Колпакова Федора Анатольевича на тему

"Компьютерное моделирование биологических систем и анализ биомедицинских данных"

Комиссия в составе академика, д.ф.-м.н. Шокина Юрия Ивановича, к.ф.-м.н. Кутумовой Елены Олеговны и Кондрахина Юрия Васильевича рассмотрела вопрос об использовании результатов диссертационной работы Колпакова Федора Анатольевича и установила, что разработанные в рамках диссертационной работы ПК BioUML и база данных GTRD внедрены и активно используются в научной работе, а именно:

- ПК BioUML активно используется в лаборатории биоинформатики для выполнения проектов по государственному заданию, в частности для построения компьютерных моделей регуляции артериального давления;
- ПК BioUML стал основой для выполнения договора № 33.2022 на выполнение научно-исследовательских работ по теме "Разработка и испытания биоинформатической платформы для хранения, анализа и графического представления данных, полученных при одномолекулярном секвенировании ДНК", заключенного с Федеральным государственным бюджетным учреждением науки Институт аналитического приборостроения Российской академии наук.
- При помощи ПК BioUML были разработаны компьютерные модели для выполнения работ по гранту РФФИ № 20-04-60355 "Разработка

мультимасштабной иммуно-эпидемиологической математической модели COVID-19 с учетом воздействия на экономику региона и сценариев действия органов власти".

- База данных GTRD и ПК BioUML активно использовались при выполнении гранта РФФИ № 19-14-00295 "Интеграция и анализ омиксных данных по регуляции транскрипции, оценка влияния SNV".

Председатель комиссии:

научный руководитель
академик, д.ф.-м.н.



Ю.И. Шокин

Члены комиссии:

с.н.с., к.ф.-м.н.



Е.О. Кутумова

с.н.с., к.б.н.



Ю.В. Кондрахин

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ



Всех знаний — Вершиной Науки

Федеральное государственное бюджетное
учреждение науки Государственный научный центр
Российской Федерации

**ИНСТИТУТ
МЕДИКО-БИОЛОГИЧЕСКИХ ПРОБЛЕМ**

Российской академии наук
(ГНЦ РФ-ИМБП РАН)

ИНН/КПП 7714038980/771401001
ОГРН 1027739333710

Хорошевское шоссе, д. 76А, г. Москва, 123007

телефон: (499) 195-15-73,

факс: (499) 195-22-53

e-mail: doc@imbp.ru

http://www.imbp.ru

11.07.2023 № 109/ *16/1694*

На № _____ от _____

«УТВЕРЖДАЮ»

Директор ГНЦ РФ ИМБП РАН

Орлов О.И.

«11» июля 2023 г.



АКТ

о внедрении результатов разработок по компьютерному моделированию биологических систем и анализу биомедицинских данных

Комиссия в составе:

- д.б.н. Попова Даниила Викторовича
- к.ф.-м.н. Боровика Анатолия Стратоновича
- к.б.н. Махновского Павла Александровича

составила настоящий акт о том, что научные результаты, полученные Колпаковым Федором Анатольевичем при подготовке его диссертационной работы, а именно:

- 1) база данных GTRD (Gene Transcription Regulation Database <https://gtrd20-06.biouml.org/>),
- 2) платформа для анализа биомедицинских данных BioUML <https://ict.biouml.org/> и разработанная на ее основе платформа для анализа биоинформатических данных geneXplain <https://genexplain.com/>

внедрены и активно использовались в работе при выполнении научных проектов:

- грант РФФИ 17-00-00242 Молекулярные механизмы регуляции экспрессии генов и белков в скелетной мышце человека в ответ на острое и хроническое стресс-воздействие физических нагрузок.

- грант РФФИ 19-315-90135 Поиск транскрипционных регуляторов, экспрессирующихся в скелетной мышце человека в ответ на аэробную физическую нагрузку, методом целевой протеомики.
- грант РФФИ 20-015-00415 Поиск транскрипционных факторов, играющих ключевую роль в регуляции митохондриального биогенеза и углеводно-жирового обмена в скелетной мышце человека при сократительной активности.
- грант РФФИ 21-15-00405 Регуляция митохондриального биогенеза в скелетной мышце при изменении уровня двигательной активности и старении.
- грант РФФИ 21-75-10146 Внутриклеточная сигнализация и генная экспрессия в скелетной мышце человека при ожирении и сахарном диабете 2 типа: поиск новых регуляторных механизмов и потенциальных биомаркеров.

д.б.н., проф. РАН, в.н.с.-зав.лаб
физиологии мышечной деятельности
ГНЦ РФ ИМБП РАН



Попов Д.В.

к.ф.-м.н., в.н.с. лаб. физиологии
мышечной деятельности ГНЦ РФ ИМБП РАН



Боровик А.С.

к.б.н., н.с. лаб. физиологии
мышечной деятельности ГНЦ РФ ИМБП РАН



Махновский П.А.

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное учреждение науки
Институт белка Российской академии наук
(ИБ РАН)

ул. Институтская, д. 4, г. Пущино, Московская обл., 142290
Тел./факс +7(495)514-02-18; факс +7(4967)318-435 E-mail: protres@vega.protres.ru
ОКПО 02699748, ОГРН 1025007773262, ИНН/КПП 5039001220/503901001

УТВЕРЖДАЮ
Заместитель директора ИБ РАН



д.х.н. А.Д. Никулин
"8" июня 2023г.

АКТ

**о внедрении результатов диссертационной работы
Колпакова Федора Анатольевича на тему
"Компьютерное моделирование биологических систем
и анализ биомедицинских данных"**

Результаты диссертационной работы Колпакова Федора Анатольевича и разработанные в рамках диссертационной работы программный комплекс (ПК) BioUML и база данных GTRD внедрены и активно используются в научной работе при биоинформатическом анализе данных по генной регуляции в Группе регуляции биосинтеза белка, а именно:

- ПК BioUML и база данных GTRD применены для поиска регуляторов транскрипции генов Y-букс-связывающих белков, в частности, результаты, полученные при помощи ПК, включены в обзор: Елисеева И.А., Согорина Е.М., Смолин Е.А., Кулаковский И.В., Лябин Д.Н. МНОГООБРАЗИЕ СПОСОБОВ РЕГУЛЯЦИИ КОЛИЧЕСТВА YB-1 И ЕГО ГОМОЛГОВ У МЛЕКОПИТАЮЩИХ Успехи Биологической Химии, т. 62, с. 99–19509 (2022).
- ПК BioUML и база данных GTRD успешно использованы для выполнения научного проекта, поддержанного грантом РНФ 20-74-10075 «Аллель-специфичная доступность хроматина и генетические детерминанты патологий» (2021-2023), при создании и обновлении серии баз данных и интернет-сервисов HOCOMOCO, ADASTRA и ANANASTRA, поддерживаемых в Группе регуляции биосинтеза белка и посвященных генной регуляции у человека.

Ответственный за внедрение:

к.ф.-м.н., д.б.н., в.н.с.
Кулаковский И.В.

Министерство науки и высшего образования Российской Федерации
 Федеральное государственное автономное образовательное учреждение
 высшего образования «Казанский (Приволжский) федеральный университет»

«УТВЕРЖДАЮ»

Проректор по биомедицинскому направлению

Д-р мед. наук, проф.

А. П. Киясов

2023 г.



АКТ

о внедрении результатов диссертационной работы

Колпакова Федора Анатольевича

на тему

**"Компьютерное моделирование биологических систем и анализ
 биомедицинских данных"**

Комиссия в составе к.б.н. Гусева Олега Александровича, к.б.н. Шагимардановой Елены Ильясовны и к.б.н. Козловой Ольги Сергеевны рассмотрела вопрос об использовании результатов диссертационной работы Колпакова Федора Анатольевича и установила, что разработанные в рамках диссертационной работы ПК BioUML, платформа Sirius-web и база данных GTRD внедрены и активно используются в научно-образовательном процессе, а именно:

1. в научном проекте, выполняемый по Соглашению о предоставлении из федерального бюджета грантов в форме субсидий в соответствии с пунктом 4 статьи 78.1 Бюджетного кодекса Российской Федерации № 075-15-2021-1344 от 04.10.2021 г.

2. программе дополнительного профессионального образования «Системная биология» Центра последипломного образования Института фундаментальной медицины и биологии КФУ.

Председатель комиссии:

Директор Научного центра «Регуляторная геномика»

к.б.н. Гусев Олега Александрович

Члены комиссии:

в.н.с. НИЛ «Молекулярная вирусология»

к.б.н. Шагимарданова Елена Ильясовна

с.н.с. НИЛ «Молекулярная вирусология»

к.б.н. Козлова Ольга Сергеевна

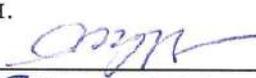
МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

630090, Новосибирская область, г. Новосибирск, ул. Пирогова, д. 1.

УТВЕРЖДАЮ

Зам. зав. кафедрой информационной
биологии ФЕН НГУ,
д.б.н.


" 27 " ИЮНЬ 2023г.



АКТ

о внедрении результатов диссертационной работы

Колпакова Федора Анатольевича на тему

**"Компьютерное моделирование биологических систем и
анализ биомедицинских данных"**

Программный комплекс BioUML, разработанный в рамках диссертационной работы Колпакова Федора Анатольевича, внедрен и активно используются в образовательном процессе в рамках курса "Математические основы системной биологии: моделирование молекулярно-генетических систем" для магистрантов кафедры информационной биологии ФЕН НГУ.

Ответственный за внедрение:



к.б.н., старший преподаватель КИБ ФЕН НГУ,

руководитель курса

Акбердин Илья Ринатович

АВТОНОМНАЯ НЕКОММЕРЧЕСКАЯ ОБРАЗОВАТЕЛЬНАЯ ОРГАНИЗАЦИЯ
 ВЫСШЕГО ОБРАЗОВАНИЯ «НАУЧНО-ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ
 «СИРИУС»

(АНО ВО «УНИВЕРСИТЕТ «СИРИУС»)

УТВЕРЖДАЮ

Проректор по научно-технологическому
 развитию - директор научного центра
 трансляционной медицины НТУ «СИРИУС»,
 канд. мед. наук



Р.А. Иванов

2023г.

о внедрении результатов диссертационной работы

Колпакова Федора Анатольевича на тему

"Компьютерное моделирование биологических систем и анализ биомедицинских данных"

Комиссия в составе к.ф.м.н. Ненашева Анатолия Сергеевича, к.б.н. Акбердина Ильи Ринатовича и к.б.н. Пинтуса Сергея Сергеевича рассмотрела вопрос об использовании результатов диссертационной работы Колпакова Федора Анатольевича и установила, что разработанные в рамках диссертационной работы ПК BioUML, платформа Sirius-web и база данных GTRD внедрены и активно используются в образовательном процессе и научной работе, а именно:

- ПК BioUML и платформа Sirius-web используется для выполнения научных проектов:
 - 1) научный проект СМВ-RND-2123 "Виртуальный пациент: оптимизация выбора лекарственной терапии на основе построения "цифрового двойника" пациента" направления "Вычислительная биология";
 - 2) грант РФФ 21-75-30020 "Исследование и моделирование метаболических и гомеостатических переменных при онкотранспорте наночастиц и доставке лекарственных средств".
- ПК BioUML стал основой для выполнения контракта № 43-03/21/44/152 от 13.12.2021 по теме "Исследование, обоснование и выбор программных решений для визуализации генетических данных и обеспечения инструментов для работы

с генетической информацией в «Национальной базе генетической информации», заключенного с Федеральным государственным бюджетным учреждением «Национальный исследовательский центр «Курчатовский институт».

- база данных GTRD используется для выполнения для работ по договору No 0.1.1.59-02/1216/22 от 27.10.2022 на выполнение работы в рамках Проекта: «Регуляторная транскриптомика для оценки потенциала модификации генома и изучения регенерации в животноводстве», заключенным с Федеральным государственным автономным образовательным учреждением высшего образования «Казанский (Приволжский) федеральный университет».
- ПК BioUML и платформа Sirius-web используется при ведении курса "Системная биология".
- Платформа Sirius-web используется для хранения и доступа к учебным курсам и результатов дипломных работ, выполненных в НТУ "Сириус".

Председатель комиссии:
исполнительный директор
Научного центра информационных технологий
и искусственного интеллекта
к.ф-м.н. Ненашев Анатолий Сергеевич



Члены комиссии:

с.н.с. направления "Вычислительная биология"
Научного центра информационных технологий
и искусственного интеллекта
к.б.н. Акбердин Илья Ринатович



с.н.с. направления "Вычислительная биология"
Научного центра информационных технологий
и искусственного интеллекта
к.б.н. Пинтус Сергей Сергеевич

