ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

кандидата физико-математических наук Штокало Дмитрия Николаевича на диссертационную работу МУСТАФИНА ЗАХАРА СЕРГЕЕВИЧА на тему «РАЗРАБОТКА КОМПЛЕКСА ПРОГРАММ ДЛЯ АНАЛИЗА ЭВОЛЮЦИОННЫХ ХАРАКТЕРИСТИК ГЕННЫХ СЕТЕЙ», представленную к защите на соискание ученой степени кандидата биологических наук по специальности 03.01.09 – математическая биология, биоинформатика.

Актуальность. Изучение эволюции молекулярно-генетических систем и их элементов является одной из основных задач биоинформатики. Более глубокое понимание законов эволюции генов и механизмов их взаимодействия обладает большой фундаментальной и прикладной ценностью. Например, разработка таргетных препаратов нового поколения для лечения заболеваний опирается на анализ генных сетей, работу которых необходимо изменить с помощью терапевтического воздействия. Достаточно информативным для анализа отдельных генов и белков показал себя филостратиграфический анализ, предложенный ученым Томиславом Домазет-Лосо в 2007 г. Этот метод позволяет дать оценку времени возникновения гена на таксономическом дереве, найдя наиболее позднего общего «предка» заданного гена и его ортологов из других организмов. Несмотря на кажущуюся простоту идеи метода, актуальной проблемой оставалась разработка программных инструментов, позволяющих рутинное и аккуратное его применение в связке с другими методами, например при процессе анализа генных сетей. Филостратиграфический анализ относится к группе методов макроэволюционного анализа. Существуют также методы микроэволюционного анализа, позволяющие оценить скорость эволюции отдельно взятого гена и степень эволюционного давления на него. Разработка программ, позволяющих легко кооперировать методы микроэволюционного анализа с методами анализ молекулярно-генетических систем, также является актуальной практической задачей биоинформатики.

Диссертационная работа Мустафина Захара Сергеевича посвящена разработке программного инструментария, позволяющего удобное применение методов макро и микроэврролюционного анализа к изучению свойств генных сетей и решению с их помощью задач системной биологии. С помощью разработанного инструментария в диссертации проведено исследование генных сетей биосинтеза стероидов, биосинтеза стероидных гормонов, 80 генных сетей болезней человека, генов *Arabidopsis thaliana,* ассоциированных со стрессом. Показана способность разработанного программного инструментария делать потенциально важные выводы, выдвигать гипотезы по интерпретации данных системной биологии и визуализировать полученные результаты.

Научная новизна. С использованием современного языка программирования Java 8 разработаны программные приложения Orthoscape и Orthoweb. Первое приложение является подключаемым к многофункциональному программному комплексу Cytoscape (http://apps.cytoscape.org). Оно предназначено для анализа эволюционной информации о генах в генных сетях, а именно: а) анализа с целью выявления, являются ли гены гомологичными, б) поиска предполагаемого этапа возникновения гена на таксономическом дереве, в) определения уровня эволюционной изменчивости гена. Интеграция приложения с Cytoscape позволяет удобно решать сопутствующие задачи работы с генными сетями и визуализацией результатов. Ярким подтверждением ценности разработанной программы является факт того, что она была скачана 9020 раз на момент середины апреля 2021 г. Второе разработанное автором приложение Orthoweb обладает схожим с Orthoscape функционалом, но ориентировано на анализ групп генов, не обязательно представляющих собой единую генную сеть. Кроме того Orthoweb, в отличие от Orthoscape, не требует локальной установки Cytoscape и может работать через стандартный веб-браузер. В обоих приложениях, среди прочего, реализован расчет PAI (phylostratigraphic age index) – индекс макроэволюционного анализа, отражающий возраст эволюционного возникновения генов и DI (divergence index) – индекс микроэволюционного анализа, отражающий уровень эволюционной изменчивости генов. Способность разработанного инструментария делать адекватные выводы была установлена путем проверки уже известных утверждений. Так, генная сеть биосинтеза стероидов оказалась «более древней», чем генная сеть биосинтеза стероидных гормонов. Далее, были сделаны оригинальные выводы. У человека эволюционно молодыми генами оказались обогащены генные сети, связанные с заболеваниями иммунной системы, а эволюционно древними – с зависимостью от веществ, вызывающих привыкание. У *Arabidopsis thaliana* генные сети, ассоциированные с реакцией на температуру, свет, соленость среды и присутствие окислителей, обогащены эволюционно древними и консервативными генами.

Структура диссертационной работы. Диссертация состоит из введения, пяти глав, заключения, выводов, списка литературы из 139 наименований и списка используемы аббревиатур и сокращений. Основное содержание диссертации изложено на 101 странице, содержит 42 иллюстрации и 6 таблиц. Полный текст диссертации занимает 116 страниц.

Первая глава посвящена обзору литературы. В ней описывается суть методики филостратиграфического. Описываются уже известные ранее индексы эволюции TAI, PAI, TDI, DI, dN/dS (Ka/Ks) и методы, лежащие в основе их расчета. Также в обзоре литературы рассматриваются современные способы реконструкции и визуализации биологических сетей Cytoscape, Pathway Studio, ANDSystem, GeneMANIA, String. Рассматриваются биоинформатические базы данных KEGG, Ensembl, TAIR, DAVID, STRING.

Во второй главе изложены материалы и методы, касающиеся разработки программной части. Перечислены все сторонние библиотеки, программные средства, СУБД (системы управления базами данных) и фреймворки, использованные при разработке приложений Orthoscape и Orthoweb. Описан принцип работы Cytoscape с подключаемыми модулями. В данной главе также описан способ применения биоинформатических баз данных и программные средства для обращения к ним.

В третьей главе подробно описаны приложения Orthoscape и Orthoweb. Дано описание функционала, ссылка доступа к исходному коду, обозначены правила лицензирования кода программ, особенности использования биоинформатических баз данных, особенности вычисления эволюционных индексов PAI и DI, дано описание способов визуализации результатов.

В четвертой главе изложены результаты исследования эволюционных характеристик генных сетей болезней человека. С помощью Orthoscape были проанализированы генные сети из KEGG Pathway, раздел Human Diseases, состоящий из групп онкология, иммунные заболевания, нейродегенеративные заболевания, зависимость от веществ, сердечно-сосудистые, эндокринные заболевания и болезни нарушения метаболизма, инфекционные заболевания, резистентность к лекарственным препаратам. В общей сложности 80 сетей и 11 групп. В результате анализа были выявлены следующие тенденции. Сети, в которых максимально повышено количество эволюционно-молодых генов (появившихся у позвоночных, PAI>=5), как правило, принадлежат к группе Иммунных заболеваний. Самое высокое значение PAI у генов сети Астма. Сети, в которых максимально повышено число эволюционно-старых генов, как правило, принадлежат к группе Зависимость от веществ. Самое низкое значение PAI (наиболее эволюционно старые гены, появившиеся на этапе формирования эукариот) у генов сети Никотиновая зависимость. Из этого следует вывод, что гены, участвующие в регуляции процессов зависимостей от веществ, содержат в себе гены, ортологи которых встречаются в большем спектре организмов. Как правило это гены, регулирующие основополагающие для организма и клеток процессы. Например, в случае с Никотиновой зависимостью, это гены, связанные с процессом дыхания. Далее автором показано, что значения индексов PAI и DI для 80 сетей заболеваний человека (рассчитанные как средние значения соответствующих индексов входящих в эти сети генов) достоверно скоррелированы между собой. Т.е. чем меньше эволюционный возраст генов, тем больше их эволюционная изменчивость. В этой же главе проведен более подробный анализ сетей болезни Паркинсона и Диабета I и II типов. Для болезни Паркинсона показано, что ключевые для развития болезни гены являются эволюционно молодыми и обладают высокой изменчивостью относительно гоминид. Процесс апоптоза регулируют эволюционно молодые гены, но в целом в сети преобладают эволюционно древние гены.

В пятой главе с помощью Orthoscape и Ortoweb были проанализированы списки генов, которые ассоциированы со стрессом у *Arabidopsis thaliana*. Рассмотрены 7 типов стресса: холодовой, тепловой, световой, осмотический, оксидативный, солевой и водный. Автор показал, что в списках генов, ассоциированных со стрессом, значительно выше доля эволюционно древних генов по сравнению с полным списком генов этого растения. Кроме того, в них больше консервативных генов и меньше изменчивых. Также более подробный анализ сетей стресса позволил предположить, что в процессе эволюции новые функции могут вносить молодые гены, в то время как в основе сети лежит кластер эволюционно древних генов. Кроме того, результаты анализа могут свидетельствовать о многофункциональности древних генов, участвующих в реакции на стресс, вплоть до участия в процессах, которые образовались у растений уже на более поздних этапах эволюции.

Разделы диссертации Заключение и Выводы адекватно отражают полученные результаты.

Теоретическая и практическая значимость работы. Результаты диссертации будут интересны для сотрудников научно-исследовательских организаций и для специалистов других организаций и предприятий, проводящих анализ молекулярно-генетических систем. Программные средства, разработанные автором, представляют интерес для тех, кто занимается интерпретацией экспериментальных данных геномики, транскриптомики, протеомики и мультиомиксным анализом в целом. Разработанные программы подходят для решения такой практической проблемы биоинформатики, как выбор нескольких штук потенциальных биомаркеров из нескольких тысяч кандидатов. С помощью представленных инструментов можно выделять наиболее интересные для дальнейшего рассмотрения гены. Как правило, это гены на разных концах спектра эволюционной консервативности, эволюционной древности, занимающие особое положение в генной сети того или иного процесса. Полученные в диссертации выводы о генах иммунных и онкологических, заболеваний, а также связанных с привыканием к веществам, болезни Паркинсона и диабета имеют научное значение.

Автореферат. Текст автореферата соответствует содержанию диссертационной работы. Содержание работы отражено в публикациях. По теме работы опубликованы 3 статьи в изданиях из списка ВАК, получено авторское свидетельство на программу Ортоскейп. Результаты работы были доложены на профильных конференциях.

Замечания. Представленная работа выполнена на достаточном научном и методическом уровне. К работе имеются следующие замечания.

1) Количество опечаток в работе автором сведено к минимуму. Обнаружена только одна опечатка в автореферате на стр. 13 и четыре в диссертации (стр. 14, 40, 71, 103). Это очень хороший показатель.
2) Автор неоднократно употребляет термин «фенотипический признак». Данное словосочетание является довольно распространенным, но есть сомнения в грамотности его построения. По определению, фенотип – это совокупность биологических свойств и признаков организма, сложившаяся в процессе его индивидуального развития. С учетом данного определения, «фенотипический признак» звучит, как тавтология. Вместо него можно просто употреблять термин «фенотип».
3) В разделе «Апробация работы» автор перечисляет конференции, симпозиумы и практические курсы. Но при этом не приводится ни одного факта выступления на научно-практическом семинаре. Если таковые были, то их тоже следовало упомянуть.
4) Для лучшего понимания, почему из формулы TAI_S следует, что на начальной и конечной стадиях онтогенеза работают «эволюционно молодые» гены на стр. 22, стоило бы упомянуть, что эволюционный возраст гена тем меньше, чем больше индекс ps_i и наоборот.
5) Главы 1 и 2 (Обзор литературы и Материалы и методы) написаны четко и понятно, но к ним имеются замечания. С одной стороны, они нагружены большим количеством технической информации о базах данных и программных средствах. С другой стороны, по результату прочтения обзора литературы возникает чувство недосказанности. А именно, стоило бы расширить обзор с целью лучшего понимания со стороны читателей степени новизны положения №2, выносимого на защиту (у человека эволюционно молодыми генами обогащены генные сети, связанные с заболеваниями иммунной системы, а эволюционно древними – с зависимостью от веществ, вызывающих привыкание). Так, например, сам же автор на стр. 24 пересказывает наблюдение, полученное в 2008 г. учеными Domazet-Loso и Tautz. Суть его в том, что гены, возникшие на более поздних стадиях эволюции являются ассоциированными с функцией иммунного ответа. Это наблюдение, полученное ранее другими учеными, очевидно, связано с положением №2, выносимым на защиту. Хотелось бы получить разъяснения, является ли более позднее утверждение прямым следствием из более раннего или нет. Также на стр. 26 автор перечисляет программные продукты и базы данных для реконструкции

генных сетей. Среди популярных инструментов стоило бы указать Ingenuity Pathway Analysis.

6) В третьей главе указывается, что программа Orthoscape может запрашивать в KEGG при помощи API доменный состав белков с целью дальнейшего анализа. Известно ли, откуда KEGG берёт информацию о доменном составе белков и как часто эта информация обновляется?
7) Самой досадной опечаткой (ошибкой?) является следующее. На стр. 20 автор написал, что 64 триплета кодируют 22(!) аминокислоты и стоп кодон, проиллюстрировав это рисунком 1.5.
8) В качестве совета автору для будущей работы, за рамками данной диссертации, хотелось бы предложить проанализировать методом Orthoscape генные сети заболеваний организма Голого землекопа *Heterocephalus glaber.*

Перечисленные замечания не снижают ценность работы. Они имеют целью обратить внимание автора на нюансы, к которым следует относиться более внимательно в будущих научных работах. Кроме того, дают идеи для дальнейших исследований.

Заключение. Диссертационная работа Мустафина Захара Сергеевича «Разработка комплекса программ для анализа эволюционных характеристик генных сетей», представленная на соискание научной степени кандидата биологических наук по специальности 03.01.09 – математическая биология, биоинформатика, является завершенной научно-исследовательской работой в области изучения эволюции молекулярно-генетических систем. Автором создано популярное программное обеспечение для микро и макроэволюционного анализа генов в составе генных сетей. Автор умело использовал имеющиеся в свободном доступе технологии и знания, проявив должную квалификацию, обогатив арсенал биоинформатиков новыми программами. В своей работе автор также продемонстрировал возможности применения созданного им инструментария для решения актуальных задач системной биологии. Автором сделаны важные выводы об особенностях эволюции молекулярно-генетических систем человека и растения *A. thaliana*. Созданный программный инструментарий и сделанные с его помощью выводы вносят ценный вклад в развитие теории и практики системного анализа генных сетей, предоставляют новые возможности для интерпретации экспериментальных данных.

Тема диссертации, публикации по работе и положения, выносимые на защиту, полностью соответствуют специальности 03.01.09. Диссертация состоит из списка используемых аббревиатур и сокращений, введения, пяти глав, заключения, выводов, списка литературы из 139 наименований. Содержание текста диссертации изложено на 116 страницах, содержит 42 иллюстрации и 6 таблиц. Структура диссертационной работы

позволяет автору в логической последовательности изложить необходимый материал, сформулировать цели работы, описать методы и результаты исследований. Стоит отметить четкость и ясность изложения материала и высокую проработанность диссертации с точки зрения правил русского языка и орфографии. Выводы диссертации сформулированы четко и адекватно отражают полученные результаты. Диссертационная работа Мустафина Захара Сергеевича апробирована на всероссийских и международных профильных научных конференциях. Текст автореферата соответствует содержанию диссертационной работы. Содержание работы в полной мере отражено в публикациях – опубликованы 3 статьи в изданиях из списка ВАК – и в тезисах конференций.

Представленная работа полностью соответствует критериям пп. 9-14 «Положения о присуждении научных степеней», утверждённого постановлением Правительства Российской Федерации № 842 от 24 сентября 2013 г. (в редакции с изменениями, утвержденными Постановлением Правительства РФ от 21 апреля 2016 г. №335), предъявляемым к диссертациям на соискание ученой степени кандидата биологических наук, а ее автор Мустафин Захар Сергеевич заслуживает присуждения ученой степени кандидата биологических наук по специальности 03.01.09 – математическая биология, биоинформатика.

Официальный оппонент, 27.09.2021
старший научный сотрудник
Лаборатории моделирования сложных систем
Института систем информатики им. А.П. Ершова СО РАН,
кандидат физико-математических наук Штокало Дмитрий Николаевич

Федеральное государственное бюджетное учреждение науки Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук
630090, Новосибирск, пр. Акад. Лаврентьева 6
+7(383)332-16-76
shtokalod2@gmail.com

Подпись Д.Н. Штокало заверяю: заместитель директора по научной работе А.В. Кроновский