МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное учреждение науки
Институт общей генетики им. Н.И. Вавилова
Российской академии наук
(ИОГен РАН)

ул. Губкина, д. 3, г. Москва, ГСП-1, 119991
Тел.: (499) 135-62-13, (499) 135-20-41
Факс: (499) 132-89-62

E-mail: iogen@vigg.ru
http: www.vigg.ru

УТВЕРЖДАЮ
Директор
ФГУН Институт общей генетики им. Н.И. Вавилова РАН
член-корреспондент РАН, д.б.н. А.М. Кудрявцев

19.01.2024

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИЙ

на диссертационную работу **Колпакова Федора Анатольевича**
«Компьютерное моделирование биологических систем и анализ биомедицинских данных»,
представленную на соискание ученой степени доктора биологических наук по специальности
1.5.8 – «Математическая биология, биоинформатика»

Актуальность исследования. Диссертационная работа Колпакова Федора Анатольевича посвящена актуальным вопросам и задачам компьютерного моделирования биологических систем и анализа биомедицинских данных.

На данный момент, в системной биологии отсутствует формализованный процесс и технология для построения больших компьютерных моделей сложных биологических систем, таких как модель целой клетки с учётом сигналов, поступающих от её окружения или модель процессов, происходящих в целом организме. Также отсутствуют программные комплексы, объединяющие процесс построения моделей с анализом омиксных данных, как это предполагает системно-биологический подход и задачи интерпретации индивидуальных геномных данных для целей персонализированной медицины.

Таким образом, решаемая в диссертационной работе задача разработки технологии и инструментария сложных модульных моделей биологических систем в интеграции с анализом омиксных данных является крайне востребованной и актуальной.

Научная новизна исследования. В диссертационной работе Колпакова Ф.А. разработана оригинальная технология моделирования для итерационного создания, тестирования и использования сложных модульных моделей биологических систем. Технология реализована в виде специализированного программного комплекса (ПК) BioUML, реализующего весь разработанный инструментарий, необходимый для успешного использования предложенной технологии. Построены новые модульные модели сложных биологических систем: процессы, происходящие на клеточном уровне (метаболизм, транскрипция, отдельные пути передачи сигнала, апоптоз); регуляцию артериального давления у человека; лекарственную терапию артериальной гипертензии; эпидемиологию COVID-19.

При построении модели 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 позволило более точно описать динамику инфицирования и уменьшить количество параметров, определяемых в ходе решения обратной задачи, что делает эти параметры более идентифицируемыми.

Большим достижением группы явилось создание новых технологий для интеграции геномных данных, в частности интеграции данных ChIP-seq экспериментов и достоверного выявления соответствующих районов связывания транскрипционных факторов. Эти технологии были использованы при создании уникальной базы данных GTRD, описанной в диссертации.

Колпаковым Ф.А. была разработана новая технология построения цифрового двойника пациента и показана ее применимость на примере оптимизации выбора лекарственной терапии для лечения артериальной гипертензии. При этом автором предложен новый подход - генерация популяции виртуальных пациентов для одного реального пациента - для решения проблемы неидентифицируемости части параметров модели.

Теоретическая и практическая значимость диссертационной работы.

В диссертационной работе Колпакова Ф.А. разработан ряд новых методов и технологий:

- технология для итерационного создания, тестирования и использования модульных моделей биологических систем;
- технология объединения данных из различных NGS экспериментов по регуляции транскрипции в виде кластеров и мета-кластеров;
- технология построения цифрового двойника пациента.

Разработанный под руководством автора ПК BioUML можно рассматривать как полнофункциональную отечественную биоинформатическую платформу, имеющую мировое значение. Платформа ПК BioUML явилась критическим компонентом для успеха ряда отечественных и международных проектов. На основе ПК BioUML созданы коммерческие программы для визуализации и анализа геномных/омиксных данных geneXplain и Genome

Enhancer, а также качественно новые модели лекарственной терапии артериального давления с учётом сердечно-сосудистой регуляции кровяного давления

Разработанная технология создания модульных моделей биологических систем и реализующий ее ПК BioUML позволили перейти на более высокий уровень сложности создания таких моделей, а также упростили и ускорили, с точки зрения создателей моделей, этот процесс.

ПК BioUML используется как основной инструмент в курсах по системной биологии, проводимых в Новосибирском государственном университете и научно-технологическом университете "Сириус".

Под руководством Колпакова Ф.А. создана новая база данных GTRD - Gene Transcription Regulation Database, которая в настоящее время является крупнейшей в России базой данных по генетической информации (общий объем, включая исходные NGS данные, составляет 600+ ТБ) и одной из крупнейших в мире по регуляторной геномике (по количеству ChIP-seq и ChIP-exo экспериментов). Использование оригинальных идей и подходов для интеграции данных ChIP-seq экспериментов и достоверного выявления соответствующих районов связывания транскрипционных факторов стало важным фактором, положительно отличающим GTRD от других аналогичных баз даных. База данных GTRD является высоко востребованной, широко используемой и цитируемой - более 400 цитирований, 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, – коллекция мотивов для сайтов связывания транскрипционных факторов человека и мыши;
- ADASTRA – коллекция данных по аллель-специфичному связыванию факторов транскрипции в геноме человека;
- ANANASTRA – веб-сервер для аннотации влияния SNP на аллель-специфичное связывание факторов транскрипции в геноме человека.

Также информация из базы данных GTRD была использована при создании международных веб-ресурсов: BaMM motif – библиотека мотивов для распознавания сайтов связывания транскрипционных факторов; mSigDB – Molecular Signatures DataBase, раздел по регуляции генов.

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», что подтверждается 7 актами внедрения, приведенными в приложении диссертационной работы.

Структура и содержание диссертационной работы. Диссертационная работа состоит из введения, 8 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.

Во введении обосновывается актуальность выбора темы исследований, приведены цели, задачи и положения, выносимые на защиту, а также приведены научная новизна, теоретическая и практическая значимость, апробация работы и личный вклад автора.

Глава 1 содержит обзор литературы. В ней рассмотрено понятие системной биологии и соответствующие экспериментальные методы. Отдельное внимание уделено воспроизводимости моделей биологических систем и используемым для этого международным стандартам. Также рассмотрены основные методы и подходы, а также программные комплексы для моделирования биологических систем. Отдельные подразделы посвящены методам и сценариям анализа генетических данных, а также их графического представления. В отдельном разделе приведен обзор предшествующих работ автора, выполненных в 1994–2000 г.г. в лаборатории теоретической генетики Института Цитологии и Генетики СО РАН под руководством академика РАН, профессора, д.б.н. Колчанова Н.А., которые послужили заделом для разработки BioUML.

Глава 2 посвящена описанию архитектуры, интерфейса пользователя и технологий использованных при создании ПК BioUML. Часть информации вынесена в отдельные приложения (этапы развития, основные модули, Java библиотека для доступа и поиска информации в базах данных, Java библиотека графических объектов).

Глава 3 посвящена технологии создания и использования модульных моделей биологических систем. Для этого автором разработана специальная мета-модели для комплексного описания, графического представления и численного моделирования широкого круга биологических систем. Также описаны использование визуального, модульного, агентного и численного моделирования, текстовое представление модели и его синхронизация с графическим представлением, автоматическая генерация кода, определение параметров моделей. Завершает главу описание создание математических моделей сложных биологических систем эволюционным путем. Часть информации также вынесена в отдельные приложения.

Глава 4 посвящена описанию возможностей ПК BioUML для анализа и графического представления биомедицинских данных. В ней описан интерфейс пользователя для поиска и запуска как отдельных методов анализа омиксных данных, так и комплексных сценариев. В ПК BioUML реализована возможность описания и запуска сценариев на языках CWL, WDL и Nexftlow. Для графического представления геномных данных разработан новый геномный браузер, который по функционалу сопоставим с ведущими геномными браузерами, а также имеет ряд уникальных возможностей.

Глава 5 посвящена описанию базы данных GTRD. Данная глава начинается с постановки задачи, далее подробно описаны использованные методы и подходы для решения этих задач. В следующем разделе описано использование базы данных GTRD для построения цистрома для человека и мыши, а также другие базы данных и веб-ресурсы, созданные на основе информации из базы данных GTRD (см. раздел "Теоретическая и практическая значимость диссертационной работы"). В заключении главы приведено сравнение GTRD с другими базами данных по ChIP-seq экспериментам, а также обсуждается вопрос полноты покрытия транскрипционных факторов и их сайтов связывания.

Глава 6 посвящена моделирование сложных биологических систем. В первом разделе детально рассмотрены основные этапы предложенной технологии на примере создания модульной модели апоптоза. Следующие разделы посвящены моделям регуляции генной экспрессии при физической нагрузке в скелетных мышцах, эпидемиологии COVID-19, регуляции артериального давления у человека.

Глава 7 посвящена описаниею технологии построения цифрового двойника пациента. В начале главы описана концепция цифрового двойника пациента. Далее идет описание генерации виртуальных популяций и моделирования лекарственной терапии артериальной гипертензии и ее валидации на основе клинических данных. Следующий раздел посвящен персонализации параметров модели виртуального пациента, т.е. настройке параметров модели под конкретного пациента. Далее описан интерфейс пользователя и примеры отчетов для прогнозирования результатов антигипертензивной терапии. Приведен пример валидации прогноза результата лечения артериальной гипертонии на основе реальных клинических данных. В обсуждении приведены направления дальнейшего развития предложенного подхода.

Глава 8 содержит краткое описание и примеры интерфейса пользователя других программных продуктов, созданных на основе ПК BioUML: компоненты прототипа «Национальной базы генетической информации"; биоинформатическая платформа для хранения, анализа и графического представления данных, полученных при одномолекулярном секвенировании ДНК; платформа Sirius-web для проектов, связанных с анализом данных и моделированием для образовательных и научных проектов; платформы geneXplain и Genome Enhancer для анализа омиксных данных.

В заключении приведены результаты независимого сравнения ПК BioUML с другими программами, поддерживающими стандарт SBML, для моделирования биологических систем: 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.

Также обсуждается применение созданной технологии, ПК BioUML и платформы Sirius-web для создания информационной среды, в которой модели могут эволюционировать подобно живым существам.

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.

Вопросы и замечания. Работа имеет огромный объем (около 400 страниц) и содержит очень разнородную информацию, включая разработку программного комплекса и биологические приложения. Собственно, одним из главных недостатков работы является попытка диссертанта включить в состав диссертации весь имеющий отношение к работе материал. В результате в работу включены огромные приложения, включающие в себя списки программных модулей на языке Java, созданные при разработке BioUML, подробную таблицу работ по моделированию клеточных и физиологических процессов и другие материалы. При всем уважении к огромной работе, проделанной диссертантом, списки программных модулей и их блок-схемы являются технической информацией и обычно информация такого рода сообщается через ссылки на репозитории. Кроме того мне вообще не очень ясна логика деления материала на основной текст и приложения: результаты некоторых исследований, например результаты численного моделирования эпидемиология COVID изложены в приложениях, включая сравнение с экспериментальными данными, что на мой взгляд является основным научным результатом. В то же время разные технические подробности создания моделей в большом объеме включены в основной текст. Все эти недостатки относятся к структуре изложения, и не сказываются на ценности проведенных под руководством диссертанта научных исследований и разработок.

Работа в целом хорошо оформлена, аккуратно написана хорошим языком. Следует отметить, что автор уделил большое внимание обозначениям: несмотря на огромный объем материала обозначения согласованы в разных разделах, понятны, и последовательно изложены, имеются таблицы обозначений.

Заключение о соответствии диссертационной работы предъявляемым требованиям. Диссертационная работа Колпакова Ф.А. является оригинальным исследованием, содержит ряд новых научных результатов, как теоретических, так и практических, и по сути открывает ряд новых направлений в вычислительной биологии, как в области анализа омиксных данных, так и в области моделирования медицинских показателей при развитии хронических заболеваний.

Научные положения, выносимые на защиту, и выводы, сформулированные в диссертации, полностью обоснованы, их достоверность и новизна не вызывает сомнений.

Автореферат диссертации полно и достоверно отражает сущность работы.

Диссертация Колпакова Федора Анатольевича «Компьютерное моделирование биологических систем и анализ биомедицинских данных», представленная на соискание ученой степени доктора биологических наук по специальности 1.5.8 – математическая биология, биоинформатика, является законченной научно-квалификационной работой, полностью

соответствующей требованиям, предъявляемым к диссертациям на соискание ученой степени доктора наук согласно п. 9-14 «Положение о присуждении ученых степеней», утвержденного Постановлением Правительства РФ от 24.09.2013 № 842, а её автор, Колпаков Федор Анатольевич, достоин присуждения искомой ученой степени доктора биологических наук по специальности 1.5.8 — математическая биология, биоинформатика.

Отзыв на диссертационную работу обсужден и утвержден на семинаре лаборатории системной биологии и вычислительной генетики Института общей генетики РАН им. Н.И. Вавилова, состоявшемся 7.11.2023.

Заведующий лабораторией функциональной геномики
Федерального государственного бюджетного учреждения науки Институт общей генетики им. Н.И. Вавилова Российской академии наук,
кандидат биологических наук по специальности
03.01.03 – молекулярная биология, доцент С.А. Брускин
19.01.2024

Подпись Брускина С.А. заверяю
Ученый секретарь Федерального государственного бюджетного учреждения науки Институт общей генетики им. Н.И. Вавилова Российской академии наук
д.б.н. И.И. Горячева

Адрес Федерального государственного бюджетного учреждения науки Институт общей генетики им. Н.И. Вавилова Российской академии наук: 119991, Москва, ул. Губкина, д. 3, (499) 135-62-13, iogen@vigg.ru, brouskin@vigg.ru