# О Т З Ы В

официального оппонента на диссертационную работу Базыкина Георгия Александровича «Положительный и эпистатический отбор в эволюции аминокислотных последовательностей», представленную на соискание ученой степени доктора биологических наук по специальности 03.01.09 – «Математическая биология, биоинформатика».

Эволюционная биоинформатика – одна из немногих областей современной молекулярной и теоретической биологии, способных развиваться с привлечением минимального репертуара экспериментальных данных. В то время как регуляторная и функциональная геномика в своем развитии задействуют все более широкий спектр экспериментов, эволюционные исследования все еще продвигаются вперед, базируясь на сравнительно простой информации о геномных последовательностях и аннотациях генов. В свою очередь, эта информация становится все более доступной благодаря удешевлению высокопроизводительного секвенирования, и спектр данных о геномах как в смысле прочтения геномов новых видов, так и в смысле индивидуальных геномов и популяционных полиморфизмов, продолжает активно расширяться. Представленная Г.А. Базыкиным работа в полной мере эксплуатирует растущее разнообразие геномных данных, и представляет множество новых подходов для исследования отбора, действующего на участки белок-кодирующих генов, и анализа динамики этого отбора в ходе эволюции. Таким образом, методы и исследования, описанные в диссертации, являются **безусловно актуальными**, соответствуют современным и активно развивающимся направлениям биоинформатики, а **востребованность предложенных методов и полученных результатов** не вызывает сомнений. Однопозиционный адаптивный ландшафт (в терминологии автора - ОПАЛ) как профиль приспособленности аминокислот в конкретном аминокислотном сайте, является центральным объектом исследования, и является связующим звеном для разнообразных задач, поставленных и решенных автором. **Широта охвата проблемы** (от принципиальных фундаментальных вопросов эволюции к конкретным вопросам регуляции трансляции и важнейшим практическим задачам по изучению вирусной адаптации, от высших эукариот до прокариот и вирусов) **и фундаментальный характер объекта исследования и постановка цели исследования** безусловно определяют сильные положительные стороны диссертации.

Результат работы диссертанта представляет собой многоплановый материал, который систематизирован и изложен в сжатой форме. Хочется особенно отметить степень разработанности темы исследования, которая

убедительно продемонстрирована в отдельном разделе. При чтении диссертации детально прослеживается **внутренняя логика исследования**, диссертант продемонстрировал глубокое понимание взаимосвязи различных процессов, отражающихся в эволюции конкретных аминокислотных сайтов, и корректно применил методы математической статистики, что позволило подкрепить принципиальные утверждения достоверными оценками. Материалы диссертации широко представлены научному сообществу на различных профильных конференциях, хорошо отражены в автореферате, опубликованы в более чем 20 печатных работах, включая чрезвычайно престижные международные журналы, в том числе *PLoS Genetics*, *Genome Research* и *Science*.

**Научная новизна** исследования обусловлена разработкой и практическим использованием принципиально новых подходов для изучения различных не рассматривавшихся ранее аспектов эпистатического отбора. **Практическая и теоретическая значимость** работы несомненны, что подтверждается уровнем публикаций по теме диссертации и активным их цитированием в рамках исследовательского сообщества, судя по данным систем Scopus и Web-of-Science.

Диссертационная работа Г.А. Базыкина с точки зрения оформления и подачи материала в целом соответствует требованиям, предъявляемым к докторским диссертациям. Представленная работа изложена на 199 страницах и включает в себя введение (где, в том числе, излагается актуальность и новизна темы работы, цели и задачи исследования), обзор литературы, 4 главы основного текста, заключение, список публикаций по теме диссертации, список литературных источников из 429 наименований и список сокращений. Работа включает 29 рисунков и 20 таблиц, что предоставляет хороший иллюстративный материал и облегчает восприятие текста.

Некоторые моменты в структуре работы вызывают удивление, например, обзор литературы (Глава 1) изложен на 10 страницах и посвящен исключительно общим вопросам изучения адаптивного ландшафт одиночных аминокислотных сайтов. Скромный объем литературного обзора, на первый взгляд, не согласуется с масштабом исследования. Однако, это объяснимо, отчасти, подробным разделом «Степень разработанности темы исследования» и независимым введением – литературным обзором к каждой главе диссертации. Такое построение работы не соответствует стандартной схеме изложения материала, но отчасти помогает сфокусироваться на литературе, релевантной каждой конкретной задаче. Объем и адекватность ссылок на литературные источники позволяет уверенно утверждать, что автор действительно хорошо ориентируется в современном состоянии исследований по теме работы. Что интересно, выделенный раздел «Материалы и методы» в работе также отсутствует, и

методические вопросы равномерно распределены по основным главам диссертации. Это отчасти удобно при прочтении материалов каждой главы, но ряд моментов (источники геномных данных и аннотаций, статистические тесты итд.) вынужденно повторяется по тексту диссертации. Тем не менее, методы выглядят адекватно поставленным задачам, а публикация основных материалов в открытом доступе в сети Интернет обеспечивает **воспроизводимость** и доступность материалов биоинформатического исследования для научного сообщества. То касается содержания основного текста работы, глава 2 посвящена изучению адаптивных ландшафтов одиночных аминокислотных сайтов, глава 3 - эпистатическимо адаптивному ландшафту для пар таких сайтов, а глава 4 специфически сфокусирована на анализе концов генов (альтернативных сайтов инициации трансляции у эукариот и эволюции 3' концов генов у прокариот).

Помимо важных методических достижений, хочется особенно отметить несколько **принципиально важных и интересных результатов** и соответствующих им выводов**, определяющих значимость работы для научного сообщества**: (1) Предложена концепция однопозиционных адаптивных ландшафтов как векторов приспособленности аминокислот в определенной позиции белка; (2) Предложен и апробирован на линии *D. melanogaster* подход к оценке доли аминокислотных замен, происходивших при поддержке положительного естественного отбора, путем сопоставления параметров внутривидовой изменчивости в важных и селективно-нейтральных сайтах, в которых недавно произошла замена того же вида; (3) Продемонстрировано, что в белок-кодирующих генах у *Drosophila* доля аллельных замещений, вызванных положительным отбором, как и сила отбора максимальны в консервативных участках; показано, что в эволюции насекомых и позвоночных отбор против восстановления предковой аминокислоты после аминокислотной замены усиливается с течением времени. **Безусловным украшением и прямым практическим результатом работы** является выявление эпистатических взаимодействий между мутациями в разных генах, т.е. межгенного эпистаза для генома вируса гриппа, с учетом возможных реассортаций. Это особенно производит впечатление, в связи с типичным восприятием эволюционной биоинформатики как чисто теоретической науки с ограниченным практическим компонентом. Результаты, полученные автором, должны стать базой для построения принципиально новых предиктивных моделей, открывающих путь к рациональному прогнозированию в дизайне вакцин. Наконец, с точки зрения регуляторной геномики высших эукариот **большой интерес** представляет изученная эволюция альтернативных сайтов инициации трансляции.

Тем не менее, не смотря на **высокий уровень представленной диссертации и прекрасный стиль изложения материала**, ряд моментов в тексте

работы хочется прокомментировать и прояснить (ниже наиболее важные моменты выделены подчеркиванием).

Введение, стр. 3: «В главе 1, которая в основном представляет собой обзор литературы…» (можно предположить, что глава 1 и есть обзор литературы?).

Введение, стр. 4: «Ранее мы показали…» (в исследованиях, предшествовавших диссертационному?).

Актуальность темы исследования, стр. 6: «Исследования покрывают широкий спектр биологических систем…» (неясно, какое отношение это имеет к актуальности темы).

Степень разработанности темы исследования, стр. 7-8: используется масса специальных терминов, которые не введены, глоссарий в диссертации также не присутствует: «генетический контекст», «синергический (сужающий) эпистаз», «генетический груз», «параллельная, ревертивная, конвергентная эволюция».

Цели и задачи исследования, стр. 12: «…вируса гриппа H3N2 после его появления в популяции человека в 1968 году» (неясно, почему конкретный год столь важен, что вынесен в задачи исследования).

Цели и задачи исследования, стр. 12: задача 8 «…находятся ли альтернативные старт-кодоны под действием отрицательного отбора…», здесь и в других задачах не указаны домены или конкретные организмы, но в задаче 9 явно говорится о прокариотах.

Научная новизна, стр. 19: «…исключение кодирующей последовательности часто приводит к образованию нижерасположенных стоп-кодонов в рамке». Обсуждается включение участков 3’ нетранслируемой области в состав кодирующей; в этом контексте неясно, что значит «исключение кодирующей последовательности».

Формальные характеристики работы, стр. 21: наблюдается несоответствие в числе публикаций по теме диссертации. Автор пишет, что материал опубликован в 40 работах, из которых 23 имеют прямое отношение к тексту диссертации. При этом в списке авторских публикаций (и в диссертации и в автореферате) присутствуют библиографические данные 29 работ.

Обзор литературы, стр. 23: «…литература по (несколько более простым и понятным) поверхностям приспособленности нуклеиновых кислот». Утверждение не подкреплено ссылками и выглядит достаточно спорным, учитывая сложную структуру некодирующих последовательностей, например длинных некодирующих РНК или энхансеров в геномах высших эукариот.

Глава 2, стр. 39: «…сравнения с единственной референтной последовательностью может приводить к ошибкам…когда референтная последовательность несет низкочастотный аллель». Неясно, почему проблема не может быть решена перестройкой референтной последовательности с заменой редких аллелей на частые по имеющимся популяционным данным.

Глава 2, стр. 55: «…мы картировали на выравнивание 13300 генов *D. melanogaster*…». Неясно, как именно (какой программой или какой процедурой) гены были картированы на выравнивание. Здесь же и далее (например, стр. 61) часто упоминаются «канонические варианты сплайсинга». Учитывая ткань-специфичность альтернативного сплайсинга, неясно какие именно изоформы транскриптов выбирались и какая именно изоформа считалась каноничной.

Глава 4, стр. 122: «использование альтернативных старт-сайтов … – это функциональный механизм, находящийся под отбором, направленным на увеличение эффективности трансляции …» (неясно, почему альтернативные сайты инициации трансляции должны приводить к увеличению эффективности трансляции в целом; этот момент затем недостаточно прояснен в тексте).

Глава 4, стр. 133: «…сбалансированной трансляции двух изоформ…». Требует пояснения, почему механизм регулируемого переключения изоформ рассматривался как менее вероятный (в частности, это могло бы объяснить малую силу эффекта, обсуждаемую в следующем абзаце).

Глава 4, стр. 135-136, табл. 14-16: неясно какие P-значения приведены для обогащений терминами генной онтологии (исходные или с поправкой на множественное тестирование). Более важный момент: в таблице 16 присутствует целая группа терминов онтологии, ассоциированных с гомеодоменами, неясно речь идет о соответствующих участках связывания в окрестности генов или о функциональной роли самих генов. Этот момент кажется важным, но никак не обсуждается в тексте.

Глава 4, стр. 139 и др.: автор обсуждает идею о том, что общепринятый взгляд предполагает наличие единственного старт-кодона в транскрипте. Эта точка зрения устарела, и изучение альтернативной инициации трансляции сегодня является одной из важных отраслей регуляторной геномики. По всей видимости, текст был напрямую взят из работы, сделанной и опубликованной авторами в 2011 году. При этом классические работы по рибосомному профилированию (Ribo-Seq) были опубликованы еще в 2009, и сегодня прямые экспериментальные данные о локализации рибосом (в том числе, на альтернативных сайтах инициации) доступны для транскриптомов большинства модельных организмов. В частности, известно, что важную регуляторную роль играют апстрим-рамки

считывания, локализованные в 5’ нетранслируемой области транскриптов. Хочется прояснить, не могли ли эффекты от наличия таких рамок внести существенных искажений в анализ альтернативных стартов для основных рамок считывания, проведенный авторами.

При всем этом, число опечаток и стилистических погрешностей в тексте работы достаточно невелико, но полностью избежать их не удалось.

Глава 2, стр. 40: «…распределени**я** частот предковых аллелей также сходн**о**…» (опечатка).

Глава 2, стр. 44: «…получены с биоинформатической платформы Galaxy…» (некорректно, нужно указывать конкретную публичную инсталляцию сервиса Galaxy, т.к. различные сервисы обладают принципиально разными инструментами и версиями баз данных); «взяты из Генбанка» (жаргон, корректнее писать, что данные получены из базы данных GenBank).

Глава 2, стр. 25: «Среди полиморфных сайтов с предковым вариантом … сравнивалось с использованием У-теста Мэнна-Уитни распределение частот предковых … вариантов» (неудачный порядок слов).

Глава 2, стр. 46: «консерват**и**в**р**вных сегментах», «аллельных **щ**амещений» (опечатки).

Глава 2, стр. 47-48: фраза «Мы сравниваем долю замен…» продублирована с некоторым изменением формулировки.

Глава 2, стр. 51 и далее: термин «аутгруппа» используется в мужском роде (аутгрупп).

Глава 2, стр. 66: «приспособленность **аллей**» (опечатка).

Глава 3, стр. 75: «…и **ища** несовместимости между ними» (опечатка).

Глава 3, стр. 77: «…этот способ картирования был … **самоочевидным**…» (смысл неясен).

Глава 3, стр. 83: «…это **невозможно в грипп** А…» (опечатка или пропущенное слово).

Глава 3, стр. 87: «…эпистатические взаимодействия широко распространены в гриппе А.» (имеется в виду геном вируса).

Глава 3, стр. 93: «…интерес представляет наше открытие…» (слово «открытие» в указанном контексте читается достаточно нескромно).

Глава 3, стр. 105: «…эффект реассортации на наши результаты…» (пропущено слово).

Глава 3, стр. 106, табл. 9: в подписи к таблице указано, что обедненные категории обозначены курсивом, но курсив в таблице не используется.

Глава 3, стр. 114, «Мы скачали все полные изоляты гриппа А … из базы данных» (по всей видимости речь идет о геномах изолятов).

Глава 3, стр. 119: математические символы не пропечатаны в тексте.

Глава 4, стр. 122: «Гены … обогащены факторами транскрипции» (имеется в виду обогащение набора генов терминами генной онтологии?)

Глава 4, стр. 122 и далее: используется нестандартное сокращение НТП (не расшифровано, по всей видимости нетранслируемая последовательность), в то время как принятое сокращение – НТО, нетранслируемая область.

У множества рисунков не подписаны оси (например Рис. 19-20 и далее) и панели (например, Рис. 22-23), или подписаны загадочным образом (Рис. 4, ось Y: «Замен под положительным отбором»). Наименования программ в разных разделах написаны с отличающейся капитализаций (например GIRAF и GiRaF).

Некоторые технические подробности выглядят неуместными, например, на стр. 89 «…на узле кластера с 512 Гб оперативной памяти». Неясно, почему важно, что расчет проводился именно на узле кластера (а не на персональном компьютере), и почему объем памяти может как-то повлиять на результаты анализа.

Раздел 4.2. (стр. 144) резко переходит с эукариот к прокариотам, этот переход требует пояснений (почему альтернативная инициация изучалась именно у эукариот, а эволюция терминации именно у прокариот).

В целом, указанные замечания относятся к стилистике изложения и оформлению текста и не снижают чрезвычайно позитивного впечатления от уровня представленной работы, а вопросы носят исключительно уточняющий и дискуссионный характер.

Таким образом, диссертация Базыкина Георгия Александровича представляет собой полноценную **завершенную научно-исследовательскую работу**, описывающее **решение крупной научной проблемы**, имеющей серьезное **теоретическое и практическое значение**. Автореферат **полноценно и достоверно** отражает содержание диссертации, выводы работы **хорошо обоснованы и подкреплены фактическим материалом**. Таким образом, диссертационная работа Базыкина Георгия Александровича «Положительный и эпистатический отбор в

эволюции аминокислотных последовательностей» соответствует п.9 Положения «О порядке присуждения ученых степеней», утвержденного Постановлением Правительства Российской Федерации от 24 сентября 2013 г. №842, с изменениями Постановления Правительства Российской Федерации от 21 апреля 2016 года №335, а ее автор заслуживает присуждения искомой степени доктора биологических наук по специальности «03.01.09 - Математическая биология, биоинформатика».

Кулаковский Иван Владимирович
Ведущий научный сотрудник, к.ф.-м.н., д.б.н.,
Лаборатория вычислительных методов системной биологии,
Федеральное государственное бюджетное учреждение науки
Институт молекулярной биологии им. В.А. Энгельгардта
Российской академии наук (ИМБ РАН),
ГСП-1, 119991, г. Москва, ул. Вавилова, д. 32. ИМБ РАН
Тел.: +7 499 1356000, Факс: +7 499 1351405
Эл.почта: ivan.kulakovskiy@eimb.ru

_____ Кулаковский И.В.
«_5_» июня 2018 г.

Подпись Кулаковского И.В. заверяю:
ученый секретарь ИМБ РАН, к.в.н. Бочаров А.А. _____