УДК 573.2:573.22

СИСТЕМНАЯ БИОЛОГИЯ

© 2014 г. **Д.А. Афонников, В.В. Миронова**

Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия;
Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия,
e-mail: ada@bionet.nsc.ru, kviki@bionet.nsc.ru

Поступила в редакцию 5 ноября 2013 г. Принята к публикации 1 февраля 2014 г.

ВВЕДЕНИЕ

Первые попытки применения теории систем к биологии относятся к 30-м годам XX в. Так, в 1932 г. Уолтер Кэнон, декан факультета физиологии Гарвардского университета, в своей книге «Мудрость тела» («The wisdom of the body») описал термином «гомеостаз» способность организмов поддерживать большое число физиологических величин на постоянном уровне, несмотря на непрерывные изменения условий внешней среды. В 1943 г. американский математик Норберт Винер вместе с соавторами предположил, что отрицательные обратные связи могут играть центральную роль в поддержании стабильности живых систем, связав, тем самым, концепции контроля и оптимума с динамикой биологических систем.

В последние годы интерес к системному подходу в биологии был вызван прорывом в технологиях секвенирования и, как результат, расшифровке геномов, траскриптомов и протеомов человека и других организмов. Наличие мощных вычислительных ресурсов (суперкомпьютеров) и скоростных Интернет-соединений также значительно облегчило доступ к огромным массивам молекулярно-биологических данных и обеспечило возможность их анализа, что в значительной степени стало основанием для современной системной биологии. Об активном развитии этой области биологии в последнее время говорит следующий факт: количество статей, представленных PubMed и содержащих фразу «systems biology», увеличилось со 140 в 2003 г. до более 10 000 в 2013 г.

Целое больше, чем сумма его частей

Центральным в системной биологии является то, что взаимодействие нескольких агентов (таких как белки в генной сети) может демонстрировать новые состояния системы (клетки), возникающие как следствие их взаимодействий. Таким образом, во главу угла поставлен принцип «Целое больше, чем сумма его частей», предложенный еще Аристотелем в труде «Метафизика».

Что такое системная биология? Есть как минимум две разных точки зрения на этот вопрос. Для исследования взаимодействия компонент в биологических системах первая группа системных биологов занимается анализом результатов полногеномных экспериментов. Например, интеграция данных о связывании транскрипционного фактора с ДНК (ChIP-Seq) с данными об экспрессии генов в мутанте по этому транскрипционному фактору (RNA-Seq, Microarray) позволяет получить более значимые данные о топологии исследуемой генной сети, нежели анализ каждого из экспериментов в отдельности. Такой подход позволяет определить структуру системы.

Вторая группа системных биологов применяет принципы теории систем к моделированию динамики в биологической системе. Исследование функционирования системы в пространстве

и времени позволяет более точно описать те или иные процессы. Например, все чаще прибегают к математическому моделированию в биологии развития, в которой учет динамики наиболее значим.

В дополнение к этим двум направлениям в системной биологии в настоящий момент есть большая заинтересованность в развитии более мощных вычислительных средств и инструментов для проектирования и управления биологическими системами. Этот инженерный подход в будущем позволит применить на практике достижения системной биологии, например, при создании «биологических компьютеров», выращивании искусственных органов и тканей.

Хироаки Китано, один из пионеров системной биологии, для объяснения этого термина использует аналогию с дорожной картой (Kitano, 2001). Карта очень полезна, для оценки расположения городов и расстояния между ними, но она не раскрывает информацию о трафике и не позволяет его контролировать. То же можно сказать и о сети молекулярно-генетических взаимодействий. Даже если такая сеть будет полностью известна, использование этой информации для решения практических задач потребует динамической системы мониторинга, моделирования и тестирования.

Задачи системной биологии

В настоящее время можно говорить о смене парадигмы в биологических исследованиях от редукционистского подхода (разложение системы на более простые части и изучение поведения каждой части по отдельности) к синтетическому, выраженному принципом Аристотеля. Без сомнения, редукционистский подход в биологии позволил получить значительные результаты. Но у него есть серьезные ограничения, так как понимание поведения сложной системы путем простого расширения свойств отдельных его частей невозможно. Применение подходов системной биологии позволит исследовать такие сложные процессы, как морфогенез, патогенез, память и многое другое.

К задачам системной биологии можно отнести:

− интеграцию и хранение экспериментальных данных и результатов их анализа;

− разработку методов, подходов и технологий для анализа биологических данных;

− анализ биологических данных большого объема – полных геномов, транскриптомов, протеомов и т. п.;

− математическое моделирование динамики биологических систем.

Рассмотрим некоторые области применения системной биологии.

БАЗЫ ДАННЫХ В БИОЛОГИИ

Базы данных (БД) создаются, чтобы обеспечить накопление, хранение, систематизацию и анализ больших объемов информации. Для этого информация в них, в отличие от обычных текстов или научных статей, представляется в формализованном виде. Таким образом, базы данных представляют формализованное описание биологических объектов. В биологии многие базы данных состоят из записей, представляющих собой единицы информации. Например, в базе данных Genbank содержатся данные о нуклеотидных последовательностях. Каждая запись в этой БД представляет информацию об отдельной последовательности. Чтобы структурировать данные, каждую запись, в свою очередь, разбивают на поля (смысловые разделы). Набор полей и их порядок определяют специфику той или иной БД. Например, в базе Genbank для последовательности указываются ее название, авторы, которые ее секвенировали, публикации, в которых приводятся сведения о ней, организм, из которого эта последовательность была выделена, краткое описание функций, а также сама последовательность в стандартной кодировке (рис. 1). Такое, хоть и упрощенное, но стандартизированное, описание информации позволяет легко представлять ее в компьютерном виде, быстро проводить по ней поиск и ее компьютерную обработку.

База данных – представленная в объективной форме совокупность самостоятельных материалов, систематизированных таким образом, чтобы эти материалы могли быть найдены и обработаны с помощью электронной вычислительной машины (ЭВМ).

Идентификатор записи, дата создания	LOCUS HSRIB1 459 bp DNA linear PRI 27 -APR-2004
Название гена	DEFINITION Homo sapiens pancreatic ribonuclease rib-1 gene.
Идентификатор последовательности	ACCESSION X79235
Версия	VERSION X79235.1 GI:488412
Ключевые слова	KEYWORDS pancreatic ribonuclease.
Название организма и таксономическая информация	SOURCE Homo sapiens (human) ORGANISM Homo sapiens Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
Ссылка на основной литературный источник (авторы, заголовок статьи, название журнала, идентификатор статьи в БД Medline)	REFERENCE 1 AUTHORS Kochetov,A.V., Lukasheva,V.V., Filipenko,M.L., Mertvetsov,N.P. and Rivkin,M.I. TITLE Primary structure of the coding part of the gene for human pancreatic ribonuclease and its chromosomal location JOURNAL Bioorg. Khim. 21 (9), 691 -694 (1995) MEDLINE 96072053 PUBMED 8588814
Дополнительные ссылки на литературные источники	REFERENCE 2 (bases 1 to 459) AUTHORS Filipenko,M.L. TITLE Direct Submission JOURNAL Submitted (13 -MAY-1994) M.L. Filipenko, Institute of Bioorganic Chemistry, Lavrentjeva 8, Novosibirsk, 630090 RUSSIA
Описание функциональных районов последовательности (локализация гена, CDS, последовательность кодируемого белка и пр.)	FEATURES Location/Qualifiers source 1..459 /organism="Homo sapiens" /mol_type="genomic DNA" /db_xref="taxon:9606" /chromosome="14" /sex="female" /tissue_type="placenta" gene 1..459 /gene="rib-1" CDS 1..459 /gene="rib-1" /codon_start=1 /product="pancreatic ribonuclease" /protein_id="CAA55817.1" /db_xref="GI:488413" /db_xref="GOA:P07998" /db_xref="UniProt/Swiss-Prot:P07998" /translation="MGLEKSLVRLLLLVLILLVLGWVQPSLGKESRAKKFQRQHMDSD SSPSSSSTYCNQMMRRRNMTQGRCKPVNTFVHEPLVDVQNVCFQEKVTCKNGQGNCYK SNSSMHITDCRLTNGSRYPNCAYRTSPKERHIIVACEGSPYVPVHFDATV"
Нуклеотидная последовательность (в стандартной кодировке с нумерацией нуклеотидов)	ORIGIN 1 atgggtctgg agaagtctct tgtccggctc cttctgcttg tcctgatact gctggtgctg 61 ggctgggtcc agccttccct gggcaaggaa tcccgggcca agaaattcca gcggcagcat 121 atggactcag acagttcccc cag cagcagc tccacctact gtaaccaaat gatgaggcgc 181 cggaatatga cacaggggcg gtgcaaacca gtgaacacct ttgtgcacga gcccctggta 241 gatgtccaga atgtctgttt ccaggaaaag gtcacctgca agaacgggca gggcaactgc 301 tacaagagca actccagcat gcacatcaca gactgccgcc tgacaaacg g ctccaggtac 361 cccaactgtg cataccggac cagcccgaag gagagacaca tcattgtggc ctgtgaaggg 421 agcccatatg tgccagtcca ctttgatgct actgtgtag //

Рис. 1. Пример записи базы данных Genbank для гена панкреатической нуклеазы человека в текстовом формате.

Слева на русском языке приведена расшифровка основных полей базы данных.

Существующие открытые базы данных условно можно разделить на три категории. К первой относятся базы, содержащие первичные данные, полученные непосредственно в результате молекулярно-биологических экспериментов. К таким базам данных можно отнести информационные ресурсы, представляющие данные о последовательностях ДНК, РНК (Genbank, www.ncbi.nlm.nih.gov/genbank/) и белков (UniProt, www.uniprot.org), пространственных структурах генетических макромолекул (PDB, Protein Data Bank, www.pdb.org); результаты экспериментов по определению уровней экспрессии генов (GEO, Gene Expression Omnibus, www.ncbi.nlm.nih.gov/gds/).

Вторая категория включает базы, основанные на компьютерной обработке первичных молекулярно-генетических данных. Например, база данных Pfam (pfam.sanger.ac.uk) содержит информацию о последовательностях белков, сгруппированных в функциональные семейства: их выравнивание и описание функций. База данных PDBSite (wwwmgs.bionet.nsc.ru/mgs/gnw/pdbsite/) содержит информацию о структурах активных центров белков, полученных путем извлечения этой информации из белковых структур, представленных в банке PDB.

К третьей категории можно отнести информационные ресурсы, содержащие разнообраз-

ную интегрированную информацию о сложных биологических объектах, как первичные данные, так и компьютерную и экспертную аннотации. К таким системам можно отнести Ensembl (www.ensembl.org) – информационный ресурс по компьютерной геномике человека и ряда модельных организмов, базы данных по генным сетям Киотского университета KEGG (www.genome.jp/kegg/). Такое интегрированное представление все более активно разрабатывается в последнее время.

Один из важнейших источников пополнения БД новой информацией – это аннотирование, т. е. формализованное описание и характеристика биологических объектов, структур, процессов и т. п. Аннотирование может быть автоматическим (за счет использования специальных компьютерных программ, например, программ по поиску генов в расшифрованных нуклеотидных последовательностях), полуавтоматическим (когда результаты автоматического аннотирования проверяются специалистами) или ручным (когда информация вводится в БД экспертом). Экспертное аннотирование, как правило, подразумевает извлечение экспериментальной информации из научных статей и ее формализованное представление в БД. В большинстве современных БД совмещаются все три типа аннотирования.

Для обращения пользователей к информации БД разрабатываются интерфейсы, которые, как правило, работают через Интернет-браузеры. Усовершенствование системы поиска данных в БД, как правило, проходит путь от простого интерфейса, обеспечивающего доступ к каждому из типов хранящихся данных по отдельности, до интерфейса системных запросов, позволяющего интегрировать данные разных типов как внутри одного информационного ресурса, так и между разными информационными ресурсами (Ng *et al*., 2006; Stein *et al*., 2006). Примерами таких интегрированных интерфейсов для работы с большим количеством разнородных баз данных является система Entrez, разрабатываемая в Национальном центре биотехнологической информации (NCBI) (www.ncbi.nlm.nih.gov/Entrez/) и SRS Европейского института биоинформатики (srs.ebi.ac.uk).

В настоящее время в Интернете представлено большое количество БД, в которых пользователь может найти информацию о геномах, транскриптомах и протеомах различных организмов. Их количество постоянно растет. Для того чтобы помочь биологам ориентироваться в этом информационном потоке, журнал «Nucleic Acids Research» (nar.oxfordjournals.org) ежегодно публикует специальный выпуск, посвященный базам данных в биологии, и поддерживает список всех ранее опубликованных в нем баз данных (www.oxfordjournals.org/nar/database/a/), который насчитывает в настоящее время 1512 ресурсов, доступных через Интернет. Ниже рассмотрим некоторые наиболее популярные из них, представляющие информацию из разных областей системной биологии.

Выше уже были упомянуты основные источники информации по нуклеотидным (Genbank) и аминокислотным последовательностям (UniProt), трехмерным структурам макромолекул (PDB), метаболическим сетям (KEGG). Существуют также специализированные базы данных, в которых пользователь может найти такие данные, как: 1) структуру и функцию микроРНК (miRBase, www.mirbase.org); 2) функциональную аннотацию генов (Gene Ontology, www.geneontology.org); 3) сети генных и белковых взаимодействий (STRING, string-db.org); 4) последовательности и аннотацию регуляторных элементов генов (TRRD, wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/).

При решении биологических задач биологу, как правило, бывает недостаточно информации только лишь одной базы. Для того чтобы обеспечить унифицированный доступ к базам данных, разрабатываются системы, обеспечивающие однообразный доступ пользователя к большому количеству баз данных, распределенных по различным серверам. К таким системам относится BioMart (www.biomart.org). Она позволяет обращаться к десяткам баз данных, используя программный язык Perl; создавать похожие интерфейсы для визуализации данных в Интернет-браузерах.

АНАЛИЗ ПОЛНОГЕНОМНЫХ ДАННЫХ ПО ЭКСПРЕССИИ ГЕНОВ

Одним из ключевых процессов в живых клетках является экспрессия генов – процесс

Экспрессия генов – процесс реализации генетической информации от генов к молекулярным объектам. Он включает несколько стадий (транскрипция гена, процессинг пре-мРНК, трансляция мРНК, посттрансляционные модификации белков).

реализации генетической информации от генов к молекулярным объектам. Он включает несколько стадий (транскрипция гена, процессинг пре-мРНК, трансляция мРНК, посттрансляционные модификации белков). Поэтому оценивать экспрессию гена можно на каждом из этих этапов экспрессионного процесса. Одним из наиболее используемых методов является оценка уровня экспрессии по концентрации мРНК. В этом направлении были разработаны несколько подходов. Часть из них направлены на оценку концентраций мРНК одного или небольшого числа генов. Однако в последние 15 лет бурно развиваются технологии, которые позволяют оценивать уровни экспрессии генов по концентрации мРНК в масштабах всего трансриптома. К этим методам относятся технологии микрочипов и RNA-Seq.

Существуют два основных типа постановки эксперимента по анализу экспрессии генов. К первому относится оценка дифференциальной экспрессии. В рамках этой задачи необходимо сравнить уровни экспрессии генов по концентрациям их мРНК для тканей в двух или нескольких состояниях ткани, как правило, это нормальные условия (контроль) и воздействие каких-либо факторов (метаболитов, химических реагентов, лекарств) или заболеваний (например, рак). В результате этих экспериментов удается выяснить, экспрессия каких генов повышается при воздействии на ткань различных факторов, а каких, наоборот, понижается.

Ко второму типу задач относится изучение экспрессии генов в серии экспериментов. Серия экспериментов может включать разные концентрации воздействующего на ткань вещества, изменение уровня экспрессии генов на разных стадиях клеточного цикла или в разные моменты времени. Такие эксперименты позволяют ответить на вопрос, какие гены имеют сходные профили экспрессии в этих экспериментах (коэкспрессируются).

Микрочиповые эксперименты

При полногеномном анализе экспрессии генов широкое распространение получили микрочиповые эксперименты. Существует несколько технологий для решения этой задачи, предлагаемых различными компаниями (Agilent, Nimblegene, Illumina и др.). Одной из популярных платформ являются олигонуклеотидные микрочипы компании Affymetrix (отметим, что компании предлагают микрочипы, основанные на несколько других технологиях, позволяющие решать задачи подобного рода, их описание выходит за пределы данного обзора и может быть получено читателем на сайтах соотвествующих компаний). Они представляют собой однонитевые молекулы ДНК, закрепленные одним концом на твердой подложке, которые называются ДНК-зондами или пробами. Длина зондов составляет 25 нуклеотидов. В процессе эксперимента ДНК-зонды гибридизуются за счет формирования комплементарных взаимодействий с присутствующими в образце фрагментами РНК транскриптома, помеченными флюоресцентной меткой (рис. 2). Интенсивность гибридизации РНК с пробами определяется оптической лазерной системой. Условия экспериментов по микрочипам устанавливаются таким образом, что интенсивность сигнала флюоресценции для спаренных зондов оказывается пропорциональной концентрации свободной РНК в образце.

Набор олигонуклеотидов для микрочипа разрабатывается с учетом генома организма, который необходимо исследовать, и набора генов, экспрессию которых необходимо определить. Микрочипы Affymetrix, имеющие одинаковые наборы олигонуклеотидов, называются «платформой». Например, для полногеномного анализа экспрессии генов человека используются микрочипы платформы U133A Plus 2.0. Эта платформа создана для анализа 18400 транскриптов и их вариантов, содержит 500 тыс. олигонуклеотидных проб. Для анализа экспрессии *A. thaliana* разработан микрочип ATH1, который содержит зонды для 22 тыс. генов этого организма.

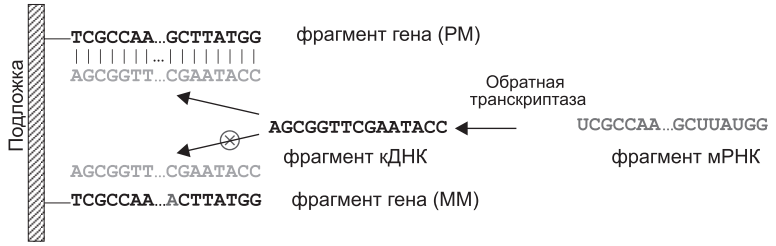

Рис. 2. Схема оценки концентрации мРНК по технологии Affymetrix.

Чтобы охарактеризовать уровень экспрессии транскрипта, микрочип содержит для него от 20 до 25 ДНК-зондов, последовательности которых удовлетворяют ряду условий: являются уникальными для данного транскрипта; при гибридизации спаривание оснований должно происходит без петель; зонды должны обладать сходными термодинамическими параметрами для всего микрочипа. Дополнительно для каждого зонда, соответствующего олигонуклеотиду, способному формировать совершенное спаривание с фрагментом транскрипта (такие типы зондов обозначаются типом PM, perfect match), на чипе располагают зонды, которые содержат в центре ДНК-зонда один неспаренный нуклеотид (рис. 2, такие типы зондов обозначаются MM, mismatch). Добавление зондов с неспаренным нуклеотидом направлено на использование сравнения интенсивности гибридизации для пар PM/MM. Это позволяет получить более точную оценку уровня концентрации РНК транскриптов за счет устранения ошибок, возникающих из-за неспецифической и фоновой гибридизации. Кроме того, для более надежной оценки уровня концентрации РНК на чипе располагают несколько наборов проб для гибридизации молекул РНК, транскрипты которых отсутствуют в исследуемом геноме. Это контрольные пробы (англ. spike-in), РНК для которых добавляют в образец искусственно в заданных концентрациях. При анализе данных для дополнительной коррекции ошибок некоторые алгоритмы используют значения интенсивности сигнала для контрольных проб.

Предварительная обработка данных микрочиповых экспериментов

Описание расположения проб на микрочипе задается в формате CDF: для каждой пробы в текстовом виде указываются ее координаты на чипе, принадлежность к набору проб для определенного транскрипта и т. п. Эти файлы поставляются производителем чипов и доступны свободно на его сайте (www.affymetrix.com). Результаты измерения интенсивности гибридизации для каждой пробы по итогам эксперимента представляются в файле формата CEL. Приводятся данные по среднему значению интенсивности свечения пикселей на цифровом изображении микрочипа для данной пробы, стандартное отклонение и количество пикселей, соответствующих ДНК-зонду на изображении. Один эксперимент представляет собой результаты обработки одного микрочипа, записанные в одном CEL файле.

Для того чтобы оценить уровень сигнала гибридизации по этим данным, существует несколько стандартных процедур их предварительной численной обработки. Цель этого заключается в нормализации данных таким образом, чтобы можно было бы сравнивать оценки уровней экспрессии генов, произведенных в сериях из нескольких экспериментов, т. е. данные из нескольких CEL файлов. Существует несколько алгоритмов нормализации микрочиповых данных, в их числе RMA, GCRMA, алгоритмы локальной регрессии LOESS. Один из них предложен разработчиком технологии, компанией Affymetrix, и называется MAS 5.0. Описание алгоритма доступно на сайте компании (http://media.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf).

Работа алгоритма производится в несколько этапов:

1. Коррекция фонового шума. На этом этапе область чипа делится на 16 прямоугольных зон одинакового размера. Интенсивность шума оценивается отдельно в каждой зоне как среднее значение уровней интенсивности 2 % проб с на-

именьшим уровнем сигнала. Далее для каждой пробы уровень шума пересчитывается как взвешенное значение шума каждой из зон, причем вес зоны тем меньше, чем дальше от этой зоны на чипе расположена проба. Тем самым, учитывается возможная неравномерность сигнала по поверхности чипа. Далее уровень шума вычитается из уровня сигнала каждой пробы.

2. Оценка уровня экспрессии (величины сигнала). Оценка уровня экспрессии производится для наборов проб, характеризующих концентрацию РНК от определенного транскрипта. Для этого используются данные по парам проб PM/MM. Для оценки уровня экспрессии используются только те пробы, для которых интенсивность сигнала зонда с идеальной гибридизацией (PM) значительно выше, чем уровень сигнала зонда с несовпадением (MM). Затем для набора проб оценивается логарифм сигнала (SLVi, signal log value) с помощью метода среднего Тьюки, устойчивого к выбросам (без 2 % наибольших и 2 % наименьших значений). После этого происходит нормализация уровня сигнала для набора проб к общему уровню так, что становится возможным сравнивать уровни экспрессии разных мРНК.

3. Оценка значимости отличия сигнала от фона (определение *p*-значений уровня экспрессии). Для того чтобы оценить значимость отличия уровня сигнала от шума, используется ранговый критерий Вилкоксона. Этот тест позволяет оценить значимость отклонения дискриминирующего параметра R_i для набора проб *i*

$$R_i = \frac{PM_i - MM_i}{PM_i + MM_i}$$

от порогового значения τ, величина которого устанавливается по умолчанию равной 0,015. Это позволяет оценить уровни значимости *p* отличия сигнала от шума.

4. Оценка наличия/отсутствия сигнала. Алгоритм MAS 5.0 определяет три категории сигнала: сигнал присутствует (P, present), сигнал незначительный (M, marginal detection) или отсутствует (A, absent).

Полученные в результате данные по нормализованному сигналу каждой из набора проб сохраняются в формате CHP вместе с параметрами нормализации (шкалирующий и нормализующий факторы).

Сравнение уровней экспрессии генов в нескольких экспериментах

Для того чтобы оценить изменение уровня экспрессии гена в различных условиях, используют серии экспериментов, один из которых является контрольным. Это данные по значениям интенсивности нескольких микрочипов с одинаковым набором проб. Требуется определить статистическую значимость изменения среднего уровня экспрессии транскрипта (набора проб) в каждом эксперименте по сравнению с контрольным экспериментом. Алгоритм MAS5.0 для каждого транскрипта *i* из эксперимента *j* и контрольного эксперимента сообщает следующие состояния: 1) существенное увеличение уровня экспрессии по сравнению с контрольным экспериментом; 2) незначительное увеличение уровня экспрессии; 3) существенное уменьшение уровня экспрессии по сравнению с контрольным экспериментом; 4) незначительное уменьшение уровня экспрессии; 5) отсутствие изменения уровня экспрессии по сравнению с контрольным экспериментом.

Другой подход, реализованный, например, в системе Gene Expression Omnibus NCBI, позволяет использовать для сравнения группы экспериментов, т. е. сравнение уровней экспрессии транскриптов в ряде повторных экспериментов, соответствующих контрастным условиям. Для этого используется *t*-статистика Стьюдента,

$$t_i = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

где m_1, m_2 – средние значения уровня экспрессии транскрипта *i* в наборе экспериментов в условиях 1 и 2 соответственно; s_1, s_2 – соответствующие стандартные отклонения; n_1, n_2 – число экспериментов первого и второго типов соответственно. На основании значений t_i можно оценить значимость отличий уровней экспрессии транскрипта в двух группах экспериментов. Она выражается в значениях *p*, оценивающих вероятность того, что наблюдаемые различия в уровне экспрессии транскрипта в двух группах экспериментов вызваны случайными причинами при условии, что экспрессия транскрипта в двух экспериментах одинакова. Чем меньше величина *p*, тем с большей вероятностью можно

утверждать, что экспрессия транскрипта в двух экспериментах различается. Принято считать, что значимой является величина $p < 0{,}01$.

ГЕННЫЕ СЕТИ, БЕЛОК-БЕЛКОВЫЕ ВЗАИМОДЕЙСТВИЯ

Системы взаимодействующих объектов можно представить в виде графов. В этом случае объекты представляются вершинами графа, а ребра соединяют вершины, объекты которых взаимодействуют между собой. В системной биологии объекты – молекулярные сущности (белки, гены, РНК, метаболиты и пр.), взаимодействия между ними также принято называть сетями. Например, сеть белок-белковых взаимодействий в качестве вершин содержит все белки организма, а ребра графа этой сети связывают два узла, если соответствующие им белки формируют белковый комплекс. Такой граф, полученный для протеома дрожжей, представлен на рис. 3.

Граф $G = (V, E)$ состоит из двух множеств: конечного множества элементов, называемых вершинами (V), и конечного множества элементов, называемых ребрами (E). Каждое ребро определяется парой вершин. Если ребра определяются упорядоченными парами вершин, то G называется *направленным*, или *ориентированным* графом. В противном случае граф G называется *ненаправленным*, или *неориентированным* графом.

В зависимости от типов молекулярных объектов и характера взаимодействий описания живых систем могут быть представлены: сетями белок-белковых взаимодействий сетями регуляторных взаимодействий (описывают регуляцию экспрессии одних белков другими), метабо-

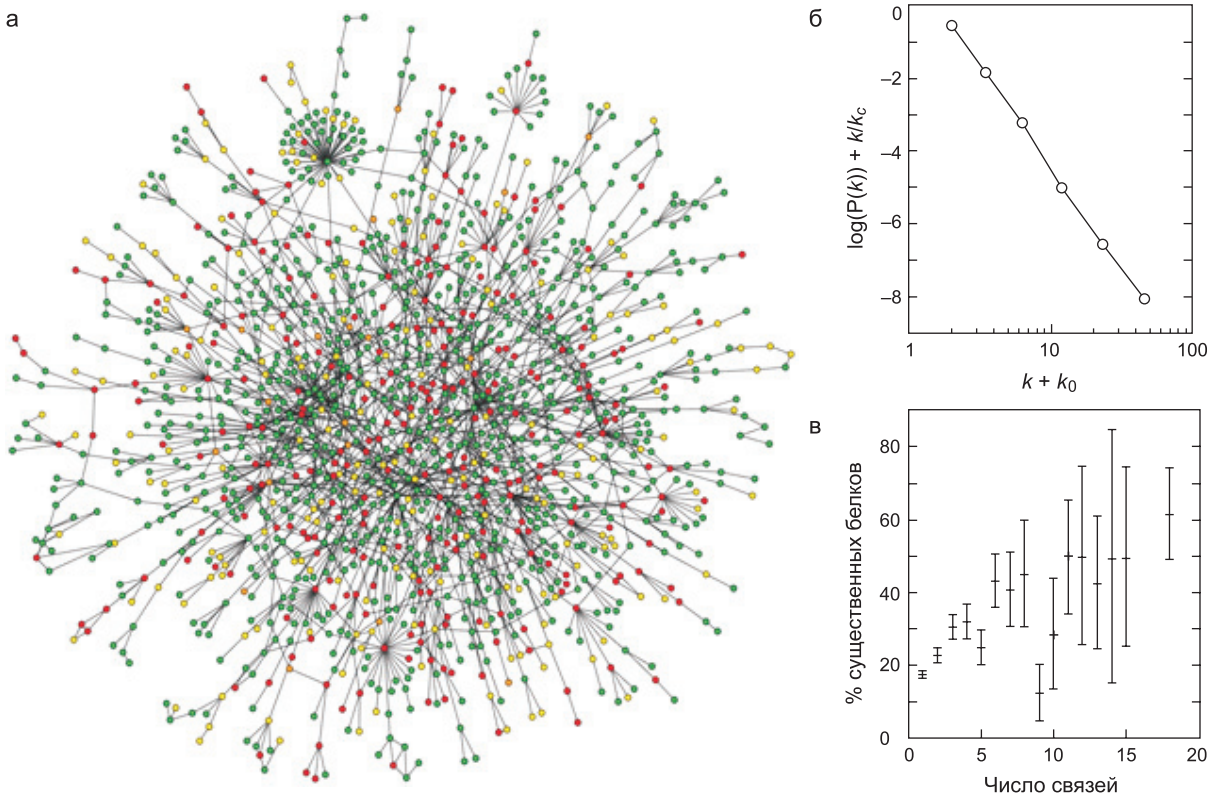

Рис. 3. Сеть белок-белковых взаимодействий дрожжей *S. serevisiae* (Jeong *et al.*, 2001).

а – наибольший по размеру фрагмент сети. На рисунке каждый узел соответствует белку, связи объединяют взаимодействующие белки. Цвет узла отражает фенотипический эффект проявления мутации белка: летальный (красный), нейтральный (зеленый), эффект неизвестен (желтый); б – зависимость числа связей узла в генной сети (ось Y) от количества таких узлов (ось X) в двойном логарифмическом масштабе; в – зависимость доли белков, мутации в которых летальны (существенные белки), и имеющих в точности k связей (ось Y) от числа связей (k, ось X).

лическими сетями (описывают взаимосвязь метаболических процессов в клетке), сетями передачи сигналов (сигнальными путями), сетями коэкспрессии (описывающими совместную экспрессию генов) и др. Интеграция указанных процессов осуществляется в рамках генных сетей, молекулярно-генетических систем, обеспечивающих формирование фенотипических характеристик организмов (молекулярных, биохимических, структурных, морфологических, поведенческих и т. д.) на основе информации, закодированной в их геномах (Колчанов и др., 2008).

Важно отметить характерные особенности организации генных сетей. Прежде всего, это наличие элементов контроля (положительных и отрицательных обратных связей), модульная организация, вырожденность структуры (когда одну и ту же функцию могут выполнять несколько ее элементов). Все это позволяет генным сетям в процессе эволюционной изменчивости геномов обеспечивать устойчивость как по отношению к изменению структуры сети, так и по отношению к изменению концентраций ее молекулярных компонентов.

Для больших биологических сетей самой разной природы такую устойчивость обеспечивает специфическая организация: эти сети содержат большое число узлов с малым числом связей и малое число узлов с большим числом связей (их обычно называют хабами, от англ. «hub» – ступица колеса). Несколько таких хабов хорошо заметны на графе сети белок-белковых взаимодействий (рис. 3, а). В двойном логарифмическом масштабе для таких «безмасштабных» сетей (Barabasi *et al*., 2009) распределение числа связей подчиняется линейному закону (рис 3, б). Безмасштабные сети обладают высокой устойчивостью к случайным ошибкам: удаление взятого наугад узла практически не влияет на связность сети (т. е. число вершин, которые необходимо посетить, чтобы из одной вершины перейти по ребрам до любой другой вершины). Анализ эффекта мутаций в белках дрожжей показал, что доля белков, мутации которых летальны, для узлов с небольшим числом связей относительно невелика. Например, белки, имеющие 5 и менее связей, составляют 93 % протеома, однако только по 21 % из них мутации являются летальными для организма. В то же время 0,7 % белков имеют число взаимодействий более 15, однако их мутации являются летальными до 61 % белков (Jeong *et al*., 2001). Интересно также отметить, что сравнение последовательностей генов из разных организмов показало, что для генов, соответствующих хабам в биологических сетях, наблюдается более частое превышение доли фиксации синонимических замен в кодонах над несинонимическими. Это свидетельствует о преимущественной эволюции генов-хабов в режиме стабилизирующего отбора и их важной функциональной роли в организме.

Устойчивая к случайным повреждениям структура сети формируется за счет дупликаций ее фрагментов (генов, отдельных блоков), при этом часть существовавших связей в сети сохраняется. В результате новые вершины с большей вероятностью оказываются связанными с вершинами, уже имеющими некоторое число связей (реализуется принцип «богатый становится еще богаче» по отношению к числу связей у вершин).

ФУНКЦИОНАЛЬНАЯ АННОТАЦИЯ ГЕНОВ

Задача описания функции генов сложна как в силу неполноты наших знаний о процессах, происходящих в живой клетке, так и в силу того, что многие гены выполняют несколько функций, участвуют в нескольких процессах жизнедеятельности клетки. Петер Карп (Karp, 2000) предложил рассматривать два уровня описания функции генов, кодирующих белки. Во-первых, эта функция определяется теми биохимическими взаимодействиями, которые способен реализовывать белок (так называемая локальная функция). Во-вторых, для белков существует еще и интегральная функция, т. е. роль, которую белок играет в общей системе функционирования клетки. Для описания локальной функции в настоящее время используются базы данных функциональной аннотации белков (по первичной и третичной структуре). К их числу относятся БД Prosite (описание функциональных сайтов в белках; http://prosite.expasy.org/); Pfam (http://pfam.sanger.ac.uk/), CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), Superfamily (http://supfam.cs.bris.ac.uk/SUPERFAMILY/),

описывающие эволюционно консервативные семейства функциональных доменов в белках; SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) и CATH (http://www.cathdb.info/), описывающие структурные домены белков.

Другим способом описания функции служат ключевые слова: наборы терминов, характеризующих функциональные свойства белков. Эти ключевые слова, как правило, в свободной форме присваиваются записям генов в БД (см. рис. 1). Расширением такого описания стало создание контролируемых иерархических словарей, в которых функция генов характеризуется на разных уровнях (онтологий). Для описания функций генов была создана база данных «Онтология генов» (Gene Ontology (GO); http://www.geneontology.org/). Эта база содержит наборы контролируемых словарей, которые классифицируют понятия, связанные с функцией гена, и устанавливают связи между этими понятиями. В базе данных GO описаны молекулярная функция гена, биологические процессы, в которых он участвует, компартмент, в котором ген функционирует. Для построения словарей существует несколько правил:

– словари строятся иерархически;

– термины в такой иерархии могут иметь одного или нескольких «родителей», ни одного, одного или несколько «потомков»;

– термины связаны дополнительными отношениями: «является» (is-a) и «часть» (part-of).

Удобство такого описания заключается в том, что все термины в словарях контролируются, иерархическая структура позволяет эффективно обрабатывать данные в компьютере, выбирать уровень детализации описания функции гена. На рис. 4 представлен пример функциональной классификации ряда генов дрожжей, дрозофилы и мыши, выполняющих функцию связывания с ДНК.

В базе данных представлена достаточно полная функциональная аннотация генов и большинства модельных организмов, а также человека. Это позволяет решать задачи массовой обработки данных по функциям десятков тысяч генов одновременно.

К типичным задачам подобного рода относятся задачи функциональной аннотации генов, имеющих повышенный уровень экспрессии в микрочиповом эксперименте. Например, если

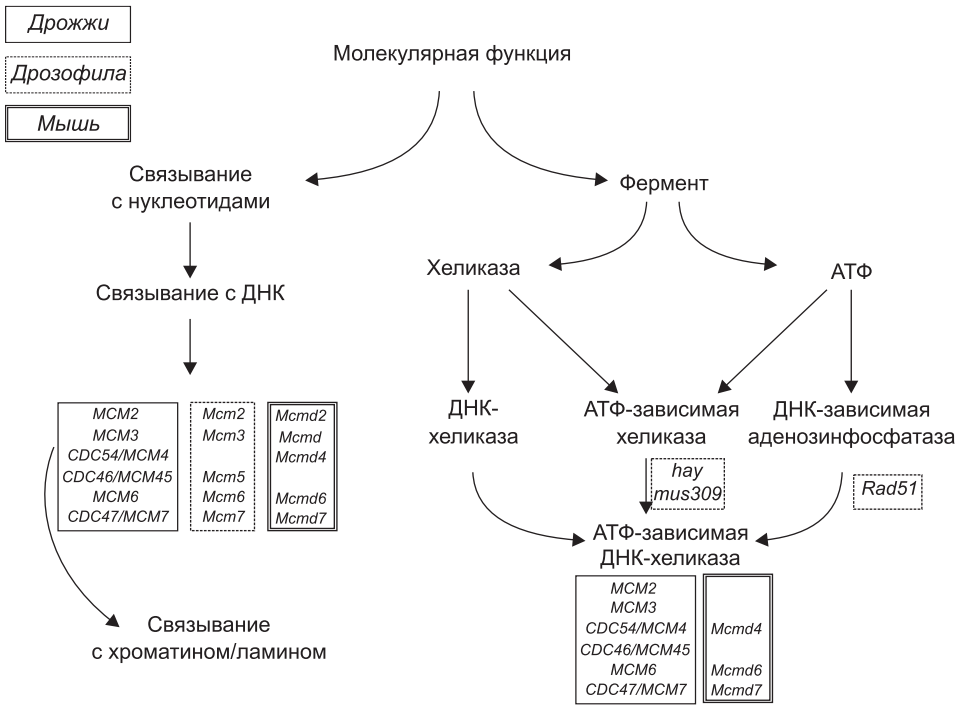

Рис. 4. Пример функциональной классификации ряда генов дрожжей, дрозофилы и мыши, выполняющих функцию связывания с ДНК в системе иерархического описания функций Gene Ontology.

в этом эксперименте из всех высоко экспрессирующихся генов 10 % составляют киназы (в геноме их доля составляет около 1 %), это может означать, что исследованный процесс может иметь отношение к функции киназ (передачи сигналов). Такой анализ (в английской терминологии он называется «gene enrichment» – анализ насыщенности генами (Huang *et al*., 2009) – позволяет ответить на вопрос: какие функции организма могут быть связаны с повышенной (или пониженной) экспрессией той или иной группы генов? Значимость обнаруженной корреляции определяется на основе ряда статистических критериев (точного критерия Фишера, критерия χ^2, перестановочного теста и т. п.). В настоящее время такого рода анализ сопровождает все эксперименты по полногеномному анализу экспрессии генов.

Одним из наиболее популярных ресурсов в этой области биоинформатики является DAVID (The Database for Annotation, Visualization and Integrated Discovery; Huang *et al*., 2008). Особенностью этого ресурса является использование интегрированной базы знаний, в которой унифицированы идентификаторы генов из самых разных молекулярно-биологических баз данных, а также микрочиповых платформ. Это позволяет пользователю в большинстве случаев (однако не всегда) не заботиться об идентификации генов в списке по идентификаторам – ресурс делает это автоматически. Кроме того, данный ресурс осуществляет разнообразное представление полученных результатов: от списка генов с аннотациями, которые можно скачать на персональный компьютер, до реконструкции функциональных связей между генами, выраженными в виде списков кластеров терминов и аннотаций. К удобствам этой базы также относится большой объем данных, бесплатный доступ и наличие аннотации для большого числа организмов.

АНАЛИЗ ДАННЫХ ПО ЭКСПРЕССИИ ГЕНОВ

Рассмотрим пример анализа экспрессии генов с помощью микрочипа фирмы Affymetrix по данным, представленным в работе Чанга (Chung *et al*., 2011). В этой работе проводились два эксперимента над корнями проростков *A. thaliana*: 1) контрольный образец тканей корня при эндогенном воздействии раствора 2,4-дихлорфеноксиацетиловой кислоты (синтетического гормона роста растений ауксина) и 2) образец тканей корня после эндогенного воздействия такого же раствора с добавлением брассиназола (ингибитора синтеза брассиностероидов, стероидных гормонов роста растений). Известно, что ауксин вызывает экспрессию гена DWARF4, кодирующего фермент DWARF4. Этот фермент участвует в осуществлении лимитирующей стадии биосинтеза брассиностероидов в тканях растений. Цель эксперимента заключалась в выявлении того, какую роль может играть ауксин в биосинтезе брассиностероидов, стимулирующих рост растений. Данные по экспрессии генов в этом эксперименте были депонированы в базу данных GEO за номером GDS3823. Рассмотрим пример выявления дифференциально экспрессирующихся генов в двух этих экспериментах и анализ функциональной аннотации этих генов.

На рис. 1 (Приложение 2)* представлен пример анализа этих данных в системе GEO NCBI. На главной странице этого ресурса необходимо ввести идентификатор данных, приведенный в статье (1), в поле Datasets необходимо ввести идентификатор и нажать кнопку «Go». В результате открывается страница с записями, связанными с данным экспериментом (2). С этой страницы по ссылке «Datasets» необходимо перейти на страницу с описанием данного эксперимента (3). Необходимо перейти по ссылке «Analyze Dataset» на страницу анализа данных (4). Всего в этом эксперименте представлены 6 результатов определения экспрессии генов: 3 в базовом эксперименте (auxin) и 3 в эксперименте с воздействием брассиназола («Auxin and brassinazole»).

На странице анализа необходимо выбрать ссылку «Compare 2 sets of samples». В этом случае появляется всплывающее окно (6), которое позволяет сгруппировать данные измерений в группы A и B (в данном случае измерение экспрессии без брассиназола (A) и с брассиназолом (B)). После формирования списка нажатие кнопки «OK» позволяет перейти к странице анализа (7), на которой нужно установить

* Приложение 2 см. по адресу: http://www.bionet.nsc.ru/vogis/pict_pdf/2014/18_1/appx_2.pdf.

уровень значимости для дифференциально экспрессирующихся генов (0,005) и выбрать правило сравнения (в данном случае выбрана опция «отобрать гены», экспрессия которых выше в образце A, т. е. чья экспрессия при воздействии брассиназола понижается). Затем необходимо нажать на ссылку «Query group A vs B». В результате будет получена страница, представленная на рис. 2 (1) (Приложение 2).

На рис. 2 (1) (Приложение 2) представлен пример анализа генов, полученных в результате отбора в системе GEO при помощи системы DAVID. Как описано выше, отбирались гены, которые имели в группе экспериментов A (воздействие ауксина) уровень экспрессии выше, чем в группе экспериментов B (воздействие ауксина и брассиназола). Результаты в виде списка генов представлены на странице результатов GEO (1); всего было определено 722 пробы на чипе, чье изменение концентраций удовлетворяет выбранному условию. Для дальнейшего анализа необходимо получить список генов, соответствующих этим пробам. Для этого необходимо на странице результатов (1) в правой панели выбрать раздел «Find related data» и в меню «Databases» этого раздела выбрать пункт «Genes». После этого нажать на кнопку «Find items». Откроется страница со списком соответствующих генов (2).

На следующем этапе необходимо скачать список отобранных генов в файл. Для этого нажать ссылку «Send to» и в появившемся окне установить опции загрузки: тип сохранения «Send to File», формат данных – «UI list». Полученный файл необходимо открыть в текстовом редакторе, как показано на рисунке (3). Выделить все полученные идентификаторы и скопировать их в буфер памяти (как это делается в любом текстовом редакторе). Данные в буфере памяти готовы для дальнейшего анализа.

Далее необходимо открыть страницу сервиса DAVID (http://david.abcc.ncifcrf.gov/). Откроется главная страница (4). Для функциональной аннотации дифференциально экспрессирующихся генов необходимо выбрать ссылку в левой панели меню «Functional Annotation». После этого откроется страница ввода данных (5). В левом поле во вкладке «Upload» в поле «A: Paste a list» необходимо вставить из буфера список идентификаторов отобранных генов. Установить следующие опции: «Step 2: Select Identifier» в меню «ENTREZ_GENE_ID», которые указывают, что нами будут использованы для анализа идентификаторы генов из БД Entrez; «Step 3: List Type» установить в положение «Gene List». Это означает, что мы будем анализировать частоты встречаемости терминов аннотации генов в нашем списке по отношению к частотам встречаемости этих терминов в полном списке генов *A. thaliana*. После этого необходимо нажать кнопку «Submit list». В результате получим страницу результатов, показанную на рисунке (6).

На этой странице есть возможность загрузить результаты в виде списка кластеров аннотаций («Functional annotation clustering»), таблицы и графиков («Functional annotation Charts») и таблицы («Functional annotation Table»). Рассмотрим для примера первую опцию. При нажатии этой кнопки загружается страница результатов, в которой аннотации выбранных генов кластеризованы по функциональным группам и для этих групп приведены наиболее информативные термины (7). Полученные данные можно загрузить в файл, выбрав ссылку «Download File».

Проведенный анализ позволил выделить гены у *A. thaliana*, которые высоко экспрессируются при воздействии ауксина, но в то же время их экспрессия понижена при ингибировании брассиностероидов. Таким образом, это те гены, которые, вероятнее всего, регулируются ауксином напрямую, а не опосредованно, через брассиностероиды. Согласно полученным данным это гены, кодирующие гликопротеины, гены, связанные с передачей сигнала, продукты которых секретируются (аннотация кластера 1, (7)). В то же время аннотация этих генов непосредственно связана с ауксином, согласно аннотации кластера 2 (7).

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ В БИОЛОГИИ

Одним из основных направлений системной биологии является математическое моделирование динамики биологических процессов. Даже если исследователь знает механизм протекания того или иного процесса, например биосинтеза аминокислоты, регуляции экспрессии сигнального пептида, дифференцировки стволовой

клетки и. т. д., исследование этого процесса в динамике позволяет ответить на множество дополнительных вопросов, таких как:

– изменяется ли концентрация результирующих веществ при изменении условий протекания реакции? Например, как меняется концентрация сахара в крови при стрессе (рис. 5, а);

– изменяется ли динамика исследуемого процесса при изменении внешних условий и патологии? Например, меняются ли периодические колебания концентраций гормонов роста при смене часового пояса (рис. 5, б);

– какой тип регуляции лучше описывает изменения концентрации веществ, наблюдаемые в различных экспериментах? Например, если известно, что регуляция филлотаксиса у растения осуществляется фитогормоном ауксином, который регулирует экспрессию своих транспортеров PIN, какой тип регуляции наиболее хорошо описывает формирование распределения ауксина в меристеме побега (рис. 5, в).

Основной целью математического моделирования является проверка гипотез об особенностях регуляции биологических процессов, а также генерация новых гипотез. Кроме того, математические модели позволяют решать практические задачи в биологии и медицине. Например, актуальной является задача оценки концентрации питательных веществ, оптимальных для устойчивого развития бактерий в ферментерах – аппаратах для промышленного выращивания бактерий.

Эксперимент *in silico*

По аналогии с экспериментами *in vivo* (в живом организме) и *in vitro* (в пробирке) биологические эксперименты, осуществленные на компьютере, в настоящее время стали называть *in silico* (компьютерный эксперимент). Рассмотрим основные этапы эксперимента *in silico* по математическому моделированию в биологии (рис. 6).

Первым этапом является сбор и анализ экспериментальных данных. Любая математическая модель в биологии должна основываться на доказанных в экспериментах фактах. В случае недостаточности экспериментальных данных и с целью упрощения расчета модели в эксперименте *in silico* могут быть приняты

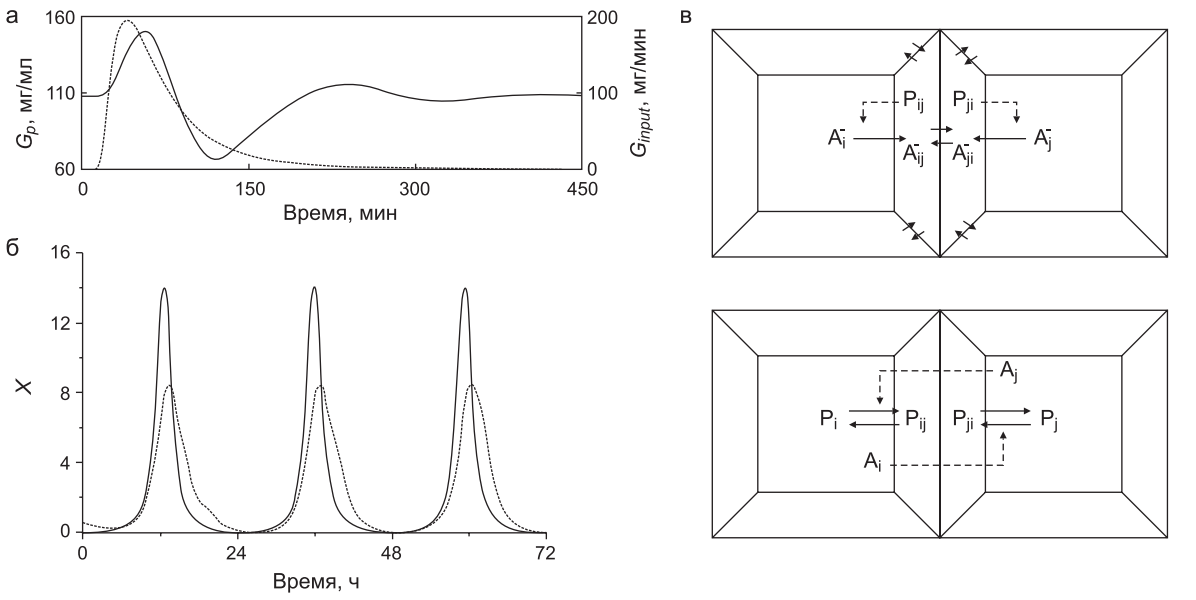

Рис. 5. Примеры исследований динамики в биологических системах, проведенных методами математического моделирования.

а – моделирование динамики изменения концентрации глюкозы в крови (черная кривая) после употребления ее в пищу (пунктирная линия) (адаптировано из: (Kang *et al*., 2012. P. 84–93)); б – моделирование динамики активности нейронов в суточных ритмах. Черная линия – зимнее время, пунктирная линия – летнее время (Bodenstein *et al*., 2012. P. 1633–1638); в – два различных механизма регуляции транспорта ауксина A через его транспортер P. В результате моделирования было доказано, что механизм на нижней схеме хорошо описывает распределение ауксина в меристеме побега (Jönsson *et al*., 2006. P. 1633–1638).

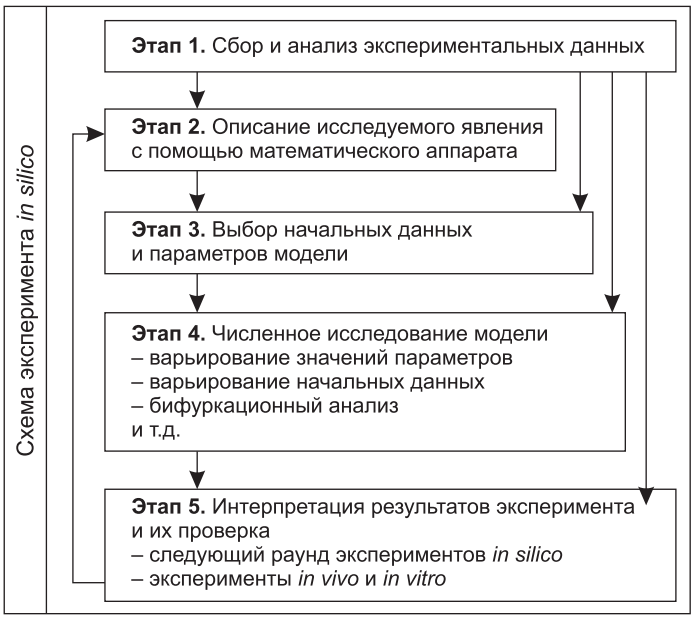

Рис. 6. Основные этапы эксперимента *in silico* по математическому моделированию биологических процессов.

некоторые допущения и упрощения. Их правомочность должна быть хорошо аргументирована. Например, для упрощения расчетов моделей клетки часто рассматривают в виде кубических ячеек. Это позволяет проще вычислять объем клетки и концентрации в ней веществ.

Вторым этапом эксперимента *in silico* является формализация исследуемого процесса в виде математических формул и уравнений. При этом учитываются основные физические законы, например закон сохранения массы. В некоторых областях математической биологии уже сложились правила описания биологических процессов в моделях, о некоторых из них будет сказано ниже. Ключевыми понятиями в математической модели являются переменные и параметры. Переменные – биологические объекты, динамика которых изучается в эксперименте, например транскрипционные факторы, биологически активные вещества и метаболиты. Переменными в математических моделях в биологии часто выступают концентрации веществ. Для их расчета в модели необходимо задать начальные данные – значения переменных в начальный момент расчета модели. Параметры – константы скоростей реакции, коэффициенты и другие величины. Например, часто параметрами выступают стационарные значения концентрации веществ, динамика которых в данной математической модели не интересует исследователей. Параметры могут быть размерными и безразмерными.

Третьим этапом эксперимента *in silico* является выбор начальных данных и параметров модели. Это наиболее важный этап, который во многом зависит от биолога, предоставляющего данные, необходимые для оценки параметров. Часто, используемые в модели параметры сложно оценить экспериментально, но существует несколько способов выхода из этой ситуации:

− упрощение (редукция) модели, как результат – уменьшение количества параметров;

− оценка значений параметров на основе косвенных данных. Часто для решения этой задачи создают дополнительные математические модели, позволяющие оценить необходимые параметры;

− оценка отрезков варьирования значений параметров и их подбор при численном анализе математической модели (model fitting). Неизвестных параметров не должно быть очень много, иначе возникнет проблема единственности решения исследуемой задачи. Для моделей с большим количеством неизвестных параметров, как правило, можно подобрать несколько наборов параметров, дающих качественно схожее решение модели.

Четвертый этап заключается в численном исследовании созданной модели с выбранными набором параметров и начальными данными. В зависимости от решаемой задачи в эксперименте *in silico* анализируют стационарное решение модели или изменение значений переменных во времени. Перечислим ниже основные методы исследования модели.

Во-первых, в случае если не все параметры модели определены, необходима процедура оценки неизвестных параметров (model fitting). Она заключается в последовательном расчете модели при различных наборах параметров, в которых неизвестный параметр варьируется в заданных пределах. Набор параметров, при котором результаты расчета наилучшим образом приближаются к экспериментальным данным, выбирается как стандартный. Остальные расчеты модели проводятся с этим стандартным набором параметров. Иногда исследователи могут выбрать несколько наборов параметров и проанализировать динамику модели с разными наборами.

Анализ модели при варьировании параметров является стандартной процедурой исследования модели. В качестве основы берется модель со стандартным набором параметров, в которой некоторые параметры последовательно изменяются от расчета к расчету. Такой эксперимент *in silico*, как правило, имитирует реальные изменения, происходящие в живых объектах. Результатом исследования являются оценка устойчивости системы, динамика изменения значений переменных, а также области варьирования параметров, при которых исследуемая система может (или не способна) функционировать. Интерпретация результатов этих расчетов является важной частью эксперимента *in silico*.

Целью численного анализа модели при варьировании начальных данных также является имитация реальных изменений *in silico*. Например, в модели метаболизма жиров уменьшение концентрации АТФ в начальный момент времени может имитировать голодание или стресс.

В некоторых экспериментах *in silico* также используют более сложные методы анализа математических моделей, такие как метод продолжения по параметру (Фадеев, 1995) или бифуркационный анализ модели.

Пятым, заключительным, этапом является проверка предложенных в результате анализа модели гипотез. Это может быть экспериментальная проверка. А также для проверки гипотез иногда проводят следующий раунд экспериментов *in silico*, который начинается с изменения исходной модели.

На последних двух этапах эксперимента *in silico* результаты численных экспериментов необходимо сравнить с имеющимися экспериментальными данными. Необходимо критически подойти к следующим вопросам. Адекватна ли постановка и формализация модели для решения данной задачи? Описывает ли модель весь набор экспериментальных данных, которые можно исследовать с помощью данной модели *in silico*? Дает ли модель новые предсказания по сравнению с имеющимися экспериментальными данными? Если ответ на все три вопроса утвердительный, то эксперимент *in silico* можно считать успешным.

Описание молекулярно-генетических процессов в математических моделях

Существует несколько уровней организации живого: молекулярно-генетический, клеточный, тканевой, организменный и популяционный. Для описания биологических процессов на каждом из этих уровней существуют свои методы и подходы. В данной главе мы рассмотрим лишь некоторые подходы к описанию молекулярно-генетических процессов.

Закон действующих масс. Предложенный еще в XIX в. закон действующих масс стал базовым для математического моделирования молекулярных процессов. Согласно этому закону, скорость ферментативной реакции пропорциональна вероятности взаимодействия фермента с субстратом. Поэтому для обратимой реакции взаимодействия двух веществ субстратов *S1* и *S2* с образованием двух молекул продукта P:

$$S1 + S2 \Leftrightarrow \mathrm{P}, \qquad (1)$$

Закон действующих масс – скорость ферментативной реакции пропорциональна вероятности взаимодействия фермента с субстратом.

скорость реакции v может быть записана как разность скоростей прямой v_+ и обратной v_- реакций:

$$v = v_+ - v_- = k_+ S1 \cdot S2 - k_- P, \qquad (2)$$

где k_+ и k_- – константы скорости прямой и обратной реакций соответственно.

Закон действующих масс является универсальным, поэтому его также можно применять для описания некоторых регуляторных или, напротив, нерегулируемых реакций. Например, скорость деградации вещества S можно описать как $v = k_d \cdot S$.

Закон действующих масс позволил развить целое направление в химии и биологии: кинетическое моделирование ферментативных реакций. Одной из ключевых моделей в этой области является модель Михаэлис–Ментен.

Закон сохранения массы. При описании молекулярно-генетических процессов ключевым является физический закон сохранения массы вещества. Крылатой стала фраза Михаила Ломоносова, обосновавшего этот закон: «Если где-то что-то убыло, значит, где-то что-то прибыло». Как соответствие с этим законом при описании любого молекулярно-генетического процесса следует учитывать, что концентрация вещества в определенный момент времени есть баланс между скоростью синтеза (притока) v_+ и скоростью деградации (оттока) v_- вещества. Изменение концентрации вещества P из реакции (1) во времени можно записать как:

$$\frac{dP}{dt} = v_+ - v_- = k_+ S1 \cdot S2 - k_- P. \qquad (3)$$

В применении закона сохранения массы к математическому описанию изменения концентрации белка X, синтез которого зависит от продукта реакции (1), уравнение будет выглядеть так:

$$\frac{dX}{dt} = f_s(P, t) - k_d \cdot X, \qquad (4)$$

где функция биосинтеза белка $f_s(P, t)$ может быть записана сложным образом на основе информации о механизме регуляции веществом P синтеза X, но с учетом закона действующих масс. Параметр k_d – константа деградации вещества X. В общем виде деградация может быть также нелинейным во времени процессом, описанным более сложной функцией.

Закон сохранения массы – концентрация вещества в определенный момент времени есть баланс между скоростью синтеза (притока) и скоростью деградации (оттока) этого вещества.

Уравнения (3) и (4) являются примером простой математической модели регуляции синтеза белка. Решение системы этих обыкновенных дифференциальных уравнений позволит наблюдать изменение концентрации вещества во времени; узнать стационарные концентрации P и X при заданных условиях реакции; исследовать чувствительность механизма синтеза белка от изменения концентрации вещества P.

Описание генетической регуляции. Математическое моделирование генетических процессов является совсем молодым направлением системной биологии, поэтому на данный момент теория генных сетей недостаточно хорошо проработана. Существует несколько методов описания генетической регуляции в математических моделях. Мы рассмотрим один из них, использующийся наиболее часто.

Рассмотрим упрощенную схему регуляции транскрипции гена X (рис. 9). Регуляция транскрипции этого гена обеспечивается двумя транскрипционными факторами – активатором A и репрессором P. Активатор облегчает посадку на промотор гена РНК-полимеразы (инициация транскрипции), репрессор же подавляет инициацию транскрипции. В случае если с промотором связываются и активатор и ингибитор, промотор остается не активным. С некоторой вероятностью инициация транскрипции может начаться и без активатора, но при отсутствии ингибитора.

Можно выписать все возможные состояния промотора гена X: 1) промотор без регуляторов; 2) промотор, связанный с активатором; 3) промотор, связанный с репрессором; 4) промотор, связанный с активатором и репрессором. Первое и второе состояния промотора способствуют инициации транскрипции, третье и четвертое, напротив, предотвращают ее. В соответствии с законом действующих масс можно описать активность промотора гена в целом (v_x), если в числитель дроби записать активирующие состояния промотора, а в знаменатель – все возможные состояния:

$$v_x = k_0 \frac{\delta_0 + \delta_1 A}{1 + \delta_2 A + \delta_3 P + \delta_4 A \cdot P}, \qquad (5)$$

где k_0 – константа скорости инициации транскрипции, δ_0 – коэффициент базальной (без регуляторов) экспрессии гена, δ_1 – коэффициент позитивного влияния A на активность промотора, δ_2 – коэффициент, отражающий эффективность связывания A с промотором, δ_3 – коэффициент, отражающий эффективность связывания P с промотором, δ_4 – коэффициент, отражающий вероятность нахождения обоих регуляторов на промоторе гена.

Уравнение (5) описывает только скорость экспрессии гена с механизмом регуляции, указанным на рис. 7. При описании регуляции другого рода функция v_x может сильно отличаться. Помимо наличия активаторов и репрессоров, учитывают также взаимное расположение сайтов в промоторе, мультимерность строения транскрипционных факторов, синергические взаимодействия между ними. Подробнее о математическом описании генетической регуляции можно прочитать в монографии «Системная компьютерная биология» (Колчанов и др., 2008).

ЗАКЛЮЧЕНИЕ

Системная биология – бурно развивающееся направление в биологии. Ее цель заключается в системном описании живых объектов на разных уровнях их организации, начиная с уровня молекул и заканчивая клетками и тканями. Ключевым при этом является всесторонний учет взаимодействий между биологическими объектами. Эти взаимодействия могут быть описаны в виде графов. В системной биологии широко используется компьютерное моделирование для интерпретации экспериментальных данных. Это позволяет на основе широкомасштабных экспериментов в области протеомики, экспрессии генов, метаболомики уточнять характер взаимодействия молекулярных объектов и совершенствовать модели живых систем.

Системная биология изучает взаимоотношения между молекулярными объектами в живых системах. Центральным в системной биологии является то, что взаимодействие нескольких агентов (таких как белки в генной

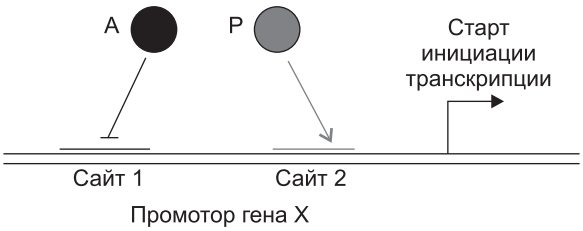

Рис. 9. Пример механизма регуляции транскрипции гена, скорость протекания которой описана в уравнении (5).

сети) может демонстрировать новые состояния системы (клетки), возникающие как следствие их взаимодействий. В последние годы интерес к системному подходу в биологии был вызван прорывом в технологиях секвенирования и, как результат, расшифровкой геномов, траскриптомов и протеомов человека и других организмов. Наличие мощных вычислительных средств и широкополосных Интернет-соединений также значительно облегчило анализ и распространение огромных наборов данных, на которых в значительной степени основаны методы системной биологии.

Работа частично поддержана грантами РФФИ-11-04-33112, грантом фонда «Династия» для молодых биологов и Министерством образования и науки (договор №14.B25.31.0033).

ЛИТЕРАТУРА

Колчанов Н.А., Гончаров С.С., Лихошвай В.А., Иванисенко В.А. Системная компьютерная биология. Новосибирск: Изд-во СО РАН, 2008.

Фадеев С.И. Минимизация свободной энергии Гиббса методом продолжения решения по параметру // Сплайны и их приложения. Вычислительные системы. 154. Сб. науч. тр. Новосибирск, 1995. С. 92–110.

Barabási A.L. Scale-free networks: a decade and beyond //Science. 2009. V. 325. No. 5939. P. 412–413.

Bodenstein C., Gosak M., Schuster S. *et al*. Modeling the seasonal adaptation of circadian clocks by changes in the network structure of the suprachiasmatic nucleus // PLoS Computat. Biol. 2012. V. 8. No. 9. e1002697.

Chung Y., Maharjan P.M., Lee O. *et al*. Auxin stimulates DWARF4 expression and brassinosteroid biosynthesis in Arabidopsis // Plant J. 2011. V. 66. No. 4. P. 564–578.

Huang D.W., Sherman B.T., Lempicki R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources // Nature Prot. 2008. V. 4. P. 44–57.

Huang D.W., Sherman, B.T., Lempicki R.A. Bioinformatics enrichment tools: paths toward the comprehensive func-

tional analysis of large gene lists // Nucl. Acids Res. 2009. V. 37. P. 1–13.

Jeong H., Mason S.P., Barabási A.L., Oltvai Z.N. Lethality and centrality in protein networks // Nature. 2001. V. 411. No. 6833. P. 41–42.

Jönsson H., Heisler M.G., Shapiro B.E. *et al*. An auxin-driven polarized transport model for phyllotaxis // Proc. Natl Acad. Sci. USA. 2006. V. 103. No. 5. P. 1633–1638.

Kang H., Han K., Choi M. Mathematical model for glucose regulation in the whole-body system // Islets. 2012. V. 4. No. 2. P. 84–93.

Karp P.D. An ontology for biological function based on molecular interactions // Bioinformatics. 2000. V. 16. No. 3. P. 269–285.

Kitano H. Foundations of Systems Biology. Cambridge, MA: MIT Press, 2001.