ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

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

Диссертационная работа Ф.А. Колпакова посвящена созданию технологии и программного инструментария для эффективного построения и использования сложных модульных моделей в биологии, главным образом в системной биологии, предметом изучения которой являются сложные взаимодействия и процессы в живых системах, происходящие одновременно на разных иерархических уровнях организации – от молекулярного, геномного и клеточного до органов, систем и взаимодействий между ними. Впрочем, созданный Ф.А. Колпаковым программный комплекс может быть также применён для решения задач мультимасштабного моделирования, выходящими за пределы одного организма, позволяя описать и рассчитать наиболее важные взаимодействия между вирусом, организмом и его иммунной системой, а также иммунно-эпидемиологические процессы, происходящие на уровне популяции, что было успешно продемонстрировано им на примере COVID-19.

Актуальность. Прогресс в области технологий секвенирования генетических последовательностей сопровождается уменьшением стоимости процесса, увеличением его скорости и повышением качества результата. Впервые определение последовательностей нуклеотидов, образующих геном человека (как мы теперь знаем – около 3 миллиардов пар азотистых оснований), было начато в 1990 г. и заняло 13 лет, завершившись в 2003 г. Второй геном человека был прочитан в 2007 г., когда была изобретена технология секвенирования нового поколения NGS (next generation sequencing) – это был геном Нобелевского лауреата Джеймса Уотсона, первооткрывателя двойной спирали ДНК. Стоимость составила около 1 миллиона долларов, а затраты по времени – около двух месяцев. Вслед за первым геномом, который был размещён в интернете и стал общедоступным, Уотсон сделал со своим то же самое со словами «*Я размещаю свой геном в интернете, чтобы дать толчок развитию новой эпохи персонализированной медицины. В этой новой эпохе информация, содержащаяся в геноме, будет помогать идентифицировать и предотвращать болезни, а также позволит создать персонализированные медицинские методы лечения*». В наши дни (по данным 2023 г.) стоимость секвенирования генома благодаря развитию технологий снизилась примерно до 600 долларов (!), время – до 5 часов (!), а число индивидуальных секвенированных геномов в мире превысило полмиллиона (и все они различны).

Однако, прочтение генома – это лишь начало пути. Знать геном организма или даже знать положение каждого атома в живой клетке – далеко не то же самое, что иметь функционирующую модель этой клетки. Геном человека включает около 20 000 генов, а

значит – не меньшее количество различных белков (выполняющих широкий спектр различных функций в живых клетках), которые могут взаимодействовать между собой и с множеством других более низкомолекулярных соединений в организме человека, образуя сложные сети взаимосвязей, которые могут по-разному функционировать в норме и при различных патологиях. Сложные взаимодействия в живых системах являются предметом изучения системной биологии, получившей активное развитие с началом XXI века. Одним из ключевых инструментов, используемых в этой междисциплинарной науке, является компьютерное моделирование, поскольку при таком количестве взаимодействующих элементов, такой детализации и таких объемах информации, которая может быть получена из современных экспериментов (и использоваться для построения моделей), для многих задач лишь высокопроизводительные вычисления и автоматизация могут скомпенсировать рост сложности исследуемых систем, с расчетом которых вручную человек уже не способен эффективно справляться.

Именно в эти области науки, которые шаг за шагом меняют жизнь человечества к лучшему, развивая биологию, медицину и сопутствующие им технологии, Федор Анатольевич внёс весьма значительный и многогранный вклад, одной из важнейших составляющих которого представляется создание отечественной биоинформатической программной платформы BioUML (https://www.biouml.org, Biological Universal Modeling Language, универсальный язык для моделирования в биологии) с открытым исходным кодом, широко используемой в настоящее время для решения реальных практических научных задач, а также в образовательных целях при подготовке специалистов в соответствующих областях.

Структура диссертации

Диссертационная работа состоит из введения, 8 глав, заключения, списка литературы и приложений. Она включает 291 страницу основного текста, имеет общий объем 395 страниц, содержит 120 рисунков, 20 таблиц и 438 ссылок в списке литературы.

Обзор литературы (**Глава 1**) написан одновременно увлекательно и с достаточно глубоким погружением в проблематику системной биологии, от её истоков до нынешнего времени, показывая сильные и слабые стороны имеющихся подходов и концепций, уверенно аргументируя и подводя к необходимости создания тех программных инструментов в области биоинформатики и системной биологии, разработке и применению которых посвящена значительная часть диссертационной работы.

Главы 2-5 посвящены проблемам создания программных средств для проектирования и реализации модульных моделей сложных биологических систем, визуализации больших данных в области геномики и транскриптомики, а также графическому представлению и анализу результатов численных экспериментов. Среди основных результатов – создание программного комплекса BioUML, разрабатываемого с 2001 г. по настоящее время и представляющего собой отечественную биоинформатическую программную платформу, широко известную и используемую в России и в мире. Масштаб разработки таков, что подробно описать все программы и сценарии, входящие в ПК BioUML, не представляется

возможным в рамках диссертационной работы, поскольку одна лишь документация по методам и сценариям анализа биомедицинских данных для платформы geneXplain (основанной на архитектуре ПК BioUML) составляет более 500 страниц.

В **Главах 6-8** представлены результаты, полученные с использованием созданного программного комплекса – построены модели актуальных сложных биологических систем, описаны численные эксперименты, произведенные на их основе, а также приведены визуализация, анализ и интерпретация полученных результатов, в том числе в сравнении достижениями предшественников. Так, например, комплексная модель регуляции артериального давления, представленная в **Главе 6**, в процессе опубликования соответствующей статьи была оценена одним из рецензентов следующим образом: «*Никто со времен Гайтона и его команды в Миссисипи (1972 г.) не сделал такого качественного скачка в моделировании регуляции и контроля кровяного давления на уровне всего организма*». Также в 6-й главе представлены модульные модели апоптоза (запрограммированной гибели клеток), регуляции генной экспрессии в скелетных мышцах при физической нагрузке, ряд моделей COVID-19 (три эпидемиологические модели, агентная модель, мультимасштабная эпидемиологическая модель и модель взаимодействия вируса с организмом и иммунной системой хозяина), а также модели регуляции артериального давления у человека. Для модели апоптоза в диссертации последовательно представлены первая, две промежуточных и четвертая финальная модель (включающая 13 модулей, 280 белков и их комплексов, 372 химических реакции и 459 параметров), на сегодняшний день являющаяся самой сложной имеющейся моделью данного процесса. **Глава 7** посвящена созданию, валидации и практическому использованию цифрового двойника пациента для помощи при подборе индивидуального лечения артериальной гипертензии на основе концепции персонализованной медицины. **Глава 8** содержит описание ряда разработок на основе программного комплекса BioUML, включая их назначения, возможности и области применения (включая платформы geneXplain, Genome Enhancer, Sirius-web, u-science, платформу для одномолекулярного секвенирования ДНК и др.). В **заключении** подводятся итоги работы.

Научная ценность и новизна

Оригинальная технология моделирования для итерационного создания, тестирования и использования сложных модульных моделей биологических систем, программный комплекс BioUML, реализующий весь инструментарий для успешного использования этой технологии, база данных регуляции транскрипции генов GTRD и геномный браузер, а также созданные на основе упомянутых разработок модульные модели сложных биологических систем, от наиболее полной модели апоптоза до модели эпидемиологического процесса COVID-19 и цифрового двойника пациента – все эти разработки основаны на множестве новых идей, подходов, моделей, архитектур и концепций, предложенных и реализованных автором. Даже краткое описание основных элементов новизны, приведенное в соответствующем разделе, занимает несколько страниц.

Теоретическая значимость диссертационной работы включает новые технологии создания, тестирования и использования модульных моделей биологических систем,

объединения данных из различных NGS экспериментов в виде кластеров, мета-кластеров и мастер-треков, построения цифровых двойников пациентов, а также собственно новые модульные модели, разработанные на основе этой технологии.

Практическая значимость, главным образом, заключается в создании программного комплекса BioUML, представляющего собой отечественную биоинформатическую программмную платформу, весьма востребованную и нашедшую широкое применение, в том числе в следующих проектах:

- "Исследование, обоснование и выбор программных решений для визуализации генетических данных и обеспечения инструментов для работы с генетической информацией в «Национальной базе генетической информации»" (2021),
- "Разработка и испытания биоинформатической платформы для хранения, анализа и графического представления данных, полученных при одномолекулярном секвенировании ДНК" (2022),
- Sirius-web (https://sirius-web.org), информационная платформы на основе ПК BioUML для проектов, связанных с анализом данных и моделированием для образовательных и научных проектов (активно используемая в Научно-технологическом университете «Сириус» для проведения исследований в области наук о жизни),
- Коммерческих программах для визуализации и анализа геномных/омиксных данных (платформы Genome Enhancer и geneXplain),
- При создании высоко востребованной и широко используемой в мире базы данных GTRD (Gene Transcription Regulation Database – база данных регуляции транскрипции генов).

В контексте практической значимости также стоит отметить, что 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 и Научно-технологический университет «Сириус».

Достоверность и обоснованность результатов

В диссертационной работе Ф.А. Колпакова вопросам достоверности, воспроизводимости, надежности и верификации результатов моделирования уделено должное внимание. Поскольку они актуальны как для науки в целом, так и для системной биологии в частности, в 2009 г. был создан и до сих пор поддерживается и развивается стандарт COMBINE (https://co.mbine.org) – COmputational Modeling in BIology NEtwork (компьютерное моделирование в биологических сетях), в разработке которого Ф.А. Колпаков принимал непосредственное участие. Этому и ряду других сопутствующих вопросов посвящен целый раздел первой главы диссертации, 1.1.3 – «Воспроизводимость и стандарты в системной биологии». В результате разработанный с учетом этих стандартов программный комплекс BioUML был охарактеризован следующим образом: «*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», и при этом он «*является самым качественным и быстрым SBML симулятором в мире*» (SBML – Systems Biology Markup Language, XML формат для представления моделей в области системной биологии). Это весьма достойный результат и уровень, которого стоило бы придерживаться во многих науках.

Замечания по диссертации и автореферату

- В тексте автореферата и диссертации имеется некоторое количество орфографических ошибок и опечаток, в основном в окончаниях слов (например, "международный семинаре «From virtual cell to virtual human and virtual patient» (Новосибирск, 2012); международной конференции по биомедицинской инженерии и компьютерных технологиях SIBIRCON»" в разделе «Апробация работы» и т.п.) – типичные издержки редактирования текста, встречающиеся практически в любом тексте большого объема. Одна опечатка закралась в название раздела – "6.2.2 Регуляциея экспрессии генов …" (с. 157 диссертации).
- В тексте автореферата (с. 26) упоминается Рис. 5.5.1, которого нет ни в автореферате, ни в диссертации. По-видимому, имеется в виду рис. 5.1 в автореферате.
- Не совсем верное название раздела «Свидетельства на регистрацию программ для ЭВМ» в автореферате на стр. 40 (всё-таки правильно «Свидетельства о регистрации программ для ЭВМ»)
- Низкое качество рисунка 6.2.3 – по-видимому, из-за артефактов сжатия изображения. Мелкие символы в формулах едва читаются (с. 157 диссертации).

Также при прочтении возник следующий вопрос. В автореферате на с. 16 сказано, что объем программного кода составляет "Java – 6 199 файлов, объем 36.8 Мб, ~ 1.2 млн. строк кода, JavaScript – 74 файла, 1.5 Мб, ~ 50 000 строк кода", а на с. 20 имеется абзац об автоматической генерации кода. Хотелось бы уточнить: 1.2 млн. строк – это с учетом автоматической генерации? Каково соотношение между объемами кода, написанного вручную и сгенерированного автоматически?

Указанные замечания незначительны, не снижают высокой оценки уровня диссертационной работы и не влияют на понимание её сути, результатов и научной ценности.

Заключение. Диссертационная работа написана подробно, в хорошем стиле. Её основные положения опубликованы в ведущих международных и отечественных научных журналах (включая Nature Communications и Nature Biotechnology), число которых в несколько раз превышает достаточное для докторской диссертации, а количество их цитирований превышает 400 (включая цитирования в журналах Nature и Science). Результаты доложены соискателем на множестве известных международных и российских научных конференций, симпозиумов и семинаров. Автореферат адекватно отражает основные результаты диссертации.

Личный вклад соискателя, подробно описанный в соответствующем разделе автореферата и диссертации, включает как полностью самостоятельно созданные концепции, технологии, модели, алгоритмы, программные средства и научные результаты, так и результаты, полученные при совместной работе с коллегами или под

руководством Федора Анатольевича. Личный вклад в совместных работах носит основополагающий, принципиальный характер. Содержание диссертации и основные положения, выносимые на защиту, отражают персональный вклад автора в опубликованные работы.

Диссертационная работа Федора Анатольевича Колпакова является завершенным научным исследованием, содержащим как фундаментальные, так и практические результаты в таких областях, как молекулярная биология, генетика, биоинформатика, эпидемиология, иммунология и ряде других, довольно тесно связанных между собой на уровне взаимодействия функциональных систем организма (объединяемые системной биологией), поэтому их можно квалифицировать как совокупность научных достижений. По моему мнению, данная работа, без преувеличения, является выдающейся и представляет собой заметное событие в отечественной и мировой науке. На основании вышеизложенного считаю, что диссертация Ф.А. Колпакова полностью удовлетворяет всем требованиям ВАК, предъявляемым к докторским диссертациям (отвечает п. 9 Положения о присуждении ученых степеней, утвержденного Постановлением Правительства РФ от 24 сентября 2013 г. № 842), а её автор заслуживает присуждения ученой степени доктора биологических наук по специальности 1.5.8 – «математическая биология, биоинформатика».

Официальный оппонент,
и.о. директора ИСИ СО РАН,
доктор физико-математических наук
«15» января 2024 г. Пальянов Андрей Юрьевич

Подпись А.Ю. Пальянова заверяю
Ученый секретарь ИСИ СО РАН
Кандидат физико-математических наук Насибулов Егор Андреевич

Пальянов Андрей Юрьевич
Доктор физико-математических наук, специальность 05.13.18 –
математическое моделирование, численные методы и комплексы программ
И.о. директора Федерального государственного бюджетного учреждения науки Института систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук (ИСИ СО РАН)
Рабочий телефон: +7(383)330-86-52
Электронная почта: palyanov@iis.nsk.su

Адрес: 630090, Россия, г. Новосибирск, проспект Академика Лаврентьева, 6
Веб-сайт: https://www.iis.nsk.su
Телефон: +7(383)330-86-52
Факс: +7(383)332-34-94