

Интернет-доступные информационные ресурсы по геномным сетям, включающие данные по человеку и животным

Е.В. Игнатиева^{1, 2}✉, Д.А. Афонников^{1, 2}, Н.А. Колчанов^{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

Геномные сети – это молекулярно-генетические системы, обеспечивающие формирование фенотипических характеристик организмов (молекулярных, биохимических, структурных, морфологических, поведенческих и т.д.) на основе информации, закодированной в их геномах. Реконструкция геномных сетей обеспечивает методическую основу современной системной биологии. Большую ценность представляет информация о структурно-функциональной организации геномных сетей, накопленная в современных базах данных. В настоящем обзоре представлена характеристика интернет-доступных информационных ресурсов, ориентированных на человека и животных и содержащих данные по геномным сетям и их функциональным модулям. Не претендуя на полноту охвата абсолютно всех информационных ресурсов, содержащих данные, относящиеся к человеку и животным по этой тематике, предложен обзор создан для того, чтобы оценить современное состояние проблемы, а также представить критерии, согласно которым целесообразно оценивать полезность информационных ресурсов для конкретных исследовательских задач. Исходя из этого нами, во-первых, была сформирована и охарактеризована подборка баз данных, содержащих сведения о метаболических и сигнальных путях, а также о путях регуляции биологических процессов на клеточном и организменном уровнях. Во-вторых, в качестве примера описаны несколько известных баз данных по межмолекулярным взаимодействиям различных типов. В обзоре рассматриваются следующие характеристики баз данных: 1) типы накопленной информации; 2) способы представления информации; 3) способы наполнения баз данных; 4) основные источники информации; 5) программные средства, позволяющие осуществлять поиск и анализ данных. Сопоставление перечисленных характеристик показало, что рассмотренные базы данных очень гетерогенны по тематике, источникам, типам и способу представления информации, а также по возможностям формировать запросы и анализировать данные. Делается вывод о том, что до начала реконструкции геномной сети определенного биологического процесса очень важно иметь представление о максимально полном наборе информационных источников, из которых может быть взята информация. Приведены примеры веб-порталов, аккумулирующих сведения о базах данных и информационных ресурсах, которые могут быть полезны для реконструкции и анализа геномных сетей.

Ключевые слова: системная биология; базы данных; геномные сети; Интернет.

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ:

Игнатиева Е.В., Афонников Д.А., Колчанов Н.А. Интернет-доступные информационные ресурсы по геномным сетям, включающие данные по человеку и животным. Вавиловский журнал генетики и селекции. 2017;21(8):895-902. DOI 10.18699/VJ17.310

HOW TO CITE THIS ARTICLE:

Ignatieva E.V., Afonnikov D.A., Kolchanov N.A. Online resources on gene networks containing human and animal data. Vavilovskii Zhurnal Genetiki i Selekcii = Vavilov Journal of Genetics and Breeding. 2017;21(8):895-902. DOI 10.18699/VJ17.310 (in Russian)

УДК 573.22+004.9

Поступила в редакцию 15.09.2017

Принята к публикации 20.11.2017

© АВТОРЫ, 2017

✉ e-mail: eignat@bionet.nsc.ru

Online resources on gene networks containing human and animal data

E.V. Ignatieva^{1, 2}✉, D.A. Afonnikov^{1, 2},
N.A. Kolchanov^{1, 2}

¹ Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

² Novosibirsk State University, Novosibirsk, Russia

Gene networks are molecular genetic systems that ensure the formation of phenotypic characteristics of organisms (molecular, biochemical, structural, morphological, behavioral, etc.) based on information encoded in their genomes. Reconstruction of gene networks provides a methodological basis for modern systems biology. In this regard, the information on the structural and functional organization of gene networks accumulated in modern databases is extremely valuable. This review characterizes a number of Internet-accessible information resources oriented to humans and animals and containing data on gene networks and their functional modules. Without pretending to fully cover all information resources containing data related to humans and animals on the subject, the current review was created to report the current status of the problem and to present the criteria according to which we propose to evaluate the utility of web-resources for specific research tasks. On this basis, we compiled and characterized a collection of databases containing information on metabolic and signaling pathways, as well as pathways of regulation of biological processes at the cellular and organismal levels. In addition, we observed the characteristics of several well-known databases containing data on interactions between biomolecules of various types. The following characteristics of databases were considered: (1) the types of information accumulated in the databases; (2) methods of data presentation; (3) methods of data collection; (4) data sources; (5) special search tools and options for data analysis. A comparison of the above characteristics showed that the databases are very heterogeneous according to their scopes, sources and types of data, interfaces, as well as according to their search options and data analysis tools. It was concluded that at the first step of the gene network reconstruction it is important to form a full set of information resources from which the data can be obtained. The web portals accumulating information about the databases that may be useful for the reconstruction and analysis of gene networks are specified.

Key words: systems biology; databases; gene networks; Internet.

При изучении генетических механизмов регуляции молекулярно-биологических процессов в современной биологической науке широко используется понятие «генные сети» (в англоязычной литературе – gene regulatory networks). Современное определение понятия «генные сети» базируется на представлениях о молекулярно-генетических системах управления (МГСУ), сформулированных профессором В.А. Ратнером. Термин МГСУ применялся для информационно-кибернетического описания и моделирования совокупности универсальных (репликация, транскрипция, трансляция, репарация, рекомбинация, сегрегация) и неуниверсальных молекулярно-биологических процессов и реакций, протекающих в клетках живых организмов (Ратнер, 1966). Позднее С. Кауфманом был предложен сходный термин «генные сети» (первоначально звучавший как genetic control networks), который и закрепился в мировой литературе (Kauffman, 1969).

В дальнейшем под генными сетями мы будем подразумевать молекулярно-генетические системы, обеспечивающие формирование фенотипических характеристик организмов (молекулярных, биохимических, структурных, морфологических, поведенческих и т. д.) на основе информации, закодированной в их геномах. Генные сети включают группы координированно функционирующих генов, которые взаимодействуют 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., 2000, 2013).

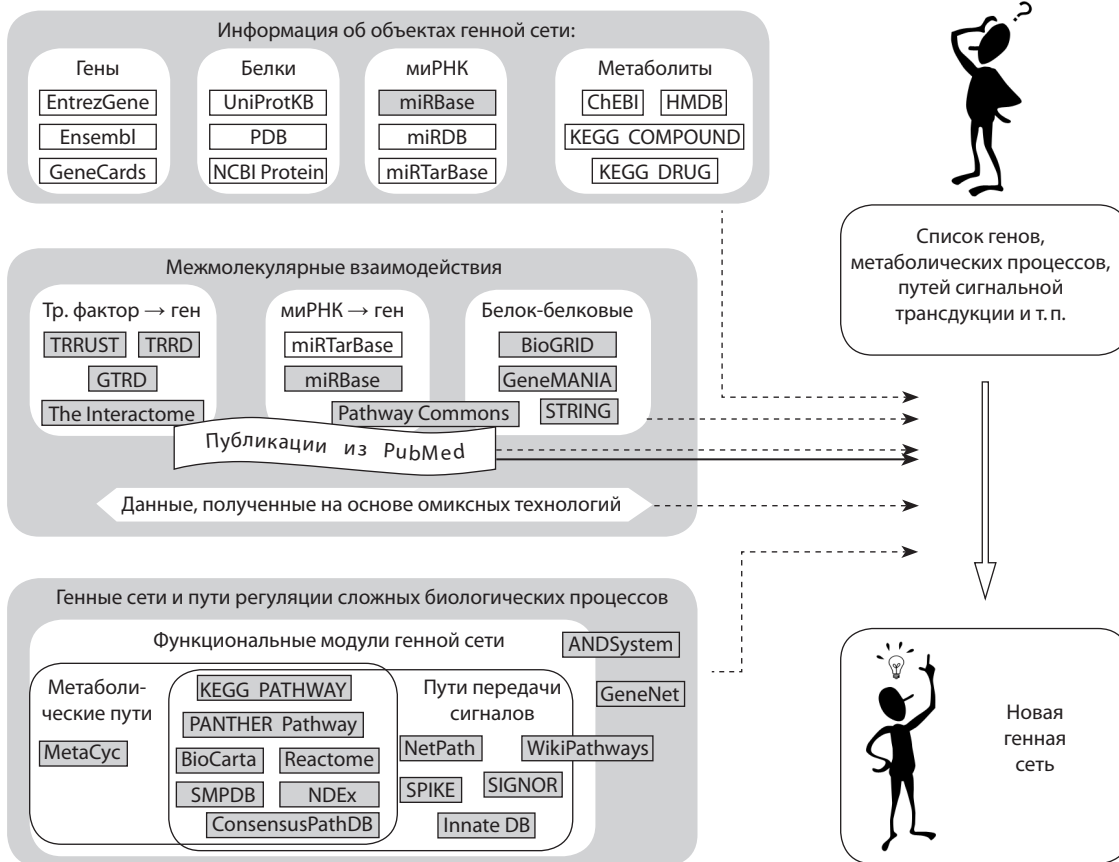
Концепция генных сетей постоянно развивается, обеспечивая тем самым методическую основу для решения все более широкого круга задач. Высокопроизводительные экспериментальные методы, применяемые в современных исследованиях, позволяют получать и анализировать все взаимодействия между определенными типами молекулярно-генетических объектов, идентифицированных в конкретном типе клеток (например, ДНК-белковые и белок-белковые взаимодействия). Массовое накопление подобных данных стало предпосылкой для выделения в рамках теории генных сетей таких сущностей, как: 1) сети транскрипционной регуляции (transcriptional regulatory networks) (Reese-Hoyes et al., 2005; Neph et al., 2012); 2) сети взаимодействий между миРНК и соответствующими генами-мишенями (miRNA regulatory networks) (Plaisier et al., 2012; Obermayer, Levine, 2014); а также 3) сети белок-белковых взаимодействий (protein-protein interaction (PPI) networks) (Lei et al., 2013; Ignatieva et al., 2016).

Рассмотрим в качестве примера сети транскрипционной регуляции (СТР). Эти сети отображают регуляторные взаимодействия типа «транскрипционный фактор → регулируемый ген-мишень». Каждая вершина СТР представляет сразу две сущности: ген и кодируемый им белок. Данные о регуляторных взаимодействиях в СТР могут быть получены на основе различных экспериментально-теоретических подходов, например: 1) компьютерного анализа регуляторных областей генов, выявленных методикой ДНКазы I футпринтинга (Neph et al., 2012); 2) сопоставления экспрессионных данных и данных, полученных методикой иммунопреципитации

хроматина с последующим секвенированием (ChIP-seq) (McMullen et al., 2014); 3) исследования эффектов выключения генов, кодирующих транскрипционные факторы, с помощью siRNA (Tomaru et al., 2014; Li et al., 2016). Как и все высокопроизводительные методы анализа, иммунопреципитация хроматина и ДНКазы I футпринтинг не позволяют получать данные со стопроцентным уровнем точности. Кроме того, компьютерные методы анализа, которые, как правило, применяют в таких исследованиях, характеризуются определенной долей ложноположительных и ложноотрицательных результатов (Kolchanov et al., 2007; Levitsky et al., 2007). Таким образом, регуляторные взаимодействия, выявленные с помощью описанных выше методов (а также ряда других высокопроизводительных методов анализа), должны быть использованы при реконструкции генных сетей с определенной осторожностью.

В настоящее время активно исследуются также ассоциативные генные сети, содержащие расширенный набор типов вершин (включая биологические процессы, заболевания, фенотипические характеристики, мутации, терапевтические воздействия). Такой подход позволяет объединять разрозненные локальные генные сети, не пересекающиеся между собой по молекулярно-генетическим объектам, и выявлять добавочные объекты (фенотипический признак, заболевание и т. д.), связывающие локальные генные сети (Glotov et al., 2015; Ivanisenko et al., 2015).

Реконструкция генных сетей – неотъемлемый этап исследования сложных молекулярно-генетических систем и процессов. При этом сначала конкретизируется объект исследования (фенотипический признак либо патологический процесс) и формируются списки элементарных биологических процессов, а также объектов (генов, белков, миРНК, метаболитов), вовлеченных в генную сеть. Данные об объектах генной сети доступны в базах данных EntrezGene, Ensembl, UniProtKB, miRBase, KEGG COMPOUND и др. (см. рисунок). Далее необходимо собрать информацию о взаимодействиях между объектами. Для этого можно обращаться как к разрозненным публикациям из PubMed, так и к специализированным информационным ресурсам, содержащим данные по различным типам межмолекулярных взаимодействий (BioGRID, GeneMANIA, miRBase, TRRUST, Pathway Commons и т. д.). Кроме того, данные о взаимодействиях между объектами (регуляторные, образование комплексов и т. д.) могут быть предсказаны на основе компьютерного анализа данных, полученных с помощью омиксных технологий (транскриптомных, протеомных, метаболомных и т. д.). При реконструкции генных сетей очень полезно использовать также сведения о функциональных модулях генных сетей (метаболических путях (МП), путях передачи сигналов (ППС) и др.), а также о путях регуляции сложных биологических процессов на клеточном и организменном уровне (ПРБП), которые представлены в современных базах данных. Наиболее крупными базами подобного рода являются KEGG PATHWAY, Reactome, MetaCyc, WikiPathways. Объем, а также разнообразие типов данных, доступных в интернет-ресурсах, чрезвычайно велики, в связи с чем планирование работ по реконструкции генных сетей обязательно должно включать этап поиска и систематизации сведений об источниках данных.



Использование различных информационных источников в процессе реконструкции геномных сетей.

Названия баз данных содержатся в прямоугольниках. Базы данных, охарактеризованные в настоящем обзоре, обозначены прямоугольниками с тенью. Сплошная стрелка обозначает процесс ручной аннотации научных публикаций, штриховые – автоматическую аннотацию данных либо экстракцию из баз данных.

Сведения об интернет-доступных информационных ресурсах по тематике «геномные сети» можно почерпнуть из современных веб-порталов. Например, PathGuide (<http://www.pathguide.org/>) отображает список из 690 баз, расклассифицированных по нескольким категориям, включая: 1) белок-белковые взаимодействия; 2) метаболические пути; 3) сигнальные пути; 4) сети транскрипционных взаимодействий. Другим очень полезным ресурсом является веб-коллекция баз данных по молекулярной биологии, размещенная на сайте журнала Nucleic Acids Research (NAR Molecular Biology Database Collection, <http://www.oxfordjournals.org/nar/database/c/>). Согласно этому ресурсу, в 2017 г. общий список баз данных по всем видам организмов, имеющих отношение к тематике «геномные сети», включал 143 наименования (36 из категории «метаболические пути», 93 из категории «белок-белковые взаимодействия» и 19 из категории «сигнальные пути») (Galperin et al., 2017).

В нашем обзоре охарактеризованы базы данных из вышеперечисленных тематических разделов веб-коллекции журнала NAR, ориентированные (в числе прочих видов) на человека и животных и находящиеся в открытом доступе в сети Интернет в 2017 г. Нами сформирована и охарактеризована подборка баз данных, содержащих сведения о метаболических и сигнальных путях, а также

о путях регуляции биологических процессов на клеточном и организменном уровнях (табл. 1 и Приложение 2¹). Чтобы познакомить читателя с предметной областью, мы представили характеристики ряда наиболее известных и значимых баз данных по межмолекулярным взаимодействиям. Таким образом, не претендуя на полноту охвата абсолютно всех баз данных, включающих сведения по геномным сетям человека и животных, представленный обзор знакомит читателя с современным состоянием проблемы, а также рассматривает критерии, согласно которым целесообразно оценивать полезность информационных ресурсов для конкретных исследовательских задач.

Типы информации и способы ее представления в базах данных

Содержащаяся в базах данных информация может быть отнесена к следующим иерархическим уровням: 1) пути регуляции сложных биологических процессов (включая заболевания и патологии), протекающих на уровне клетки, ткани, органа, а также целого организма; 2) функциональные модули геномных сетей – метаболические или сигнальные пути; 3) межмолекулярные взаимодействия. Информация, соответствующая первым двум иерархиче-

¹ Приложения 1–4 см. по адресу: <http://www.bionet.nsc.ru/vogis/download/pict-2017-21/appx13.pdf>

Таблица 1. Интернет-доступные базы данных по геномным сетям, метаболическим и сигнальным путям, путям регуляции других биологических процессов и молекулярно-генетическим взаимодействиям (представлены в алфавитном порядке)

Название базы	Способ наполнения или источник данных	Тип данных	Ссылка
ANDSystem	БД, ААТ	ББВ, РВ, коэкспрессия, участие белков в биологических процессах по данным UniProt-GOA, использование вещества для лечения заболевания, ассоциации «ген-заболевание»	http://www-bionet.sccc.ru/andvisio/
BiGG Models	Р	МП	http://bigg.ucsd.edu/
BioCarta	Р	МП, ППС, ПРБП, заболевания	https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways
BioCyc	Р, БД, КП	МП	https://biocyc.org/
BioGRID	Р	ББВ, ГВ	https://thebiogrid.org/
ConsensusPathDB	Р, БД	МП, ППС, ПРБП, ББВ, РВ, ГВ	http://cpdb.molgen.mpg.de/
GeneMANIA	ААТ, БД, КП	ББВ, ГВ, коэкспрессия, связи между генами/белками, предсказанные компьютерными методами	http://genemania.org/
GeneNet	Р	Структурно-функциональная организация ГС, ППС, МП	http://wwwmgs.bionet.nsc.ru/mgs/gnw/genenet/
GTRD	КАОД	ДНК-белковые взаимодействия	http://gtrd.biouml.org
InnateDB	Р, БД	ППС	http://www.innatedb.ca/
The Interactome	КАОД	Тканеспецифические СТР	http://www.regulatorynetworks.org/
KEGG PATHWAY	Р	МП, ППС, ПРБП, заболевания, лекарства	http://www.genome.ad.jp/kegg
miRBase	КАОД, КП, Р	РВ	http://www.mirbase.org/
MetaCyc	Р	МП	http://metacyc.org/
NDEx	БД, Р	МП, ППС, ПРБП, ББВ, РВ	http://www.ndexbio.org/#/
NetPath	Р	ППС человека	http://www.netpath.org/
PANTHER Pathway	Р, БД	МП, ППС	http://pantherdb.org/
Pathway Commons	БД	ББВ, РВ	http://www.pathwaycommons.org/
Reactome	Р	МП, ППС, пути транспорта молекул в клетке и репликации ДНК	http://www.reactome.org/
SIGNOR	Р	ППС и регуляторные взаимодействия между их участниками, ПРБП, заболевания	http://signor.uniroma2.it/
SMPDB	Р	МП, ППС, ПРБП, заболевания	http://smpdb.ca/
SPIKE	Р, БД	ППС человека	http://www.cs.tau.ac.il/~spike/
STRING	Р, ААТ, БД, КП	ББВ, коэкспрессия, связи между генами/белками, предсказанные компьютерными методами	string-db.org/
TRED	Р, БД, КП	СТР, РВ	http://rulai.cshl.edu/TRED
TRRD	Р	РВ, ДНК-белковые взаимодействия	http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/
TRRUST	ААТ, Р	СТР, РВ	http://www.grnpedia.org/trrust
WikiPathways	Р	МП, ППС, ПРБП, заболевания	https://www.wikipathways.org/index.php/WikiPathways

Примечание. Способы наполнения: Р – ручное аннотирование научных публикаций; ААТ – автоматический анализ текстов; БД – импорт из других баз данных; КП – компьютерные предсказания; КАОД – компьютерный анализ данных, полученных высокопроизводительными экспериментальными методами. Типы данных: ББВ – белок-белковые взаимодействия; ГВ – генетические взаимодействия; МП – метаболические пути; ППС – пути передачи сигналов; ПРБП – пути регуляции биологических процессов на клеточном и организменном уровне; РВ – регуляторные взаимодействия (транскрипционный фактор – регулируемый ген либо мРНК – регулируемый ген); СТР – сети транскрипционной регуляции. Более подробное описание баз данных со ссылками на наиболее свежие публикации представлено в Приложении 2.

ским уровням, как правило, представлена в базах данных в виде диаграмм. Данные, соответствующие третьему иерархическому уровню (межмолекулярные взаимодействия), отображаются в виде сетей взаимодействий между объектами.

Диаграммы. В таких базах, как KEGG PATHWAY, MetaCyc, WikiPathways, Reactome, SIGNOR, GeneNet и др., основными единицами информации являются диаграммы – статические изображения, содержащие фиксированный набор объектов (генов, белков, метаболитов и т. д.) и отображающие метаболические либо сигнальные пути, а также схемы регуляции сложных биологических процессов. Объекты диаграмм, а также узлы реакций и регуляторных процессов, как правило, интерактивны, обеспечивая переход к их текстовым описаниям.

В каждой базе данных используется специально разработанный графический язык представления информации. В качестве примера описан графический язык базы KEGG PATHWAY (Приложение 1).

Известны базы данных, диаграммы которых отображают только сигнальные пути (SIGNOR, SPIKE, InnateDB) (см. аббревиатуру ППС в табл. 1). Другие базы (KEGG PATHWAY, WikiPathways, GeneNet, BioCarta, PANTHER Pathway) представляют на диаграммах как метаболические, так и сигнальные пути, а также пути регуляции других процессов (МП, ППС, ПРБП в табл. 1). Базы KEGG PATHWAY, BioCarta, WikiPathways содержат также диаграммы, отображающие механизмы развития патологических процессов.

Помимо того, что диаграммы являются единицами информации в базах данных, они выполняют еще одну очень важную функцию. Диаграммы создаются аннотаторами-биологами и отражают их представления о регуляции биологических процессов. Эти представления формировались в ходе многолетних исследований. Таким образом, наряду с функциональной аннотацией генов терминами из словаря Gene Ontology, диаграммы биологических процессов служат еще одним курируемым источником данных о функциональной роли гена.

Сети взаимодействий. Данные о парных либо множественных межмолекулярных взаимодействиях (регуляторных, белок-белковых, формировании комплексов, биохимических реакциях и т. д.) очень важны для реконструкции генных сетей. Известен целый ряд баз данных (STRING, GeneMANIA, Pathway Commons, The Interactome, TRRUST и др.), накапливающих сведения о межмолекулярных взаимодействиях (табл. 2). Базы этой категории не содержат готовых статических диаграмм. Сети взаимодействий между заданным набором объектов (генов или белков) генерируются в этих базах на основании запроса (списка генов/белков, заданного пользователем). Результат запроса отображается в графическом виде, в виде интерактивной сети, включающей объекты (гены, белки и т. д.) и связи различных типов – ассоциации. В каждом случае имеется возможность получить данные об ассоциациях в текстовом виде. В качестве запроса может быть задан список генов/белков (STRING, GeneMANIA, Interactome, ConsensusPathDB и т. д.) или только один ген/белок (BioGRID, TRRUST). Основные типы межмолекулярных взаимодействий/ассоциаций

между объектами генных сетей, накопленные в соответствующих информационных ресурсах, представлены в табл. 2. Большинство из рассмотренных ресурсов содержат данные об ассоциациях различных типов. Отдельную группу составляют базы The Interactome, TRRUST, TRRD, TRED и GTRD, содержащие только один тип ассоциаций – регуляторные взаимодействия типа «транскрипционный фактор → регулируемый ген».

Способы наполнения баз данных и источники данных

Данные могут быть внесены в базу путем ручной аннотации научных публикаций, автоматической аннотации различных источников данных (текстов статей, результатов компьютерного анализа, данных экспериментов и т. д.), а также экстракции данных из других ресурсов и их интеграции в едином формате.

Ручная аннотация научных публикаций использовалась при создании баз данных KEGG PATHWAY, MetaCyc, WikiPathways, Reactome, SIGNOR, GeneNet, TRRD, а также некоторых других (см. табл. 1, базы с пометкой «Р»). Ручная аннотация – трудоемкий процесс, однако он обеспечивает более высокое качество информации по сравнению с данными, внесенными с помощью автоматической аннотации. Именно поэтому такие базы данных очень востребованы. Источниками информации могут быть как научные публикации, описывающие результаты экспериментов с отдельными генами и белками (основная масса данных в базах KEGG PATHWAY, MetaCyc, WikiPathways, SIGNOR), так и публикации, представляющие результаты современных высокопроизводительных экспериментов (подобные сведения накапливаются в BioGRID и ConsensusPathDB) и их анализа (накоплены в BiGG Models). Команды аннотаторов многих баз достаточно многочисленны, например, в наполнении базы WikiPathways участвовало более 400 экспертов из разных стран (Kutmon et al., 2016).

Автоматическая аннотация была использована (наряду с другими методами) при создании таких информационных ресурсов, как STRING, GeneMANIA, The Interactome, TRED и др. Этот способ позволяет получать данные о парных взаимодействиях (ассоциациях) между объектами (генами, белками и т. д.). Источниками данных при таком способе наполнения могут быть: 1) сведения, извлеченные из рефератов и текстов научных публикаций методами текст-майнинга (см. табл. 1, базы AAT); 2) ассоциации, выявленные (предсказанные) на основе компьютерного анализа геномов и протеомов (см. табл. 1, базы КП); 3) результаты компьютерного анализа данных широкомасштабных геномных, протеомных и транскриптомных исследований (см. табл. 1, базы КАОД).

Интеграция информации из интернет-доступных информационных источников является еще одним возможным способом организации новых информационных ресурсов. Таким путем создана база Pathway Commons (см. табл. 1, обозначение «БД»).

Кроме того, при создании информационных ресурсов широко применяется комбинация трех вышеперечисленных подходов (PANTHER Pathway, STRING, GeneMANIA, ConsensusPathDB, TRED, TRRUST, ANDSysystem).

Таблица 2. Информационные ресурсы, включающие данные по межмолекулярным взаимодействиям между объектами генных сетей/ассоциациями, и накопленные в них типы данных

Название ресурса	Тип межмолекулярных взаимодействий										
	Белок-белковые	Совместная встречаемость в публикациях	Совместная встречаемость в метаболических/сигнальных путях	Гомология белков	Кок-прессия	Локализация в геноме	Регуляторные TF → гены-мишени	миРНК → гены-мишени	Генетические*	Связь «ген-заболевание»	Связь «лекарство-молекула-мишень» или «лекарство-заболевание»
ANDSystem	☑	☑	☑**	×	☑	×	☑	☑	×	☑	☑
BioGRID	☑	×	×	×	×	×	×	☑	×	×	×
ConsensusPathDB	☑	×	☑	×	×	×	☑	×	☑	×	☑
GeneMANIA	☑	×	☑	☑	☑	×	×	☑	×	×	×
GTRD	×	×	×	×	×	×	☑	×	×	×	×
miRBase	×	×	×	×	×	×	×	☑	×	×	×
The Interactome	×	×	×	×	×	×	☑	×	×	×	×
Pathway Commons	☑	×	☑	×	×	×	☑	×	×	×	☑
STRING	☑	☑	☑	☑	☑	☑	×	×	×	×	×
TRED	×	×	×	×	×	×	☑	×	×	×	×
TRRD	×	×	×	×	×	×	☑	×	×	×	×
TRUST	×	×	×	×	×	×	☑	×	×	×	×

* Нарушение одного гена отражается на функционировании другого.

** По данным UniProt-GOA.

Программные инструменты баз данных

Программные средства баз данных обеспечивают поиск данных, их анализ, а также экспорт в различных форматах.

Практически все базы снабжены специальными поисковыми системами, позволяющими проводить поиск по названию объекта (гена, белка, метаболита и др.). В базах KEGG PATHWAY, Reactome, WikiPathways и др. результатом поиска являются диаграммы биологических процессов. В базах по межмолекулярным взаимодействиям (STRING, GeneMANIA, BioGRID, The Interactome, ANDSysSystem) на основе такого запроса (как уже отмечено выше) выполняется реконструкция сети взаимодействий, включающая заданный объект (или объекты, если их было несколько в запросе). В информационных системах STRING, GeneMANIA и ANDSysSystem есть возможность включить в сеть добавочные объекты (гены/белки), имеющие максимальное количество связей с объектами из сети. Таким образом, можно выявить наиболее короткий путь, связывающий два объекта, если изначальный результат запроса не содержал прямой связи между ними. В системе ConsensusPathDB имеется специальная опция «найти самый короткий путь».

Экспорт данных из баз можно осуществить: 1) в виде графических файлов, содержащих изображения диаграмм либо сетей взаимодействий между объектами (форматы png, pdf и др.); 2) в виде текстовых файлов (таблиц), содержащих списки объектов или бинарных взаимодействий между ними; 3) в специальных форматах (BioPAX, PSI-MITAB, SBML, SBGN, gpml, owl, pwl и т. п.), представляющих описание объектов и связей на диаграмме. Более подробно возможности экспорта данных описаны в Приложении 2. У некоторых баз (KEGG PATHWAY, Reactome, ConsensusPathDB, PANTHER Pathway, NDEx, BiGG Models) имеются программные интерфейсы APIs (application programming interfaces), позволяющие формировать и выполнять запросы ко всей информации, содержащейся в базах.

Наиболее распространенный и востребованный вариант анализа данных, предоставляемый веб-сайтами информационных ресурсов, – это анализ, проводимый с целью функциональной аннотации списка генов. В англоязычной литературе он именуется как «pathway enrichment analysis», а также «over-representation analysis, ORA» (Khatgi et al., 2012). Входными данными является список генов. В результате анализа пользователь получает сведения о диаграммах базы, в которых присутствуют гены из исходного списка, а также статистическую оценку повышенной (пониженной) представленности генов из исходного набора в диаграммах. Этот вид анализа широко используется для интерпретации функций групп дифференциально экспрессирующихся генов, полученных на основе микрочиповых технологий, а также для интерпретации данных транскриптомных и протеомных исследований. Программные средства для функциональной аннотации наборов генов имеются у баз PANTHER Pathway, ConsensusPathDB, Reactome, SPIKE, InnateDB и некоторых других.

Известны и другие программные инструменты, предоставляемые отдельными базами. Например, в базе KEGG PATHWAY есть возможность выделить на диаграмме объ-

екты из списка, введенного пользователем. В базах WikiPathways, SPIKE, NDEx имеются программные средства, позволяющие создавать диаграммы. Анализ информации из баз ConsensusPathDB, GeneMANIA, NDEx, InnateDB (и ряда других) можно осуществлять через специальные приложения (плагины) системы Cytoscape.

Заключение

Реконструкция генных сетей обеспечивает методическую основу для исследования закономерностей организации и механизмов генетического контроля функционирования живых систем (Neph et al., 2012; Stergachis et al., 2014), их эволюции (Mustafin et al., 2017), роли генов в развитии патологий (Ivanisenko et al., 2015; Ignatieva et al., 2016; Saik et al., 2016), а также для разработки математических моделей биологических процессов (Подколотный и др., 2016; Podkolodnaya et al., 2017).

В настоящем обзоре представлены основные сведения о ряде информационных ресурсов, содержащих данные по генным сетям и их функциональным модулям, которые находятся в открытом доступе в сети Интернет в 2017 г. и ориентированы на человека и животных. Известны и коммерческие базы данных по этой тематике (см. Приложение 4), которые могут быть полезны для реконструкции молекулярно-генетических сетей. Однако они доступны через Интернет на платной основе (через логин и пароль). Сведения об этих коммерческих системах предоставляются фирмами-разработчиками.

Сопоставление основных характеристик интернет-ресурсов (см. табл. 1 и 2, а также Приложение 2 и 3) выявило большое разнообразие предлагаемых ими возможностей. Кроме того, очевидно, что каждый ресурс содержит уникальный набор данных и использует оригинальные подходы к представлению информации. В связи с этим на этапе планирования работы по реконструкции генной сети определенного биологического процесса очень важно иметь представление о максимально полном наборе информационных источников, из которых может быть взята информация. Стремительный рост объема новых экспериментальных данных по молекулярно-генетическим взаимодействиям побуждает исследователей к созданию новых информационных ресурсов и выпуску более современных версий (релизов) уже известных баз. Поэтому для получения наиболее свежих сведений о базах по тематике «генные сети» целесообразно обращаться к ежегодным выпускам журнала *Nucleic Acids Research*, посвященным базам данных (первый номер каждого года) (Galperin et al., 2017), к другим веб-порталам, а также обзорным публикациям по этой тематике из специальных научных журналов (*Database: The Journal of Biological Databases and Curation*, *BMC Bioinformatics* и т. п.).

Благодарности

Работа выполнена при частичной поддержке проекта № 14.В25.31.0033 Правительства Российской Федерации (анализ возможностей баз данных, содержащих диаграммы путей передачи сигналов, метаболизма и регуляции других биологических процессов), а также бюджетного проекта № 0324-2016-0008 (анализ функциональных характеристик информационных ресурсов, содержащих

данные по межмолекулярным взаимодействиям между объектами генных сетей).

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов

Список литературы

- Колчанов Н.А., Ананько Е.А., Колпаков Ф.А., Подколodная О.А., Игнатиева Е.В., Горячковская Т.Н., Степаненко И.Л. Генные сети. Молекуляр. биология. 2000;34(4):533-544.
- Колчанов Н.А., Игнатиева Е.В., Подколodная О.А., Лихошвай В.А., Матушкин Ю.Г. Генные сети. Вавиловский журнал генетики и селекции, 2013;17(4/2):833-850.
- Подколodный Н.Л., Твердохлеб Н.Н., Подколodная О.А. Математическая модель циркадного осциллятора млекопитающих: взаимодействие с системой NAD⁺/SIRT1 и возрастные изменения экспрессии генов циркадного осциллятора. Вавиловский журнал генетики и селекции. 2016;20(6):848-856. DOI 10.18699/VJ16.201.
- Ратнер В.А. Генетические управляющие системы. Новосибирск: Наука, 1966.
- Galperin M.Y., Fernández-Suárez X.M., Rigden D.J. The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. Nucleic Acids Res. 2017;45(D1):D1-D11. DOI 10.1093/nar/gkw1188.
- Glotov A.S., Tiys E.S., Vashukova E.S., Pakin V.S., Demenkov P.S., Saik O.V., Ivanisenko T.V., Arzhanova O.N., Mozgovaya E.V., Zainulina M.S., Kolchanov N.A., Baranov V.S., Ivanisenko V.A. Molecular association of pathogenetic contributors to pre-eclampsia (pre-eclampsia associome). BMC Syst. Biol. 2015;9(Suppl.2):S4. DOI 10.1186/1752-0509-9-S2-S4.
- Ignatieva E.V., Afonnikov D.A., Saik O.V., Rogaev E.I., Kolchanov N.A. A compendium of human genes regulating feeding behavior and body weight, its functional characterization and identification of GWAS genes involved in brain-specific PPI network. BMC Genet. 2016;17(Suppl.3):158. DOI 10.1186/s12863-016-0466-2.
- Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. BMC Syst. Biol. 2015;9(Suppl.2):S2. DOI 10.1186/1752-0509-9-S2-S2.
- Kauffman S. Homeostasis and differentiation in random genetic control networks. Nature. 1969;224(5215):177-178.
- Khatri P., Sirota M., Butte A.J. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput. Biol. 2012;8(2):e1002375. DOI 10.1371/journal.pcbi.1002375.
- Kolchanov N.A., Merkulova T.I., Ignatieva E.V., Ananko E.A., Oshchepkov D.Y., Levitsky V.G., Vasiliev G.V., Klimova N.V., Merkulov V.M., Hodgman T.C. Combined experimental and computational approaches to study the regulatory elements in eukaryotic genes. Brief. Bioinform. 2007;8(4):266-274.
- Kutmon M., Riutta A., Nunes N., Hanspers K., Willighagen E.L., Bohler A., Mélius J., Waagmeester A., Sinha S.R., Miller R., Coort S.L., Cirillo E., Smeets B., Evelo C.T., Pico A.R. WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Res. 2016;44(D1):D488-494. DOI 10.1093/nar/gkv1024.
- Lei X., Wu S., Ge L., Zhang A. Clustering and overlapping modules detection in PPI network based on IBFO. Proteomics. 2013;13(2):278-290. DOI 10.1002/pmic.201200309.
- Levitsky V.G., Ignatieva E.V., Ananko E.A., Turnaev I.I., Merkulova T.I., Kolchanov N.A., Hodgman T.C. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. BMC Bioinformatics. 2007;8:481. DOI 10.1186/1471-2105-8-481.
- Li J.R., Suzuki T., Nishimura H., Kishima M., Maeda S., Suzuki H. Asymmetric regulation of peripheral genes by two transcriptional regulatory networks. PLoS ONE. 2016;11(8):e0160459. DOI 10.1371/journal.pone.0160459.
- McMullen P.D., Bhattacharya S., Woods C.G., Sun B., Yarborough K., Ross S.M., Miller M.E., McBride M.T., LeCluyse E.L., Clewell R.A., Andersen M.E. A map of the PPAR α transcription regulatory network for primary human hepatocytes. Chem. Biol. Interact. 2014;209:14-24. DOI 10.1016/j.cbi.2013.11.006.
- Mustafin Z.S., Lashin S.A., Matushkin Y.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(Suppl.1):1427. DOI 10.1186/s12859-016-1427-5.
- Neph S., Stergachis A.B., Reynolds A., Sandstrom R., Borenstein E., Stamatoyannopoulos J.A. Circuitry and dynamics of human transcription factor regulatory networks. Cell. 2012;150(6):1274-1286. DOI 10.1016/j.cell.2012.04.040.
- Obermayer B., Levine E. Exploring the miRNA regulatory network using evolutionary correlations. PLoS Comput. Biol. 2014;10(10):e1003860. DOI 10.1371/journal.pcbi.1003860.
- Plaisier C.L., Pan M., Baliga N.S. A miRNA-regulatory network explains how dysregulated miRNAs perturb oncogenic processes across diverse cancers. Genome Res. 2012;22(11):2302-2314. DOI 10.1101/gr.133991.111.
- Podkolodnaya O.A., Tverdokhle N.N., Podkolodny N.L. Computational modeling of the cell-autonomous mammalian circadian oscillator. BMC Syst. Biol. 2017;11(Suppl.1):379. DOI 10.1186/s12918-016-0379-8.
- Reece-Hoyes J.S., Deplancke B., Shingles J., Grove C.A., Hope I.A., Walhout A.J. A compendium of *Caenorhabditis elegans* regulatory transcription factors: a resource for mapping transcription regulatory networks. Genome Biol. 2005;6(13):R110.
- Saik O.V., Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. Interactome of the hepatitis C virus: Literature mining with ANDSystem. Virus Res. 2016;218:40-48. DOI 10.1016/j.virusres.2015.12.003.
- Stergachis A.B., Neph S., Sandstrom R., Haugen E., Reynolds A.P., Zhang M., Byron R., Canfield T., Stelting-Sun S., Lee K., Thurman R.E., Vong S., Bates D., Neri F., Diegel M., Giste E., Dunn D., Vierstra J., Hansen R.S., Johnson A.K., Sabo P.J., Wilken M.S., Reh T.A., Treuting P.M., Kaul R., Groudine M., Bender M.A., Borenstein E., Stamatoyannopoulos J.A. Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature. 2014;515(7527):365-370. DOI 10.1038/nature13972.
- Tomaru Y., Hasegawa R., Suzuki T., Sato T., Kubosaki A., Suzuki M., Kawaji H., Forrest A.R., Hayashizaki Y., FANTOM Consortium, Shin J.W., Suzuki H. A transient disruption of fibroblastic transcriptional regulatory network facilitates *trans*-differentiation. Nucleic Acids Res. 2014;42(14):8905-8913. DOI 10.1093/nar/gku567.