ПРИЛОЖЕНИЯ

К статье Е.В. Игнатьевой, Д.А. Афонникова, Н.А. Колчанова «Интернет-доступные информационные ресурсы по генным сетям, включающие данные по человеку и животным»

Приложение 1

Графический язык отображения информации о путях регуляции биологических процессов в базах данных

Рассмотрим способ представления информации в базах данных, содержащих диаграммы, на примере базы KEGG PATHWAY. Для графического отображения объектов и процессов используются специальные условные обозначения (рис. 1, *а*). Однако создатели диаграмм не ограничиваются стандартными обозначениями и часто включают в диаграммы другие графические изображения. Например, диаграмма Rheumatoid arthritis включает такие объекты, как схематическое изображение кровеносного сосуда, а также изображения специализированных клеток организма (макрофаги, остеобласты, остеокласты, Т- и В-клетки) (рис. 1, *б*).

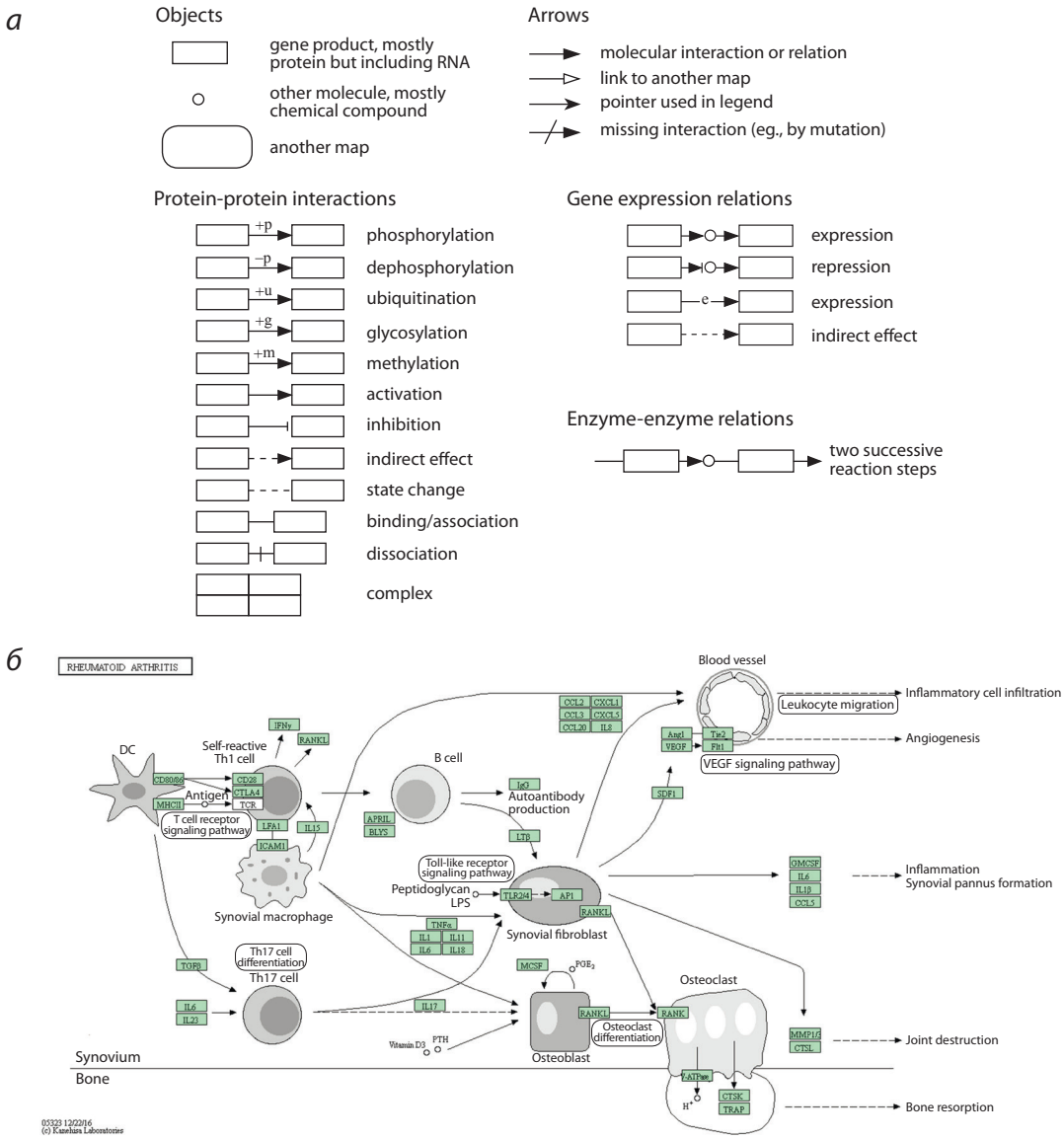

Рис. 1. Представление информации в базе KEGG PATHWAY.

а – условные обозначения для отображения объектов и их взаимодействий (приведены на веб-странице, доступной по адресу: http://www.genome.jp/kegg/document/help_pathway.html); *б* – диаграмма «Rheumatoid arthritis – *Homo sapiens* (human)», отнесенная разработчиками базы KEGG PATHWAY к категории иммунных заболеваний (см. разд. 6.3 *Immune diseases* в браузере базы KEGG PATHWAY). По данным KEGG PATHWAY, у человека в регуляции этого процесса участвуют 90 генов. Интерфейс базы обеспечивает возможность получения информации по сети ревматоидного артрита у 49 видов организмов.

Приложение 2

Характеристика баз данных

KEGG PATHWAY

KEGG PATHWAY – одна из самых крупных и известных баз по генным сетям, метаболическим и сигнальным путям. База развивается в рамках проекта Kyoto Encyclopedia of Genes and Genomes (г. Киото, Япония). В 2017 г. KEGG PATHWAY содержала созданные экспертами вручную графические представления и текстовые описания 496 метаболических либо сигнальных путей, схем регуляции биологических процессов, заболеваний, классификаций лекарственных веществ (Kanehisa et al., 2017). Каждая диаграмма представляет обобщенные данные по многим видам организмов. Специальная опция интерфейса базы позволяет настроить диаграмму на конкретный вид организма, причем количество видов зависит от того, насколько универсальный биологический процесс отражен на диаграмме.

В KEGG PATHWAY имеется иерархический классификатор диаграмм, включающий следующие разделы: (1) метаболические пути (178 диаграмм); (2) реализация генетической информации (транскрипция, трансляция) (22 диаграммы); (3) взаимодействие с внешней средой (сигнальная трансдукция) (38 диаграмм); (4) клеточные процессы (клеточный цикл, цитоскелет, хемотаксис) (24 диаграммы); (5) организменные процессы (иммунная, эндокринная система и др.) (78 диаграмм); (6) заболевания (81 диаграмма); 7) лекарства (75 диаграмм).

Информацию о каждой диаграмме базы KEGG PATHWAY можно скачать в графическом виде и в виде текстового описания (включая список идентификаторов генов). Имеется программный интерфейс KEGG REST API, позволяющий формировать и выполнять запросы ко всей информации, содержащейся в базе KEGG.

GeneNet

База GeneNet (ИЦиГ СО РАН, г. Новосибирск) является одной из первых баз по генным сетям, реконструированным с помощью ручной аннотации (Kolpakov et al., 1998). В GeneNet используется оригинальный метод формализованного графического представления генных сетей. Все объекты генной сети (гены, РНК, белки, небелковые молекулы и т. д., а также взаимодействия и реакции) характеризуются набором свойств (атрибутов), определяющих способ графического представления (цвет или форму) объектов на диаграммах (Kolpakov, Ananko, 1999; Ananko et al., 2002).

В интернет-доступной версии базы GeneNet представлены генные сети как прокариот (21 диаграмма), так и эукариот (55 диаграмм), включающие сведения о генах и белках 93 видов организмов. Эти генные сети регулируют: (1) клеточный цикл; (2) липидный метаболизм; (3) процессы в эндокринной системе; (4) созревание эритроцитов; (5) иммунный ответ; (6) процессы, протекающие в растительных клетках; (7) ответ клеток на стресс; (8) редокс-регуляцию; (9) метаболические реакции у прокариот (как правило, у *E. coli*). Поиск информации в GeneNet обеспечивается средствами поисковой системы SRS (Sequence Retrieval System), а также браузера диаграмм. Текстовые описания объектов, отображенных на диаграммах, доступны в xml формате (Ananko et al., 2005).

MetaCyc и BioCyc

MetaCyc и BioCyc – базы данных по метаболическим путям, разработанные группой исследователей из различных университетов США. MetaCyc и BioCyc имеют общий формат представления информации, общие поисковые системы и программные инструменты для анализа данных.

MetaCyc – одна из самых крупных баз данных по метаболическим путям, биохимическим реакциям и участвующим в них ферментам, собранных экспертами-биологами из научных публикаций. В 2017 г. MetaCyc содержала сведения о более чем 2400 метаболических путей 2816 видов организмов (как прокариот, так и эукариот), полученные из 46000 публикаций (Caspi et al., 2016).

Родственная база BioCyc накапливает данные, полученные на основе компьютерного анализа геномов 5700 видов организмов (большинство из которых – прокариоты). Для всех этих видов были сконструированы интегральные сети метаболизма, включающие как уже известные, так и предсказанные ферменты, биохимические реакции, метаболические пути, системы транспорта, а также опероны (для прокариот) (Caspi et al., 2016).

В рамках проекта MetaCyc–BioCyc сформированы специальные базы для человека и ряда модельных видов: HumanCyc (*Homo sapiens*), EcoCyc (*Escherichia coli*), AraCyc (*Arabidopsis thaliana*), LeishCyc (*Leishmania major*), YeastCyc (*Saccharomyces cerevisiae*). Эти базы включают данные по конкретному виду как из MetaCyc, так и из BioCyc.

Программные средства баз MetaCyc и BioCyc позволяют проводить сравнительный анализ метаболических путей различных организмов, а также конструировать интегральные схемы из заданного набора диаграмм. Информацию из баз MetaCyc и BioCyc можно получить через интерфейс прикладного программирования (APIs).

Reactome

База знаний Reactome содержит курируемые экспертами сведения по метаболическим и сигнальным путям, процессам транспорта молекул в клетке, репликации ДНК. Reactome разработана объединенными усилиями исследователей из разных стран (Англия, Канада, США, Китай). В Reactome накоплены данные о процессах и реакциях 15 видов эукариот и 4 видов прокариот. Наибольшее количество информации собрано по человеку и мыши (2148 и 1613 диаграмм соответственно).

Данные представлены в виде единой метаболической карты. Навигация по карте осуществляется с помощью иерархически организованного браузера. Специальная программа позволяет находить метаболические и сигнальные пути, содержащие определенный ген или набор генов, а также оценивать обогащенность найденных путей генами из данного набора (Fabregat et al., 2016). Данные базы Reactome можно скачать в форматах BioPAX, PSI-MITAB, SBML и SBGN.

WikiPathways

WikiPathways содержит схемы регуляции биологических процессов, построенные вручную, на основе анализа научных публикаций. WikiPathways является открытой платформой для сбора и распространения данных, поскольку ввод информации осуществляется зарегистрированными пользователями через специальную интернет-доступную программу. Благодаря такой системе WikiPathways чрезвычайно активно пополняется, и к настоящему моменту в ней насчитывается 2400 диаграмм по 25 видам организмов (в том числе 800 диаграмм по человеку и 200 по мыши), которые были созданы при участии более 400 экспертов из разных стран (Kutmon et al., 2016). Поиск данных в базе WikiPathways осуществляется по названию гена или любого другого объекта, а также по названию диаграммы. Все диаграммы можно скачать в различных форматах (gpml, svg, txt, owl, pwf, png, pdf).

SIGNOR

Ресурс SIGNOR (SIGnaling Network Open Resource) содержит информацию по сигнальным путям трех видов организмов (человек, мышь, крыса), которая внесена в систему на основе ручного анализа 6800 научных публикаций. В базе содержатся экспериментально подтвержденные данные о регуляторных взаимодействиях между объектами сигнальных путей: образовании комплексов, активации либо ингибировании активности белков на основе посттранскрипционных модификаций, регуляции транскрипции. В 2017 г. в базе имелись данные о 18200 взаимодействиях, в которых участвовало 4000 биологических объектов, в том числе 3800 белков. Диаграммы сигнальных путей интерактивны, их можно просматривать в нескольких режимах. В SIGNOR предоставляется доступ к текстовым описаниям объектов и взаимодействий, снабженным ссылками на экспериментальные статьи. Имеется поисковик по названию объектов, а также браузер диаграмм, которые сгруппированы по трем разделам: сигнальные

пути (60 входов), заболевания (4 входа), опухоли (8 входов). Данные можно скачать в текстовом виде в различных форматах (xls, csv и др.) (Perfetto et al., 2016).

SPIKE

SPIKE (Signaling Pathways Integrated Knowledge Engine) содержит данные о сигнальных путях человека. В 2017 г. в базе содержались 23 диаграммы, описывающие процессы регуляции клеточного цикла, программируемой клеточной гибели, ответа клетки на повреждение ДНК, а также регуляторные процессы в клетках внутреннего уха. Данные были внесены в базу SPIKE экспертами на основе анализа научных публикаций, а также экстракции и тщательной проверки данных из других ресурсов по генным сетям (Reactome, KEGG PATHWAY, NetPath, The Transcription Factor Encyclopedia, IntAct и MINT). Данные доступны для скачивания в текстовом виде (форматы XML, BioPAX и SIF) (Paz et al., 2011).

PANTHER Pathway

Ресурс PANTHER database (Protein Analysis THrough Evolutionary Relationships) содержит данные по функциям и эволюции белок-кодирующих генов, а также базу по метаболическим и сигнальным путям 104 видов организмов (Mi, Thomas, 2009). В 2017 г. в системе PANTHER содержались данные по 177 метаболическим и сигнальным путям, включающие 3092 объекта со ссылками на 6002 публикации. Сведения о метаболических и сигнальных путях были взяты как из других известных информационных источников, так и из научных публикаций (Mi et al., 2017). Доступ к диаграммам базы PANTHER осуществляется через браузер. Специальные программные средства позволяют осуществлять функциональную аннотацию списков генов, поданных на вход пользователем, предоставляя статистическую оценку их обогащенности генами из метаболических и сигнальных путей базы PANTHER.

NetPath

NetPath – база сигнальных путей человека (преимущественно вовлеченных в регуляцию иммунных процессов), построенных на основе ручного анализа 5500 публикаций. NetPath содержит 36 диаграмм, включающих 2800 объектов (метаболиты, белки и гены, для которых имеются данные о транскрипционных факторах, регулирующих активность этих генов) и 1600 реакций между ними. Данные базы NetPath можно загрузить в форматах BioPAX, PSI-MI, SBM (Kandasamy et al., 2010).

InnateDB

InnateDB – база, включающая информацию по сигнальным путям врожденного иммунного ответа у трех видов организмов (человека, мыши и крупного рогатого скота). InnateDB содержит: (1) экспериментально подтвержденные данные, полученные на основе ручной аннотации научных публикаций (18780 взаимодействий из 5235 источников); (2) сведения о межмолекулярных взаимодействиях из других баз данных (352782 взаимодействия); (3) данные о межмолекулярных взаимодействиях у крупного рогатого скота, предсказанные компьютерными программами, на основе сведений об ортологичных генах коровы и человека. Интерфейс InnateDB позволяет проводить поиск по названию гена, получать данные о сигнальных путях иммунной системы (в том числе 11 сигнальных путях человека), реконструировать сети взаимодействий для заданного набора генов. Визуализации сетей взаимодействий осуществляется программой Cerebral (Java плагин системы Cytoscape) с учетом локализации объектов в клетке. Кроме того, в InnateDB осуществляется анализ перепредставленности генов в сигнальных путях, содержащихся в базе. Данные могут быть экстрагированы в различных форматах (tab, csv, xls, sif, PSI-MI XML 2.5, MITAB) (Breuer et al., 2013).

BioCarta

Информационный ресурс BioCarta содержит более 300 схем метаболических и сигнальных путей, путей регуляции биологических процессов (включая патологии) у человека и мыши. В настоящее время

данные доступны только в графическом виде через браузер диаграмм, так как проект закрыт и пополнение базы новыми данными прекратилось.

SMPDB

SMPDB (Small Molecule Pathway Database) содержит 618 диаграмм путей регуляции различных процессов человека: (1) метаболических и сигнальных путей; (2) патологических процессов; (3) метаболизма лекарственных веществ и механизмов их действия; (4) сложных биологических процессов на клеточном и организменном уровне. Диаграммы созданы вручную, на основе аннотации научных публикаций. Данные можно загрузить в форматах BioPax SVG+BioPax SBGN SBML PWML (Jewison et al., 2014).

Pathway Commons

Pathway Commons – информационный ресурс, содержащий информацию о биохимических реакциях, образовании белковых комплексов, процессах транспорта и катализа, а также физических взаимодействиях, в которых участвуют белки, ДНК, РНК, их комплексы, а также метаболиты (Cerami et al., 2011). Pathway Commons интегрирует данные из 22 различных информационных источников и представляет их в едином формате, что обеспечивает быстрый поиск и загрузку данных. Источниками данных по связям и взаимодействиям в сетях являются такие известные базы, как: (1) BioGRID, IntAct, IntAct Complex, BIND, CORUM (белок-белковые взаимодействия); (2) TRANSFAC Public (взаимодействия между транскрипционными факторами и генами-мишенями); (3) MiRTarBase (взаимодействия между миРНК и генами-мишенями); (4) KEGG, Reactome, HumanCyc, Panther, WikiPathways, NetPath, PID (взаимодействия в пределах биологических путей). В базе Pathway Commons (версия 9) представлена информация о более чем 4000 биологических путей регуляции у человека, включающих 1300000 взаимодействий между объектами. Данные можно загрузить в различных форматах (BioPAX, SIF and GMT и т. д.).

ConsensusPathDB

ConsensusPathDB – ресурс, содержащий данные по трем видам (человеку, мыши и дрожжам) (Herwig et al., 2016). Разделы, содержащие данные по человеку и мыши, созданы на основе интеграции данных из 32 и 16 различных интернет-доступных информационных источников (соответственно), а также на основе аннотации научных публикаций. В ресурс интегрировано 5068 диаграмм человека и 2173 диаграммы мыши, включающие 534634 и 34064 взаимодействия между объектами (соответственно). ConsensusPathDB не содержит статических диаграмм, однако статистика по данному показателю имеется, поскольку специальная программа-вьюер ресурса ConsensusPathD позволяет отображать отдельные диаграммы из других баз и объединять диаграммы из нескольких баз. В базе представлены следующие типы взаимодействий: (1) регуляция экспрессии генов; (2) белок-белковые взаимодействия; (3) генетические взаимодействия; (4) биохимические реакции; (5) взаимодействие лекарств с их мишенями. Ресурс снабжен программными средствами, позволяющими выполнять функциональный анализ списков генов и метаболитов. Данные о белок-белковых взаимодействиях, а также о наборах генов, белков и метаболитов, функционирующих в биологических путях, можно скачать в текстовом виде (Herwig et al., 2016).

TRED

TRED (Transcriptional Regulatory Element Database) включает данные о сайтах связывания транскрипционных факторов и соответствующих сетях транскрипционной регуляции. В базе TRED представлены сети транскрипционной регуляции для транскрипционных факторов из 36 семейств, вовлеченных в развитие опухолевых процессов. Объектами сетей являются сами транскрипционные факторы, члены их семейств и регулируемые ими гены. В TRED включена информация о генах трех видов организмов – человека, мыши и крысы. Сети реконструированы как на основе данных об экспериментально идентифицированных сайтах связывания транскрипционных факторов, так и с использованием данных сайтах, предсказанных компьютерными программами.

База TRED представляет сети транскрипционной регуляции в виде диаграмм, снабженных краткими текстовыми описаниями (списками генов). В базе предусмотрены опции для анализа: поиск ортологичных генов, выявление консервных мотивов и потенциальных сайтов связывания и т. д. (Jiang et al., 2007).

NDEx

NDEx (the Network Data Exchange) – ресурс, накапливающий информацию о генных сетях разных видов организмов, включая человека. NDEx объединяет данные, загруженные из таких известных баз, как NCI – Pathway Interaction Database (PID), SIGNOR, Reactome (v46), Cancer Cell Map Initiative, NetPath. Другим источником данных являются сведения, напрямую загружаемые зарегистрированными пользователями. NDEx поддерживает загрузку данных в различных форматах (SIF, XGMML, BioPAX3, OpenBEL и др.), сохраняя при этом семантику исходных форматов. NDEx включает программные средства, обеспечивающие накопление, визуализацию и анализ данных различных типов (моделей биологических путей, сетей взаимодействий, а также новых знаний, полученных на основе анализа данных). В частности, активная ранее база NCI – Pathway Interaction Database (PID) в настоящее время доступна только через сервер NDEx. Согласно данным веб-браузера NDEx, в системе содержится 3 165 диаграмм различных видов организмов, 380 из которых включают гены человека. Поиск и анализ данных, накопленных в NDEx, осуществляется через веб-интерфейс, а также через API сервер и плагин системы Cytoscape. Экспорт данных возможен в форматах различных типов (CX, SIF, Microsoft Excel, TSV, GSEA) (Pratt et al., 2015; Pillich et al., 2017).

The Interactome

База Interactome содержит данные о тканеспецифичных транскрипционных регуляторных сетях. Эти данные были получены на основе данных эксперимента по ДНК-аза I футпринтингу в клетках человека и мыши. Далее с помощью теоретических методов анализа в регуляторных районах генов были выявлены сайты связывания 475 различных транскрипционных факторов. Таким образом были построены транскрипционные регуляторные сети для 41 типа клеток человека и 88 типов клеток мыши (Neph et al., 2012; Stergachis et al., 2014). Программа-визуализатор системы Interactome предоставляет возможность одновременно просматривать изображения двух тканеспецифических сетей, включающих заданный набор транскрипционных факторов, построенных для двух различных типов клеток человека либо мыши. Данные базы можно скачать в текстовом виде.

BiGG Models

BiGG Models (Biochemical, Genetic and Genomic knowledge base) содержит интегральные (т. е. уровня полного генома) схемы метаболизма для 23 видов организмов, 11 459 реакций и 4 040 метаболитов. Данные взяты из публикаций, представляющих результаты экспериментального (например, протеомное профилирование) и теоретического (например, аннотация генома) анализа, позволяющие охарактеризовать полный набор метаболических рекаций, протекающих в различных клетках. В BiGG Models представлены данные для одноклеточных организмов (*E. coli*, *B. subtilis*, *H. pylori*, *S. cerevisiae* и др.), также имеются данные по человеку и мыши. BiGG Models содержит три глобальные метаболические карты человека, две из которых ориентированы на специфические типы клеток (эритроциты и тромбоциты). Поиск данных базы BiGG Models осуществляется через веб-интерфейс, а также через API сервер. Данные можно загрузить в виде файлов (форматы txt и json) (King et al., 2016).

STRING

Система STRING включает данные о белок-белковых взаимодействиях, как подтвержденных экспериментально, так и выявленных на основе различных компьютерных методов. STRING включает также данные о связях между объектами (ассоциациях), выявленных на основе совместной встречаемости в метаболических или сигнальных путях, близости расположения в геноме, сходства филогенетических

профилей либо экспрессионных характеристик (коэкспрессия), а также на основе методов автоматического анализа текстов (Szklarczyk et al., 2017). В STRING содержатся данные об 1380 млн белок-белковых взаимодействий, участниками которых являются 9600 тыс. белков, относящихся к 2031 виду организмов. Данные о взаимодействиях доступны для скачивания в текстовом виде. Система STRING предоставляет возможность построить сеть взаимодействий между заданным набором генов/белков. Сеть отображается в графическом виде, специальные опции программы позволяют фильтровать данные по типу связей, находить добавочные гены/белки, имеющие наибольшее количество связей с объектами из сети, а также сохранять полученные данные в текстовом и графическом виде.

GeneMANIA

GeneMANIA – информационный ресурс, интегрирующий данные из различных источников (публикации, базы данных, компьютерные предсказания) о следующих типах связей между генами/белками: (1) коэкспрессия; (2) колокализация в геноме; (3) генетические взаимодействия; (4) общие биологические пути; (5) прямые белок-белковые взаимодействия; (6) наличие сходных белковых доменов (Zuberi et al., 2013). GeneMANIA содержит информацию о 597392998 взаимодействиях между 163599 генами/белками, принадлежащими 9 видам организмов (включая человека, мышь, крысу). GeneMANIA имеет две реализации: (1) интернет-доступная версия; (2) плагин системы Cytoscape. Интерфейс интернет-доступной версии системы GeneMANIA предоставляет такой же набор опций, как и у охарактеризованной выше системы STRING.

ANDSystem

Система ANDSystem (Ivanisenko et al., 2015) разработана для автоматической реконструкции ассоциативных генных сетей. Система ANDSystem включает три основных модуля. Первые два модуля являются серверными компонентами системы: (1) программы, осуществляющие экстракцию данных из PubMed и других баз; (2) база знаний ANDCell. Третий модуль, ANDVisio, является клиентским, он позволяет строить и визуализировать ассоциативные генные сети в различных режимах.

Данные внесены в базу знаний ANDCell двумя способами. Первый способ – автоматический анализ текстов (технология text-mining). На основе анализа 15 млн рефератов публикаций из PubMed за 1990–2015 гг. в базу ANDCell (версия 20160414) была включена информация о 9177866 межмолекулярных взаимодействиях между 16300426 объектами. Второй способ – экстракция данных из различных баз. Из этого источника были получены данные о 5546319 взаимодействиях. Белок-белковые взаимодействия были экстрагированы из баз IntAct и MINT, регуляторные взаимодействия – из базы TRRD, данные об участии белков в биологических путях – из базы InterPro, соответствия между генами и кодируемыми белками – из EntrezGene, взаимодействия между миРНК и регулируемыми генами/белками – из базы miRBase, а данные об участии белков в биологических процессах – из UniProt-GOA. Данные о взаимодействиях между объектами ассоциативных генных сетей человека, мыши и крысы составляют 23, 18 и 12 % от общего количества информации. В системе содержатся данные о взаимодействиях следующих типов:

(1) участие в биологических процессах по данным UniProt-GOA;
(2) участие белков в образовании комплексов;
(3) экспрессия (связь ген–белок);
(4) регуляторное взаимодействие (транскрипционный фактор/миРНК → ген-мишень);
(5) регуляция биохимического пути;
(6) регуляция транспорта;
(7) использование вещества для лечения заболевания;
(8) участие в каталитической реакции;
(9) регуляция активности белков;
(10) регуляция стабильности и деградации молекул;

(11) регуляция экспрессии белков;
(12) коэкспрессия генов;
(13) участие в расщеплении белков;
(14) катализ посттрансляционной модификации белков;
(15) ассоциации (взаимосвязи между генами и заболеваниями).

Для того чтобы использовать систему ANDSystem, необходимо зарегистрироваться в системе по адресу http://www-bionet.sscc.ru/andvisio/, скачать и установить на персональный компьютер клиентское программное приложение ANDVisio. Программа ANDVisio содержит интерфейс, позволяющий вводить списки объектов, из которых необходимо построить ассоциативную генную сеть, а также настраивать параметры запроса (выбирать вид организма, типы связей между объектами в сети и т. д.). ANDVisio позволяет также настраивать режим визуализации (раскладку сети), осуществлять поиск объекта в сети, добавлять или удалять объекты. Экспорт данных об объектах и связях сети возможен в виде графического изображения, а также в текстовом виде в форматах AND и TSV (Ivanisenko et al., 2015).

BioGRID

BioGRID (Biological General Repository for Interaction Datasets) – активно развивающаяся база данных, которая в настоящее время содержит сведения о 1 495 320 белковых и генетических взаимодействиях, 27 785 химических ассоциациях связях и 38 559 посттранскрипционных модификациях, выявленных у человека и 65 основных модельных организмов. Информация внесена в базу на основе ручного анализа 63 487 публикаций, содержащих данные, полученные различными экспериментальными методами, включая высокопроизводительные омиксные технологии. Поисковая система BioGRID позволяет находить информацию по названию гена либо публикации. Данные можно скачать в текстовом виде (форматы mitab, psi25, tab2) (Chatr-Aryamontri et al., 2017).

TRRUST

TRRUST (Transcriptional Regulatory Relationships Unravelled by Sentence-based Text-mining) – база транскрипционных регуляторных взаимодействий у человека и мыши. В настоящее время TRRUST (версия 2) содержит информацию о 8 908 и 7 382 регуляторных взаимодействиях, в которых участвуют 821 транскрипционный фактор человека и 859 транскрипционных факторов мыши (соответственно). Эти данные были экстрагированы из 11 237 статей, доступных в базе PubMed. Для 60 % взаимодействий имеется информация о типе влияния транскрипционного фактора на ген-мишень (активация или подавление транскрипционной активности). Данные получены на основе автоматического анализа текстов более 20 млн абстрактов из информационной системы PubMed и последующей ручной верификации. Поиск по базе TRRUST осуществляется по названию гена, результат выдается в графическом и текстовом виде. Данные базы могут быть также загружены в виде текстового файла (Han et al., 2015).

TRRD

TRRD (Transcription Regulatory Regions Database) разработана в ИЦиГ СО РАН (г. Новосибирск) на основе ручной аннотации экспериментальных данных из научных публикаций. TRRD содержит комплексное описание районов, регулирующих транскрипцию генов эукариот. Это описание включает данные о ДНК-белковых взаимодействиях между транскрипционными факторами, регулирующими транскрипцию гена, и их сайтах связывания в регуляторных районах генов. В базе также указываются эффекты транскрипционных факторов на транскрипцию генов (активация или подавление). Эти сведения можно получить по запросу к базе через стандартную поисковую систему SRS (Sequence Retrieval System). Поиск можно осуществлять по названию транскрипционного фактора либо по названию гена. Соответственно результатом будет либо список генов, регулируемых фактором, либо список факторов, регулирующих ген. В базе содержится информация о регуляции транскрипции генов различных видов эукариот, включая гены человека (763), мыши (528) и крысы (336) (Kolchanov et al., 2002).

GTRD

GTRD (Gene Transcription Regulation Database) – база данных, которая содержит сведения о сайтах связывания транскрипционных факторов в геномах человека и мыши, идентифицированных методикой ChIP-seq. Данные были получены на основе компьютерного анализа результатов экспериментов ChIP-seq из проектов ENCODE и SRA. Одним из этапов компьютерного анализа было распознавание потенциальных сайтов связывания транскрипционных факторов в районе ChIP-seq на основе PWM матрицы из базы HOCOMOCO.

GTRD содержит данные о сайтах посадки 476 транскрипционных факторов человека и 257 факторов мыши. Описание сайтов посадки транскрипционных факторов в GTRD включает позицию в геноме, а также информацию о клеточных линиях и условиях эксперимента.

Веб-интерфейс базы GTRD позволяет проводить поиск всех сайтов связывания транскрипционных факторов, находящихся в окрестностях гена, а также поиск всех генов, потенциально регулируемых транскрипционным фактором (Yevshin et al., 2017).

miRBase

miRBase – база, включающая информацию о миРНК и их мишенях в геноме. В miRBase (v20), содержатся данные о 24521 предшественнике и 30424 зрелых microRNA, выявленных у 206 видов организмов. Для вида *Homo sapiens* имеются сведения о 1881 предшественнике и 2588 зрелых миРНК.

Данные внесены в базу на основе сообщений от авторов статей, которые впервые идентифицируют миРНК. Регистрация миРНК в базе miRBase является необходимым условием опубликования таких статей. Существенная доля всех миРНК выявлена методом глубокого секвенирования.

Данные о взаимодействиях миРНК с последовательностями-мишенями в мРНК регулируемых генов накоплены в базе miRBase не только на основе данных экспериментов, но и на основе предсказания различными компьютерными программами.

Поиск миРНК возможен по названию миРНК, по ее первичной последовательности, ткани, в которой экспрессируется миРНК, а также публикации. Данные о взаимодействиях «миРНК → регулируемый ген» содержатся в каждом входе базы, описывающем конкретную миРКН (информационные поля «предсказанные мишени» и «доказанные мишени») (Kozomara, Griffiths-Jones, 2014).

Приложение 3

Объемы информации, представленной в базах по генным сетям и их функциональным модулям

Сопоставление объемов информации, накопленных в базах данных по метаболическим путям, путям передачи сигналов либо путям регуляции других биологических процессов, достаточно проблематично. Это связано в первую очередь с тем, что каждая база представляет статистику информационного содержания с разной степенью детализации, вследствие чего сложно выбрать показатель, который был бы представлен в статистических отчетах всех баз.

В настоящем исследовании нами было проведено сопоставление баз по одному из возможных критериев – количеству содержащихся в них диаграмм. Мы осознаем, что данный показатель не является полностью корректным, поскольку при реконструкции диаграмм разработчики различных баз руководствовались различными концепциями. Например, часть диаграмм в базах KEGG PATHWAY и HumanCyc представляет интегральные схемы всех метаболических путей, в то же время диаграммы многих баз включают только несколько десятков объектов (например, диаграммы путей сигнальной трансдукции в базе SIGNOR). Кроме того, количество объектов даже в одноименных диаграммах из разных баз может различаться в несколько раз (Stobbe et al., 2011; Chowdhury et al., 2015). Таким образом, сравнение количества диаграмм в базах позволяет получить лишь приблизительные оценки (рис. 2). Наибольшие показатели выявлены у ресурсов ConsensusPathDB, Pathway Commons, Reactome. Первые два ресурса созданы на основе интеграции информации из других баз (см. описание баз выше), что и позволило накопить максимальное количество данных.

Необходимо отметить, что базы данных из представленного на рис. 2 набора очень гетерогенны. Некоторые базы (SPIKE, SIGNOR) содержат информацию только по сигнальным путям, в то время как диаграммы других баз (KEGG PATHWAY, WikiPathways, SMPDB) отображают также метаболические пути и пути регуляции различных биологических процессов (включая патологии). Кроме того, ряд баз специализированы по узким тематикам: (1) процессы в иммунной системе (NetPath, InnateDB); (2) регуляция генов транскрипционными факторами, имеющими отношение к канцерогенезу (TRED); (3) клеточный цикл и программируемая клеточная гибель (SPIKE).

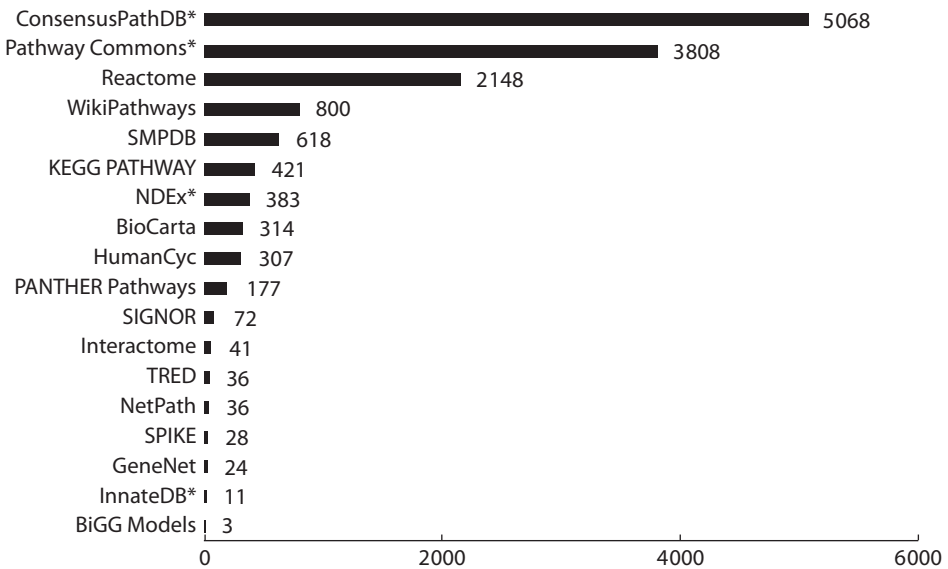

Рис. 2. Объемы данных по виду *Homo Sapiens* в публичных информационных ресурсах.

Цифры на горизонтальной оси соответствуют количеству диаграмм, содержащих сведения по виду *Homo Sapiens*, отображающих биологические процессы всех типов (метаболические и сигнальные пути, а также любые схемы регуляции процессов на клеточном и организменном уровне). Звездочкой помечены информационные ресурсы, созданные на основе интеграции информации из баз, представляющих данные в виде диаграмм.

Приложение 4

Примеры коммерческих баз данных по генным сетям, метаболическим путям, путям передачи сигналов и путям регуляции сложных биологических процессов на клеточном и организменном уровнях

Название базы	Фирма	Адрес
Pathway studio	Elsevier	http://www.pathwaystudio.com/
MetaCore	Thomson Reuters	https://clarivate.com/products/metacore/
Ungenital pathway (IPA)	QIAGEN Bioinformatics	https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/

Список литературы

Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. GeneNet: a database on structure and functional organisation of gene networks. Nucleic Acids Res. 2002;30(1):398-401.

Ananko E.A., Podkolodny N.L., Stepanenko I.L., Podkolodnaya O.A., Rasskazov D.A., Miginsky D.S., Likhoshvai V.A., Ratushny A.V., Podkolodnaya N.N., Kolchanov N.A. GeneNet in 2005. Nucleic Acids Res. 2005;33(Database issue):D425-7.

Breuer K., Foroushani A.K., Laird M.R., Chen C., Sribnaia A., Lo R., Winsor G.L., Hancock R.E., Brinkman F.S., Lynn D.J. InnateDB: systems biology of innate immunity and beyond-recent updates and continuing curation. Nucleic Acids Res. 2013;41(Database issue):D1228-33. DOI 10.1093/nar/gks1147.

Caspi R., Billington R., Ferrer L., Foerster H., Fulcher C.A., Keseler I.M., Kothari A., Krummenacker M., Latendresse M., Mueller L.A., Ong Q., Paley S., Subhraveti P., Weaver D.S., Karp P.D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2016;44(D1):D471-80. DOI 10.1093/nar/gkv1164.

Cerami E.G., Gross B.E., Demir E., Rodchenkov I., Babur O., Anwar N., Schultz N., Bader G.D., Sander C. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011;39(Database issue):D685-90.

Chatr-Aryamontri A., Oughtred R., Boucher L., Rust J., Chang C., Kolas N.K., O'Donnell L., Oster S., Theesfeld C., Sellam A., Stark C., Breitkreutz B.J., Dolinski K., Tyers M. The BioGRID interaction database: 2017 update. Nucleic Acids Res. 2017;45(D1):D369-D379. DOI 10.1093/nar/gkw1102.

Chowdhury S., Sarkar R.R. Comparison of human cell signaling pathway databases – evolution, drawbacks and challenges. Database (Oxford). 2015;2015. pii: bau126. DOI 10.1093/database/bau126.

Fabregat A., Sidiropoulos K., Garapati P., Gillespie M., Hausmann K., Haw R., Jassal B., Jupe S., Korninger F., McKay S., Matthews L., May B., Milacic M., Rothfels K., Shamovsky V., Webber M., Weiser J., Williams M., Wu G., Stein L., Hermjakob H., D'Eustachio P. The Reactome pathway knowledgebase. Nucleic Acids Res. 2016;44(D1):D481-7. DOI 10.1093/nar/gkv1351.

Han H., Shim H., Shin D., Shim J.E., Ko Y., Shin J., Kim H., Cho A., Kim E., Lee T., Kim H., Kim K., Yang S., Bae D., Yun A., Kim S., Kim C.Y., Cho H.J., Kang B., Shin S., Lee I. TRRUST: a reference database of human transcriptional regulatory interactions. Sci. Rep. 2015;5:11432. DOI 10.1038/srep11432.

Herwig R., Hardt C., Lienhard M., Kamburov A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. Nat. Protoc. 2016;11(10):1889-907. DOI 10.1038/nprot.2016.117.

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.

Jewison T., Su Y., Disfany F.M., Liang Y., Knox C., Maciejewski A., Poelzer J., Huynh J., Zhou Y., Arndt D., Djoumbou Y., Liu Y., Deng L., Guo A.C., Han B., Pon A., Wilson M., Rafatnia S., Liu P., Wishart D.S. SMPDB 2.0: big improvements to the Small Molecule Pathway Database. Nucleic Acids Res. 2014;42(Database issue):D478-84. DOI 10.1093/nar/gkt1067.

Jiang C., Xuan Z., Zhao F., Zhang M.Q. TRED: a transcriptional regulatory element database, new entries and other development. Nucleic Acids Res. 2007;35(Database issue):D137-40.

Kandasamy K., Mohan S.S., Raju R., Keerthikumar S., Kumar G.S., Venugopal A.K., Telikicherla D., Navarro J.D., Mathivanan S., Pecquet C., Gollapudi S.K., Tattikota S.G., Mohan S., Padhukasahasram H., Subbannayya Y., Goel R., Jacob H.K., Zhong J., Sekhar R., Nanjappa V., Balakrishnan L., Subbaiah R., Ramachandra Y.L., Rahiman B.A., Prasad T.S., Lin J.X., Houtman J.C., Desiderio S., Renauld J.C., Constantinescu S.N., Ohara O., Hirano T., Kubo M., Singh S., Khatri P., Draghici S., Bader G.D., Sander C., Leonard W.J., Pandey A. NetPath: a public resource of curated signal transduction pathways. Genome Biol. 2010;11(1):R3. DOI 10.1186/gb-2010-11-1-r3.

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.

King Z.A., Lu J., Dräger A., Miller P., Federowicz S., Lerman J.A., Ebrahim A., Palsson B.O., Lewis N.E. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 2016;44(D1):D515-22. DOI 10.1093/nar/gkv1049.

Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucleic Acids Res. 2002;30(1):312-7.

Kolpakov F.A., Ananko E.A. Interactive data input into the GeneNet database. Bioinformatics. 1999;15(7-8):713-4.

Kolpakov F.A., Ananko E.A., Kolesov G.B., Kolchanov N.A. GeneNet: a gene network database and its automated visualization. Bioinformatics. 1998;14(6):529-37.

Kozomara A., Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42(Database issue):D68-73. DOI 10.1093/nar/gkt1181.

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):D488-94. DOI 10.1093/nar/gkv1024.

Mi H., Huang X., Muruganujan A., Tang H., Mills C., Kang D., Thomas P.D. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 2017;45(D1):D183-D189. DOI 10.1093/nar/gkw1138.

Mi H., Thomas P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. Methods Mol. Biol. 2009;563:123-40. DOI 10.1007/978-1-60761-175-2_7.

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):1274-86. DOI 10.1016/j.cell.2012.04.040.

Paz A., Brownstein Z., Ber Y., Bialik S., David E., Sagir D., Ulitsky I., Elkon R., Kimchi A., Avraham K.B., Shiloh Y., Shamir R. SPIKE: a database of highly curated human signaling pathways. Nucleic Acids Res. 2011;39(Database issue):D793-9. DOI 10.1093/nar/gkq1167.

Perfetto L., Briganti L., Calderone A., Cerquone Perpetuini A., Iannuccelli M., Langone F., Licata L., Marinkovic M., Mattioni A., Pavlidou T., Peluso D., Petrilli L.L., Pirrò S., Posca D., Santonico E., Silvestri A., Spada F., Castagnoli L., Cesareni G. SIGNOR: a database of causal relationships between biological entities. Nucleic Acids Res. 2016;44(D1):D548-54. DOI 10.1093/nar/gkv1048.

Pillich R.T., Chen J., Rynkov V., Welker D., Pratt D. NDEx: A community resource for sharing and publishing of biological networks. Methods Mol. Biol. 2017;1558:271-301. DOI 10.1007/978-1-4939-6783-4_13.

Pratt D., Chen J., Welker D., Rivas R., Pillich R., Rynkov V., Ono K., Miello C., Hicks L., Szalma S., Stojmirovic A., Dobrin R., Braxenthaler M., Kuentzer J., Demchak B., Ideker T. NDEx, the Network Data Exchange. Cell Syst. 2015;1(4):302-305.

Stergachis A.B., Neph S., Sandstrom R., Haugen E., Reynolds A.P., Zhang M., Byron R., Canfield T., Stelhing-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):365-70. DOI 10.1038/nature13972.

Stobbe M.D., Houten S.M., Jansen G.A., van Kampen A.H., Moerland P.D. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. BMC Syst. Biol. 2011;5:165. DOI 10.1186/1752-0509-5-165.

Szklarczyk D., Morris J.H., Cook H., Kuhn M., Wyder S., Simonovic M., Santos A., Doncheva N.T., Roth A., Bork P., Jensen L.J., von Mering C. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017;45(D1):D362-D368. DOI 10.1093/nar/gkw937.

Yevshin I., Sharipov R., Valeev T., Kel A., Kolpakov F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. Nucleic Acids Res. 2017;45(D1):D61-D67. DOI 10.1093/nar/gkw951.

Zuberi K., Franz M., Rodriguez H., Montojo J., Lopes C.T., Bader G.D., Morris Q. GeneMANIA prediction server 2013 update. Nucleic Acids Res. 2013;41(Web Server issue):W115-22. DOI 10.1093/nar/gkt533.