На правах рукописи

Мустафин Захар Сергеевич

РАЗРАБОТКА КОМПЛЕКСА ПРОГРАММ ДЛЯ АНАЛИЗА ЭВОЛЮЦИОННЫХ ХАРАКТЕРИСТИК ГЕННЫХ СЕТЕЙ

03.01.09 – Математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание учёной степени

кандидата биологических наук

Новосибирск – 2021

Работа выполнена в Федеральном государственном бюджетном научном учреждении «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук» в секторе компьютерного анализа и моделирования биологических систем, г. Новосибирск

Научный руководитель	**Лашин Сергей Александрович** к.б.н., в.н.с., и.о. заведующего сектором компьютерного анализа и моделирования биологических систем, ФГБНУ ФИЦ «Институт цитологии и генетики СО РАН», г. Новосибирск
Официальные оппоненты	**Щербаков Дмитрий Юрьевич** д.б.н., заведующий лабораторией геносистематики, ФГБУН Лимнологический институт СО РАН, г. Иркутск **Штокало Дмитрий Николаевич** к.ф.-м.н., старший научный сотрудник лаборатории моделирования сложных систем, Институт Систем Информатики им. А.П. Ершова СО РАН, г. Новосибирск

Ведущее учреждение: ФБУН ГНЦ ВБ «Вектор» Роспотребнадзора, г. Новосибирск

Защита диссертации состоится «__» _____ 2021 г. на утреннем заседании диссертационного совета 24.1.239.01 (Д 003.011.01) на базе ФГБНУ «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук» в конференц-зале Института по адресу:
пр. ак. Лаврентьева 10, г. Новосибирск, 630090,
тел +7 (383) 3634906, факс +7(383) 3331278.
e-mail: dissov@bionet.nsc.ru.

С диссертацией можно ознакомиться в библиотеке ИЦиГ СО РАН и на сайте Института: www.bionet.nsc.ru.

Автореферат разослан «__» _____ 2021 г.

Ученый секретарь
диссертационного совета, Т.М. Хлебодарова
доктор биологических наук

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Изучение эволюции молекулярно-генетических систем – одна из глобальных задач биоинформатики. Одним из развивающихся подходов является филостратиграфический анализ, он позволяет определить возраст гена через оценку времени его возникновения на основе изучения распределения ортологичных рассматриваемому генов в геномах организмов, принадлежащих к различным таксономическим группам. Наряду с методами микроэволюционного анализа (например, оценка соотношения *dN/dS*), филостратиграфические методы всё больше входят в методический арсенал эволюционных биоинформатиков.

В настоящий момент разработано несколько приложений для филостратиграфического анализа генов, однако, известно, что формирование фенотипических признаков, обеспечивающих адаптацию организмов к условиям окружающей среды, контролируется не отдельными генами, а генными сетями – группами координированно функционирующих генов и продуктов их работы (РНК, белками, метаболитами и др.). Анализ сетей начинает играть всё более важную роль в различных областях биологии и на данный момент наблюдается дефицит программного обеспечения для эволюционного анализа генных сетей.

На основе филостратиграфического анализа списков генов человека, ассоциированных с онкологическими заболеваниями, показно, что большинство таких генов являются эволюционно древними. Однако, к моменту начала этой работы, филостратиграфический анализ обширной группы генов и генных сетей, связанных с заболеваниями человека различной природы, не проводился.

Одним из самых многофункциональных комплексов для работы с сетями, в т.ч. биологическими, является Cytoscape (http://apps.cytoscape.org/). Важное достоинство Cytoscape заключается в том, что пользователи могут реализовывать собственный функционал в виде подключаемых приложений. В этой работе представлены приложения Orthoscape и Orthoweb. Orthoscape – приложение, подключаемое к Cytoscape, направленное на анализ эволюционной информации о генах в генных сетях, а именно: (1) анализ с целью выявления, являются ли гены гомологичными; (2) поиск предполагаемого этапа возникновения гена на таксономическом дереве; (3) определение уровня эволюционной изменчивости гена. Orthoweb – веб-приложение со схожей с Orthoscape

функциональностью, ориентированное на анализ функционально связанных групп генов, не объединенных в генную сеть (т.е. без ребер).

Цели и задачи диссертационной работы. Целью данной работы является разработка комплекса компьютерных методов филостратиграфического анализа и определения уровня эволюционной изменчивости генов и генных сетей, и его применение.

Для достижения цели были поставлены следующие задачи:

1. Разработка компьютерных программ Orthoscape и Orthoweb для определения эволюционного возраста генов, кодирующих белки, в составе генных сетей и анализа особенностей эволюционной изменчивости этих генов.
2. Анализ эволюционных особенностей генных сетей заболеваний человека, представленных в базе данных KEGG.
3. Анализ эволюционных особенностей генов, ассоциированных с различными типами абиотического стресса у *Arabidopsis thaliana*.

Научная новизна работы. Впервые филостратиграфический анализ был применен для анализа генных сетей. Впервые разработаны и реализованы программы для филостратиграфического анализа генных сетей, и проведен филостратиграфический анализ генных сетей заболеваний человека различной природы.

Теоретическая и практическая значимость работы. Разработанные программные средства Orthoscape и Orthoweb могут быть использованы для анализа таких эволюционных характеристик, как возраст гена и степень давления отбора на ген, что позволяет определить, какие гены в тех или иных процессах являются наиболее эволюционно древними/молодыми и в то же время, являются ли эти гены консервативными или же, наоборот, изменчивыми. В сочетании с генной сетью эта информация позволяет выделить целые кластеры генов, интересных для более подробного исследования. На данный момент Orthoscape является самым скачиваемым приложением к Cytoscape с тегом evolution (9020 скачиваний на середину апреля 2021 года).

Положения, выносимые на защиту.

1. Программы Orthoscape и Orthoweb позволяют проводить анализ эволюционных особенностей генных сетей у различных видов организмов на основе определения таких характеристик, как возраст генов и уровень их изменчивости.

2. У человека эволюционно молодыми генами обогащены генные сети, связанные с заболеваниями иммунной системы, а эволюционно древними – с зависимостью от веществ, вызывающих привыкание.
3. У *A. thaliana* генные сети, ассоциированные с реакцией на температуру, свет, соленость среды и присутствие окислителей, обогащены эволюционно древними и консервативными генами.

Апробация работы. Основные результаты работы были представлены на следующих конференциях:

1. The 12th International Young Scientists School «Systems Biology and Bioinformatics» (SBB 2020) (Ялта/Севастополь, 2020).
2. «Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB 2020) (Новосибирск, 2020).
3. VII съезд Вавиловского общества генетиков и селекционеров, посвященный 100-летию кафедры генетики СПбГУ, и ассоциированные симпозиумы (Санкт-Петербург, 2019).
4. III Российская мультидисциплинарная конференция с международным участием «Сахарный диабет-2019: от мониторинга к управлению» (Новосибирск, 2019).
5. The 11th International Young Scientists School «Systems Biology and Bioinformatics» (SBB 2019) (Новосибирск, 2019).
6. V Международная конференция. «Постгеном 2018». В поисках моделей персонализированной медицины (Казань, 2018).
7. «Bioinformatics of Genome Regulation and Structure/Systems Biology» (BGRS/SB 2018) (Новосибирск, 2018).
8. Международная конференция, посвященная 100-летию со дня рождения академика АН СССР Дмитрия Константиновича Беляева (Новосибирск, 2017).
9. Международный форум «Биотехнология: состояние и перспективы развития» (Москва, 2017).
10. «Bioinformatics of Genome Regulation and Structure/Systems Biology» (BGRS/SB 2016) (Новосибирск, 2016).
11. The 8th International Young Scientists School «Systems Biology and Bioinformatics» (SBB 2016) (Новосибирск, 2016).

Объем и структура диссертации. Диссертация изложена на 116 страницах машинописного текста, содержит 42 рисунка и 6 таблиц. Список литературы включает 139 ссылок. Диссертация состоит из введения, литературного

обзора, описания материалов и методов, главы с описанием разработанных программ, двух глав с описанием полученных на их основе результатов, заключения, выводов и списка литературных источников.

Публикации. По теме диссертации опубликовано 16 работ, из них 3 статьи в рецензируемых научных журналах, входящих в перечень ВАК, 1 авторское свидетельство и 12 тезисов конференций.

Личный вклад автора. Автором были реализованы приложение Orthoscape для анализа эволюционных характеристик генных сетей, импортированных в Cytoscape, и веб-приложение Orthoweb для анализа групп функционально связанных генов, не объединенных в сеть. Проведен анализ генных сетей заболеваний человека, представленных в KEGG и генов *A. thaliana*, ассоциированных с различными типами стресса.

Благодарности. Автор выражает благодарность научному руководителю к.б.н. Лашину С.А., а также к.б.н. Клименко А.И., к.б.н. Казанцеву Ф.В. и академику Колчанову Н.А. за плодотворные научные дискуссии.

Содержание работы

Глава 1. Обзор литературы

В обзоре литературы описывается суть методики филостратиграфического анализа, современное состояние работ в этой области. Описываются уже сформулированные ранее эволюционные характеристики генов и полученные на их основе результаты.

Например, с помощью разработанных ранее эволюционных характеристик TAI (ур. 1) и TDI (ур. 2) для *Arabidopsis thaliana* и *Danio rerio* продемонстрировано, что на раннем и позднем этапе онтогенеза экспрессируются эволюционно молодые и изменчивые гены, а на промежуточном – эволюционно древние и консервативные. Данное явление назвали паттерном песочных часов.

$$TAI_s = \frac{\sum_{i=1}^{n} ps_i e_i}{\sum_{i=1}^{n} e_i} \quad (1)$$

где ps_i – целое число, отражающее возраст для гена с индексом i, e_i – уровень экспрессии гена с номером i, n – общее число генов.

$$TDI_s = \frac{\sum_{i=1}^{n} DI_i e_i}{\sum_{i=1}^{n} e_i} \quad (2)$$

где DI_i – результат вычисления общепринятого отношения dN/dS (Ka/Ks), отражающего отношение несинонимичных и синонимичных замен при сравнении последовательностей для гена с индексом i и его ортолога, e_i – уровень экспрессии гена с номером, n – общее число генов.

Также в обзоре литературы рассмариваются современные способы реконструкции биологических сетей, в частности, Cytoscape как программный комплекс для реконструкции и визуализации генных сетей и подключаемые приложения к Cytoscape. Рассматриваются базы данных, с которыми велась работа:

1) KEGG (Kyoto Encyclopedia of Genes and Genomes) - ресурс, разработанный в Японии, в котором содержится биологическая информация самых разных категорий, в том числе и генные сети, представленные в собственном формате – KGML (KEGG Markup Language).

2) Ensembl, позиционирует себя, как геномный браузер для работ по сравнительной геномике, эволюции, изменчивости последовательностей и регуляции транскрипции.

3) TAIR (The Arabidopsis Information Resource) - база данных, сконцентрированная на *A. thaliana*. В данные TAIR включены как генетические данные о самом растении, так и публикации.

4) DAVID (The Database for Annotation, Visualization and Integrated Discovery) - база знаний, предоставляющая инструменты для определения функциональных особенностей генов, объединенных в список.

5) STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) - база данных с информацией об известных и предсказанных белок-белковых взаимодействиях. На основе данных в базе STRING позволяет реконструировать генные сети по имеющемуся списку генов.

Глава 2. Материалы и методы

В главе перечислены все сторонние библиотеки и программные средства, использованные при разработке приложений Orthoscape и Orthoweb. Описан принцип работы Cytoscape с подключаемыми модулями. Описаны использованные при разработке Orthoweb фреймворки - Spring и Webix (фреймворк – программное обеспечение, облегчающее разработку и объединение разных компонент большого программного проекта). С помощью Spring в работе создается RESTful (основанный на архитектуре REST (Representational State Transfer)) веб-сервис. Сервер содержит базу с данными и проводит анализ этих данных, а клиентом выступает компьютер,

с которого пользователь запускает веб-браузер и веб-приложение Orthoweb. Webix используется для реализации клиентской части, на его основе сконструирован графический интерфейс, с которым пользователь работает в браузере. Данные на сервере хранятся в MongoDB – нереляционной базе данных (т.е. записи хранятся без явных и структурированных механизмов для связывания друг с другом). Записи в базе хранятся в формате json.

Описана область применения используемых в работе баз данных. KEGG использовался для получения вручную реконструированных сетей заболеваний человека, а также таких данных, как списки ортологичных генов, таксономические ряды организмов, гены которых рассматриваются в анализе, нуклеотидные последовательности генов и аминокислотные последовательности белков, кодируемых рассматриваемыми генами, белковые домены. Раздел Ensembl Biomart – для установления связей между генами *H. sapiens* и ассоциированными с ними терминами генной онтологии. TAIR и DAVID – для установления связей между генами *A. thaliana* и ассоциированными с ними терминами генной онтологии. STRING – для построения генных сетей на основе списков генов *A. thaliana,* ассоциированных со стрессом.

Глава 3. Приложения для анализа эволюционных характеристик генных сетей и генов.

Разработанные приложения и рассчитываемые эволюционные индексы.

В главе 3 представляется Orthoscape - приложение к Cytoscape для анализа эволюционных характеристик генных сетей. Приложение Orthoscape написано на языке Java, распространяется по лицензии GPL (General Public License, лицензия на открыто распространяемое программное обеспечение) с использованием сторонних библиотек и плагинов. Приложение доступно для скачивания в Cytoscape App Store (http://apps.cytoscape.org/apps/orthoscape), исходный код выложен в репозитории github (https://github.com/ZakharM/Orthoscape). Orthoscape позволяет искать гены, гомологичные друг другу, анализировать эволюционные характеристики генов, входящих в сеть, как с учетом ее топологии, так и без, т.е. работать с множеством генов. Главным источником информации о генах и их характеристиках является база данных KEGG.

Первая возможность Orthoscape – поиск генов, гомологичных генам из заданного входного набора генов/генной сети. Для поиска гомологов используется база данных KEGG, в которой представлены сведения как для паралогов, так и для ортологов. Отбор происходит по параметрам: идентичность аминокислотных последовательностей кодируемых генами белков, результат алгоритма Смита-Ватермана по выравниванию последовательностей, сходство доменного состава белков.

Вторая возможность Orthoscape – анализ эволюционных характеристик. Эволюционные характеристики генных сетей выражаются с помощью эволюционных характеристик генов, входящих в рассматриваемые сети. Было реализовано две эволюционных характеристики генов – PAI и DI, которые позволяют с разных сторон взглянуть на эволюцию генов, содержащихся в генной сети. Главная задача индексов - помочь выделить в сети именно те гены, на которые следует обратить внимание эксперту для более подробного анализа.

Для исследования эволюции на макроэволюционном уровне был реализован индекс PAI (phylostratigraphic age index). Индекс PAI принимает значение таксона, который является узлом филостратиграфического дерева, наиболее отдаленным от его корня и фигурирующим в таксономических группах гена и всех его ортологов. Таким образом, значение PAI отражает позицию последнего общего предка для гена и всех найденных ортологов на таксономическом дереве. Например, при анализе гена человека *INSIG-2* было найдено 6 ортологов у следующих организмов: *Gorilla gorilla gorilla (ggo), Nomascus leucogenys (nle), Oryctolagus cuniculus (ocu), Pan paniscus (pps), Pan troglodytes (ptr), Saimiri boliviensis boliviensis (sbq)* (рисунок 1).

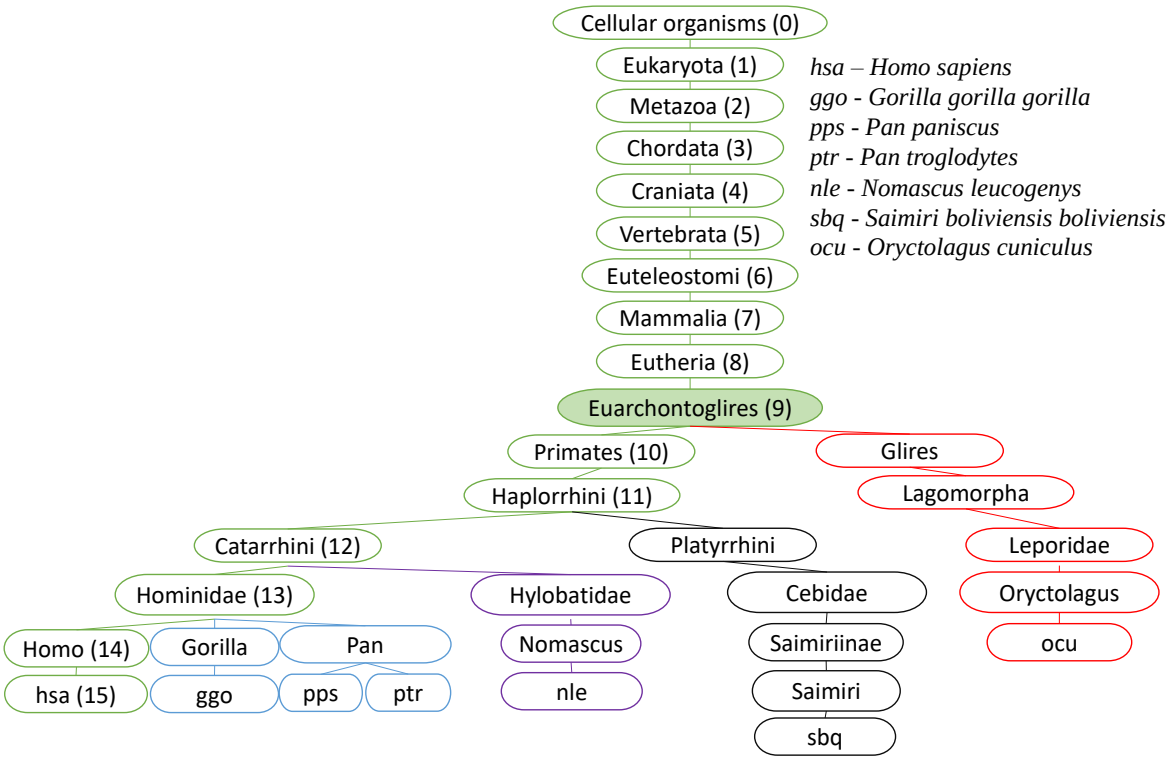

Рис. 1: *Пример определения PAI для гена человека с шестью найденными ортологами. PAI принимает значение таксона, описанного в узле, наиболее отдаленном от корня дерева и являющемся общим для гена и всех ортологов, т. е. полученное значение PAI = 9 (Euarchontoglires).*

Для исследования эволюции на микроэволюционном уровне был реализован индекс DI (divergence index). Данный индекс основан на популярном в среде эволюционных биологов отношении dN/dS. DI отражает степень давления естественного отбора, путем вычисления отношения несинонимичных замен между двумя сравниваемыми нуклеотидными последовательностями к синонимичным:

$$DI_s = \frac{\sum_{i=1}^{n} dnds_i}{n}, \qquad (3)$$

где $dnds_i$ – значение dN/dS отношения для последовательности гена и ортолога с номером i, n – число ортологов, попавших в анализ.

DI следует рассчитывать при анализе ортологов организмов, расположенных близко к исследуемому организму на таксономическом дереве. Основная задача DI – выделить консервативные гены, для которых индекс близок к нулю, и изменчивые гены, для которых индекс обладает наибольшими значениями.

На основе алгоритмов, заложенных в Orthoscape, разработано веб-приложение Orthoweb, не требующее установки Cytoscape и направленное на анализ функционально связанных групп генов, не объединенных в сеть. Для работы с Orthoweb достаточно стандартного веб-браузера. Схема работы приведена на рисунке 2.

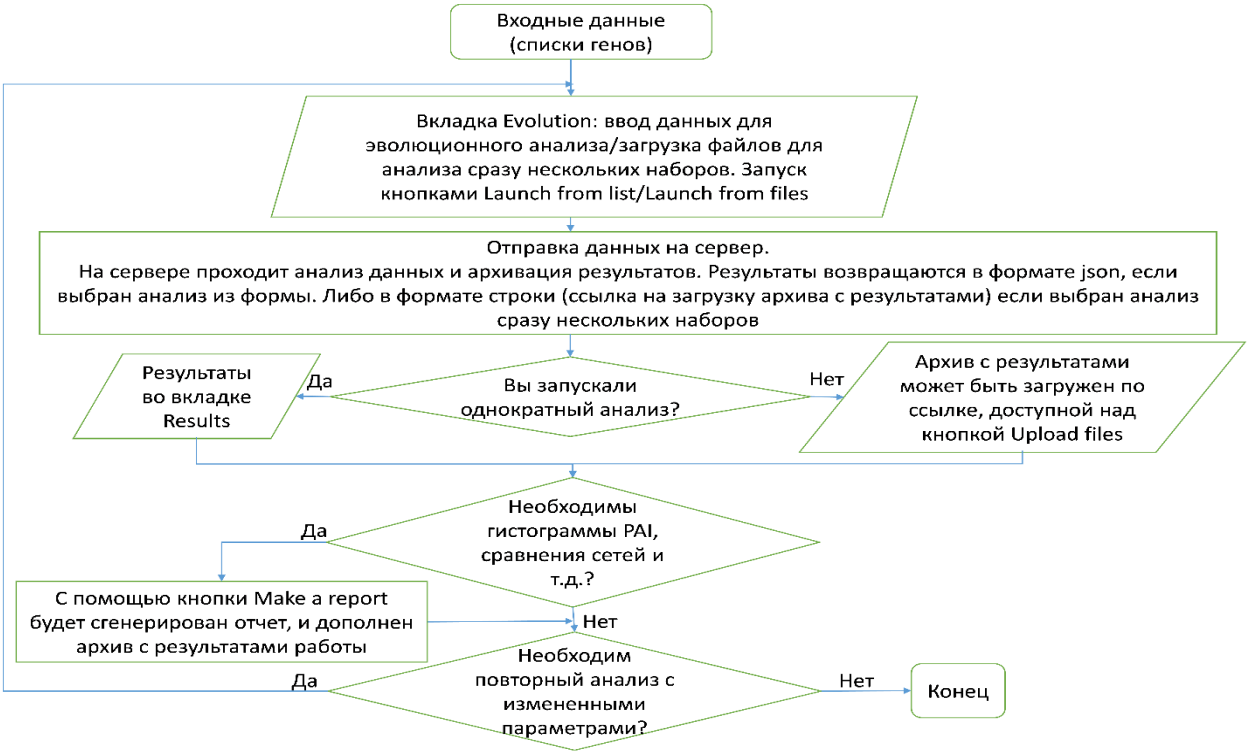

Рис. 2 *Схема работы с Orthoweb.*

Визуализация результатов

Для визуализации результата было реализовано два стиля: тепловая карта (heatmap на основе таких цветов как красный, желтый, зеленый, голубой, синий) и градиентная схема (на основе синего и красного цветов). Схемы можно применять для окрашивания по PAI (программой будет создана дискретная схема) и DI (программой будет создана непрерывная схема).

Для визуализации результатов, которые не требуется отображать на генной сети, был сделан отдельный модуль создания отчетов. Для каждой сети визуализируются следующие результаты: графики значений PAI сети (по оси абсцисс идентичность/SW-Score, по оси ординат среднее значение PAI); гистограммы распределения PAI для всех генов сети; графики violin plot, построенные на основе гистограмм результатов всех проанализированных сетей.

Глава 4. Исследование эволюционных характеристик генных сетей болезней человека

С помощью Orthoscape были проанализированы сети из KEGG Pathway, раздел Human Diseases, состоящий из 11 групп (Cancers: Overview, Cancers: Specific types, Immune diseases, Neurodegenerative diseases, Substance dependence, Cardiovascular diseases, Endocrine and metabolic diseases, Infectious diseases: Bacterial, Infectious diseases: Viral, Infectious diseases: Parasitic, Drug resistance: Antineoplastic), содержащих в общей сложности 80 сетей.

На основе результатов анализа были выделены следующие тенденции: сети, в которых максимальное (по сравнению с другими сетями) среднее значение PAI, как правило, принадлежат к группе Immune diseases, а самое высокое значение PAI у сети Asthma. Таким образом, в заболевания человека, связанные с работой иммунной системы, вовлечены гены, ортологи которых найдены только в отдельных группах наиболее эволюционно молодых организмов и не найдены в остальных, по сравнению с большинством других генов, ортологи которых встречаются в существенно более широком спектре видов.

Сети, в которых минимальное относительно других сетей среднее значение PAI, как правило, принадлежат к группе Substance dependence, а самое низкое значение PAI у сети Nicotine addiction. Из этого можно сделать вывод, что гены, участвующие в регуляции процессов зависимостей от веществ, вызывающих привыкание, 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 Nicotine addiction эти гены связаны с таким биологическим процессом как дыхание.

В таблице 1 приведен результат анализа генов, разбитых уже не на сети, а на группы. Дубли генов, встречающихся сразу в нескольких сетях из группы, удалялись. Результаты оказались сопоставимы с индивидуальным анализом каждой сети из группы.

Таблица 1 Результаты анализа индексов PAI и DI сетей заболеваний человека, представленных в KEGG, сгруппированных по категориям.

Категория	PAI	DI	Среднее число генов в сети	Число сетей	Число уникальных генов
Substance dependence	1,03	0,16	24	5	78
Cancers Specific types	1,77	0,24	43	16	241
Cancers Overview	2,08	0,29	110	7	542
Neurodegenerative diseases	2,08	0,24	38	5	153
Endocrine and metabolic diseases	2,22	0,28	40	6	184
Drug resistance Antineoplastic	2,46	0,28	40	4	130
Cardiovascular diseases	2,63	0,28	41	5	142
Infectious diseases Viral	2,84	0,31	95	9	482
Infectious diseases Bacterial	2,89	0,29	41	10	270
Infectious diseases Parasitic	4,08	0,37	42	6	158
Immune diseases	5,21	0,46	29	8	162

На рисунке 3 можно видеть значения индексов PAI и DI для всех сетей заболеваний человека, представленных в KEGG.

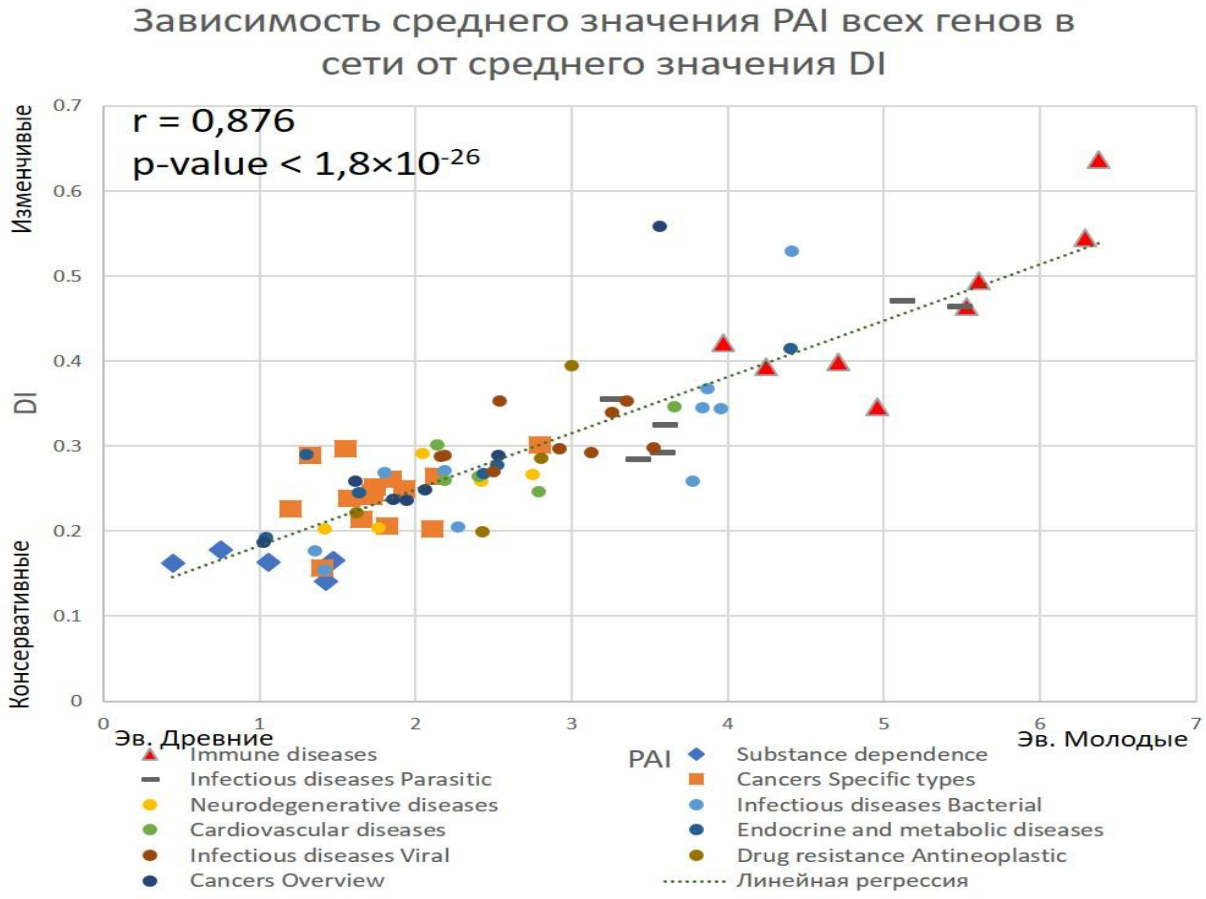

Рис. 3 *Диаграмма рассеяния для средних значений индексов PAI и DI для 80 генных сетей заболеваний человека, описанных в базе KEGG Pathway, Human Diseases.*

Между индексами PAI и DI наблюдается высокая и достоверная корреляция ($r = 0{,}876$, p-value $< 1{,}8 \times 10^{-26}$), т.е., чем меньше эволюционный возраст генов, тем больше уровень их генетической изменчивости. Кроме того, большинство (1396 из 1436) генов, входящих в состав исследованных генных сетей, были идентифицированы, как эволюционно консервативные ($DI < 1$).

Гены, выделяющиеся по PAI и DI, могут оказаться наиболее интересными мишенями для более подробного анализа. Был проведен анализ сетей болезни Паркинсона, Диабета 1 и 2 типа. Анализ генной сети болезни Паркинсона выявил низкий показатель PAI, т.е. гены данной сети являются эволюционно древними, тем не менее, данное заболевание свойственно человеку и другим приматам. Подробный анализ сети показал, что:

1) Процесс апоптоза (клеточной гибели) регулируют эволюционно молодые гены.

2) Наибольшей изменчивостью по отношению к гоминидам обладают именно те гены, по которым уже проведены отдельные исследования и показан их важный вклад в регуляцию болезни Паркинсона (*Parkin, Pink1, LRRK2*).

3) В сети преобладают эволюционно древние гены, которые, в свою очередь, не отличаются повышенной изменчивостью относительно гоминид, за счет чего сильно уменьшается средний возраст генов в сети, но ключевые для развития заболевания гены в большинстве своем являются эволюционно молодыми и обладают высокой изменчивостью относительно гоминид, вероятно, оказывая большое влияние на развитие заболевания.

Глава 5. Исследование эволюционных характеристик генов, ассоциированных со стрессом у *A. thaliana*.

С помощью Orthoscape и Orthoweb были проанализированы списки генов, которые ассоциированы со стрессом у *A. thaliana*. Рассмотрены 7 типов стресса: “холодовой”, “тепловой”, “световой”, “осмотический”, “оксидативный”, “солевой”, “водный”. Списки были сформированы на основе терминов генной онтологии, ассоциированных со стрессом, и данных из базы TAIR. С помощью ресурса String на основе составленных списков были построены генные сети и проанализированы с помощью Orthoscape. С помощью Orthoweb аналогичная работа была проведена для исходных списков генов, ассоциированных со стрессом. PAI и DI анализ списков генов, ассоциированных со стрессом.

Анализ показал, что в списках генов, ассоциированных со стрессом, значительно выше доля эволюционно древних генов по сравнению с полным списком генов *A. thaliana*, кодирующих белки (рис. 4). Кроме того, в них больше консервативных генов и меньше изменчивых, чем в полном списке генов *A. thaliana*, кодирующих белки.

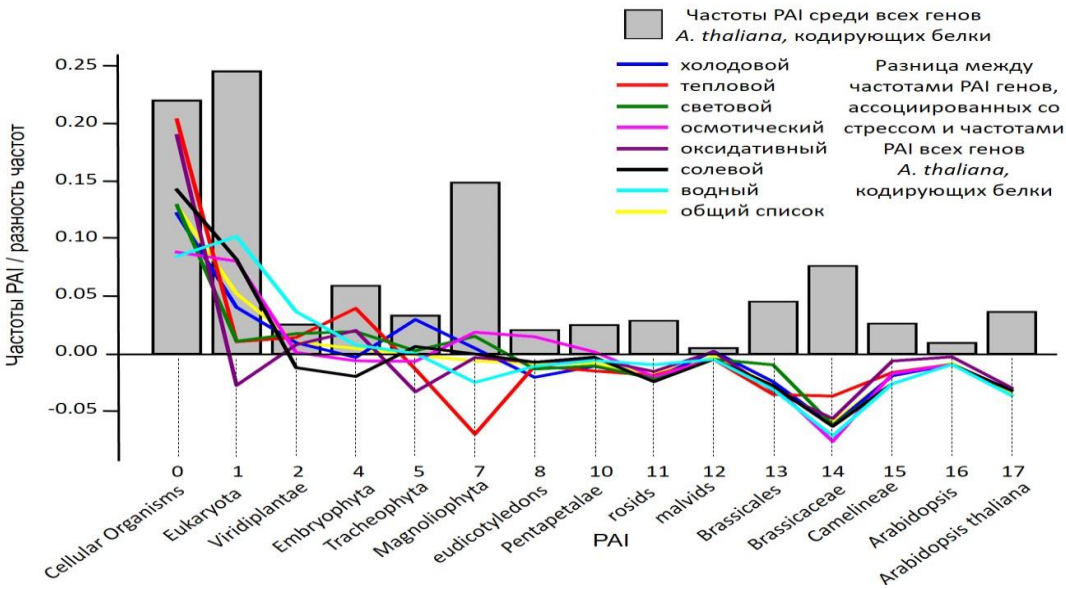

Рис. 4 *Серыми столбцами показано распределение частот встречаемости PAI среди всех генов A. thaliana, кодирующих белки. Цветными линиями показана разница между частотой PAI соответствующего цвету типа стресса и частотой PAI среди всех генов A. thaliana, кодирующих белки. Если линия выше нуля – гены с соответствующем значением PAI преобладают в стрессовой выборке, а если ниже нуля – среди всех генов A. thaliana, кодирующих белки.*

С помощью средства String были реконструированы и разбиты на кластеры генные сети для каждого типа стресса. Анализ показал, что в большинстве случаев различные кластеры описывают молекулярные механизмы стрессового ответа. В сетях, связанных с холодовым, солевым, осмотическим и водным стрессами регуляторная компонента четко выделяется, включает в себя большое число генов и связей между ними. Такие компоненты ассоциированы хорошо известными гормонами абиотического стресса, абсцизовой кислотой и этиленом, содержат много генов, общих для этих типов стресса, что соответствует степени сходства стрессов, рассчитанной с помощью критерия Очиаи. В сети теплового

стресса регуляторная компонента представлена слабее, а в сетях светового и оксидативного стрессов практически отсутствует.

Кроме того, была обнаружена отрицательная корреляция между степенью узла сети и PAI этого узла для трех типов стрессов: теплового, осмотического и солевого. В данных сетях отмечено большое число эволюционно древних генов с высоким числом связей. В сети осмотического стресса это такие гены, как *ABI1/AT4G26080* (степень 23) и *ABI2/AT5G57050* (степень 22), принадлежащие семейству белковой Фосфотазы 2C, содержат большое число связей с другими генами эукариот. Следует отметить, что не было обнаружено положительной корреляции ни для одного типа стресса, что говорит о том, что эволюционно молодые гены не обладают значимо большим числом связей с другими генами.

Подобные результаты позволяют предположить, что в процессе эволюции новые функции могут вносить молодые гены, в том время как в основе сети лежит кластер эволюционно древних генов. Были проанализированы термины генной онтологии, с которыми ассоциированы рассмотренные гены, и показано, что многие гены функционально связаны сразу с несколькими типами стресса. Например, для теплового стресса наблюдаются термины, функционально связанные с холодовым стрессом (“response to freezing, “response to cold”). С ними ассоциирован ген *TIL/AT5G58070* (temperature-induced lipocalin), важный компонент регуляции температурного режима. *TIL1* локализован в плазматической мембране и экспрессируется в ответ на воздействие холода, таким образом, предположительно, выполняя защитную роль в условиях вызванной холодом дегидротации. *TIL1* перемещается под действием солевого стресса и защищает хлоропласты от ионной токсичности. За счет подобных многофункциональных генов обеспечивается связь различных типов стресса общими терминами генной онтологии. Полученные результаты свидетельствуют о многофункциональности древних генов, участвующих в реакции на стресс, вплоть до их участия в процессах, которые образовались у растений уже на более поздних этапах эволюции.

Выводы

1. Разработаны компьютерные программы Orthoscape и Orthoweb для анализа эволюции генных сетей на основе оценки таких характеристик, как: (а) филостратиграфический индекс генов (PAI), отражающий возраст их эволюционного возникновения и, (б) индекс дивергенции генов (DI), отражающий уровень их эволюционной изменчивости.

2. На основе анализа 80 генных сетей заболеваний человека из базы данных KEGG показано:

а. Генные сети категории «Иммунные заболевания» содержат наибольшую долю эволюционно молодых генов (65%) по сравнению с другими генными сетями, а генные сети категории «Зависимость от веществ, вызывающих привыкание» – наибольшую долю эволюционно древних (88%).

Генные сети категории «Инфекционные заболевания, вызванные паразитами» достоверно обогащены эволюционно молодыми генами.

Генные сети категории «Специфические типы рака» достоверно обогащены эволюционно древними генами.

б. Подавляющее большинство (1396 из 1436) генов, задействованных в исследованных генных сетях, являются эволюционно консервативными (DI < 1).

в. Наблюдается достоверная ($r = 0{,}876$, p-value < $1{,}8 \times 10^{-26}$) зависимость между средним эволюционным возрастом генов в генных сетях и уровнем их эволюционной изменчивости: чем меньше эволюционный возраст генов, тем больше уровень их эволюционной изменчивости.

3. Проведен анализ генов *A. thaliana*, ассоциированных с ответом на холодовой, солевой, тепловой, осмотический, оксидативный, водный, световой стрессы. Показано, что выборки генов *A. thaliana*, ассоциированных с реакцией на абиотические стрессы, достоверно обогащены эволюционно древними и консервативными генами.

Публикации

Статьи в рецензируемых журналах, входящих в перечень ВАК:

1. **Мустафин З.С.**, Лашин С.А., Матушкин Ю.Г. Филостратиграфический анализ генных сетей заболеваний человека // Вавиловский журнал генетики и селекции. – 2021. 25(1). - 46-56.
2. **Mustafin Z.S.,** Zamyatin V.I., Konstantinov D. K., Doroshkov A. V., Lashin S. A., Afonnikov D. A. Phylostratigraphic Analysis Shows the Earliest Origination of the Abiotic Stress Associated Genes in *A. thaliana* // Genes. – 2019. 10(12). - 963.
3. **Mustafin Z.S.**, Lashin S.A., Matushkin Yu.G., Gunbin K.V., Afonnikov D.A. Orthoscape: a cytoscape application for grouping and visualization kegg based gene networks by taxonomy and homology principles // BMC Bioinformatics. – 2017. 18:427.

Авторские свидетельства:

1. Лашин С.А., Афонников Д.А., **Мустафин З.С.**, Матушкин Ю.Г., Гунбин К.В. Программа для анализа эволюционных характеристик генных сетей (Ортоскейп) / Application for evolutionary analysis of gene networks (Orthoscape), 2016.

Тезисы конференций:

1. **Mustafin Z.S.**, Mukhin A.M., Afonnikov D.A., Matushkin Yu.G., Lashin S.A. OrthoWeb – web application for macro-and microevolutionary analysis of genes // Bioinformatics of Genome Regulation and Structure/Systems Biology. – Novosibirsk. – 2020. - p.228.

2. Zamyatin V., **Mustafin Z.S.**, Matushkin Yu.G., Afonnikov D.A., Klimontov V.V., Lashin S.A. Gene network of type 2 diabetes: reconstruction and analysis // Bioinformatics of Genome Regulation and Structure/Systems Biology. – Novosibirsk. – 2020. - p.196.
3. Лашин С.А., **Мустафин З.С.**, Замятин В.И., Константинов Д.К., Дорошков А.В., Афонников Д.А. Эволюционный анализ генных сетей абиотического стресса растений // VII съезд Вавиловского общества генетиков и селекционеров, посвященный 100-летию кафедры генетики СПбГУ, и ассоциированные симпозиумы. - Санкт-Петербург. – 2019. - стр. 134.
4. Замятин В.И., **Мустафин З.С.**, Матушкин Ю.Г., Климонтов В.В., Лашин С.А. Реконструкция и анализ генной сети сахарного диабета 2 типа // III Российская мультидисциплинарная конференция с международным участием

«Сахарный диабет-2019: от мониторинга к управлению». – Новосибирск. - 2019. - стр. 33-35.

5. Zamyatin V., **Mustafin Z.S.**, Matushkin Yu.G., Klimontov V.V., Lashin S.A. Gene networks of type 2 diabetes and Alzheimer's disease. Reconstruction and analysis // The 11th international young scientists school «Systems biology and bioinformatics». – Novosibirsk. – 2019. - p.51.

6. Лашин С.А., **Мустафин З.С.**, Замятин В.И., Афонников Д.А., Матушкин Ю.Г., Колчанов Н.А. Программные средства для комплексного анализа генных сетей // V Международная конференция. «Постгеном 2018». В поисках моделей персонализированной медицины. – Казань. – 2018. - стр. 90.

7. Lashin S.A., **Mustafin Z.S.**, Manevich V.A., Afonnikov D.A., Ignatieva E.V., Matushkin Yu.G., Klimontov V.V. Evolutionary Analysis and Mathematical Modeling of Gene Networks of Energy Metabolism Disorders // Bioinformatics of Genome Regulation and Structure/Systems Biology. – Novosibirsk. – 2018. - p.78.

8. **Mustafin Z.S.**, Afonnikov D.A., Matushkin Yu.G., Lashin S.A. On evolutionary analysis of gene networks by the Orthoscape software // Bioinformatics of Genome Regulation and Structure/Systems Biology. – Novosibirsk. – 2018. - p.41.

9. **Мустафин З.С.**, Афонников Д.А., Гунбин К.В., Матушкин Ю.Г., Лашин С.А. Orthoscape: Cytoscape приложение для анализа эволюционных характеристик генных сетей // Belyaev conference: a triumphant event in commemoration of the centenary of the birth of academician Dmitri Belyaev. – Novosibirsk. – 2017. - p.177.

10. Лашин С.А., **Мустафин З.С.,** Матушкин Ю.Г., Гунбин К.В., Афонников Д.А. Анализ эволюционных характеристик генных сетей с помощью программы Orthoscape // Материалы международного конгресса «Биотехнология: состояние и перспективы развития». – Москва. – 2017. - с. 364-365.

11. **Mustafin Z.S.**, Afonnikov D.A., Gunbin K.V., Matushkin Yu.G., Lashin S.A. Orthoscape: a cytoscape plugin for evolutionary analysis of gene networks // Bioinformatics of Genome Regulation and Structure/Systems Biology. – Novosibirsk. – 2016. - p. 195.

12. **Mustafin Z.S.**, Afonnikov D.A., Gunbin K.V., Matushkin Yu.G., Lashin S.A. Orthoscape: a cytoscape plugin for evolutionary analysis of gene networks // The 8th international young scientists school «Systems biology and bioinformatics». – Novosibirsk. – 2016. - p. 49.