Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2024;28(8):874-881

doi 10.18699/vjgb-24-95

Orthoweb: программный комплекс для эволюционного анализа генных сетей

Р.А. Иванов [1], А.М. Мухин [1, 2], Ф.В. Казанцев [1, 2], З.С. Мустафин [2], Д.А. Афонников [1, 2], Ю.Г. Матушкин [1], С.А. Лашин [1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
ivanovromanart@bionet.nsc.ru

Аннотация. В данной статье описывается Orthoweb (https://orthoweb.sysbio.cytogen.ru/) – программный комплекс, разработанный для вычисления эволюционных индексов, включая филостратиграфические индексы и индексы дивергенции (K_a/K_s) как отдельных генов, так и генных сетей. Индекс филостратиграфического возраста (PAI) позволяет оценить эволюционную стадию появления гена (при этом косвенно оценив приблизительное время его возникновения – так называемый эволюционный возраст) на основе анализа ортологичных генов у близкородственных и дальнородственных таксонов. Кроме того, Orthoweb поддерживает расчет индексов возраста транскриптома (TAI) и дивергенции транскриптома (TDI). Эти индексы важны для понимания динамики экспрессии генов и ее последствий для развития и адаптации организмов. Orthoweb содержит также дополнительные аналитические функции, такие как возможность анализа терминов Gene Ontology (GO), что позволяет проводить функциональное обогащение и связывать эволюционное происхождение генов с биологическими процессами. Помимо этого, доступна возможность анализа обогащения по однонуклеотидным полиморфизмам (SNP), который помогает исследовать эволюционное значение генетических вариантов в конкретных геномных регионах. Одной из ключевых особенностей Orthoweb является интеграция перечисленных индексов с анализом генетических сетей. Программный пакет предлагает расширенные средства визуализации, такие как картирование генетических сетей и графическое представление распределения филостратиграфических индексов элементов сетей, что облегчает интуитивную интерпретацию сложных эволюционных связей. Для упрощения рабочих процессов в Orthoweb включена база данных с предварительно рассчитанными индексами для множества таксонов, доступная через API. Эта функция позволяет эффективно получать готовые данные по филостратиграфическим индексам и индексам дивергенции, значительно сокращая время вычислений.

Ключевые слова: генные сети; эволюция; филостратиграфия.

Для цитирования: Иванов Р.А., Мухин А.М., Казанцев Ф.В., Мустафин З.С., Афонников Д.А., Матушкин Ю.Г., Лашин С.А. Orthoweb: программный комплекс для эволюционного анализа генных сетей. *Вавиловский журнал генетики и селекции*. 2024;28(8):874-881. doi 10.18699/vjgb-24-95

Финансирование. Работа выполнена при поддержке бюджетного проекта № FWNR-2022-0020.

Orthoweb: a software package for evolutionary analysis of gene networks

R.A. Ivanov [1], A.M. Mukhin [1, 2], F.V. Kazantsev [1, 2], Z.S. Mustafin [2], D.A. Afonnikov [1, 2], Y.G. Matushkin [1], S.A. Lashin [1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
ivanovromanart@bionet.nsc.ru

Abstract. This article introduces Orthoweb (https://orthoweb.sysbio.cytogen.ru/), a software package developed for the calculation of evolutionary indices, including phylostratigraphic indices and divergence indices (K_a/K_s) for individual genes as well as for gene networks. The phylostratigraphic age index (PAI) allows the evolutionary stage of a gene's emergence (and thus indirectly the approximate time of its origin, known as "evolutionary age") to be assessed based on the analysis of orthologous genes across closely and distantly related taxa. Additionally, Orthoweb supports the calculation of the transcriptome age index (TAI) and the transcriptome divergence index (TDI). These indices are important for understanding the dynamics of gene expression and its impact on the development and adaptation of organisms. Orthoweb also includes optional analytical features, such as the ability to explore Gene Ontology (GO) terms associated with genes, facilitating functional enrichment analyses that link evolutionary origins of genes to biological processes. Furthermore, it offers tools for SNP enrichment analysis, enabling the users to assess the evolutionary significance of genetic variants within specific genomic regions. A key feature of Orthoweb is its ability to

integrate these indices with gene network analysis. The software offers advanced visualization tools, such as gene network mapping and graphical representations of phylostratigraphic index distributions of network elements, ensuring intuitive interpretation of complex evolutionary relationships. To further streamline research workflows, Orthoweb includes a database of pre-calculated indices for numerous taxa, accessible via an application programming interface (API). This feature allows the users to retrieve pre-computed phylostratigraphic and divergence data efficiently, significantly reducing computational time and effort.

Key words: gene networks; evolution; phylostratigraphy.

For citation: Ivanov R.A., Mukhin A.M., Kazantsev F.V., Mustafin Z.S., Afonnikov D.A., Matushkin Y.G., Lashin S.A. Orthoweb: a software package for evolutionary analysis of gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii* = *Vavilov Journal of Genetics and Breeding*. 2024;28(8):874-881. doi 10.18699/vjgb-24-95

Введение

Эволюционный анализ генных сетей позволяет исследовать происхождение и развитие биологических систем в контексте эволюции. Один из ключевых аспектов этого анализа – изучение индексов возраста генов, которые позволяют определить временные рамки появления и диверсификации генов в различных филогенетических линиях. Филостратиграфия, методология, основанная на оценке эволюционного возраста генов, предоставляет возможность для поиска древних и недавно возникших генов, а также для понимания их функциональной значимости в биологических процессах (Domazet-Lošo, Tautz, 2008; Tautz, Domazet-Lošo, 2011; Šestak et al., 2013; Xie et al., 2017). Целью филостратиграфического анализа является определение возраста возникновения гена-основателя на основе оценки распределения гомологичных ему генов в геномах организмов, принадлежащих к различным таксономическим группам. Для оценки времени происхождения генов в филостратиграфии используется индекс филостратиграфического возраста (Phylostratigraphic Age Index, PAI), соответствующий самой древней филострате, в которую входят гомологичные последовательности искомых генов.

Поиск генов с ограниченной в пределах таксонов гомологией особенно интересен с точки зрения эволюционной биологии, поскольку в ряде исследований было показано, что новые гены могут играть значительную роль в появлении новых эволюционных признаков и быть связаны с формированием новых морфологических признаков наземных растений (Bowles et al., 2020) и многоклеточных животных (Paps, Holland, 2018). Кроме того, продемонстрировано, что эволюционно новые гены связаны с каскадами развития органов, в частности с развитием мозговых тканей (An et al., 2023), и что гены, специфичные для таксона, перепредставлены в системах борьбы со стрессом и иммунной системе (Dornburg, Yoder, 2022). Некоторые исследователи предполагают, что гены, специфичные для таксона, связаны с экологической специализацией в различных таксонах (Baalsrud et al., 2018).

Однако классический подход филостратиграфии сталкивается с рядом ограничений, связанных с увеличением объема получаемых геномных данных и недостаточной точностью алгоритма BLASTP в определении гомологов. Эти факторы, а также высокая вычислительная сложность приводят к тому, что филостратиграфический анализ полногеномных данных с использованием BLASTP может занимать до нескольких недель (Buchfink et al., 2021). В связи с этим возникла потребность в разработке новых программных решений для филостратиграфического анализа.

Современные программные инструменты, такие как fagin (Arendsee et al., 2019), GenEra (Barrera-Redondo et al., 2023) и oggmap (Ullrich, Glytnasi, 2023), предлагают альтернативные подходы к филостратиграфическому анализу, позволяя преодолеть некоторые из ограничений классических методов. Программа fagin, написанная на языке R, использует подход к поиску гомологов, основанный на определении синтенных геномных интервалов в целевом геноме и дальнейшем поиске гомологии как в аминокислотных, так и в нуклеотидных последовательностях. Разработчики программного комплекса GenEra внесли ряд изменений в классический метод определения гомологов в филостратиграфии, заменив традиционный метод поиска BLASTP на алгоритм DIAMOND v2, что способствует лучшему определению отдаленных гомологий из-за снятия ограничений по числу лучших совпадений последовательностей при выравнивании. Также была добавлена возможность проверки ошибки поиска гомологии и возможность оценить таксономическую репрезентативность – показатель, который учитывает наличие гомологов гена хотя бы у одного репрезентативного вида для каждого промежуточного таксономического уровня между наиболее отдаленно родственным родом и искомым видом. Программа oggmap (Ullrich, Glytnasi, 2023), реализованная в виде пакета на языке программирования Python, предназначена для получения карт ортологий (ортокарт) или, другими словами, значений индекса филостратиграфического возраста заданных групп ортологов из результатов, полученных с помощью таких инструментов, как OrthoFinder (Emms, Kelly, 2019) и eggNOG (Huerta-Cepas et al., 2019). В отличие от классического метода филостратиграфии, данный подход не подразумевает шага поиска ортологов при помощи инструментов выравнивания. Вместо этого предполагается использование готовых результатов поиска ортологов в виде ортокарт и дальнейшее их применение в расчете возраста генов. Такие ортокарты содержат информацию о возрасте генов в каждой группе ортологов.

Для полноценного эволюционного анализа эти инструменты и подходы требуют знаний языков программирования. Кроме того, большая часть программных средств использует алгоритмы выравнивания BLAST, продолжительность работы которых в ряде случаев существенно замедляет анализ. Наконец, существующие на сегодняшний день программные реализации метода расчета филостратиграфических индексов не способны на полноценный и быстрый эволюционный анализ компонентов генных сетей. В этой статье мы представляем Orthoweb – программный комплекс для эволюционного

анализа генных сетей и отдельных генов, реализованный в веб-приложении по адресу https://orthoweb.sysbio.cytogen.ru.

Материалы и методы

Orthoweb разработан на языке программирования Java с использованием фреймворка Spring для реализации серверной части приложения и фреймворков Vue.JS и webix для реализации клиентской части. Для визуализации сетей используется группа библиотек cytoscape.js. В качестве СУБД для хранения данных из базы KEGG (таксоны, список ортологов, кодирующие последовательности и т. д.) и промежуточных результатов анализа применяется MongoDB, что значительно увеличивает скорость последующей работы с этими данными.

Рассчитанные индексы хранятся в базе данных на основе СУБД PostgreSQL. Доступ к этим данным реализован по технологии REST API с применением библиотеки FLASK (flask.palletsprojects.com). Благодаря этому программному интерфейсу возможно запросить данные при помощи разных инженерных сред моделирования (Matlab, Octave, Statistica) или стандартными библиотеками скриптовых языков программирования (R, Python).

Результаты

Функциональность Orthoweb

Расчет эволюционных индексов возраста единичных генов. Основная функция Orthoweb – определение индексов филостратиграфического возраста (PAI) генов.

В Orthoweb реализованы два метода определения PAI: 1) на основе анализа метрик идентичности гомологических последовательной; 2) с использованием классификации белков на ортологические группы БД KEGG (KEGG Orthology, KO). На основе информации KO из БД KEGG (Kanehisa et al., 2016) Orthoweb позволяет для каждой последовательности белка найти соответствующие ей ортологи и определить виды, в геномах которых они были найдены. Таксономические линии выявленных видов последовательно сравниваются с линией исследуемого вида, чтобы определить их эволюционную родословную и установить последнего общего предка для рассматриваемого гена, чей номер от корня таксономического дерева рассчитывается как PAI (рис. 1). Таксономические линии ортологов уже отобраны в базе данных KEGG и практически не требуют дополнительных настроек от пользователя. Рассчитанные индексы PAI сохраняются в регулярно обновляемую базу данных, о которой более подробно пойдет речь ниже. Поскольку данные по ортогруппам KEGG обновляются часто, то в Orthoweb реализована также возможность рассчитывать индексы PAI по группам ортологов напрямую из базы данных KEGG для доступа к самой актуальной информации. Однако не для всех генов такая информация доступна. В частности, для человека только примерно 2/3 генов, представленных в KEGG, ассоциированы с KO-группой.

Вторым способом расчета PAI является использование таблицы сходства последовательностей (Best Similarity Table), доступной для большинства генов, представленных в KEGG (Kanehisa et al., 2016). При помощи этого способа пользователь может отобрать гомологичные гены на основе таких параметров, как сходство аминокислотных последовательностей кодируемых генами белков и результат работы алгоритма Смита–Ватермана по локальному выравниванию последовательностей.

Расчет индексов дивергенции. Orthoweb поддерживает также опцию расчета соотношения несинонимичных и синонимичных замен (соотношение d_N/d_S) между

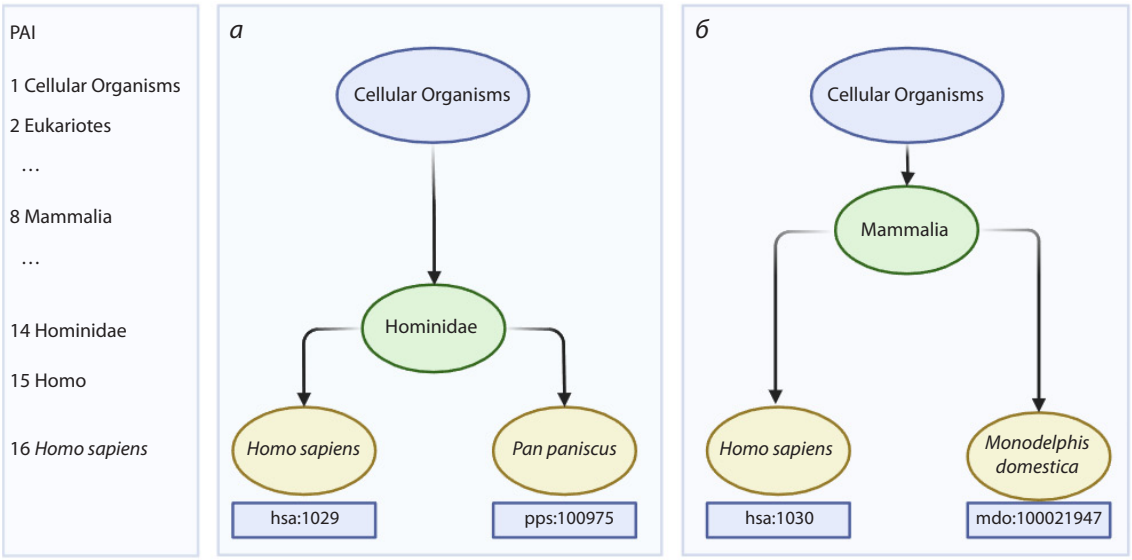

Рис. 1. Пример определения PAI для двух генов *Homo sapiens*.

а – эволюционно молодой ген hsa:1029 (CDKN2A); наиболее отдаленный от исследуемого организм, у которого был найден ортолог этого гена, – *Pan paniscus* (шимпанзе бонобо); *б* – эволюционно более древний ген hsa:1030 (CDKN2B); наиболее отдаленный от исследуемого организм, у которого был найден ортолог этого гена, – *Monodelphis domestica* (домовый опоссум). Можно заключить, что ген на примере (*а*) эволюционно моложе гена на примере (*б*). Шкала слева показывает индекс PAI, который соответствует глубине узла таксономического дерева, адаптировано из (Mustafin et al., 2021).

последовательностью изучаемого гена и каждым из его гомологов у близкородственных видов, отраженного в индексе дивергенции (divergence index, DI). Этот индекс позволяет определить тип отбора, которому подвержен ген. Индекс вычисляется на основании отношения d_N/d_S (в литературе также известен как K_a/K_s), где d_N – доля несинонимичных замен в последовательностях исследуемого гена и его гомологов, т. е. таких замен, которые приводят к смене кодируемой данным триплетом аминокислоты; d_S – доля синонимичных замен, не приводящих к замене кодируемой аминокислоты. Считается, что значение индекса меньше 1 указывает на то, что ген подвержен стабилизирующему отбору, значения близкие к 1 – нейтральной эволюции, а значения больше 1 – движущему отбору (Yang, Nielsen, 2000).

В случае сравнения с одной гомологичной последовательностью DI равен d_N/d_S. При наличии нескольких гомологов DI равен среднему значению d_N/d_S для всех сравнений. При подсчете индекса DI пользователи Orthoweb могут выбрать таксономическую глубину анализа, чтобы учитывать эволюционную изменчивость гена среди более или менее эволюционно отдаленных организмов. Для подсчета отношения d_N/d_S используется программа PAML (Yang, 2007).

Расчет насыщенности гена однонуклеотидными полиморфизмами и анализ терминов генной онтологии. Orthoweb также интегрирует информацию об ассоциированных с генами терминах генной онтологии и насыщенности исследуемых генов однонуклеотидными полиморфизмами (SNP). Для получения информации по терминам генной онтологии используется ресурс http://geneontology.org/ (Ashburner et al., 2000; Carbon et al., 2021). Загрузка происходит с помощью предоставляемого API (application programming interface – интерфейс программного доступа к ресурсам). Например, запрос для гена TBP формируется как http://api.geneontology.org/api/bioentity/gene/NCBIGene:6908/function, в нем указана база данных и идентификатор гена в ней. Orthoweb предоставляет эту информацию самостоятельно, для большинства модельных организмов используются связанные с ними базы данных (например, TAIR для *Arabidopsis thaliana*, FlyBase для *Drosophila melanogaster* и т. д.), а для остальных организмов – база UniProt. Если в Gene Ontology есть данные по исследуемому гену и в KEGG имеется необходимый идентификатор, что выполняется практически для всех хорошо изученных генов, то будут загружены идентификаторы ассоциированных с геном терминов GO и их наименования.

Для получения данных о насыщенности искомых генов однонуклеотидными полиморфизмами реализована автоматическая система запросов к базе данных NCBI SNP (Sayers et al., 2022). Запрос конструируется на основании идентификатора гена и, например, для гена TBP с идентификатором hsa:6908 имеет следующий вид: https://www.ncbi.nlm.nih.gov/snp/?term=6908. В результате запроса пользователю предоставляется количество найденных SNP. Следует отметить, что в текущей версии Orthoweb поиск SNP осуществляется только для генов человека.

Расчет эволюционных индексов группы генов. Еще одной функцией Orthoweb является ввод данных об экспрессии генов для расчета филотранскриптомных индексов. Анализ филотранскриптомных индексов – это подход, объединяющий информацию об эволюционном возрасте генов и данные об уровне их экспрессии. Этот анализ позволяет исследовать взаимосвязь между индексом PAI генов и изменениями в их активности в контексте различных физиологических состояний, адаптивных ответов или этапов развития организмов. Используя филотранскриптомный анализ, можно выявить, как эволюционные характеристики генома связаны с транскрипционной регуляцией и функциональной динамикой генов в разнообразных биологических контекстах. К филотранскриптомным индексам относятся два эволюционных индекса: Transcriptome Age Index (Domazet-Lošo, Tautz, 2010) и Transcriptome Divergence Index (Quint et al., 2012).

Transcriptome Age Index (TAI) – это индекс возраста транскриптома, который отражает взвешенный средний возраст транскриптома в выбранном биологическом процессе. Данные об экспрессии выступают в качестве дополнительного множителя и используются для нормализации результата, так что чем выше итоговое значение TAI/TDI, тем больше вклад эволюционно молодых/изменчивых генов. Формулы для расчета индексов выглядят следующим образом:

$$\mathrm{TAI} = \frac{\sum_{i=1}^{n} ps_i e_i}{\sum_{i=1}^{n} e_i},$$

где ps_i – целое число, отражающее PAI для гена с индексом i; e_i – значение уровня экспрессии, полученное с транскриптомных данных по анализу экспрессии гена i; n – общее число генов.

Transcriptome Divergence Index (TDI) – индекс дивергенции транскриптома, количественно отражает консервативность транскриптома в искомом процессе и помогает выявить биологические процессы или стадии развития организма, в которых наблюдается повышенная экспрессия более консервативных или более молодых генов:

$$\mathrm{TDI} = \frac{\sum_{i=1}^{n} \mathrm{DI}_i e_i}{\sum_{i=1}^{n} e_i}.$$

Здесь DI_i – индекс дивергенции для гена i; e_i – уровень экспрессии гена i; n – количество генов.

Примеры использования Orthoweb

Для иллюстрации работы Orthoweb мы опишем порядок действий и примеры его применения в филостратиграфическом анализе.

Анализ характеристик единичных генов. В анализе эволюционных индексов отдельных генов Orthoweb принимает несколько форматов входных файлов: список генов из формы на сайте, список генов из файла или файл со взаимодействиями между элементами генной сети в форматах .txt или .tsv. Пользователи могут выбрать желаемый формат входных данных в соответствующей форме – *Choose the type of input data* (рис. 2). Для корректного анализа в Orthoweb требуется подавать идентификаторы генов KEGG.

Следующим шагом является выбор режима анализа в форме *The type of orthology*. В этой форме можно выбрать одну из опций: расчет филостратиграфических индексов с

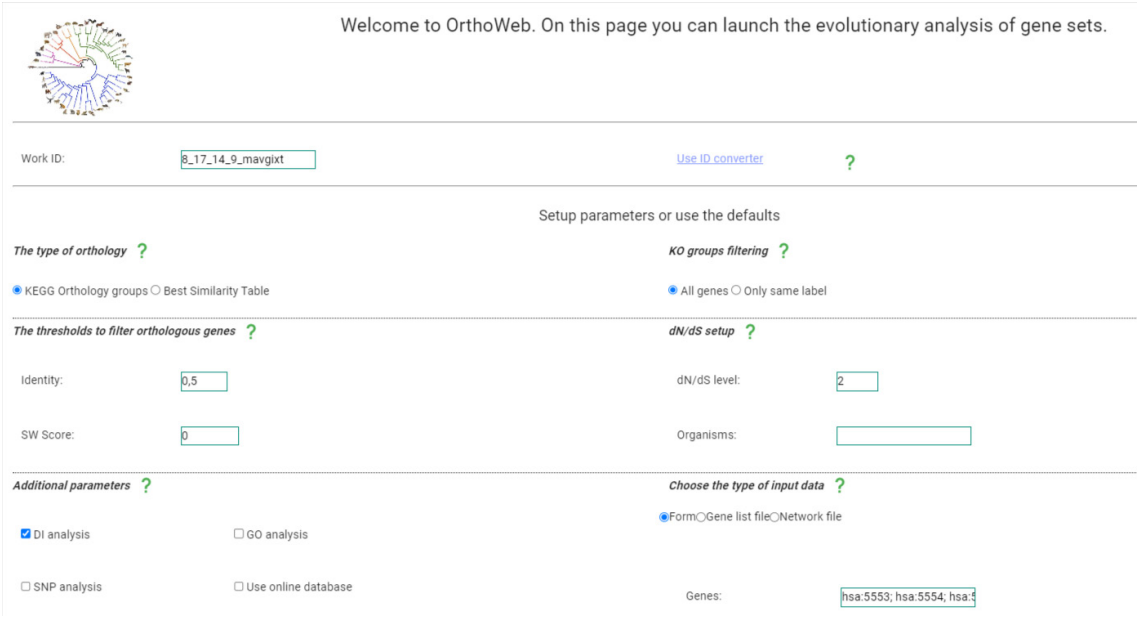

Рис. 2. Стартовая страница веб-сервиса Orthoweb.

использованием анализа семейств ортологов и KO-групп (опция *KEGG Orthology groups*) или с применением анализа гомологичных последовательностей (опция *Best Similarity Table*).

При выборе режима *KEGG Orthology groups* необходимо также принять решение относительно того, включать ли паралогичные гены в анализ в окне *KO groups filtering*.

При выборе режима расчета филостратиграфических индексов генов с применением анализа гомологов необходимо ввести в Orthoweb пороговые значения параметров идентичности аминокислотных последовательностей (по умолчанию установлено значение 0.5) и показателя алгоритма Смита–Ватермана для отбора гомологичных генов в опции *The thresholds to filter orthologous genes*.

В окне *Additional parameters* можно задать ряд дополнительных параметров для анализа: расчет индекса дивергенции DI в опции *DI analysis*, оценку обогащения однонуклеотидными полиморфизмами и идентификацию терминов Gene Ontology. Для расчета индексов дивергенции можно дополнительно настроить группы организмов, относительно которых вы хотите рассчитать значение индекса, в окне d_N/d_S *setup*. В этой опции есть два варианта настройки. Первый параметр, d_N/d_S *level*, определяет таксономический уровень, на котором проводится анализ d_N/d_S. В основном этот вид анализа используется для сравнения последовательностей близкородственных организмов. Значение параметра, равное 1, ограничивает анализ организмами в пределах одного рода. Например, при анализе генов человека значение, равное 2, означает, что d_N/d_S будет рассчитываться относительно других организмов из семейства Hominidae. Второе поле, *Organisms*, позволяет вводить коды конкретных видов из базы данных KEGG. В частности, для сравнения последовательности изучаемого гена человека не со всеми гоминидами, а только с гориллой, в поле нужно ввести код "*ggo*".

Результатом работы Orthoweb в этих режимах анализа является файл архива, в котором будет табличный текстовый файл со столбцами данных: Gene – с идентификаторами генов KEGG ID; Label – с идентификаторами Entrez ID; PAI – со значениями индексов филостратиграфического возраста; и столбцы со значениями из дополнительных режимов анализа: DI, SNP и GO label.

Анализ характеристик групп генов. Для расчета индексов возраста (TAI) и дивергенции транскриптомов (TDI) необходимо выбрать опцию формата входных данных *Network file – Use expression*. В этом режиме пользователь должен предоставить табличный текстовый файл, содержащий колонку с названиями генов и несколько колонок с нормализованными значениями экспрессии, названными по условию, в котором проводился анализ экспрессии. Подаваемый файл может быть как файлом генной сети, так и файлом только со списком генов.

На выходе программа Orthoweb выдаст табличный текстовый файл с тремя столбцами: Data – с наименованиями условий, заданных во входном файле; TAI – со значениями индекса возраста транскриптома выбранного набора генов; TDI – со значениями индекса дивергенции транскриптома в заданных условиях.

Анализ генных сетей. Помимо анализа индексов отдельных генов и списков генов, в Orthoweb реализованы филостратиграфический анализ и визуализация генных сетей. Пользователи могут анализировать как сети, импортированные из баз данных KEGG Pathway (Kanehisa et al., 2017) и WikiPathways, так и сети, загруженные из текстовых файлов. Доступ к анализу сетей из баз данных предоставляется по ссылке: https://orthoweb.sysbio.cytogen.ru/pathway.html.

Orthoweb поддерживает импорт и анализ сетей из двух основных баз данных. Первая база данных – KEGG Pathway, содержит множество генных сетей и метаболических путей, классифицированных по различным критериям,

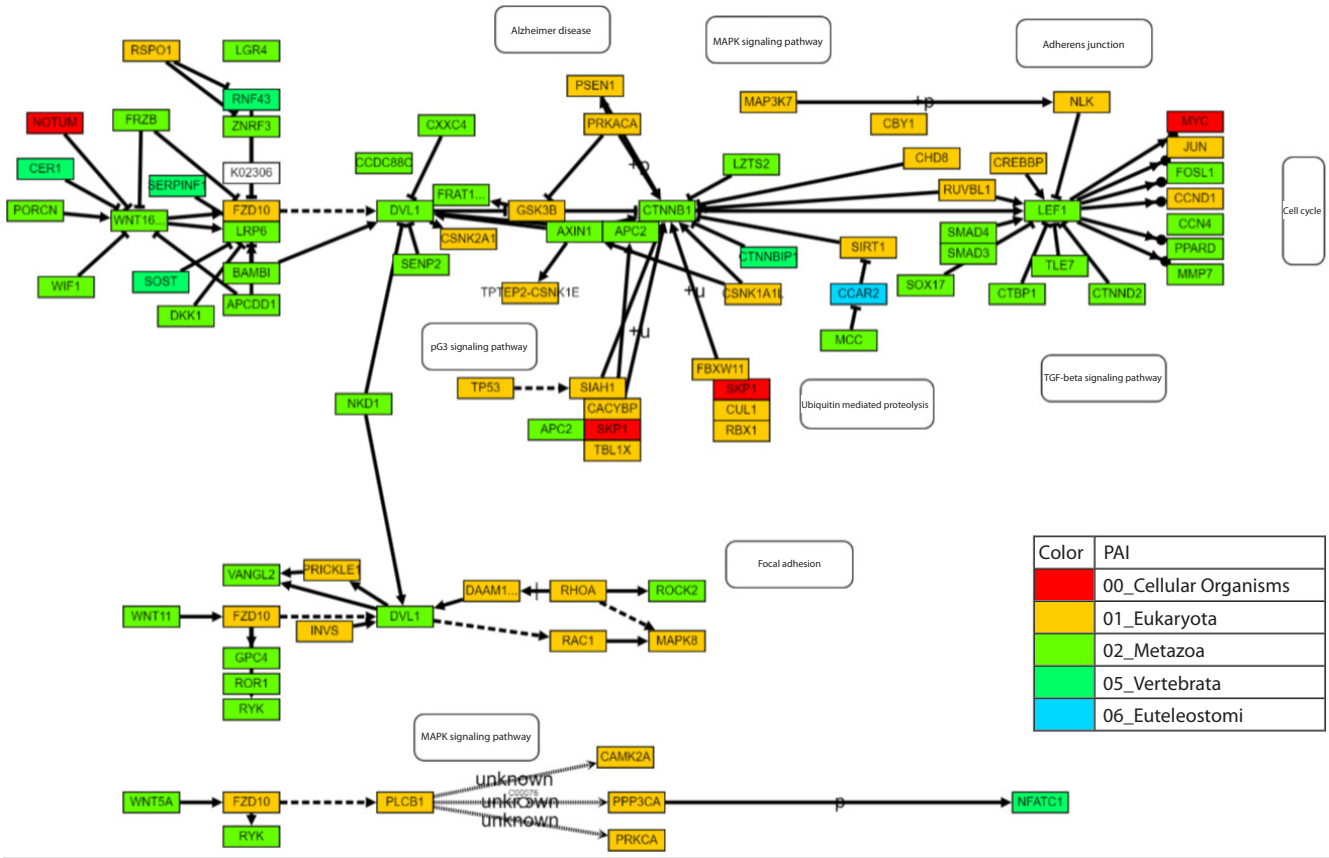

Рис. 3. Пример визуализации сети из базы данных KEGG pathway "Wnt signalling pathway" (https://www.kegg.jp/pathway/hsa 04310), проанализированной с помощью Orthoweb.

Цвет узла соответствует индексу PAI гена (белые элементы – пути и химические соединения). По умолчанию импортируется стандартная структура сети с сохранением координат элементов, но масштаб сети может быть изменен пользователем, и с каждым элементом можно взаимодействовать.

таким как метаболизм, функционирование разных систем организма и заболевания человека. Для запуска анализа пользователю необходимо указать код метаболического пути и организм, для которого будет импортирована сеть. В результате анализа сети из KEGG Pathway пользователь Orthoweb получит генную сеть, на узлах которой цветом будут отражены значения PAI, определенные на основании представленных в сети KO групп. Поскольку все элементы, представленные в сетях KEGG, описаны в самой базе KEGG, то импорт и анализ такой сети максимально удобны для Orthoweb, который большую часть информации для анализа загружает из KEGG.

В качестве примера работы этого режима Orthoweb мы проанализировали сеть Wnt/β-catenin сигнального каскада (рис. 3). WNT/β-catenin путь участвует в контроле клеточного цикла, адгезии, миграции и дифференцировке клеток. Активация пути начинается со связывания лигандов WNT с рецепторами Frizzled и LRP на поверхности клетки, что приводит к стабилизации и накоплению β-catenin в цитоплазме и его последующему перемещению в ядро, где он взаимодействует с факторами транскрипции и стимулирует экспрессию целевых генов (Davidson et al., 2009). Дисрегуляция данного пути связана с развитием множества типов рака (Zhan et al., 2017). Этот сигнальный каскад является одним из самых древних и включает в себя преимущественно гены, возникшие на этапе происхождения многоклеточных организмов и эукариот (PAI = 1, 2).

Вторая база данных, для которой реализован импорт сетей, – это WikiPathways. Сети, представленные в WikiPathways, содержат больше деталей, сущностей и вариантов взаимодействий, чем в KEGG, что усложняет их полноценный импорт и требует учета идентификаторов из множества различных баз данных.

Orthoweb предлагает пошаговый процесс импорта и анализа генных сетей, подготовленных пользователем. Пользователь может сначала импортировать сеть в формате TSV (текстовый табличный файл, где значения разделены табуляцией), а затем взаимодействовать с ней, например, перемещать элементы перед запуском анализа. Этот формат применяется в широко используемом инструменте STRING (von Mering et al., 2005), что обеспечивает совместимость и простоту интеграции данных из STRING в Orthoweb без дополнительной обработки. В случае сети, импортированной из STRING, колонка combined_score содержит достоверность найденного взаимодействия с весом от 0 до 1. По завершении анализа цвет генов изменяется в соответствии с показателем PAI (рис. 4). При включении опций дополнительных режимов анализа, описанных ранее в тексте, те также будут отражены в визуализации.

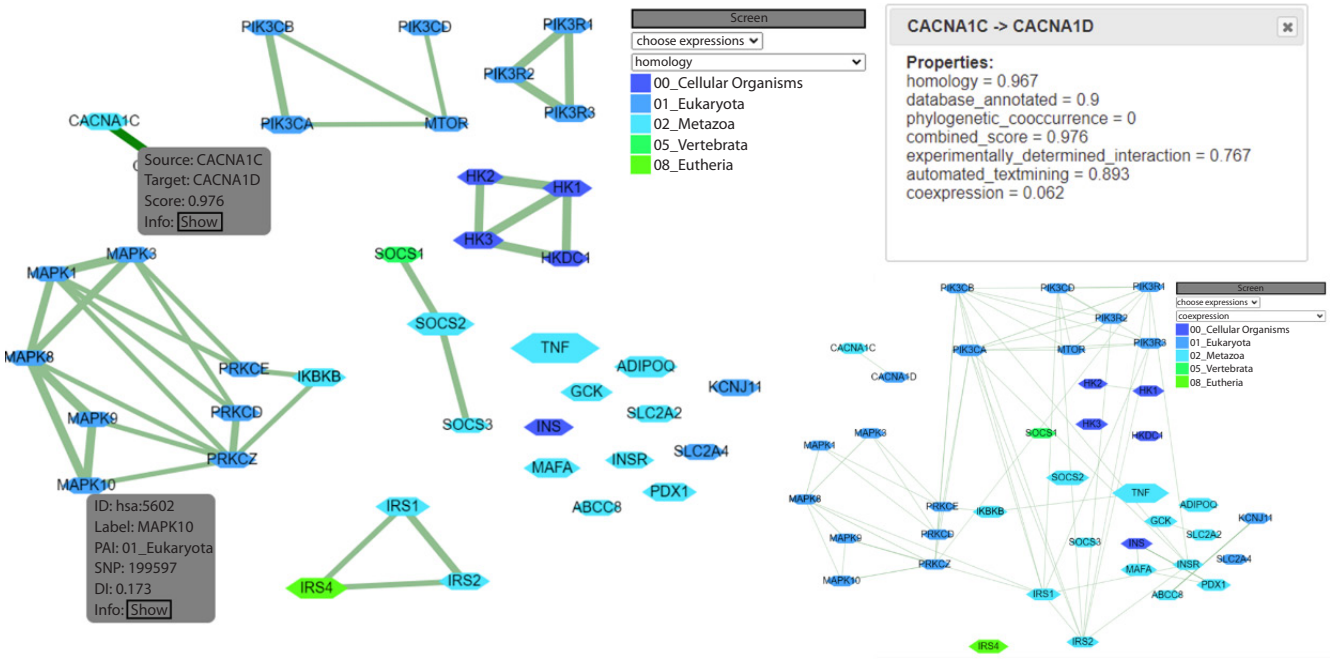

Рис. 4. Пример сети, импортированной из инструмента STRING, в котором цвет узла соответствует PAI индексу, а толщина дуги между ребрами – значению combined_score STRING. При выборке конкретной связи в сети предоставляется информация по уровням достоверностям этой связи в STRING.

База данных для хранения результатов

Для ускорения расчетов индексов и исключения многократных повторных вычислений в Orthoweb разработана база данных, содержащая таблицы: Организмы; Гены; Индексы PAI; Индексы DI; Термины генной онтологии (идентификаторы и наименования); SNP; Индексы PAI, определенные на основе KO-групп. Помимо использования этой базы при работе в интерактивном режиме, возможно ее использование через API (программный инструмент доступа) в рамках инженерных сред моделирования или популярных скриптовых языков программирования (Matlab, Octave, R, Python…). Таким образом, предоставляется доступ к имеющейся информации по всем посчитанным PAI и DI индексам для генов конкретных организмов, на основе которых можно выстроить сценарии отбора и визуализации данных. Благодаря API обращение к данным базы осуществляется через обычные URL-адреса специальной структуры. Результатом запроса является текстовый файл в структурированном формате JSON. С описанием ключей для API запросов и примером запроса к базе данных можно ознакомиться в электронном Приложении[1].

Заключение

В работе представлен программный комплекс Orthoweb, предназначенный для анализа филостратиграфических индексов и индексов дивергенции как отдельных генов, так и генных сетей. Orthoweb позволяет также интегрировать значения эволюционных индексов с данными по уровням экспрессии генов в различных условиях.

Одной из ключевых особенностей Orthoweb является его расширенная функциональность по визуализации данных. Инструменты для отображения эволюционных индексов на генных сетях значительно упрощают интерпретацию сложных эволюционных взаимосвязей, делая результаты анализа доступными для широкого круга исследователей.

[1] Приложение см. по адресу:
https://vavilovj-icg.ru/download/pict-2024-28/appx30.pdf

Список литературы / References

An N.A., Zhang J., Mo F., Luan X., Tian L., Shen Q.S., Li X., Li C., Zhou F., Zhang B., Ji M., Qi J., Zhou W.-Z., Ding W., Chen J.-Y., Yu J., Zhang L., Shu S., Hu B., Li C.-Y. De novo genes with an lncRNA origin encode unique human brain developmental functionality. *Nat. Ecol. Evol.* 2023;7(2):264-278. doi 10.1038/s41559-022-01925-6

Arendsee Z., Li J., Singh U., Bhandary P., Seetharam A., Wurtele E.S. fagin: synteny-based phylostratigraphy and finer classification of young genes. *BMC Bioinformatics*. 2019;20(1):440. doi 10.1186/s12859-019-3023-y

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., Sherlock G. Gene Ontology: tool for the unification of biology. *Nat. Genet.* 2000;25(1):25-29. doi 10.1038/75556

Baalsrud H.T., Tørresen O.K., Solbakken M.H., Salzburger W., Hanel R., Jakobsen K.S., Jentoft S. *De novo* gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Mol. Biol. Evol.* 2018;35(3):593-606. doi 10.1093/molbev/msx311

Barrera-Redondo J., Lotharukpong J.S., Drost H.-G., Coelho S.M. Uncovering gene-family founder events during major evolutionary transitions in animals, plants and fungi using GenEra. *Genome Biol.* 2023;24(1):54. doi 10.1186/s13059-023-02895-z

Bowles A.M.C., Bechtold U., Paps J. The origin of land plants is rooted in two bursts of genomic novelty. *Curr. Biol.* 2020;30(3):530-536.e2. doi 10.1016/j.cub.2019.11.090

Buchfink B., Reuter K., Drost H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods*. 2021;18(4):366-368. doi 10.1038/s41592-021-01101-x

Carbon S., Douglass E., Good B.M., Unni D.R., Harris N.L., Mungall C.J., Basu S., Chisholm R.L., Dodson R.J., Hartline E., … Stein L., Howe D.G., Toro S., Westerfield M., Jaiswal P., Cooper L., Elser J. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* 2021;49(D1):D325-D334. doi 10.1093/nar/gkaa1113

Davidson G., Shen J., Huang Y.-L., Su Y., Karaulanov E., Bartscherer K., Hassler C., Stannek P., Boutros M., Niehrs C. Cell cycle control of Wnt receptor activation. *Dev. Cell*. 2009;17(6):788-799. doi 10.1016/j.devcel.2009.11.006

Domazet-Lošo T., Tautz D. An ancient evolutionary origin of genes associated with human genetic diseases. *Mol. Biol. Evol.* 2008;25(12):2699-2707. doi 10.1093/molbev/msn214

Domazet-Lošo T., Tautz D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*. 2010;468(7325):815-819. doi 10.1038/nature09632

Dornburg A., Yoder J.A. On the relationship between extant innate immune receptors and the evolutionary origins of jawed vertebrate adaptive immunity. *Immunogenetics*. 2022;74(1):111-128. doi 10.1007/s00251-021-01232-7

Emms D.M., Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019;20(1):238. doi 10.1186/s13059-019-1832-y

Huerta-Cepas J., Szklarczyk D., Heller D., Hernández-Plaza A., Forslund S.K., Cook H., Mende D.R., Letunic I., Rattei T., Jensen L.J., von Mering C., Bork P. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 2019;47(D1):D309-D314. doi 10.1093/nar/gky1085

Kanehisa M., Sato Y., Kawashima M., Furumichi M., Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 2016;44(D1):D457-D462. doi 10.1093/nar/gkv1070

Kanehisa M., Furumichi M., Tanabe M., Sato Y., Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017;45(D1):D353-D361. doi 10.1093/nar/gkw1092

Mustafin Z.S., Lashin S.A., Matushkin Y.G. Phylostratigraphic analysis of gene networks of human diseases. *Vavilov J. Genet. Breed.* 2021;25(1):46-56. doi 10.18699/VJ21.006

Paps J., Holland P.W.H. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. *Nat. Commun.* 2018;9(1):1730. doi 10.1038/s41467-018-04136-5

Quint M., Drost H.G., Gabel A., Ullrich K.K., Bönn M., Grosse I. A transcriptomic hourglass in plant embryogenesis. *Nature*. 2012;490(7418):98-101. doi 10.1038/nature11394

Sayers E.W., Bolton E.E., Brister J.R., Canese K., Chan J., Comeau D.C., Connor R., Funk K., Kelly C., Kim S., Madej T., Marchler-Bauer A., Lanczycki C., Lathrop S., Lu Z., Thibaud-Nissen F., Murphy T., Phan L., Skripchenko Y., Tse T., Wang J., Williams R., Trawick B.W., Pruitt K.D., Sherry S.T. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 2022;50(D1):D20-D26. doi 10.1093/nar/gkab1112

Šestak M.S., Božičević V., Bakarić R., Dunjko V., Domazet-Lošo T. Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems. *Front. Zool.* 2013;10(1):18. doi 10.1186/1742-9994-10-18

Tautz D., Domazet-Lošo T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 2011;12(10):692-702. doi 10.1038/nrg3053

Ullrich K.K., Glytnasi N.E. oggmap: a Python package to extract gene ages per orthogroup and link them with single-cell RNA data. *Bioinformatics*. 2023;39(11):btad657. doi 10.1093/bioinformatics/btad657

von Mering C., Jensen L.J., Snel B., Hooper S.D., Krupp M., Foglierini M., Jouffre N., Huynen M.A., Bork P. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* 2005;33(D1):D433-D437. doi 10.1093/nar/gki005

Xie L., Draizen E.J., Bourne P.E. Harnessing big data for systems pharmacology. *Annu. Rev. Pharmacol. Toxicol.* 2017;57(1):245-262. doi 10.1146/annurev-pharmtox-010716-104659

Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 2007;24(8):1586-1591. doi 10.1093/molbev/msm088

Yang Z., Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 2000;17(1):32-43. doi 10.1093/oxfordjournals.molbev.a026236

Zhan T., Rindtorff N., Boutros M. Wnt signaling in cancer. *Oncogene*. 2017;36(11):1461-1473. doi 10.1038/onc.2016.304

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 08.11.2024. После доработки 21.11.2024. Принята к публикации 22.11.2024.