

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
НАУЧНОЕ УЧРЕЖДЕНИЕ
«ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР ИНСТИТУТ
ЦИТОЛОГИИ И ГЕНЕТИКИ
СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК»

На правах рукописи

Мустафин Захар Сергеевич

**РАЗРАБОТКА КОМПЛЕКСА ПРОГРАММ ДЛЯ
АНАЛИЗА ЭВОЛЮЦИОННЫХ ХАРАКТЕРИСТИК
ГЕННЫХ СЕТЕЙ**

Специальность 03.01.09

Математическая биология, биоинформатика

Диссертация на соискание учёной степени

Кандидата биологических наук

Научный руководитель:

к.б.н. Лашин С. А.

Новосибирск

2021

Оглавление

Список используемых аббревиатур и сокращений.....	4
Введение	5
Цели и задачи диссертационной работы	7
Научная новизна работы	8
Теоретическая и практическая значимость работы.....	8
Положения, выносимые на защиту	8
Апробация работы.....	9
Объем и структура диссертации.....	10
Публикации.....	10
Личный вклад автора	13
Благодарности	13
1. Обзор литературы	14
1.1. Эволюционные характеристики генов.....	14
1.1.1 Гомология	15
1.1.2 Филостратиграфические деревья	17
1.1.3 Индексы молекулярной эволюции.....	20
1.2. Генные сети	24
1.3. Базы данных и знаний в биологии	30
1.3.1. KEGG.....	32
1.3.2. Ensembl.....	34
1.3.3. TAIR	35
1.3.4. DAVID.....	36
1.3.5. STRING	37
1.4. Заключение к главе 1	39
2. Материалы и методы	40
2.1. Веб-приложение и использованные при разработке технологии	41
2.1.1. Spring.....	41

2.1.2. Webix	43
2.1.3. MongoDB	46
2.2. Cytoscape и подключаемые к Cytoscape приложения	46
3. Приложения для анализа эволюционных характеристик генных сетей и генов.	49
3.1. Функциональные возможности Orthoscape и Orthoweb.....	49
3.1.1 Поиск гомологов	49
3.1.2 Анализ эволюционных характеристик	51
3.1.3 Визуализация результатов	55
3.2. Отличия Orthoscape и Orthoweb	58
3.3. Интерпретация результатов, полученных с помощью Orthoscape	64
3.4. Заключение к главе 3	67
4. Исследование эволюционных характеристик генных сетей болезней человека	68
4.1. Анализ генной сети болезни Паркинсона	76
4.2. Анализ генных сетей Диабета I и II типов	80
4.3. Заключение к главе 4	85
5. Исследование эволюционных характеристик генов, ассоциированных со стрессом у <i>A. thaliana</i>.	86
5.1. Подготовка списков генов и генных сетей.....	87
5.2. PAI и DI анализ списков генов, ассоциированных со стрессом.	88
5.3. PAI и DI анализ генных сетей, реконструированных на основе ассоциированных со стрессом генов.....	94
Заключение.....	102
Выводы.....	104
Список литературы	105

Список используемых аббревиатур и сокращений

API – Application Programming Interface (Интерфейс Прикладного Программирования)

DI – Divergence Index (Индекс Изменчивости)

EMBL - European Molecular Biology Laboratory (Европейская Молекулярно-Биологическая Лаборатория)

IDDM – Insulin Dependent Diabetes Mellitus (Инсулин Зависимый Сахарный Диабет)

KEGG - Kyoto Encyclopedia of Genes and Genomes (Киотская Энциклопедия Генов и Геномов)

NCBI - National Center for Biotechnological Information (Национальный Центр Биотехнологической Информации)

PAI – Phylostratigraphic Age Index (Индекс Филостратиграфического Возраста)

PAML - Phylogenetic Analysis by Maximum Likelihood (Филогенетический Анализ методом Максимального Правдоподобия)

TAI - Transcriptome Age Index (Индекс Возраста Транскриптома)

TDI - Transcriptome Divergence Index (Индекс Изменчивости Транскриптома)

ВЗ – Внедрение Зависимостей

ГО – Генная Онтология

ГС – Генная Сеть

Введение

Изучение эволюции молекулярно-генетических систем – одна из глобальных задач биоинформатики. Одним из видов макроэволюционного анализа является филостратиграфический анализ, предложенный в середине 2000-х годов Томиславом Домазетом-Лошо и Дитхардом Таутцем [Domazet-Lošo, Brajković, Tautz, 2007]. Целью филостратиграфического анализа является определение времени возникновения гена на основе оценки распределения ортологичных ему генов в геномах организмов, принадлежащих к различным таксономическим группам. Наряду с методами микроэволюционного анализа (например, оценкой соотношения dN/dS [Hurst, 2002]), филостратиграфические методы всё больше входят в методический арсенал эволюционных биоинформатиков. Несмотря на достигнутые успехи, ряд важных методических проблем в этой области до сих пор не решён. В частности, к настоящему моменту не существует общепризнанного метода расчёта филостратиграфического индекса, имеющего стабильную программную реализацию.

Отметим также, что до последнего времени центральным объектом молекулярной эволюции являлись отдельные гены и белки. Между тем, 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.) [Baptiste, Huneman, 2018; Ignatieva, Afonnikov, Kolchanov, 2017; Колчанов и др., 2013], таким образом, сетевой анализ начинает играть всё более важную роль в различных областях биологии [Baptiste, Huneman, 2018; Barabási, Gulbahce, Loscalzo, 2011; Kovács и др., 2019]. Среди сетей в биологии можно наблюдать очень высокое разнообразие, например, сети белок-белковых взаимодействий, регуляторные сети, сети коэкспрессии генов, метаболические и сигнальные

пути, нейронные сети, сети экологических взаимодействий, например, трофические сети и сети биологических сообществ, и многие другие. Обобщая различные способы представления данных в биологии, можно сказать, что сеть является объектом, из которого во многих случаях возможно получить полезную информацию даже с первого взгляда. Однако, использование разнородных биологических данных в комбинированных сетях может приводить к существенному затруднению их понимания. Компьютерный анализ сетей помогает выявить неочевидные, неожиданные и контринтуитивные результаты [Barabási, Gulbahce, Loscalzo, 2011; Cheng, Kovács, Barabási, 2019]. В настоящее время в свободном доступе находятся сотни ресурсов, предоставляющих сами биологические сети и средства работы с ними (<http://pathguide.org/>). В то же время, наблюдается дефицит программного обеспечения для эволюционного анализа генных сетей, что ограничивает возможности их макро- и микроэволюционного анализа.

Одним из самых многофункциональных программных комплексов для работы с сетями, в частности, биологическими, является Cytoscape (<http://apps.cytoscape.org/>) [Shannon и др., 2003]. Важное достоинство Cytoscape заключается в том, что пользователи могут расширять его функциональность за счет создания собственных подключаемых модулей, расширяя базовые возможности Cytoscape по визуализации, окрашиванию и компоновке сетей. На середину апреля 2021 года на официальном ресурсе по распространению подключаемых к Cytoscape приложений (Cytoscape App Store) содержалось 238 приложений с поддержкой Cytoscape 3.0 и 127 приложений, работающих только с предыдущей версией Cytoscape. Но только шесть из них (включая описываемый в работе Orthoscape) были доступны по запросу «evolution». При этом несколько приложений направлены на работу с базой данных KEGG (<http://www.kegg.jp/>) [Kanehisa и др., 2017], которая содержит собственную коллекцию генных сетей и метаболических путей. KEGG снабжена полезной информацией обо всех элементах генной сети, а

также реализованным механизмом API запросов, что делает ее удобным ресурсом для разработки подключаемых модулей. Три самых популярных подключаемых модуля в Cytoscape: (ClueGO [Bindea и др., 2009], BinGO [Maere, Heymans, Kuiper, 2005] и CluePedia [Bindea, Galon, Mlecnik, 2013]) осуществляют работу с данными проекта «Генная Онтология» (ГО), что говорит об актуальности работы и с этими данными.

В диссертации представлены приложения Orthoscape и Orthoweb. Orthoscape – приложение, подключаемое к Cytoscape, направленное на анализ эволюционных характеристик генов 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 наборов генов, не объединенных в генную сеть (т.е. без ребер).

Цели и задачи диссертационной работы

Цель работы:

Разработка комплекса компьютерных методов филогенетического анализа и определения уровня эволюционной изменчивости генов и генных сетей, и его применение.

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 филогенетический анализ был применен для анализа генных сетей. Впервые разработаны и реализованы программы для филогенетического анализа генных сетей, и проведен филогенетический 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, вовлеченные в те или иные процессы, являются эволюционно более древними или новоприобретенными и, в то же время, являются ли эти гены консервативными или же, наоборот, изменчивыми. Эта информация расширяет стандартный репертуар методов анализа генных сетей и позволяет выделять эволюционно близкие кластеры генов, что может представлять интерес для решения биологических задач широкого профиля. 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 научных конференциях, симпозиумах и практических курсах:

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. Диссертация состоит из введения, пяти глав, заключения и выводов, а также списка литературы. В первой главе представлен обзор литературы (охватывающий такие темы, как «эволюционные характеристики генов», «генные сети», «базы данных и знаний в биологии»). Во второй главе приведены материалы и методы, использованные в работе. Третья глава содержит описание приложений Orthoscape и Orthoweb. В четвертой главе описаны результаты анализа генных сетей заболеваний человека. В пятой главе представлены результаты анализа генов, ассоциированных с различными типами стресса у *A. thaliana*. В заключении сформулированы основные выводы из проделанной работы. Список литературы включает 139 ссылок.

Публикации

По теме диссертации опубликовано 3 статьи в рецензируемых журналах из списка ВАК, 1 авторское свидетельство, 12 тезисов международных и всероссийских конференций:

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.

Личный вклад автора

Автором были реализованы приложение Orthoscape для анализа эволюционных характеристик генных сетей, импортированных в Cytoscape, и веб-приложение Orthoweb для анализа эволюционных характеристик функционально связанных групп генов. 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.

Глава 1. Обзор литературы

Развитие компьютерных технологий в 21-ом веке существенно расширило потенциал научных исследований во всех областях биологии, в том числе дало возможность создавать компьютерные методы для анализа эволюционных процессов. Одним из видов макроэволюционного анализа, предложенного в середине 2000-х годов Томиславом Домазетом-Лошо и Дитхардом Таутцем [Domazet-Lošo, Brajković, Tautz, 2007; Tautz, Domazet-Lošo, 2011] и развивающегося по сей день, является филостратиграфический анализ. Целью филостратиграфического анализа является определение времени возникновения гена на основе оценки распределения ортологичных ему генов в геномах организмов, принадлежащих к различным таксономическим группам. Наряду с методами микроэволюционного анализа (например, оценкой соотношения dN/dS), филостратиграфические методы всё больше входят в методический арсенал эволюционных биоинформатиков.

1.1. Эволюционные характеристики генов

Понятие филостратиграфического возраста генов было впервые введено в работе [Domazet-Lošo, Brajković, Tautz, 2007]. Это макроэволюционная характеристика, основанная на анализе таксономических деревьев. В настоящее время используют как полное название, так и, для краткости, часто пишут просто возраст гена [Chen и др., 2014; D'Antonio, Ciccarelli, 2011; Domazet-Lošo, Brajković, Tautz, 2007]. Имея два ортологичных гена, с помощью сравнения срезов таксономического дерева от корня до таксона, описывающего организм, можно определить таксон, который был последним общим для организмов, чьи гены рассматриваются в анализе.

На основании найденного таксона можно определить этап, с которого гены-ортологи дивергировали. Глубина этого таксона относительно корня дерева и будет отражать возраст гена. Данное понятие не является

общепринятым в русскоязычной литературе, тем не менее, уже вводилось Евгением Куниным в книге «Логика случая», где возраст гена соответствовал самому «древнему» таксономическому узлу, в котором могут быть определены гомологи для белка, производимого данным геном [Кунин, 2014]. При работе с возрастом генов важными этапами являются анализ гомологии генов и построение филостратиграфического дерева на основе таксономического.

1.1.1 Гомология

В классической биологии гомологичными называют «органы, являющиеся производными от общего предка» [Хаубольд Б., 2011]. Для того, чтобы определить, являются ли гены гомологичными, сравниваются аминокислотные последовательности белков, кодируемых рассматриваемыми генами. Процедура сравнения состоит из выравнивания последовательностей и вычисления оценки сходства между получившимися выравненными последовательностями. Сходство последовательностей в биологии может быть выражено на основе простого вычисления доли совпавших элементов к их общему числу у выравненной последовательности, либо на основе алгоритмов с задействованием весовых матриц для различных сочетаний нуклеотидов/аминокислот, позиций триплетов, штрафов за пропуск позиции и т.д.

Ключевыми алгоритмами в данной области являются алгоритм глобального выравнивания (т.е. по всей длине последовательности) Нидлмана-Вунша [Needleman, Wunsch, 1970] и алгоритм локального выравнивания (т.е. поиск сходных участков в последовательностях) Смита-Ватермана [Smith, Waterman, 1981]. В результате работы алгоритмов исследователь получает выравнивание — две последовательности сопоставляются друг другу и, на основе этого сопоставления и матрицы весов получается итоговый «score» (результат выравнивания в виде численной характеристики). Выравнивание может содержать пропуски, образованные

процессами вставки и делеции. Участки выравнивания с пропуском позиции на одной из последовательностей называют «гэпы» (от англ. gap – разрыв). Они уменьшают score выравнивания, при чем важно не только количество разрывов, но и их длина.

Гомологичные гены, как правило, делятся на два типа – гены-ортологи и гены-паралоги (рисунок 1.1). Ортологи это гены различных организмов, разошедшиеся в процессе видообразования и зачастую выполняющие одну и ту же функцию. Паралоги это гены одного организма, разделившиеся в результате дупликации [Gerlt, Babbitt, 2000; Studer, Robinson-Rechavi, 2009; Tekaiia, 2016]. Иногда выделяют также гены-ксенологи, которые получают посредством передачи генетической информации в результате горизонтального переноса. Выявление гомологии может служить разным целям, например, определению генов, вовлеченных в различные заболевания [Dickerson и др., 2011; Kann, 2010].

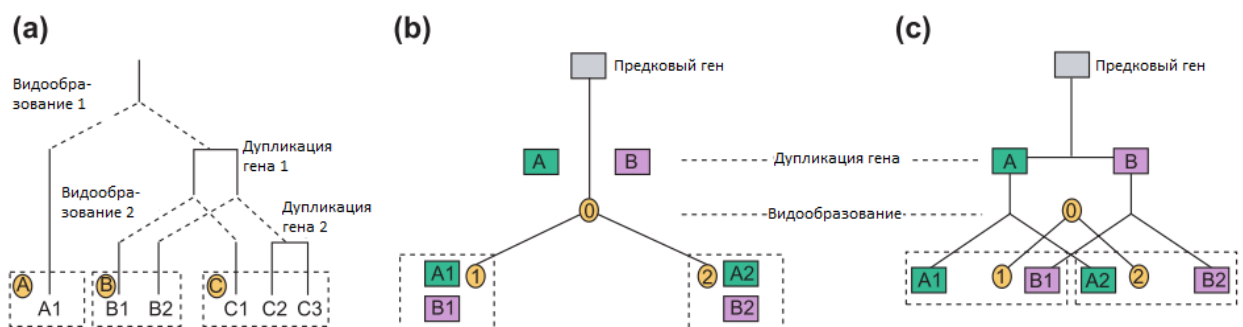


Рис. 1.1 Диаграмма основных типов гомологии, адаптированная из [Gerlt, Babbitt, 2000]. **(а)** Линии A, B и C получены с помощью видообразования. Гены A1, B1, B2, C1, C2 и C3 получены из предкового гена за счет видообразования и дупликации. **(б, в)** Образование паралогов и ортологов из предкового гена. Дупликация создает гены-паралоги A и B. После видообразования образуются две пары паралогов A1, B1 и A2, B2, и две пары ортологов A1, A2 и B1, B2.

Одним из распространенных способов поиска гомологов является поиск с помощью BLAST (Basic Local Alignment Search Tool) [Altschul и др., 1997]. Данная методика позволяет сравнивать имеющиеся

нуклеотидные/аминокислотные последовательности генов/белков с последовательностями из базы данных NCBI, выявляя в результате гены, гомологичные данному. Широкая распространенность подхода привела и к некоторой критике в его адрес. Так, на протяжении достаточно большого временного периода различными коллективами авторов отмечается особенность работы BLAST с поиском наилучшего совпадения, приводящая к некорректному результату [Koski, Golding, 2001; Shah и др., 2019], а в работе [Moyers, Zhang, 2015] в целом проанализированы частоты ошибок при поиске гомологов с помощью BLAST и отмечена высокая их частота, что стимулирует к использованию других средств для поиска гомологов. Кроме BLAST для поиска гомологов могут быть использованы, например, EggNOG [Chen и др., 2013; D'Antonio, Ciccarelli, 2011] и KEGG [Mustafin и др., 2017].

1.1.2 Филостратиграфические деревья

Анализ макроэволюционных особенностей, основываясь на определении возраста генов и поиске гена-основателя на таксономическом дереве впервые был назван «*филостратиграфическим*» (от слова филостратиграфия, в оригинале - “phylostratigraphy”) в статье Домазета-Лошо и соавторов [Domazet-Lošo, Brajković, Tautz, 2007]. Ген-основатель — это ген, дающий начало новому семейству генов (рисунок 1.2). Возникновение таких генов может быть скоррелировано с новыми функциями организма. Возникновение новых генов может произойти за счет неправильной рекомбинации, активности вирусов и транспозонов, горизонтального переноса генетического материала. При этом горизонтальный перенос затрудняет филостратиграфический анализ и не дает реальной информации об эволюционном этапе происхождения гена [Tautz, Domazet-Lošo, 2011]. Также на дереве часто находятся «орфанные гены» (в оригинале – “orphan genes”) – гены, ограниченные одной филогенетической линией. Как правило, такие гены эволюционируют быстрее других генов, но они также сильнее

подвержены исчезновению [Arendsee, Li, Wurtele, 2014; Palmieri, Kosiol, Schlötterer, 2014].

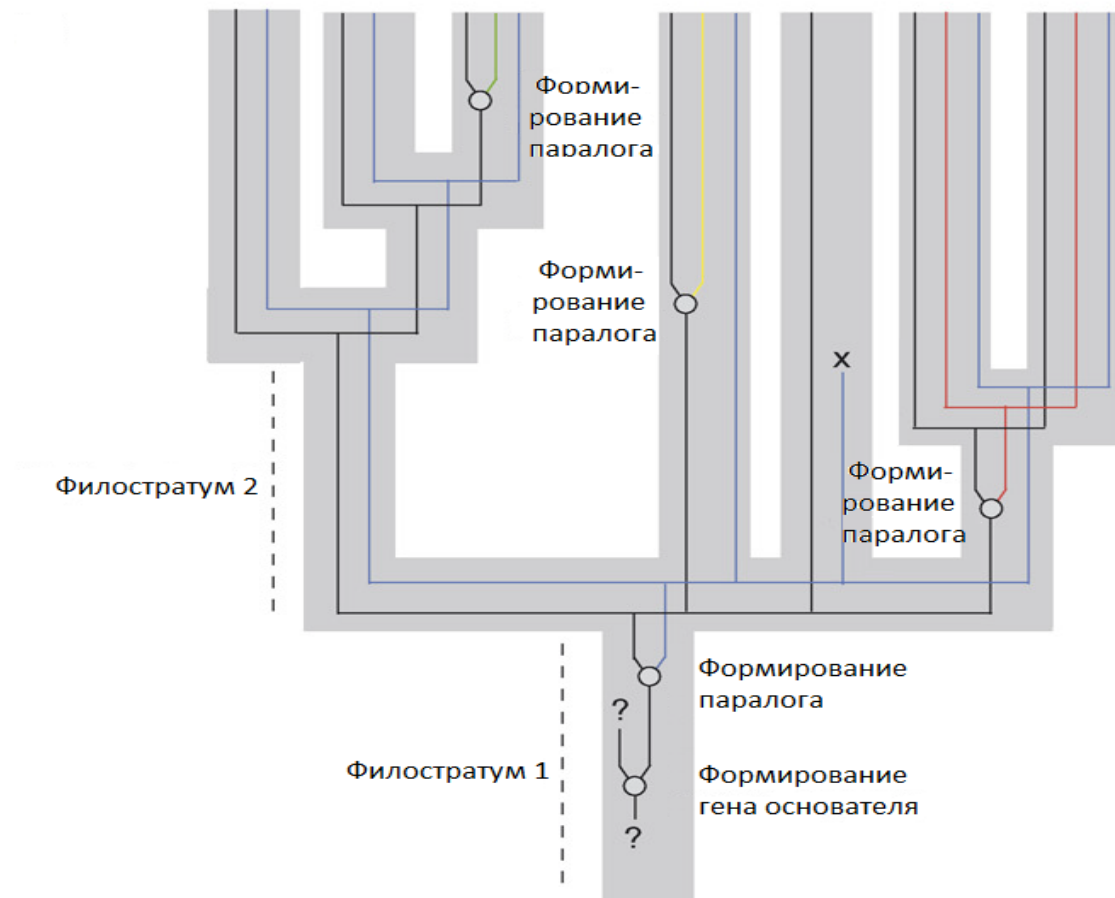


Рис. 1.2 Модель возникновения семейств белков, адаптировано из [Domazet-Lošo, Brajković, Tautz, 2007]. Серое дерево описывает филогенетические отношения между таксонами, внутренние деревья (разноцветные линии) описывают эволюцию семейств после возникновения гена-основателя. Круги – точки возникновения паралогов. “x” – потеря гена. ? – другие этапы развития, опущенные на этом изображении.

Дерево, построенное за счет выборки определенных таксонов из таксономического дерева рассматриваемого организма, называют филостратиграфическим деревом. Филостратиграфический анализ начинается с построения такого дерева. Этот этап заключается в выделении тех таксонов, которые будут учтены при анализе гомологии генов. Например, в одной из

работ по *Drosophila melanogaster* [Domazet-Lošo, Brajković, Tautz, 2007] авторами было построено филогенетическое дерево, показанное на рисунке 1.3. Таксоны на этом дереве выделены как компромисс между попыткой отразить все важнейшие события в эволюции дрозофилы, надежностью филогенетических отношений и доступными исследователям данными. В выделенном авторами дереве можно насчитать 12 таксонов (включая узел Cellular Organism, соответствующий всем клеточным организмам, в том числе бактериям и археям, который служит корнем филогенетического дерева). В следующей публикации авторами также были представлены данные по работе с дрозофилой, но на новом дереве было выделено уже 14 таксонов (рисунки 1.4). В первую очередь это изменение связано с тем, что в этой работе проводился анализ не только дрозофилы, но и рыбы *Danio rerio*, для которой на дереве было выделено 14 таксонов и для проведения сравнительного анализа дерево дрозофилы было расширено до 14 таксонов путем добавления узлов Endopterygota и Holozoa.

Таким образом, филогенетическое дерево создается для локализации самых важных для анализа таксонов. Не существует стандартного филогенетического дерева для осуществления филогенетического анализа, независимо от организма. Исследователь должен самостоятельно определять степень точности выделения таксонов в зависимости от имеющихся у него данных и прочих факторов.

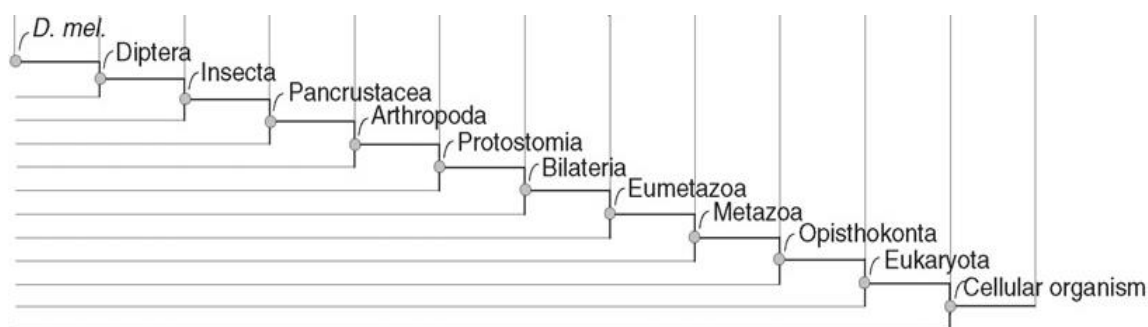


Рис. 1.3 Филогенетическое дерево для *D. melanogaster*, адаптировано из статьи [Domazet-Lošo, Brajković, Tautz, 2007].

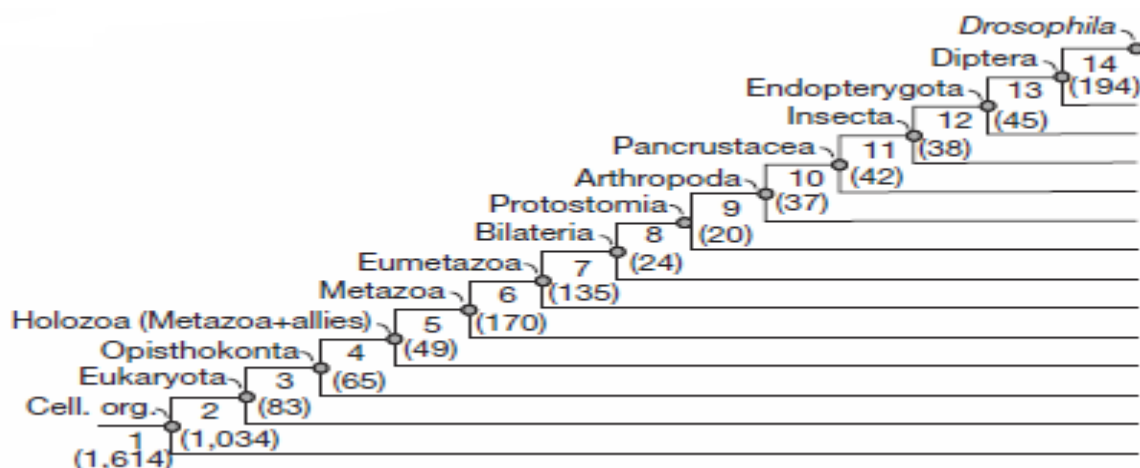


Рис. 1.4 Филостратиграфическое дерево для *D. melanogaster*, адаптировано из статьи [Domazet-Lošo, Tautz, 2010a].

1.1.3 Индексы молекулярной эволюции

Данные о степени сходства генов, выраженные в виде численного значения (score) в результате работы одного из алгоритмов сравнения последовательностей гомологичных генов, позволяют определять группы схожих друг с другом генов, между которыми можно проводить эволюционный анализ. Одним из распространенных микроэволюционных индексов, который практически не связан с таксономической классификацией, а базируется на сравнении нуклеотидных последовательностей, является dN/dS индекс [Kryazhimskiy, Plotkin, 2008; Nekrutenko, Makova, Li, 2001]. При расчете этого индекса используются нуклеотидные последовательности генов и аминокислотные последовательности белков, кодируемых рассматриваемыми генами, каждый триплет нуклеотидов кодирует определенную аминокислоту или стоп кодон в соответствии с генетическим триплетным кодом. Данный код вырожден, поскольку 64 триплета кодируют 23 позиции (22 аминокислоты и стоп кодон). Если замена нуклеотида в триплете привела к изменению кодируемой им аминокислоты, то такая замена называется несинонимической, если не привела – синонимической (рисунок 1.5).

1-е основание	2-е основание								3-е основание
	U		C		A		G		
U	UUU	Фенилаланин	UCU	Серин	UAU	Тирозин	UGU	Цистеин	U
	UUC		UCC		UAC		UGC		C
	UUA		UCA		UAA	Стоп кодон	UGA	Стоп кодон	A
	UUG		UCG		UAG		UGG	Триптофан	G
C	CUU	Лейцин	CCU	Пролин	CAU	Гистидин	CGU	Аргинин	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Глутамин	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Изолейцин	ACU	Треонин	AAU	Аспарагин	AGU	Серин	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Лизин	AGA	Аргинин	A
	AUG	Метионин	ACG		AAG		AGG		G
G	GUU	Валин	GCU	Аланин	GAU	Аспарагиновая кислота	GGU	Глицин	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Глутаминовая кислота	GGA		A
	GUG		GCG		GAG		GGG		G

Рис. 1.5 Генетический триплетный код. Если два триплета соответствуют одной аминокислоте (например, UUU и UUC), то замена, приводящая от одного такого триплета к другому, является синонимической. Если триплетам до и после замены соответствуют разные аминокислоты, то замена является несинонимической.

В dN/dS индексе числитель dN отражает число несинонимичных замен, а знаменатель dS – синонимичных. Итоговое значение индекса отражает влияние отбора на эволюцию, и в соответствии с его значением выделяется режим движущего отбора, стабилизирующего отбора или нейтральной эволюции. Как правило, значение индекса изменяется в пределах от 0 до 1, иногда наблюдается значение индекса выше 1, сигнализирующее о наличии «полезных» мутаций. Существует несколько мер/алгоритмов расчета итогового значения индекса, с учетом позиции триплетов или весовых коэффициентов. Например, в программе PAML [Yang, 2007], на данный момент используемой в большинстве публикаций по расчету индекса, используются методы Nei-Gojobori [Nei M, Gojobori T, 1986], Yang & Nielsen [Yang, Nielsen, 2000], LWL85 [Li, Wu, Luo, 1985], LWLm [Li, 1993], LPB93 [Pamilo, Bianchi, 1993].

Методы различаются по тому, как учитывается позиция триплета, веса триплетов и т.д. Например, формула расчета методом Nei-Gojobori [Nei M, Gojobori T, 1986] имеет следующий вид для синонимичных замен:

$$d_s = -(3/4) * \log_e(1 - (4/3) * p_s)$$

где $p_s = S_d/S$, S_d – сумма весов кодонов с заменами, S – среднее суммы числа синонимичных и несинонимичных сайтов. Аналогично для несинонимичных.

Другая категория индексов связана с анализом филогенетических деревьев, такие индексы относятся к группе филогенетических [Domazet-Lošo, Tautz, 2010a]. Один из них – TAI (transcriptome age index, индекс возраста транскриптома) показывает связь возраста гена с уровнем его экспрессии. TAI может быть вычислен по формуле:

$$TAI_s = \frac{\sum_{i=1}^n p_{s_i} e_i}{\sum_{i=1}^n e_i}$$

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

Данный индекс применялся для анализа *D. rerio*, *D. melanogaster*, *A. thaliana* [Domazet-Lošo, Tautz, 2010a; Quint и др., 2012]. В работе [Domazet-Lošo, Tautz, 2010a] с помощью TAI устанавливается связь между филогенией и онтогенезом и впервые описывается паттерн «песочных часов». Данный паттерн характеризуется высоким показателем TAI на начальных и конечных этапах развития организма и низким на промежуточном этапе (рисунок 1.6), т.е. TAI показывает, что на начальной/конечной стадиях онтогенеза работают «эволюционно молодые» гены, а на промежуточной – «эволюционно древние».

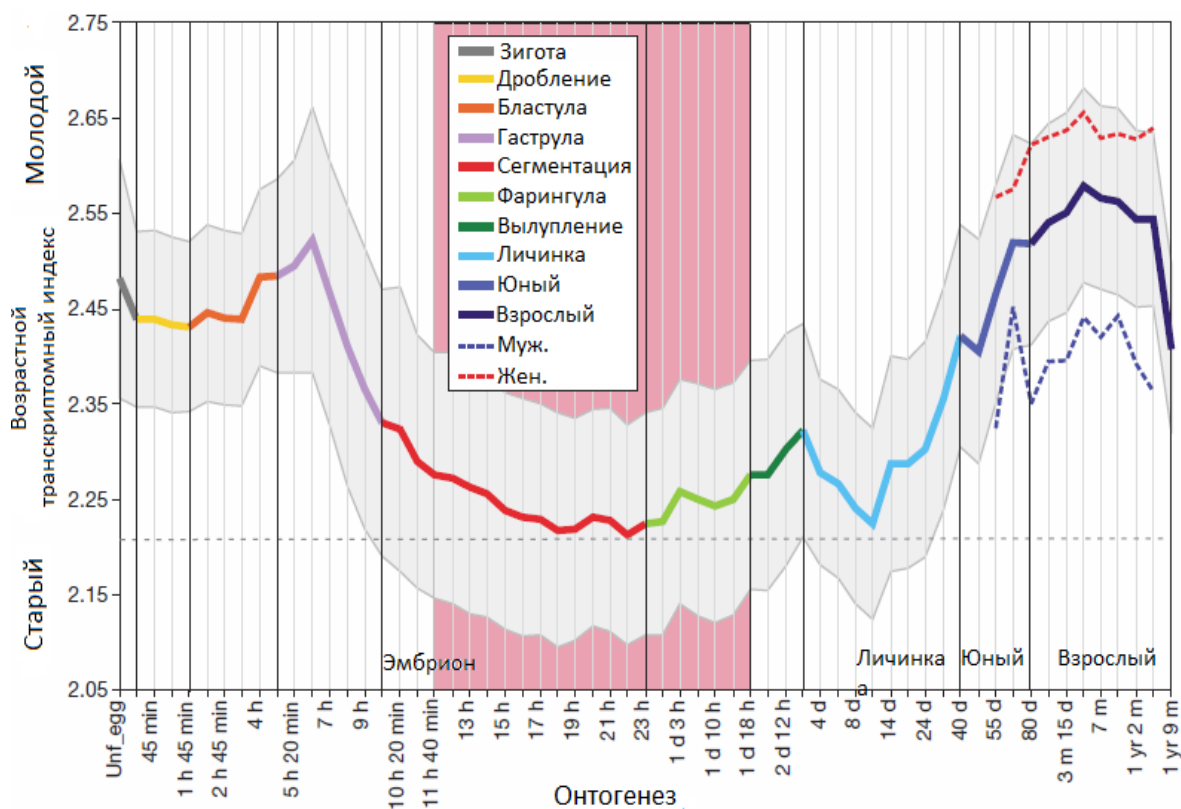


Рис. 1.6 Паттерн песочных часов, получен при расчете индекса TAI на разных этапах развития *D. rerio*, адаптировано из [Domazet-Lošo, Tautz, 2010a].

В работе [Quint и др., 2012] применили филостратиграфический анализ для исследования растений, а именно для анализа *A. thaliana*. В этой работе используются два дополняющих друг друга подхода, основанных на анализе макроэволюционных характеристик (TAI, transcriptome age index) и микроэволюционных характеристик (TDI, transcriptome divergence index) показывающих, что и для эволюции *A. thaliana* характерен паттерн «песочных часов». По аналогичной методике авторами этой работы позднее также был получен паттерн песочных часов для *D. rerio* и *D. melanogaster* [Drost и др., 2015].

Другой интересной задачей, решаемой с помощью филостратиграфического анализа, является поиск функциональных особенностей генов, различающихся по эволюционному возрасту. Есть данные о том, что в фундаментальных процессах в клетках задействованы

более «древние» гены. Например, в работе [Domazet-Loso, Tautz, 2008] показано, что наиболее древние гены человека (ортологичные с ними гены находятся у других эукариот и бактерий), в основном ассоциированы с базовыми клеточными функциями (метаболические процессы, регуляция транскрипции), тогда как гены, возникшие на более поздних стадия эволюции, ассоциированы с генами иммунного ответа и размножения. В работе [Zhang и др., 2019] с помощью филогенетического анализа были проанализированы генные сети, выявлены важные детали их эволюции и найдены функциональные модули. Такой подход был использован для крупномасштабного анализа эволюции сетей коэкспрессии генов *A. thaliana* [Ruprecht и др., 2017]. Авторы показали, что гены, происходящие в одном и том же эволюционном периоде, имеют тенденцию быть связанными между собой, при этом древние и молодые гены, наоборот, не связаны друг с другом.

Несмотря на достигнутые успехи, ряд важных методических проблем в этой области до сих пор не решён. В частности, к настоящему моменту не существует общепризнанного метода расчёта филогенетического индекса, имеющего стабильную программную реализацию. Кроме того, в большинстве работ объектами гипотез становятся именно гены, безотносительно их связей между собой, в то время как интересным представляется не только анализ эволюционных характеристик генов, но и того, как связаны гены с различными значениями этих характеристик.

1.2. Генные сети

До последнего времени центральным объектом молекулярной эволюции являлись отдельные гены и белки. Между тем, формирование фенотипических признаков, обеспечивающих адаптацию организмов к условиям окружающей среды, контролируется не отдельными генами, а генными сетями - группами координированно-функционирующих генов, взаимодействующих друг с другом как через свои первичные продукты (РНК и белки), так и через разнообразные метаболиты и другие вторичные продукты функционирования

генных сетей [Колчанов и др., 2013]. Изначально понятие генной сети зародилось в науке еще в 60-е годы прошлого века [Kauffman, 1969; Ратнер В.А., 1966], но широкое распространение получило только в начале 2000-х годов. В работе с сетями перекликаются математические дисциплины (теория графов, комбинаторика, теория вероятностей), информатика (алгоритмы генерации сетей, оптимизация комбинаторных алгоритмов) и биология [Lesne, 2006]. В биологии сетевая организация элементов и их взаимодействий является наглядным, удобным и в то же время емким способом представления данных [Alon, 2003; Gehlenborg и др., 2010]. Существует множество различных типов сетей биологических данных, как правило, классифицированных по своему функциональному назначению. Одни сети описывают белок-белковые взаимодействия, другие – регуляцию экспрессии генов [Hakes и др., 2008]. Различные типы элементов и взаимодействий сочетаются в метаболических и сигнальных путях. Сетевую архитектуру используют, в том числе, и для выражения связей и отношений в популяции, будь то трофические взаимодействия (пищевые цепи), сети взаимодействия представителей сообществ на экологическом уровне и т.д. [Bascompte, Melián, 2005; Dunne, Williams, Martinez, 2002; Kéfi и др., 2015]. Генная сеть, позволяет выделять центральные гены и белки, которые являются ключевыми и координируют работу остальных генов сети. Как правило, такие гены отличаются увеличенным числом связей, а также выделяются своей позицией в топологии сети, что может быть видно благодаря её удобному визуальному представлению. Использование генных сетей позволило исследователям, во-первых, строить гипотезы еще до осуществления компьютерного анализа, основываясь на удобном визуальном представлении данных, а во-вторых, непосредственно проводить компьютерный анализ сетей, которые легко формализуются в виде списка взаимодействий входящих в них сущностей.

В настоящее время получение новых данных зачастую идет быстрее, чем обработка уже имеющихся, и возникает проблема анализа этих данных и

структурированного представления результатов этого анализа [Chen и др., 2014; Sonnhammer и др., 2014; Несговорова Г.П., 2012]. Реконструкция генных сетей является одной из задач, которая дала толчок развитию новых компьютерных методик и программных комплексов в этой области [Несговорова Г.П., 2012]. Сейчас сети могут быть реконструированы вручную с помощью многофункциональных программных комплексов (таких как Cytoscape [Shannon и др., 2003]), могут быть воссозданы автоматически на основе знаний о взаимодействиях элементов, извлеченных из научных публикаций (например, с помощью систем ANDSystem [Ivanisenko и др., 2015] или Pathway Studio [Nikitin и др., 2003]), а также на основе данных о взаимодействии конкретных генов и белков, взятых из баз биологических данных (например, с помощью приложений GeneMANIA [Montejo и др., 2010] и String [Szklarczyk и др., 2019]).

После реконструкции сети естественным образом возникает задача её анализа. В рамках этой задачи используются разные подходы. Один из них - анализировать элементы сети, тем или иным образом учитывая связи между ними и топологию сети. В таком случае вывод о характеристиках сети, заключаемый по результату анализа, опирается на анализ входящих в нее элементов. Другой подход – анализировать эволюцию генной сети с помощью реконструкции элементов и их взаимодействий на различных временных этапах. В этом случае вывод делается не на основе результатов анализа отдельных элементов, а именно на основе их взаимодействия. Для анализа сетей и входящих в них элементов, для визуализации готовых сетей на основе широкого выбора возможностей компоновки, окрашивания и т.д. создаются компьютерные приложения. Одним из таких является Cytoscape (<http://www.cytoscape.org/>) - платформа для визуализации сложных генных сетей, расширяемая за счет пользовательских приложений. На данный момент Cytoscape является крайне популярной и при этом активно развивающейся платформой, как самим разработчиками, так и пользовательским

сообществом, которым написано уже несколько сотен подключаемых приложений.

Самыми популярными приложениями к Cytoscape на данный момент являются ClueGO [Bindea и др., 2009] (<http://apps.cytoscape.org/apps/cluego>, для создания и визуализации сетей на основе словарей генной онтологии (ГО) или метаболических путей) и BiNGO [Maere, Heymans, Kuiper, 2005] (<http://apps.cytoscape.org/apps/bingo>, для расчета перепредставленности терминов ГО и создания сети со значимыми словарями. Классификация генов по терминам ГО возникла в связи с ростом объемов биологических данных, чтобы исследователь имел способы выбора конкретных данных для анализа конкретных процессов. Классификация ГО набрала большую популярность [Gene Ontology Consortium, 2004; Gene Ontology Consortium, 2015]. Гены в ГО объединяются по принципу функциональной близости, но не имеют явно выраженных связей между собой. В ГО реализовано три словаря – «молекулярные функции», «биологические процессы», «клеточные компоненты». Каждый отдельный словарь состоит из множества поднаборов (называемых терминами), например, в словаре «биологические процессы» содержится термин «биологическая регуляция». Термины, в свою очередь, сами состоят из терминов, образуя таким образом древовидную структуру от одного корня с большим числом генов до множества листьев с меньшим числом генов. Два термина могут быть связаны различными отношениями, такими как «является частным случаем», «является частью», «включает», «позитивно/негативно регулирует», «встречается при». Проект активно развивается, создаются браузеры для работы со структурированными данными, например, AmiGO (рисунок 1.7).

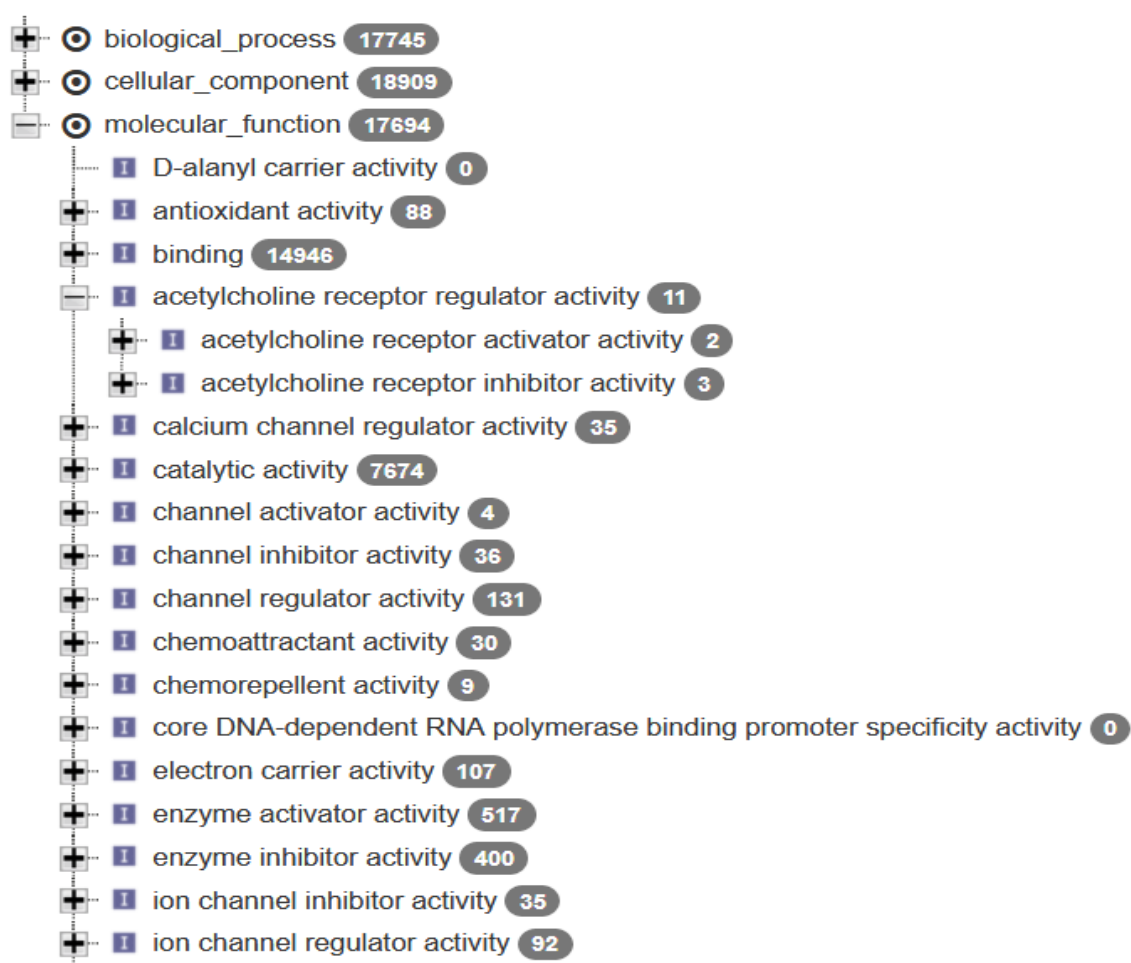


Рис. 1.7 Фрагмент словаря «молекулярные функции» на первом уровне погружения от корня древовидной структуры. Включен фильтр по организму *Homo sapiens*.

Некоторые приложения для построения генных сетей реализованы как в Cytoscape, так и вне его. Одним из таких приложений является GeneMANIA [Montojo и др., 2010] (<http://apps.cytoscape.org/apps/genemania>, также входит в четверку популярных Cytoscape приложений, после ClueGO, BinGO и CluePedia), с помощью которой можно по заданному списку генов импортировать из открытых баз данные о взаимодействиях между этими генами, как прямых, так и через гены посредники, конструируя в результате импорта генную сеть. Связи между генами реконструируются на основе данных из таких ресурсов, как BIOGRID [Chatr-aryamontri и др., 2017], GEO [Barrett и др., 2009], I2D [Brown, Jurisica, 2005] и Pathway Commons [Cerami и др., 2009].

др., 2011]. Приложение воспринимает идентификаторы и синонимы из Entrez, Ensembl, RefSeq, TAIR и UniProt.

Для работы с генными сетями в Cytoscape создаются также полезные приложения не для создания, но для импорта и конвертации сетей из других источников. Существует несколько приложений для импорта сетей, некоторые из них сконцентрированы на базе KEGG. Импортируются как предварительно загруженные сети в формате kgml, так и загруженные в онлайн-режиме (cyKEGGParser (<http://apps.cytoscape.org/apps/cykeggparser>) [Nersisyan, Samsonyan, Arakelyan, 2014], CytoKegg (<http://apps.cytoscape.org/apps/cytokegg>), KEGGscape). Доступны и приложения для импорта из других баз, например CyPath2 (<http://apps.cytoscape.org/apps/cypath2>) позволяет импортировать сети из базы PathwayCommons в формате BioPAX (level 3) [Cerami и др., 2011; Demir и др., 2010].

CyKEGGParser осуществляет адаптацию данных из сети, подходящих для нескольких организмов, с учетом выбранного исследователем организма (например, раскрывает kegg orthology группы и подставляет в сеть ген конкретного организма вместо целой группы), импортирует сети в формате KGML и позволяет сохранять сети в форматах KGML и BioPAX (level2, level3). На рисунке 1.8 можно видеть пример генной сети, взятой из базы данных KEGG Pathway, категория Human Diseases, сеть Parkinson's disease (https://www.genome.jp/kegg-bin/show_pathway?hsa05012), которую можно импортировать с помощью CyKEGGParser. Данная сеть реконструирована экспертами вручную, на ней можно видеть не только привычные элементы сети и их взаимодействия, но и их расположение с учетом их функциональности. Например, в правом нижнем углу можно видеть гены, объединенные в блок "Mitochondria". Сети, реконструированные автоматически на основе анализа публикаций или записей в базах данных, могут быть легко получены исследователем с помощью разработанного

программного обеспечения, но, как правило, такие сети не отличаются такой проработанностью, как сети, реконструированные экспертами вручную.

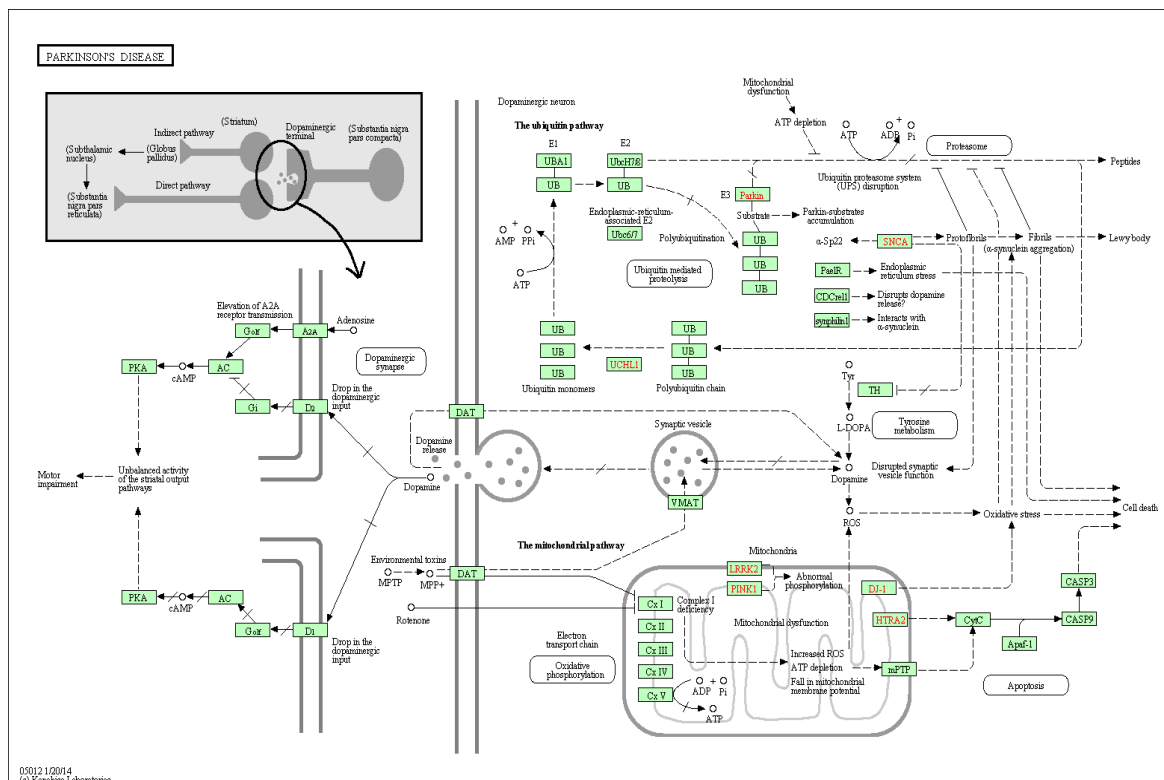


Рис. 1.8 Генная сеть болезни Паркинсона из базы данных KEGG (https://www.genome.jp/kegg-bin/show_pathway?hsa05012). Зелеными прямоугольниками обозначены гены, маленькими белыми кругами – химические соединения, большими белыми скругленными прямоугольниками – ссылки на другие пути.

1.3. Базы данных и знаний в биологии

Современная биология уже прошла тот исторический этап, когда получение данных было лимитирующим фактором в выполнении работ (<https://www.ncbi.nlm.nih.gov/genbank/statistics/>). Благодаря развитию компьютерных и интернет технологий, объемы данных, как уже проанализированных, так и еще только ожидающих анализа, достигли колоссальных размеров. Для структурирования всего многообразия данных были организованы консорциумы, например, NCBI в США и EMBL в Европе. Одними из наиболее универсальных и часто упоминаемых в других работах

являются такие базы, как Genbank [Benson и др., 2018], Uniprot [O'Donovan и др., 2002], Ensembl [Zerbino и др., 2018], KEGG [Kanehisa и др., 2017]. Базы данных, как правило, разбиваются на первичные, в которых содержатся данные, полученные из экспериментов, и вторичные – в них помещаются компьютерно-обработанные данные первичных баз [Афонников, Миронова, 2014]. Некоторые базы содержат и экспериментальные данные, и результаты компьютерной обработки. Базы могут содержать в себе информацию по любым организмам, а могут быть сконцентрированы на конкретных представителях. Модельными организмами являются *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* и *Saccharomyces cerevisiae*. Некоторые базы хранят информацию только о конкретных типах генетических объектов, как правило, это гены или белки. Некоторые пробуют сочетать в себе максимально разнородную информацию. Наиболее подробную информацию о современном состоянии баз данных, а также о динамике их развития, можно увидеть в серии ежегодных публикаций от Nucleic Acids Research, последняя из которых это работа [Rigden, Fernández, 2020].

Иногда в базах данных разделяется информация, полученная современными компьютерными методами и экспертами вручную. Примером такой базы служит UniProt [O'Donovan и др., 2002]. В UniProt данные, полученные экспертами вручную, без задействования компьютерной техники, содержатся в разделе UniProt/Swiss-Prot, а данные, полученные с помощью компьютерной техники и дополненные с помощью компьютерной аннотации в разделе UniProt/TrEMBL. При этом данные в TrEMBL также делятся на две категории – для данных в SP-TrEMBL планируется перенос в Swiss-Prot, а для данных REM-TrEMBL – нет. Вручную созданные аннотации белка проходят контроль качества перед включением в базу.

Отдельную нишу занимает ресурс PubMed (<https://www.ncbi.nlm.nih.gov/pubmed/>). PubMed это база данных, содержащая

в себе информацию о медицинских и биологических публикациях. По своей функциональности PubMed стоит в одном ряду с такими многофункциональными сервисами по поиску публикаций, как Google Scholar, Scopus и Web of science. Более того, показано, что в качестве сервиса по поиску именно биомедицинских публикаций, а также в поиске по ключевым словам, PubMed в этом списке играет наиболее важную роль [Falagas и др., 2008].

1.3.1. KEGG

Kyoto Encyclopedia of Genes and Genomes (<http://www.genome.jp/kegg/>) – ресурс, разработанный в Японии, в котором содержится биологическая информация самых разных категорий [Kanehisa и др., 2017]. KEGG является одним из самых известных представителей современных ресурсов, в которых содержится максимально разнородная информация, как с точки зрения разнообразия представленных организмов, так и с точки зрения типа описывающих эти организмы данных. Здесь можно найти вручную реконструированные сети (KEGG Pathway), группы ортологичных генов (KEGG Orthology), подробную информацию о генах со ссылками на прочие ресурсы (KEGG Genes) и т.д. (рисунок 1.9).

Entry	7077 CDS T01001
Gene name	TIMP2, CSC-21K, DDC8
Definition	(RefSeq) TIMP metalloproteinase inhibitor 2
Organism	hsa Homo sapiens (human)
SSDB	Ortholog Paralog GFIIT
Motif	Pfam: TIMP NTR Motif
Other DBs	NCBI-GeneID: 7077 NCBI-ProteinID: NP_003246 OMIM: 188825 HGNC: 11821 Ensembl: ENSG00000035862 Vega: OTTHUMG00000154517 Pharos: P16035 (Tbio) UniProt: P16035 A0A140VK57
LinkDB	All DBs
Structure	PDB: 4ILW 1BR9 1GXD 2TMP Thumbnail  Jmol
Position	17q25.3
AA seq	220 aa AA seq DB search MGAAARTLRLALGLLLATLLRPADACSCSPVHPQQAFCNADVIRAKAVSEKEVDSGND IYGNPIKRIQYEIKQIKMFKGPEKDIEFIYTAPSSAVCGVSLDVGGKKEYLIAGKAEGDG KMHITLCDFIIVPWTLSLTQKSLNHRVQMGCECKITRCPMIPCYISSPDECLWMDWVTE KNINGHQAKFFACIKRSDGSCAWYRGAAPPKQEFLDIEDP
NT seq	663 nt NT seq atggggcgccgcccgcacccctgcccgtggcgctcgggcctcctgctgctggcgacgctg cttcgccccggcgacgctgcagctgctccccggtgcacccgcaacaggcggttttgcaat gcagatgtagtgatcagggccaaagcggtcagtgagaaggaagtggactctggaaacgac atztatggcaaccctatcaagaggatccagtatgagatcaagcagataaagatgttcaaa gggcctgagaaggatatagagtttatctacacggccccctcctcggcagtggtggtggtc tcgctggacggttgaggaaagaaggaatatctcattgcaggaaaggccgagggggacggc aagatgcacatcacccctctgtgacttcacgtgcctgggacacccctgagcaccacccag aagaagagcctgaaccacaggtaccagatgggctgcgagtgcaagatcacgcgctgcccc atgatcccggtgctacatctcctccccggacgagtgccctctggatggactgggtcacagag aagaacatcaacgggcaccaggccaagttcttcgctgcatcaagagaagtgcgggtcc tgtgcgtggtaccgcggcgccggcggcccccaagcaggagtttctcgacatcgaggaccca taa

Рис. 1.9 Пример карточки гена в KEGG (https://www.kegg.jp/dbget-bin/www_bget?hsa:7077).

Сети, представленные в KEGG, хранятся в собственном формате – KGML (KEGG Markup Language). Элементами сети могут быть гены, химические соединения (нуклеозиды, нейромедиаторы и т.д.), ортологичные группы, метаболические пути. Для удобного доступа и загрузки информации реализован API, а для визуализации генных сетей в рамках ресурса реализована система покраски прямо в браузере.

1.3.2. Ensembl

Для работы с данными позвоночных создан многофункциональный ресурс Ensembl [Flicek и др., 2012; Zerbino и др., 2018]. Ensembl позиционирует себя, как геномный браузер для работ по сравнительной геномике, эволюции, изменчивости последовательностей и регуляции транскрипции. Наиболее популярными и хорошо описанными являются такие организмы, как *D. rerio*, *H. sapiens*, *M. musculus*. С помощью браузера можно посмотреть экспрессию генов в различных тканях, нуклеотидные последовательности генов, найти в них однонуклеотидные полиморфизмы или найти гены ортологи. С помощью раздела Biomart можно загрузить исходные данные базы в виде таблицы с огромными вариантами идентификации и описания гена. Ensembl/Biomart на данный момент является одним из самых удобных словарей биологической информации, с помощью которого можно перевести один вид биологического идентификатора в другой (рисунок 1.10).

External References (max 3)

- | | |
|--|--|
| ☐ CCDS ID | ☐ MIM morbid accession |
| ☐ ChEMBL ID | ☐ miRBase ID |
| ☐ Clone-based (Ensembl) gene ID | ☐ miRBase accession |
| ☐ Clone-based (Ensembl) transcript ID | ☐ miRBase transcript name ID |
| ☐ DataBase of Aberrant 3' Splice Sites name | ☐ NCBI gene ID |
| ☐ DataBase of Aberrant 3' Splice Sites ID | ☐ PDB ID |
| ☐ DataBase of Aberrant 5' Splice Sites name | ☐ Reactome ID |
| ☐ DataBase of Aberrant 5' Splice Sites ID | ☐ Reactome gene ID |
| ☐ Ensembl Human Transcript ID | ☐ Reactome transcript ID |
| ☐ Ensembl Human Translation ID | ☐ RefSeq mRNA ID |
| ☐ EntrezGene transcript name ID | ☐ RefSeq mRNA predicted ID |
| ☐ European Nucleotide Archive ID | ☐ RefSeq ncRNA ID |
| ☐ Expression Atlas ID | ☐ RefSeq ncRNA predicted ID |
| ☐ GeneDB ID | ☐ RefSeq peptide ID |
| ☐ HGNC symbol | ☐ RefSeq peptide predicted ID |
| ☐ HGNC ID | ☐ RFAM ID |
| ☐ HGNC transcript name ID | ☐ RFAM transcript name ID |
| ☐ Human Protein Atlas ID | ☐ RNAcentral ID |
| ☐ INSDC protein ID | ☐ UCSC Stable ID |
| ☐ KEGG Pathway and Enzyme ID | ☐ UniGene ID |
| ☐ LRG display in Ensembl gene ID | ☐ UniParc ID |
| ☐ LRG display in Ensembl transcript ID | ☐ UniProtKB Gene Name ID |
| ☐ MEROPS - the Peptidase Database ID | ☐ UniProtKB/Swiss-Prot ID |
| ☐ MetaCyc ID | ☐ UniProtKB/TrEMBL ID |
| ☐ MIM gene description | ☐ WikiGene description |
| ☐ MIM gene accession | ☐ WikiGene name |
| ☐ MIM morbid description | ☐ WikiGene ID |

Рис. 1.10 Фрагмент выбора фильтров в Biomart. Здесь указаны идентификаторы из внешних баз, которые можно запросить для генов человека.

1.3.3. TAIR

TAIR (The Arabidopsis Information Resource) – база данных, сконцентрированная на *A. thaliana* [<https://www.arabidopsis.org/>][Lamesch и др., 2012]. Информация, которую можно найти в базе, описана на рисунке 1.11.

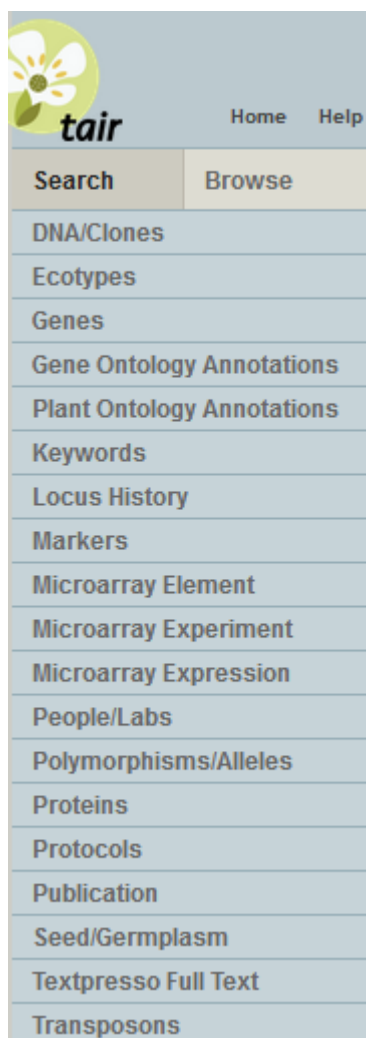


Рис. 1.11 Фрагмент меню TAIR, опция поиска. Во всплывающем окне описан обширный набор данных об *A. thaliana*.

В данные TAIR включены полная геномная последовательность, структура генов, экспрессия генов, продукты генов, ДНК, семенные запасы, геномные карты, генетические и физические маркеры, публикации, информация о сообществе исследователей, занимающихся арабидопсисом. Данные обновляются каждую неделю на основе последних публикаций и запросов сообщества.

1.3.4. DAVID

DAVID (The Database for Annotation, Visualization and Integrated Discovery) – база знаний, предоставляющая инструменты для определения функциональных особенностей генов [https://david.ncifcrf.gov/home.jsp][Huang, Sherman, Lempicki, 2009]. На основании списка генов, предоставленного исследователем, DAVID может: определить и кластеризовать обогащенные термины ГО, выделить обогащенные функционально связанные группы генов, визуализировать гены на путях из BioCarta и KEGG pathway, визуализировать отношения “много генов – много терминов” на 2-D изображении, искать функционально связанные гены с генами из исходно представленного списка, конвертировать идентификаторы генов, находить ассоциации генов с болезнями, выделять функциональные белковые домены и мотивы, находить связанные с генами научные публикации и др.

65 chart records [Download File](#)

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	GOTERM_BP_DIRECT	response to cold	RT		54	76,1	1,4E-82	3,6E-80
☐	GOTERM_BP_DIRECT	cold acclimation	RT		28	39,4	1,7E-54	2,1E-52
☐	GOTERM_BP_DIRECT	response to salt stress	RT		23	32,4	1,7E-18	1,5E-16
☐	GOTERM_BP_DIRECT	response to water deprivation	RT		19	26,8	7,8E-18	5,1E-16
☐	GOTERM_BP_DIRECT	response to abscisic acid	RT		17	23,9	9,1E-13	4,7E-11
☐	GOTERM_BP_DIRECT	response to osmotic stress	RT		11	15,5	2,6E-11	1,1E-9
☐	GOTERM_BP_DIRECT	cellular response to cold	RT		5	7,0	5,2E-6	1,9E-4
☐	GOTERM_BP_DIRECT	response to water	RT		4	5,6	1,1E-5	3,5E-4
☐	GOTERM_BP_DIRECT	regulation of seed germination	RT		5	7,0	1,5E-5	4,2E-4
☐	GOTERM_BP_DIRECT	leaf senescence	RT		6	8,5	2,4E-5	6,2E-4
☐	GOTERM_BP_DIRECT	response to freezing	RT		4	5,6	6,3E-5	1,5E-3
☐	GOTERM_BP_DIRECT	DNA duplex unwinding	RT		4	5,6	7,3E-5	1,6E-3
☐	GOTERM_BP_DIRECT	response to cadmium ion	RT		8	11,3	2,8E-4	5,6E-3
☐	GOTERM_BP_DIRECT	circadian rhythm	RT		5	7,0	4,7E-4	8,7E-3

Рис. 1.12 Пример результата DAVID по поиску наиболее представленных терминов геной онтологии в исходном наборе генов.

1.3.5. STRING

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) – база данных с информацией об известных и предсказанных белок-белковых взаимодействиях [<https://string-db.org/>][Szklarczyk и др., 2017]. В базу включены физические и функциональные ассоциации. STRING также позволяет выделять функциональные обогащения в представленных списках белков с использованием таких систем классификации, как GO, Pfam (Protein families), KEGG и др. (Рис. 1.13). Возможности STRING позволяют реконструировать сети на основе имеющихся списков генов, задавая уровень уверенности в связях, которые ищутся между генами и становятся ребрами сети. STRING позволяет пользователю добавить интересующее его число генов, не входящих в исходный список, на основании имеющихся в базе данных о взаимодействиях между ними, чтобы расширить исходную сеть. Два гена в STRING могут быть объединены ребром или ребрами, обозначающими одно или более из восьми следующих отношений, их связывающих: упоминаются в литературе, найдены в базах данных, получены из эксперимента, коэкспрессируются, гомологичны, совпадает филогенетический профиль, кодируют гибридный белок, сходятся по числу нуклеотидов в последовательностях, кодируемых генами.

Cellular Component (GO)			
<i>GO-term</i>	<i>description</i>	<i>count in gene set</i>	<i>false discovery rate</i>
GO:0005622	intracellular	65 of 10570	1.74e-18
GO:0044424	intracellular part	63 of 10448	8.45e-17
GO:0044464	cell part	66 of 12106	1.56e-16
GO:0043229	intracellular organelle	56 of 9362	4.93e-13
GO:0043227	membrane-bounded organelle	54 of 9036	3.00e-12
(more ...)			
Reference publications			
<i>publication</i>	<i>(year) title</i>	<i>count in gene set</i>	<i>false discovery rate</i>
PMID:23466881	(2013) Cold signaling and cold response in plants.	14 of 62	6.49e-19
PMID:28533787	(2017) Genome-Wide Analysis of Gene Expression Provides...	12 of 44	4.09e-17
PMID:20387039	(2010) Gene regulation during cold stress acclimation in pl...	13 of 64	4.09e-17
PMID:21886453	(2011) Engineering cold stress tolerance in crop plants.	12 of 49	9.29e-17
PMID:25479236	(2014) RNA-seq analysis of oil palm under cold stress reve...	9 of 17	1.30e-14
(more ...)			
KEGG Pathways			
<i>pathway</i>	<i>description</i>	<i>count in gene set</i>	<i>false discovery rate</i>
ath04712	Circadian rhythm - plant	4 of 36	0.00011
ath04016	MAPK signaling pathway - plant	5 of 127	0.00039
ath03040	Spliceosome	5 of 184	0.0014
ath03018	RNA degradation	4 of 112	0.0019
ath04070	Phosphatidylinositol signaling system	3 of 70	0.0060
(more ...)			
Reactome Pathways			
<i>pathway</i>	<i>description</i>	<i>count in gene set</i>	<i>false discovery rate</i>
ATH-72165	mRNA Splicing - Minor Pathway	4 of 78	0.0062
ATH-8953854	Metabolism of RNA	7 of 569	0.0084
ATH-881907	Gastrin-CREB signalling pathway via PKC and MAPK	2 of 14	0.0084
ATH-72163	mRNA Splicing - Major Pathway	5 of 194	0.0084
ATH-450294	MAP kinase activation	3 of 47	0.0084
(more ...)			
UniProt Keywords			
<i>keyword</i>	<i>description</i>	<i>count in gene set</i>	<i>false discovery rate</i>
KW-0346	Stress response	25 of 520	7.31e-23
KW-0694	RNA-binding	16 of 701	1.79e-09
KW-0597	Phosphoprotein	23 of 1757	1.86e-09
KW-0539	Nucleus	31 of 3725	1.80e-08
KW-0963	Cytoplasm	21 of 1842	1.13e-07
(more ...)			
PFAM Protein Domains			
<i>domain</i>	<i>description</i>	<i>count in gene set</i>	<i>false discovery rate</i>
PF00257	Dehydrin	4 of 8	1.55e-06
PF00098	Zinc knuckle	5 of 32	1.59e-06
PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	8 of 243	7.29e-06
PF00313	'Cold-shock' DNA-binding domain	3 of 4	1.12e-05
PF00512	His Kinase A (phospho-acceptor) domain	3 of 11	9.19e-05
(more ...)			

Рис. 1.13 Пример (фрагмент результата) STRING по обогащенности представленными генами в различных системах.

1.4. Заключение к главе 1

Анализ эволюционных характеристик генов потенциально может позволить построить картину развития различных процессов в организме на основании полученных данных о генах, вовлеченных в генные сети, эти процессы описывающие. В частности, анализ такой макроэволюционной характеристики, как возраст гена, позволяет определить последнего общего предка с помощью анализа ортологов гена. В то же время, анализ такой микроэволюционной характеристики, как отношение dN/dS , позволяет понять, какому типу отбора данный ген подвергается, на основании сравнения последовательности кодируемого им белка с последовательностями белков, кодируемыми его ортологами у эволюционно ближайших родственных организмов. Данные два метода вместе позволяют определить, какие гены в тех или иных процессах являются наиболее 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.

Глава 2. Материалы и методы

В работе представлен Orthoscape – подключаемое приложение к программному комплексу Cytoscape. Orthoscape написан на языке Java 8, 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>). Графический интерфейс реализован с помощью библиотеки swing, результаты анализа визуализируются помощью бесплатной сторонней библиотеки JFreeChart (<http://www.jfree.org/jfreechart/>) и скриптов на языке программирования R (<https://www.r-project.org/>).

Для выравнивания нуклеотидных последовательностей в Orthoscape была адаптирована реализация алгоритма Нидлмана-Вунша с ресурса (<https://zhanglab.ccmb.med.umich.edu/NW-align/>). Для расчета dN/dS индекса на основе выравненных последовательностей используется пакет PAML (<http://abacus.gene.ucl.ac.uk/software/paml>), который может быть запущен напрямую из Orthoscape. В процессе работы Orthoscape обращается к базе данных KEGG для получения следующей информации: гомологи рассматриваемых генов и соответствующие им значения идентичности последовательностей, результат (score) работы алгоритма Смита-Ватермана, название гена, полная таксономическая принадлежность, согласно классификации KEGG, домен (данные в KEGG дублируют данные Pfam с соответствующими ссылками), нуклеотидные и аминокислотные последовательности. Информация получается с помощью API (<https://www.kegg.jp/kegg/rest/keggapi.html>) на основе идентификаторов генов.

Для получения анализируемых в работе генных сетей использовалась база KEGG Pathway, раздел Human Diseases. Для получения данных ГО и формирования на их основе древовидной структуры использовался веб-браузер AmiGO (http://amigo.geneontology.org/amigo/dd_browse). Для связи идентификаторов генов из KEGG с данными ГО и названиями генов использовался ресурс Ensembl/Biomart (<http://ensembl.org/biomart>). Для получения списка генов *A. thaliana*, ассоциированного с конкретными терминами генной онтологии, использовался ресурс TAIR (<https://www.arabidopsis.org/>). Для определения обогащенности конкретными терминами ГО по заданному списку генов использовался ресурс DAVID (<https://david.ncifcrf.gov/>). Для определения связей между генами при реконструкции генных сетей использовался ресурс STRING (<https://string-db.org/>).

2.1 Веб-приложение и использованные при разработке технологии

Для решения отдельного класса задач, не связанного с генными сетями и не задействующего Cytoscape, на основе алгоритмов, используемых в Orthoscape, было создано веб-приложение Orthoweb, запускаемое на веб-сервере и позволяющее пользователю работать с ним через веб-браузер. Задачи, связанные с веб-разработкой, как правило, решаются с помощью множества сторонних библиотек и фреймворков. Фреймворк – это программное обеспечение, облегчающее разработку и объединение разных компонентов большого программного проекта. При построении веб сервиса были использованы фреймворки Spring (spring.io) и Webix (webix.com).

2.1.1. Spring

Фреймворк Spring объединяет в себе множество функций, упрощающих проектирование и разработку приложений на Java. Spring автоматически осуществляет конфигурацию проекта, регулирует зависимости и связи

различных модулей программы. Объекты, которые находятся внутри фреймворка Spring и управляются им, называются «бинами», т.е. термин «бин» в Spring используется для ссылки на любой компонент. В бинах Spring реализовано все необходимое для создания частей программного обеспечения, отвечающих за такие важные этапы, как веб-программирование и создание веб-сервисов, работу с базами данных, загрузку файлов, обработку ошибок и т.д. Зависимости между бинами регулируются с помощью Dependency Injection контейнера (контейнер внедрения зависимостей, далее «ВЗ»). ВЗ-контейнер также регулирует процессы предоставления внешних зависимостей к программным компонентам. С помощью ВЗ разработчик может загружать конфигурацию зависимостей из внешних xml-файлов, из конфигурационных классов Spring и с помощью аннотаций в коде. ВЗ предоставляет возможность бинам использовать любую реализацию программных интерфейсов для учета зависимости.

С помощью Spring в работе создается RESTful веб-сервис – веб-сервис, основанный на архитектуре REST (representational state transfer, передача состояния представления). Данная архитектура подразумевает наличие отдельно реализованных клиентской и серверной частей и регулирует взаимодействие между клиентом и сервером. Сервер содержит базу с данными и проводит за счет своих мощностей анализ этих данных, а клиентом выступает компьютер, с которого пользователь запускает веб-браузер и веб-приложение Orthoweb. Для загрузки страницы с формами ввода исходных данных ее адрес вводится в браузер, в котором осуществляется GET запрос к серверу. Затем, с помощью POST запроса, данные, введенные пользователем в формы, передаются на сервер для проведения анализа. GET и POST запросы являются самыми распространенными и простыми типами запросов данных в сети интернет. Запрос GET является открытым и позволяет получать данные с сайта как с помощью подготовленного на странице инструментария, так и с помощью ручного ввода текста запроса в адресную строку. Запрос POST же

скрывает передаваемые данные, таким образом, он может быть вызван только с помощью специально подготовленного программистом для пользователя функционала. Данные, передаваемые между клиентом и сервером, представлены в одном из заранее описанных форматов. Одним из самых популярных, в том числе, и используемым в данной работе является формат json (JavaScript Object Notation). В большинстве языков программирования уже реализованы функции для автоматической обработки данных, представленных в json, к тому же такие данные могут быть легко интерпретированы и человеком. Ниже приведен пример данных на языке json:

```
{
  "Entry": "10518",
  "Organism": "hsa",
  "Pfam Motifs": [
    "EF-hand_7",
    "EF-hand_1"
  ],
  "Other DBs": {
    "NCBI-GeneID": "10518",
    "Ensembl": "ENSG00000136425",
    "UniProt": "075838"
  }
}
```

Результаты анализа отправляются сервером клиенту в формате json и могут быть как просмотрены на веб странице, так и загружены с сервера в заархивированном виде, с помощью отдельной опции.

2.1.2. Webix

Для реализации клиентской части использовался фреймворк Webix (<https://webix.com/>). Webix позволяет конструировать виджеты и графические интерфейсы, ориентирован на визуализацию и редактирование данных, создание макетов, организацию навигации по странице. Webix поддерживает такие возможности, как перетаскивание элементов (drag and drop), динамическую загрузку данных, интеграцию со сторонними виджетами и т.д. Еще одним плюсом Webix является наличие полноценной документации и

рабочего примера для практически каждого виджета, реализованного во фреймворке, с доступным для просмотра исходных кодом, что существенно упрощает и ускоряет процесс изучения фреймворка и его возможностей.

Для того, чтобы html-страница имела доступ к возможностям Webix, необходимо либо предварительно загрузить и указать путь к файлам `webix.js` и `webix.css`, либо указать в html-странице требование загрузить эти данные онлайн, с помощью следующих строк:

```
<link rel="stylesheet" href="http://cdn.webix.com/edge/webix.css"
type="text/css"></link>
<script src="http://cdn.webix.com/edge/webix.js"
type="text/javascript"></script>
```

Элементы графического интерфейса располагаются внутри блоков `webix.ui({`
`// Здесь реализация графических форм, таблиц и т.д.`
`})`

Расположение элементов на странице (layout) регулируется с помощью расположения содержимого страницы в заранее заготовленной сетке (grid) и с помощью разбиения представления содержимого на строки (rows) и столбцы (cols) (рисунок 2.1).

Рис. 2.1 Пример построения графического интерфейса в Webix на основе Orthoweb – исходная форма состоит из двух столбцов, а содержимое левого столбца, в свою очередь, состоит из нескольких строчек с данными.

Для переключения между вкладками различных типов анализа и вкладкой результатов анализа используется `tabbar` (https://docs.webix.com/desktop_tabbar.html). Для хранения динамически погружаемого результата анализа используется `datatable` (https://docs.webix.com/datatable_index.html).

Для реакции на действия пользователя в Webix реализованы функции. Код функции, например, содержащий запрос на отправку данных серверу, помещается в следующий блок:

```
function funcName() {
    // Здесь описывается функциональность
}
```

2.1.3. MongoDB

Для существенного ускорения эволюционного анализа, а также для обеспечения возможности осуществлять анализ даже без доступа к ресурсу с данными, загруженные ранее данные хранятся на сервере в локальной базе. Данные хранятся в различных структурах, поэтому, для хранения была выбрана нереляционная база данных MongoDB (т.е. записи хранятся без явных и структурированных механизмов для связывания друг с другом). Конкретная запись в таблице, например, карточка одного гена с данными о его идентификаторе, доменах, нуклеотидной последовательности и т.д., обладающая своим уникальным идентификатором, называется документом. Для хранения документов в базе используется формат json.

Данные в MongoDB легко могут быть добавлены и обновлены, полностью или частично, что позволяет поддерживать базу в актуальном состоянии, добавляя новые данные и обновляя устаревшие. База занимает на порядок меньше места, чем занимали бы записи, сохраненные в отдельных текстовых файлах.

2.2 Cytoscape и подключаемые к Cytoscape приложения

Основным программным комплексом, используемым в работе, является Cytoscape. Cytoscape реализован на языке программирования Java и содержит инструментарий для поддержки разработки подключаемых приложений. Разработчиками Cytoscape написано руководство по созданию подключаемых приложений (<https://github.com/cytoscape/cytoscape/wiki/Cytoscape-App-Ladder>), которое ориентировано на среды разработки Eclipse (<http://www.eclipse.org/>) и IntelliJ (<https://www.jetbrains.com/ru-ru/idea/>) с использованием платформы Apache Maven (<http://maven.apache.org/>) для сборки проектов.

Для написания программного кода приложения Orthoscape использовалась среда разработки Eclipse. Eclipse является свободно распространяемой средой разработки (распространяется по лицензии Eclipse Public License) и обладает полным набором функциональных возможностей для написания, сборки и отладки проектов. Orthoscape написан на языке Java. Java является объектно-ориентированным языком, одной из особенностей которого является тот факт, что любой код программы принадлежит классам. Точкой входа (т.е. местом, с которого начинается выполнение программы) в работу подключаемого приложения является класс `CyActivator`, функция `start(BundleContext bc)`. В данном классе разработчиком формируется главное меню приложения и регистрируются обработчики для каждой опции из главного меню. Обработчиками являются классы вида `TaskFactory`, создаваемые для каждой функциональной ветви главного меню приложения, код которых вызывается при выборе соответствующей им опции. Классы `TaskFactory` используются для создания конкретных задач `Task` (класс с конкретной реализацией соответствующей задачи). Таким образом, архитектура обработки задач выглядит следующим образом (рисунок 2.2).

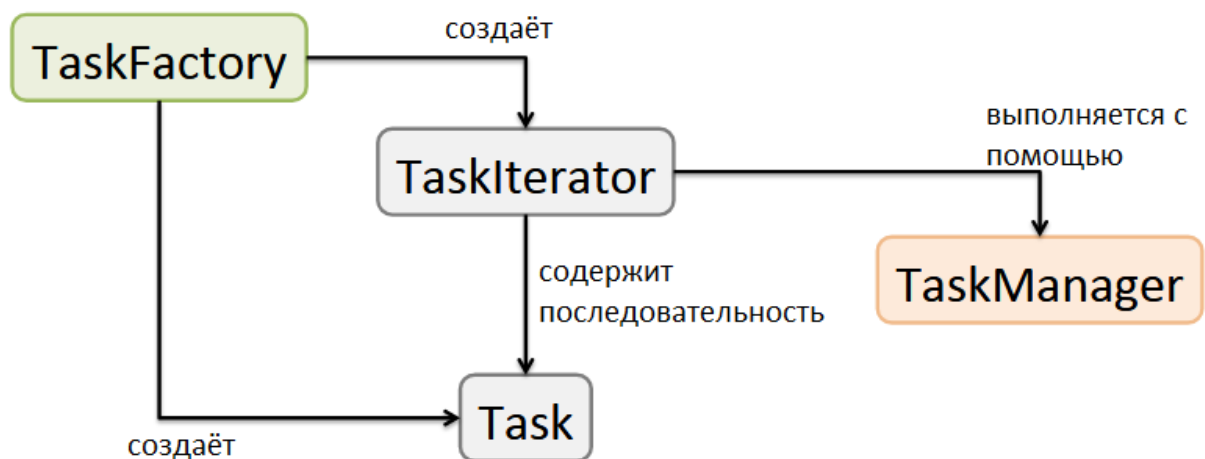


Рис. 2.2 Схема управления задачами в подключаемых модулях Cytoscape (класс `TaskFactory` начинает свою работу при нажатии пользователем на одну из опций модуля внутри Cytoscape).

Установленные приложения отображаются в меню Cytoscape -> Apps. В данной работе использовалось три ранее разработанных приложения:

1) Приложение CyKEGGParser (<http://apps.cytoscape.org/apps/cykeggparser>) [Nersisyan, Samsonyan, Arakelyan, 2014] используется для импорта сетей из базы данных KEGG.

2) Приложение GeneMANIA (<http://apps.cytoscape.org/apps/genemania>) [Montejo и др., 2010] используется для реконструкции сети на основе заданного набора генов.

3) Приложение CyPath2 (<http://apps.cytoscape.org/apps/cypath2>) [Cerami и др., 2011] используется для импорта сети из базы PathwayCommons в формате BioPAX (level 3).

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

В работе представлены приложения Orthoscape и Orthoweb. Orthoscape - приложение к программному комплексу Cytoscape, предназначенное для анализа эволюционных характеристик генных сетей [Mustafin и др., 2017]. Orthoscape позволяет искать гены, гомологичные друг другу, анализировать эволюционные характеристики генов, входящих в сеть. Orthoweb - самостоятельное веб-приложение, не требующее установки Cytoscape. Orthoweb работает в стандартном веб-браузере, имеет локальную базу, реализованную с помощью системы управления базами данных MongoDB и направлен на анализ функционально связанных групп генов, не объединенных в сеть.

3.1 Функциональные возможности Orthoscape и Orthoweb

3.1.1 Поиск гомологов

Главным источником информации о генах и их характеристиках является база данных KEGG. В процессе работы приложения с помощью API базы запрашивают такую информацию, как списки гомологов, доменный состав белков, кодируемых анализируемыми генами, таксономический ряд организма (ген которого подвергается анализу), нуклеотидные/аминокислотные последовательности гена/белка. Данные могут быть сохранены в локальную базу, что значительно ускорит повторную работу приложений с ними. На основе полученных данных разработанные приложения осуществляют поиск генов, гомологичных генам из заданного входного набора генов/генной сети. В KEGG представлены данные об ортологах и паралогах, которые имеют следующий вид (рисунок 3.1):

SSDB Best Search Result

KEGG ID : [hsa:51141](#) (225 a.a.)
 Definition: insulin induced gene 2
 Update status: [T01001](#) (aman,aoa,bgi,blat,boe,bwh,cann,cbra,ccjz,cdk,cet,ceu,chv,colw,dfg,dori,eai,epz,exf,fib,lcv,magn,mbry,mcak,mih,naf,nta,ofo,osg,paln)
 Show : ☐ Best-best ☒ Best ☐ Paralogs ☐ Gene clusters
 Threshold:

Top10

Search Result : 307 hits

Entry	KO	len	SW-score(margin)	bits	identity	overlap	best(all)
☒ ggo:101148382 insulin induced gene 2		225	1529 (394)	354	1.000	225	<-> 2
☒ nle:100603825 insulin induced gene 2		225	1529 (386)	354	1.000	225	<-> 2
☒ ocu:100355934 insulin induced gene 2		225	1529 (397)	354	1.000	225	<-> 3
☒ pps:100991290 insulin induced gene 2		225	1529 (394)	354	1.000	225	<-> 2
☒ ptr:743385 insulin induced gene 2		225	1529 (394)	354	1.000	225	<-> 2
☒ sbq:101044190 insulin-induced gene 2 protein isoform X1		225	1529 (406)	354	1.000	225	<-> 2
☒ csab:103216939 insulin induced gene 2		225	1525 (382)	353	0.996	225	<-> 2
☒ mcc:693654 insulin induced gene 2		225	1525 (390)	353	0.996	225	<-> 2
☒ mcf:102140508 insulin induced gene 2		225	1525 (390)	353	0.996	225	<-> 2
☒ pon:100173587 insulin induced gene 2		225	1524 (381)	353	0.996	225	<-> 3
☐ ssc:100170127 insulin induced gene 2		225	1520 (400)	352	0.991	225	<-> 2
☐ ecb:100049957 insulin induced gene 2		225	1519 (384)	352	0.991	225	<-> 2

Рис. 3.1 Пример выдачи списка ортологов гена из базы данных KEGG. На данном фрагменте можно видеть параметры SW-score (результат работы алгоритма локального выравнивания Смита-Ватермана, изменяется от 1529 до 1519 для ортологов этого фрагмента) и identity (изменяется от 1 до 0.991). Эти параметры используются разработанными приложениями для отбора ортологов.

Для каждого гена, описанного в базе данных KEGG, предоставляется список гомологичных ему генов, упорядоченных по результату работы алгоритма локального выравнивания Смита-Ватермана [Smith, Waterman, 1981]. Это первый параметр, по которому пользователь может фильтровать гомологи из списка, предоставленного KEGG. Пользователь может указать минимальное значение результата работы алгоритма Смита-Ватермана (SW score), гомологи из списка со значением ниже указанного будут отброшены. Это позволяет контролировать степень сходства генов, которую готов допустить исследователь в своей работе. Вторым параметром, по которому ведётся отбор, является параметр identity. Этот параметр также указан в исходных данных KEGG и отражает степень посимвольной идентичности аминокислотных последовательностей белков, кодируемых последовательностями рассматриваемого гена и его гомолога (таким образом,

identity является числом из интервала $[0; 1]$). Третьим параметром для отбора гомологов является указание степени сходства доменного состава белков, кодируемых геном и гомологом. Для каждого гена KEGG предоставляет данные по таким доменам, и пользователь может указать количество доменов, которое должно совпадать у гена и гомолога, или же указать в текстовом виде название конкретных доменов, и приложения будут отбирать гомологов, у которых есть все указанные домены. Все три параметра проверяются по очереди, но возможно пропустить проверку по любому из них указав его значение равным 0.

В результате работы алгоритма поиска гомологов для каждого гена будет создано поле, указывающее его принадлежность к гомологичной группе. Все гены с одинаковым значением этого поля принадлежат к одной гомологичной группе, то есть у каждого участника группы есть хотя бы один гомолог среди других генов этой группы.

3.1.2 Анализ эволюционных характеристик

Разработанные приложения могут проводить анализ макро- и микроэволюционных характеристик. Эволюционные характеристики генных сетей выражаются с помощью эволюционных характеристик генов, входящих в рассматриваемые сети. Было разработано и реализовано две эволюционных характеристики генов – PAI и DI, которые позволяют с разных сторон взглянуть на эволюцию генов, содержащихся в генной сети, либо в списке генов, объединенных по какому-либо признаку. Следует отметить, что главная задача индексов помочь выделить именно те гены, на которые следует обратить внимание эксперту для более подробного анализа.

3.1.2.1 PAI

Для исследования эволюции на макроэволюционном уровне был разработан 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 древе. 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 3.2).

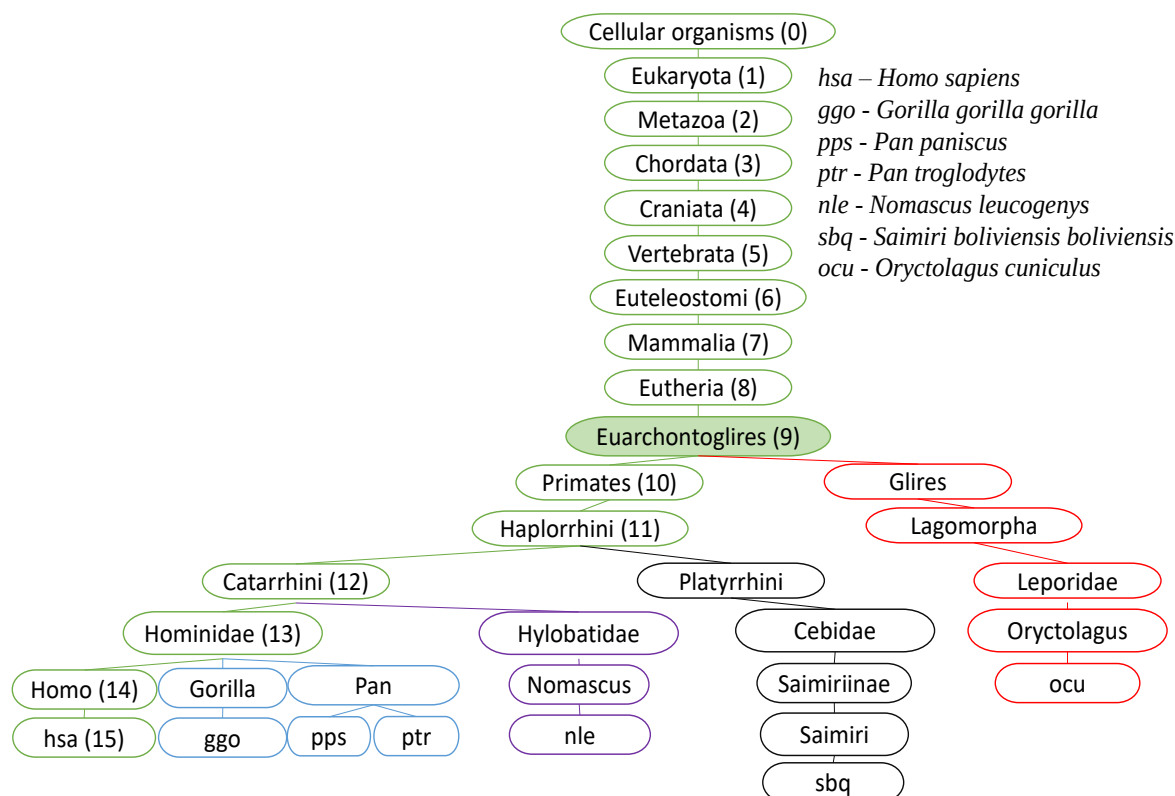


Рис. 3.2: Пример определения PAI для гена человека с шестью найденными ортологами. PAI принимает значение таксона, являющегося последним общим предком для гена и всех ортологов, т. е. полученное значение $PAI = 9$ (*Euarchontoglires*).

Алгоритм расчета PAI имеет следующий вид:

1) За основу берется таксономический ряд организма, которому принадлежит ген, из базы KEGG (например, для человека это ряд Cellular Organism; Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates;

Haplorrhini; Catarrhini; Hominidae; Homo; Homo sapiens (human)). Если ортологов найдено не будет, то значением PAI будет 15 (*Homo sapiens* (human)).

2) Проводится последовательное сравнение этого ряда с аналогичными рядами всех ортологов, на каждом этапе сохраняется только общая часть обоих рядов.

3) После анализа последнего ортолога последний таксон в ряду будет являться итоговым значением PAI.

Значение PAI опеределется таксоном и его порядковым номером от корня дерева, изменяющимся от 0 (Cellular organisms) до N, где N = число узлов на филостратиграфическом дереве - 1 (вычитаем единицу потому что счет ведется с нуля). PAI генной сети рассчитывается с помощью PAI всех входящих в сеть генов, с возможностью учета связности сети и без её учета.

3.1.2.2. DI

Для исследования эволюции на микроэволюционном уровне был разработан индекс DI (divergence index). Данный индекс является модификацией другого популярного среди эволюционных биологов индекса – dN/dS (K_a/K_s). DI отражает степень давления естественного отбора, путем вычисления отношения несинонимичных замен между двумя сравниваемыми нуклеотидными последовательностями к синонимичным (рисунок 3.3). DI следует рассчитывать при анализе гомологов организмов, расположенных близко к исследуемому организму на таксономическом дереве [Quint и др., 2012]. Пользователь может вручную указать на сколько слоев подниматься вверх от листа таксономического дерева при отборе ортологов. Кроме того, можно перечислить конкретные организмы, и только их ортологи будут рассматриваться в анализе. В результате расчета индекса для конкретного гена пользователь получает сводную характеристику (в виде среднего значения

индекса для всех ортологов гена) и значение индекса для каждой пары ген-ортолог.

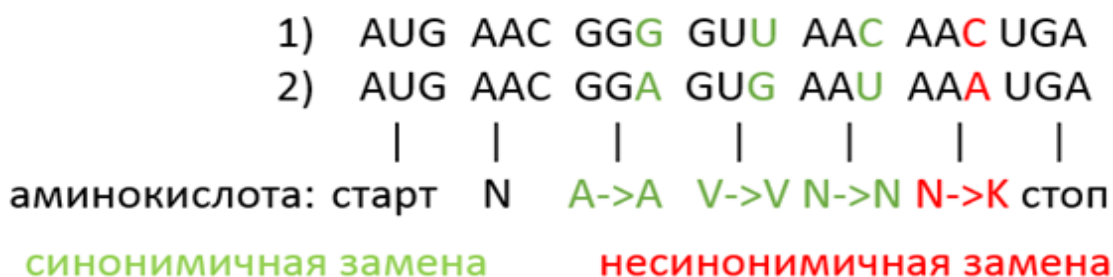


Рис. 3.3: Пример определения синонимичности замены. Основываясь на количестве синонимичных/несинонимичных замен, а также положении триплетов и прочих факторах, происходит итоговый расчет dN/dS .

Для расчета индекса необходимо подать на вход выравненные нуклеотидные последовательности гена и ортолога. Исходные последовательности загружаются из KEGG. Далее последовательности выравниваются с помощью Java реализации онлайн инструмента NW-align (<http://zhanglab.ccmb.med.umich.edu/NW-align/>) и подаются либо на вход KaKs Calculator [Zhang и др., 2006] (алгоритмы которого встроены в Orthoscape), либо на вход PAML (который должен быть загружен дополнительно и вызывается Orthoscape как самостоятельный процесс).

DI может быть вычислен для гена с помощью усреднения результата dN/dS для каждого проанализированного ортолога. Основная задача DI – выделить консервативные гены, для которых индекс близок к нулю, и вариабельные гены, для которых индекс обладает наибольшими значениями.

$$DI_s = \frac{\sum_{i=1}^n dnds_i}{n}$$

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

3.1.3 Визуализация результатов

Cytoscape отлично справляется с визуализацией результата, который можно непосредственно отобразить на генной сети. Но зачастую возникает необходимость создания статистик, на которых сравнивались бы результаты анализа нескольких сетей между собой (рисунок 3.4). Для этих целей была реализована возможность выведения дополнительных данных и создания отчетов вне Cytoscape, доступная в разработанных приложениях. В отчётах так же содержится существенная часть вспомогательных результатов, которые проблематично поместить в таблицу генной сети (например, отчет программы PAML по результатам анализа dN/dS для каждой уникальной пары ген-ортолог).

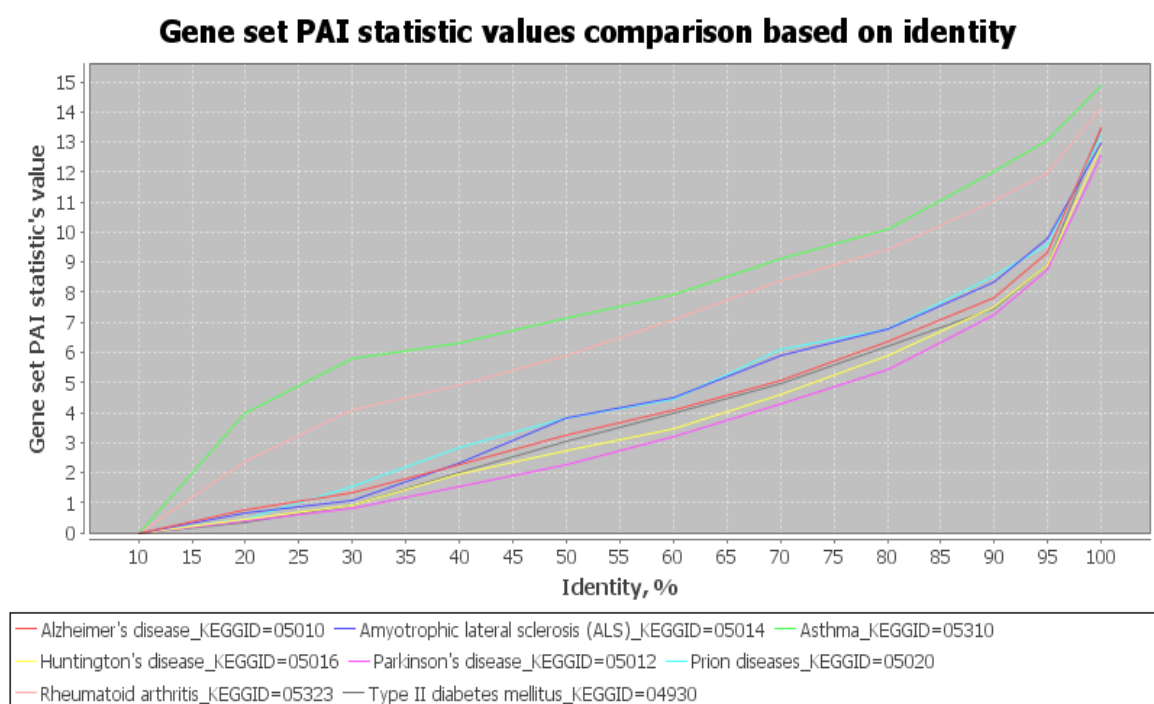


Рис. 3.4 Сравнение средних значений PAI по набору генов для нескольких рассмотренных генных сетей заболеваний человека при разных значениях параметра *identity*. Например, по рисунку видно, что при любых параметрах идентичности гена и ортологов (чем меньше параметр, тем больше ортологов берется для анализа) гены сети *Asthma* моложе генов любой другой сети (PAI сети *Asthma* (Астма) выше PAI других сетей). Сгенерировано Orthoscape.

В результате работы вычислительных модулей Orthoscape выводятся следующие вспомогательные результаты:

1) Директория *Aligned Sequences* (Выравненные Последовательности). Содержит нуклеотидные последовательности гена и ортолога, полученные после работы алгоритма Нидлмана-Вунша, которые далее используются для расчета DI. Для каждой уникальной пары ген-ортолог выводится собственный файл.

2) Директория *Full Ages Data* (Полные Данные о Возрастах). Содержит срезы таксономического дерева исходного организма и всех его ортологов. Это позволит вручную восстановить значение PAI. Для каждого рассматриваемого гена сети выводится собственный файл.

3) Директории *Kaks Base/Paml Base* (База Kaks, База Paml) содержат подробные отчеты по результатам вычисления отношения dN/dS . Для каждой уникальной пары ген-ортолог выводится собственный файл.

4) Директория, одноименная с рассматриваемой сетью, в которой находятся результаты анализа эволюционных индексов, гистограмма PAI (рисунок 3.5), скрипты для создания визуализации в пакете R (рисунок 3.6).

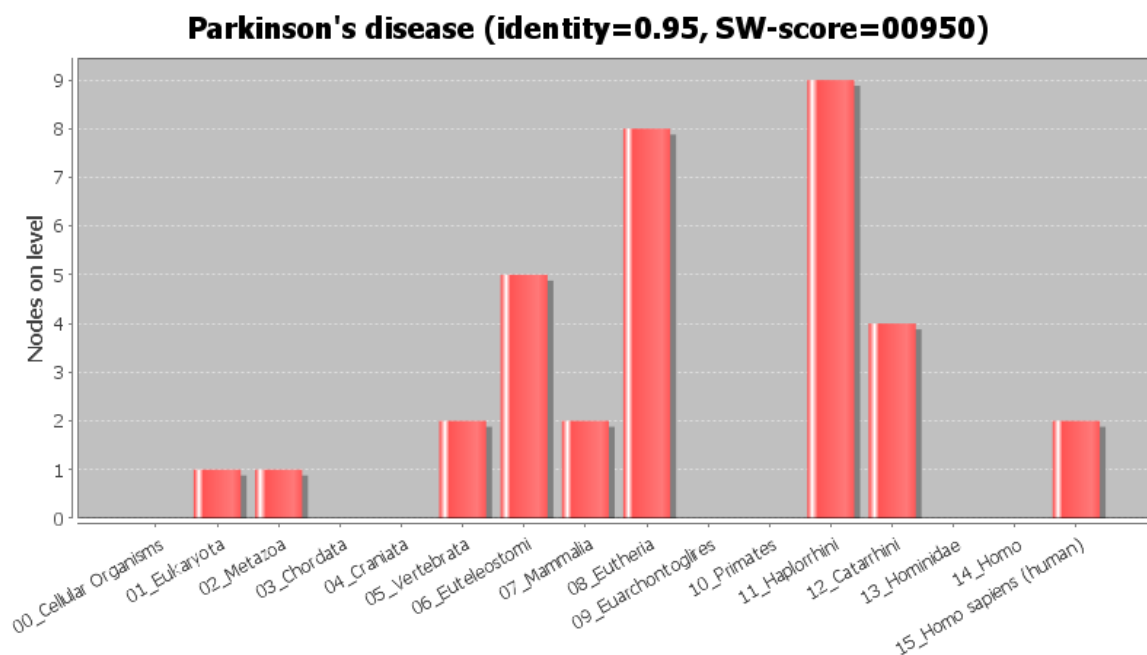


Рис. 3.5 Гистограмма распределения PAI в генной сети Parkinson's disease (болезнь Паркинсона) при значениях пороговых параметров identity = 0.95, SW-score = 950. Результаты можно условно разбить на 3 блока – «древние» гены (2 гена), гены «среднего возраста» (15 генов) и «молодые» гены (15 генов). Сгенерировано Orthoscape.

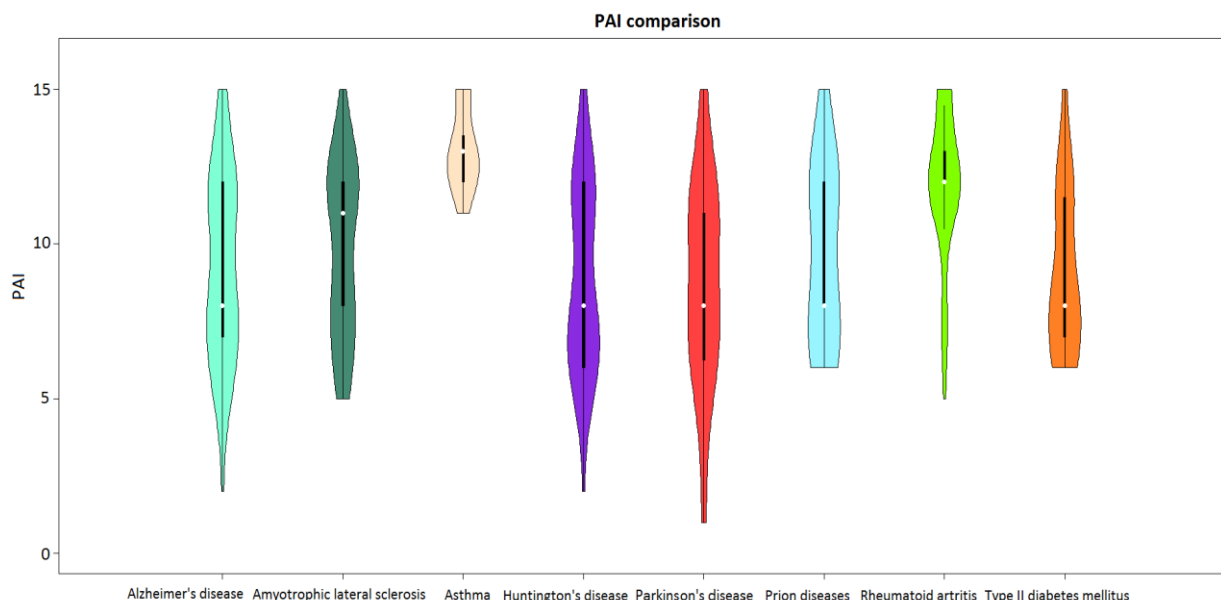


Рис. 3.6 Профили сетей нескольких заболеваний, построенные на основе гистограмм (рисунок 3.5). Чем «длиннее» профиль по оси ординат – тем

больший разброс PAI среди его генов. Чем «шире» профиль по оси абсцисс – тем больше генов соответствует PAI, указанному на оси ординат. Белая точка – среднее значение. Черная полоса – квартили распределения. Например, на рисунке видно, что в сети *Asthma*, в отличие от всех остальных сетей нет генов с низким значением PAI (эволюционно древних генов). Построено в R на основе R скрипта, сгенерированного Orthoscape.

Для создания итогового отчета по каждой сети отдельно и по всем сетям вместе используется самостоятельный модуль создания отчетов в Orthoscape. В результате его работы создаются html-файлы, общий и собственные для каждой сети, которые связаны гиперссылками между собой, в которых можно посмотреть упомянутую ранее информацию по сетям отдельно, а также сравнительную информацию для всех сетей (рисунок 3.5).

Для каждой обработанной сети рассчитываются следующие статистики:

а) Gene Set PAI. Эта статистика отражает среднее значение PAI всех генов в сети без учета ребер, т.е. вычисляется следующим образом:

$$Gene\ Set\ PAI = \frac{\sum_{genes} PAI}{N}$$

б) Median – значение статистической медианы в упорядоченной выборке из PAI всех генов.

с) Oldest – минимальное значение PAI (самый «древний» таксон).

д) Youngest – максимальное значение PAI (самый «молодой» таксон).

Кроме того, данные отчеты содержат гиперссылки на сети в базе KEGG для всех сетей, взятых из KEGG.

3.2 Отличия Orthoscape и Orthoweb

Orthoscape и Orthoweb обладают рядом отличительных особенностей:

- Локальная база данных в Orthoscape хранится на компьютере пользователя. Для хранения требуется место на жестком диске, но база позволяет работать с теми же данными без подключения к интернету. База Orthoweb хранится на сервере, не требует от пользователя места на его компьютере, но для использования Orthoweb, в любом случае, необходимо подключение к интернету.

- Orthoscape работает с генными сетями Cytoscape. Сети могут быть получены четырьмя способами:

- 1) Сети, импортированные из KEGG.

- 2) Сети, созданные по заданному набору генов с помощью плагина GeneMANIA. Перед работой с такой сетью потребуется однократно запустить специально реализованную в Orthoscape функцию конвертации для записи данных GeneMANIA в соответствующие сетям KEGG столбцы.

- 3) Сети в формате BioPAX, которые могут быть импортированы из базы Pathway Commons с помощью плагина CyPath2. Перед работой с такой сетью также понадобится однократная конвертация.

- 4) Сети, импортированные в качестве таблицы исходных данных. Для создания сети требуется импортировать в таблицу Node Table данные, разделенные табуляцией, в которых содержатся следующая информация:

- a) Столбец “name” (содержит в себе данные вида “*organism code:NCBI ID*”, например *hsa:51141*)

- b) Столбец “type” (на данный момент анализируются только гены, т.е. значение этого поля всегда должно быть равно “gene”).

- c) Опционально – столбец “Label”, в котором содержится текстовое название гена, которое будет выводиться рядом с кодом NCBI ID (например, *INSIG-2* для гена *hsa:51141*). Пользователь может пропустить этот этап, это не

скажется на результате работы программы, а лишь затруднит его интерпретацию.

Кроме того, потребуется вручную создать два столбца в таблице Network Table, необходимых для создания отчетов.

а) “title” – в этом столбце указывается название сети.

б) “org” – в этом столбце указывается код организма, чьи гены рассматриваются (например, hsa для человека).

Как следствие работы с сетями, есть и дополнительная статистика для отчета - Network PAI. Она отличается от Gene Set PAI тем, что вклад каждого гена прямо пропорционален числу ребер, ассоциированных с этим геном, т.е.

$$Network\ PAI = \frac{\sum_{nodes} (PAI * d)}{2N}$$

где N – число генов (узлов), d – степень узла.

- В Orthoscape доступны средства визуализации самого Cytoscape. Это компоновка на плоскости, группировка генов и покраска. Cytoscape позволяет реализовывать собственные визуальные стили и использовать их в своих плагинах. Для визуализации результата было создано 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 как по PAI (программой будет создана дискретная схема), так и по DI (программой будет создана непрерывная схема) (рисунок 3.7).

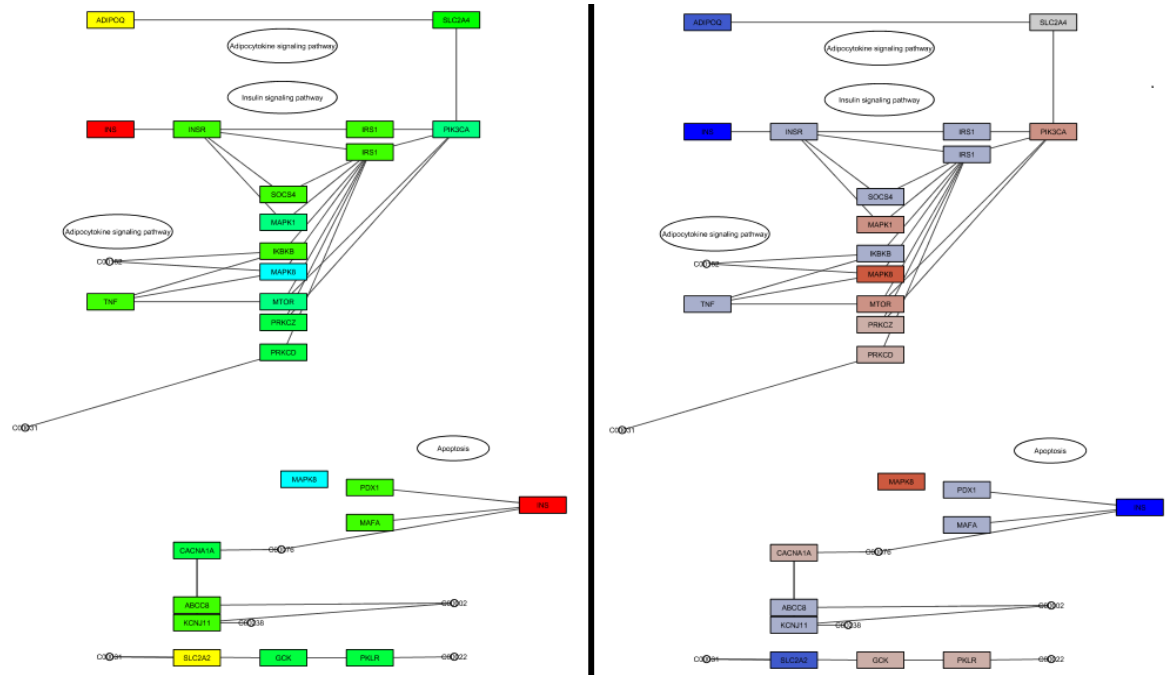


Рис. 3.7 Пример окрашивания с помощью цветовых схем, реализованных в Orthoscape: *heatmap* (слева) и *blue-red* (справа). В случае, когда результаты анализа подразумевают большой разброс данных, лучше использовать *heatmap* схему, которая за счет своей яркости обеспечивает лучшую читаемость. В случае большей концентрации результатов на коротком интервале, как на текущем рисунке, лучше использовать *blue-red* схему, поскольку *heatmap* обладает переизбытком оттенков зеленого для центральных узлов PAI.

Схемы работы Orthoscape приведена на рисунке 3.8.

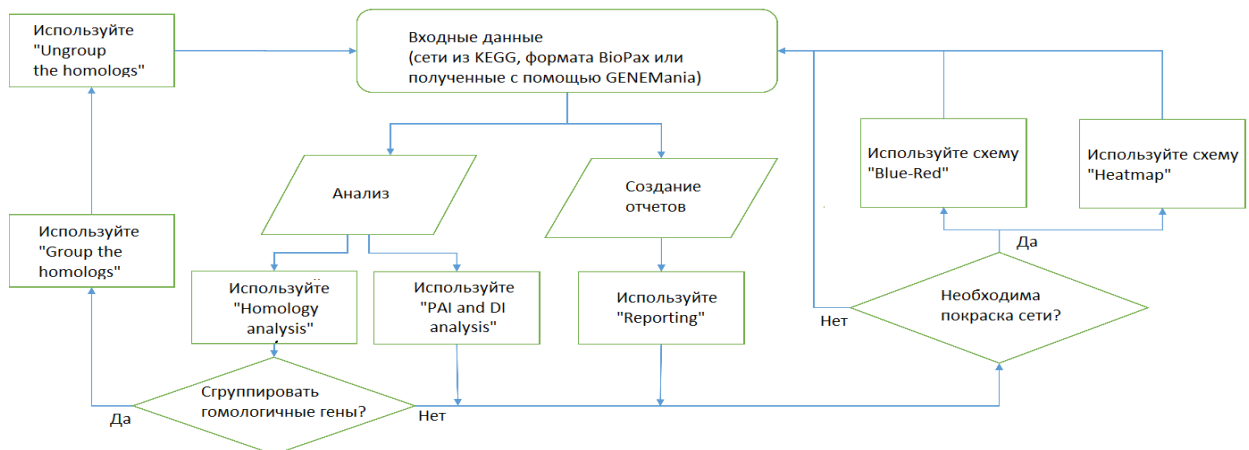


Рис. 3.8 Схема работы с Orthoscape.

• Orthoweb отличается архитектурой, имеет клиентскую и серверную части. Серверная часть реализована с помощью фреймворка Spring, а клиентская с помощью фреймворка Webix (рисунок 3.9). Доступен по адресу <https://orthoweb.sysbio.cytogen.ru/>.

The screenshot displays the 'Evolution' tab in Orthoweb, divided into 'Analysis' and 'Results' sections. The 'Analysis' section contains two sub-sections: 'Main parameters' and 'Divergence index analysis parameters'. In 'Main parameters', 'Use KEGG for a...' is checked, 'Identity' is set to 1, 'SW score' is 0, and 'Domains number' is 0. In 'Divergence index analysis parameters', 'Launch DI analysis' is unchecked and 'Taxonomy distance' is 2. The 'Results' section shows a 'Session ID' of 'default', a 'Make a report' button, and a large blue 'Upload files' button. At the bottom of the 'Analysis' section are two buttons: 'Launch from list' and 'Launch from files'.

Рис. 3.9 Стартовый вид вкладки *Evolution* в *Orthoweb*, на которой сосредоточен функционал блока, ответственного за расчет эволюционных индексов.

В левой части сосредоточены параметры, имеющие отношение непосредственно к эволюционному анализу. Все эти поля выполняют ту же функцию, что и в Orthoscape, за исключением поля Genes – данного поля в Orthoscape нет. Здесь пользователь указывает список генов, которые необходимо проанализировать. На анализ может быть подано сразу несколько списков, в таком случае, пользователю необходимо предварительно загрузить файлы с генами с помощью кнопки "Upload files" и они появятся в правой секции страницы. При каждом запуске программы пользователю вместе с результатом выдается идентификатор сессии, сгенерированный на основании

текущего времени системы, что позволяет избежать дублирования сессий. В данном случае под сессией понимается, главным образом, директория, в которой хранятся результаты анализа. Если пользователь хочет провести анализ нескольких разных сетей и сравнить их результаты между собой, то ему требуется указать для всех сетей один идентификатор сессии в поле "Session ID" вместо слова "default". В любой момент, с помощью кнопки "Make a report" пользователь может запустить процесс генерации отчета, аналогично Orthoscape, указав идентификатор сессии, и загрузить эти результаты, заархивированные приложением в zip-архив, с помощью появившейся ссылки "Download results". Результаты работы хранятся на сервере в течении месяца. Пользователь может изначально указать свой собственный идентификатор сессии, который он может легко запомнить, но, в таком случае, он должен быть в достаточно степени уникальным, чтобы другие пользователи не могли случайно начать работу в той же сессии. Схемы работы Orthoscape приведена на рисунке 3.10.

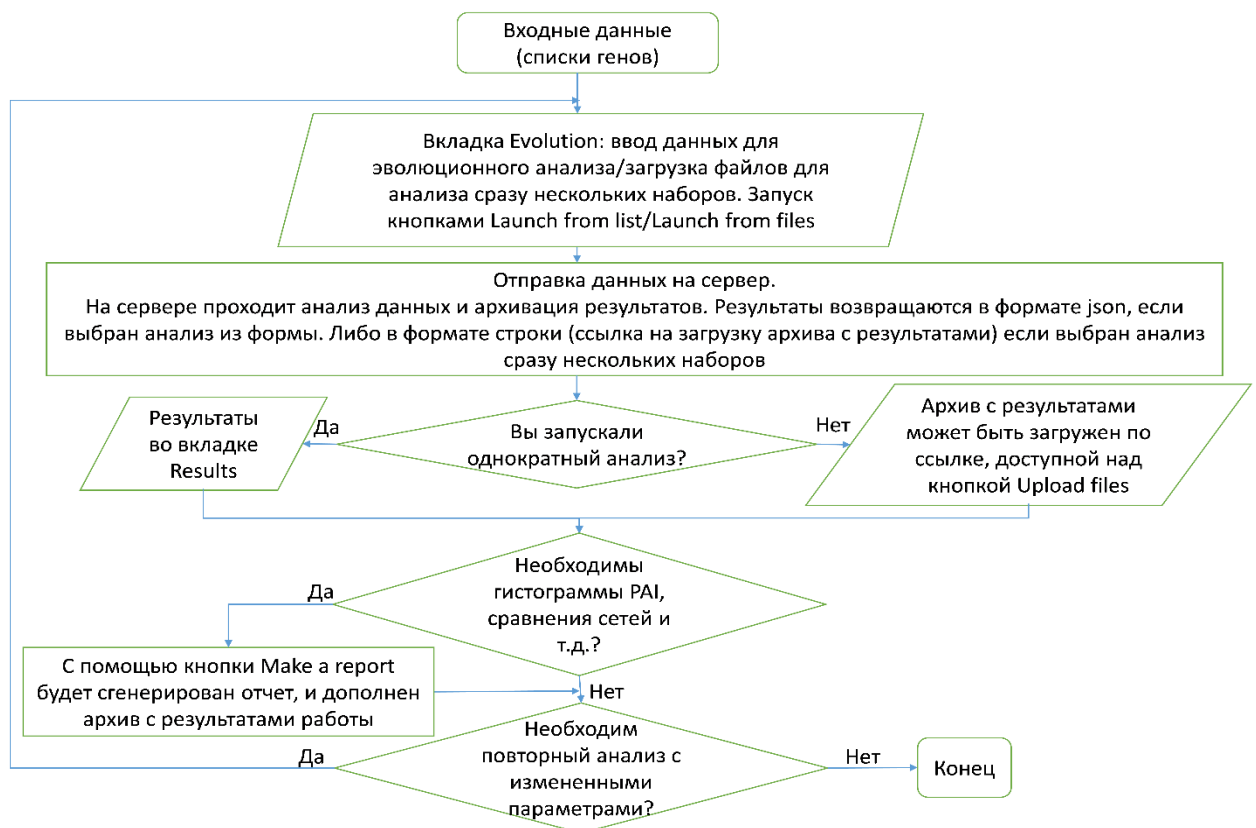


Рис. 3.10 Схема работы с Orthoweb.

3.3 Интерпретация результатов, полученных с помощью Orthoscape

В качестве проверки возможностей Orthoscape были выбраны сети биосинтеза стероидов (рисунок 3.11) и биосинтеза стероидных гормонов (рисунок 3.12) [Mustafin и др., 2017].

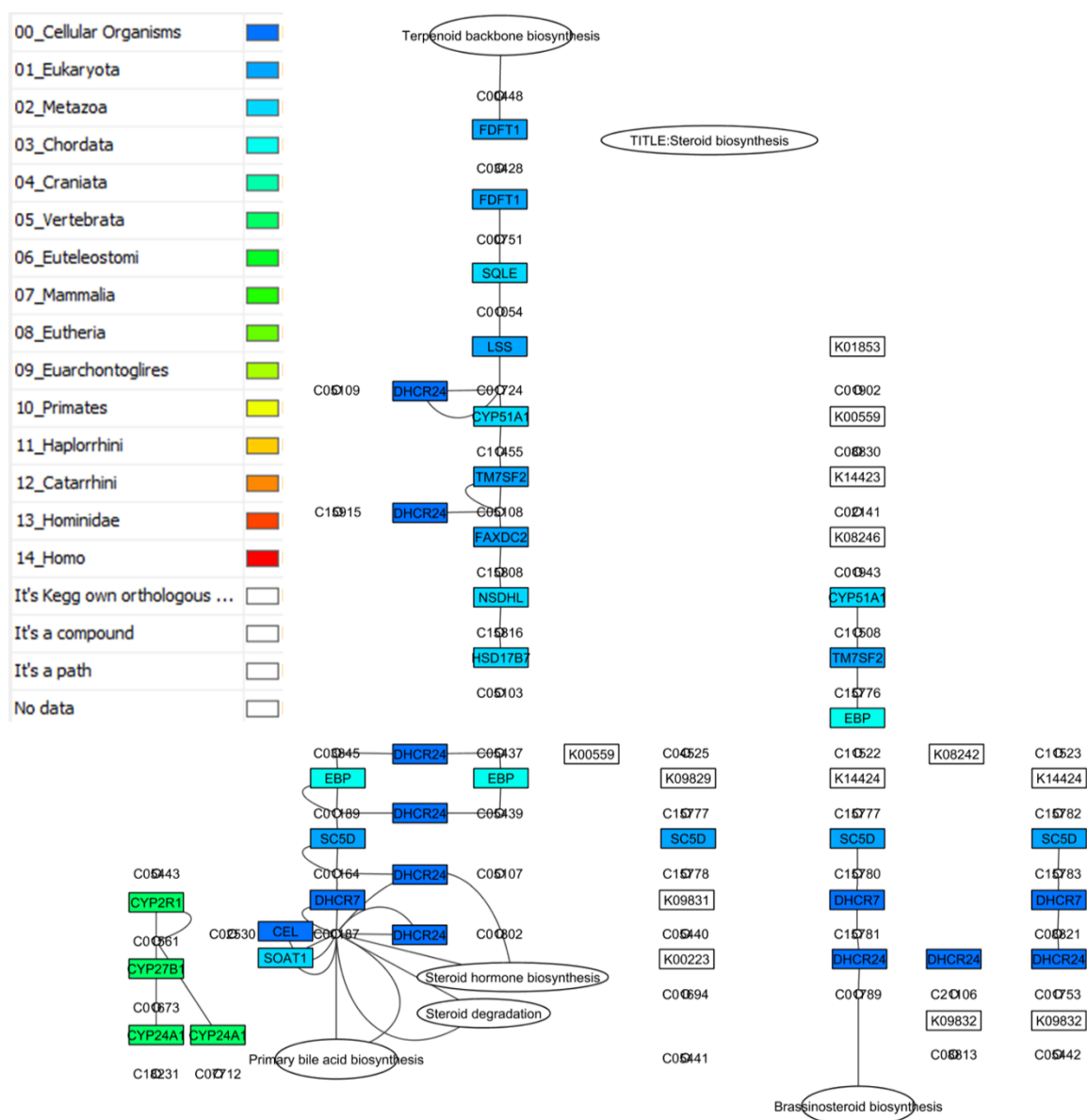


Рис. 3.11 Сеть биосинтеза стероидов

(<https://www.genome.jp/kegg/pathway/hsa/hsa00100.html>), покрашенная Orthoscape с помощью цветовой схемы «тепловая карта», т.е. возраст генов растет от зеленого цвета к синему. В данной сети преобладают «древние» гены.

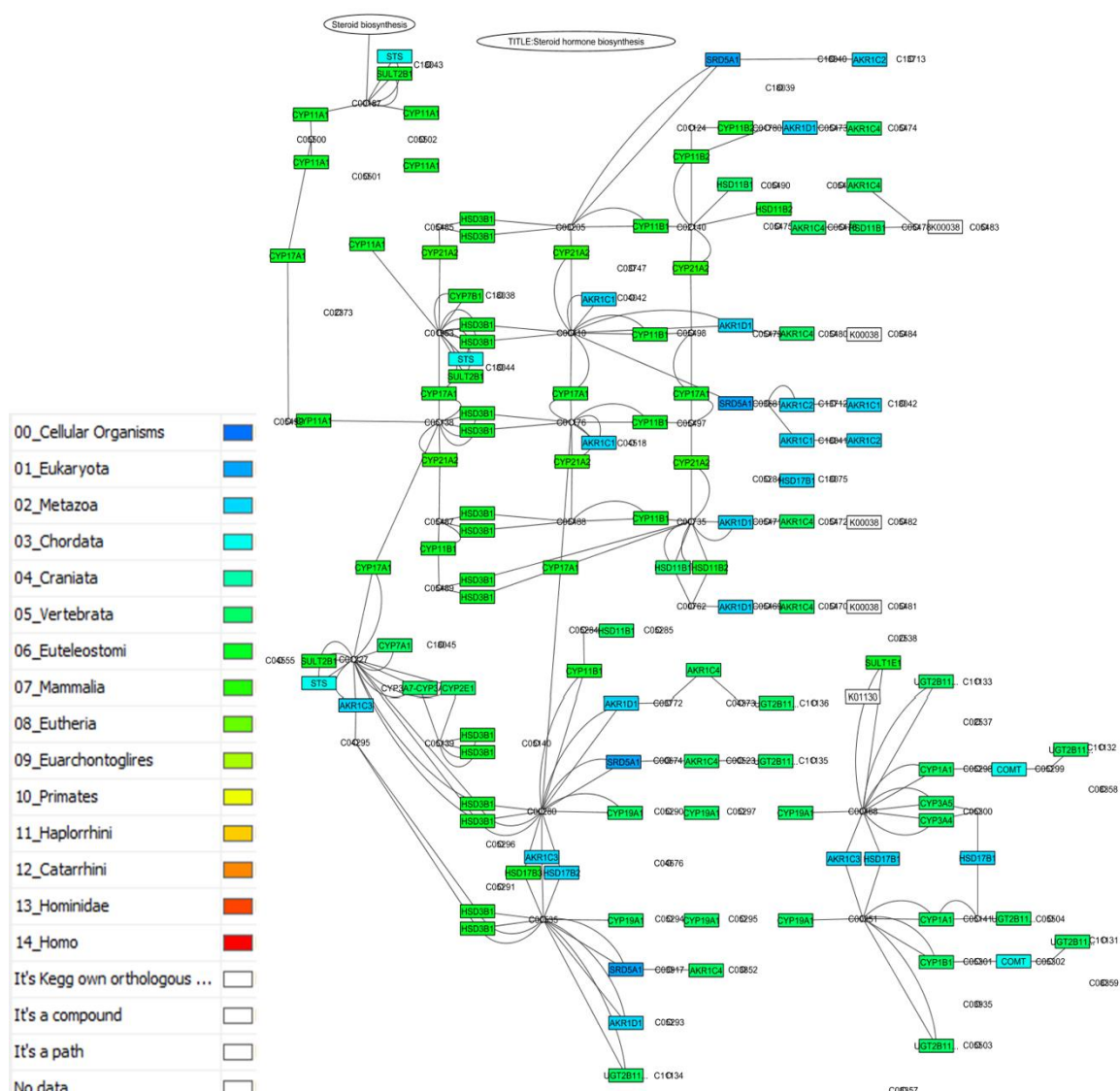


Рис. 3.12 Сеть биосинтеза стероидных гормонов (<https://www.genome.jp/kegg/pathway/hsa/hsa00140.html>), покрашенная Orthoscape с помощью цветовой схемы «тепловая карта», т.е. возраст генов растет от зеленого цвета к синему. В данной сети преобладают гены «среднего возраста».

Рисунки 3.11 и 3.12 показывают, что сеть биосинтеза стероидных гормонов содержит большее число генов «среднего возраста» (от позвоночных до млекопитающих, окрашены различными оттенками зеленого цвета, $PAI \geq 5$) и меньшее число «древних» генов (от клеточных организмов до хордовых, окрашены оттенками голубого и синего цветов, $PAI < 5$), чем сеть биосинтеза стероидов. Достоверность этого результата была проверена с

помощью точного теста Фишера (для двух сетей и двух признаков ($RAI < 5$ и $RAI \geq 5$)) было получено $p\text{-value} = 7.32 \times 10^{-12}$). Данный эффект сохраняется при использовании различных порогов идентичности нуклеотидных последовательностей гена и ортологов, т.е. в зависимости от установленного порога при отборе ортологов меняется возраст генов, но не меняется тот факт, что в сети биосинтеза стероидов средний возраст генов выше, чем в сети биосинтеза стероидных гормонов. Анализ показал, что большинство генов пути биосинтеза стероидных гормонов отличались от своих предков и приобрели свою новую функцию позже по сравнению с генами из пути биосинтеза стероидов.

Данные результаты согласуются с текущими знаниями об эволюции эукариот. Стероиды, например, холестерин, синтезируются почти во всех эукариотических клетках, которые используют тритерпеноидные липиды для контроля текучести и гибкости клеточных мембран. Стероиды участвуют в формировании мембраны, которая является основной клеточной структурой эукариот. Таким образом, биосинтез стероидов является фундаментальной особенностью эукариотических клеток, и общепризнанно, что путь биосинтеза стероидов появился после появления кислородного фотосинтеза и оксигенации атмосферы и океанов [Summons и др., 2006].

Стероидные гормоны регулируют разнообразные физиологические функции, такие как репродукция, водно-солевой баланс крови, поддержание вторичных половых признаков, реакция на стресс, функции нейронов, различные метаболические процессы [Hu и др., 2010] и реакция на факторы окружающей среды [Nau, 2007]. Анализ эволюции рецепторов стероидных гормонов показывает, что они являются общими для позвоночных [Baker, 2003], но могли возникать до расхождения позвоночных, что указывает на древнее происхождение некоторых систем стероидных гормонов (как было показано для передачи сигналов эстрогена [Thornton, 2003]). Следует, однако, отметить, что гормоны являются важным источником фенотипической

пластичности организма [Dufty, 2002]. Например, плацента, эволюционная инновация млекопитающих, демонстрирует удивительное разнообразие [Strauss, Martinez, Kiriakidou, 1996]. Эти фенотипические инновации должны требовать эволюционных изменений в пути биосинтеза гормонов, например, увеличивая его сложность для производства различных типов гормонов. Это вероятная причина относительно скромного возраста генов, участвующих в биосинтезе стероидных гормонов.

3.4 Заключение к главе 3

В этой главе представлены приложения Orthoscape и Orthoweb, предоставляющие пользователю возможность анализа генных сетей или самостоятельных наборов генов с целью установления между генами отношений гомологичности, осуществления филогенетического анализа и определения типа отбора, которому подвержены гены в процессе эволюции организма. Подобный анализ может помочь выделить эволюционно отличающиеся от других гены среди имеющихся, отделить кластеры гомологичных или наиболее изменчивых генов и решать другие задачи.

Глава 4. Исследование эволюционных характеристик генных сетей болезней человека

Перспективным подходом к пониманию природы заболеваний человека является их изучение в эволюционном аспекте. Так, например, несколько попыток было предпринято для объяснения закономерностей формирования энергетического метаболизма и связанных с ним заболеваний именно с позиций эволюционной биологии. Более 50 лет назад генетик из Университета Мичигана Джеймс Нил выдвинул гипотезу «сберегающего генотипа» ("thrifty genotype"), согласно которой гены, позволяющие человеку эффективно использовать ограниченные пищевые ресурсы в условиях голода, в условиях изобилия пищи приводят к развитию ожирения и сахарного диабета. Согласно этой гипотезе, гены, которые обеспечивают эффективное запасание энергии (накопление жира), были выгодны в условиях частых и длительных периодов голода, случавшихся на протяжении эволюции человека в палеолите, но стали «диабетогенными» в современных условиях [NEEL, 1962]. Таким образом, подход к изучению механизмов и распространенности болезней в эволюционном контексте, постепенно складывается в новое междисциплинарное направление на стыке фундаментальной медицины, генетики человека и эволюционной биологии [Степанов, 2016].

База KEGG Pathways содержит большое количество созданных вручную генных сетей. Для анализа был выбран раздел Human Diseases, сети в котором разделены на категории. В работе были рассмотрены категории: Immune diseases (Заболевания, связанные с иммунной системой, 8 сетей), Endocrine and metabolic diseases (Эндокринные заболевания и нарушения метаболизма, 6 сетей), Infectious diseases: Bacterial (Инфекционные заболевания, вызванные бактериями, 10 сетей), Infectious diseases: Viral (Инфекционные заболевания, вызванные вирусами, 9 сетей), Infectious diseases: Parasitic (Инфекционные

заболевания, вызванные паразитами, 6 сетей), Drug resistance: Antineoplastic (Лекарственная устойчивость к противоопухолевым препаратам, 4 сети), Cancers: Overview (Рак: обобщение, 7 сетей), Cancers: Specific types (Специфические типы рака, 15 сетей), Substance dependence (Зависимость от веществ, вызывающих привыкание к химическим соединениям, 5 сетей), Neurodegenerative diseases (Нейродегенеративные заболевания, 5 сетей), Cardiovascular diseases (Сердечно-сосудистые заболевания, 5 сетей). Суммарно в 11 категориях рассмотрено 80 сетей. Данные сети были проанализированы с помощью Orthoscape.

Таблица 4.1 *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 основе результатов анализа (таблица 4.1) 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 (Астма). Таким образом, в заболевания человека, связанные с иммунной системой, вовлечены гены, ортологи которых найдены только в отдельных группах современных 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 (Зависимость от веществ, вызывающих привыкание), а самое низкое значение PAI у сети Nicotine addiction (Никотиновая зависимость). Из этого можно сделать вывод, что гены, участвующие в регуляции процессов зависимостей, 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.

Также была выявлена высокая корреляция между PAI и DI индексами - 0.876 ($p\text{-value}=1,8 \times 10^{-26}$), что говорит о том, что гены, которые в рамках нашего анализа считаются «молодыми», как правило, отмечаются высокой изменчивостью по сравнению с группой гоминид. На рисунке 4.1 представлен график пары (PAI, DI) для всех сетей заболеваний человека.

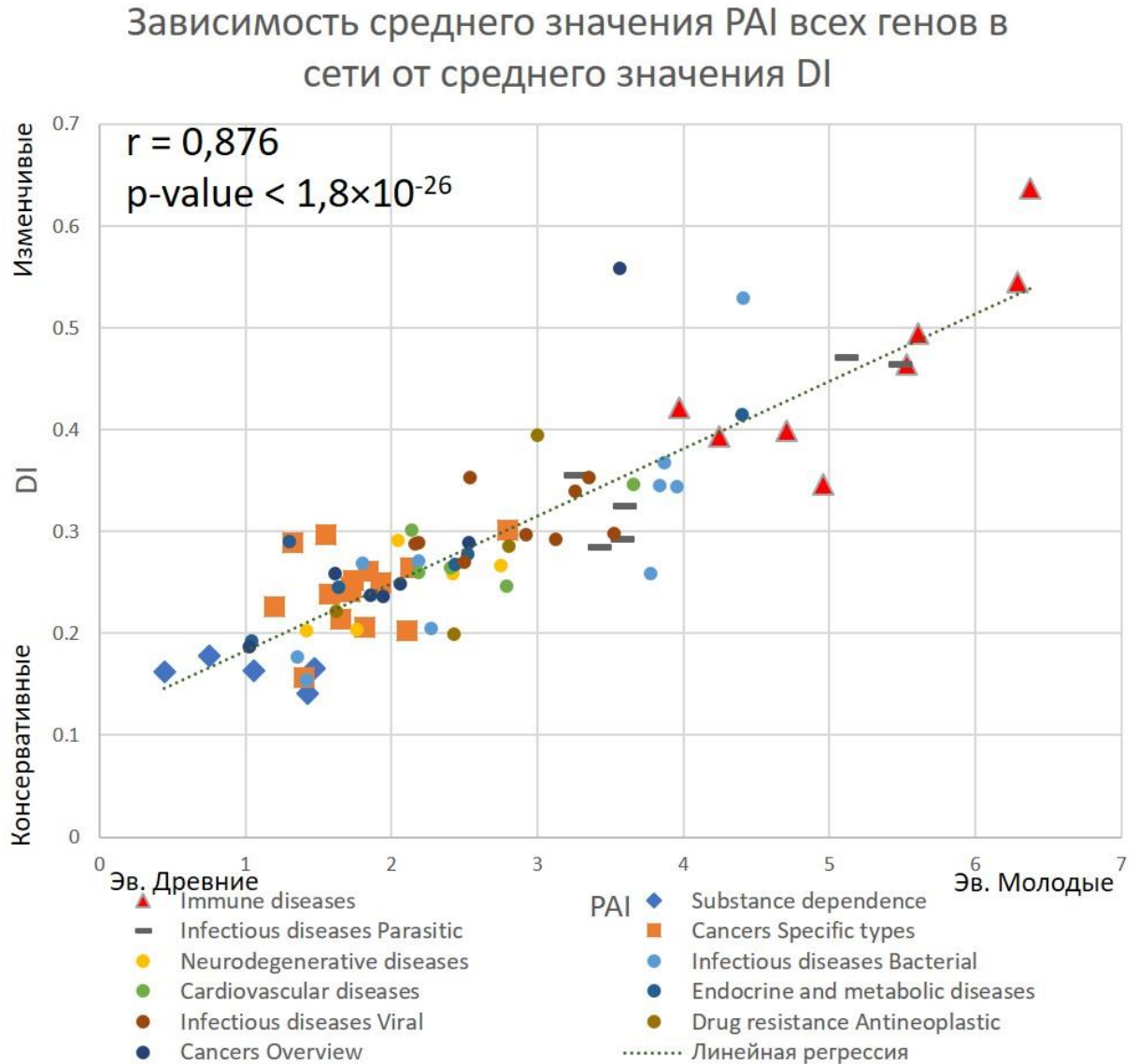


Рис. 4.1 Диаграмма рассеяния для средних значений индексов PAI и DI для 80 генных сетей заболеваний человека, описанных в базе KEGG Pathway, Human Diseases.

Отдельно стоит остановиться на 4 категориях. По высокому показателю PAI и DI выделяются категории Immune diseases и Infectious diseases: Parasitic. Низкий показатель PAI и DI характерен для заболеваний из Cancers: Specific types и Substance dependence. Гены из этих категорий, а также полный набор проанализированных генов (т.е. все гены, встреченные хотя бы в одной сети, всего таких генов получилось 1436), были разбиты на две группы: 1) группа эволюционно древних генов с $PAI < 5$ (в этом случае расхождение

исследуемого вида с наиболее отдаленным родственным таксоном, в котором обнаружен ортолог рассматриваемого гена, произошло на этапе от клеточных организмов до хордовых); 2) группа эволюционно молодых генов с $PAI \geq 5$ (расхождение исследуемого вида с наиболее отдаленным родственным таксоном, в котором обнаружен ортолог рассматриваемого гена, произошло на этапе возникновения Craniata (плеченогие) или позднее). Были составлены таблицы сопряженности и с помощью точного теста Фишера оценивалось, можно ли говорить о достоверном отличии разбиения генов на группы в категории от разбиения в полном списке генов. С учетом поправки Бонферрони на множественную проверку гипотез, было выделено 4 категории заболеваний, представлены в таблице 4.2.

Таблица 4.2. *Результаты точного теста Фишера по сравнению распределения по группам эволюционно древних и эволюционно молодых генов среди всех генов, описанных в генных сетях заболеваний человека из KEGG Pathway, Human Diseases, и среди генов в рамках одной категории. В колонках для каждой категории из строки указаны: название, количество эволюционно древних генов, количество эволюционно молодых генов, значение PAI, p-value теста.*

Категория	Эв. древних	Эв. молодых	PAI	p-value
Immune diseases	56	106	5,21	$8,84 \times 10^{-05}$
Infectious diseases Parasitic	74	84	4,08	$2,79 \times 10^{-06}$
Cancers Specific types	187	54	1,77	$4,41 \times 10^{-04}$
Substance dependence	69	9	1,03	$1,75 \times 10^{-05}$
Всего из 1436 генов	952	484	2,49	-

Среднее значение PAI всех 1436 задействованных генов составило 2,49. Генные сети, связанные с заболеваниями иммунной системы, обладают не только самым высоким значением PAI (5,21), но и достоверно отличным

распределением доли молодых и древних генов от аналогичной доли среди всех проанализированных генов (достоверность приведена в последнем столбце таблицы).

Доля молодых генов в категории заболеваний, связанных с иммунной системой, составила 65%. При этом наибольшая доля генов приходится на позвоночных (Vertebrata+Euteleostomi), что соответствует современным представлениям о развитии специфического иммунитета: он существует у хрящевых рыб (акул и скатов) и, следовательно, появился по крайней мере 400–500 млн лет назад. У этих рыб есть гены, родственные генам вариабельной области Ig (IgV) или генам рецепторов Т-клеток (ТкР). При этом еще более примитивные позвоночные – круглоротые (миксины и миноги) – не имеют системы приобретенного иммунитета, у них нет ни IgV-, ни ТкР-генов [Галактионов, 2004]. Анализ выявил также и некоторую долю эволюционно древних генов в категории заболеваний, связанных с иммунной системой. Это соответствует сложившемуся представлению о том, что некоторые функции иммунной системы возникали еще у одноклеточных, например, способность к фагоцитозу; клетки, имеющие маркер Т-лимфоцита, впервые обнаруженные у кольчатых червей, а система гистосовместимости - у губок [Хаитов, 2016].

С другой стороны, наибольшая доля эволюционно древних генов характерна для категории зависимостей от веществ, вызывающих привыкание (88%). Большинство рассмотренных генов вовлечены в фундаментальные процессы организма, в функционирование нервной системы, включая нейротрансмиттерные функции.

Достоверным отличием доли эволюционно древних и эволюционно молодых генов от аналогичной среди всех проанализированных генов обладает категория инфекционных заболеваний, вызванных паразитами (53% эволюционно молодых генов). Высокая доля эволюционно молодых генов может быть напрямую связана с высокой долей эволюционно молодых генов и высокой эволюционной изменчивостью генов, найденной в категории

заболеваний, связанных с иммунной системой. Именно инфекционные заболевания являются одним из важнейших движущих факторов эволюции иммунной системы. При этом инфекционные заболевания различной природы и иммунная система коэволюционируют в процессе формирования механизмов борьбы друг с другом [K. Sasaki, A. Tsutsumi, 2000; Khakoo, 2004; Zheleznikova, 2014].

Категория *Cancers Specific types* включает гены, ассоциированные с различными типами рака (рак желудка, рак поджелудочной железы, рак груди и т.д.). Для этой категории наблюдается достоверное превышение доли древних генов над молодыми в сравнении с их распределением в полной выборке проанализированных генов. Этот результат соответствует современным представлениям о том, что генные сети, вовлеченные в процессы развития рака, формировались на этапах возникновения многоклеточных организмов [Domazet-Lošo, Tautz, 2010b].

Большая часть из 1436 рассмотренных генов, при отборе ортологов с идентичностью аминокислотных последовательностей в 70% и выше, попадает в группы *Vertebrata*, *Euteleostomi*, *Mammalia*, *Eutheria*. Это означает, что они сформировались еще на этапе развития позвоночных. Таким образом, на основе этого результата, при проведении сравнительно-геномных исследований можно сузить круг рассматриваемых организмов без потери общности.

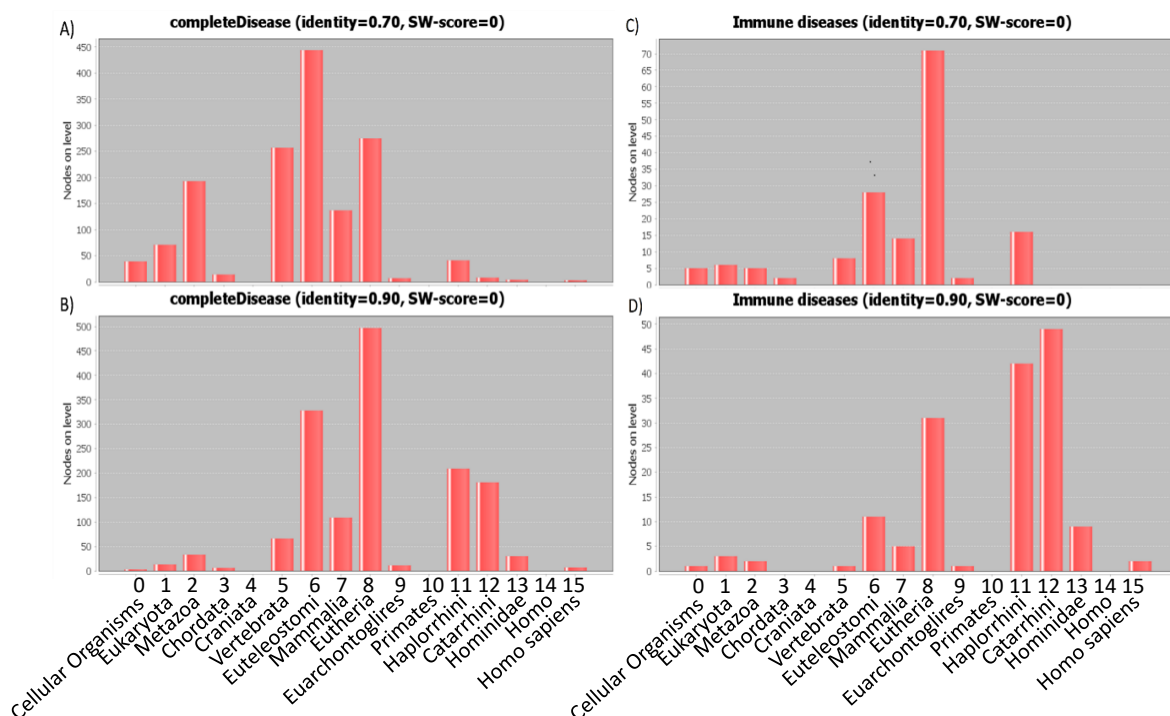


Рис. 4.2 Распределение возраста генов среди всех генов, собранных на основе 80 рассмотренных сетей (A, B) и среди генов группы Immune Diseases (C, D). Порога идентичности 0.9 достаточно для захвата всех генов семейства Hominidae и полноценного сравнения DI и PAI на максимально соответствующем им пороге идентичности. Порог 0.7 обладает немного меньшей корреляцией с DI, тем не менее, позволяет захватить существенно больше ортологов и проанализировать изменение возраста генов.

На рисунке 4.2B, при пороге идентичности 0.9, можно видеть разбиение на три группы по PAI (от 0 до 3, группа «древних» генов; от 5 до 9, группа генов «среднего возраста»; от 11 до 15, группа «молодых» генов). Если сравнить число генов третьей группы с аналогичным числом на рисунке 4.2D, на котором проанализированы уникальные гены группы Immune diseases, то видно, что около четверти всех «молодых» генов (102 из 427, ~24%) описываются этой группой, это наивысший показатель доли «молодых» генов среди всех групп заболеваний. Более того, если понизить порог идентичности при отборе гомологов и провести аналогичный анализ, т.е. обратиться к рисункам 4.2A и 4.2C, то доля таких генов уже будет относительно близка к

трети (16 из 56, ~29%), т.е. среди новых, добавленных в анализ ортологов (генов других организмов, идентичных рассматриваемым генам человека с идентичностью аминокислотных последовательностей между 0.7 и 0.9), лишь малая часть ортологов была обнаружена у генов группы Immune Diseases. Именно группа Immune diseases является основным источником «молодых» генов и генов «среднего возраста» (т.е. генов в середине и правой половине гистограммы) среди всех проанализированных генов заболеваний человека, представленных в KEGG.

Кроме того, большинство генов, входящих в состав исследованных генных сетей, были идентифицированы, как эволюционно консервативные ($DI < 1$) (рисунок 4.3).

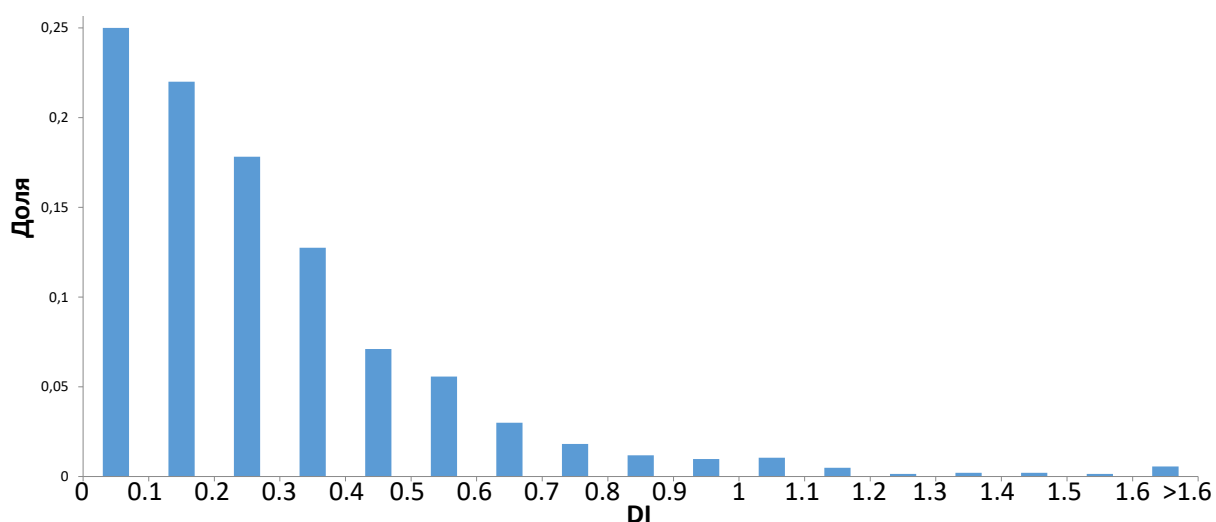
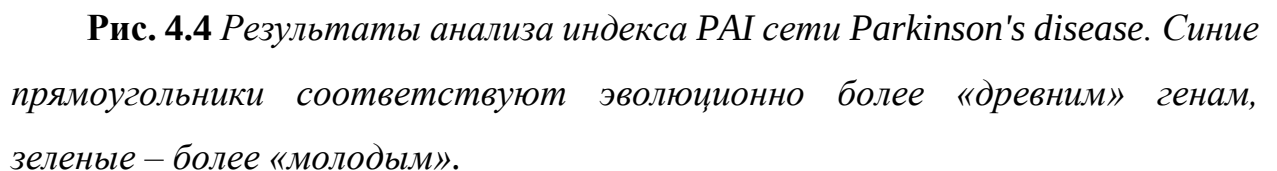


Рис. 4.3 Распределение DI среди всех генов, задействованных в генных сетях из KEGG Pathway, Human Diseases.

Тем не менее, независимо от группы, гены, которые выделяются и по PAI , и по DI , могут оказаться наиболее интересными мишенями для более подробного анализа. Для некоторых сетей был проведен такой анализ.

4.1 Анализ генной сети болезни Паркинсона

Результат анализа генной сети Parkinson's disease (болезнь Паркинсона) выявил низкий показатель PAI , что говорит о том, что гены данной сети



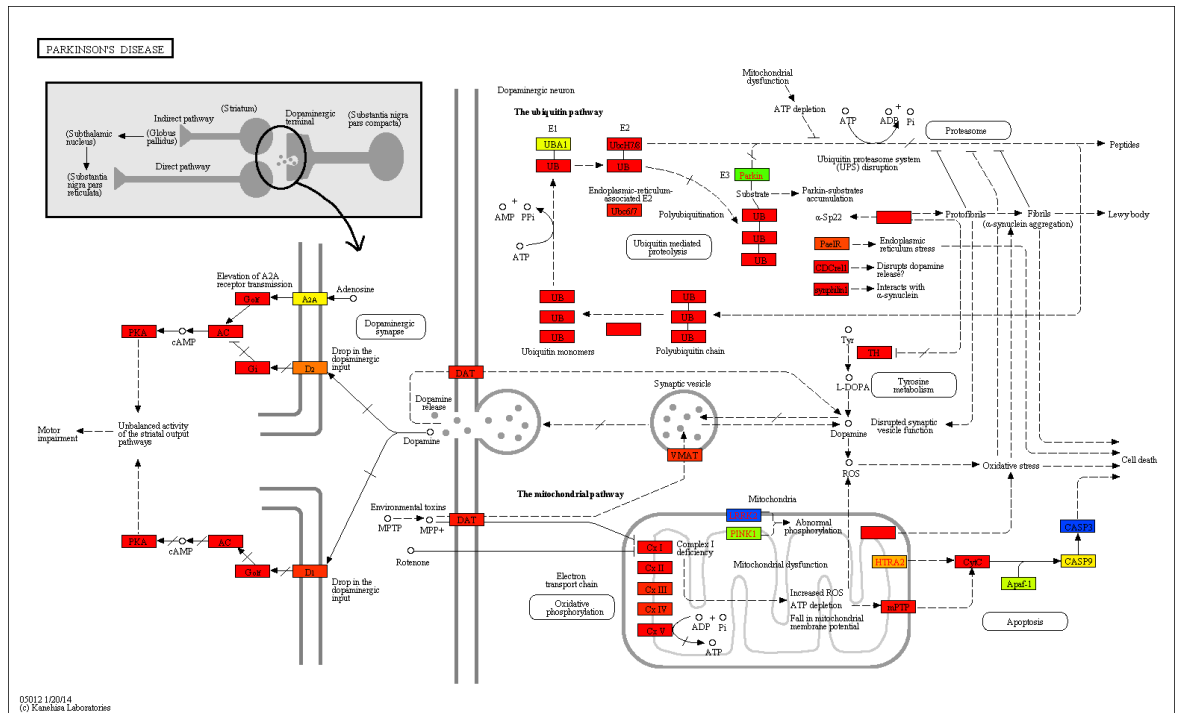


Рис. 4.5 Результаты анализа индекса DI сети Parkinson's disease. Красные прямоугольники соответствуют генам, нуклеотидные последовательности которых не имеют либо синонимичных, либо несинонимичных замен относительно последовательностей генов-ортологов, а также генам, у которых не найдено ортологов. Оранжевый, желтый, зеленый и синий прямоугольники соответствуют генам с изменчивостью (с нарастанием изменчивости вдоль описанной линии цветов) по отношению к генам-ортологам группы гоминид.

На рисунках 4.4-4.5 можно видеть два интересных результата:

1) На рисунке 4.4 (PAI) видно, что более «молодые» гены, как правило, располагаются в области мембраны, а также регулируют процесс апоптоза – программируемой гибели клетки.

2) На рисунке 4.5 (DI) можно видеть, что 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*, *Casp9*)[Deng и др., 2008; Gilks и др., 2005; Valente и др., 2004; Zimprich и др., 2004]. Ген *PARK2* (на сети

обозначен Parkin) кодирует белок parkin. В начале двухтысячных был достигнут прогресс в изучении функций этого белка. Было выяснено, что он является убиквитинлигазой, компонентом убиквитиновой системы, важным в процессе деградации аденозинтрифосфат зависимых белков. При этом, у пациентов с диагностированной болезнью Паркинсона он был подвержен множеству мутаций. В первую очередь мутации parkin связывают с возникновением аутосомно-рецессивной ювенильной болезни Паркинсона [Mizuno и др., 2001]. Мутации гена *PINK1* (*PARK6*), предназначенного для защиты клеток от митохондриальной дисфункции, также приводят к болезни Паркинсона. В серии работ Ганди с соавторами исследовали кальциевый гомеостаз и митохондриальную функцию в дефицитных по *PINK1* нейронах млекопитающих и подробно описали механизм, приводящий к гибели нейронов [Gandhi и др., 2006; Gandhi и др., 2009]. Ген *LRRK2* (*PARK8*) играет важную роль в связанных с дофамином функциях, мутации в нем ассоциированы с болезнью Паркинсона (type 8) [Dächsel, Farrer, 2010; Mata и др., 2006]. Ген *CASP9* является представителем большого семейства генов, кодирующих каспазы – семейства протеаз, которые играют важную роль в процессе апоптоза. Различные гены из этого семейства можно наблюдать в генных сетях множества заболеваний, но экспрессия *CASP9* может быть связана именно с гибелью нейронов [Hartmann и др., 2000; Tatton, 2000].

Таким образом, сеть Parkinson's disease попала в группу к более «древним» сетям за счет преобладающего числа «древних» генов, которые, в свою очередь, не отличаются повышенной изменчивостью в пределах семейства гоминид. При этом ключевые для развития заболевания гены в большинстве своем являются более «молодыми» и обладают высокой изменчивостью у гоминид.

4.2 Анализ генных сетей Диабета I и II типов

Сеть диабета первого типа, представленная в KEGG, содержит малое число генов, тем не менее на ней хорошо проявляется кластеризация на комплексы «древних» и «молодых» генов (рисунки 4.6 и 4.7).

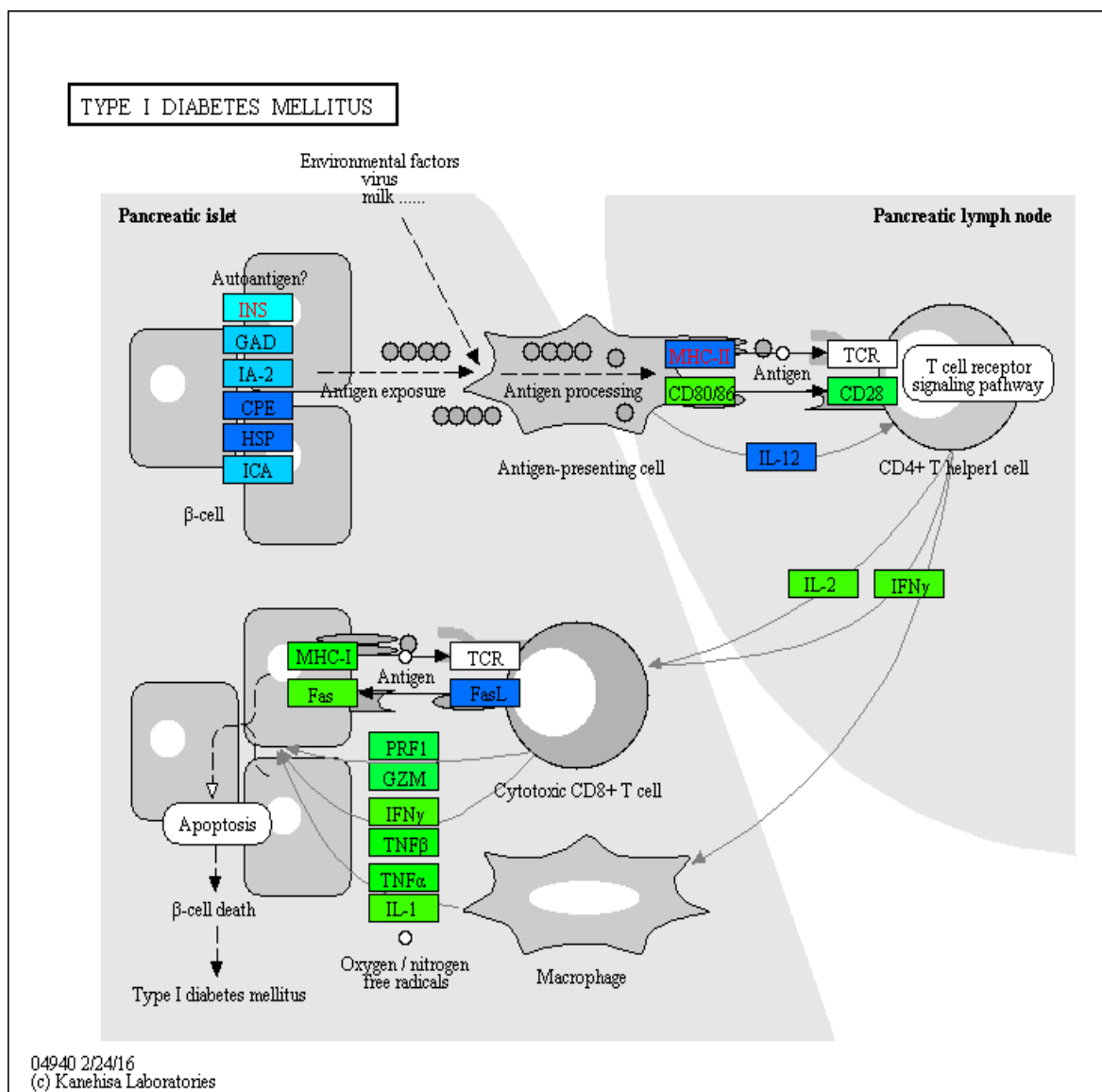


Рис. 4.6 Результаты анализа индекса PAI сети Type I diabetes mellitus. На данном рисунке хорошо видно кластер «древних» (голубых и синих) генов, регулирующих выработку антигенов, и кластер относительно «молодых» (покрашенных оттенками зеленого) генов, регулирующих процессы выживания клетки и клеточной гибели.

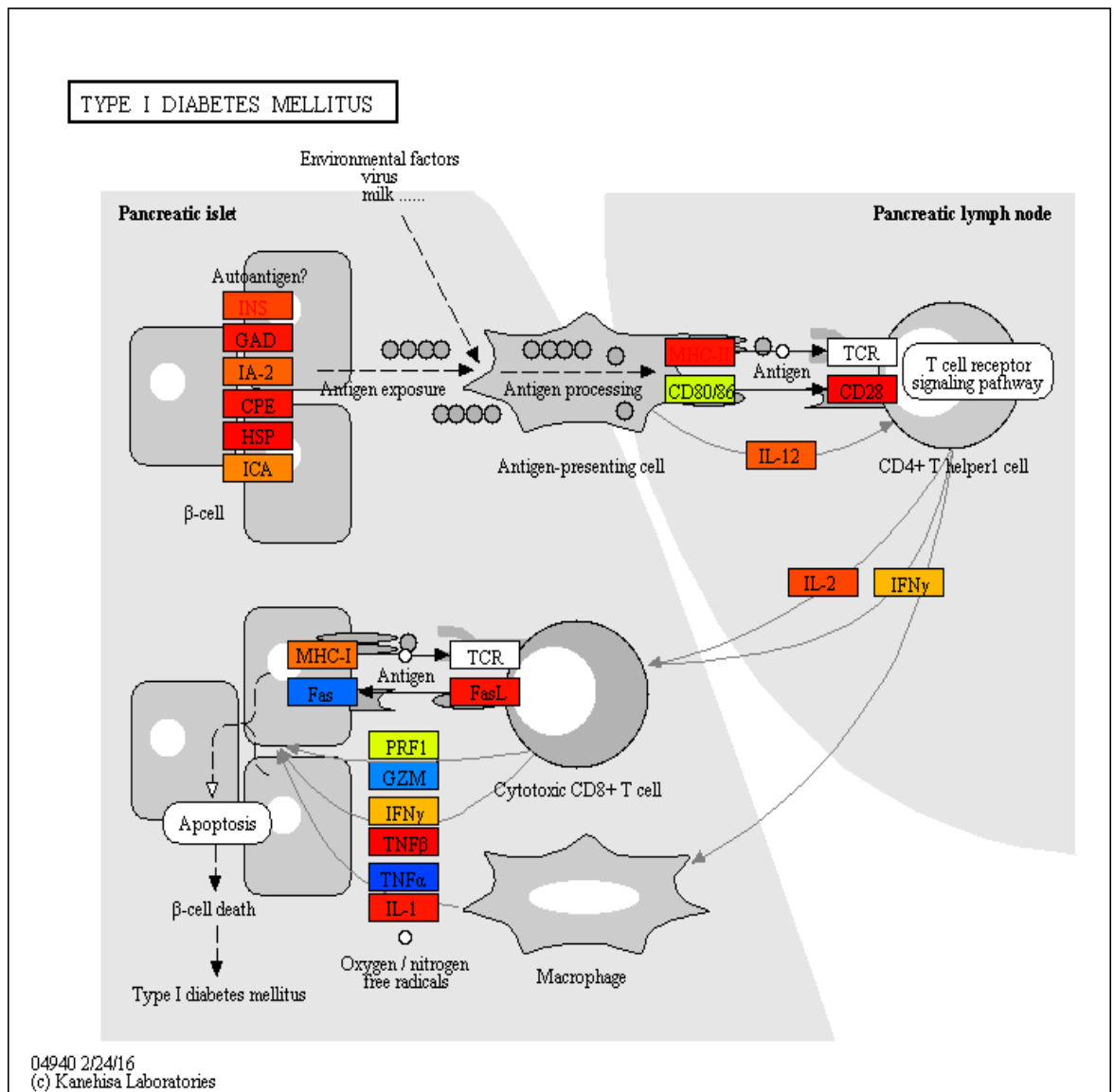


Рис. 4.7 Результаты анализа индекса DI сети Type I diabetes mellitus. Красные прямоугольники соответствуют генам, нуклеотидные последовательности которых идентичны последовательностям генов-ортологов, либо генам, у которых не найдено ортологов. Оранжевый, желтый, и синий прямоугольники соответствуют генам, показывающим повышенную изменчивость (с нарастанием изменчивости вдоль описанной линии цветов) по отношению к генам группы гоминид.

Рисунки 4.6 и 4.7 хорошо выделяют кластеры «древних» и «молодых» генов, а также гены с наибольшей изменчивостью относительно гоминид. Обычно именно такие гены наиболее интересны и важны при анализе сетей. Патогенез инсулин-зависимого диабета хорошо изучен, и именно гены

нижнего комплекса играют в нем ключевую роль [Yoon, Jun, 2006]. Ген *FAS* связан с регулированием апоптоза клеток. В исследованиях пациентов с диабетом первого типа он усиленно экспрессируется β -клетками и приводит к апоптозу клеток, расположенных в непосредственной близости к Fas-лиганд-экспрессирующим Т-лимфоцитам, проникающим в островки IDDM [Stassi и др., 1997]. Ген *GZMB* – уникальный для млекопитающих ген, также индуцирующий процесс апоптоза. Каких-либо специфических особенностей работы этого гена в рамках комплекса генов, связанных с диабетом первого типа, на данный момент не найдено, тем не менее, анализ показал высокую степень эволюционной изменчивости этого гена у человека относительно приматов. Если проанализировать все метаболические пути KEGG, в которых встречается этот ген, то можно увидеть, что они в большинстве своем связаны с процессом апоптоза и аутоиммунными заболеваниями, что сходится с нашими результатами о том, что в эти процессы вовлечены наиболее «молодые» и активно эволюционирующие гены.

В сети диабета второго типа оказалось гораздо меньше генов, на которые можно было бы обратить особое внимание (рисунки 4.8 и 4.9)

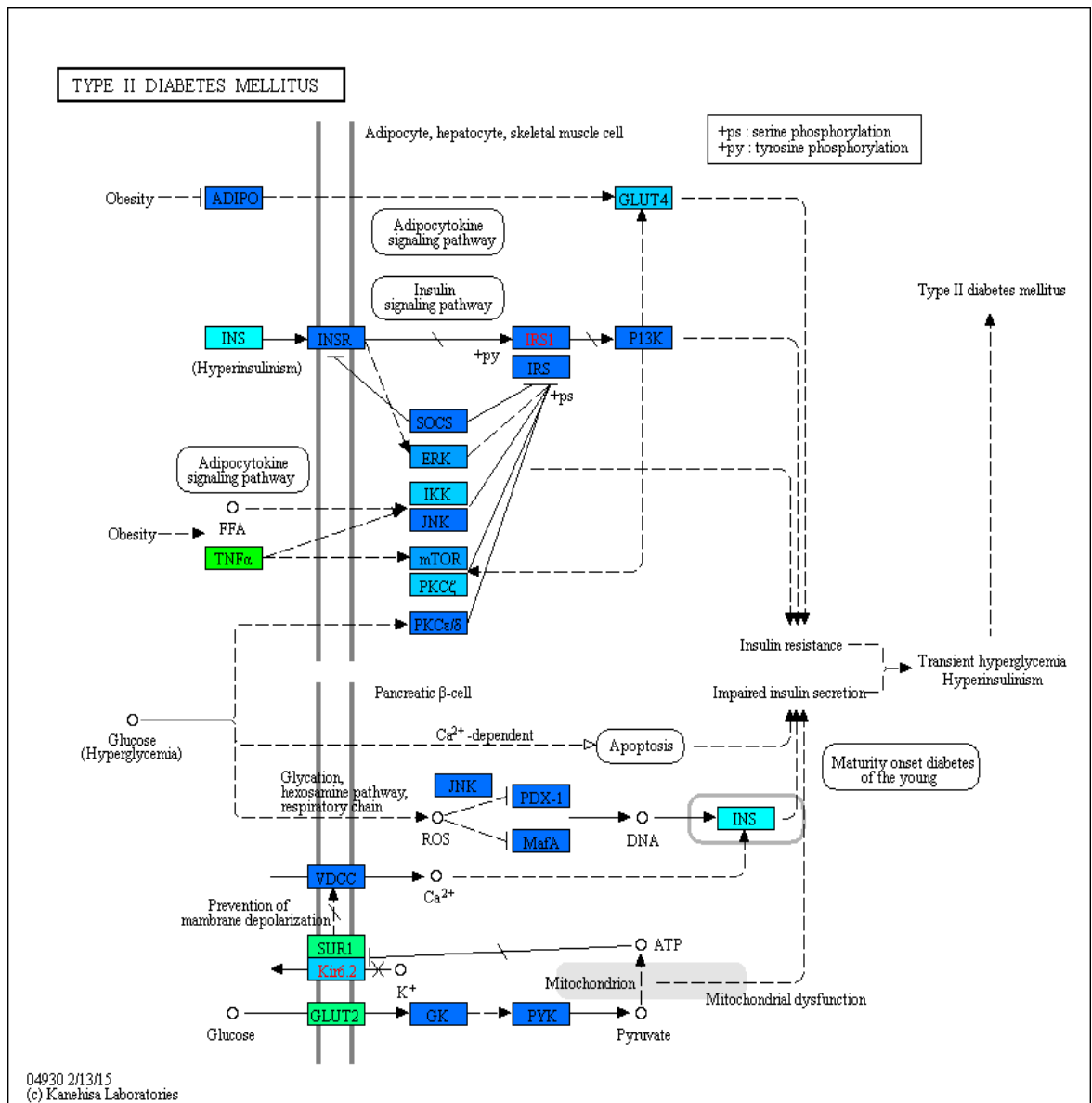


Рис. 4.8 *Результаты анализа индекса PAI сети Type II diabetes mellitus. Большинство генов обладают достаточно низким значением индекса PAI (покрашены голубым и синим цветами), выделяются гены TNF alpha, SUR1 и GLUT2.*

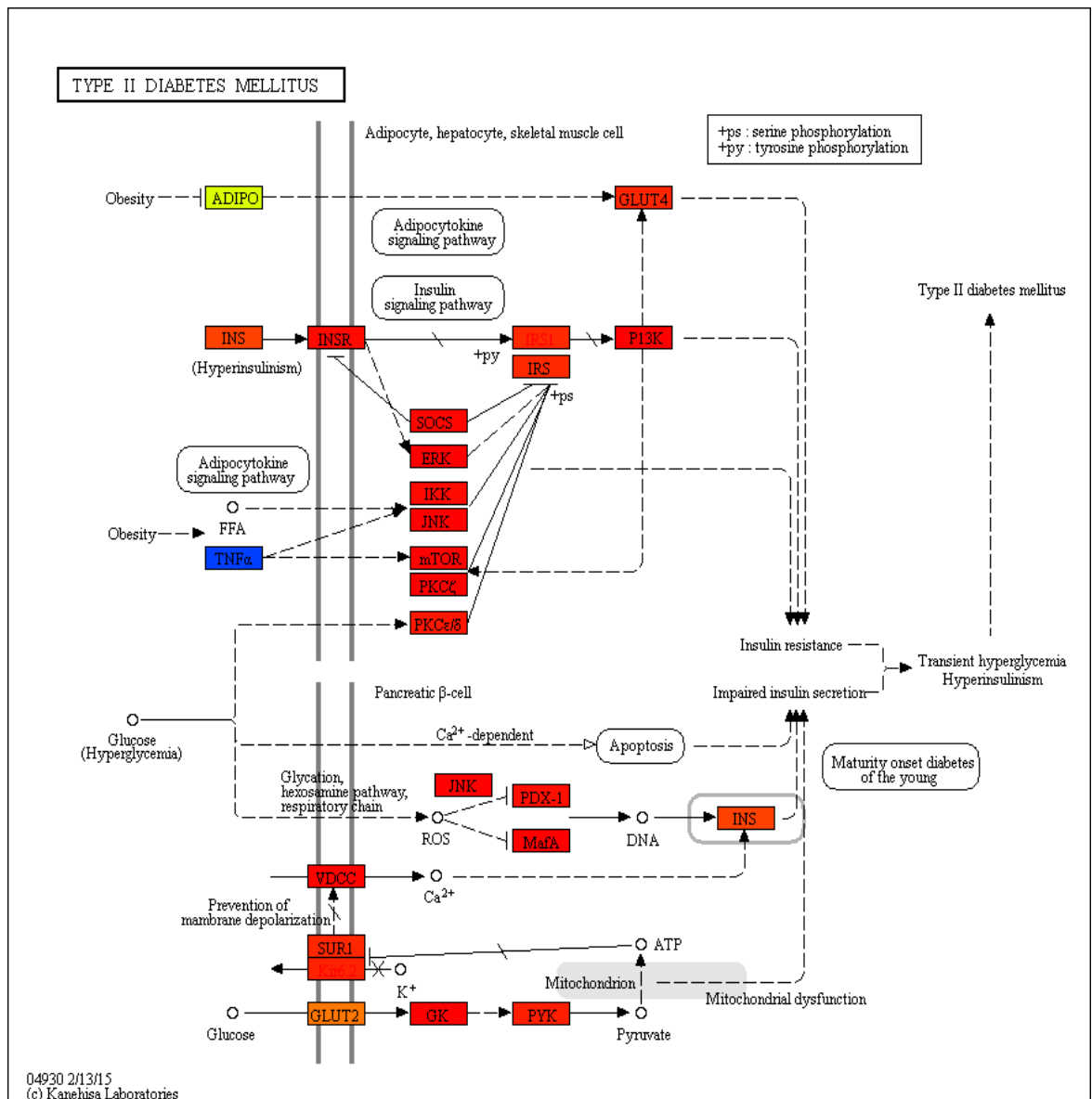


Рис. 4.9 Результаты анализа индекса DI сети Type II diabetes mellitus. Большинство генов обладают достаточно низким значением индекса DI (покрашены красным), выделяются гены ADIPO, TNF alpha и GLUT2.

Таким образом, с помощью Orthoscape из общей массы удалось выделить 4 гена. Белок TNF alpha - фактор некроза опухоли уже замечен в патогенезе различных заболеваний [Conway и др., 1997; Gao, Bataller, 2011; Palladino и др., 2003; Plevy и др., 1997], в том числе он был в диабете первого типа. Это один из самых распространенных элементов как среди сетей болезней человека (найден в 30 сетях этой группы), так и в базе KEGG Pathway в целом (найден в 53 метаболических путях). Ген ADIPO – адипонектин обладает

высоким уровнем сходства с *TNF alpha* и также уже отмечен как важный фактор в развитии диабета второго типа [Chu и др., 2013; Daimon и др., 2003; Maeda и др., 2002; Sanghera и др., 2010]. Ген *GLUT2* – ген из группы мембранных транспортных белков – *SLC* транспортеров. В нескольких работах на мышах показано, что слабая экспрессия гена *GLUT2* может привести к развитию диабета [Guillam и др., 1997; Valera и др., 1994]. Наконец, другой транспортный ген, *SUR1*, из семейства ABC транспортеров, пока не так активно обсуждается как важнейший участник в развития диабета второго типа, тем не менее, есть работы в этом направлении, в том числе и по мутациям гена *SUR1*, приводящим к развитию заболевания [Gloyn и др., 2003; Rafiq и др., 2008].

4.3 Заключение к главе 4

Анализ генных сетей болезней человека показал, что в сетях группы Immune diseases в большинстве своем задействованы гены, во-первых, обладающие высоким значением PAI, что говорит о том, что развитие их текущей функции началось относительно недавно в масштабе эволюции. Во-вторых, эти же гены обладают высоким значением DI при сравнении с генами группы гоминид, что говорит о том, что именно в них наблюдается максимальное отличие человека от других гоминид, т.е. гены активно эволюционируют. Аналогичный, но диаметрально противоположный, результат получен и для «древних» генов, в группе Substance dependence, что говорит о консервативности генов, описывающих процессы этой группы заболеваний.

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

Растения не могут избежать воздействия стрессовых условий. Они приспосабливаются к таким стрессовым факторам, как высокая или низкая температура, засуха и наводнения, загрязнение и засоленность почвы. У растений выработан набор механизмов устойчивости к стрессу, которые представляют собой различные процессы, включающие физиологические и биохимические изменения, приводящие к адаптивным морфологическим изменениям. Изучение механизмов адаптации растений к неблагоприятным факторам внешней среды, их реакции на стресс, представляет большой интерес для селекции сортов, приспособленных к конкретным условиям [Kumar и др., 2014; Matsukura и др., 2010]. Реакция растений на стрессовые факторы является сложной, как с точки зрения физиологических, так и молекулярных систем.

С помощью Orthoscape и Orthoweb были проанализированы списки генов, которые ассоциированы со стрессом у *A. thaliana* [Mustafin и др., 2019]. Работа проведена по следующим этапам:

- 1) На основе терминов генной онтологии, ассоциированных со стрессом, из базы данных TAIR v. 10 [Lamesch и др., 2012] были извлечены списки генов, ассоциированных со следующими типами стресса: “холодовой”, “тепловой”, “световой”, “осмотический”, “оксидативный”, “солевой”, “водный”.
- 2) С помощью ресурса String [Szklarczyk и др., 2017] на основе составленных списков были построены генные сети.
- 3) С помощью Orthoscape были проанализированы генные сети и посчитаны индексы PAI и DI. С помощью Orthoweb аналогичная работа была проведена для исходных списков генов, ассоциированных со стрессом.

4) С помощью языка программирования python и сервиса David 6.8 были проанализированы данные по содержанию рассматриваемых генов в терминах генной онтологии.

5.1 Подготовка списков генов и генных сетей

Исходные списки генов отбирались с использованием базы данных TAIR v. 10 [Lamesch и др., 2012]. Для каждого из семи рассматриваемых типов стресса было отобрано от 4 до 48 ассоциированных с ним терминов генной онтологии. На основе ассоциированных с терминами генов был составлен список всех встречающихся генов (Таблица 5.1).

Таблица 5.1 Число терминов генной онтологии, ассоциированных с каждым типом стресса, количество уникальных генов среди всех этих терминов и количество найденных для них карточек гена в KEGG.

Тип стресса	Число ГО терминов	Число генов	Число генов в KEGG
Холодовой	4	150	144
Тепловой	14	102	102
Световой	48	155	141
Осмотический	23	116	114
Окислительный	28	154	152
Солевой	17	231	230
Водный	27	215	211

Следующим этапом стал анализ генов с помощью ресурса String [<https://string-db.org/>] с целью выявить взаимодействия между генами.

Взаимодействия отбирались с настройками по умолчанию: 1) Внутренний отбор взаимодействий с «порогом уверенности», равным 0.7, что соответствует описанию "high confidence" (высокая уверенность). 2) К каждому списку добавляется 10 генов, взаимодействующих с исходно запрашиваемыми. На основании результатов String в Cytoscape были воспроизведены генные сети для каждого из семи типов стресса.

Кроме 7 списков генов, ассоциированных со стрессом, был построен восьмой список, содержащий гены *A. thaliana*, описанные в KEGG и кодирующие белки. Размер списка составил 27636 генов.

5.2 PAI и DI анализ списков генов, ассоциированных со стрессом.

На первом этапе были проанализированы исходные списки генов. Анализ PAI показал, что в списках генов, ассоциированных со стрессом, значительно выше доля эволюционно древних генов по сравнению с полным списком генов *A. thaliana*, кодирующих белки. Соответственно, для молодых генов была отмечена обратная ситуация (Рисунок 5.1).

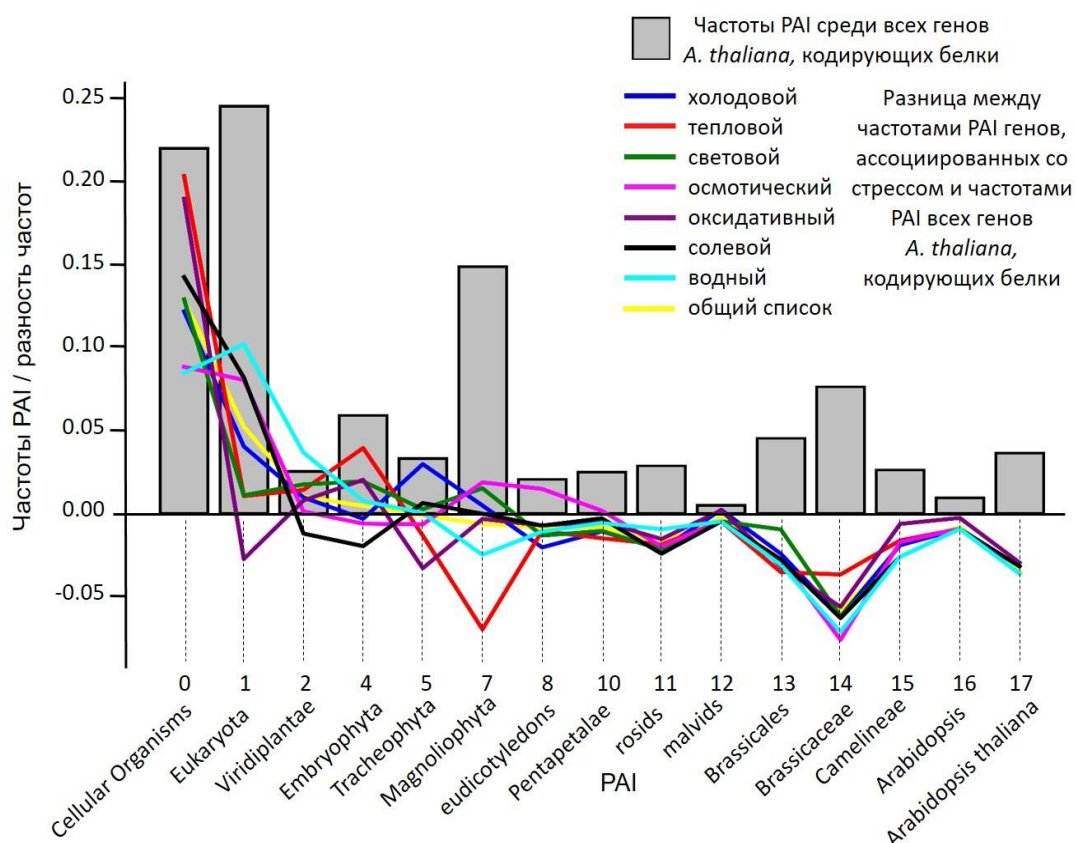


Рисунок 5.1. Серыми столбцами показано распределение частот встречаемости PAI среди всех генов *A. thaliana*, кодирующих белки. Цветными линиями показана разница между частотой PAI соответствующего цвету типа стресса и частотой аналогичного PAI среди всех генов *A. thaliana*.

Для определения достоверности анализа для каждого списка стрессовых генов было сгенерировано 100000 случайных выборок из всех генов *A. thaliana*, кодирующих белки. Размер каждой выборки соответствовал числу генов, ассоциированных со стрессом. Была определена доля случайно сгенерированных выборок, в которых 1) PAI оказался выше, чем у выборки генов, ассоциированных со стрессом 2) Значение критерия Хи-Квадрат распределения PAI значений генов случайной выборки выше аналогичного для выборки генов, ассоциированных со стрессом 3) Количество генов с конкретным значением PAI в случайной выборке выше, чем в выборке генов, ассоциированных со стрессом (для каждого PAI) (Таблица 5.2).

Таблица 5.2 Сравнение распределения PAI списков генов, ассоциированных со стрессом с распределением случайной выборки такого же размера из всех генов *A. thaliana*, кодирующих белки. В первой строке указан тип стресса. Во второй доля случайных выборок (из 100000), для которых среднее значение PAI_{rand} случайной выборки генов выше, чем PAI_{stress} стрессовой. В третьей доля случайных выборок, для которых значение критерия Хи-квадрат $ChiSq_{rand}$ выше, чем у аналогичной доли стрессовых выборок $ChiSq_{stress}$. В пятой и последующих строках доля случайных выборок, в которых число генов с соответствующим строке значением PAI превосходит аналогичное для стрессового набора генов. Значения долей умножены на 10^5 (число повторов эксперимента) для облегчения читаемости. Достоверными ($p\text{-value} < 0.05$) принимаются значения в таблице < 5000 (т.е. по данному показателю стрессовая выборка обладает достоверно меньшим значением) и > 95000 (достоверно большим).

Стресс	Холодовой	Тепловой	Световой	Осмотический	Окислительный	Солевой	Водный	Общий список
$p(PAI_{stress} < PAI_{rand})$	<u>100000</u>	<u>100000</u>	<u>100000</u>	<u>100000</u>	<u>100000</u>	<u>100000</u>	<u>100000</u>	<u>100000</u>
$p(ChiSq_{stress} < ChiSq_{rand})$	<u>163</u>	<u>153</u>	<u>303</u>	<u>2132</u>	<u>20</u>	<u>1</u>	<u>1</u>	<u>0*</u>
$p(dfPAI_{stress} < dfPAI_{rand})$ для конкретного таксона								
00_Cellular Organisms	<u>28</u>	<u>0*</u>	<u>11</u>	<u>1051</u>	<u>0*</u>	<u>0*</u>	<u>153</u>	<u>0*</u>
01_Eukaryota	11218	35449	34137	<u>2009</u>	75372	<u>194</u>	<u>34</u>	<u>78</u>
02_Viridiplantae	15729	11571	6603	32387	18710	83664	<u>96</u>	6394
04_Embryophyta	47489	<u>3778</u>	12708	51099	11320	87444	26046	25790
05_Tracheophyta	<u>2181</u>	65732	32104	51865	99386	22936	39481	53618
07_Magnoliophyta	38170	<u>97432</u>	26294	23881	49013	45825	82207	<u>100000</u>

08_eudicotyledons	94830	62021	78520	8542	60328	69323	80882	<u>100000</u>
10_Pentapetalae	69619	72175	68242	31311	52357	50365	60197	<u>97992</u>
11_rosids	91840	79304	91447	83906	81256	<u>99084</u>	72339	<u>99935</u>
12_malvids	15610	39144	49460	42420	16775	67429	63996	78703
13_Brassicales	89388	94844	61746	89433	<u>96971</u>	<u>97955</u>	<u>98746</u>	<u>100000</u>
14_Brassicaceae	<u>99913</u>	89314	<u>99872</u>	<u>99981</u>	<u>99765</u>	<u>99998</u>	<u>100000</u>	<u>100000</u>
15_Camelinaeae	89271	74850	<u>97699</u>	80248	56525	<u>99796</u>	<u>99627</u>	<u>100000</u>
16_Arabidopsis	74084	61724	73454	65695	41687	88476	86494	<u>99736</u>
17_A. thaliana	<u>99520</u>	<u>97657</u>	<u>99446</u>	<u>98525</u>	<u>97458</u>	<u>99811</u>	<u>99948</u>	<u>100000</u>

Таким образом показано, что древние гены достоверно перепредставлены, а молодые недопредставлены в выборках генов, ассоциированных со стрессом у *A. thaliana*.

Результаты аналогичного анализа для DI показаны на рисунке 5.2 и в таблице 5.3.

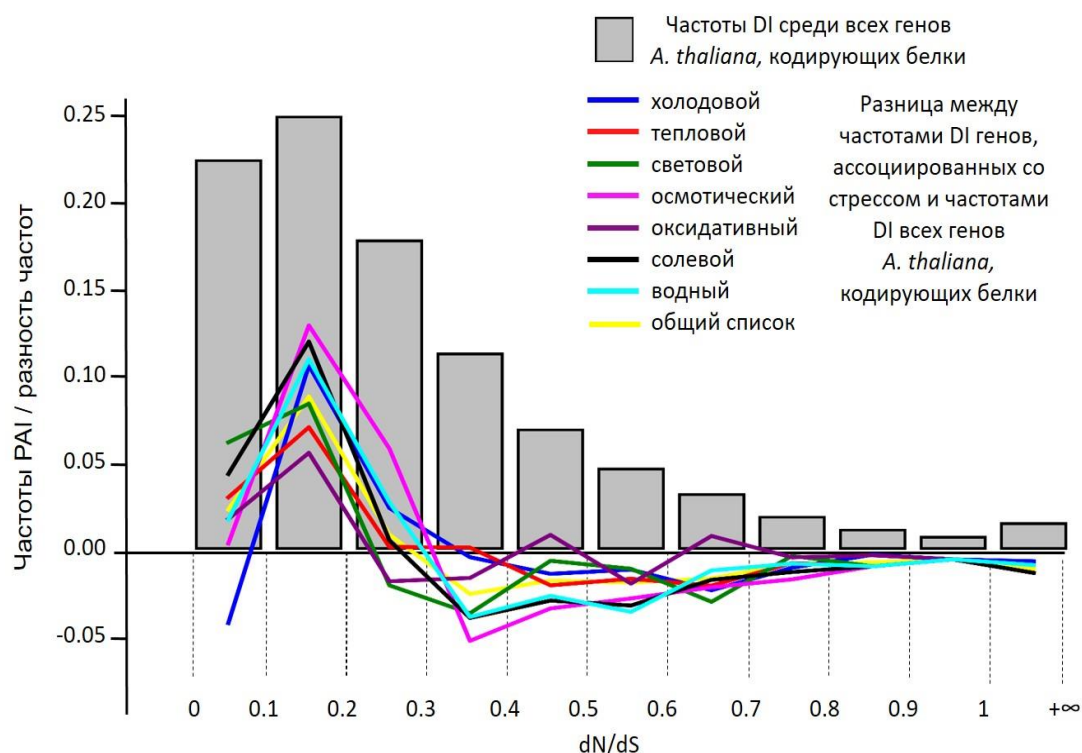


Рисунок 5.2 Серыми столбцами показано распределение частот встречаемости DI среди всех генов *A. thaliana*, кодирующих белки *A. thaliana*. Цветными линиями показана разница между аналогичной частотой DI соответствующего цвету типа стресса и частотой всех генов *A. thaliana*.

Таблица 5.3 Сравнение распределения DI списков генов, ассоциированных со стрессом с распределением случайной выборки такого же размера из всех генов *A. thaliana*, кодирующих белки той. В первой строке указан тип стресса. Во второй доля случайных выборок (из 100000), для которых среднее значение DI_{rand} случайной выборки генов выше, чем DI_{stress} стрессовой. В третьей доля случайных выборок, для которых значение критерия Хи-квадрат $ChiSq_{rand}$ выше, чем у аналогичной доли стрессовых выборок $ChiSq_{stress}$. В пятой и последующих строках доля случайных выборок, в которых число генов с соответствующим строке значением DI превосходит аналогичное для стрессового набора генов. Значения долей умножены на 10^5 (число повторов эксперимента) для облегчения читаемости. Достоверными ($p\text{-value} < 0.05$) принимаются значения в таблице < 5000 (т.е. по данному показателю

стрессовая выборка обладает достоверно меньшим значением) и >95000 (достоверно большим).

Стресс	Холодовой	Тепловой	Световой	Осмотический	Окислительный	Солевой	Водный	Общий список
$p(DI_{\text{stress}} < DI_{\text{rand}})$	<u>96272</u>	<u>99700</u>	<u>99993</u>	<u>99993</u>	<u>98140</u>	<u>100000</u>	<u>99999</u>	<u>100000</u>
$p(\chi^2_{\text{stress}} < \chi^2_{\text{rand}})$	7010	43398	<u>4707</u>	<u>589</u>	52740	<u>10</u>	<u>77</u>	<u>0*</u>
$p(N_{DI \text{ stress}} < N_{DI \text{ rand}})$ для конкретного DI столбца								
[0,0.1]	93375	28064	6137	53230	38498	10701	38527	<u>3109</u>
(0.1,0.2]	<u>67</u>	<u>2567</u>	<u>440</u>	<u>25</u>	<u>2633</u>	<u>0*</u>	<u>3</u>	<u>0*</u>
(0.2,0.3]	15832	38676	66383	<u>3174</u>	64490	32404	9731	18452
(0.3,0.4]	49562	40493	90569	<u>96058</u>	69027	<u>97027</u>	<u>96333</u>	<u>99395</u>
(0.4,0.5]	71783	75719	58004	92193	30113	<u>96979</u>	94603	<u>99451</u>
(0.5,0.6]	70023	73560	68603	92023	86198	<u>99567</u>	<u>99808</u>	<u>99973</u>
(0.6,0.7]	95009	84956	99071	88993	22528	93999	82333	<u>99957</u>
(0.7,0.8]	76327	58418	51318	88964	55802	93241	77350	<u>98435</u>
(0.8,0.9]	47813	67878	79538	72465	50554	92166	90674	<u>99617</u>
(0.9,1]	61938	49902	61515	53707	64025	78219	75853	<u>99712</u>
(1,+∞)	62703	78089	87708	81801	89300	<u>96544</u>	82213	<u>99985</u>

Таким образом показано, что древние гены достоверно перепредставлены, а молодые недопредставлены в выборках генов, ассоциированных со стрессом у *A. thaliana*.

с холодным, 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 Оchiaи [Ochiai, 1957] (Рисунок 5.4). В сети теплового стресса регуляторная компонента представлена слабее, а в сетях светового и оксидативного стрессов практически отсутствует.



Рисунок 5.4 Степень сходства наборов генов, ассоциированных с семью типами стресса. Дерево реконструировано с помощью метода UPGMA (unweighted pair group method with arithmetic mean), расстояния вычислены на основе коэффициентов Оchiaи. Снизу показан масштаб расстояния.

Была проанализирована и структура генной сети, а именно связь PAI гена и степени связности соответствующего ему узла сети k (Таблица 5.4).

Таблица 5.4 Коэффициент корреляции Пирсона $r(k, PAI)$ между PAI гена и степени связности соответствующего ему узла сети k для семи сетей стресса. Первая колонка – название стресса. Вторая – значение коэффициента корреляции. Третья – достоверность. Значения $p\text{-value} < 0.05$ выделены жирным шрифтом.

Стресс	$r(k, PAI)$	$p\text{-value}$
Холодовой	0.004	0.974
Тепловой	-0.361	0.003
Световой	-0.125	0.248
Осмотический	-0.379	0.006
Окислительный	0.019	0.875
Солевой	-0.266	0.006
Водный	-0.061	0.524

Была обнаружена отрицательная корреляция между степенью узла и PAI для трех типов стрессов. Это сети теплового, осмотического и солевого стрессов. В данных сетях отмечено большое число эволюционно древних генов с высоким числом связей. Для каждого типа стресса были построены точечные диаграммы, и диаграмма осмотического стресса лучше всего демонстрирует найденную зависимость (Рисунок 5.5).

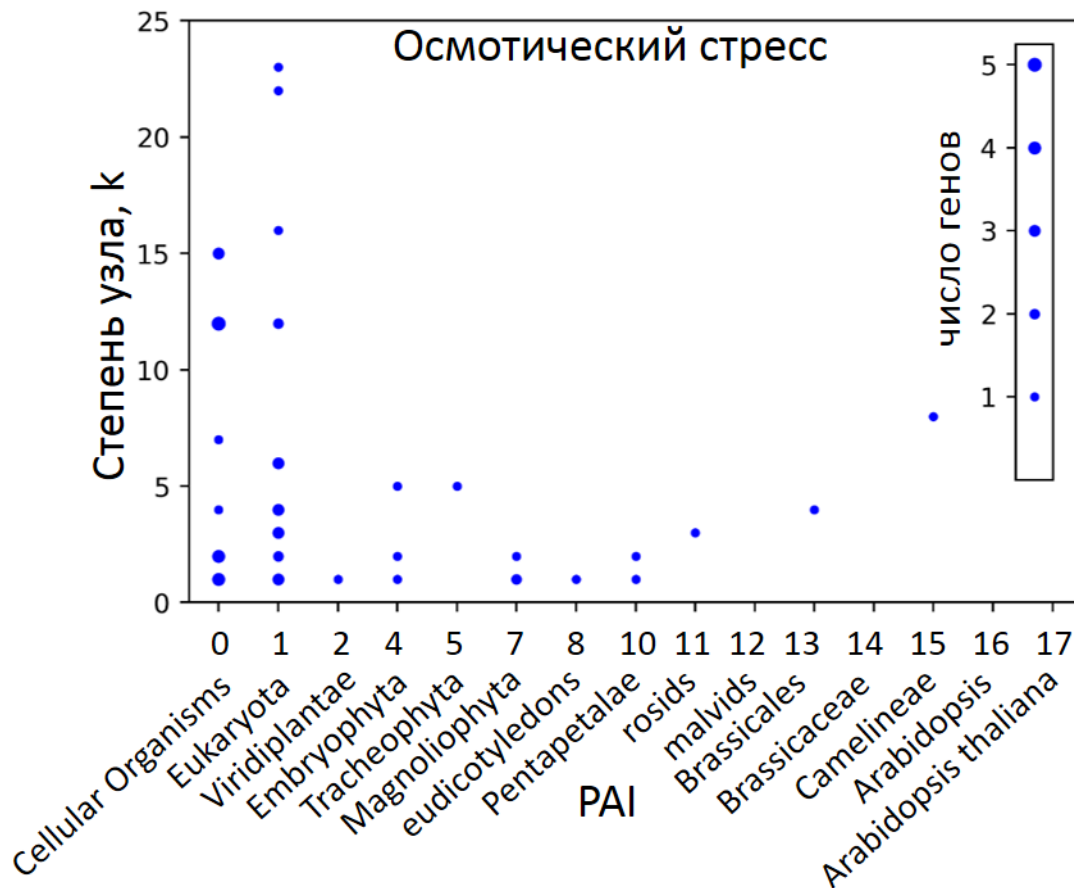


Рисунок 5.5 Точечная диаграмма распределения генов по признакам PAI и связности узлов в сети, эти гены описывающих, для осмотического стресса.

В сети осмотического стресса такие гены, как *ABI1/AT4G26080* gene ($k=23$) и *ABI2/AT5G57050* gene ($k=22$), принадлежащие семейству белковой Фосфотазы 2С, содержат большое число связей с другими генами (таксон Эукариоты).

Следует отметить, что не было обнаружено положительной корреляции ни для одного типа стресса, что говорит о том, что молодые гены не обладают большим числом связей с другими генами.

Подобные результаты позволяют предположить, что в процессе эволюции новые функции могут выполнять «молодые» гены, в том время как в основе сети лежит кластер эволюционно древних генов. В рамках данной гипотезы были проанализированы термины генной онтологии, с которыми

ассоциированы рассматриваемые гены, и PAI распределение этих генов. Т.е. как изменяется количество генов каждого таксона в рамках термина генной онтологии.

Для отбора терминов использовался сервис DAVID v6.8 (<https://david.ncifcrf.gov/>). Были отобраны все термины, с которыми значимо ассоциирован хотя бы один ген из списка. Значимость определялась по «суммарному» критерию DAVID и по критерию Бенджамини [Ferreira, Zwinderman, 2006], отбирались только значения с $p\text{-value} < 0.05$. На основе полученных результатов для каждого термина строились диаграммы распределения PAI ассоциированных с ним генов. Например, такая диаграмма для генов из списка *heat* отражена на рисунке 5.6

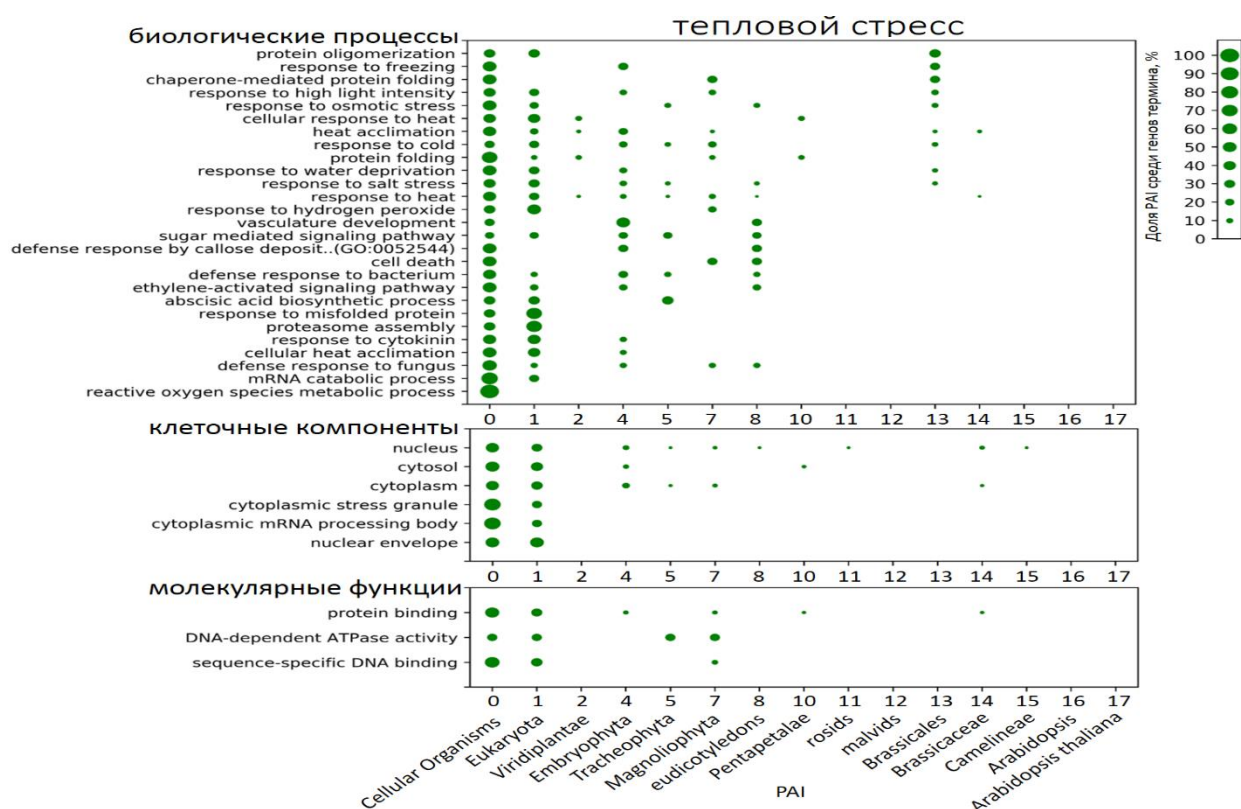


Рисунок 5.6 Частоты распределения PAI генов, ассоциированных с термином генной онтологии. Диаметр круга отражает долю значения PAI (по оси абсцисс) среди всех генов термина (по оси ординат). Термины сгруппированы по трем словарям: биологические процессы, клеточные компоненты, молекулярные функции.

Распределение показывает, что гены, входящие в список ассоциированных с одним типом стресса, функционально связаны сразу с несколькими. Например, для теплового стресса наблюдаются термины, функционально связанные с холодным стрессом (“response to freezing”, “response to cold”). С ними ассоциирован ген *TIL/AT5G58070* (temperature-induced lipocalin), важный компонент регуляции температурного режима, вероятно, действующий против перекисного окисления липидов, вызванного сильным тепловым стрессом [СНІ и др., 2009]. Кроме того, он ассоциирован с терминами “response to water deprivation”, “response to cytokinin”, “response to high light intensity”, “heat acclimation”, “response to heat”. Продукт гена *TIL1* – температурно-индуцированный липокалин-1 (ТИЛ1) – локализуется в плазматической мембране [Frenette Charron и др., 2005], его концентрация увеличивается в ответ на воздействие холода, таким образом, предположительно, выполняя защитную роль в условиях вызванной холодом дегидротации [Uemura и др., 2006]. Также показано, что ТИЛ1 перемещается под действием солевого стресса и защищает хлоропласты от ионной токсичности [Abo-Ogiala и др., 2014]. За счет подобных многофункциональных генов обеспечивается связь различных типов стресса общими терминами генной онтологии.

Также из рисунка 5.6 видно, что все термины, ассоциированные с генами теплового стресса, обязательно содержат в себе древние гены с $PAI = 0$ (Cellular Organisms) и почти все термины с $PAI = 1$ (Eukaryota). При этом генов со значениями PAI выше 8 значительно меньше. Таким образом, большая часть функций, ассоциированных с работой этих генов, связана с древними генами.

Наиболее необычными выглядят следующие результаты: например, для термина “vasculature development” найдены гены и со значением таксона (Embryophyta, eudicotyledons), т.е. описывающие высшие растения, и со значением Cellular organism. Т.е. гены, ассоциированные с процессами

развития сосудов уже существовали на стадии организмов, у которых сосуды еще не были сформированы. Посмотрев список генов, ассоциированных с термином, можно увидеть такие гены, как *ETR1/AT1G66340* (кодирует сигнальную трансдукцию гистидинкиназы), *TE1/AT2G26330* (кодирует лицин-обогащенный рецептор-подобный белок семейства протеинкиназ), *XRN4/AT1G54490* (кодирует экзорибонуклеазу 4), *EIN2/AT5G03280* (кодирует NRAMP белок семейства переносчиков ионов металлов). *TE1* и *ETR1* имеют PAI=4 (Embryophyta), *EIN2* имеет PAI=8 (eudicotyledons), и *XRN4* имеет PAI=0 (Cellular Organisms). Ген *XRN4*, также известный как *EIN5*, является эндогенным супрессором посттранскрипционного сайленсинга генов путем преимущественной деградации выбранных субстратов [Rymarquis, Souret, Green, 2011]. Он вовлечен в путь регуляции ответа на этилен [Potuschak и др., 2006], а также в регуляцию ответа на тепловой стресс. С использованием анализа микрочипов и распада мРНК, потеря функции *AtXRN4* приводит к снижению деградации фактора теплового шока A2 (*HSFA2*) и фактора 1 (*ERF1*) ответа на этилен [Nguyen и др., 2015]. Это обеспечивает увеличение выживаемости, когда растения подвергаются кратковременному сильному тепловому стрессу. В то же время *XRN4* необходим для регуляции теплообмена растений при длительном воздействии умеренно высокой температуры. Он участвует в деградации транскрипта *A. thaliana*, которая происходит на ранних этапах реакции на тепловой стресс [Merret и др., 2013]. В то же время, показано участие этого гена в регуляции деления сосудистых клеток [Etchells, Provost, Turner, 2012]. Таким образом, многофункциональный ген *XRN4* принимает участие в специфической деградации РНК, вовлечен в несколько различных биологических процессов, относящихся к реакции на стресс и развитию растительных тканей.

Аналогичные диаграммы были построены для всех 7 типов стресса. На всех диаграммах наблюдается тенденция вовлеченности многофункциональных эволюционно древних генов в ответ на стресс.

Полученные результаты свидетельствуют о многофункциональности эволюционно древних генов, участвующих в реакции на стресс. Эти гены участвуют в процессах формирования сосудов (гены, ассоциированные с тепловым стрессом), регуляции развития цветка (гены, ассоциированные со световым стрессом), развития семени (гены, ассоциированные с осмотическим стрессом), регуляция покоя семян (гены, ассоциированные с солевым стрессом). При этом известно, что эти процессы сформировались у растений на более позднем этапе эволюции [Linkies и др., 2010].

Заключение

В настоящей работе реализованы на языке Java приложение Orthoscape к программному комплексу Cytoscape и веб-приложение Orthoweb. Orthoscape позволяет проводить макро- и микроэволюционный анализ генных сетей и входящих в них генов внутри программного комплекса Cytoscape, Orthoweb позволяет анализировать списки генов, не объединенные в сеть. Приложения позволяют определить «возраст» гена на основании филогенетического анализа и поиска последнего общего предка для гена и его ортологов. В работе предложен филогенетический индекс гена (PAI), который в численном эквиваленте выражает «возраст» гена. Также приложения позволяют определять уровень эволюционной изменчивости гена с помощью вычисления отношения dN/dS , посчитанного для гена и его ортологов. В работе был предложен индекс дивергенции (DI), который в численном эквиваленте выражает совокупность значений dN/dS для гена и всех его ортологов.

С помощью Orthoscape были проанализированы сети биосинтеза стероидов (<https://www.genome.jp/kegg/pathway/hsa/hsa00100.html>) и биосинтеза стероидных гормонов (<https://www.genome.jp/kegg/pathway/hsa/hsa00140.html>) и показано, что сеть биосинтеза стероидов содержит большее число генов «среднего возраста» и меньшее число «древних» генов, чем сеть биосинтеза стероидов при использовании различных порогов идентичности нуклеотидных последовательностей гена и ортологов, что соответствует известным данным об эволюции эукариот.

С помощью Orthoscape были проанализированы генные сети различных заболеваний человека из базы KEGG Pathway, раздела Human Disease. Было показано, что сети заболеваний из категории Immune diseases (Иммунные заболевания) содержат наиболее «молодые» гены, в частности, самым

высоким показателем PAI характеризуется сеть Asthma (Астма). Сети категории Immune Disease также обладают наивысшим показателем DI (в анализе рассматривались только ортологи других гоминид), что говорит о том, что гены, вовлеченные в эти заболевания, эволюционируют быстрее остальных генов по сравнению с гомологичными генами наиболее близкородственных организмов.

Было показано, что сети заболеваний из категории Substance dependence (Зависимость от веществ, вызывающих привыкание) описаны наиболее «древними» генами в частности, самым низким показателем PAI характеризуется сеть Nicotine addiction (Никотиновая зависимость). Таким образом, сети этой категории, содержат в себе гены, ортологи которых найдены у эволюционно более древних 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.

С помощью Orthoweb были проанализированы гены *A. thaliana*, ассоциированные с ответом на абиотический стресс. Анализ PAI и DI показал, что в выборках генов *A. thaliana*, ассоциированных с реакцией на абиотические стрессы, эволюционно древние и консервативные гены достоверно перепредставлены в сравнении со всеми генами *A. thaliana*, кодирующими белки.

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 ($r = 0,876$, $p\text{-value} < 1,8 \times 10^{-26}$) 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.

Список литературы

1. Abo-Ogiala A. и др. Temperature-induced lipocalin (TIL) is translocated under salt stress and protects chloroplasts from ion toxicity // J. Plant Physiol. 2014. T. 171. № 3–4. С. 250–259.
2. Alon U. Biological Networks: The Tinkerer as an Engineer // Science (80-.). 2003. T. 301. № 5641.
3. Altschul S. F. и др. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. // Nucleic Acids Res. 1997. T. 25. № 17. С. 3389–402.
4. Arendsee Z. W., Li L., Wurtele E. S. Coming of age: orphan genes in plants // Trends Plant Sci. 2014. T. 19. № 11. С. 698–708.
5. Baker M. E. Evolution of adrenal and sex steroid action in vertebrates: a ligand-based mechanism for complexity // BioEssays. 2003. T. 25. № 4. С. 396–400.
6. Baptiste E., Huneman P. Towards a Dynamic Interaction Network of Life to unify and expand the evolutionary theory // BMC Biol. 2018. T. 16. № 1. С. 56.
7. Barabási A.-L., Gulbahce N., Loscalzo J. Network medicine: a network-based approach to human disease // Nat. Rev. Genet. 2011. T. 12. № 1. С. 56–68.
8. Barrett T. и др. NCBI GEO: archive for high-throughput functional genomic data // Nucleic Acids Res. 2009. T. 37. № Database. С. D885–D890.
9. Bascompte J., Melián C. J. SIMPLE TROPHIC MODULES FOR COMPLEX FOOD WEBS // Ecology. 2005. T. 86. № 11. С. 2868–2873.
10. Benson D. A. и др. GenBank // Nucleic Acids Res. 2018. T. 46. № D1. С. D41–D47.
11. Bindea G. и др. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks // Bioinformatics. 2009. T. 25. № 8. С. 1091–1093.
12. Bindea G., Galon J., Mlecnik B. CluePedia Cytoscape plugin: pathway

insights using integrated experimental and in silico data // Bioinformatics. 2013. T. 29. № 5. С. 661–663.

13. Brown K. R., Jurisica I. Online Predicted Human Interaction Database // Bioinformatics. 2005. T. 21. № 9. С. 2076–2082.

14. Cerami E. G. и др. Pathway Commons, a web resource for biological pathway data // Nucleic Acids Res. 2011. T. 39. № SUPPL. 1. С. 685–690.

15. Chatr-aryamontri A. и др. The BioGRID interaction database: 2017 update // Nucleic Acids Res. 2017. T. 45. № D1. С. D369–D379.

16. Chen M. и др. Big Data Analysis. : Springer, Cham, 2014. С. 51–58.

17. Chen W.-H. и др. Human Monogenic Disease Genes Have Frequently Functionally Redundant Paralogs // PLoS Comput. Biol. 2013. T. 9. № 5. С. e1003073.

18. Cheng F., Kovács I. A., Barabási A.-L. Network-based prediction of drug combinations // Nat. Commun. 2019. T. 10. № 1. С. 1197.

19. CHI W.-T. и др. Temperature-induced lipocalin is required for basal and acquired thermotolerance in Arabidopsis // Plant. Cell Environ. 2009. T. 32. № 7. С. 917–927.

20. Chu H. и др. *AdipoQ* polymorphisms are associated with type 2 diabetes mellitus: a meta-analysis study // Diabetes. Metab. Res. Rev. 2013. T. 29. № 7. С. n/a-n/a.

21. Conway D. J. и др. Scarring trachoma is associated with polymorphism in the tumor necrosis factor alpha (TNF-alpha) gene promoter and with elevated TNF-alpha levels in tear fluid. // Infect. Immun. 1997. T. 65. № 3. С. 1003–6.

22. D'Antonio M., Ciccarelli F. D. Modification of Gene Duplicability during the Evolution of Protein Interaction Network // PLoS Comput. Biol. 2011. T. 7. № 4. С. e1002029.

23. Dächsel J. C., Farrer M. J. LRRK2 and Parkinson Disease // Arch. Neurol. 2010. T. 67. № 5. С. 542–547.

24. Daimon M. и др. Decreased Serum Levels of Adiponectin Are a Risk Factor for the Progression to Type 2 Diabetes in the Japanese Population // Diabetes

Care. 2003. T. 26. № 7.

25. Demir E. и др. The BioPAX community standard for pathway data sharing // Nat. Biotechnol. 2010. T. 28. № 12. С. 1308–1308.

26. Deng H. и др. The Parkinson's disease genes pink1 and parkin promote mitochondrial fission and/or inhibit fusion in Drosophila. // Proc. Natl. Acad. Sci. U. S. A. 2008. T. 105. № 38. С. 14503–8.

27. Dickerson J. E. и др. Defining the Role of Essential Genes in Human Disease // PLoS One. 2011. T. 6. № 11. С. e27368.

28. Domazet-Lošo T., Brajković J., Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages // Trends Genet. 2007. T. 23. № 11. С. 533–539.

29. Domazet-Lošo T., Tautz D. An Ancient Evolutionary Origin of Genes Associated with Human Genetic Diseases // Mol. Biol. Evol. 2008. T. 25. № 12. С. 2699–2707.

30. Domazet-Lošo T., Tautz D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns // Nature. 2010a. T. 468. № 7325. С. 815–818.

31. Domazet-Lošo T., Tautz D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa // BMC Biol. 2010b. T. 8. № 1. С. 66.

32. Drost H.-G. и др. Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis // Mol. Biol. Evol. 2015. T. 32. № 5. С. 1221–1231.

33. Dufty A. Hormones, developmental plasticity and adaptation // Trends Ecol. Evol. 2002. T. 17. № 4. С. 190–196.

34. Dunne J. A., Williams R. J., Martinez N. D. Food-web structure and network theory: The role of connectance and size. // Proc. Natl. Acad. Sci. U. S. A. 2002. T. 99. № 20. С. 12917–22.

35. Eaton S. L., Wishart T. M. Bridging the gap: large animal models in neurodegenerative research // Mamm. Genome. 2017. T. 28. № 7–8. С. 324–337.

36. Emborg M. E. Nonhuman primate models of Parkinson's disease. // ILAR J. 2007. T. 48. № 4. С. 339–55.
37. Etchells J. P., Provost C. M., Turner S. R. Plant Vascular Cell Division Is Maintained by an Interaction between PXY and Ethylene Signalling // PLoS Genet. 2012. T. 8. № 11. С. e1002997.
38. Falagas M. E. и др. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses // FASEB J. 2008. T. 22. № 2. С. 338–342.
39. Ferreira J. A., Zwinderman A. H. On the Benjamini–Hochberg method // Ann. Stat. 2006. T. 34. № 4. С. 1827–1849.
40. Flicek P. и др. Ensembl 2012 // Nucleic Acids Res. 2012. T. 40. № D1. С. D84–D90.
41. Frenette Charron J.-B. и др. Identification, Expression, and Evolutionary Analyses of Plant Lipocalins // Plant Physiol. 2005. T. 139. № 4. С. 2017–2028.
42. Gandhi S. и др. PINK1 protein in normal human brain and Parkinson's disease // Brain. 2006. T. 129. № 7. С. 1720–1731.
43. Gandhi S. и др. PINK1-Associated Parkinson's Disease Is Caused by Neuronal Vulnerability to Calcium-Induced Cell Death // Mol. Cell. 2009. T. 33. № 5. С. 627–638.
44. Gao B., Bataller R. Alcoholic Liver Disease: Pathogenesis and New Therapeutic Targets // Gastroenterology. 2011. T. 141. № 5. С. 1572–1585.
45. Gehlenborg N. и др. Visualization of omics data for systems biology // Nat. Methods. 2010. T. 7. № 3s. С. S56–S68.
46. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource // Nucleic Acids Res. 2004. T. 32. № 90001. С. 258D – 261.
47. Gene Ontology Consortium. Gene Ontology Consortium: going forward // Nucleic Acids Res. 2015. T. 43. № D1. С. D1049–D1056.
48. Gerlt J. A., Babbitt P. C. Orthologs and paralogs - we need to get it right // Genome Biol. 2000. T. 1. № 5. С. reviews0005.1.
49. Gilks W. P. и др. A common LRRK2 mutation in idiopathic Parkinson's

disease // *Lancet*. 2005. T. 365. № 9457. C. 415–416.

50. Gloyn A. L. и др. Large-Scale Association Studies of Variants in Genes Encoding the Pancreatic β -Cell KATP Channel Subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) Confirm That the KCNJ11 E23K Variant Is Associated With Type 2 Diabetes // *Diabetes*. 2003. T. 52. № 2.

51. Guillam M.-T. и др. Early diabetes and abnormal postnatal pancreatic islet development in mice lacking Glut-2 // *Nat. Genet.* 1997. T. 17. № 3. C. 327–330.

52. Hakes L. и др. Protein-protein interaction networks and biology—what's the connection? // *Nat. Biotechnol.* 2008. T. 26. № 1. C. 69–72.

53. Hartmann A. и др. Caspase-3: A vulnerability factor and final effector in apoptotic death of dopaminergic neurons in Parkinson's disease. // *Proc. Natl. Acad. Sci. U. S. A.* 2000. T. 97. № 6. C. 2875–80.

54. Hau M. Regulation of male traits by testosterone: implications for the evolution of vertebrate life histories // *BioEssays*. 2007. T. 29. № 2. C. 133–144.

55. Hu J. и др. Cellular cholesterol delivery, intracellular processing and utilization for biosynthesis of steroid hormones // *Nutr. Metab. (Lond)*. 2010. T. 7. № 1. C. 47.

56. Huang D. W., Sherman B. T., Lempicki R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources // *Nat. Protoc.* 2009. T. 4. № 1. C. 44–57.

57. Hurst L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution // *Trends Genet.* 2002. T. 18. № 9. C. 486–487.

58. Ignatieva E. V., Afonnikov D. A., Kolchanov N. A. Online resources on gene networks containing human and animal data // *Vavilovskii Zhurnal Genet. Seleksii*. 2017. T. 21. № 8.

59. Ivanisenko V. A. и др. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology // *BMC Syst. Biol.* 2015. T. 9. № Suppl 2. C. S2.

60. K. Sasaki, A. Tsutsumi N. W. Mannose-Binding Lectin Polymorphisms in Patients with Hepatitis C Virus Infection // *Scand. J. Gastroenterol.* 2000. T. 35. №

9. C. 960–965.

61. Kanehisa M. и др. KEGG: New perspectives on genomes, pathways, diseases and drugs // *Nucleic Acids Res.* 2017. Т. 45. № D1. С. D353–D361.

62. Kann M. G. Advances in translational bioinformatics: computational approaches for the hunting of disease genes // *Brief. Bioinform.* 2010. Т. 11. № 1. С. 96–110.

63. Kauffman S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. // *J. Theor. Biol.* 1969. Т. 22. № 3. С. 437–67.

64. Kéfi S. и др. Network structure beyond food webs: mapping non-trophic and trophic interactions on Chilean rocky shores // *Ecology.* 2015. Т. 96. № 1. С. 291–303.

65. Khakoo S. I. HLA and NK Cell Inhibitory Receptor Genes in Resolving Hepatitis C Virus Infection // *Science (80-.).* 2004. Т. 305. № 5685. С. 872–874.

66. Koski L. B., Golding G. B. The Closest BLAST Hit Is Often Not the Nearest Neighbor // *J. Mol. Evol.* 2001. Т. 52. № 6. С. 540–542.

67. Kovács I. A. и др. Network-based prediction of protein interactions // *Nat. Commun.* 2019. Т. 10. № 1. С. 1240.

68. Kryazhimskiy S., Plotkin J. B. The population genetics of dN/dS // *PLoS Genet.* 2008. Т. 4. № 12.

69. Kumar T. и др. Genetic Improvement of Sugarcane for Drought and Salinity Stress Tolerance Using Arabidopsis Vacuolar Pyrophosphatase (AVP1) Gene // *Mol. Biotechnol.* 2014. Т. 56. № 3. С. 199–209.

70. Lamesch P. и др. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools // *Nucleic Acids Res.* 2012. Т. 40. № D1. С. D1202–D1210.

71. Lesne A. Complex Networks: from Graph Theory to Biology // *Lett. Math. Phys.* 2006. Т. 78. № 3. С. 235–262.

72. Li W.-H. Unbiased estimation of the rates of synonymous and nonsynonymous substitution // *J. Mol. Evol.* 1993. Т. 36. № 1. С. 96–99.

73. Li W. H., Wu C. I., Luo C. C. A new method for estimating synonymous

and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. // *Mol. Biol. Evol.* 1985. Т. 2. № 2. С. 150–174.

74. Linkies A. и др. The evolution of seeds // *New Phytol.* 2010. Т. 186. № 4. С. 817–831.

75. Maeda N. и др. Diet-induced insulin resistance in mice lacking adiponectin/ACRP30 // *Nat. Med.* 2002. Т. 8. № 7. С. 731–737.

76. Maere S., Heymans K., Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks // *Bioinformatics.* 2005. Т. 21. № 16. С. 3448–3449.

77. Mata I. F. и др. LRRK2 in Parkinson's disease: protein domains and functional insights // *Trends Neurosci.* 2006. Т. 29. № 5. С. 286–293.

78. Matsukura S. и др. Comprehensive analysis of rice DREB2-type genes that encode transcription factors involved in the expression of abiotic stress-responsive genes // *Mol. Genet. Genomics.* 2010. Т. 283. № 2. С. 185–196.

79. Merret R. и др. XRN4 and LARP1 Are Required for a Heat-Triggered mRNA Decay Pathway Involved in Plant Acclimation and Survival during Thermal Stress // *Cell Rep.* 2013. Т. 5. № 5. С. 1279–1293.

80. Mizuno Y. и др. Parkin and Parkinson's disease // *Curr. Opin. Neurol.* 2001. Т. 14. № 4. С. 477–482.

81. Montojo J. и др. GeneMANIA cytoscape plugin: Fast gene function predictions on the desktop // *Bioinformatics.* 2010. Т. 26. № 22. С. 2927–2928.

82. Moyers B. A., Zhang J. Phylostratigraphic Bias Creates Spurious Patterns of Genome Evolution // *Mol. Biol. Evol.* 2015. Т. 32. № 1. С. 258–267.

83. Mustafin Z. S. и др. Orthoscape: a cytoscape application for grouping and visualization KEGG based gene networks by taxonomy and homology principles // *BMC Bioinformatics.* 2017. Т. 18. № S1. С. 1–9.

84. Mustafin Z. S. и др. Phylostratigraphic Analysis Shows the Earliest Origination of the Abiotic Stress Associated Genes in *A. thaliana* // *Genes (Basel).* 2019. Т. 10. № 12. С. 963.

85. Needleman S. B., Wunsch C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins // J. Mol. Biol. 1970. T. 48. № 3. C. 443–453.
86. NEEL J. V. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? // Am. J. Hum. Genet. 1962. T. 14. № 4. C. 353–62.
87. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutionsojobori 1986 // Mol. Biol. Evol. 1986. T. 3. C. 418–426.
88. Nekrutenko A., Makova K. D., Li W. The K A / K S Ratio Test for Assessing the Protein-Coding Potential of Genomic Regions : An Empirical and Simulation Study // Genome Res. 2001. C. 198–202.
89. Nersisyan L., Samsonyan R., Arakelyan A. CyKEGGParser: tailoring KEGG pathways to fit into systems biology analysis workflows // F1000Research. 2014. T. 145. № May. C. 1–11.
90. Nguyen A. H. и др. Loss of Arabidopsis 5'–3' Exoribonuclease AtXRN4 Function Enhances Heat Stress Tolerance of Plants Subjected to Severe Heat Stress // Plant Cell Physiol. 2015. T. 56. № 9. C. 1762–1772.
91. Nikitin A. и др. Pathway studio--the analysis and navigation of molecular networks // Bioinformatics. 2003. T. 19. № 16. C. 2155–2157.
92. O'Donovan C. и др. High-quality protein knowledge resource: SWISS-PROT and TrEMBL // Brief. Bioinform. 2002. T. 3. № 3. C. 275–284.
93. Ochiai A. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions // Bull. Japanese Soc. Sci. Fish. 1957. T. 22. C. 526–530.
94. Palladino M. A. и др. Anti-TNF-alpha therapies: The next generation // Nat. Rev. Drug Discov. 2003. T. 2(9). C. 736–746.
95. Palmieri N., Kosiol C., Schlötterer C. The life cycle of Drosophila orphan genes // Elife. 2014. T. 3. C. e01311.
96. Pamilo P., Bianchi N. O. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. // Mol. Biol. Evol. 1993. T. 10. № 2. C. 271–

281.

97. Plevy S. E. и др. A role for TNF-alpha and mucosal T helper-1 cytokines in the pathogenesis of Crohn's disease. // J. Immunol. 1997. Т. 159. № 12.

98. Porras G., Li Q., Bezard E. Modeling Parkinson's disease in primates: The MPTP model. // Cold Spring Harb. Perspect. Med. 2012. Т. 2. № 3. С. a009308.

99. Potuschak T. и др. The Exoribonuclease XRN4 Is a Component of the Ethylene Response Pathway in Arabidopsis // Plant Cell. 2006. Т. 18. № 11. С. 3047–3057.

100. Quint M. и др. A transcriptomic hourglass in plant embryogenesis // Nature. 2012. Т. 490. № 7418. С. 98–101.

101. Rafiq M. и др. Effective Treatment With Oral Sulfonylureas in Patients With Diabetes Due to Sulfonylurea Receptor 1 (SUR1) Mutations // Diabetes Care. 2008. Т. 31. № 2.

102. Rigden D. J., Fernández X. M. The 27th annual Nucleic Acids Research database issue and molecular biology database collection // Nucleic Acids Res. 2020. Т. 48. № D1. С. D1–D8.

103. Ruprecht C. и др. Phylogenomic analysis of gene co-expression networks reveals the evolution of functional modules // Plant J. 2017. Т. 90. № 3. С. 447–465.

104. Rymarquis L. A., Souret F. F., Green P. J. Evidence that XRN4, an Arabidopsis homolog of exoribonuclease XRN1, preferentially impacts transcripts with certain sequences or in particular functional categories // RNA. 2011. Т. 17. № 3. С. 501–511.

105. Sanghera D. K. и др. PPARG and ADIPOQ gene polymorphisms increase type 2 diabetes mellitus risk in Asian Indian Sikhs: Pro12Ala still remains as the strongest predictor // Metabolism. 2010. Т. 59. № 4. С. 492–501.

106. Shah N. и др. Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows // Bioinformatics. 2019. Т. 35. № 9. С. 1613–1614.

107. Shannon P. и др. Cytoscape: A software Environment for integrated models of biomolecular interaction networks // Genome Res. 2003. Т. 13. № 11. С.

2498–2504.

108. Smith T. F., Waterman M. S. Identification of common molecular subsequences // *J. Mol. Biol.* 1981. T. 147. № 1. С. 195–197.

109. Sonnhammer E. L. L. и др. Big data and other challenges in the quest for orthologs // *Bioinformatics.* 2014. T. 30. № 21. С. 2993–2998.

110. Stassi G. и др. Nitric oxide primes pancreatic beta cells for Fas-mediated destruction in insulin-dependent diabetes mellitus. // *J. Exp. Med.* 1997. T. 186. № 8. С. 1193–200.

111. Strauss J. F., Martinez F., Kiriakidou M. Placental Steroid Hormone Synthesis: Unique Features and Unanswered Questions1 // *Biol. Reprod.* 1996. T. 54. № 2. С. 303–311.

112. Studer R. A., Robinson-Rechavi M. How confident can we be that orthologs are similar, but paralogs differ? // *Trends Genet.* 2009. T. 25. № 5. С. 210–216.

113. Summons R. E. и др. Steroids, triterpenoids and molecular oxygen // *Philos. Trans. R. Soc. B Biol. Sci.* 2006. T. 361. № 1470. С. 951–968.

114. Szklarczyk D. и др. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible // *Nucleic Acids Res.* 2017. T. 45. № D1. С. D362–D368.

115. Szklarczyk D. и др. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets // *Nucleic Acids Res.* 2019. T. 47. № D1. С. D607–D613.

116. Tatton N. A. Increased Caspase 3 and Bax Immunoreactivity Accompany Nuclear GAPDH Translocation and Neuronal Apoptosis in Parkinson's Disease // *Exp. Neurol.* 2000. T. 166. № 1. С. 29–43.

117. Tautz D., Domazet-Lošo T. The evolutionary origin of orphan genes // *Nat. Rev. Genet.* 2011. T. 12. № 10. С. 692–702.

118. Tekaiia F. Inferring Orthologs : Open Questions and Perspectives Fredj Tekaiia Inferring Orthologs : Open Questions and Perspectives // 2016. С. 17–28.

119. Thornton J. W. Resurrecting the Ancestral Steroid Receptor: Ancient

Origin of Estrogen Signaling // *Science* (80-.). 2003. Т. 301. № 5640. С. 1714–1717.

120. Uemura M. и др. Responses of the plasma membrane to low temperatures // *Physiol. Plant.* 2006. Т. 126. № 1. С. 81–89.

121. Valente E. M. и др. Hereditary Early-Onset Parkinson's Disease Caused by Mutations in PINK1 // *Science* (80-.). 2004. Т. 304. № 5674.

122. Valera A. и др. Expression of GLUT-2 antisense RNA in beta cells of transgenic mice leads to diabetes. // *J. Biol. Chem.* 1994. Т. 269. № 46. С. 28543–6.

123. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood // *Mol. Biol. Evol.* 2007. Т. 24. № 8. С. 1586–1591.

124. Yang Z., Nielsen R. Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models // *Mol. Biol. Evol.* 2000. Т. 17. № 1. С. 32–43.

125. Yoon J.-W., Jun H.-S. Cellular and Molecular Pathogenic Mechanisms of Insulin-Dependent Diabetes Mellitus // *Ann. N. Y. Acad. Sci.* 2006. Т. 928. № 1. С. 200–211.

126. Zerbino D. R. и др. Ensembl 2018 // *Nucleic Acids Res.* 2018. Т. 46. № D1. С. D754–D761.

127. Zhang L. и др. Phylostratigraphic analysis of gene co-expression network reveals the evolution of functional modules for ovarian cancer // *Sci. Rep.* 2019. Т. 9. № 1. С. 2623.

128. Zhang Z. и др. KaKs_Calculator: Calculating Ka and Ks Through Model Selection and Model Averaging // *Genomics. Proteomics Bioinformatics.* 2006. Т. 4. № 4. С. 259–263.

129. Zheleznikova G. F. Infection and immunity: strategies from both sides // *Med. Immunol.* 2014. Т. 8. № 5–6. С. 597.

130. Zimprich A. и др. Mutations in LRRK2 Cause Autosomal-Dominant Parkinsonism with Pleomorphic Pathology // *Neuron.* 2004. Т. 44. № 4. С. 601–607.

131. Афонников Д. А., Миронова В. В. СИСТЕМНАЯ БИОЛОГИЯ //

Вавиловский журнал генетики и селекции. 2014. Т. 18. № 1. С. 175–192.

132. Галактионов В. Г. Иммунология: учебник для студентов вузов, обучающихся по направлению 510600 «Биология» и биол. специальностям. , 2004. Вып. М.: Академ.

133. Колчанов Н. и др. Генные сети // Вавиловский журнал генетики и селекции. 2013. Т. 17. № 4/2. С. 833–850.

134. Кунин Е. В. Логика Случая. : Центрполиграф, 2014. С. 528.

135. Несговорова Г.П. Биоинформатика: пути развития и перспективы // Информатика в науке и образовании. 2012. С. 71–89.

136. Ратнер В.А. ГЕНЕТИЧЕСКИЕ УПРАВЛЯЮЩИЕ СИСТЕМЫ // 1966.

137. Степанов В. А. ЭВОЛЮЦИЯ ГЕНЕТИЧЕСКОГО РАЗНООБРАЗИЯ И БОЛЕЗНИ ЧЕЛОВЕКА // Генетика. 2016. Т. 52. № 7. С. 852–864.

138. Хаитов Р. М. Иммунология: учебник для студентов медицинских вузов. , 2016. Вып. М.: ГЭОТАР.

139. Хаубольд Б. Введение в вычислительную биологию: эволюционный подход. , 2011. Вып. М. Ижевск.