ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

ОСНОВАН В 1997 г.

Том 16
4/1
Октябрь 2012

VAVILOV JOURNAL OF GENETICS AND BREEDING

FOUNDED IN 1997

Vol. 16
4/1
October 2012

«Вавиловский журнал генетики и селекции» / «Vavilov Journal of Genetics and Breeding» до 2011 г. выходил под названием «Информационный вестник ВОГиС» / «The Herald of Vavilov Society for Geneticists and Breeding Scientists».

«Вавиловский журнал генетики и селекции» включен ВАК Минобрнауки России в Перечень ведущих рецензируемых научных журналов и изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени доктора и кандидата наук (по биологическим наукам).
(Редакция 17 июня 2011 г.: http://vak.ed.gov.ru)

«Вавиловский журнал генетики и селекции» включен в федеральный почтовый Объединенный каталог «ПРЕССА РОССИИ».
Персональный подписной индекс № 42153.

Адрес редакции:

«Вавиловский журнал генетики и селекции»,
ИЦиГ СО РАН,
Проспект Академика Лаврентьева, 10,
Новосибирск, 630090

Факс: (383) 3331278
e-mail: vavilov_journal@bionet.nsc.ru

Ответственный секретарь редакции:
С.В. Зубова,
тел. 363-4922 *1351

Регистрационное свидетельство ПИ № ФС77-45870 выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций 20 июля 2011 г.

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

ОСНОВАН В 1997 г.

Том 16
4/1
Октябрь 2012

Содержание

VAVILOV JOURNAL OF GENETICS AND BREEDING

FOUNDED IN 1997

Vol. 16
4/1
October 2012

Content

ПРЕДИСЛОВИЕ

Бурное развитие экспериментальных технологий в таких областях современной биологии, как геномика, транскриптомика, протеомика, молекулярная и клеточная биология, генетика, физиология, биомедицина, биотехнология и других, привело к появлению огромных объемов информации о различных аспектах организации и функционирования живых систем.

Выявление общих принципов, стоящих за единичными фактами, изучение тех или иных отношений между индивидуальными феноменами, систематизация и интерпретация беспрецедентно огромных объемов данных, генерируемых современной экспериментальной биологией, невозможны без привлечения современных информационных технологий, эффективных методов компьютерного анализа данных и математического моделирования биологических систем и процессов на различных уровнях организации живой материи: от молекулярно-генетического до экосистемного и биосферного. В связи с этим критически возрастает роль таких научных направлений, как биоинформатика и системная компьютерная биология.

К числу ключевых задач биоинформатики относятся:

– создание баз данных и баз знаний;

– интеграция геномной, транскриптомной, протеомной, метаболомной информации, получаемой из гетерогенных распределенных источников экспериментальной данных;

– автоматическое извлечение знаний из текстов научных публикаций и фактографических баз данных (text and data mining);

– разработка алгоритмов и программного обеспечения для анализа биологических экспериментальных данных;

– разработка программного обеспечения для суперкомпьютерных вычислений;

– создание компьютерных систем конвейерного анализа сложных данных.

Биоинформатика имеет важнейшее значение для ассемблирования геномов и метагеномов, их функциональной аннотации, предсказания генов и их регуляторных районов, реконструкции пространственной структуры белков, предсказания их функции и др.

Следует подчеркнуть, что ни один, даже самый совершенный, экспериментальный подход сам по себе не может дать целостного представления об изучаемых биологических системах. В ответ на этот вызов возникло новое междисциплинарное направление исследований – системная компьютерная биология (СКБ). Ее задача – получение целостного комплексного представления о структурной организации и механизмах функционирования живых систем. Важнейшей составляющей исследовательского процесса в биоинформатике и системной компьютерной биологии стал вычислительный эксперимент. Системная компьютерная биология сформировалась как наука, когда она приобрела способность к количественному предсказанию и планированию эксперимента (рис.).

К числу важнейших задач СКБ относятся: реконструкция генных сетей на основе экспериментальных данных геномики, протеомики, транскриптомики и других экспериментальных технологий современной молекулярной и клеточной биологии; математическое моделирование генных сетей и молекулярно-генетических систем и процессов, контролирующих формирование фенотипических характеристик организмов на биохимическом, клеточном, физиологическом, морфологическом уровнях организации живых систем; математическое моделирование процессов морфогенеза, а также компьютерный анализ и математическое моделирование процессов молекулярной эволюции сложных молекулярно-генетических систем. Системная компьютерная биология имеет важнейшее значение для решения широкого круга прикладных задач в области биомедицины (моделирование механизмов возникновения патологий) и биотехнологии (моделирование и оптимизация метаболических путей при создании бактериальных штаммов-суперпродуцентов, планирование экспериментов по созданию генетически

Рис. Цикл исследований в системной компьютерной биологии.

модифицированных растений и животных с заданными целевыми свойствами).

В настоящем выпуске журнала представлены результаты исследований, проводимых в СО РАН по различным направлениям биоинформатики и системной компьютерной биологии, включая разработку программно-информационных систем в области компьютерной геномики, компьютерной транскриптомики, исследования механизмов регуляции транскрипции и трансляции, компьютерной протеомики, реконструкции и моделирования генных сетей, моделирования пространственно распределенных процессов морфологии, моделирования эволюции бактериальных сообществ, а также разработку специализированных Web-порталов для поддержки экспериментальных исследований и программных средств для управления сценариями конвейерной обработки данных и проведения вычислительных экспериментов.

Н.А. Колчанов
Приглашенный редактор Н.Л. Подколодный

УДК 577.21:004.02:004.94

ICGenonics: ПРОГРАММНЫЙ КОМПЛЕКС АНАЛИЗА СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ГЕНОМИКИ

© 2012 г. **Ю.Л. Орлов[1, 2], А.О. Брагин[1], И.В. Медведева[1], К.В. Гунбин[1], П.С. Деменков[1], О.В. Вишневский[1], В.Г. Левицкий[1], Д.Ю. Ощепков[1], Н.Л. Подколодный[1], Д.А. Афонников[1, 2], И. Гроссе[3], Н.А. Колчанов[1, 2, 4]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: orlov@bionet.nsc.ru;
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия;
[3] Институт информатики, Университет Мартина Лютера, Халле, Германия;
[4] НИЦ «Курчатовский институт», Москва, Россия

Поступила в редакцию 10 июля 2012 г. Принята к публикации 10 августа 2012 г.

Экспериментальный образец программного комплекса анализа символьных последовательностей геномики (ЭОПК АСПГ) ICGenomics предназначен для хранения, передачи, обработки и анализа данных о символьных последовательностях, полученных в рамках теоретической и прикладной геномики с целью повышения качества вычислительной обработки биологических данных, используемых в биомедицине и биотехнологии. В комплексе реализованы новые оригинальные методы обработки первичных данных высокопроизводительного секвенирования, в том числе данных ChIP-seq, предсказания регуляторных участков генов в нуклеотидных последовательностях, модели расположения нуклеосом, структурно-функциональной аннотации белков, включая их аллергенные свойства и особенности эволюции. Рассмотрено применение комплекса к анализу последовательностей паразитического червя *O. felineus*, данным ChIP-seq по профилям связывания транскрипционных факторов в геномах мыши и человека. Комплекс доступен по адресу: http://www-bionet.sscc.ru/icgenomics.

Ключевые слова: геномика, программный комплекс, высокопроизводительное секвенирование, последовательности ДНК, анализ данных, ChIP-seq.

ВВЕДЕНИЕ

Программный комплекс ICGenomics предназначен для компьютерной поддержки исследований в геномике, молекулярной биологии, биотехнологии и биомедицине. Основное назначение – функциональная аннотация геномных последовательностей, получаемых в результате массового высокопроизводительного секвенирования на уровне нуклеотидных и аминокислотных последовательностей. Рабочее название – экспериментальный образец программного комплекса анализа символьных последовательностей геномики (ЭОПК АСПГ).

Важная технологическая проблема обработки и анализа данных высокопроизводительного геномного секвенирования требует разработки специализированных компьютерных средств. Развитие новых экспериментальных методов геномики, прежде всего, секвенирования, привело к стремительному росту объемов экспериментальных данных, «информационному взрыву».

Основная задача компьютерного анализа геномных данных состоит в их функциональной аннотации, интеграции результатов с молекулярно-биологическими информационными ресурсами. В связи с этим большую актуальность приобретает разработка информационно-компьютерных технологий автоматического анализа и функциональной аннотации геномных последовательностей. Для решения

задачи разработан ряд программ для извлечения и интеграции данных, а также визуального представления накопленной информации в форме геномных профилей, представленных на серверах крупнейших международных научных центров NCBI (http://www.ncbi.nlm.nih.gov/), UCSC Genome Browser (http://genome.ucsc.edu/), EBI (http://www.ebi.ac.uk/).

Важнейшим объектом анализа являются молекулярно-генетические системы, координирующие функцию геномов, генов, РНК, белков, генных и метаболических путей на различных иерархических уровнях жизни: клеточном, тканевом, органном, организменном, популяционном. Основным источником данных являются нуклеотидные последовательности, получаемые в результате массовых экспериментов высокопроизводительного секвенирования (Ivanisenko *et al.*, 2012). Несмотря на доступность компьютерных программ биоинформатики, в связи со все возрастающими объемами данных остается ряд направлений, важных для более детальной разработки. Можно выделить следующие направления исследования геномных последовательностей:

1. Разработка конвейерного подхода (pipeline) для первичной обработки, процессинга, картирования на референсный геном последовательностей, полученных в ходе масштабного параллельного секвенирования.

2. Функциональная аннотация геномных последовательностей (генома человека и модельных организмов) с целью разметки регуляторных районов, сайтов формирования нуклеосом и определения структуры хроматина. Сюда входят аннотация потенциальных микроРНК и анализ промоторных последовательностей генов.

3. Разработка программ для разметки функциональных сайтов белков, определения свойств белковых фрагментов, кодируемых в нуклеотидных последовательностях, оценки потенциальной аллергенности кодируемых белков с использованием оригинальных баз данных и методов.

4. Сравнение функциональных свойств вновь секвенированных генов различных организмов. Исследование адаптивного режима эволюции на уровне отдельных семейств генов и на геномном уровне.

Решение перечисленных задач необходимо для обеспечения технической поддержки геномных исследований. Соответствующие технические средства реализованы в разработанном программном комплексе. Особое внимание было уделено оригинальным методам, не повторяющим стандартные алгоритмы для уже достаточно рутинных задач, таких, как выделение кодирующей последовательности или предсказание сайтов связывания транскрипционных факторов (ССТФ) только по нуклеотидной последовательности (с помощью весовых матриц), стандартные решения для которых представлены на серверах NCBI, UCSC, EBI.

Конкретная задача компьютерного анализа геномных последовательностей включала реализацию следующих независимых конкретных процедур, объединяемых общими типами данных:

– Обработка последовательностей ДНК из геномных фрагментов, полученных с помощью установок геномного секвенирования нового поколения.

– Функциональная аннотация геномных нуклеотидных последовательностей с возможностями аннотации нуклеосом, поиска экзонов, поиска промоторов генов миРНК.

– Предсказание аллергенности белков по их структурным и функциональным свойствам на основе метода функциональной аннотации пространственных структур белков, предсказание функциональных сайтов в пространственных структурах белков и предсказание специфической активности белков по их первичной и пространственной структуре.

– Реализованная в виде конвейера обработки данных процедура анализа режимов эволюции белок-кодирующих генов с возможностями реконструкции эволюционной истории белков на основе предсказания ортологов в секвенированных геномах, филогенетического анализа, а также изучения режимов отбора.

Программный комплекс ICGenomics (http://www-bionet.sscc.ru/icgenomics) был реализован и протестирован на вычислительном оборудовании ЦКП «Биоинформатика» СО РАН.

МАТЕРИАЛЫ И МЕТОДЫ

Программный комплекс ICGenomics позволяет выполнять следующие логически различные функции:

– процессинг (обработку) протяженных последовательностей нуклеотидов из данных секвенирования, полученных с помощью установок секвенирования нового поколения, в том числе: процессинг данных секвенирования платформ 454 и Illumina, процессинг данных секвенирования платформы SOLiD и обработку полногеномных профилей ChIP-seq, включая выделение пиков и предсказание ССТФ;

– аннотацию геномных нуклеотидных последовательностей, включая разметку положения нуклеосом на основе вейвлет-преобразования полногеномных профилей предсказания и распознавание сайтов формирования нуклеосом с помощью данных полногеномного секвенирования линкерной ДНК; поиск экзонов во вновь секвенированных последовательностях; поиск промоторов генов миРНК в нуклеотидных последовательностях на основе специфичных структурных мотивов;

– предсказание аллергенности белков по их структурным и функциональным свойствам на основе метода функциональной аннотации пространственных структур белков, в том числе предсказания функциональных сайтов в пространственных структурах белков;

– исследование режимов эволюции белок-кодирующих генов, включая реконструкцию эволюционной истории белков на основе предсказания ортологов в секвенированных геномах, филогенетический анализ и исследование режимов эволюционного отбора.

Каждая из перечисленных выше функций реализована в соответствующем программном компоненте (рис. 1).

Программный комплекс состоит из модуля управления (программной компоненты ICGenomics-Web и управляющей программы ICGenomics-start) и 4 программных компонент: ICGenomics-Processing, ICGenomics-Genome Annotation, ICGenomics-Allergen и ICGenomics-Evolution (рис. 1).

Общий интерфейс представлен на рис. 2. Рассмотрим компоненты более подробно:

1. ICGenomics-Processing – программный компонент, осуществляющий обработку последовательностей ДНК из фрагментов, полученных с помощью установок геномного секвенирования нового поколения, обладающий функционалом процессинга исходных («сырых») данных секвенирования, обработки полногеномных профилей ChIP-seq, выделения пиков и предсказания ССТФ.

2. ICGenomics-GenomeAnnotation – программный компонент функциональной аннотации геномных нуклеотидных последовательностей, обладающий возможностями:

– функциональной аннотации нуклеосом;
– поиска экзонов;
– поиска промоторов генов миРНК (рис. 3).

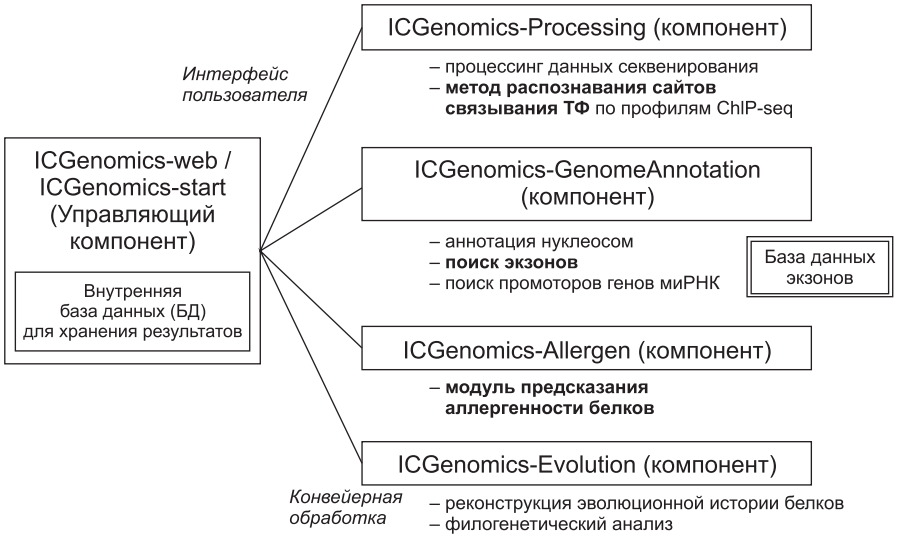

Рис. 1. Структура программного комплекса ICGenomics.

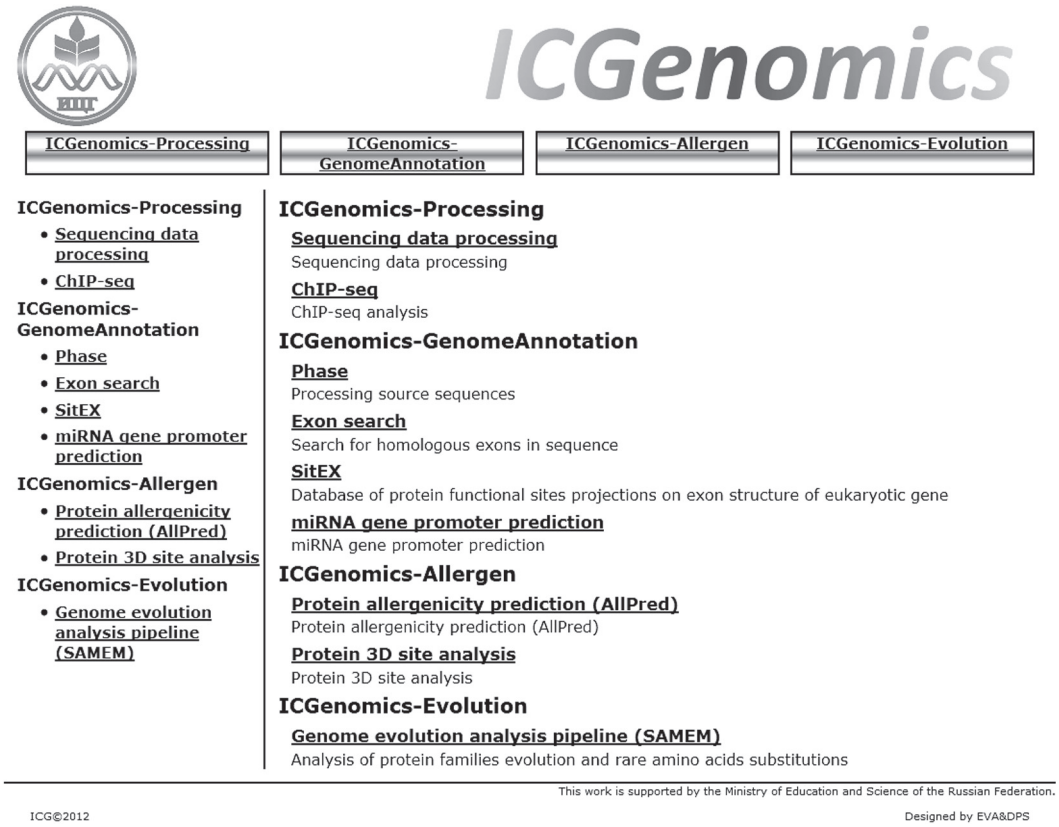

Рис. 2. Пример интерфейса управляющего модуля, содержащего функциональные компоненты.

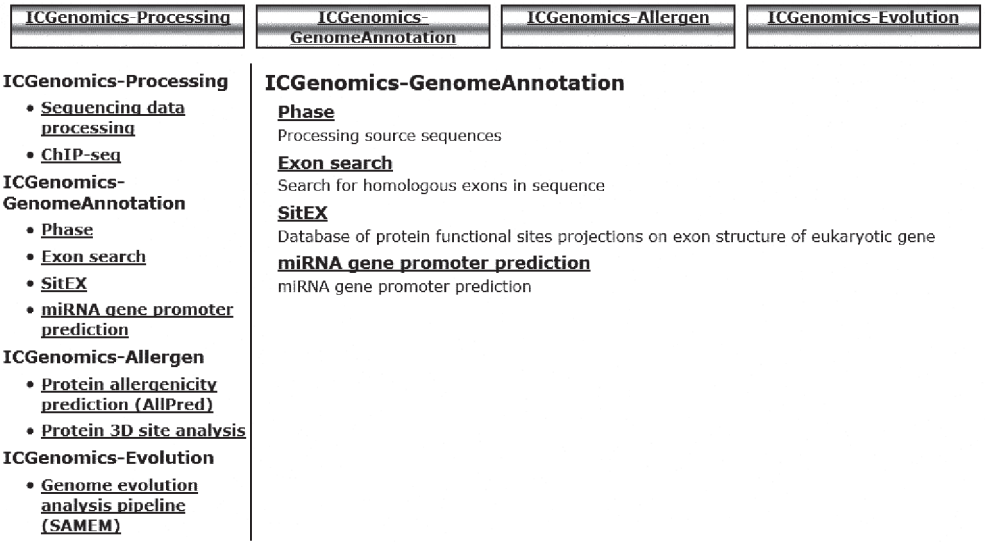

Рис. 3. Пример интерфейса ICGenomics-GenomeAnnotation.

3. ICGenomics-Allergen – программный компонент предсказания аллергенности белков по их структурным и функциональным свойствам.

4. ICGenomics-Evolution – программный компонент исследования режимов эволюции белок-кодирующих генов, обладающий функционалом: реконструкции эволюционной истории белков на основе предсказаний ортологов в секвенированных геномах; филогенетического анализа и исследования режимов отбора. Компонент реализован в виде конвейера обработки данных.

Входными данными для системы служат файлы нуклеотидных и аминокислотных последовательностей в формате FASTA, а также данные секвенирования в форматах платформ секвенирования Illumina, SOLiD. Возможно использование форматов геномных профилей bed (геномные координаты), wig (численный профиль). В комплексе используются базы данных SiteEx (Medvedeva *et al*., 2012), и PDBSite (Ivanisenko *et al*., 2005), содержащие скомпилированную ранее информацию об экзонах и пространственных сайтах белков.

Компонент ICGenomics-Processing включает в себя модули процессинга данных, в том числе конвертации форматов и фильтрации сигнала секвенирования ДНК, процессинга данных секвенирования платформ 454 и Illumina (исходные форматы fastq, qseq), процессинга данных секвенирования платформы SOLiD (в цветовой кодировке color-space – исходный формат csfasta) и конвейерной обработки задач картирования данных SOLiD. Этот компонент (модуль ChIP-seq pipeline) также выполняет обработку полногеномных профилей ChIP-seq, выделение пиков профиля и предсказание ССТФ в геноме.

Типичные задачи, которые решаются на этапе предобработки – преобразование данных, полученных в результате эксперимента, в стандартные форматы; анализ качества последовательностей и фильтрация по качеству; подготовка результата по проведенным операциям. Метод распознавания реализован в программе ChIP-seq pipeline и предназначен для конвейерной обработки выходных данных эксперимента по массовому секвенированию функциональных сайтов. Массовость означает полногеномный характер анализа и большие объемы данных. В качестве функциональных сайтов исследуются ССТФ различных типов. Программный комплекс преследует две основные задачи: а) обработку данных и привязку их к геномным картам; б) верификацию обнаруженных геномных локусов с помощью различных биоинформатических средств (программ распознавания ССТФ). Подобный подход позволяет: а) исключить из рассмотрения ошибки и артефакты, присутствующие в данных эксперимента ChIP-seq; б) правильно настроить параметры на различных этапах обработки данных (картирование на полный геном, выбор минимального числа прочтений сайта в геномном локусе и т. д.); в) получить в итоге исчерпывающий список генов-мишеней исследуемого ССТФ для полного генома (Lee *et al*., 2011).

В качестве программы для картирования прочтений (первичных данных эксперимента ChIP-seq) использовался рекомендованный производителем SOLiD™ BioScope™ Software с настройками по умолчанию. Далее в соответствии с рекомендациями производителя с помощью этого же программного обеспечения (SOLiD™ BioScope™ Software) производилась конвертация выходного формата файла с результатами картирования (формат «.ma» в формат «.bam») для последующей подачи на вход программы MACS (Zhang *et al*., 2008). MACS предназначена для проведения процедуры поиска пиков ChIP-seq (ChIP-seq peak calling) и является одной из самых широко используемых программ, кроме того, обладает наибольшей точностью в определении локализации сайта связывания (Malone *et al*., 2011).

Результатом работы программы является полногеномный профиль в формате wig, который представляет собой список пар «позиция»–«покрытие». «Позиция» – хромосомная локализация, включает в себя номер хромосомы и собственно позицию от начала хромосомы. «Покрытие» – число прочтений – это число зафиксированных взаимодействий исследуемого белка (ТФ) с ДНК в рассматриваемой хромосомной локализации. Далее могут быть определены нуклеотидные последовательности, содержащие ССТФ, проанализированы частоты олигонуклеотидов с помощью разработанных ранее программ (Putta *et al*., 2011).

Используемые для секвенирования фрагментов ДНК технологии компаний Illumina и ABI SOLiD характеризуются особенностями, связанными с проведением экспериментальных процедур, что отражено в форматах входных данных используемых ICGenomics. Технология компании Illumina (Solexa) (http://www.illumina.com) использует оптическое сканирование флюоресценции меченых нуклеотидов в клонированных колониях молекул ДНК на твердой поверхности, в то время как технология секвенирования ABI (Applied Biosystems) SOLiD (Sequencing by Oligonucleotide Ligation) использует лигирование и, соответственно, кодировку

по двум нуклеотидам. Для процессинга данных секвенирования реализована возможность использования следующих форматов геномных данных: FASTA, fastq, clustal.

Используя тот же формат FASTA, компонент ICGenomics-GenomeAnnotation функциональной аннотации геномных нуклеотидных последовательностей (рис. 3) решает задачи:

– функциональной аннотации нуклеосом (включая применение вейвлет-преобразования для анализа полногеномных профилей предсказания сайтов формирования нуклеосом и распознавания сайтов формирования нуклеосом с помощью данных полногеномного секвенирования линкерной ДНК);

– поиска экзонов во вновь секвенированных последовательностях для более подробной аннотации генов и кодируемых ими белков, а также входящих в их состав доменов на основе базы данных последовательностей экзонов и структур полипептидов, кодируемых единственным экзоном;

– поиска промоторов генов миРНК в нуклеотидных последовательностях на основе специфичных структурных мотивов.

Вызов отдельных модулей выполняется из общего интерфейса пошагово. На рис. 4 приведен пример вызова программы Phase предсказания положения нуклеосом в нуклеотидной последовательности. Программа Phase успешно применялась для анализа генома дрожжей и сравнения эффективности транскрипции генов в зависимости от предсказанной локализации нуклеосом в промоторах генов (Matushkin *et al*., 2012).

Таким же образом могут вызываться модуль анализа экзонов, в том числе база данных SiteEx (Medvedeva *et al*., 2012) и модуль предсказания промоторов генов миРНК (Vishnevsky *et al*., 2010).

Программный компонент ICGenomics-Allergen предсказания аллергенности белков по их структурным и функциональным свойствам выполняет предсказание аллергенности белков (пептидов) с использованием конформационных пептидов (Bragin *et al*., 2012). Кроме того, модуль может передавать данные функциональных сайтов в пространственных структурах белков (рис. 5). Программа вычисляет значения аллергенности по заданной последовательности

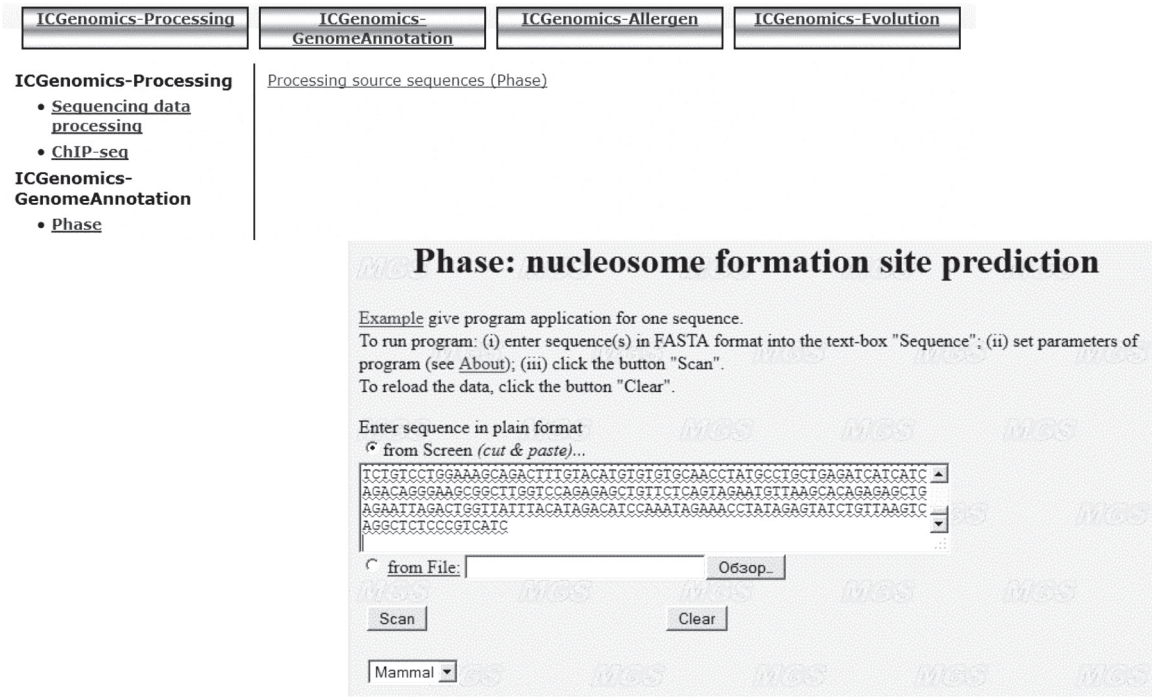

Рис. 4. Пример вызова модуля предсказания положения нуклеосом из компонента ICGenomics-GenomeAnnotation (верхняя панель) и интерфейс вызванной программы Phase предсказания положения нуклеосом по нуклеотидной последовательности (нижняя панель).

пептида. Результатом работы является числовое значение аллергенности и текстовое описание. Те же последовательности могут быть переданы на анализ гомологии с последовательностями экзонов в соответствующем модуле и на сравнительный анализ семейства белков, приводящих к появлению свойств аллергенности.

Точность предсказания аллергенности белков разработанным модулем сравнивалась с точностью предсказания стандартных программ (Bragin *et al*., 2012). Точность метода была оценена на выборке белков-аллергенов из работы H.C. Muh и соавт. (Muh *et al*., 2009), создавших программу AllerHunter (http://tiger.dbs.nus.edu.sg/AllerHunter/). Использование одной только программы BLAST (белок считался аллергеном, если значение E-value сходства его последовательности с известными аллергенами было ниже 10^{-21}) позволяет точно предсказать аллергенность только у 84 % белков. В то время как метод, применяющий поиск гомологов при помощи BLAST и поиск пептидов в анализируемых белках, правильно предсказывает 92 % белков-аллергенов из этой же выборки.

Программный компонент ICGenomics-Evolution исследования режимов эволюции белок-кодирующих генов выполняет задачи реконструкции эволюционной истории белков на основе предсказания ортологов в секвенированных геномах и филогенетического анализа и исследования режимов отбора. Компонент реализован в виде конвейера обработки данных (рис. 6).

Методы анализа режимов эволюции, входящие в данный программный компонент, были успешно использованы в работах Gunbin с соавт. (2010, 2011); Гунбин и др. (2011).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Разработан комплекс ICGenomics, использующий ряд уникальных модулей. Программа позволяет выполнять следующие функции обработки и анализа геномных последовательностей:

– процессинг (обработку) протяженных последовательностей нуклеотидов из данных секвенирования, полученных с помощью установок секвенирования нового поколения, в том числе обработку полногеномных профилей ChIP-seq;

– аннотацию геномных нуклеотидных последовательностей, включая: разметку положения нуклеосом на основе вейвлет-преобразования полногеномных профилей предсказания, сайтов формирования нуклеосом; поиск экзонов во вновь секвенированных последовательностях; поиск промоторов генов миРНК в нуклеотидных последовательностях;

– предсказание аллергенности белков по их структурным и функциональным свойствам;

– исследование режимов эволюции белок-кодирующих генов, включая реконструкцию эволюционной истории белков на основе на предсказания ортологов в секвенированных геномах, филогенетический анализ и исследование режимов эволюционного отбора.

Реализованные в проекте авторские методы уникальны, что подтверждено регистрационными свидетельствами на программы, входящие в компоненты предсказания аллергенности и

Рис. 5. Интерфейс программы ICGenomics-Allergen предсказания аллергенности белков.

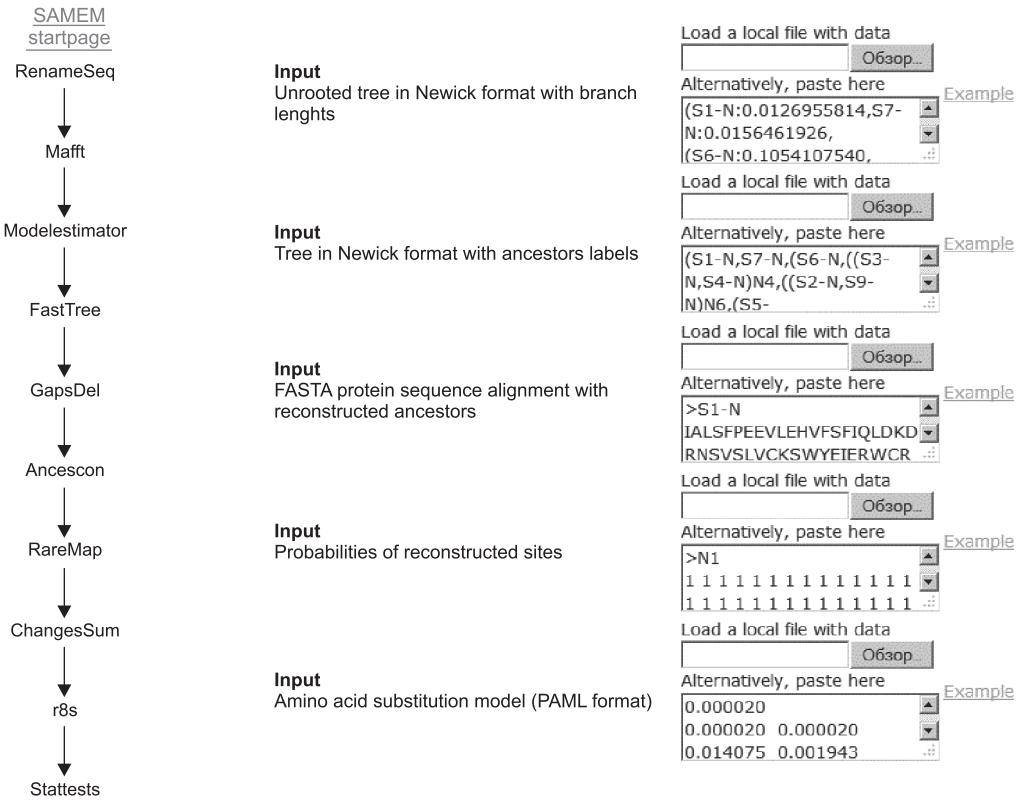

Рис. 6. Схема конвейера (левая панель) и выбор основных параметров (правая панель) в модуле исследования режимов эволюции ICGenomics-Evolution (SAMEM).

анализа режимов эволюции белков. Комплекс применялся к анализу геномных последовательностей паразитического червя *O. felineus* и к данным ChIP-seq по профилям связывания транскрипционных факторов в геномах мыши и человека (для фактора FoxA) (Левицкий и др., 2011).

Были исследованы три образца ткани *O. felineus* (стадии: марита без яиц, марита с яйцами, метацеркарий), а также препараты тканей *O. viverini* и *C. sinensis* на стадии марит. Картирование осуществляли на геном паразитического плоского червя шистосомы (*Schistosoma japonicum*). Шистосома – паразитический червь, который поражает кровеносную систему организма. Это ближайший родственный вид, геномная последовательность которого расшифрована практически полностью.

Для генома *S. japonicum* ранее была проведена функциональная аннотация генома и идентифицированы 55 последовательностей микроРНК. Эти гены принимают участие в регуляции стадий развития организма червя. Анализ локализации нуклеотидных фрагментов трех организмов позволил нам установить, что из этих 55 микроРНК 17 представлены и в геномах *O. felineus*, *O. viverini* и *C. sinensis*. При этом число картированных последовательностей для этих генов зависит как от стадии развития организма, так и от вида.

В ЭОПК АСПГ использовались разработанные в ИЦиГ СО РАН методы предсказания аллергенности белков по аминокислотным последовательностям (конформационным пептидам), предсказания позиций нуклеосом, предсказания сайтов. Конструктивными характеристиками разработанных методов являются возможности обрабатывать большие объемы данных секвенирования и возможность обмена данными в FASTA формате.

БЛАГОДАРНОСТИ

Авторы выражают благодарность В.А. Иванисенко и М.П. Пономаренко за научную дискуссию по данному проекту.

Разработка программного комплекса поддержана госконтрактом Минобрнауки РФ № 07.514.11.4003. Тестирование выполнялось на суперкомпьютерном кластере ССКЦ СО РАН, ЦКП «Биоинформатика».

ЛИТЕРАТУРА

Гунбин К.В., Суслов В.В., Афонников Д.А. Генетическая основа макроэволюционных преобразований: исследование режимов молекулярной эволюции ортологичных белков позвоночных и беспозвоночных // Тр. Междунар. конф. «Современные проблемы математики, информатики и биоинформатики», посвященной 100-летию со дня рождения чл.-корр. А.А.Ляпунова. 11–14 октября 2011 г. Новосибирск, Россия. 2011. ПП. 4.7. С. 52–53.

Левицкий В.Г., Ощепков Д.Ю., Ершов Н.И. и др. Разработка методов распознавания сайтов связывания транскрипционных факторов FoxA, их экспериментальная верификация и использование для анализа данных массовой иммунопреципитации хроматина // Докл. АН. 2011. Т. 436. № 3. С. 417–421.

Bragin A.O., Demenkov P.S., Kolchanov N.A., Ivanisenko V.A. Accuracy of protein allergenicity prediction can be improved by taking into account data on allergenic protein discontinuous peptides // J. Biomol. Struct. Dyn. 2012. Jul. 18. [Epub ahead of print]

Gunbin K.V., Genaev M. A., Afonnikov D. A., Kolchanov N.A. A computer system for the analysis of molecular evolution modes of protein-encoding genes (SAMEM): The relationship between molecular evolution and phenotypic traits // Mosc. Univ. Biol. Sci. Bull. 2010. V. 65. No. 4. P. 142–144.

Gunbin K.V., Suslov V.V., Turnaev I.I. *et al*. Molecular evolution of cyclin proteins in animals and fungi // BMC Evol. Biol. 2011. V. 11. P. 224.

Ivanisenko V.A., Demenkov P.S., Pintus S.S. *et al.* Computer analysis of metagenomic data-prediction of quantitative value of specific activity of proteins // Dokl. Biochem. Biophys. 2012. V. 443. P. 76–80.

Ivanisenko V.A., Pintus S.S., Grigorovich D.A., Kolchanov N.A. PDBSite: a database of the 3D structure of protein functional sites // Nucl. Acids Res. 2005. V. 33. Database, P. 183–187.

Lee K.L., Lim S.K., Orlov Y.L. *et al*. Graded Nodal/Activin signaling titrates conversion of quantitative phospho-Smad2 levels into qualitative embryonic stem cell fate decisions // PLoS Genet. 2011. V. 7. No. 6. e1002130.

Malone B.M., Tan F., Bridges S.M., Peng Z. Comparison of four ChIP-Seq analytical algorithms using rice endosperm H3K27 trimethylation profiling data // PLoS One. 2011. V. 6. No. 9. e25260.

Matushkin Y.G., Levitsky V.G., Orlov Y.L. *et al*. Translation efficiency in yeasts correlates with nucleosome formation in promoters // J. Biomol. Struct. Dyn. 2012. Jul. 18. [Epub ahead of print].

Medvedeva I., Demenkov P., Kolchanov N., Ivanisenko V. SitEx: a computer system for analysis of projections of protein functional sites on eukaryotic genes // Nucl. Acids Res. 2012. V. 40 (Database issue). P. 278–83.

Muh H.C., Tong J.C., Tammi M.T. AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins // PLoS One. 2009. V. 4. No. 6. e5861.

Putta P., Orlov Yu.L., Podkolodnyy N.L., Mitra C.K. Relatively conserved common short sequences in transcription factor binding sites and miRNA // Вавилов. журн. генет. и селекции. 2011. Т. 15. № 4. С. 750–756.

Vishnevsky O.V., Gunbin K.V., Bocharnikov A.V., Berezikov E.V. Analysis of degenerate motifs in the promoters of miRNA genes expressed in different mammalian tissues // Mosc. Univ. Biol. Sci. Bull. 2010. V. 65. No. 4. P. 193–195.

Zhang Y., Liu T., Meyer C.A. *et al*. Model-based Analysis of ChIP-Seq (MACS) // Genome Biol. 2008. V. 9. No. 9. R137.

ICGenomics: A PROGRAM COMPLEX FOR ANALYSIS OF SYMBOL SEQUENCES IN GENOMICS

Y.L. Orlov[1, 2], A.O. Bragin[1], I.V. Medvedeva[1], K.V. Gunbin[1], P.S. Demenkov[1], O.V. Vishnevsky[1], V.G. Levitsky[1], D.Y. Oshchepkov[1], N.L. Podkolodnyy[1], D.A. Afonnikov[1, 2], I. Grosse[3], N.A. Kolchanov[1, 2, 4]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: orlov@bionet.nsc.ru;
[2] Novosibirsk National Research State University, Novosibirsk, Russia;
[3] Institute of Computer Science, Martin Luther University, Halle, Germany;
[4] National Research Centre «Kurchatov Institute», Moscow, Russia

Summary

The pilot program complex for analysis of symbol sequences in genomics, ICGenomics, has been designed for storage, mining, and analysis of sequences related to theoretical and applied genomics. ICGenomics enables wet-lab biologists to perform high-quality processing of data in the fields of genomics, biomedicine, and biotechnology. ICGenomics implements both conventional and modern methods for processing, analyzing, and visualizing sequence data. They include novel methods of the processing of initial high-throughput sequencing data. Examples are: ChIP-seq analysis; functional annotation of gene regulatory regions in nucleotide and amino acid sequences; prediction of nucleosome positioning; and structural and functional annotation of proteins, including their allergenicity and evolution features. Application of ICGenomics to the analysis of genomic sequences of the parasite *Opisthorchis felineus* and to ChIP-seq data on the mouse and human is considered. The system is available at http://www-bionet.sscc.ru/icgenomics.

Key words: genomics, program complex, high-throughput sequencing, nucleotide sequences, data analysis, ChIP-seq.

УДК 577.214:004.822

ИНФОРМАЦИОННАЯ ПОДДЕРЖКА ИССЛЕДОВАНИЯ МЕХАНИЗМОВ РЕГУЛЯЦИИ ТРАНСКРИПЦИИ: ОНТОЛОГИЧЕСКИЙ ПОДХОД

© 2012 г. **Н.Л. Подколодный[1, 2], Е.В. Игнатьева[1], О.А. Подколодная[1], Н.А. Колчанов[1, 3, 4]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия;
[2] Институт вычислительной математики и математической геофизики СО РАН, Новосибирск, Россия, e-mail: pnl@bionet.nsc.ru;
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия;
[4] НИЦ «Курчатовский институт», Москва, Россия

Поступила в редакцию 15 июля 2012 г. Принята к публикации 31 августа 2012 г.

В настоящее время накоплен колоссальный объем данных в области регуляции транскрипции генов эукариот, которая контролируется при участии большого количества белков, выполняющих различные функции в зависимости от стадии процесса транскрипции, что создает возможность реализации большого разнообразия механизмов регуляции. В данной работе представлены подходы к построению онтологии предметной области, формализации описания механизмов регуляции транскрипции и разработке на этой основе методов интеграции гетерогенной информации об особенностях регуляции экспрессии генов эукариот и базы знаний по механизмам регуляции транскрипции. Описана пилотная версия базы знаний по регуляции транскрипции генов эукариот, которая включает понятия, связанные с процессом регуляции транскрипции; иерархическую классификацию регуляторов транскрипции; классификацию этапов и стадий транскрипции, а также базу данных транскрипционных регуляторов трех видов млекопитающих (человека, мыши, крысы) и словари по молекулярным процессам, обеспечивающим регуляцию транскрипции. База знаний предназначена для информационной поддержки исследования механизмов тканеспецифичной регуляции транскрипции генов. Рассмотрены подходы к построению гипотез о механизмах регуляции транскрипции генов эукариот с использованием информации из базы знаний.

Ключевые слова: биоинформатика, регуляция транскрипции, базы знаний, онтология.

ВВЕДЕНИЕ

Транскрипция генов эукариот – сложный процесс, который осуществляется при участии РНК-полимераз трех типов: Pol I, Pol II, и Pol III, каждая из которых обеспечивает транскрипцию определенного набора генов со специфическими механизмами регуляции (Carey, Smale, 2000; Kolchanov *et al*., 2002, 2008). Транскрипционная активность конкретного гена многоклеточного эукариотического организма зависит от типа клетки, ткани и органа, стадии развития организма, стадии клеточного цикла, этапа дифференцировки клеток, воздействия многочисленных индукторов и репрессоров и т. д. Кроме того, транскрипция генов эукариот контролируется большим количеством регуляторных белков, выполняющих различные функции на разных стадиях регуляции транскрипции и работающих в тесной кооперации в составе сложных комплексов (рис. 1).

Например, процесс инициации транскрипции, т. е. позиционирование РНК-полимеразы в районе старта транскрипции гена с последующим образованием короткой цепи РНК (2–9 оснований), осуществляется при участии ба-

Рис. 1. Основные классы белков, участвующих в регуляции транскрипции генов эукариот.

ПИК – предынициаторный комплекс, включающий РНК-полимеразу и базальные транскрипционные факторы.

зальных транскрипционных факторов, которые являются общими для всех генов, транскрибируемых конкретной РНК-полимеразой. Еще один класс регуляторных белков составляют транскрипционные факторы (ТФ), каждый из которых специфичным образом осуществляет регуляцию определенных групп генов в соответствии с клеточной ситуацией (Lemon, Tjian, 2000). ТФ взаимодействуют с определенными участками ДНК в регуляторных районах генов – сайтами связывания транскрипционных факторов (ССТФ) и влияют на интенсивность транскрипции. Обязательным атрибутом ТФ является наличие ДНК-связывающего домена, участвующего в распознавании специфических сигналов (сайтов связывания) в регуляторных районах генов, регулируемых конкретным ТФ.

Помимо ТФ к числу регуляторов транскрипции относятся белки-медиаторы и корегуляторные (кофакторные) белки, которые, как правило, не имеют ДНК-связывающих доменов и участвуют в регуляции транскрипции без непосредственного специфического взаимодействия с ДНК. Белки-медиаторы в составе медиаторного комплекса взаимодействуют с РНК-полимеразой II в области ее С-концевого домена и стабилизируют контакт РНК-полимеразы II с ДНК (Hahn, 2004). К числу корегуляторов транскрипции относятся белки, ковалентно модифицирующие гистоны и белки, осуществляющие АТФ-зависимую реорганизацию (ремоделирование) хроматина.

Процесс регуляции транскрипции можно разбить на этапы, каждый из которых характеризуется набором регуляторных событий и их участников.

В настоящее время накоплен колоссальный объем данных в области регуляции экспрессии генов эукариот, наблюдается их непрерывный рост. В связи с этим большую актуальность приобретают формализация описания механизмов регуляции транскрипции и разработка на этой основе методов интеграции гетерогенной информации об особенностях регуляции экспрессии генов.

Проблемы разработки онтологии регуляции экспрессии генов

Одним из основных этапов семантической интеграции гетерогенных данных является согласование понятий предметной области, их определений и атрибутов, отношений между ними, способов их описания и использования, а также связанных с ними аксиом и правил вывода. Такое согласованное описание конкретной предметной области называют онтологией (Smith *et al*., 2005).

В настоящее время онтологическое моделирование и построение онтологии становятся существенной частью современной биоинформатики и активно применяются при накоплении, сравнении, интеграции и анализе больших объемов гетерогенных данных, полученных с использованием высокопроизводительных экспериментальных исследований в масштабе генома (Подколодный, 2011).

В качестве примера можно привести онтологии, представленные в Open Biological Ontologies (OBO) (http://obo.sourceforge.net). Здесь содержится описание более 70 онтологий по различным направлениям, включая анатомию, биохимию, биологические процессы, биологические функции, биологические последовательности, здоровье, окружающую среду, экспериментальные доказательства, фенотип, белки, таксономии и др. (Schober *et al*., 2009). В рамках проекта OBO разрабатываются унифицированные подходы для разработки онтологий, методы интеграции онтологий, а также инструментальные средства для работы с онтологиями (Smith *et al*., 2007).

Одним из самых успешных проектов создания онтологии является Gene Ontology (GO) (http://www.geneontology.org/), которая включает 3 раздела: биологические процессы (biological process), биологические структуры (cellular component) и молекулярные функции (molecular function), которые выполняют гены, РНК или белки, локализованные в определенных клетках или клеточных структурах в том или ином биологическом процессе (Gene Ontology Consortium, 2010).

Онтологии позволяют представить понятия в таком виде, что они становятся пригодными для машинной обработки и вследствие этого используются в качестве посредника между пользователем и информационной системой или между членами научного сообщества при обмене данными.

Формально онтология включает набор понятий (терминов) предметной области, их определений и атрибутов, а также связанных с ними аксиом и правил вывода. Таким образом, формальная модель онтологии – это упорядоченная тройка конечных множеств:

$$\mathbf{O} = < \mathbf{T, R, F} >,$$

где **T** – конечное и непустое множество классов и концептов (понятий, терминов) предметной области, которую описывает онтология **O**; **R** – конечное множество отношений между концептами заданной предметной области; **F** – конечное множество функций интерпретации, заданных на понятиях и/или отношениях онтологии **O**, или аксиом, используемых для моделирования утверждений, которые всегда являются истинными, что ограничивает интерпретацию и обеспечивает корректное использование понятий.

Одним из наиболее продуктивных подходов к описанию и использованию знаний о предметной области являются дескриптивные логики (ДЛ), которые определяют формальный язык для описания понятий (концéпт, класс, категория или сущность) и отношений между понятиями (называемых ролями), утверждений о фактах и запросов к ним. Кроме этого, в ДЛ входят конструкторы (операции) для понятийных выражений, включающие конъюнкцию, дизъюнкцию и определение отношений.

Базы знаний предметной области с позиции дескриптивной логики подразделяются на общие знания о понятиях и их взаимосвязях (T-Box) и знания об индивидуальных объектах, их свойствах и связях с другими объектами (A-Box).

T-box (terminological knowledge) – это набор утверждений, описывающих множество классов понятий предметной области, их свойства и отношения между ними. Эти знания более стабильны и постоянны. Именно эти знания соответствуют онтологии предметной области.

A-box (assertional knowledge) содержит утверждения об экземплярах понятий, т. е. описывает предметную область на уровне конкретных данных (база данных). В базе знаний обе компоненты взаимосвязаны.

Разработка онтологии регуляции транскрипции является сложным и затратным процессом.

Первый этап этого процесса – онтологический анализ предметной области регуляции транскрипции генов эукариот, включая создание словаря терминов, точных их определений и взаимосвязей между ними, описание правил и ограничений, согласно которым на базе введенной терминологии формируются достоверные утверждения о состоянии системы.

С использованием стандартов OBO разрабатывается онтология регуляции генов Gene Regulation Ontology (GRO) (Beisswanger *et al*., 2008), которая включает 508 классов, в том числе и классы, описывающие процессы различных типов воздействия, биологические процессы, экспериментальные воздействия, молекулярные процессы, мутации, регуляторные процессы и т. д.

Следует отметить, что знания о механизме регуляции транскрипции генов основываются на интеграции гетерогенных знаний о биологических объектах (белках, генах, РНК и др.), вовлеченных в регуляторный процесс, их структурно-функциональной организации и ролях, которые они играют на различных стадиях регуляции. Поэтому механизм регуляции транскрипции можно описывать на разном уровне детальности, и полнота описания зависит от наших знаний и возможностей.

Механизм регуляции транскрипции удобно характеризовать с помощью таких понятий, как событие, действие, процесс. В этом случае для описания механизма регуляции транскрипции необходимо выделить основные подпроцессы, из которых складывается это биологическое явление; описать основных участников этих процессов и их ролевые функции. В качестве участников процесса регуляции транскрипции выступают гены и регуляторы транскрипции различного типа, включая транскрипционные факторы, корегуляторные белки, белки-медиаторы и т. п.

Пространство описания понятий предметной области определяется необходимостью отвечать на вопросы: ЧТО? (описание ситуации или события, например уровень экспрессии генов в клетках конкретного биоматериала), ГДЕ и КОГДА? (локализация события во времени (возможно с точностью до отношения к моменту другого события) и в пространстве (возможно с точностью до компартмента), описание биоматериалов и клеточной ситуации: вид организма, состояние организма, индукторы, органы, ткани, клетки, их стадии развития), КАК (механизмы регуляции транскрипции и их нарушение) и ПОЧЕМУ? (множество событий, которые необходимы для понимания семантики конкретного события).

Роль объекта определяется в контексте реализации конкретного события, которое изменяет значения атрибутов объектов, определяющих ситуацию. Поэтому для каждого события необходимо определить роли участников этих событий, которые необходимы для реализации события.

Таким образом, представление механизма регуляции транскрипции включает описание структуры системы и участников процесса регуляции транскрипции, множества возможных состояний системы, множества взаимосвязанных событий, которые определяют поведение системы и меняют состояние системы, а также роли, которые играют отдельные элементы системы в реализации тех или иных событий.

В общем случае ситуация может быть охарактеризована предикатом, который является истинным или ложным в зависимости от того, наблюдается или нет данная ситуация. Ситуации являются абстрактными сущностями и могут обладать различными свойствами. В этом отношении ситуации могут быть простыми (т. е. не иметь внутренней структуры и быть пределом точности описания в данной модели внешнего мира) и сложными, имеющими определенную структуру, включающими подмножество ситуаций.

Основой для формирования онтологии регуляции транскрипции генов является формальное представление следующих понятий:

− физические сущности (Physical_Entity), в частности ген РНК, белок, белковый комплекс, геномная последовательность, район регуляции транскрипции, промотор, сайт связывания транскрипционного фактора, нуклеосома, транскрипционный фактор, регулятор транскрипции и т. д.;

− механизм регуляции транскрипции;

− стадии регуляции транскрипции;

− регуляторные события, обусловливающие реализацию механизмов регуляции транскрип-

ции и роли, которые играют участники в этих событиях;

− описание клеточных ситуаций, в которых получены экспериментальные данные по экспрессии генов;

− свойства регуляторов транскрипции, которые коррелируют с их функциональными возможностями; компьютерное предсказание этих свойств позволяет, например, делать выводы о возможности участия конкретного белка в регуляции транскрипции на определенной стадии, т. е. выполнении определенной роли на этой стадии;

− структурно-функциональные закономерности организации регуляторных районов генов (регуляторные структурные модули), обусловливающих особенности регуляции экспрессии генов, коэкспрессирующихся в разных клеточных ситуациях (Подколодный и др., 2010).

К отношениям верхнего уровня относятся базовые отношения (например is_a/has_subclass, part_of/has_part, part_for, instance_of/has_instance, includes/include_of, composed_of, consist_of), пространственные отношения (например located_in, contained_in, includes, composed_of, adjacent_to), временные, или темпоральные, отношения (например transformation_of, derives_from, preceded_by), отношения участия (например has_participant, has_agent, regulates и т. д. (Özgövde *et al*., 2010).

Ниже приведены определения и примеры использования некоторых базовых отношений между классами, которые применяются нами при описании предметной области:

Отношение ***is_a*** (класс–подкласс):

X ***is_a*** $Y =_{def} \forall x : x$ ***instance_of*** $X \Rightarrow y$ ***instance_of*** Y.

Пример. P_1 ***is_a*** P_2 – любой белок из класса P_1 входит в класс P_2.

Отношение ***part_of***:

X ***part_of*** $Y =_{def} \forall x, t : x$ ***instance_of*** X *at* $t \Rightarrow \exists y : (y$ ***instance_of*** Y *at* t *&* x ***part_of*** y *at* $t)$.

Пример. P_1 ***part_of*** P_2 – для любого белкового комплекса из класса P_2 независимо от времени t существует белок из класса P_1, который входит в этот белковый комплекс, а также для любого белка из класса P_1 существует белковый комплекс из класса P_2, в который входит этот белок. Таким образом, P_1 – класс белков, которые образуют комплексы, а P_2 – класс белковых комплексов, которые образуют белки из P_1.

На рис. 2 в качестве примера представлена схема фрагмента раздела «Biomaterials» онтологии регуляции транскрипции.

На рис. 3 представлен фрагмент раздела «Genome_Entity» онтологии регуляции транскрипции.

В качестве основных типов молекулярно-генетических событий, которые играют важную роль в регуляции транскрипции, можно выделить:

− связывание (bind);

− освобождение (release);

− расщепление (cleavage);

− модификации (modify), включая:

− модификации, связанные с появлением новых связей, например phosphorylate, glycosylate, methylate, hydroxylate, acetylate, acylate, ubiquitinize и др.;

− модификации, связанные с разрушением связей, например dephosphorylate, glycosylate, demethylate, dehydroxylate, deacetylate, deacylate, deubiquitinize и др.;

− транспорт (transport).

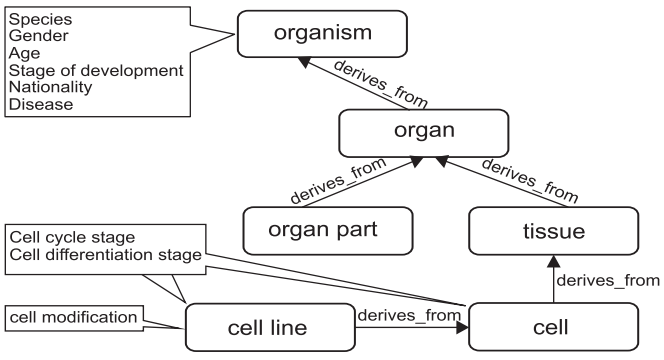

Рис. 2. Фрагмент раздела «Biomaterials» онтологии регуляции транскрипции.

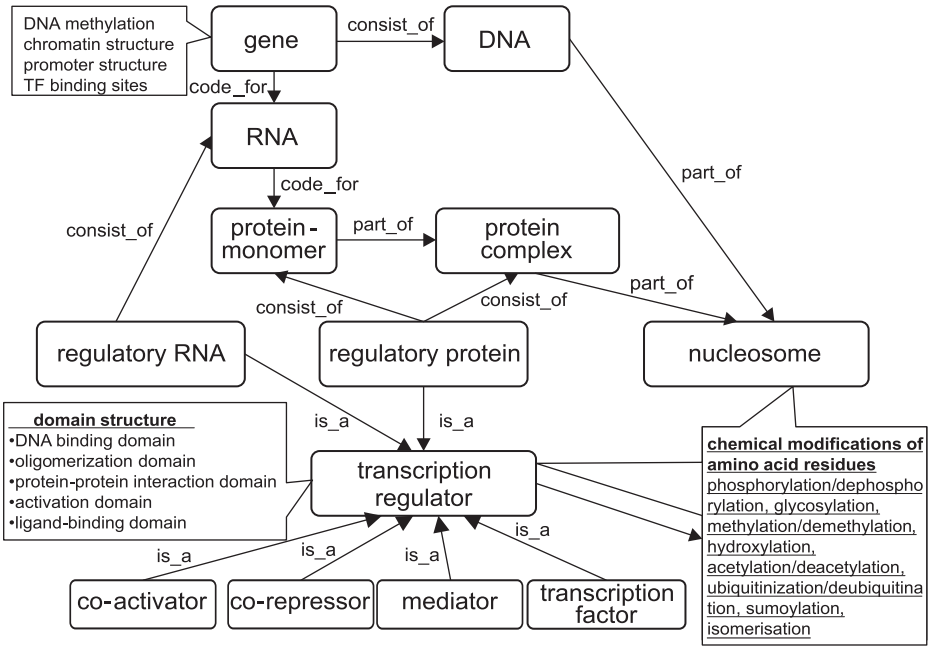

Рис. 3. Фрагмент раздела «Genome_Entity» онтологии регуляции транскрипции.

База знаний по регуляции транскрипции

С целью систематизации данных о механизмах регуляции транскрипции генов эукариот нами разработана база знаний по регуляции транскрипции, включающая иерархически организованные словари (классификаторы) белков, регулирующих транскрипцию, стадий транскрипции, молекулярных процессов, обеспечивающих регуляцию транскрипции, а также сведения о транскрипционных регуляторах трех видов млекопитающих: человека, мыши, крысы (табл. 1) (Shipra *et al*., 2006; Podkolodnyy *et al*., 2008; Schaefer *et al*., 2011).

Таблица 1

Количество записей
по транскрипционным регуляторам
раздела A-box базы знаний

Организм	Человек	Мышь	Крыса
Транскрипционные факторы	1365	1301	1267
Транскрипционные кофакторы	536	521	508
Факторы, участвующие в ремоделировании хроматина	64	63	63

Построена классификация регуляторов транскрипции (рис. 4), имеющая иерархическую структуру (до 4-го уровня иерархии). Классы белков в классификации характеризуются через их функциональные роли в процессе транскрипции либо его регуляции. Понятия первого уровня иерархии соответствуют терминам, обозначающим активности РНК-полимераз (DNA-directed RNA polymerase activity) и транскрипционных регуляторов (transcription regulator activity). Понятия второго уровня описывают функциональные роли белков, либо указывая на тип РНК-полимеразы, в комплексе с которой работает белок, либо характиризуя белок как кофакторный (корегуляторный). Понятия третьего и четвертого уровней более детально характеризуют активности функциональных подклассов белков, регуляторов транскрипции.

Например, понятие второго уровня «транскрипционные кофакторы» (transcription cofactor activity) включает термин следующего, третьего, уровня, обозначающий активность белков, модифицирующих гистоны (histone modification activity). Данный термин, в свою очередь, имеет подчиненные понятия четвертого уровня иерархии, уточняющие механизм функционирования транскрипционных регуляторов конкретных подклассов: histone kinase activity, histone acetyl-

1. DNA-directed RNA polymerase activity
   1.1. DNA-directed RNA polymerase I activity
   1.2. DNA-directed RNA polymerase II activity
   1.3. DNA-directed RNA polymerase III activity
2. transcription regulator activity
   2.1. RNA polymerase I transcription factor activity
   2.2. RNA polymerase II transcription factor activity
      2.2.1. general RNA polymerase II transcription factor activity
      2.2.2. specific RNA polymerase II transcription factor activity
   2.3. RNA polymerase III transcription factor activity
   2.4. transcription cofactor activity
      2.4.1. histone modification activity*
         2.4.1.1. histone kinase activity
         2.4.1.2. histone acetyltransferase activity
         2.4.1.3. histone methyltransferase activity
         2.4.1.4. histone deacetylase activity
         2.4.1.5. histone demethylase activity
      2.4.2. ATP-dependent chromatin remodeling activity*

Рис. 4. Фрагмент иерархической классификации функциональных ролей белков, участвующих в процессе транскрипции и ее регуляции.

* Понятия, отсутствовавшие в GO и включенные на основе анализа научных публикаций.

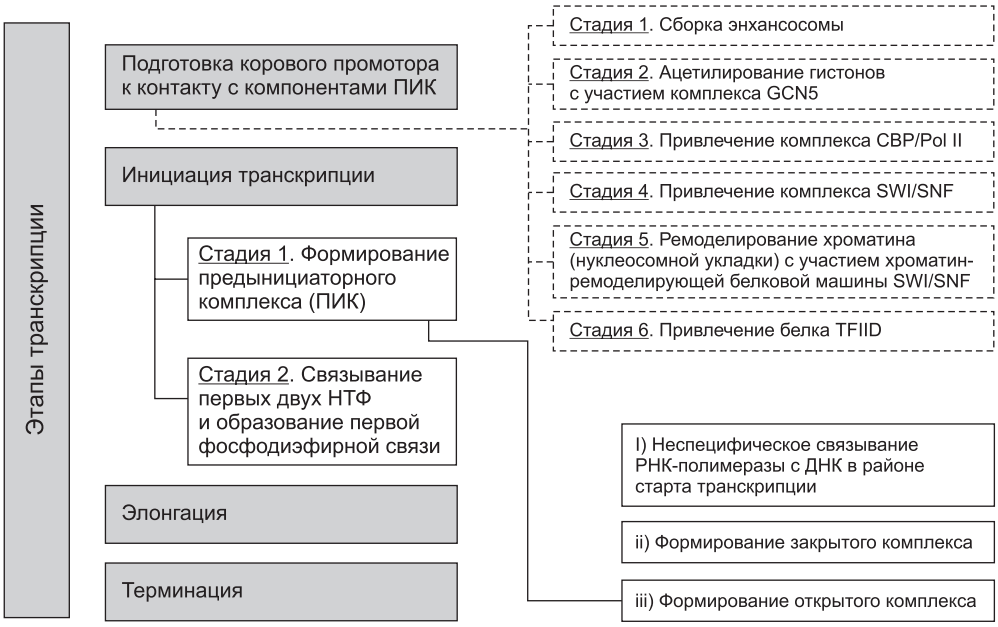

Рис. 5. Этапы и стадии процесса транскрипции генов эукариот.

Этапы обозначены серыми прямоугольниками. В качестве примера для этапа «инициация транскрипции» приведены понятия второго (стадии) и третьего уровней (процессы), которые обозначены белыми прямоугольниками со сплошной границей. Стадии, специфичные для этапа подготовки корового промотора конкретного гена (интерферона β человека), обозначены прямоугольниками, ограниченными пунктиром.

transferase activity, histone methyltransferase activity и др.

Классификация этапов и стадий транскрипции включает понятия нескольких иерархических уровней, соответствующих упорядоченным по времени (т. е. следующих друг за другом) процессам. Понятия верхнего уровня иерархии соответствуют основным этапам, посредством которых осуществляется транскрипция генов эукариотических организмов (рис. 5). Необходимо отметить, что этап подготовки корового промотора к контакту с компонентами предынициаторного комплекса (ПИК) специфичен для процесса транскрипции генов эукариот, в силу того что геномная ДНК эукариот находится в комплексе с белками (хроматин), что затрудняет взаимодействие белков транскрипционной машины с ДНК в области промотора (Разин, 2007; Berger, 2007).

Понятия следующего уровня иерархии соответствуют стадиям, посредством которых реализуется конкретный этап. Например, этап инициации транскрипции включает две стадии. На первой стадии происходит формирование предынициаторного комплекса, включающего РНК-полимеразу и вспомогательные белки (базальные транскрипционные факторы), на второй стадии осуществляются связывание первых двух нуклеотидтрифосфатов и образование первой фосфодиэфирной связи вновь синтезированного транскрипта РНК (Hahn, 2004).

В свою очередь, стадия формирования ПИК включает следующие процессы:

– первоначально РНК-полимераза связывается с двухцепочечной ДНК неспецифически;

– затем РНК-полимераза в составе ПИК связывается с ДНК в районе промотора специфически, благодаря электростатическим взаимодействиям и формирует закрытый комплекс, в котором ДНК сохраняет двухспиральную структуру.

– далее закрытый комплекс превращается в открытый, в котором РНК-полимераза расплетает двойную спираль ДНК в районе точки инициации транскрипции (Hahn, 2004).

Классификация этапов и стадий транскрипции включает понятия двух типов:

1) общие для всех генов;

2) специфичные для конкретного гена либо группы генов и выполняющиеся в каждом конкретном случае в определенном порядке.

Например, подготовка корового промотора к взаимодействию с машиной может осуществляться различными механизмами, комбинации которых обеспечивают разнообразие паттернов экспрессии генов, специфичных для определенной стадии развития организма, ткани либо типа клеток (Blanchette *et al.*, 2006). Для гена интерферона β человека этот этап включает 6 стадий (рис. 5) (Agalioti *et al*., 2000).

Раздел A-box базы знаний включает также данные по транскрипционным факторам, транскрипционным кофакторам и белкам, участвующим в ремоделировании хроматина (табл. 1). В настоящее время в базе знаний содержатся сведения по транскрипционным регуляторам трех видов млекопитающих (человека, мыши, крысы) (Ignatieva, 2012).

Словарь по молекулярным процессам, обеспечивающим регуляцию транскрипции, составлен на основе анализа терминов из системы Gene Ontology (раздел «biological_process»), а также анализа научных публикаций и включает около 40 терминов, распределенных по 4 иерархическим уровням. Например, термин, обозначающий «регуляцию транскрипции путем реорганизации хроматина», является одним из понятий верхнего уровня иерархии. Термины следующих двух уровней иерархии представляют более детальное описание возможных механизмов реализации процесса (рис. 6).

Представление и использование знаний о механизмах регуляции транскрипции

В нашей онтологии механизм регуляции транскрипции является классом, который соответствует потенциально всем возможным вариантам реализации процессов регуляции. В зависимости от контекста, предусловия и источника реализуется конкретный вариант процесса.

На рис. 7 представлена упрощенная схема зависимостей между компонентами базы знаний T-box и A-box, включающая описание понятий Process, Event, Role, Type_of_object, конкретные реализации (экземпляры): process, event, object соответственно.

Процесс регуляции транскрипции является комплексным процессом, который состоит из множества взаимосвязанных подпроцессов и/или элементарных событий.

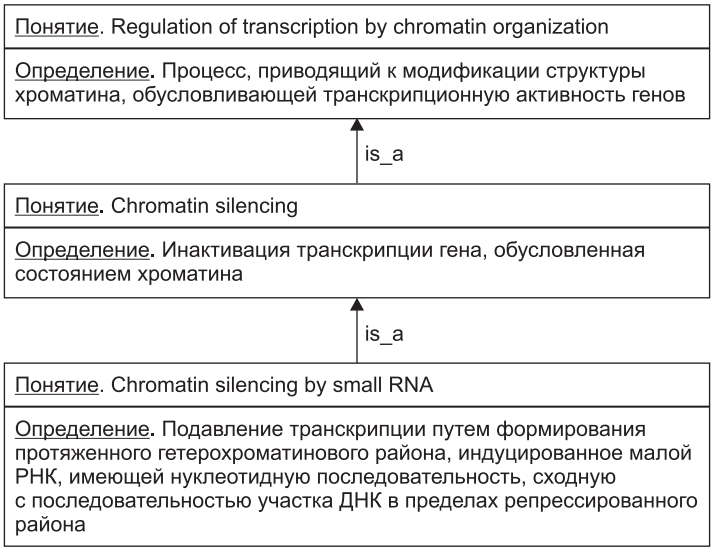

Рис. 6. Группа иерархически подчиненных терминов из словаря по молекулярным процессам, обеспечивающим регуляцию транскрипции.

Для описания механизма регуляции транскрипции необходимо выделить основные подпроцессы и регуляторные события, из которых складывается это биологическое явление; описать основных участников этих процессов и их ролевые функции.

Событие считается неделимым, и все процессы описываются с точностью до события. Для каждого класса события Event указывается набор ролей объектов (Role), которые необходимы для реализации этого события или могут участвовать в нем.

За основу описания понятия «процесс» нами взяты стандартные ситуативные роли Дж. Совы (http://www.jfsowa.com/ontology/) и работа N. Baumgartner с соавт. (2006). В описание понятия «процесс» входят следующие компоненты.

– **Контекст**. Какая клеточная ситуация (состояние системы) может обусловливать реализацию данного механизма? Где (гены, виды, органы, ткани, типы клеток и т. д.), когда (стадии клеточного цикла, стадии развития и т. д.) может возникать такая клеточная ситуация.

– **Предусловие** – условия, необходимые для старта процесса, реализующего механизм: наличие участников процесса, внешние сигналы и т. д.

– **Сценарий процесса** – сеть взаимосвязанных событий с частичным порядком по времени. Фиксированный источник порождает дерево событий с частичным порядком. Корнем этого дерева будет **Источник**.

– **Инициатор** – детерминирующий участник (участник, определяющий направление процесса или цель).

– **Источник** – внешний сигнал, запуск процесса, начало процесса (должен присутствовать в начале процесса, но не обязан принимать участие во всем процессе).

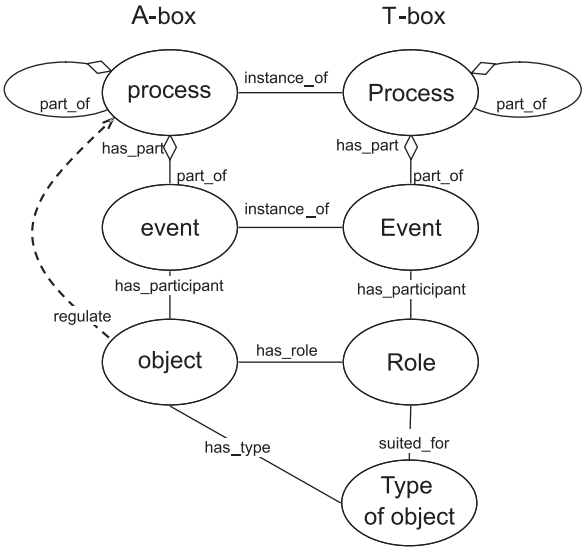

Рис. 7. Схема зависимостей между компонентами базы знаний.

Таблица 2

Формальное описание механизма активации эукариотического гена на примере интерферона β человека

1. Характеристика объекта		
Ген (G_k)	Вид организма	Клеточная ситуация (C_i)
IFNB1	Homo sapiens	Virus-infected HeLa cells
2. Характеристика стадий транскрипции и событий		
Этапы транскрипции (S_i)	Стадии и события (E_i)	Регуляторные белки (комплексы) (S_i, K_i), функционирующие на данной стадии
S_1: Подготовка корового промотора к контакту с компонентами ПИК	E_1 – сборка энхансосомы	Транскрипционные факторы ATF2, NF-KB, IRF1, HMGI(Y)
	E_2 – ацетилирование гистонов с участием комплекса GCN5	Комплекс GCN5
	E_3 – привлечение комплекса CBP/ Pol II	Комплекс CBP/ Pol II
	E_4 – привлечение комплекса SWI/SNF	Комплекс SWI/SNF
	E_5 – ремоделирование хроматина (нуклеосомной укладки) с участием хроматин-ремоделирующей белковой машины SWI/SNF	Комплекс SWI/SNF
	E_6 – привлечение белка TFIID	TFIID

3. Выборочная характеристика функциональных ролей регуляторных белков – участников определенной стадии транскрипции (на примере двух белков)

Регуляторный белок (P_i)	Функциональный класс	Стадия процесса транскрипции	Событие	Функциональная роль регуляторного белка (R_i)
HMG(Y)	Транскрипционный фактор	S_1	E_1 – сборка энхансосомы	Связывание с ДНК и белок-белковые взаимоедйствия в пределах энхансосомы
Комплекс GCN5	Корегулятор транскрипции	S_1	E_2 – ацетилирование гистонов с участием комплекса GCN5	Ацетилирование гистонов

− **Продукт** – продукт (может появляться в конце процесса, но не обязан принимать участие во всем процессе).

− **Постусловие** – условие окончания процесса.

Описание элементарного события (Event) включает описание ролей (Role) участников этого события, которые могут иметь определенный тип (Type of object). Например, множество участников события типа (класса) «Метаболическая реакция» включает типы объектов или роли: субстраты, продукты, ферменты, коферменты и регуляторы. Конкретная реакция (реализация или экземпляр класса) идентифицируется набором конкретных субстратов, продуктов и ферментов. Регуляторы реакции влияют на скорость реакции, т. е. меняют значения параметров этого события.

Иерархические классификаторы и сведения о транскрипционных регуляторах, накопленные в базе знаний, обеспечивают возможность формализованного описания механизмов регуляции транскрипции.

Например, механизм активации гена интерферона β человека, реконструированный на основе экспериментальных данных в работе Agalioti с соавт. (2000), может быть представлен в виде стадий регуляции транскрипции и набора регуляторных событий (табл. 2). Такое описание включает:

– характеристику объекта и клеточной ситуации;

– характеристику этапов (стадий) процесса транскрипции, а также регуляторных событий с указанием их участников (регуляторных белков);

– характеристику функциональных ролей регуляторных белков, участвующих в регуляторных событиях на конкретном этапе.

Для данных, представленных в формате OWL/RDF, можно сделать запрос на языке SPARQL (SPARQL Query Language for RDF, 1998).

```
SELECT ?stadia ?event ?role ?objectType ?objectName
WHERE {
      ?gene rdf:has_type gene.
      ?gene rdf:species ?species.
      ?gene rdf:name ?geneName.
      ?process rdf:has_type «regulation of transcription».
      ?process rdf:has_context ?context.
      ?context rdf:gene ?gene.
      ?process rdf:has_part ?stadia.
      ?stadia  rdf:has_part ?event.
      ?event   rdf:has_participant ?object.
      ?object  rdf:has_type ?objectType.
      ?object  rdf:has_role ?role.
      ?object  rdf:has_name ?objectName.
       FILTER (?gene = «IFN beta», ?species = «human»)
      }
```

В результате запроса будет выдана таблица, включающая список всех стадий регуляции транскрипции гена интерферона β человека, с указанием всех событий, которые происходят на каждой стадии, участников этих событий с указанием роли, типа и имени объекта.

Знания, полученные из гетерогенных источников, могут быть неполными, фрагментарными, нечеткими, косвенными и противоречивыми. В частности, может оказаться известным только то, что белок в составе некоторого неизвестного комплекса участвует в регуляции транскрипции. Знания о составе белкового комплекса тоже могут быть неполными. Например, не все субъединицы комплекса известны, или неизвестно, сколько всего субъединиц входит в комплекс.

В некоторых случаях имеется возможность генерации правдоподобных гипотез, которые не противоречат известным фактам. Такого рода гипотетические знания с указанием относительного уровня достоверности полезны при дальнейшем анализе и построении непротиворечивых знаний (Ponomaryov *et al*., 2011).

Пусть, например, известно, что некоторый белок в составе неизвестного комплекса участвует в регуляции транскрипции. Среди множества белковых комплексов, в состав которых входит этот белок, те комплексы, в состав которых входят другие белки, обладающие способностью регулировать транскрипцию, с большой вероятностью могут быть транскрипционными факторами.

Примером косвенных знаний могут быть знания о взаимодействии между субъединицами белков, участвующих в регуляции транскрипции. Эти знания дают основание предположить, что участие обоих этих субъединичных белков в регуляции транскрипции может осуществляться через образование транскрипционного комплекса, в который входят оба белка.

Ниже приводится пример логического вывода новых знаний о регуляции транскрипции на основании фактов о связывании белков и образовании белкового комплекса и участии их в регуляции транскрипции.

∀ proteinA, proteinB
// 1. Белок proteinA участвует в регуляции транскрипции через образование
// неизвестного регуляторного комплекса proteinX
$\exists e_1$: e_1 ***instance_of*** TranscriptionRegulationProcess and e_1 ***has_participant*** proteinX,
∃proteinX : proteinA ***part_of*** proteinX,

// 2. Белок proteinB участвует в регуляции транскрипции через образование
// неизвестного регуляторного комплекса proteinY
$\exists e_1 : e_2$ ***instance_of*** TranscriptionRegulationProcess,
$\exists$proteinY : e_2 ***has_participant*** proteinY and proteinB ***part_of*** proteinY,
// 3. Белок proteinA связывается с белком proteinB, образуя белковый комплекс ProteinAB.
proteinA ***part_of*** proteinAB & proteinB ***part_of*** proteinAB
Вывод (гипотеза):
proteinX ≡ proteinY ≡ proteinAB & $e_1 \equiv e_1 \equiv e$ & e ***has_participant*** proteinAB.

Это предположение становится более правдоподобным, если известно, что действие этих белков на транскрипцию одинаково (подавление либо усиление транскрипции). Это позволяет задать частичный порядок на множестве гипотез по уровням относительной достоверности.

В ряде случаев можно распространять свойства через мереологические иерархии (часть–целое). В качестве примера вывода гипотетических свойств белкового комплекса по свойствам субъединиц можно привести связывание с ДНК (DNA_binding). Наличие ДНК-связывающего домена в субъединице позволяет сделать предположение о возможности связывания белкового комплекса, в который входит эта субъединица. Безусловно, это предположение может рассматриваться только как гипотеза, и только экспериментальная проверка может подтвердить этот факт.

Анализ реализаций механизмов регуляции транскрипции для конкретных генов в конкретной клеточной ситуации дает информацию о возможных этапах регуляции транскрипции, множестве возможных элементарных событий, составляющих их, и типах участников событий.

Аналогичные этапы могут присутствовать в регуляции транскрипции других генов. Можно предположить, что в реализации элементарных событий, возможно, будут участвовать другие объекты, но того же типа, т. е. играть ту же роль в процессе. Поэтому рассуждение по аналогии позволяет делать выводы о возможных вариантах реализации механизма регуляции конкретного гена, перебирая на роль участника процесса все объекты соответствующего типа, которые могут присутствовать в данной клеточной ситуации.

Предположим, что нам не известен механизм регуляции транскрипции некоторого гена G, т. е. не известны этапы, стадии регуляции транскрипции, составляющие их события, и участники этих событий. Однако известно, что белок P участвует в регуляции транскрипции этого гена. Тогда возможен следующий вариант логического анализа и вывода гипотез о возможных механизмах регуляции транскрипции для этого гена:

Шаг 1. Анализ ролей $\{R_i\}$ белка P, которые этот белок имел (has_role) в регуляторных событиях (Event).

Шаг 2. Выделение классов событий $\{E_i\}$, для которых необходимы участники с ролями $\{R_i\}$.

Шаг 3. Выявление других ролей $\{Rs_k\}$, которые необходимы для реализации этих классов событий $\{E_i\}$.

Шаг 4. Поиск в базе знаний других объектов $\{Ps_i\}$, участников событий этих классов $\{E_i\}$ и анализ возможности участия этих объектов в регуляции гена G.

Шаг 5. Выявление класса процессов (этапов и стадий регуляции транскрипции), которые включают (part_of) эти события.

Шаг 6. Реконструкция гипотетического механизма регуляции транскрипции с типами стадий регуляции транскрипции, которые включают события $\{E_i\}$ с участниками $\{Ps_i\}$, выполняющими роли $\{Rs_k\}$.

Если роли объекта не известны, то для их предсказания возможно использование знаний о типе объекта и его свойствах. Предполагается, что класс Type_of_object включает описание типов объектов и их свойств, которые важны для предсказания возможности выполнения определенных ролей.

Необходимо отметить, что большинство этапов логического анализа может быть выполнено путем запроса к базе знаний.

ЗАКЛЮЧЕНИЕ

Разработаны подходы к построению онтологии регуляции транскрипции. На основе разработанной онтологической модели создана

пилотная версия базы знаний по регуляции транскрипции генов эукариот, включающей знания (T-Box) об онтологических понятиях и их взаимосвязях в области регуляции транскрипции (иерархически организованные словари (классификаторы) белков, регуляторов транскрипции, этапов транскрипции, молекулярных механизмов) и базу данных (A-box) по регуляторам транскрипции человека, мыши и крысы. На примере описания этапов и стадий активации конкретного эукариотического гена продемонстрирована применимость онтологической модели и базы знаний для формализованного описания механизмов регуляции транскрипции генов и построения гипотез о механизмах регуляции транскрипции генов эукариот с привлечением информации из базы знаний. Таким образом, предложенные подходы могут использоваться при реконструкции гипотетических механизмов регуляции транскрипции с учетом информации о строении регуляторных районов генов и функциях регуляторных белков, присутствующих в заданных клетках или тканях на определенной стадии развития.

В дальнейшем нами планируется развитие базы знаний по регуляции транскрипции генов эукариот с целью ее использования для интерпретации закономерностей строения регуляторных районов коэкспрессирующихся генов, функциональной интерпретации микрочиповых и протеомных данных, отражающих уровни экспрессии генов, и выявления регуляторных составляющих генных сетей, контролирующих фенотипические признаки организма.

БЛАГОДАРНОСТИ

Работа выполнена при частичной поддержке Президиума РАН (проекты А.II.6.8, 30.29), СО РАН (проект фундаментальных исследований VI.50.1.2. «Биоинформатика и системная биология молекулярно-генетических систем и процессов»), Совета по грантам Президента Российской Федерации (НШ-5278.2012.4).

ЛИТЕРАТУРА

Подколодный Н.Л. Онтологическое моделирование в биоинформатике и системной биологии // Онтологическое моделирование. ИПИ РАН, 2011. С. 233–269.

Подколодный Н.Л., Игнатьева Е.В., Рассказов Д.А. и др. Интегрированная система для информационной поддержки исследования механизмов регуляции транскрипции // Тр. 12-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010. Казань, Россия, 2010. С. 69–75.

Разин С.В. Хроматин и регуляция транскрипции // Молекуляр. биология. 2007. Т. 41. № 3. С. 387–394.

Agalioti T., Lomvardas S., Parekh B. *et al*. Ordered recruitment of chromatin modifying and general transcription factors to the IFN-b promoter // Cell. 2000. V. 103. P. 667–678.

Baumgartner N., Retschitzegger W. A survey of upper ontologies for situation awareness // Proc. of the 4th IASTED Intern. Conf. on Knowledge Sharing and Collaborative Engineering, St. Thomas, US VI. 2006. P. 1–9.

Beisswanger E., Lee V., Kim Jung J. Gene regulation ontology (GRO): design principles and use cases // II Proc. 21st. Intern. Congr. of the Europ. Federation for Med. Inform. (MIE 2008). 2008. P. 9–14.

Berger S.L. The complex language of chromatin regulation during transcription // Nature. 2007. V. 447. No. 7143. P. 407–412.

Blanchette M., Bataille A.R., Chen X. *et al*. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression // Genome Res. 2006. V. 16. No. 5. P. 656–668.

Carey M., Smale S.T. Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniques. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. 2000. 639 p.

Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements // Nucl. Acids Res. 2010. V. 38. P. D331–335.

Gene Regulatory aontology (GRO), version 0.5, 1.09.2011 – http://bioportal.bioontology.org/ontologies/1106.

Hahn S. Structure and mechanism of the RNA Polymerase II transcription machinery // Nat. Struct. Mol. Biol. 2004. V. 11. No. 5. P. 394–403.

Ignatieva E.V. TrDB: a database of the human, mouse, and rat transcriptional regulators and its potential applications in systems biology // The Eighth Intern. Conf. on Bioinformatics of Genome Regulation and Structure / Systems Biology (BGRS/SB'12). Novosibirsk, Russia, June 25–29. 2012. P. 125.

Kolchanov N.A., Ignatieva E.V., Ananko E.A. *et al*. Transcription Regulatory Regions Database (TRRD): its status in 2002 // Nucl. Acids Res. 2002. V. 30. No. 1. P. 312–317.

Kolchanov N.A., Ignatieva E.V., Podkolodnaya O.A. *et al*. TRRD: Technology for extraction, storage, and use of knowledge about the structural-functional organization of the transcriptional regulatory regions in the eukaryotic genes // Intell. Data Anal. 2008. V. 12. No. 5. P. 443–461.

Lemon B., Tjian R. Orchestrated response: a symphony of transcription factors for gene control // Genes Dev. 2000. V. 14. No. 20. P. 2551–2569.

Özgövde A., Grüninger M. Foundational process relations in bio-ontologies // Proc. of the Sixth Intern. Conf. on Formal Ontology in Information Systems (FOIS 2010). IOS Press Amsterdam, The Netherlands, 2010. P. 243–256.

Podkolodnyy N.L., Nechkin S.S., Ignatieva E.V. *et al*. A database for analysis of the organizational features of the promoter regions in the co-expressed groups of genes // Proc. of the Sixth Int. Conf. on Bioinformatics of Genome Regulation and Structure, 2008.

Ponomaryov D., Omelianchuk N., Mironova V. *et al*. From published expression and phenotype data to structured knowledge: The Arabidopsis gene net supplementary database and its applications // Lecture Notes in Artificial Intelligence. 2011. P. 101–120.

Schaefer U., Schmeier S., Bajic V.B. TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins // Nucl. Acids Res. 2011. V. 39. P. D106–D110.

Schober D., Smith B., Lewis S. *et al*. Survey-based naming conventions for use in OBO foundry ontology development // BMC Bioinformatics. 2009. 10(125). P. 1–9.

Shipra A., Chetan K., Rao M.R.S. CREMOFAC – a database of chromatin remodeling factors // Bioinformatics. 2006. V. 22. No. 23. P. 2940–2944.

Smith B., Ashburner M., Rosse C. *et al*. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration // Nat. Biotech. 2007. 25(11). P. 1251–1255.

Smith B., Ceusters W., Klagges B. *et al*. Relations in biomedical ontologies // Genome Biology. 2005. V. 6. No. R46.

SPARQL Query Language for RDF. 1998 – http://www.w3.org/TR/rdf-sparql-query/

INFORMATION SUPPORT OF RESEARCH ON TRANSCRIPTIONAL REGULATORY MECHANISMS: AN ONTOLOGICAL APPROACH

N.L. Podkolodnyy[1, 2], E.V. Ignatieva[1], O.A. Podkolodnaya[1], N.A. Kolchanov[1, 3, 4]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;
[2] Institute of Computational Mathematics and Mathematical Geophysics, Novosibirsk, Russia,
e-mail: pnl@bionet.nsc.ru;
[3] Novosibirsk National Research State University, Novosibirsk, Russia;
[4] National Research Centre «Kurchatov Institute», Moscow, Russia

Summary

By now, a huge body of experimental data on gene transcription regulation has been accumulated. Transcription is controlled by a great number of proteins acting at various steps of the process; thus, a diversity of regulatory mechanisms can be realized. This paper presents approaches to building knowledge domain ontology, formalized description of the mechanisms of transcriptional regulation and the development of methods for integration of heterogeneous information on the features of the regulation of gene expression on this base. The pilot version of the knowledge base on the transcriptional regulation of eukaryotic genes includes: (1) description of basic terms related to transcription regulation and relationships between them; (2) hierarchical classification of transcription regulators; (3) classification of phases and steps of transcription; (4) a database of transcriptional regulators of three mammalian species (human, mouse, and rat); and (5) dictionaries for molecular processes involved in transcriptional regulation. The knowledge base is designed for information support of computer analysis of transcriptional regulatory mechanisms. Approaches to reconstruction of eukaryotic transcriptional regulatory mechanisms with the new knowledge base are presented.

Key words: bioinformatics, transcription regulation, knowledge base systems, ontology.

УДК 577.24;57.084.1;57.088.1;57.088.3;577.214.3;575.22;004.9

RatDNA: БАЗА ДАННЫХ МИКРОЧИПОВЫХ ИССЛЕДОВАНИЙ НА КРЫСАХ ДЛЯ ГЕНОВ, АССОЦИИРОВАННЫХ С ЗАБОЛЕВАНИЯМИ СТАРЕНИЯ

© 2012 г. **О.С. Кожевникова[1], М.К. Мартыщенко[1], М.А. Генаев[1], Е.Е. Корболина[1], Н.А. Муралева[1], Н.Г. Колосова[1, 2], Ю.Л. Орлов[1, 2]**

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

Поступила в редакцию 15 июля 2012 г. Принята к публикации 1 августа 2012 г.

Целью создания базы данных является разработка новых экспериментально-технологических подходов для фундаментальных молекулярно-биологических исследований возрастных заболеваний человека на лабораторных животных (крысах) с использованием микрочиповых технологий оценки экспрессии генов. Несмотря на очевидную связь продолжительности жизни с наследственностью и огромное количество биомедицинских исследований процесса старения, сведения о генетических факторах детерминации процессов преждевременного старения крайне ограниченны. Было установлено, что характерные для старения структурно-функциональные изменения сетчатки аналогичны тем, что происходят на ранних стадиях ВМД и лежат в основе патогенеза этого заболевания, но не всегда приводят к его развитию. В базе данных RatDNA собрана информация о генах, ассоциированных с заболеваниями старения, в частности ВМД, и экспериментальные данные об их экспрессии в различных тканях модельной линии крыс. База доступна по адресу: http://pixie.bionet.nsc.ru/ratdna/rat/index.php.

Ключевые слова: заболевания старения, микрочипы, база данных, крысы линии OXYS.

ВВЕДЕНИЕ

ДНК-чипы

Целью работ был поиск экспериментально-технологических подходов для фундаментальных молекулярно-биологических исследований возрастных заболеваний человека на лабораторных животных (крысах) с использованием микрочиповых технологий оценки экспрессии генов. Кроме комплексного характера общей проблемы изучения заболеваний старения, независимую ценность имеют молекулярно-биологические, технические и биоинформационные методы исследования этой проблемы. Объектом исследования комплексных заболеваний являются молекулярно-генетические системы, координирующие функцию генов, РНК, белков, генных и метаболических путей на различных иерархических уровнях – клеточном, тканевом, органном, организменном. Основой управления такими системами является регуляция работы генов – их экспресссия, т. е. транскрипция и получение белкового продукта. Измерение экспрессии гена может быть выполнено индивидуально с помощью ПЦР в реальном времени или на микрочипах.

ДНК-чип или ДНК-микрочип (DNA microarray) – это комплексная технология, используемая в молекулярной биологии и медицине (Gibson, 2003). ДНК-микрочип может содержать варьирующее число (от десятков до тысяч) микроскопических точек на пластинке (чипе), соответствующих пробам. Каждая точка содержит несколько пикомолей ДНК специфической последовательности (олигонуклеотида).

Олигонуклеотид может быть коротким участком гена или другого компонента ДНК и используется для гибридизации с кДНК (кодирующей ДНК) или мРНК (матричной РНК). В основе технологии ДНК-чипов лежит использование комплементарного связывания нуклеотидов (Katagiri, Glazebrook, 2009).

В целом ДНК-чипы представляют собой уникальный аналитический инструмент, позволяющий определять наличие в анализируемом образце заданных последовательностей ДНК (так называемый гибридизационный анализ). Проведение анализа с помощью ДНК-чипов обходится в несколько раз дешевле, чем при использовании других технологий анализа экспрессии генов (электрофорез, ПЦР в реальном времени), и допускает работу вне лаборатории (Katagiri, Glazebrook, 2009).

Проблемы заболеваний старения

В условиях глобального постарения населения планеты первостепенное значение приобретает поиск путей замедления процессов старения, снижения риска возникновения возрастных заболеваний и увеличения продолжительности жизни. В России на фоне низкой продолжительности жизни имеет место преждевременное старение населения, которое проявляется «омоложением» возрастных заболеваний. В мире ведется огромное количество биомедицинских исследований, направленных на изучение процесса старения во всех возможных проявлениях. Несмотря на очевидную связь продолжительности жизни с наследственностью, сведения о генетических факторах детерминации процессов преждевременного старения крайне ограничены. Очевидно, что выяснение механизмов преждевременного старения, разработка подходов к раннему выявлению его генетически детерминированных предпосылок необходимы для создания эффективных способов профилактики и лечения, обеспечивающих существенное продление периода здоровой жизни человека.

Люди не умирают от «здорового» старения: в любом возрасте причиной их смерти становятся патологии, вероятность развития которых увеличивается с возрастом. Это рак, атеросклероз, гипертония, инсульты, сердечная недостаточность, остеопороз, диабет типа I, нейродегенеративные заболевания и др. Их развитие в более раннем возрасте рассматривается как проявление ускоренного старения, а более позднее становится основой успешного старения – долголетия (Blagosklonny, 2010; Vaupel, 2010). Исследования, цель которых – управление процессами старения, сосредоточены на выяснении молекулярно-генетических основ, с одной стороны, долгожительства (Christensen *et al*., 2009), с другой – преждевременного старения (Maier, Westendorp, 2009). Несомненную актуальность имеет поиск путей выявления предпосылок развития и способов диагностики заболеваний старения уже в молодом возрасте (Sander *et al*., 2008).

Множественность проявлений старения не позволяет четко отделять его причины от эффектов, определять темпы и давать прогноз развития заболеваний. Это связано с тем, что структурно-функциональные изменения на организменном, клеточном и субклеточном уровнях при старении аналогичны изменениям, происходящим на ранних стадиях возрастных заболеваний, лежат в основе, но не всегда приводят к их развитию (Ehrlich *et al*., 2009). Механизмы, запускающие переход обычных возрастных изменений в патологический процесс и лежащие в основе развития большинства ассоциированных со старением заболеваний, до настоящего времени не известны. В значительной степени это обусловлено невозможностью проведения исследований на ранних стадиях заболеваний, которые протекают у людей бессимптомно.

Возрастная макулярная дегенерация

Одно из распространенных заболеваний старения – возрастная макулярная дегенерация (ВМД) – становится основной причиной нарушения и потери зрения у людей старше 50 лет в развитых странах. Количество больных ВМД растет на фоне увеличения продолжительности жизни людей. В России по разным данным ВМД страдают от 14 до 46 % населения в возрасте старше 65 лет (Либман, 2006). Катаракта – одно из самых распространенных заболеваний глаза, но в развитых странах она не становится причиной слепоты. При этом около 60 % всех офтальмологических операций проводятся по

поводу различных форм помутнения хрусталика. На оперативное лечение катаракты в США тратится 12 % средств, затрачиваемых на здравоохранение. Подсчитано, что задержка развития катаракты на 10 лет уменьшает потребность в операции вдвое.

Накапливаются данные, свидетельствующие о том, что характерные для старения структурно-функциональные изменения сетчатки аналогичны тем, что происходят на ранних стадиях ВМД и лежат в основе патогенеза этого заболевания, но не всегда приводят к его развитию (Smith, Steinle, 2007). Как одно из неизбежных проявлений старения, нередко сопровождающих течение ВМД, многими исследователями рассматривается катаракта (Yoshida *et al*., 2002). Значительное количество вовлеченных генов и симптомы, характерные и для других комплексных заболеваний, усложняют диагностику ВМД.

Использование модели лабораторных животных для анализа функций генов, связанных со старением

Создание моделей возрастных заболеваний затруднено, поскольку они развиваются у людей зачастую одновременно на фоне комплексных проявлений старения, возраст манифестации которых, как и «набор» самих заболеваний, существенно различаются. В мире по-прежнему остается одна общепризнанная модель преждевременного старения: созданная японскими учеными линия мышей SAM (senescence accelerated mouse), которая представлена сегодня 12 сублиниями, различающимися проявлениями старения и характерными для него заболеваниями (Takeda, 2009). Потребность в моделях преждевременного старения с комплексной манифестацией его признаков растет. В этом убеждает растущая востребованность мышей SAM: количество выполненных с их использованием работ в последние три года резко возросло. Как показали наши исследования, комплексное проявление признаков преждевременного старения отличает и созданную в ИЦиГ СО РАН линию крыс OXYS. Линия была создана в 70-е годы прошлого века отбором крыс Вистар по признаку ранней спонтанной катаракты. В пяти первых поколениях развитие катаракты провоцировали нагрузкой галактозой, в дальнейшем проводился отбор по ранней спонтанной катаракте, сцеплено с которой животные унаследовали синдром преждевременного старения. Помимо катаракты он проявляется снижением максимальной продолжительности жизни и ранним развитием ассоциированных со старением заболеваний: ретинопатии, остеопороза, артериальной гипертензии (Колосова и др., 2003; Bobko *et al*., 2005; Muraleva *et al*., 2010), ускоренной инволюцией тимуса (Obukhova *et al*., 2009) и проявлений преждевременного старения мозга, в том числе нейродегенеративных изменений. Доказано, что линия крыс OXYS соответствует основным критериям модели таких возрастных заболеваний человека, как сенильная катаракта, остеопороз и возрастная макулярная дегенерация (Muraleva *et al*., 2011; Жданкина и др., 2008; Markovets *et al*., 2011; Korbolina *et al*., 2012). Проявления этих заболеваний у крыс OXYS воспроизводятся на клиническом и морфологическом уровнях, они отвечают на стандартную терапию.

Фенотипическим проявлениям катаракты и ВМД предшествуют изменения экспрессии генов, однако вклад изменений транскриптома в процесс нормального физиологичного старения и тем более в развитие этих заболеваний, особенно на ранних стадиях, остается не ясным в силу сложности проведения таких исследований на людях. Существуют единичные работы, авторы которых исследовали изменения профиля экспрессии генов в сетчатке при развитии ВМД (Booij *et al*., 2009; Kurji *et al*., 2010). Большинство исследователей работает с культурами различных клеток сетчатки. Это существенно ограничивает возможности интерпретации результатов и переноса их на уровень организма.

Исследования транскриптома проводятся и на моделях ВМД – на животных, развитие ретинопатии у которых вызывают, как правило, воздействием различных физических факторов (УФ- или лазерным излучением, гипероксией), которое только частично воспроизводит картину развития ВМД (Ishikawa *et al*., 2010). Систематическое исследование раннего развития заболевания невозможно на пациентах на доклинических стадиях, что обусловливает необходимость испытаний на лабораторных животных.

Компьютерные подходы. Портал проекта

Для систематизации информации об экспрессии генов, связанных с ВМД, в рамках работ по технологической платформе «Медицина будущего» в ИЦиГ СО РАН был разработан ДНК-чип для исследований экспрессии генов крыс, созданы база данных экспрессии генов и Web-портал с ассоциированной информацией по данной проблеме.

Портал разработан на PHP, база данных RatDNA – на MySQL. Сервер MySQL управляет доступом к данным, позволяя работать с ними одновременно нескольким пользователям, обеспечивает быстрый доступ к данным (Веллинг, Томсон, 2008). Разработанные базы данных (БД), функциональное описание генов и информация о проекте в целом доступны по адресу http://pixie.bionet.nsc.ru/ratdna/rat/index.php.

Стартовой страницей портала является страница с информацией по проекту исследования заболеваний старения на лабораторных животных –http://pixie.bionet.nsc.ru/ratdna/index.php. С главной страницы можно совершить переход в соответствующие разделы: «Общая информация о проекте», «Этапы», «Результаты», «Литературные источники», «Рабочий сайт проекта» (рис. 1, а, б).

База данных генов крысы для микрочипа

При переходе в раздел базы данных RatDNA (фрагмент интерфейса представлен на рис. 2) в навигационном меню доступен раздел Help, в котором представлена справочная информация по данной таблице. По таблице можно осуществить поиск, введя название искомого гена в поле поиска. Помимо возможности просмотра и работы с таблицами на сайте, они доступны для загрузки.

Методы поиска генов включали процессинг данных экспрессии генов на микрочипах в

Рис. 1. Структура Web-портала.

а – стартовая страница (верхняя панель), б – переход на вкладку «Результаты» (нижняя панель).

Главная | RatDNA | Help | RatDNA-OMIM | RatDNA-AMD | ЦКП "Биоинформатика" | Функциональное описание селектированных генов

Набор генов крысы, связанных с заболеваниями старения

<Скачать таблицу>

№	Идентификатор RGD_ID	Символ гена	Описание гена	Хромосома	Начало	Конец	ID транскрипта	Ориентация гена в геноме	Число экзонов	
1	68358	Acan	aggrecan	1	134787341	134848992	NM_022190	+	18	CCAACACCTACAAGCA
2	1305051	Aen	apoptosis enhancing nuclease	1	134615998	134625367	NM_001108487	+	4	agtgtactgtgagaaatcagcttgtttg
3	619885	Ak3	adenylate kinase 3	1	232658879	232684083	NM_013218	-	5	TTTCCTAAGACTTCTCT
4	620844	Apbal	amyloid beta (A4) precursor protein-binding, family A, member 1	1	227106828	227309416	NM_031779	+	13	ATAACCACTGGCAGGTA
5	620845	Apba2	amyloid beta (A4) precursor protein-binding, family A, member 2	1	118970882	119156605	NM_031780	+	14	ATGTATAATGATGACCT
6	2122	Apbb1	amyloid beta (A4) precursor protein-binding, family A, member 1 (Fe65)	1	163282918	163299333	NM_080478	-	13	tgtttgaggtggagcaggaggaactg
7	628763	Aqp11	aquaporin 11	1	154973796	154983962	NM_173105	-	3	TTGTTCTTTTGAGTGAT

Рис. 2. Фрагмент интерфейса. Таблица генов крысы и олигонуклеотидных проб «RatDNA-chip».

тканях сетчатки глаз, опубликованных в литературе. Анализ баз данных и литературных источников и QTL-анализ (Korbolina *et al*., 2012) позволили установить несколько списков генов, дифференциально экспрессирующихся в сетчатке глаза, которые были использованы при дизайне специализированного ДНК-чипа. Было отобрано 113 генов, подобраны олигонуклеотидные зонды. База данных RatDNA содержит информацию об этом наборе генов микрочипа (рис. 2).

На странице портала представлена таблица RatDNA-AMD (рис. 3). В ней собрана информация по генам крысы, связанным с возрастной макулярной дегенерацией (ВМД, – англ. AMD). По таблице также можно осуществить поиск, введя название искомого гена в поле поиска.

Данные по экспрессии генов в ткани сетчатки глаза крыс, полученные с помощью специализированного микрочипа, разработанного в ИЦиГ СО РАН, находятся на странице «RatDNA-Экспрессия генов».

Объектами в базе данных являются гены крысы, их нуклеотидные последовательности и функциональная аннотация. База данных в целом включает 5 таблиц (рис. 4):

1) таблица генов крысы RatDNA-chip предназначена для описания генов и олигонуклеотидных проб для микрочипа;

2) таблица генов крысы, гомологичных генам человека, связанным с наследственными заболеваниями человека «RatDNA-OMIM», предназначена для исследования ассоциаций заболеваний старения на крысах с аналогичными заболеваниями человека;

3) таблица генов крысы и соответствующих генов человека, ассоциированных с возрастной макулярной дегенерацией, «RatDNA-AMD», построена на основе анализа литературных данных и предназначена для последующего изучения экспрессии генов на ДНК-чипах и полногеномных данных транскриптомного секвенирования;

4) таблица «Группа генов» содержит списки генов, селектированных по дифференциальной экспрессии в тканях крысы, она построена в результате анализа экспериментальных данных микрочипов и является производной для анализа генных онтологий;

5) таблица «Экспрессия» содержит экспериментальные данные, полученные с помощью специализированного ДНК-чипа по генам крысы из таблицы RatDNA-chip.

Связи между таблицами БД RatDNA осуществляются по идентификатору гена крысы.

<Скачать таблицу>

RGD_ID	Название гена	Описание	Хромосома	Начало	Конец	Transcript_ID	Статья	PubMed_ID
1564153	Plekha1	pleckstrin homology domain containing, family A (phosphoinositide binding specific) member 1	1	190187202	190238309		Conley YP, Jakobsdottir J, Mah T, et al. CFH, ELOVL4, PLEKHA1, and LOC387715 genes and susceptibility to age-related maculopathy: AREDS and CHS cohorts and meta-analyses. Hum Mol Genet 2006;15:3206–3218. PubMed: 17000705	PubMed:17000705
2138	Apoe	apolipoprotein E, plays a role in plasma lipoprotein transport	1	79003634	79006387		Baird PN, Richardson AJ, Robman LD, Dimitrov PN, Tikellis G, et al. 2006. Apolipoprotein (APOE) gene is associated with progression of age-related mascular degeneration (AND). Hum. Mutant. 27:337–42.	PMID:16453339
3963	Vldlr	very low density lipoprotein reseptor, encodes a protein exhibiting protein tyrosine kinase activator activity, very-low-density lipoprotein receptor activity (human ortholog), protein binding (Dab1; mouse ortholog) and other functions (inferred); involve	1	230666736	230697748		Haines JL, Schnetz-Boutaud N. Functional candidate genes in age-related macular degeneration: significant association with VEGF, VLDLR, and LRP6.Invest Ophthalmol Vis Sci. 2006 Jan;47(1):329-35	PMID:16384981

Рис. 3. Фрагмент интерфейса. База данных генов человека «RatDNA-AMD» (и их гомологов), связанных заболеванием ВМД.

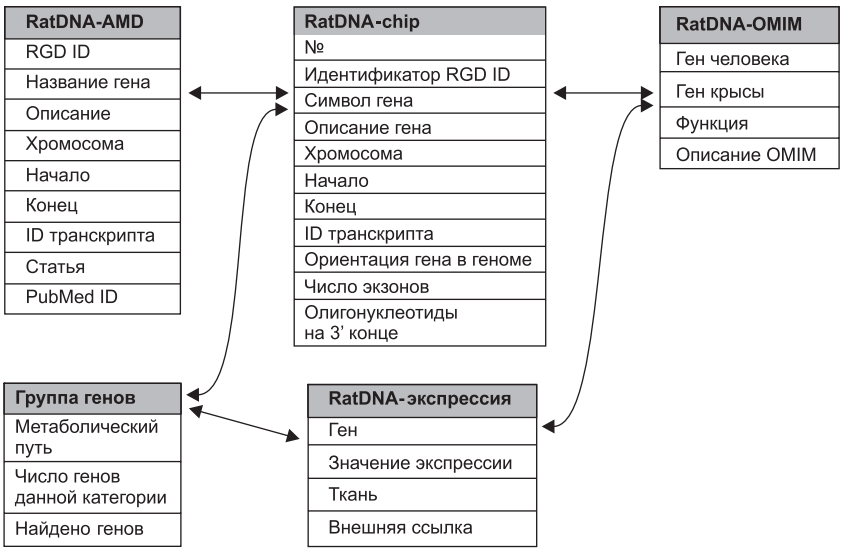

Рис. 4. Структура базы данных RatDNA и связь таблиц.

Исследование функций генов крысы, представленных в базе данных

Выбор генов, ассоциированных с заболеваниями старения, выполнялся по литературным данным, представленным в базах данных GEO NCBI (http://www.ncbi.nlm.nih.gov/gds) и OMIM (Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/omim). При сравнении данных OMIM по генам человека и генам крысы использовалось соответствие идентификаторов (например, HIF1A у человека и Hif1a у крысы), при этом использовались публикации, содержащие данные по экспрессии генов в тканях крысы, и БД Retinobase (Kalathur *et al*., 2008).

Для анализа функций выбранных генов, относящихся к заболеваниям старения человека, было выполнено сравнение списка генов крысы с генами человека, описанными в базе данных наследственных заболеваний OMIM. Было выделено 254 категории, относящиеся к старению. По названиям генов было установлено соответ-

Таблица

Соответствие найденных генов крысы и генов человека, связанных с заболеваниями старения и продолжительностью жизни (по базе данных OMIM)

Ген человека	Ген крысы	Функция	Описание OMIM
ARNTL	Arntl	Связан с циркадными ритмами, экспрессируется в ретине у мыши	*602550. ARYL HYDROCARBON RECEPTOR NUCLEAR TRANSLOCATOR-LIKE
BAD	Bad	Регуляция апоптоза – программируемой клеточной смерти	*603167. BCL2 ANTAGONIST OF CELL DEATH
BCL2	Bcl2	Онкоген	+151430. B-CELL CLL/LYMPHOMA 2
COQ7	Coq7	Регуляция базовых метаболических процессов, включая биосинтез, дыхание, продолжительность жизни у *C. elegans*	*601683. COQ7, S. CEREVISIAE, HOMOLOG OF
HIF1A	Hif1a	Фактор ответа на гипоксии	*603348. HYPOXIA-INDUCIBLE FACTOR 1, ALPHA SUBUNIT
IGF1R	Igf1r	Рецептор ростового фактора	*147370. INSULIN-LIKE GROWTH FACTOR I RECEPTOR
POLG	Polg	Комплекс транскрипции	*174763. POLYMERASE, DNA, GAMMA
SIRT3	Sirt3	Митохондриальная деацитилаза. Семейство белков-сиртуинов	*604481. SIRTUIN 3
TPH1	Tph1	Триптофан гидроксилаза, биосинтез серотонина	*191060. TRYPTOPHAN HYDROXYLASE 1

ствие, найдены гены, связанные с оксидативным стрессом, например Hif1a.

Данные соответствия с OMIM представлены в табл. («RatDNA-OMIM»).

Был проведен анализ функций генов крысы, представленных в таблице RatDNA-chip с помощью категорий генных онтологий. Для проанализированных генов было установлено соответствие 98 идентификаторов геномной аннотации RefSeq. С помощью ресурса анализа генных онтологий PANTHER была выполнена оценка обогащенности данной группы генов категориями, относящимися к метаболическим путям, молекулярным функциям и биологическим процессам. Результаты для метаболических путей также представлены в таблице на странице БД RatDNA.

Интересно отметить присутствие категорий, связанных с оксидативным стрессом (Hypoxia response), передачей сигнала FGF (FGF signaling pathway), а также метаболическими путями белков, вовлеченных в болезнь Альцгеймера и болезнь Паркинсона, развивающиеся с возрастом. Категории генных онтологий для генов, отобранных по данным экспрессии на микрочипах в геноме крысы, подтверждают литературные данные об ассоциациях с заболеваниями старения.

Набор генов был протестирован на предмет выявления регуляторных и белок-белковых взаимодействий по базе данных STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) (http://string-db.org/). Было выявлено большое число взаимодействий между белками исследуемой группы, реконструирована генная сеть. Выявлено несколько узлов сети (не менее 4 контактов). Такими узлами являются Bcl2, Bax, Timp3, Nos3, Hif1a, Igfr1, Fgdr2, Epas1, Usp3.

Многие из этих генов человека относятся к генам, связанным с заболеваниями старения, согласно базе данных наследственности человека OMIM. Так, BCL2 (B-cell Cell/Lymphoma 2) – это известный онкоген; HIF1A (Hypoxia-Inducible Factor 1, Alpha subunit) – фактор ответа на гипоксию (недостаток кислорода), Igfr1 (Insulin-Like Growth Factor I Receptor) – ростовой фактор.

Сравнение экспрессии изучаемых генов на микрочипе для крыс исследуемой линии OXYS и контрольной Вистар (Wistar) показало разли-

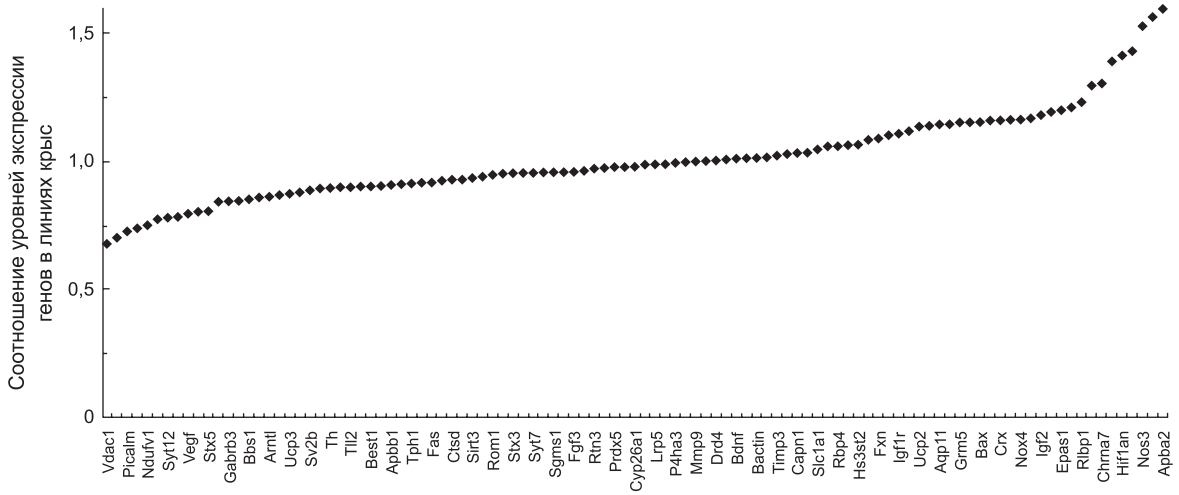

Рис. 5. Соотношение уровней экспрессии генов OXYS/ Wistar на разработанном чипе.

Ось абсцисс – гены крысы на микрочипе, ось ординат – нормализованное соотношение уровней экспрессии в тканях линий крыс.

чие уровней экспрессии в 1,3–1,4 раза (по 4 репликатам), что является достаточно небольшим диапазоном (рис. 5).

Наименьшие значения соотношения уровней экспрессии в исследуемых группах (понижение уровня экспрессии) имеют гены Vdac1 (в 0,68 раза), Cdc37l1 (0,7), Picalm (0,73); наибольшие соотношения – гены Nos3 (1,53 раза), Fgf4 (1,56) и Apba2 (1,6). Данные этих экспериментальных измерений также представлены в БД RatDNA.

ЗАКЛЮЧЕНИЕ

Таким образом, разработана база данных генов крысы, ассоциированных с заболеваниями старения, и создан веб-портал для продолжения исследований. Записи базы данных содержат экспериментальную информацию об уровнях экспрессии генов на специализированном ДНК-чипе. Разработанная интернет-доступная база данных предназначена для хранения информации о селектированных генах крысы, связанных с заболеваниями старения, их функциональной аннотации, включая олигонуклеотидные пробы для измерения экспрессии генов на микрочипах и экспериментально определенные значения экспрессии этих генов. БД интегрирует геномные данные с функциональной аннотацией генов и их связи с заболеваниями старения, включая возрастную макулярную дегенерацию, с экспериментальными данными об экспрессии этих генов в тканях лабораторных животных – крыс. Использование ссылочных таблиц на экспериментальные данные, полученные с помощью микрочипов, позволяет соотносить экспрессию генов крысы с их ролью в заболеваниях старения человека и других организмов (модельных животных), дает возможность расширения базы данных на исследования по крысам с помощью других технологий (Shevelev *et al.*, 2012).

Результаты непосредственного измерения содержания белка Aβ (амилоид β) в сетчатке на разных стадиях развития дегенеративных изменений подтверждают выявленную с помощью целевого ДНК-микрочипа общность механизмов патогенеза болезни Альцгеймера и ретинопатии у крыс OXYS и, можно полагать, ВМД у людей. Следует подчеркнуть принципиально важный момент: изменения экспрессии генов пути болезни Альцгеймера выявлены на ранних стадиях развития признаков преждевременного старения и предшествовали усиленному накоплению Aβ в сетчатке и гиппокампе, формированию в них выраженных нейродегенеративных изменений.

Исследовать у людей ранние доклинические стадии заболеваний невозможно, так же, как получать образцы сетчатки и мозга пациентов для масштабных исследований молекулярно-

генетических механизмов патогенеза ВМД и нейродегенеративных изменений мозга. Наши исследования показали, что целевые ДНК-чипы целесообразно создавать для фундаментальных исследований на биологических моделях заболеваний человека, для выяснения их этиологии и патогенеза, поиска новых мишеней для патогенетически обоснованных терапевтических воздействий на них. При этом использование биоинформационных технологий и баз данных при отборе генов-кандидатов для целевых ДНК-чипов существенно повышает эффективность работы (Yang *et al*., 2011).

БЛАГОДАРНОСТИ

Авторы благодарны С.И. Татькову, А.А. Швалову, П.С. Деменкову за помощь и научное обсуждение. Работа поддержана госконтрактом Минобрнауки РФ № 16.513.11.3107. Тестирование и установка базы данных проводились с использованием оборудования суперкомпьютерного кластера ССКЦ СО РАН, ЦКП «Биоинформатика» СО РАН.

ЛИТЕРАТУРА

Веллинг Л., Томсон Л. Разработка Web-приложений с помощью PHP и MySQL. 3-е изд. М.: Издат. дом «Вильямс», 2008. 880 с.

Жданкина А.А., Фурсова А.Ж., Логвинов С.В., Колосова Н.Г. Клинико-морфологические особенности хориоретиальной дегенерации у преждевременно стареющих крыс линии OXYS // Бюл. эксперим. биол. и медицины. 2008. V. 146. № 10. P. 435–438.

Колосова Н., Лебедев П., Фурсова А. и др. Преждевременно стареющие крысы OXYS как модель сенильной катаракты человека // Усп. геронтологии. 2003. Т. 12. С. 143–148.

Либман Е.С., Толмачев Р.А., Шахова Е.В. Эпидемиологическая характеристика инвалидности вследствие основных форм макулопатий // Матер. II Всерос. семинара «Макула-2006» / Под ред. Ю.А. Иванишко, Ростов-на-Дону, 2006. С. 15–22.

Blagosklonny M.V. Why human lifespan is rapidly increasing: solving «longevity riddle» with «revealed-slow-aging» hypothesis // Aging (Albany NY). 2010. V. 2. No. 4. P. 177–182.

Bobko A., Sergeeva S., Bagryanskaya E. *et al*. 19F NMR measurements of NO production in hypertensive ISIAH and OXYS rats // Biochem. Biophys. Res. Commun. 2005. V. 330. No. 2. P. 367–370.

Booij J.C., van Soest S., Swagemakers S.M. *et al*. Functional annotation of the human retinal pigment epithelium transcriptome // BMC Genomics. 2009. V. 20. No. 10. 164.

Christensen K., Doblhammer G., Rau R., Vaupel J.W. Ageing populations: the challenges ahead // Lancet. 2009. V. 374. No. 9696. P. 1196–1208.

Ehrlich R., Kheradiya N.S., Winston *et al*. Age-related ocular vascular changes // Graefes Arch. Clin. Exp. Ophthalmol. 2009. V. 247. No. 5. P. 583–591.

Gibson G. Microarray analysis: genome-scale hypothesis scanning // PLoS Biol. 2003. V. 1. No. 1. E15.

Ishikawa K., Yoshida S., Kadota K. *et al*. Gene expression profile of hyperoxic and hypoxic retinas in a mouse model of oxygen-induced retinopathy // Invest. Ophthalmol. Vis. Sci. 2010. V. 51. No. 8. P. 4307–4319.

Kalathur R.K., Gagniere N., Berthommier G. *et al*. RETINOBASE: a Web database, data mining and analysis platform for gene expression data on retina // BMC Genomics. 2008. V. 9. 208.

Katagiri F., Glazebrook J. Pattern discovery in expression profiling data // Curr. Protoc. Mol. Biol. 2009. Chapter 22:Unit 22.5.

Korbolina E.E., Kozhevnikova O.S., Stefanova N.A., Kolosova N.G. Quantitative trait loci on chromosome 1 for cataract and AMD-like retinopathy in senescence-accelerated OXYS rats // Aging (Albany NY). 2012. V. 4. No. 1. P. 49–59.

Kurji K.H., Cui J.Z., Lin T. *et al*. Microarray analysis identifies changes in inflammatory gene expression in response to amyloid-beta stimulation of cultured human retinal pigment epithelial cells // Invest. Ophthalmol. Vis. Sci. 2010. V. 51. No. 2. P. 1151–1163.

Maier A.B., Westendorp R.G. Relation between replicative senescence of human fibroblasts and life history characteristics // Ageing Res Rev. 2009. V. 8. No. 3. P. 237–243.

Markovets A.M., Fursova A.Z., Kolosova N.G. Therapeutic action of the mitochondria-targeted antioxidant SkQ1 on retinopathy in OXYS rats linked with improvement of VEGF and PEDF gene expression // PLoS One. 2011. V. 6. No. 7. e21682.

Muraleva N.A., Sadovoĭ M.A., Kolosova N.G. Effect of alendronate on bone tissue status of senescence-accelerated OXYS rats // Adv. Gerontol. 2011. V. 24. No. 1. P. 143–146.

Muraleva N.A., Sadovoĭ M.A., Kolosova N.G. Development of osteoporosis in prematurely aging OXYS rats // Adv. Gerontol. 2010. V. 23. No. 2. P. 233–242.

Obukhova L.A., Skulachev V.P., Kolosova N.G. Mitochondria-targeted antioxidant SkQ1 inhibits age-dependent involution of the thymus in normal and senescence-prone rats // Aging (Albany NY). 2009. V. 1. No. 4. P. 389–401.

Sander M., Avlund K., Lauritzen M. *et al*. Aging-from molecules to populations // Mech. Ageing Dev. 2008. V. 129. No. 10. P. 614–623.

Shevelev O.B., Rykova V.I., Fedoseeva L.A. *et al*. Expression of Ext1, Ext2, and heparanase genes in brain of senescence-accelerated OXYS rats in early ontogenesis and during development of neurodegenerative changes // Biochemistry (Mosc). 2012. V. 77. No. 1. P. 56–61.

Smith C.P., Steinle J.J. Changes in growth factor expression in normal aging of the rat retina // Exp. Eye Res. 2007. V. 85. No. 6. P. 817–824.

Takeda T. Senescence-accelerated mouse (SAM) with special references to neurodegeneration models, SAMP8 and

SAMP10 mice // Neurochem. Res. 2009. V. 34. No. 4. P. 639–659.
Vaupel J.W. Biodemography of human ageing // Nature. 2010. V. 464. No. 7288. P. 536–542.
Yang L., Nie Y.H., Zhou L.H. *et al*. Microarray profiles on age-related genes in the earlier postnatal rat visual cortex // Chin. Med. J. (Engl). 2011. V. 124. No. 10. P. 1545–1550.
Yoshida S., Yashar B.M., Hiriyanna S., Swaroop A. Microarray analysis of gene expression in the aging human retina // Invest. Ophthalmol. Vis. Sci. 2002. V. 43. No. 8. P. 2554–2560.

RatDNA: DATABASE ON MICROARRAY STUDIES OF RATS BEARING GENES ASSOCIATED WITH AGE-RELATED DISEASES

O.S. Kozhevnikova[1], M.K. Martyschenko[1], M.A. Genaev[1], E.E. Korbolina[1], N.A. Muraleva[1], N.G. Kolosova[1, 2], Y.L. Orlov[1, 2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: oidopova@bionet.nsc.ru;
[2] Novosibirsk National Research State University, Novosibirsk, Russia

Summary

The purpose of the RatDNA database is the development of experimental methods for basic molecular studies of human age-related diseases in rats involving microarray tests of gene expression. Despite the obvious correlation between life expectancy and heredity and numerous biomedical studies on aging, little is known about genetic factors determining aging processes. People do not die of «healthy» aging: at any age conditions whose probability increases with age become the cause of their death. Age-related macular degeneration (AMD) becomes the main cause of vision problems and sight loss in people aged above 50. Structural and functional changes in the retina characteristic of aging are similar to those observed at early stages of AMD. They underlie the pathogenesis of this disease, but not always lead to its development. RatDNA database contains information on genes associated with age-related diseases, in particular AMD, and experimental data about their expression in tissues of a model rat strain. The database is available at http://pixie.bionet.nsc.ru/ratdna/rat/index.php.

Key words: age-related diseases, microarray, database, OXYS rats.

УДК 004.65

ИНФОРМАЦИОННАЯ ПОДДЕРЖКА ЭКСПЕРИМЕНТОВ ПО ТРАНСГЕНЕЗУ РАСТЕНИЙ В БАЗЕ ДАННЫХ ТРАНСЛЯЦИОННЫХ ЭНХАНСЕРОВ

© 2012 г. **О.Г. Смирнова, Д.А. Рассказов, А.В. Кочетов**

Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: ak@bionet.nsc.ru

Поступила в редакцию 15 июля 2012 г. Принята к публикации 31 августа 2012 г.

Разработана база данных для подбора трансляционных энхансеров, обеспечивающих дополнительный контроль экспрессии чужеродного гена в растениях на уровне трансляции мРНК. База данных содержит структурированную информацию о локализованных в мРНК трансляционных энхансерах, которые контролируют экспрессию генов на посттранскрипционном уровне. Эта информация полезна для планирования генно-инженерных экспериментов. Использование платформы и интерфейса Sequence Retrieval System позволяет проводить быстрый поиск трансляционных энхансеров с определенными характеристиками и получать нуклеотидные последовательности выбранных энхансеров. База данных доступна по адресу http://wwwmgs.bionet.nsc.ru/mgs/dbases/trsig/.

Ключевые слова: база данных, информационный ресурс, трансляционный энхансер, генетическая инженерия, трансгенные растения.

ВВЕДЕНИЕ

Генетическая конструкция, которая используется для переноса чужеродного генетического материала при трансгенезе, помимо целевого гена, содержит дополнительные регуляторные последовательности, обеспечивающие транскрипцию и трансляцию этого гена. Для обеспечения необходимого паттерна экспрессии перенесенного гена необходимо использовать адекватные регуляторные последовательности и сигналы экспрессии. Дизайн генетической конструкции включает ряд последовательных этапов и требует эффективного планирования. Помимо сигналов транскрипционного контроля, большое значение для эффективного проведения научно-исследовательских работ в этой области может иметь оптимизация экспрессии трансгена на посттранскрипционном уровне. В составе эукариотических мРНК часто содержатся сигналы, контролирующие эффективность трансляции или цитоплазматическую стабильность матрицы. Применение таких сигналов в дизайне трансгена может существенно увеличить эффективность трансгенеза. Обычно генетические конструкции не содержат сайтов сплайсинга, и в них используются аутентичные сигналы полиаденилирования, взятые из генов организма-реципиента, поэтому с этой фазой экспрессии трансгена возникает меньше проблем. Таким образом, практически важным становится адекватное планирование эффективности трансляции мРНК.

Одним из способов управления экспрессией гена являются трансляционные сигналы экспрессии, обычно расположенные в составе 5′- или 3′-нетранслируемых районов мРНК (Liu *et al.*, 2009). Некоторые из таких сигналов могут определять общую трансляционную активность мРНК: например, если в составе генетической конструкции используется 5′-НТП мРНК вируса табачной мозаики (размером 68 нуклеотидов), то это в большинстве случаев увеличивает уровень синтеза белкового продукта в несколько раз (Gallie, 2002). Этот энхансер активно используется в биотехнологии растений. Кроме неспецифических усилителей трансляционной активности мРНК, известны специфические

сигналы (например, IRE, регулирующий трансляцию мРНК гена ферритина в зависимости от присутствия железа; с этим сигналом связывается специфический белок IRP, конформация и активность которого зависят от присутствия железа, что функционально эквивалентно ситуации транскрипционного контроля и сайтам связывания транскрипционных факторов в промоторах). Однако, в отличие от сайтов связывания транскрипционных факторов, сигналы трансляционного контроля экспрессии мало изучены, что связано со специфическими особенностями одноцепочечных молекул РНК, способных в цитоплазме существовать в виде различных конформеров. Большинство сигналов такого типа представляют собой комбинацию контекстных и структурных элементов, что делает их предсказание очень сложным.

Информационные ресурсы в этой области представлены базами данных (БД) UTRsite (Grillo *et al.*, 2010) и TransTerm (Jacobs *et al.*, 2009). В этих БД аннотировано крайне малое количество (всего несколько десятков) трансляционных сигналов разной таксономической принадлежности (по некоторым оценкам сигналы этого типа могут присутствовать у 10–15 % мРНК генов эукариот). Причина такой недостаточной аннотации заключается в том, что в этих БД рассматриваются только те трансляционные сигналы, у которых точно известна тонкая структура (контекстные и конформационные элементы сигнала). При этом для планирования биотехнологических опытов, как правило, достаточно знать, как изменяется паттерн трансляционной активности мРНК трансгена в присутствии тех или иных участков мРНК с известной трансляционной активностью. Таким образом, представленная в существующих базах данных по трансляционным сигналам информация, с одной стороны, недостаточна, а с другой – чрезмерно детализирована для целей планирования трансгенетических экспериментов. В то же время следует отметить, что экспериментальной информации такого рода в литературе достаточно много, что говорит о возможности создания специализированного информационного ресурса.

В настоящей работе описывается информационный ресурс БДТЭ (База данных трансляционных энхансеров), который разработан нами для решения задачи сбора, хранения и систематизации информации о трансляционных энхансерах. Представленная информация и удобный интерфейс позволяют осуществлять выбор такого важного элемента генетической конструкции, как энхансер трансляции, способствуя тем самым эффективному созданию форм растений с улучшенными и качественно новыми свойствами.

Формат и логическая структура БД трансляционных энхансеров

Согласно проведенному нами анализу литературных данных, специализированные БД трансляционных энхансеров для планирования генно-инженерных экспериментов отсутствуют. Имеющиеся аналоги приспособлены для решения более широкого круга задач (в основном фундаментального характера) и не могут быть эффективно использованы для этой цели. С нашей точки зрения, основным недостатком существующих информационных ресурсов является тот факт, что процедура аннотации энхансеров трансляции на основе анализа литературных данных в них высокоизбирательна, т. е. необходимы детальные знания об их тонкой организации.

Однако для решения биотехнологических задач эта информация не нужна. Для выбора трансляционного энхансера в качестве элемента биотехнологической генетической конструкции необходимо и достаточно знать, какой уровень и паттерн трансляции мРНК репортерного гена обеспечивает определенная нуклеотидная последовательность (прототип энхансера) в генетически модифицированном организме.

База данных трансляционных энхансеров позволяет специалисту в области генной инженерии выбирать потенциальные энхансеры по следующим полям: 1) организм – донор энхенсера; 2) организм – реципиент энхансера (в котором была оценена его экспрессия); 3) паттерн трансляционной активности (тканe-, органо-, стадиеспецифичность наработки белка-репортера); 4) уровень трансляционной активности мРНК гена репортера, содержащей данный энхансер.

БДТЭ включает в себя две составляющие:

1) таблицу объектов (TRANSIG_OBJ), содержащую информацию о прототипе исследуемого объекта (обычно – природном варианте

трансляционного энхансера) и о его свойствах, полученную при анализе научных статей. Эта информация помогает пользователю выбрать тип регуляторного сигнала с нужными свойствами;

2) таблицу энхансеров (TRANSIG_ENH), содержащую информацию об энхансерах, полученную при анализе научных статей. Под энхансером понимаются нуклеотидные последовательности, обычно представляющие собой модифицированный вариант объекта (прототипа, природного варианта), содержащий делеции или инсерции, активность которого была описана в аннотированной статье. Характерной особенностью энхансера является привязка к конкретным экспериментальным данным (нуклеотидная последовательность, вид растений, в котором была проведена проверка свойств этой нуклеотидной последовательности, репортерный ген и т. п.). Каждому прототипу (объекту) может соответствовать несколько вариантов энхансеров (экспериментальных последовательностей, как минимум, один вариант).

Структура БДТЭ в графическом виде представлена на рис. 1.

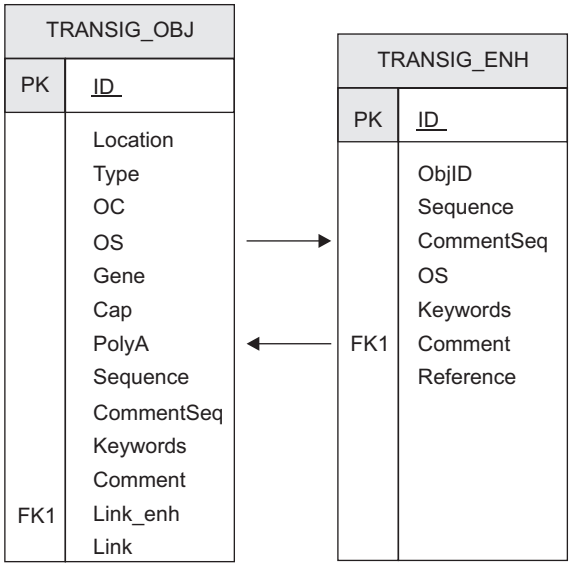

Рис. 1. Структура базы данных ТЭ.

Описание таблицы TRANSIG_OBJ включает 15 полей (табл. 1). Описание таблицы TRANSIG_ENH включает 9 полей (табл. 2). В табл. 3 и 4 приведены примеры записей в информационных таблицах TRANSIG_OBJ

Таблица 1

Структура записи в таблице TRANSIG_OBJ (+ обозначает индексируемое SRS поле, по которому пользователь может осуществлять поиск информации)

ID	Идентификатор записи TRANSIG_OBJ (+)
LOCATION	Расположение энхансера (5′UTR, 3′UTR, CDS) (+)
TYPE	Тип энхансера (stress-specific и др.) (+)
OC	Таксономическая классификация (+)
OS	Название вида (+)
GENE	Название гена (+)
CAP	Наличие кепа на 5′-конце мРНК (+)
POLYA	Наличие поли(А)-участка на 3′-конце мРНК (+)
SQ	Собственно нуклеотидная последовательность энхансера
COMMENTSEQ	Комментарий о происхождении нуклеотидной последовательности и ее расположении в составе генетических конструкций (+)
KEYWORD	Ключевые слова (+)
COMMENT	Развернутый комментарий о специфичности и активности энхансера, эффективности его использования в различных видах организмов реципиентов (+)
LINK_ENH	Ссылка на идентификатор записи таблицы TRANSIG_ENH (+)
LINK	Ссылка на банк данных нуклеотидных последовательностей (+)

Таблица 2

Структура записи в таблице TRANSIG_ENH (+ обозначает индексируемое SRS поле, по которому пользователь может осуществлять поиск информации)

ID	Идентификатор записи TRANSIG_ENH (+)
OBJID	Ссылка на идентификатор записи таблицы TRANSIG_OBJ (+)
SEQUENCE	Нуклеотидная последовательность функционального района (5′-НТП, 3′-НТП), содержащего энхансер
COMMENTSEQ	Комментарий к структуре экспериментальной конструкции (+)
ORGANISM	Видовое название организма, на котором проводили эксперименты (+)
KEYWORD	Ключевые слова (+)
COMMENT	Развернутый комментарий о специфичности и активности энхансера, эффективности его использования в различных видах организмов реципиентов (+)
REFERENCE	Название статьи и ссылка на БД PubMed (+)

Таблица 3

Пример заполнения записи в таблице TRANSIG_OBJ

ID	ADHZM5
LOCATION	5′UTR
TYPE	Stress-specific enhancer
OC	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; euphyllophytes; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; Zea
OS	Zea mays
GENE	ADH1, alcoholdehydrogenase I
CAP	Capped
POLYA	Polyadenylated
SQ	ATTTTCTCGCTCCTCACAGGCTCATCTCGTTTGGATCGATTG GTTTCGTAACTGGTGAAGGACTGAGGGTCTCGGAGTGGATCG ATTTGGGATTCTGTTCGAAGATTTGCGGAGGGGGGCA
COMMENTSEQ	5′UTR of ADH1 gene mRNA
KEYWORD	Enhancer, hypoxia, anoxia, anaerobiosis, stress
COMMENT	It was found that translation of alcoholdehydrogenase mRNA was efficient under oxygen deprivation conditions whereas translation of many other mRNAs was stopped. No changes in mRNA stability were detected so the effect observed could result from the changes in stress-specific translation rate. Deletions of ADH 5’UTR decreased stress-specific translatability: the influence of possible changes in secondary structure was not tested or discussed…
LINK_ENH	ADHZM5a
LINK	EMBL_AC X00580

Таблица 4

Пример заполнения записи в таблице TRANSIG_ENH

ID	ADHZM5a
OBJID	ADHZM5
SEQUENCE	ATAGGGAGACCGAATTCGAGCTCATTTTCTCGCTCCTCACAGGCTCATCT CGTTTGGATCGATTGGTTTCGTAACTGGTGAAGGACTGAGGGTCTCGGAG TGGATCGATTTGGGATTCTGTTCGAAGATTTGCGGAGGGGGGCA
COMMENTSEQ	Design of mRNA 5′UTR of GUS reporter gene: first 23 nt were taken from vector sequence followed by 108-nt long 5′UTR of ADH1. In this construct CDS consisted from 18 codons of ADH1 CDS fused to 8 codons derived from vector polylinker sequence and GUS CDS downstream (see LONG). 3′UTR was represented by ADH
ORGANISM	Zea mays
KEYWORD	Hypoxia, stress, 5'UTR, enhancer
COMMENT	Translational efficiencies of reporter mRNAs containing UTR sequences of maize alcoholdehydrogenase gene mRNA were tested in maize protoplasts under normal or oxygen deprivation conditions. No changes in mRNA stability were detected so the effect observed resulted from the changes in stress- specific translation rate. Interestingly, the presence of ADH 5'UTR did not affect translation under aerobic conditions. Generally, deletions of ADH-derived fragments decreased stress-specific translatability: either deletion of first 18 ADH-derived codons or fragments of 5'UTR or 3′UTR. Note, that 5′portion of 5′UTR was presented in all constructions. The influence of possible changes in secondary structure was not tested or discussed. As was found ADH1 3′UTR mRNA increase hypoxia-specific translation 3.5-fold but decrease aerobic translation 3-fold.
REFERENCE	Bailey-Serres J., Dawe R.K. Both 5′ and 3′ sequences of maize adh1 mRNA are required for enhanced translation under low-oxygen conditions. Plant Physiol. 1996. 112. 685-695 PMID:8883381

и TRANSIG_ENH. Структура базы данных позволяет расширять список аннотируемых энхансеров по мере появления новых литературных данных.

Технологии реализации БДТЭ

БДТЭ управляется средствами информационно-поисковой системы Sequence Retrieval System (SRS) 6.1, которая развернута на сервере баз данных под управлением Red Hat Enterprise Linux 5.7. Система SRS (Zdobnov *et al*., 2002) специально разработана для формализованного описания биологических данных по заказу European Bioinformatics Institute. Средства SRS позволяют индексировать большинство информационных полей и эффективно осуществлять перекрестную связь полей в таблицах баз данных, что необходимо для построения эффективных пользовательских запросов и свободной навигации между полями и записями в различных таблицах. Эта система также автоматически генерирует Web-интерфейс для обеспечения поиска и визуализации информации в БД (формы запроса, визуализация данных, гиперссылки на документы в базе и Интернет-ресурсы, настройки способа визуализации).

Пользовательский интерфейс

Модуль интерфейса для БДТЭ представляет собой программный компонент, обеспечивающий интерфейс пользователя с БД. Доступ к БДТЭ может быть осуществлен по адресу http://wwwmgs.bionet.nsc.ru/mgs/dbases/trsig/ (рис. 2).

Пользователю предоставляется возможность просматривать списки генотипов, ключевых слов, типов энхансеров, которые содержатся в базе. Кроме этого, пользователь имеет возможность осуществлять поиск по большому числу

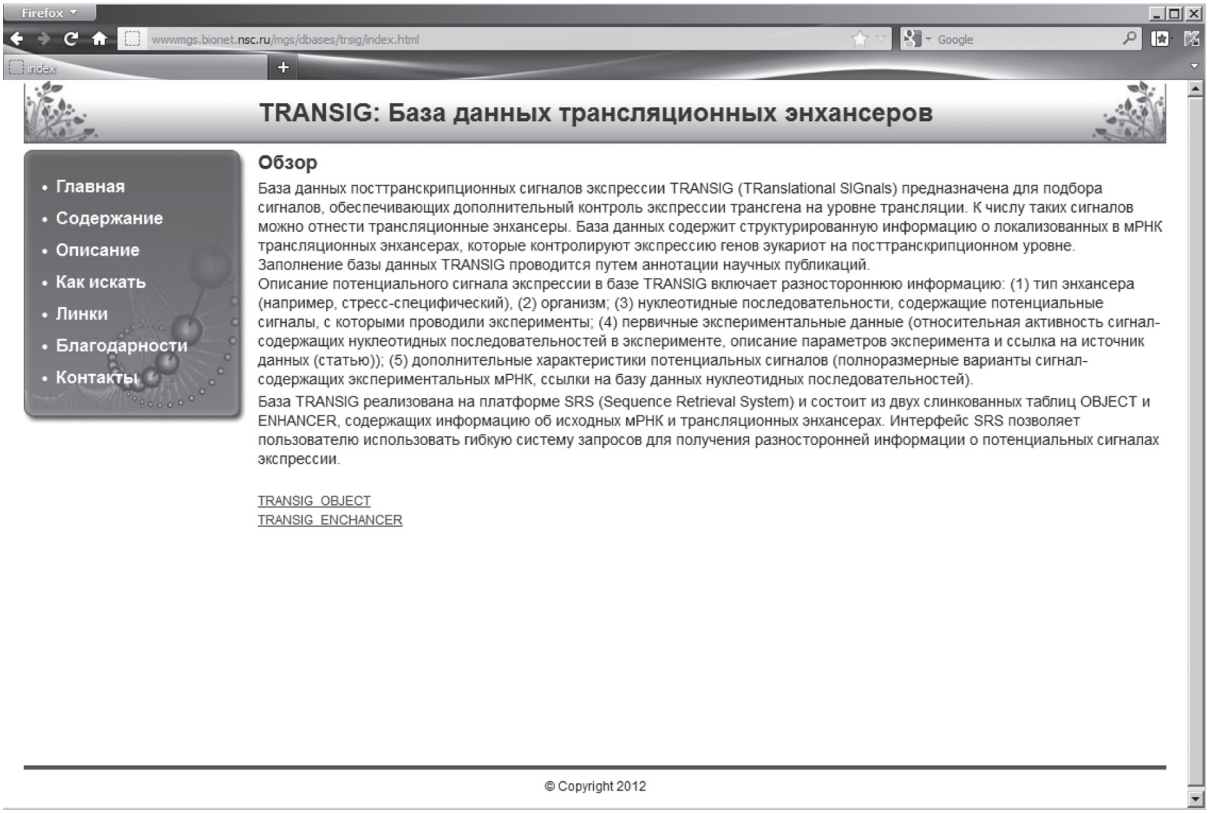

Рис. 2. Скриншот титульной страницы интерфейса БДТЭ.

поисковых полей таблиц TRANSIG_ OBJ и TRANSIG_ENH.

Поиск проводится отдельно для таблиц TRANSIG_OBJ и TRANSIG_ENH. Поиск в таблице TRANSIG_OBJ позволяет решать следующие задачи:

– найти сигналы, локализованные в 5′UTR, 3′UTR или internal fragment (поиск по полю LOCATION);

– найти тканеспецифические сигналы (поиск по полям TYPE, KEYWORDS и COMMENT);

– найти энхансеры, принадлежащие определенному организму (поиск по полю OS);

– найти энхансеры, расположенные в мРНК определенного гена (поиск по полю GENE);

– найти энхансеры, выделенные из мРНК с кепом на 5′-конце (поиск по полю CAP);

– найти энхансеры, выделенные из мРНК с поли(А)-участком на 3′-конце (поиск по полю POLYA).

Если найденный сигнал удовлетворяет требованиям пользователя, нуклеотидная последовательность из поля «Sequence» таблицы TRANSIG_OBJ может в дальнейшем использоваться как специфический сигнал экспрессии при дизайне регуляторных элементов трансгена.

Чтобы найти сигнал по названию гена, в мРНК которого он был обнаружен, в поисковой таблице необходимо выбрать название поля «Gene» и ввести часть названия гена со звездочкой (чтобы сделать расширенный поиск). Например, ген «alcoholdehydrogenase I» может быть обозначен как ADH1. В результате выполнения запроса система выдает список генов TRANSIG_OBJ: ADHZM3 и TRANSIG_OBJ:ADHZM5.

Чтобы найти сигналы, которые влияют на уровень мРНК в конкретном виде растений и зависят от конкретного регулятора, необходимо провести комплексный поиск. Для этого необходимо перейти на страницу сложного запроса, нажав кнопку «Search» на странице с типовым описанием структуры полей таблицы TRANSIG_ENH.

На этой странице меню «combine searches with» задает логическую операцию для выполнения совместных запросов нескольких полей. Например, для поиска энхансеров табака, которые реагируют на освещенность, необходимо

установить значение этого меню «AND», в левом столбце формы запроса выбрать поле «OS», а в правом столбце ввести название вида («tobacco»). В следующей строке формы в левом столбце выбрать «KEYWORDS», а в правом ввести условие («light») и выполнить запрос. Будет получен список сигналов, действие которых в растениях трансгенного табака зависело от освещения: TRANSIG_ENH:PSILRE5a.

ЗАКЛЮЧЕНИЕ

Разработана БДТЭ, предназначенная для накопления информации, необходимой для планирования генно-инженерных экспериментов с целью создания организмов с качественно новыми или улучшенными свойствами. БДТЭ содержит структурированную информацию о локализованных в мРНК трансляционных энхансерах, контролирующих экспрессию генов растений на посттранскрипционном уровне. Текущая версия БДТЭ содержит 58 аннотированных объектов (природных прототипов энхансеров) и 68 экспериментально исследованных энхансеров.

Формат представления данных обеспечивает быстрый поиск трансляционных энхансеров, чувствительных к различным индукторам, с целью обеспечения дополнительного к действию промотора паттерна транскрипции трансгена, необходимого для решения конкретных генно-инженерных задач.

БДТЭ является одним из модулей информационного портала «Биотехнология растений», разрабатываемого в ИЦиГ СО РАН (Кочетов и др., 2012). Этот информационный ресурс представляет собой платформу для решения различных задач в области генной инженерии и биотехнологии растений, содержащую модули разного типа. В частности, модуль БДТЭ планируется использовать в комбинации с модулем БДП, представляющим собой базу данных промоторов для трансгенеза (Смирнова и др., 2012).

Работа поддержана грантом Министерства образования и науки РФ в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 гг.» (07.514.11.4052).

ЛИТЕРАТУРА

Кочетов А.В., Смирнова О., Ибрагимова С. и др. Информационный портал «Биотехнология растений» – Интернет-ресурс для поддержки экспериментов в области генной инженерии растений, генетики и селекции пшеницы // Вавилов. журн. генет. и селекции. 2012. Т. 16. № 4/1. С. 838–848.

Смирнова О.Г., Рассказов Д.А., Афонников Д.А., Кочетов А.В. TGP – база данных промоторов для трансгенеза растений // Матем. биология и биоинформатика. 2012. Т. 7. № 2. С. 444–460.

Gallie D.R. The 5’-leader of tobacco mosaic virus promotes translation through enhanced recruitment of eIF4F // Nucl. Acids Res. 2002. V. 30. No. 15. P. 3401–3411.

Grillo G., Turi A., Licciulli F. *et al*. UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs // Nucl. Acids Res. 2010. V. 38. Database issue. P. D75–D80.

Jacobs G.H., Chen A., Stevens S.G. *et al*. Transterm: a database to aid the analysis of regulatory sequences in mRNAs // Nucl. Acids Res. 2009. V. 37. Database issue. P. D72–D76.

Liu Y., Wimmer E., Paul A.V. Cis-acting RNA elements in human and animal plus-strand RNA viruses // Biochim. Biophys. Acta. 2009. V. 1789. No. 9/10. P. 495–517.

Zdobnov E.M., Lopez R., Apweiler R., Etzold T. The EBI SRS server – recent developments // Bioinformatics. 2002. V. 18. P. 368–373.

A DATABASE ON TRANSLATIONAL ENHANCERS TO SUPPORT EXPERIMENTS WITH TRANSGENIC PLANTS

O.G. Smirnova, D.A. Rasskazov, A.V. Kochetov

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: ak@bionet.nsc.ru

Summary

A database on translational enhancers providing additional control of foreign gene expression at the mRNA translation level has been developed. It contains structured information on the presence of enhancers located within mRNAs, which control gene expression at the posttranscriptional stage. These data can be used to design genetic constructs for plant transgenesis. The database is based on the platform of the Sequence Retrieval System (SRS), allowing users to make a rapid search for enhancers with defined properties and retrieve corresponding nucleotide sequences. The database is available at http://wwwmgs.bionet.nsc.ru/mgs/dbases/trsig/.

Key words: database, information resource, translational enhancer, genetic engineering, transgenic plants.

УДК 577.322.2:004.94

КОМПЬЮТЕРНЫЕ МЕТОДЫ ИССЛЕДОВАНИЯ ТЕРМОСТАБИЛЬНОСТИ БЕЛКОВ И ИХ ПРИМЕНЕНИЕ В БИОЛОГИИ

© 2012 г. **Н.А. Алемасов, Э.С. Фомин**

Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия,
e-mail: alemasov@bionet.nsc.ru

Поступила в редакцию 15 июля 2012 г. Принята к публикации 31 августа 2012 г.

Проведена классификация существующих теоретических подходов к исследованию термостабильности белков. Методы компьютерного моделирования динамики белков позволяют наиболее полно оценить микро- и макропараметры исследуемых молекул, но без уточнений (например, в части учета распределения заряда) они ограничены в точности получаемых результатов. Перспективными ввиду низкой требовательности к вычислительным ресурсам являются также методы выделения жестких фрагментов белков. Они позволяют напрямую определить изменения в гибкости структуры при совершении аминокислотных замен. Но их недостатком является сложность оценки влияния окружающего растворителя и его термодинамических параметров.

Ключевые слова: термостабильность белков, свободная энергия, молекулярная динамика, «жесткие» фрагменты, электростатические и гидрофобные взаимодействия, термофилы, мезофилы.

ВВЕДЕНИЕ

Термостабильность белков отражает способность сохранять уникальную пространственную структуру полипептидной цепи под действием высокой температуры. Численно термостабильность определяется через разности энтальпии и энтропии нативного и денатурированного состояния белков (Matthews *et al.*, 1987; Talluri, 2011) и может рассматриваться с точки зрения термодинамических свойств составляющих белки аминокислот (Tanford, 1962). Существуют экспериментальные методы улучшения термостабильности белка: путем стабилизации его нативной формы, дестабилизации его денатурированной формы или комбинацией обоих подходов (Matthews, 1993).

Понимание факторов стабилизации белков, прежде всего, термофильных организмов, важно не только для выявления физико-химических принципов белкового фолдинга или механизмов стабилизации структуры и взаимодействия белков. Это также является необходимым условием для эффективного проектирования новых ферментов, способных работать при высоких температурах, что полезно в промышленности (Vogt *et al.*, 1997).

В фундаментальных исследованиях термостабильные белки часто используются в процессе биологического анализа, например в качестве биосенсоров. Кроме того, большое значение представляет оптимизация ДНК-полимераз (Talluri, 2011), которые используются в ПЦР-реакциях (реакциях полимеразной цепной реакции) (Saiki *et al.*, 1985), в направлении их большей термостабильности, специфичности и процессивности. Существуют и эволюционные аспекты, обусловливающие повышенную термостабильность у термофильных организмов (Afonnikov *et al.*, 2001; Афонников и др., 2011).

Экспериментальные методы исследования термостабильности белков могут предоставить весьма подробную информацию о связи

структурных модификаций молекулы и соответствующих ей изменений термостабильности. Но и здесь существуют ограничения (Tidor *et al.*, 1991). Рассматривая кристаллические структуры как таковые, в эксперименте сложно измерить возможный энтропийный вклад в свободную энергию, обусловленный внутренними движениями и влиянием аминокислотных замен на денатурированное состояние. Кроме того, изменения в стабильности могут быть обусловлены различными взаимодействиями, дающими как положительный, так и отрицательный вклад в свободную энергию, который также сложно отделить экспериментально.

Для преодоления данных ограничений применяют теоретические методы, которые позволяют более детально рассмотреть термодинамику белков, а также влияние замен на термостабильность. Настоящая работа направлена на классификацию существующих теоретических подходов к исследованию термостабильности белков. По каждому классу методов сделаны выводы о применимости, ограничениях и получаемых с помощью методов этого класса результатах.

НАПРАВЛЕНИЯ ИССЛЕДОВАНИЙ

В работе A. Mozo-Villiarias и E. Querol (2006) предлагается рассматривать три (возможно, пересекающихся) направления исследований термостабильности белков: сопоставление белков термофильных и мезофильных организмов; *ab initio* подходы; «доминирующие» взаимодействия. В настоящей работе предлагается расширить этот список еще одним классом методов.

Ab initio подходы (например метод молекулярной динамики (Alder, Wainwright, 1957)), а также менее трудоемкие методы, основанные на учете только «доминирующих» взаимодействий, требуют высоких вычислительных затрат. Поэтому в последние годы развиваются методы, основанные на выделении жестких фрагментов молекулы и изучении связи «жесткость/гибкость → термостабильность». Поскольку последние предполагают лишь статический анализ, они серьезно выигрывают по производительности у методов молекулярного моделирования.

Таким образом, список рассматриваемых в настоящей работе классов методов выглядит следующим образом.

1. Сопоставление белков термофильных и мезофильных организмов напрямую, с тем чтобы выявить закономерности, ведущие к их температурной стабилизации (Perutz, Raidt, 1975; Daniel *et al*., 1982; Vogt *et al*., 1997; Kumar *et al*., 2000, Szilagyi, Zavodszky, 2000);

2. Теоретические методы, учитывающие «доминирующие» эффекты (электростатические и гидрофобные взаимодействия), как наиболее существенные среди всех взаимодействий, влияющих на термостабильность (Beveridge, DiCapua, 1989; Sharp, Honig, 1990; Zhu, Elcock, 2010; Nicholls *et al*., 1991; Xiao, Honig, 1999);

3. Компьютерные алгоритмы, которые учитывают физические аспекты, влияющие на температурные характеристики белков (Tidor, Karplus, 1991; Guerois *et al*., 2002; Gromiha, 2003; Khechinashvili *et al*., 2006; Dehouck *et al*., 2009; Potapov *et al*., 2009; Seeliger, DeGroot, 2010);

4. Выделение «жестких фрагментов» для определения связи между гибкостью пространственной структуры белков и ее стабильностью (Radestock, Gohlke, 2011; Rader *et al*., 2012).

Сравнение белков мезофилов и термофилов

Исторически сложилось так, что наиболее ранние работы для изучения термостабильности белков были направлены на сравнение белков мезофильных организмов и их термофильных аналогов. Одной из первых теоретических работ является работа M. Perutz и H. Raidt (1975). Интересные результаты получены в работах R.M. Daniel с соавт. (1982), G. Vogt с соавт. (1997).

На наш взгляд, одной из наиболее показательных работ по направлению сравнения белков термофильных и мезофильных организмов является работа A. Szilagyi и P. Zavodszky (2000). В ней была исследована зависимость структурных характеристик белков от температуры, при которой существуют содержащие эти белки организмы. Авторы использовали наборы белков, представляющих 25 семейств, и рассчитали 13 структурных параметров, основывающихся на атомных координатах. Все параметры были объединены в следующие группы.

– «Вогнутость» белковой поверхности. Рассчитывалась с помощью программы Molecular Surface Package (Connolly, 1993).

– Водородные связи. Оценивались их число, количество непарных доноров и акцепторов. Расчет производился с помощью алгоритма Hb2 пакета молекулярного моделирования WHAT IF (Vriend *et al*., 1990).

– Ионные пары. Рассчитывалось число ионных пар. Два противоположно заряженных остатка считались ионной парой, если кратчайшее расстояние между их противоположно заряженными атомами не превышало заранее установленного значения (4, 6 или 8 Å).

– Параметры вторичной структуры: доля α-спиралей, β-листов, нерегулярных областей. Рассчитывались с помощью программы DSSP (Kabsch, Sander, 1983).

– Полярные/неполярные, внешние/внутренние области белка. Рассматривалось отношение площадей поверхности полярных и неполярных областей для внешних и внутренних областей белка. Все параметры подсчитывались как функции температуры оптимального роста соответствующих организмов. Для расчета использовалась программа WHAT IF. Атомы N и O считались полярными, остальные – неполярными. Внутренние области отдельно рассчитывались для денатурированной и нативной конформации полипептидной цепи, полученные площади вычитались.

Исходя из рассчитанных значений для всего набора белков, A. Szilagyi и P. Zavodszky (2000) сделали вывод о корреляции каждого из параметров с рядом характеристик. Среди них: оптимальная температура роста организма, из которого выделен исследуемый белок; изменение структурных параметров в белках умеренно термофильных организмов; изменение структурных параметров в белках крайне термофильных организмов.

Таким образом, были выявлены параметры, позволяющие отделить друг от друга умеренно и крайне термофильные организмы. Так, показано, что с ростом температуры число ионных пар увеличивается для обоих типов организмов. Кроме того, количество «вогнутостей» в белках крайне термофильных организмов значительно снижено, в то время как для умеренно термофильных организмов заметного изменения по этому параметру не наблюдается в сравнении с белками мезофилов.

На наш взгляд, рассмотренный подход к выявлению связи термостабильности белков с изменениями их структурных характеристик при сравнении термофилов и мезофилов способен дать только качественную информацию, которая может быть использована (с невысокой точностью) для нахождения общих закономерностей при разработке более точных методов. Ввиду того что данный подход основан на изучении ограниченного числа белков семейств крайне и умеренно термофильных организмов, он не способен оценить количественно изменение термостабильности конкретных белков, т. е. заранее предсказать это изменение.

Доминирующие взаимодействия

Ряд исследователей полагают, что доминирующее влияние на стабильность пространственной структуры (т. е. на термостабильность) белков оказывают электростатические и гидрофобные силы (Beveridge, DiCapua, 1989; Sharp, Honig, 1990; Zhu, Elcock, 2010).

Электростатические взаимодействия. В работе L. Xiao и B. Honig (1999) подробно рассматриваются электростатические взаимодействия и их связь с термостабильностью. Авторы используют методы статистической механики для оценки множества состояний ионизации в нативном и денатурированном белке. Так, средний заряд каждой ионизируемой группы вычислен из среднего значения, полученного из всех состояний ионизации белка. Электростатическая энергия каждого ионизационного состояния, в свою очередь, определялась с помощью решения конечно-разностной формы уравнения Пуассона–Больцмана.

Кроме того, в рассматриваемой работе влияние электростатических взаимодействий складывалось из трех частей: $\Delta\Delta E(solv)$ – дестабилизирующий вклад в стабильность белка из-за десольватации ионизующихся групп в процессе фолдинга; $\Delta\Delta E(hb)$ – сумма взаимодействий от водородных связей между заряженными и полярными группами; $\Delta\Delta E(cc)$ – кулоновские взаимодействия между заряженными группами. Следовательно, $\Delta\Delta E(elec) = \Delta\Delta E(solv) + \Delta\Delta E(hb) + + \Delta\Delta E(cc)$.

Хотя диэлектрическая проницаемость воды при 20 °C равна 80 Ф/м, авторы проводили расчеты с ε = 4 и ε = 20. Под влиянием соли и при изменении диэлектрической константы раствора величина электростатической свободной энергии значительно изменяется. Результаты расчетов показывают, что при ε = 20 полное изменение свободной энергии электростатических взаимодействий в процессе фолдинга для большинства случаев отрицательно. Кроме того, практически для всех белков термофильных организмов внутри одного семейства значение $\Delta\Delta G(elec)$ меньше, чем для мезофильных, что позволяет разделять соответствующие белки независимо от диэлектрической константы ε. Таким образом, действительно, электростатические взаимодействия вносят большой вклад в термостабильность белков гипертермофильных организмов при pH = 7 в каждом из четырех изученных семейств белков.

Другими словами, в гипертермофильных белках вклад электростатических взаимодействий в свободную энергию фолдинга более существенный, чем в их мезофильных аналогах. Этот вклад зависит от значения pH и, как считают L. Xiao и B. Honig (1999), может исчезнуть в результате изменения распределения заряда белка при определенных pH.

Гидрофобные взаимодействия. В соответствии с работой H. Dong с соавт. (2008) влияние гидрофобности на стабильность белков сводится к следующему: при замене гидрофобного остатка в гидрофобном ядре белка на полярный стабильность всей структуры падает, а при заполнении «вогнутостей» поверхности белка гидрофобной аминокислотой структура становится более стабильной. Особенно этот эффект характерен для замен в гидрофобном ядре белка. Изучение с разных сторон влияния гидрофобных взаимодействий на стабильность белковых структур проводилось в работах D. Eisenberg с соавт. (1986); S. Kumar, M. Meenatchi (2011).

Детальное изучение гидрофобных взаимодействий проводилось в работе S. Zhu, A. Elcock (2010), в которой с помощью метода молекулярной динамики сделана попытка напрямую выявить термодинамические особенности гидрофобных взаимодействий на примере ассоциации пар «ацетат–метиламмоний» и «метан–метан». Для моделирования использовалась программа GROMACS 3.3 (van der Spoel *et al.*, 2005). Значения температуры устанавливались в ряд от –12,5 °C до 112,5 °C с шагом 12,5 °C. Использовался явный растворитель – водный раствор в рамках моделей TIP3P и TIP5P. Моделируемая область имела размеры $25 \times 25 \times 25$ Å^3. Рассчитывалась динамика системы в течение 500 нс для каждого значения температуры.

Из полученных «снимков» рассчитывались гистограммы: для пары «ацетат–-метиламмоний» вычислялось расстояние между атомом углерода карбоксильной группы молекулы ацетата и атомом азота аминогруппы молекулы метиламмония; для пары «метан–метан» – расстояние между двумя углеродами. Кроме того, для пары ацетат–метиламмоний также рассчитывалась гистограмма по двум размерностям: по расстояниям между заряженными группами двух молекул (описывают электростатические взаимодействия) и между их метильными группами (описывают гидрофобные взаимодействия). Для 1D и 2D-гистограмм рассчитывались свободные энергии взаимодействия. Причем для расчетов было смоделировано два набора гистограмм с взаимодействующими между собой молекулами и с «выключенными» межмолекулярными взаимодействиями. Это было сделано для того, чтобы вычесть вклад конфигурационной энтропии в свободную энергию.

Авторы показали, что результаты предсказаний изменений свободной энергии в рамках моделей TIP3P и TIP5P хорошо согласуются между собой для температур от 25 до 100 °C. Кроме того, в рамках модели воды TIP5P относительная прочность гидрофобных взаимодействий и солевых мостиков по большей части не изменяется в промежутке от 0 до 40 °C. При более высоких температурах солевые мостики дают более весомый вклад в стабильность.

На наш взгляд, основным недостатком примененного подхода является то, что для расчета только доминирующей части свободной энергии используется весьма ресурсоемкий метод молекулярной динамики. В то же время поправки, связанные с вкладами гидрофобных и электростатических взаимодействий, могут быть введены в методы компьютерного моделирования, рассмотренные ниже, с целью учета всех вкладов в свободную энергию.

Компьютерное моделирование

Задача расчета термостабильности в данном подходе сводится к вычислению разности свободных энергий между различными состояниями моделируемой системы. Так, например, расчет термостабильности мутантной формы белка требует вычисления величины двойной разности $\Delta\Delta G = \Delta G_3 - \Delta G_2$, где ΔG_3 – разность свободных энергий между глобулярной и денатурированной формами нативного белка, а ΔG_2 – разность свободных энергий между глобулярной и денатурированной формой мутанта (см. рис. 1). Данный подход успешно был применен в работе B. Tidor, M. Karplus (1991).

Свободная энергия является функцией состояния и не зависит от пути, который проходит система. Поэтому исходя из условия равенства нулю изменения свободной энергии вдоль всего цикла искомая величина $\Delta\Delta G$ может быть рассчитана также через разность $\Delta\Delta G = \Delta G_1 - \Delta G_4$. Величины ΔG_1 и ΔG_4 связаны с нефизическим, т. е. нереализуемым в природе, непрерывным процессом «превращения» одних аминокислотных остатков в другие (см. рис. 1). Ключевой особенностью этого «превращения» (или «алхимического» процесса) является использование метода λ-динамики, в котором в один и тот же момент существуют физические взаимодействия от заменяемой и заменяющей части молекулы (Kong, Brooks, 1996).

В работе D. Seeliger, B. De Groot (2010), которая посвящена «алхимическим» расчетам свободной энергии, с помощью распространенной программы GROMACS (Hess *et al.*, 2008) и силового поля amber99sb (Hornak *et al.*, 2006) было исследовано свыше 100 мутантных белков барназы. Результаты расчетов показали лучшее согласие с экспериментальными данными, чем в работе R. Guerois с соавт. (2002). Корреляция рассчитанных и экспериментальных данных составила 0,86, а средняя абсолютная ошибка – 3,31 кДж/моль, что говорит о большем потенциале подходов «алхимических» расчетов.

Одним из подходов, реализующих расчет свободной энергии, является использование эмпирических энергетических функций (Guerois *et al.*, 2002). Но подобные функции часто дают неточные результаты при оценке энергии. Это происходит в случае расчетов свободной энергии белков, не входящих в обучающий набор или при оценке энергии замен, приводящих к значительному изменению конформации белка. Другой вариацией подходов к расчету свободной энергии является метод, основанный на упрощенных моделях, например, представляющий растворитель неявно (Vorobjev, 2011).

В нашей недавней работе (Фомин, Алемасов, 2012) были представлены результаты дальнейшего развития методов «алхимических» преобразований: в расчетную схему были включены эффекты поляризации. Установлено, что учет данных эффектов улучшает точность расчетов разностей свободных энергий для мутаций, включающих изменение заряда аминокислоты. Так, при расчете влияния замены R72G в барназе была существенно улучшена величина изменения свободной энергии в сравнении со значением, полученным в работе D. Seeliger, B. De Groot (2010). Полученное нами значение –19,8 кДж/моль находится заметно ближе к экспериментальному для этой мутации значению –10,4 кДж/моль, чем величина –57,67 кДж/моль, рассчитанная в работе D. Seeliger, B. De Groot (2010).

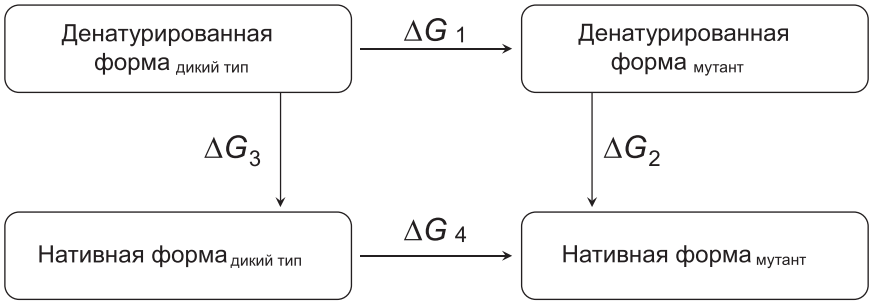

Рис. 1. Термодинамический цикл. По: D. Seeliger, B. De Groot (2010).

Выделение жестких фрагментов

В работе A.T. Rader с соавт. (2012) аналогично подходу, использованному в работе S. Radestock, H. Gohlke (2011), было сделано исследование связи аллостерии и жесткости структуры белка с его термостабильностью. Ключевым вопросом в этой работе было рассмотрение мутации A35V в GH12 эндоглюканазе *Trichoderma reesei*. Примечательно то, что эта замена делает белок способным выдерживать температуры на 8 °C выше, чем белок дикого типа. При этом разница в RMSD между этими двумя формами белка составляет всего порядка 0,4 Å.

Гибкость и стабильность биомолекулярных систем в этой работе рассчитывалась с помощью программы FIRST (Jacobs *et al.*, 2001). Модель, реализованная в FIRST, рассматривает физические взаимодействия между парами атомов в качестве ограничений (например ковалентные связи). Из этих ограничений формируется трехмерный граф, включающий полный набор взаимодействий, присутствующих в нативном состоянии белка. Жесткие фрагменты соответствуют тем наборам атомов, которые соединены так, что любое внутреннее движение нарушает хотя бы одно из ограничений. Остальные области молекулы рассматриваются как гибкие.

В FIRST каждая связь помечается как гибкая (способная вращаться) или жесткая (фиксированная относительно вращения). В соответствии с гибкостью связей определяется гибкость/жесткость аминокислотных остатков. С помощью данных авторы сравнили нативную структуру со структурой мутанта (на примере эндоглюканазы) и определили остатки, меняющие свою жесткость в мутантной структуре.

Так как в модели не учитывается температура, авторы ввели параметр E_{cut} – энергия разрыва водородной связи. Параметр рассчитывается для всех неявных водородных связей, основываясь на локальной геометрии конкретного донора, акцептора и атомов водорода. В зависимости от E_{cut} вводится условная температура как $\Theta = 300 - 20E_{cut}$.

Выбрав температуру денатурации нативного белка в качестве температуры для модели (346 K), Rader с соавт. (2012) рассчитали долю атомов в самом крупном жестком фрагменте (см. рис. 2). Очевидным недостатком данного подхода является невозможность оценить влияние растворителя на термостабильность белка, а также pH раствора и других физических характеристик (давление и т. п.).

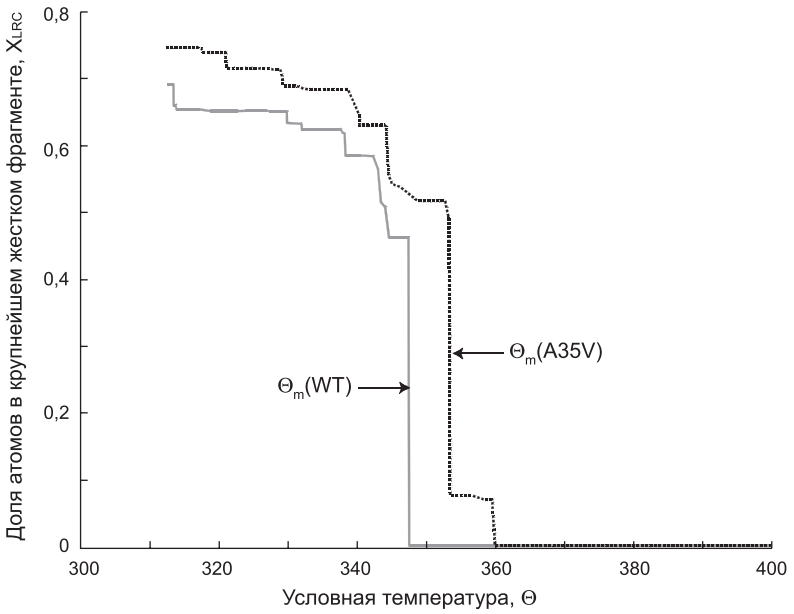

Рис. 2. Доля атомов в наиболее крупном жестком фрагменте в зависимости от условной температуры. По: Rader с соавт. (2012).

Таблица

Сведения о рассматриваемых подходах

Класс	Метод	Объект	Результаты
1	Сравнение последовательностей и пространственных структур гомологичных белков мезофилов и термофилов (Perutz, Raidt, 1975)	Бактериальный ферредоксин, гемоглобины A и A2 человека	В термостабильном ферредоксине найдены замены, формирующие водородные связи и солевые мостики. Гемоглобин A2 более стабилен, чем гемоглобин A за счет дополнительной связи между α1 и δ1 субъединицами
	Расчет доли полярных областей белка. Расчет числа водородных связей и солевых мостиков (Vogt *et al*., 1997)	16 семейств белков мезофилов и термофилов	В 80 % случаев есть корреляция внутри семейства между термостабильностью и числом водородных связей. В 67 % случаев есть корреляция термостабильности с формированием ионных пар
	Расчет структурных параметров на основе атомных координат (Szilagyi, Zavodszky, 2000)	25 семейств белков умеренно и крайне термофильных организмов	Найдены совокупности параметров, позволяющие выделять термофильные организмы: например, с ростом температуры число ионных пар увеличивается
2	Статистическая механика для расчета изменения электростатической свободной энергии в процессе фолдинга (Xiao, Honig, 1999)	GDH, GAPDH, ферредоксин, CheY мезофилов и термофилов	Для термофильных белков внутри одного семейства свободная энергия меньше, чем для мезофильных. Опровергнута корреляция термостабильности и числа ионных пар
	Расчет площадей, доступных растворителю для аминокислот, расчет свободной энергии сольватации (Eisenberg *et al*., 1986)	20 аминокислот	Корреляция рассчитанного изменения свободной энергии сольватации с экспериментом, равная 0,88
	Молекулярная динамика (GROMACS 3.3), потенциал средней силы (Zhu, Elcock, 2010)	Ацетат, метиламмоний, метан	Прочность гидрофобных взаимодействий и солевых мостиков не изменяется от 0 до 40 °C. При более высоких температурах солевые мостики доминируют в термостабильности
	Расчет кинетических и термодинамических величин (Kumar, Meenatchi, 2011)	20 аминокислот, цитохром b	Чем выше количество гидрофобных аминокислот, тем выше становится вклад в энергию Гиббса при кинетических и термодинамических расчетах
3	Молекулярная динамика (CHARMM), λ-динамика (Tidor, Karplus, 1991)	Лизоцим бактериофага T4, мутация R96H	Для исследуемой мутации получено значение –7,9 кДж/моль, что качественно соответствует эксперименту (–13,4 +/–5 ккал/моль)
	Энергетическая функция (Guerois *et al*., 2002)	1088 замен в различных белках	Наклон прямой линейной регрессии – 0,64, коэффициент корреляции с экспериментом – 0,73, стандартное отклонение – 4,27 кДж/моль
	Молекулярная динамика (GROMACS 4), λ-динамика (Seeliger, De Groot, 2010)	109 замен барназы	Корреляция рассчитанных и экспериментальных данных – 0,86, средняя абсолютная ошибка – 3,31 кДж/моль
4	Анализ сети с ограничениями (CNA), расчет доли атомов в жесткой части сети (Radestock, Gohlke, 2011)	19 семейств белков мезофильных и термофильных организмов	Около 67 % семейств белков имели температуру денатурации выше для термофильных гомологов
	Анализ жесткости структуры (FIRST), нахождение наибольшего жесткого фрагмента (Rader *et al*., 2012)	GH12 эндоглюканаза	Экспериментальное значение изменения температуры денатурации (7,7 °C) качественно подтверждается рассчитанным значением (5,9 °C)

ВЫВОДЫ

В настоящей работе представлен обзор наиболее распространенных методов теоретического исследования термостабильности. Уточнена классификация существующих подходов и представлены результаты, которые могут быть получены с их использованием. Основные сведения о рассмотренных подходах представлены в таблице. Так, наиболее распространенным методом исследования термостабильности белков является метод сравнения гомологов термофильных и мезофильных организмов. Этот подход естественным образом развился из экспериментальных методов и позволяет выявить самые существенные характеристики, которые влияют на способность белков функционировать при более высоких температурах.

В последнее время развиваются методы компьютерного моделирования динамики белков, которые позволяют наиболее полно оценить микро- и макропараметры исследуемых молекул, что обусловлено использованием базовых физических соотношений и уравнений движения. Методы молекулярного моделирования, однако, ограничены в точности и в будущем их прямое, без модификаций, использование вряд ли позволит получать более близкие к эксперименту результаты.

Необходимость модификаций методов моделирования также вытекает из существования доминирующего вклада в свободную энергию электростатических и гидрофобных взаимодействий. Уточнение параметризации электростатических и гидрофобных взаимодействий в потенциалах, определяющих движение атомов, является перспективным улучшением моделей динамики белков.

Также перспективными ввиду низкой требовательности к вычислительным ресурсам являются методы выделения жестких фрагментов белков. Они позволяют напрямую определить изменения в гибкости структуры при совершении аминокислотных замен. Их недостатком является сложность оценки влияния окружающего растворителя и его термодинамических параметров. Однако это не препятствует применению подобных методов для оценки термостабильности крупных систем, которые сложно поддаются моделированию, например, методами молекулярной динамики.

БЛАГОДАРНОСТИ

Данная работа выполнена в рамках Госконтракта 07.514.11.4011 с Министерством образования и науки РФ.

Авторы выражают признательность С.А. Лашину за критическое рассмотрение рукописи данной статьи и за ценные советы по ее улучшению.

ЛИТЕРАТУРА

Афонников Д.А., Медведев К.Е., Гунбин К.В., Колчанов Н.А. Важная роль гидрофобных взаимодействий при адаптации белков к высоким давлениям // Докл. АН. 2011. Т. 438. № 3. С. 412–415.

Фомин Э.С., Алемасов Н.А. Программный комплекс l-molkern для расчетов разностей свободных энергий с учетом эффектов перераспределения заряда // Математ. биол. и биоинформатика. 2012. Т. 7. № 2. С. 398–409.

Afonnikov D.A., Oshchepkov D.Y., Kolchanov N.A. Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with coadaptive substitutions // Bioinformatics. 2001. V. 17. No. 11. P. 1035–1046.

Alder B., Wainwright T. Phase transition for a hard sphere system // J. Chemical Physics. 1957. V. 27. P. 1208.

Beveridge D., DiCapua F. Free energy via molecular simulation: applications to chemical and biomolecular systems // Annu. Rev. Biophys. Biophys. Chem. 1989. V. 18. No. 1. P. 431–492.

Connolly M. The molecular surface package // J. Mol. Graphics. 1993. V. 11. No. 2. P. 139–141.

Daniel R.M., Cowan D.A., Morgan H.W., Curran M.P. A correlation between protein thermostability and resistance to proteolysis // Biochem. J. 1982. V. 207. No. 3. P. 641.

Dehouck Y., Grosfils A., Folch B. *et al.* Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: Popmusic-2.0 // Bioinformatics. 2009. V. 25. No. 19. P. 2537–2543.

Dong H., Mukaiyama A., Tadokoro T. *et al.* Hydrophobic effect on the stability and folding of a hyperthermophilic protein // J. Mol. Biol. 2008. V. 378. No. 1. P. 264–272.

Eisenberg D., McLachlan A. Solvation energy in protein folding and binding // Nature. 1986. V. 319. No. 6050. P. 199–203.

Gromiha M. Factors influencing the thermal stability of buried protein mutants // Polymer. 2003. V. 44. No. 14. P. 4061–4066.

Guerois R., Nielsen J., Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations // J. Mol. Biol. 2002. V. 320. No. 2. P. 369–387.

Hess B., Kutzner C., van der Spoel D., Lindahl E. Gromacs 4:

Algorithms for highly efficient, load-balanced, and scalable molecular simulation // J. Chem. Theory and Computat. 2008. V. 4. No. 3. P. 435–447.

Hornak V., Abel R., Okur A. *et al.* Comparison of multiple amber force fields and development of improved protein backbone parameters // Proteins: Structure, Function, and Bioinformatics. 2006. V. 65. No. 3. P. 712–725.

Jacobs D.J., Rader A.J., Kuhn L.A., Thorpe M.F. Protein flexibility predictions using graph theory // Proteins: Structure, Function, and Bioinformatics. 2001. V. 44. No. 2. P. 150–165.

Kabsch W., Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features // Biopolymers. 1983. V. 22. No. 12. P. 2577–2637.

Khechinashvili N.N., Fedorov M.V., Kabanov A.V. *et al.* Side chain dynamics and alternative hydrogen bonding in the mechanism of protein thermostabilization // J. Biomol. Struct. Dyn. 2006. V. 24. No. 3. P. 255–262.

Kong X., Brooks III C.L. λ-dynamics: A new approach to free energy calculations // J. Chem. Phys. 1996. V. 105. No. 6. P. 2414–2423.

Kumar S., Meenatchi M. Virtual quantification of protein stability using applied kinetic and thermodynamic parameters // IIOAB Lett. 2011. V. 1. No. 1. P. 21–28.

Kumar S., Tsai C., Nussinov R. Factors enhancing protein thermostability // Protein Engineering. 2000. V. 13. No. 3. P. 179–191.

Matthews B. Structural and genetic analysis of protein stability // Annu. Rev. Biochem. 1993. V. 62. No. 1. P. 139–160.

Matthews B., Nicholson H., Becktel W. Enhanced protein thermostability from sitedirected mutations that decrease the entropy of unfolding // Proc. Natl Acad. Sci. USA. 1987. V. 84. No. 19. P. 6663–6667.

Mozo-Villiarias A., Querol E. Theoretical analysis and computational predictions of protein thermostability // Curr. Bioinformatics. 2006. V. 1. No. 1. P. 25–32.

Nicholls A., Sharp K., Honig B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons // Proteins: Structure, Function, and Bioinformatics. 1991. V. 11. No. 4. P. 281–296.

Perutz M., Raidt H. Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2 // Nature. 1975. V. 255. P. 256–259.

Potapov V., Cohen M., Schreiber G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details // Protein Engineering Design and Selection. 2009. V. 22. No. 9. P. 553–560.

Rader A.J., Yennamalli R.M., Harter A.K., Sen T.Z. A rigid network of long-range contacts increases thermostability in a mutant endoglucanase // J. Biomol. Struct. Dynam. 2012. P. 1–10.

Radestock S., Gohlke H. Protein rigidity and thermophilic adaptation // Proteins: Structure, Function, and Bioinformatics. 2011. V. 79. No. 4. P. 1089–1108.

Saiki R.K., Scharf S., Faloona F. *et al.* Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia // Science. 1985. V. 230. No. 4732. P. 1350–1354.

Seeliger D., De Groot B. Protein thermostability calculations using alchemical free energy simulations // Biophys. J. 2010. V. 98. No. 10. P. 2309–2316.

Sharp K., Honig B. Electrostatic interactions in macromolecules: theory and applications // Annu. Rev. Biophys. Biophys. Chem. 1990. V. 19. No. 1. P. 301–332.

Szilagyi A., Zavodszky P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey // Structure. 2000. V. 8. No. 5. P. 493–504.

Talluri S. Advances in engineering of proteins for thermal stability // Intern. J. Adv. Biotechnol. Res. 2011. V. 2. No. 1. P. 190–200.

Tanford C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins // J. Amer. Chem. Soc. 1962. V. 84. No. 22. P. 4240–4247.

Tidor B., Karplus M. Simulation analysis of the stability mutant r96h of t4 lysozyme // Biochemistry. 1991. V. 30. No. 13. P. 3217–3228.

van der Spoel D., Lindahl E., Hess B. *et al.* Gromacs: fast, flexible, and free // J. Computat. Chem. 2005. V. 26. No. 16. P. 1701–1718.

Vogt G., Woell S., Argos P. Protein thermal stability, hydrogen bonds, and ion pairs1 // J. Mol. Biol. 1997. V. 269. No. 4. P. 631–643.

Vorobjev Y.N. Advances in implicit models of water solvent to compute conformational free energy and molecular dynamics of proteins at constant pH // Computat. Chem. Methods Struct. Biol. 2011. V. 85. P. 281–322.

Vriend G. What if: a molecular modeling and drug design program // J. Mol. Graphics. 1990. V. 8. No. 1. P. 52.

Xiao L., Honig B. Electrostatic contributions to the stability of hyperthermophilic proteins1 // J. Mol. Biol. 1999. V. 289. No. 5. P. 1435–1444.

Zhu S., Elcock A. A complete thermodynamic characterization of electrostatic and hydrophobic associations in the temperature range 0 to 100 °C from explicit solvent molecular dynamics simulations // J. Chem. Theory Computat. 2010. V. 6. No. 4. P. 1293–1306.

THEORETICAL METHODS FOR INVESTIGATING PROTEIN THERMOSTABILITY AND THEIR APPLICATIONS IN BIOLOGY

N.A. Alemasov, E.S. Fomin

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: alemasov@bionet.nsc.ru

Summary

Classification of existent theoretical approaches for investigating protein thermal stability is performed. Computer simulations allow to fully estimate micro- and macro properties of the molecules. But those approaches are limited in accuracy and therefore require certain improvements e.g. making them to allow for molecular charge distribution. Also promising methods are those dealing with rigid regions of proteins. They do not require a huge amount of computations and allow to directly determine changes in molecular structure flexibility caused by mutated amino acid residues. But using the only structure it is impossible to explicitly estimate an effect of solvent and its thermodynamical properties.

Key words: protein thermal stability, free energy, molecular dynamics, rigid regions, semi-empiric energy function, electrostatic and hydrophobic interactions, thermophiles, mesophiles.

УДК 616-056.3:379.0:004.4

КОМПЬЮТЕРНЫЙ АНАЛИЗ ВЗАИМОСВЯЗИ АЛЛЕРГЕННОСТИ МИКРООРГАНИЗМОВ И СРЕДЫ ИХ ОБИТАНИЯ

© 2012 г. **А.О. Брагин[1], П.С. Деменков[1], Е.С. Тийс[1], Р. Хофештадт[2], В.А. Иванисенко[1], Н.А. Колчанов[1, 3, 4]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: ibragim@bionet.nsc.ru;
[2] Университет Билефельда, Билефельд, Германия;
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия;
[4] НИЦ «Курчатовский институт», Москва, Россия

Поступила в редакцию 1 августа 2012 г. Принята к публикации 31 августа 2012 г.

В течение 20-го столетия наблюдался стремительный рост числа аллергических заболеваний. На данный момент в индустриальных странах от аллергии страдает значительная часть населения, что делает анализ аллергенных свойств белков важной задачей. Ранее выдвигались предположения, что аллергенность белков зависит от их размера, ферментативных свойств, гомологии с белками человека и т. д. Однако анализ взаимосвязи аллергенности белков и среды обитания организма, к которому принадлежат данные белки, до сих пор не проводился. Нами были предсказаны белки-аллергены из протеомов более 500 видов микроорганизмов. Показано, что количество белков-аллергенов в протеомах микроорганизмов статистически значимо связано с патогенностью, ареалом, температурными условиями среды обитания и потребностью в кислороде этих микроорганизмов.

Ключевые слова: предсказание аллергенности белков, аллергенность протеомов микроорганизмов, среда обитания микроорганизмов, экстремофилы.

ВВЕДЕНИЕ

Аллергия является распространенной проблемой со здоровьем в наши дни. От различных аллергических заболеваний в мире страдает более одной трети населения индустриально развитых стран (WAO White Book, 2011). Элиминационная терапия остается одним из наиболее эффективных средств борьбы с появлением симптомов аллергии (Platts-Mills *et al*., 2000). В связи с этим возникает потребность в анализе аллергенных свойств белков. Теоретические знания о потенциальной аллергенности белков из окружающих нас микроорганизмов имеют большое значение для прогнозирования и выработки стратегий предотвращения аллергических заболеваний.

Кроме экспериментальных методов оценки аллергенности белков существуют еще и компьютерные методы оценки перекрестной реактивности между анализируемым белком и аллергенами. Так, Всемирная организация здравоохранения (WHO) и Продовольственная и сельскохозяйственная организация (FAO) предложили проводить сравнение последовательности анализируемого белка с аминокислотными последовательностями известных аллергенов (FAO/WHO, 2003). Согласно предложенным этими организациями критериям, считалось, что белок имеет перекрестную реактивность с известным аллергеном, если у них имеется идентичный участок длиннее 6 аминокислотных остатков или фрагмент последовательности из 80 аминокислотных остатков анализируемого

белка идентичен последовательности белка-аллергена более чем на 35 %. Но такой подход давал большое количество ложноположительных результатов (Goodman *et al*., 2008).

Кроме метода оценки аллергенности белков, предложенного WHO и FAO, были разработаны и другие методы предсказания аллергенности. При анализе аллергенных свойств белков было предложено использовать поиск мотивов (Stadler M.B., Stadler B.M., 2003; Kong *et al*., 2007), вейвлет-преобразования (Li *et al*., 2004), классификаторы, такие, как метод k-ближайших соседей (Zorzet *et al*., 2002), метод опорных векторов (Saha *et al*., 2006; Muh *et al*., 2009) и др. Нами также был разработан метод предсказания аллергенности белков с использованием конформационных пептидов (Брагин и др., 2011).

В настоящее время активно ведутся работы по изучению молекулярных механизмов IgE-опосредованных аллергических реакций. Считается, что в норме IgE-антитела играют одну из важных ролей в реакциях иммунитета на паразитарные инфекции, аллергия же является нежелательным побочным эффектом (MacDonald *et al*., 2002; Bischoff *et al*., 2007). Обычно выделяют два этапа развития аллергии: сенсибилизация к определенному аллергену живого организма и выделение медиаторов воспаления при повторном контакте с аллергеном. После первого контакта с аллергеном при IgE-опосредованной аллергии начинается активная выработка антител, что приводит к сенсибилизации организма. При повторном контакте с аллергеном происходит дегрануляция тучных клеток и базофилов, при этом выделяются медиаторы, действие которых обусловливает клинические проявления аллергии (Cianferoni, Spergel, 2009; Locksley, 2010). Для дегрануляции тучных клеток требуется контакт аллергена с двумя IgE-антителами на поверхности тучных клеток, поэтому вызывать аллергию могут белки, размер которых позволяет осуществить сшивание двух реагиновых молекул на мембране тучных клеток или базофилов (Huby *et al*., 2000). Кроме того, размер аллергена должен позволять ему проникать через защитные слои, например, через слизистый слой. Обычно аллергенами являются белки массой от 10 кДа до 70 кДа (Jeebhay *et al*., 2001; Puc, 2003). Было также показано, что протеазная активность аллергенов может оказывать адъювантный эффект (Sudha *et al*., 2009). Кроме того, на аллергенность белков влияют их структурные и физико-химические свойства (Breiteneder, Mills, 2005).

Известно, что у некоторых видов организмов количество аллергенных для человека белков намного выше, чем у других. В связи с этим выдвигались гипотезы о том, что аллергенность белков организма зависит от того, насколько этот организм эволюционно близок к человеку. Так, в работе T.A. Platts-Mills показано, что более «аллергенными» являются те виды, которые имеют меньшее родство с млекопитающими (Platts-Mills, 2012). В работе J.A. Jenkins с соавт. было показано, что белки животных, вызывающие пищевую аллергию у человека, как правило, не имеют гомологов среди белков человека (Jenkins *et al*., 2007).

Особое место в изучении аллергенности белков занимают белки различных микроорганизмов. Микроорганизмы широко распространены в природе и используются в биотехнологии (Antranikian *et al*., 2005), биомедицине и ветеринарии (Irwin, 2010), фармацевтике (Van den Burg, 2003; Antranikian *et al*., 2005), текстильной промышленности (Pennisi, 1997) и т. д. Поскольку человек находится в тесном контакте с множеством видов различных микроорганизмов, то данное направление исследований является весьма актуальным. Но, несмотря на это, вопрос об аллергенности белков микроорганизмов до сих пор остается малоизученным. Одной из причин слабой изученности аллергенных свойств большинства микроорганизмов, существующих в природных условиях, является трудность их культивирования.

В связи с этим целью настоящей работы были предсказание белков-аллергенов в протеомах микроорганизмов и исследование связи аллергенности протеома с факторами среды обитания микроорганизма. Благодаря достижениям в массовом секвенировании геномов, мы смогли проанализировать многие микроорганизмы, чей геном был полностью секвенирован. Мы рассмотрели микроорганизмы, которые распространены в среде обитания, позволяющей легко контактировать с человеком, а также различные экстремофилы, для которых контакты с человеком маловероятны.

Было проанализировано более 500 видов архей и бактерий, включая экстремофилы. Показано, что аллергенность микроорганизмов (доля аллергенных белков в протеоме) статистически значимо связана с их патогенностью, ареалом, температурными условиями среды обитания и потребностью в кислороде. Оказалось, что фактором, наиболее тесно связанным с аллергенностью, является патогенность микроорганизмов. Таким образом, полученные результаты могут быть использованы для планирования экспериментов по изучению аллергенных свойств белков микроорганизмов и их связи с патогенезом заболеваний.

МАТЕРИАЛЫ И МЕТОДЫ

Предсказание аллергенности проводилось для 546 видов бактерий и архей, геномы которых были полностью секвенированы. Последовательности белков, а также характеристики микроорганизмов, включая их патогенность, подвижность, среду обитания и т. д. (см. табл. 1), были взяты из баз данных GenBank и BioProject (Benson *et al*., 2011; Barrett *et al*., 2012).

Для предсказания аллергенности белков рассматриваемых микроорганизмов использовался двухшаговый подход (рис.). На первом шаге методом выравнивания аминокислотных последовательностей отбирались потенциальные белки-аллергены, которые имели высокую гомологию с известными белками-аллергенами. На втором шаге количество потенциальных белков-аллергенов увеличивалось за счет предсказания аллергенности среди белков, не имеющих высокой гомологии с известными аллергенами, с использованием разработанного нами ранее метода, основанного на анализе конформационных пептидов (Брагин и др., 2011). В методе проводился поиск конформационных и линейных пептидов из аллергенных белков в аминокислотной последовательности анализируемого белка с учетом физико-химических свойств аминокислот.

Таблица 1

Характеристики микроорганизмов

Группа	Характеристика микроорганизма	Количество рассматриваемых видов с данной характеристикой
Царство	Бактерии	493
	Археи	53
Потребность в кислороде	Факультативные анаэробы	156
	Микроаэрофилы	20
	Аэробы	198
Патогенность	Патогенность по отношению к млекопитающим	139
Способ существования	Наземные	56
	Специализированные	78
	Множественные	140
Температурные факторы	Психрофилы	12
	Гипертермофилы	38
	Термофилы	50
	Мезофилы	432
Подвижность	Неподвижные	188
	Малоподвижные	1
	Подвижные	296

Выравнивание проводилось с помощью программы BLAST версии 2.2.26+ (Altschul *et al*., 1990). Использовались установленные по умолчанию параметры выравнивания программы BLAST. Порог E-value, показывающей уровень значимости сходства последовательностей, для отбора гомологов брался равным 10^{-21}. Значение порога было подобрано таким образом, чтобы ошибка перепредсказания (доля ложнопредсказанных как аллергены белков среди неаллергенных белков) была близка к нулю при максимальном значении порога для E-value. Выборка известных белков-аллергенов формировалась из базы данных SDAP (Ivanciuc *et al*., 2003). Из базы данных были экстрагированы 960 последовательностей известных белков-аллергенов.

Расчеты проводились на высокопроизводительном кластере центра коллективного пользования «Биоинформатика» СО РАН. Использование кластера обусловлено размером поставленной задачи. Требовалось предсказать аллергенность более 1,76 млн белков с известной последовательностью из 546 видов микроорганизмов. Использовалось распараллеливание по данным. Все белки были сгруппированы по их принадлежности к видам микроорганизмов. Вычисления для каждой группы проводились параллельно на 24 узлах с 8 ядрами на каждом. Общее время счета составило около 63 ч. В связи с суперпараллельностью задачи было достигнуто практически линейное ускорение.

Аллергенность микроорганизма рассчитывалась как отношение количества белков, предсказанных как аллергены, к общему числу белков в протеоме данного микроорганизма.

Для установления связи между характеристиками микроорганизмов (табл. 1) и их аллергенностью использовались методы множественной линейной регрессии и дисперсионный анализ, представленные в стандартной библиотеке языка R (R Development Core Team, 2011). Характеристики микроорганизмов, представленные в табл. 1, рассматривались как независимые

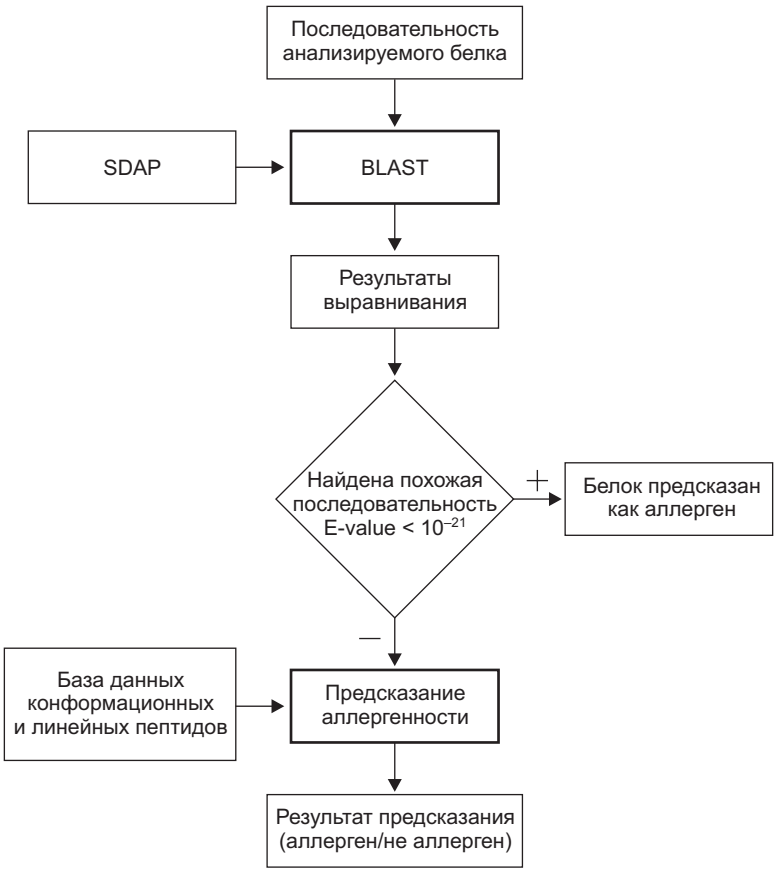

Рис. Блок-схема подхода для анализа аллергенности белков микроорганизмов.

переменные, а аллергенность – как зависимая переменная. Независимые переменные принимали значения 0 или 1. Например, если вид архей или бактерий относился к термофилам, то характеристика «термофильность» этого микроорганизма задавалась как 1, в противном случае – как 0.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Результаты дисперсионного анализа связи между характеристиками микроорганизмов и их аллергенностью приведены в табл. 2.

С помощью дисперсионного анализа были выявлены 4 наиболее значимые характеристики микроорганизмов, связанные с аллергенностью (см табл. 2). Среди них на первом месте по значимости оказалась патогенность микроорганизма. При этом, согласно нашим предсказаниям, патогенные микроорганизмы обладают повышенной аллергенностью. *Staphylococcus aureus* является одним из примеров патогенных микроорганизмов, для которых есть данные, свидетельствующие об их аллергенности. Так, в работе K. Reginald с соавт. показано, что IgE-антитела человека, страдающего атопическим дерматитом, могут взаимодействовать с белками *S. aureus*, что является показателем возможной аллергенности таких белков (Reginald *et al*., 2011). Известно также, что у людей, страдающих аллергией, могут наблюдаться изменения в составе биоценоза кишечника, в котором могут присутствовать условно-патогенные бактерии родов *Staphylococcus*, *Klebsiella*, *Proteus* (Макарова, Боровик, 2008). В связи с этим интерес могут представлять дальнейшие исследования, направленные на выяснение связи между аллергенностью и патогенностью микроорганизмов.

Таблица 2

Характеристики микроорганизмов, статистически значимо связанные с аллергенностью

Характеристика микроорганизма*	Направление связи с аллергенностью	Статистика Фишера	Значимость
Патогенность**	+	31,7	3,447e–08
Наземный способ существования	–	21,1	5,683e–06
Аэробы	–	12,9	0,0003571
Мезофилы	–	12,6	0,0004324

* Приведены характеристики с уровнем значимости менее 0,001. ** Патогенность рассматривалась по отношению к млекопитающим.

Другие выявленные нами менее значимые факторы (наземный способ существования, принадлежность к аэробам и мезофилам) оказались отрицательно связанными с аллергенностью. Среди микроорганизмов, ведущих наземный способ существования, а также мезофилов и аэробов, согласно предсказанию, аллергенные белки встречаются реже по сравнению с остальными микроорганизмами. Следует отметить, что такие микроорганизмы могут чаще других микроорганизмов, обитающих в экстремальных условиях, контактировать с людьми.

Согласно нашим предсказаниям, наиболее аллергенными среди патогенных для человека микроорганизмов могут считаться *Bacillus*

Таблица 3

Примеры некоторых видов патогенных для человека бактерий, для которых предсказано наибольшее значение показателя аллергенности

Вид	Заболевание	Показатель аллергенности
Bacillus weihenstephanensis KBAB4	Пищевое отравление	0,1009
Streptococcus pneumoniae 670-6B	Пневмония	0,0841
Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305	Инфекции мочевыводящих путей	0,0819
Staphylococcus haemolyticus JCSC1435	Широкий спектр оппортунистических инфекций	0,0812

weihenstephanensis, *Streptococcus pneumonia*, *Staphylococcus saprophyticus* и *Staphylococcus haemolyticus*, которые могут вызывать пищевое отравление (Lapidus *et al*., 2008), пневмонию (Jauneikaite *et al*., 2012), инфекции мочевыводящих путей (Kuroda *et al*., 2005) и оппортунистические инфекции (Takeuchi *et al*., 2005) соответственно (см. табл. 3).

Полученные результаты могут быть использованы при планировании экспериментов по проверке аллергенных свойств белков микроорганизмов.

БЛАГОДАРНОСТИ

Работа была поддержана грантом РФФИ «Компьютерный анализ взаимосвязи аллергенности микроорганизмов и условий их обитания» № 12-04-31892-мол_а, DAAD Leonard Euler Program grant Nr. 50024820 (AB, ET, TI) и Министерством образования и науки РФ (Госконтракт № 14.740.11.0001). Коллектив авторов выражает благодарность Ольге Владимировне Сайк за помощь в написании статьи.

ЛИТЕРАТУРА

Брагин А.О., Деменков П.С., Иванисенко В.А. Предсказание аллергенности белков с использованием информации о конформационных пептидах // Вавилов. журн. генет. и селекции. 2011. Т. 15. № 3. С. 462–468.

Макарова С.Г., Боровик Т.Э. Дисбиоз кишечника у детей с пищевой аллергией: патогенетические аспекты и современные методы коррекции // Вопр. соврем. педиатрии. 2008. Т. 7. № 2. С. 82–92.

Altschul S.F., Gish W., Miller W. *et al*. Basic local alignment search tool // J. Mol. Biol. 1990. V. 215. No. 3. P. 403–410.

Antranikian G., Vorgias C.E., Bertoldo C. Extreme environments as a resource for microorganisms and novel biocatalysts // Adv. Biochem. Eng. Biotechnol. 2005. V. 96. P. 219–262.

Barrett T., Clark K., Gevorgyan R. *et al*. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata // Nucl. Acids Res. 2012. V. 40. (Database issue) D57–63.

Benson D.A., Karsch-Mizrachi I., Lipman D.J. *et al*. GenBank // Nucl. Acids Res. 2011. V. 39. (Database issue) D32–37.

Bischoff S.C., Krämer S. Human mast cells, bacteria, and intestinal immunity // Immunol. Rev. 2007. V. 217. No. 1. P. 329–337.

Breiteneder H., Mills E.N. Molecular properties of food allergens // J. Allergy Clin. Immunol. 2005. V. 115. No. 1. P. 14–23.

Cianferoni A., Spergel J.M. Food allergy: review, classification and diagnosis // Allergol. Int. 2009. V. 58. P. 457–466.

FAO/WHO. Codex Principles and Guidelines on Foods Derived from Biotechnology. 2003.

Goodman R.E., Vieths S., Sampson H.A. *et al*. Allergenicity assessment of genetically modified crops – what makes sense? // Nat. Biotechnol. 2008. V. 26. No. 1. P. 73–81.

Huby R.D., Dearman R.J., Kimber I. Why are some proteins allergens? // Toxicol. Sci. 2000. V. 55. No. 2. P. 235–246.

Irwin J.A. Extremophiles and their application to veterinary medicine // Environ. Technol. 2010. V. 31. P. 857–869.

Ivanciuc O., Schein C.H., Braun W. SDAP: database and computational tools for allergenic proteins // Nucl. Acids Res. 2003. V. 31. No. 1. P. 359–362.

Jauneikaite E., Jefferies J.M., Hibberd M.L., Clarke S.C. Prevalence of *Streptococcus pneumoniae* serotypes causing invasive and non-invasive disease in South East Asia: a review // Vaccine. 2012. V. 30. No. 24. P. 3503–3514.

Jeebhay M.F., Robins T.G., Lehrer S.B., Lopata A.L. Occupational seafood allergy: a review // Occup. Environ. Med. 2001. V. 58. No. 2. P. 553–562.

Jenkins J.A., Breiteneder H., Mills E.N. Evolutionary distance from human homologs reflects allergenicity of animal food proteins // J. Allergy Clin. Immunol. 2007. V. 120. No. 6. P. 1399–1405.

Kong W., Tan T.S., Tham L., Choo K.W. Improved prediction of allergenicity by combination of multiple sequence motifs // In Silico Biol. 2007. V. 7. No. 1. P. 77–86.

Kuroda M., Yamashita A., Hirakawa H. *et al*. Whole genome sequence of *Staphylococcus saprophyticus* reveals the pathogenesis of uncomplicated urinary tract infection // Proc. Natl Acad. Sci. USA. 2005. V. 102. No. 37. P. 13272–13277.

Lapidus A., Goltsman E., Auger S. *et al*. Extending the *Bacillus cereus* group genomics to putative food-borne pathogens of different toxicity // Chem. Biol. Interact. 2008. V. 171. No. 2. P. 236–249.

Li K.B., Issac P., Krishnan A. Predicting allergenic proteins using wavelet transform // Bioinformatics. 2004. V. 20. No. 16. P. 2572–2578.

Locksley R.M. Asthma and allergic inflammation // Cell. 2010. V. 140. No. 6. P. 777–783.

MacDonald A.S., Araujo M.I., Pearce E.J. Immunology of parasitic helminth infections // Infect. Immun. 2002. V. 70. No. 2. P. 427–433.

Muh H.C., Tong J.C., Tammi M.T. AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins // PLoS One. 2009. V. 4. No. 6. e5861.

Pennisi E. In industry, extremophiles begin to make their mark // Science. 1997. V. 276. No. 5313. P. 705–706.

Platts-Mills T.A. Allergy in evolution // Chem. Immunol. Allergy. 2012. V. 96. P. 1–6.

Platts-Mills T.A., Vaughan J.W., Carter M.C., Woodfolk J.A. The role of intervention in established allergy: avoidance of indoor allergens in the treatment of chronic allergic disease // J. Allergy Clin. Immunol. 2000. V. 106. No. 5. P. 787–804.

Puc M. Characterisation of pollen allergens // Ann. Agric. Environ. Med. 2003. V. 10. No. 2. P. 143–149.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical

Computing. Vienna, Austria. 2011. URL http://www.R-project.org/.

Reginald K., Westritschnig K., Werfel T. *et al*. Immunoglobulin E antibody reactivity to bacterial antigens in atopic dermatitis patients // Clin. Exp. Allergy. 2011. V. 41. No. 3. P. 357–369.

Saha S., Raghava G.P. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes // Nucl. Acids Res. 2006. 34(Web Server issue):W202–209.

Stadler M.B., Stadler B.M. Allergenicity prediction by protein sequence // FASEB J. 2003. V. 17. No. 9. P. 1141–1143.

Sudha V.T., Arora N., Singh B.P. Serine protease activity of Per a 10 augments allergeninduced airway inflammation in a mouse model // Eur. J. Clin. Invest. 2009. V. 39. No. 6. P. 507–516.

Takeuchi F., Watanabe S., Baba T. *et al*. Whole-genome sequencing of staphylococcus haemolyticus uncovers the extreme plasticity of its genome and the evolution of human-colonizing staphylococcal species // J. Bacteriol. 2005. V. 187. No. 21. P. 7292–7308.

Van den Burg B. Extremophiles as a source for novel enzymes // Curr. Opin. Microbiol. 2003. V. 6. No. 3. P. 213–218.

WAO White Book on Allergy / Eds R. Pawankar, G.W. Canonica, S.T. Holgate, R.F. Lockey. Milwaukee, Wisconsin: World Allergy Organization, 2011. P. 12.

Zorzet A., Gustafsson M., Hammerling U. Prediction of food protein allergenicity: a bioinformatic learning systems approach // In Silico Biol. 2002. V. 2. No. 4. P. 525–534.

COMPUTERIZED ANALYSIS OF THE RELATIONSHIP BETWEEN ALLERGENICITY OF MICROORGANISMS AND THEIR HABITATS

A.O. Bragin[1], P.S. Demenkov[1], E.S. Tiys[1], R. Hofestädt[2], V.A. Ivanisenko[1], N.A. Kolchanov[1, 3, 4]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: ibragim@bionet.nsc.ru;
[2] Bio-Medical Informatics Department, Bielefeld University, Bielefeld, Germany;
[3] Novosibirsk National Research State University, Novosibirsk, Russia;
[4] National Research Centre «Kurchatov Institute», Moscow, Russia

Summary

The prevalence of allergic diseases was rapidly increasing in the 20th century. Currently, many people suffer from allergy in industrial countries. Therefore, analysis of allergenic properties of proteins is an urgent task. The following factors were formerly hypothesized to determine the allergenicity of a protein: size, enzymatic properties, and similarity to human proteins. However, no analysis of the relationship between allergenicity of proteins and the habitat of the organisms producing them has been conducted hitherto. We predict allergenicity of proteins from proteomes of more than 500 species of microorganisms. It is shown that the number of allergenic proteins in the proteomes of microorganisms is significantly associated with their pathogenicity, habitat, temperature conditions of the habitat, and oxygen demand.

Key words: protein allergenicity prediction, allergenicity of microbial proteomes, habitat of microorganisms, extremophiles.

УДК 577.38 577.3.0 577.322.4

РАСПРЕДЕЛЕННАЯ СИСТЕМА RESTful-WEB-СЕРВИСОВ ДЛЯ РЕКОНСТРУКЦИИ И АНАЛИЗА ГЕННЫХ СЕТЕЙ

© 2012 г. **Н.Л. Подколодный[1], А.В. Семенычев[1], Д.А. Рассказов[1], В.Г. Боровский[1], Е.А. Ананько[1], Е.В. Игнатьева[1], Н.Н. Подколодная[1], О.А. Подколодная[1], Н.А. Колчанов[1, 2, 3]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: pnl@bionet.nsc.ru;
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия;
[3] НИЦ «Курчатовский институт», Москва, Россия

Поступила в редакцию 5 июля 2012 г. Принята к публикации 25 июля 2012 г.

В данной работе описывается распределенный программный комплекс на основе RESTful-Web-сервисов, который ориентирован на решение задач реконструкции генных сетей на основе интеграции данных из гетерогенных источников информации, включая базы данных о молекулярно-генетических взаимодействиях, метаболических и сигнальных путях, генных сетях и т. д.

Ключевые слова: распределенные системы, RESTful-Web-сервисы, генные сети, интеграция данных, анализ графов генных сетей.

ВВЕДЕНИЕ

Молекулярно-генетические системы характеризуются огромным разнообразием молекулярных механизмов, обеспечивающих их функционирование, включая транскрипцию, процессинг (созревание) РНК, трансляцию (синтез полипептидных цепей), процессинг белков, ДНК-белковые, РНК-белковые, белок-белковые, лиганд-белковые и другие взаимодействия, процессы метаболизма, передачи сигналов, транспорта, деградации и т. д.

К настоящему времени в области биоинформатики и системной биологии мировым сообществом разработано более 1400 баз данных, многие из которых полезны при описании молекулярно-генетических систем и генетических механизмов их функционирования, включая базы данных по молекулярным объектам (гены, РНК, белки), молекулярно-генетическим взаимодействиям и процессам, онтологиям, метаболическим путям и путям передачи сигналов в клетке, генетической регуляции молекулярных процессов и систем, генным сетям, экспрессии генов в различных клеточных условиях и под действием различных индукторов и т. д. (Galperin, 2012).

Интеграция данных из этих Интернет-доступных гетерогенных источников информации о молекулярно-генетических взаимодействиях и генных сетях и реконструкция на этой основе генных сетей являются важнейшей задачей биоинформатики и системной биологии.

Генные сети – это молекулярно-генетические системы, обеспечивающие формирование фенотипических характеристик организмов (молекулярных, биохимических, физиологических, морфологических, поведенческих и т. д.) на основе информации, закодированной в их геномах (Kolpakov *et al*., 1998; Ananko, 2005). Обычно генные сети состоят из сотен и тысяч элементов, объединенных сложными процессами взаимодействия.

Анализ структуры генных сетей, выявление закономерностей структурно-функциональной организации генных сетей, выделение подсистем, редукция описания молекулярно-генети-

ческих систем и построение на этой основе структурной модели генной сети являются важнейшими этапами в исследовании генных сетей и первым шагом в создании математических моделей динамики генных сетей (Newman, 2006).

Проблемы реконструкции и анализа структурно-функциональной организации генных сетей

Современные технологии реконструкции генных сетей основываются на:

1) использовании специализированных графических редакторов, обеспечивающих возможность пользователю вводить и редактировать информацию о молекулярно-генетических объектах, реакциях, генетических регуляциях, генных сетях;

2) интеграции данных о молекулярно-генетических взаимодействиях из различных источников информации (баз данных);

3) использовании методов теоретического предсказания молекулярно-генетических взаимодействий.

Для этих целей крайне важной информацией является онтологическое описание молекулярно-генетических объектов, систем и процессов, включая классификацию генов, белков, ферментов, метаболических путей и генных сетей.

Обычно процесс реконструкции генной сети или метаболического пути начинается с постановки задачи, которая включает описание проблемы, выявление целей, определение границ задачи, класса реконструируемой генной сети, формирование запроса на поиск начального множества элементов генов сети и т. п.

При реконструкции генной сети необходимо учитывать, что генная сеть – это сетевая модель взаимодействий функционирующих генов, которая описывает молекулярно-генетическую систему, обеспечивающую выполнение определенной функции клетки или ее подсистем при определенных условиях, состояниях организма или клетки, при воздействии внешних индукторов, взаимодействии с другими клетками или организмами, реализации определенных молекулярных событий на определенных стадиях молекулярно-генетических процессов и т. д.

Последовательность ДНК дает только комбинаторику возможных вариантов функционирования генов. Следующие уровни регуляции работы генов в клетке (состояние хроматина, различные модификации ДНК, гистонов, структура и локализация хромосомы и т. д.) задают специфические ограничения на эти возможности. Для каждой клетки эти ограничения в общем случае различаются и могут динамически изменяться с собственными характерными временами. Таким образом, работа генов существенно зависит от типа и состояния клетки. Поэтому при реконструкции генных сетей необходимо заранее определить тип клеток, тканей либо вид организма, в которых она реализуется. Информация, полученная на других типах клеток, тканях и организме, также может быть полезна при реконструкции генной сети. Однако необходимо обосновать такую возможность и оценить степень адекватности этой информации. Более того, внешние условия, при которых были получены экспериментальные данные, также должны быть строго определены. Только в этом случае можно оценить возможность их интеграции и совместного анализа.

Генные сети можно разделить на пять основных типов (Колчанов и др., 2000): контролирующие гомеостаз; регулирующие циклические процессы; обеспечивающие стрессовый ответ; контролирующие необратимые процессы и генные сети-интеграторы.

Разные типы генных сетей имеют различные структуру и состав, которые необходимо учитывать в процессе реконструкции. Например, генные сети, описывающие поддержание некоторого гомеостатического состояния, должны включать сенсорный элемент, отслеживающий состояние гомеостатируемого параметра, а также пути поступления и утилизации продуктов, определяющих состояние этого параметра.

Генная сеть, описывающая реакцию клетки на некоторое воздействие или сигнал, должна включать рецепторы, через которые этот сигнал передается компонентам генной сети, сам путь передачи сигнала и исполняющие элементы.

Одним из первых шагов в реконструкции генной сети может быть определение множества генов, которые задают ядро реконструируемой генной сети или метаболические пути, которые должны входить в генную сеть.

Проблемы анализа структурно-функциональной организации генных сетей

Реализация RESTful-Web-сервисов для реконструкции и анализа генных сетей REST (REpresentational State Transfer) представляет собой архитектурный стиль для создания ресурс-ориентированных распределенных программных систем, основанных на архитектуре клиент–сервер и, как правило, используется для построения Web-сервисов или RESTful-Web-сервисов (Richardson, 2007).

Архитектурный стиль REST включает ряд рекомендаций или ограничений, налагаемых на архитектуру, оставляя реализацию индивидуальных компонентов свободной (Richardson, 2007; Valverde, 2009; Subbu Allamaraju, 2010; Schreier, 2011).

Адресуемость. Основным понятием в REST-архитектуре являются ресурсы как источники конкретной информации, каждый из которых определяется ссылкой с глобальным идентификатором URI. Ресурсом является все, что имеет ссылку.

Отсутствие состояния сервиса. Вся информация, необходимая для выполнения запроса, содержится в самом запросе. Информация о предыдущих запросах сервером не сохраняется и не используется. Сервер не поддерживает сеанс и не фиксирует его состояние. Вся информация о состоянии сессии поддерживается при необходимости клиентом. Однако сам ресурс имеет определенное состояние, которое может изменяться в результате выполнения запросов клиента или по другим внешним причинам. Состояние на стороне сервера адресуемо через URI как ресурс. Это делает серверы не только более видимыми для мониторинга, но и более надежными в случае частичного отказа сети, а также дополнительно улучшает их масштабируемость.

Связность. Информация о связях между ресурсами может использоваться клиентом для обнаружения идентификаторов других связанных с запросом ресурсов, в том числе ссылок на автоматически созданные ресурсы, которые являются результатом обработки данных.

Унифицированный интерфейс. REST требует использования унифицированного интерфейса, включающего множество операций или методов с известной семантикой, которые изменяют состояние ресурса. Интерфейс зависит от схемы URI. Для http это методы GET, POST, PUT, DELETE, OPTIONS. Методы являются внешними по отношению к ресурсам и включают посылку стандартных сообщений Web-серверу, указывая URI запрашиваемого ресурса, метод, передаваемые данные или метаданные.

Ресурс может иметь множественное представление, которое соответствует стандартизованному формату или типу (MIME-type), и может предоставляться Web-сервером.

Понятие «RESTful» употребляют для описания сервисов, которые построены с учетом архитектуры REST и не нарушают ни одну из ее нотаций. Соблюдение этих ограничений и, следовательно, соответствие архитектурному стилю REST позволят любой распределенной системе иметь требуемые свойства, такие, как производительность, масштабируемость, простота, модифицируемость, видимость, мобильность и надежность.

Архитектура разработанной системы представляет собой модульную систему, основанную на центральном модуле, так называемом «ядре системы». Он представляет собой основу системы и содержит в себе модель данных, набор программных инструментов для работы с ней, а также прямой доступ к интегрированной базе данных (БД) и доступ к внешним БД. При сборке программного комплекса (ПК) ядро целиком помещается в программный компонент, тем самым делая его независимым приложением. Схематическое описание архитектуры ПК изображено на рис. 1.

Одним из основных преимуществ данного подхода является поддержка расширяемости системы. Добавление новых компонент в систему требует лишь использовать универсальное ядро системы для получения всего основного функционала, необходимого при создании новой компоненты.

Интегрированная база включает в себя словари, содержащие унифицированные имена сущностей для описания генных сетей (гены, белки, метаболиты и др.), и информацию о молекулярно-генетических взаимодействиях, реакциях и регуляциях в генных сетях из гетерогенных источников информации, которая представлена в виде нескольких уровней описания. Первый уровень – «сырые данные», представленные в

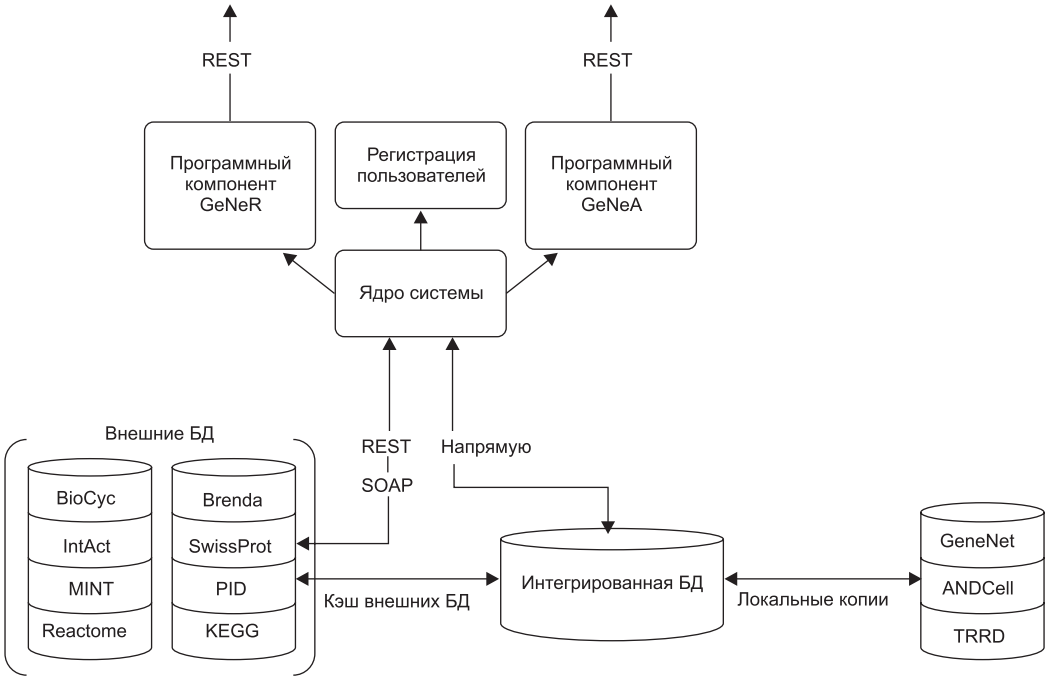

Рис. 1. Архитектура системы реконструкции и анализа генных сетей.

форматах источника данных. Второй уровень – привязанные к унифицированным именам и идентификаторам предварительно обработанные данные, включающие проекцию на схему интегрированной базы данных. БД по генным сетям реализована под управлением СУБД Oracle 11g (Гринвальд, 2009; Кайт, 2011).

Программный компонент GeNeR для реконструкции генных сетей

Программный компонент GeNeR отвечает за реконструкцию генных сетей в ПК. Схематическое описание архитектуры модуля представлено на рис. 2.

Программный компонент GeNeR предоставляет три типа сервисов в стиле REST для внешнего доступа:

– Сервис доступа к внешним источникам данных по молекулярно-генетическим взаимодействиям.

– Сервис интеграции данных и реконструкции генных сетей.

– Сервис доступа к интегрированной БД (есть по умолчанию во всех компонентах).

Данный компонент активно использует менеджер доступа к внешним базам данных по молекулярно-генетическим взаимодействиям, метаболическим путям и генным сетям. Для каждой базы данных пишется свой уникальный драйвер, который обрабатывает унифицированные запросы от системы и запрашивает данные во внешнем источнике. Каждый драйвер несет в себе подробную метаинформацию по базе данных, такую, как: название БД, краткое описание, список поддерживаемых методов доступа и форматов. Указав соответствующий заголовок запроса к сервису на получение метаинформации, можно получить ответ в формате XML.

Программный комплекс позволяет выполнять унифицированные запросы к внешним базам данных по молекулярно-генетическим взаимодействиям, включая базы данных GeneNet (Ananko, 2005), TRRD (Kolchanov, 2008), KEGG (Wrzodek *et al*., 2011), SWISS-PROT, Pathway Interaction Database (Schaefer, 2009), IntAct (Aranda, 2009), REACTOME (Croft, 2011), MINT (Chatr-Aryamontri, 2007), bioCyc (Caspi *et al*., 2010), BRENDA (Scheer, 2011).

Реализованы средства унифицированных запросов к интегрированной базе данных, данным из внешних источников, сервису интеграции данных и реконструкции генных сетей.

Общий формат запроса к сервису доступа к данным из внешних источников: https://host/rws/extdbs/{dbname}/{entity}?{query}.

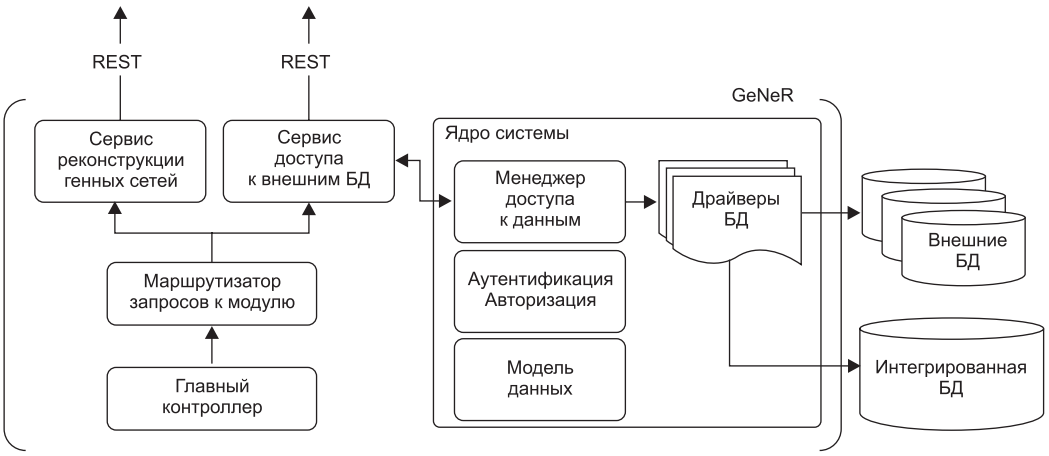

Рис. 2. Архитектура модуля GeNeR.

Приведем некоторые примеры использования сервиса в формате REST запросов:

GET https://host/rws/extdbs/ – выдает список всех внешних источников, представленных в системе.

GET https://host/rws/extdbs/{dbname}/ – выдает информацию по конкретному внешнему источнику данных. В набор информации входят: название БД, краткое описание, список поддерживаемых сущностей, организмов и типов данных.

GET https://host/rws/extdbs/{dbname}/{entity}?{query} – выдает список конкретных записей из указанной БД. Дополнительные параметры для фильтрации записей передаются в {query}.

Общий формат запроса к сервису доступа к интегрированной базе данных: https://host/rws/idb/pathways/{pathway}?{query}.

Подробное описание использования сервиса в формате REST запросов:

GET https://host/rws/idb/ – выдает информацию по интегрированной БД.

GET https://host/rws/idb/pathways/ – выдает список генных сетей, содержащихся в интегрированной БД.

POST https://host/rws/idb/pathways/ – создает новую генную сеть и возвращает ссылку на ее ресурс.

GET https://host/rws/idb/pathways/{pathway}?{query} – выдает конкретную генную сеть из интегрированной БД.

В качестве примеров можно привести следующие запросы к ПК.

Пример. Получить список биохимических реакций у человека, в которых участвует нитрооксид («Nitric oxide»).

Запрос: GET https://host/rws/dbs/kegg/hsa/reaction?metabolite=”Nitric oxide”.

Результаты запроса представлены на рис. 3.

Программный компонент GeNeA для анализа графов генных сетей

Программный компонент GeNeA для анализа графов генных сетей представляет собой набор Web-сервисов, реализованных на языке Java и обеспечивающих запуск прикладных программных модулей анализа структуры генных сетей. Основными данными, которыми оперирует прикладной программный модуль, являются графы генных сетей, представленные как список ребер (соединенных вершин), векторы и матрицы.

Для передачи входных и выходных данных при запуске модуля используются стандартные потоки ввода/вывода. Это обеспечивает возможность быстрого подключения новых программных модулей в систему анализа графов генных сетей. Прикладной программный модуль для анализа графа генных сетей реализован на языке Java и поддерживает унифицированный интерфейс вызова процедур анализа графов, который позволяет подключать внешние библиотеки, реализованные на различных языках программирования, и использовать различные форматы представления графа. В частности,

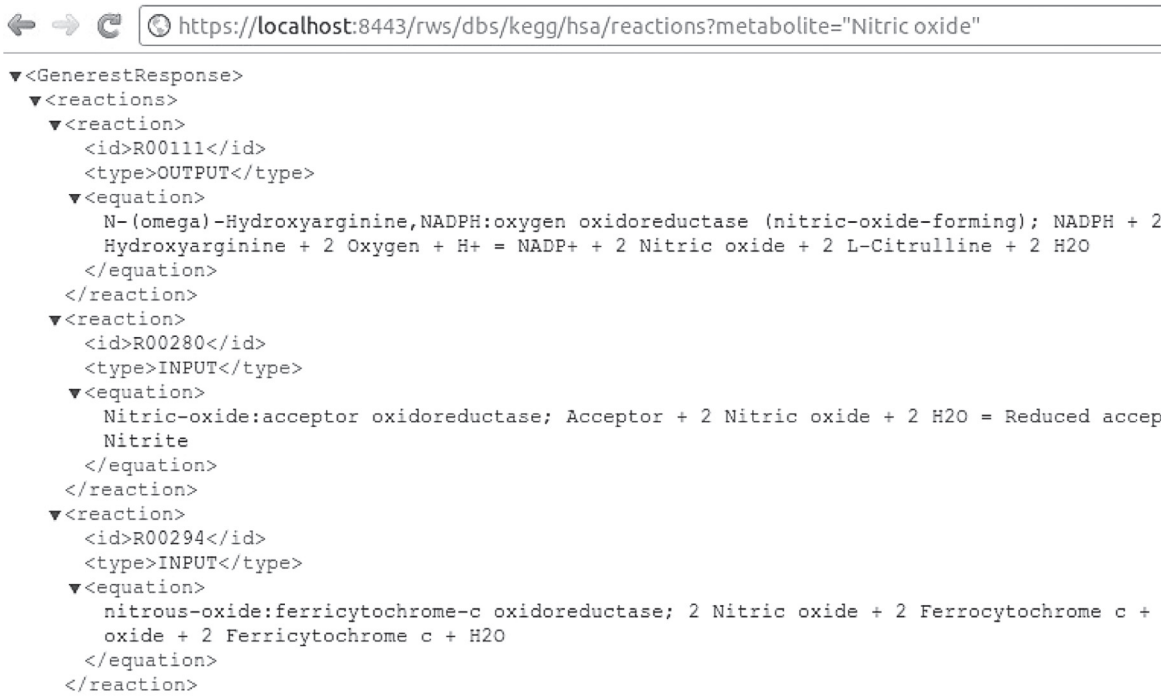

Рис. 3. Результат выполнения запроса списка реакций у человека, в которых участвует нитрооксид («Nitric oxide») (3 реакции).

нами используется библиотека igraph, в которой включено большое число типовых алгоритмов анализа графов (Csárdi, Nepusz, 2006a, b).

Схематическое описание архитектуры модуля GeNeA изображено на рис. 4. Здесь представлены следующие типы сервисов в стиле REST для внешнего доступа:

– Сервис анализа графов генных сетей.

– Сервис визуализации генных сетей.

– Сервис отправки заданий на вычислительный кластер.

– Сервис доступа к интегрированной БД (есть по умолчанию во всех компонентах).

В рамках программного компонента реализован унифицированный доступ к проблемно-ориентированным вычислительным сервисам для анализа графов генных сетей, которые обеспечивают следующие возможности:

– поддержка интроспекции (получение описания сервисов по запросу клиента);

– поддержка асинхронной обработки запросов, требующих длительных вычислений;

– поддержка передачи параметров запроса и результатов в виде файлов;

– использование архитектурного стиля REST и распространенных форматов представления данных XML, JSON.

ЗАКЛЮЧЕНИЕ

Разработан распределенный программный комплекс на основе RESTful-Web-сервисов, который ориентирован на решение задач реконструкции генных сетей на основе интеграции данных из гетерогенных источников информации, включая базы данных о молекулярно-генетических взаимодействиях, метаболических и сигнальных путях, генных сетях. Программный комплекс включает модули для расчета различного рода характеристик графа генных сетей, в частности: распределение степеней вершин, коэффициенты кластеризации, диаметр графа, плотность графа, индекс центральности, индекс Боначича, индекс Фримана, спектр графа, поиск структурных мотивов, поиск циклов в графе генных сетей и др.

Анализ структуры генных сетей, выявление закономерностей структурно-функциональной организации генных сетей, выделение подсистем, редукция описания молекулярно-генетических и построение на этой основе структурной модели генной сети являются важнейшими этапами исследования генных сетей и первым шагом в создании математических моделей динамики генных сетей.

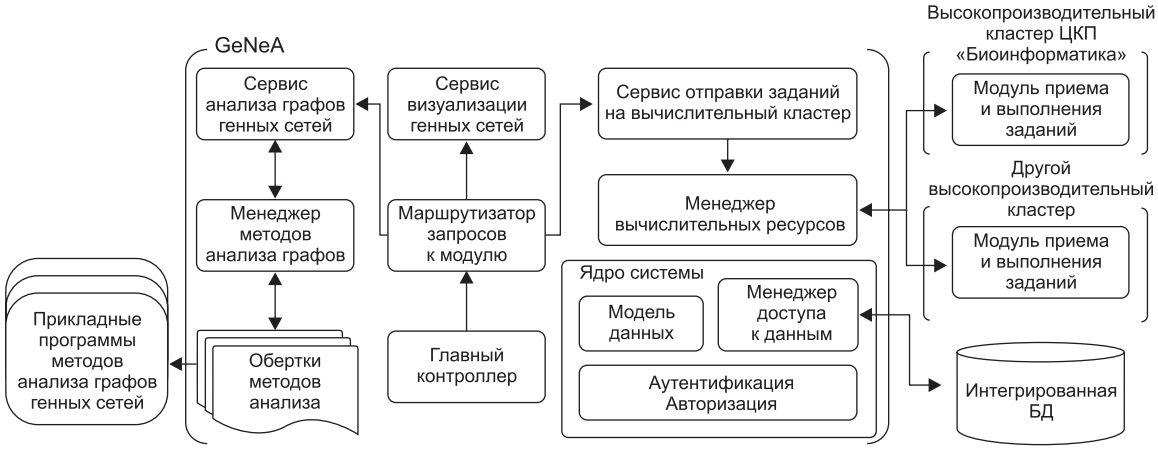

Рис. 4. Архитектура модуля GeNeA.

Работа поддержана Министерством образования и науки РФ (Госконтракт № 07.514.11.4023 по теме «Проектирование и разработка RESTful-Web-сервисов для создания распределенной инфраструктуры, ориентированной на решение задач реконструкции и анализа генных сетей»).

ЛИТЕРАТУРА

Гринвальд Р., Стаковьяк Р., Стерн Д. Oracle 11g. Основы. СПб.: Символ-плюс, 2009. 464 с.

Кайт Т. Oracle для профессионалов: архитектура, программирование и особенности версий 9i, 10g и 11g. «ВИЛЬЯМС», 2011. 848 с.

Колчанов Н.А., Ананько Е.А., Колпаков Ф.А. и др. Генные сети // Молекуляр. биология. 2000. Т. 34. № 4. С. 533–544.

Ananko E.A., Podkolodny N.L., Stepanenko I.L. *et al*. GeneNet in 2005 // Nucl. Acids Res. 2005. V. 33. D425–D427.

Aranda B., Achuthan P., Alam-Faruque Y. *et al.* The IntAct molecular interaction database in 2010 // Nucl. Acids Res. 2009. V. 38. D525–D531.

Caspi R., Altman T., Dale J.M. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases // Nucl. Acids Res. 2010. V. 38. P. 473–479.

Chatr-Aryamontri A., Ceol A., Palazzi L.M. *et al.* MINT: the Molecular INTeraction database // Nucl. Acids Res. 2007. V. 35. P. 572–574.

Croft D., O'Kelly G., Wu G., Haw R. *et al.* Reactome: a database of reactions, pathways and biological processes // Nucl. Acids Res. 2011. V. 39. P. 691–697.

Csárdi G., Nepusz T. The igraph software package for complex network research // Intern. J. Complex Syst. 2006a. V. 1695.

Csárdi G., Nepusz T. igraph Reference Manual // 29-33 Konkoly-Thege Miklys road, Budapest H-1121, Hungary, 509 p. 2006b. http://igraph.sourceforge.net/doc/igraph-docs.pdf

Galperin M.Y., Fernández-Suárez X.M. The 2012 nucleic acids research database issue and the online molecular biology database collection // Nucl. Acids Res. 2011. V. 40. D1-D8.

Kolchanov N.A., Ignatieva E.V., Podkolodnaya O.A. *et al.* TRRD: Technology for extraction, storage, and use of knowledge about the structural-functional organization of the transcriptional regulatory regions in the eukaryotic genes // Intell. Data Anal. 2008. V. 12. No. 5. P. 443–461.

Kolpakov F.A., Ananko E.A., Kolesov G.B., Kolchanov N.A. GeneNet: a database for gene networks and its automated visualization // Bioinformatics. 1998. V. 14. No. 6. P. 529–537.

Newman M.E.J. Finding community structure in networks using the eigenvectors of matrices // Phys. Rev. 2006. E 74, 036104.

Richardson L., Ruby S. RESTful Web Services. O'Relly Media. Inc., 1005 Gravensein Highway North, Sebastopol, CA, 2007. 420 c.

Schaefer C.F., Anthony K., Krupa S. *et al.* PID: the pathway interaction database // Nucl. Acids Res. 2009. V. 37. P. 674–679.

Scheer M., Grote A., Chang A. *et al.* BRENDA, the enzyme information system in 2011 // Nucl. Acids Res. 2011. V. 39. P. 670–676.

Schreier S. Modeling RESTful applications // Proc. WS-REST '11 Proceedings of the Second Intern. Workshop on RESTful Design. ACM, NY, USA, 2011. P. 15–21.

Subbu Allamaraju RESTful Web Servises Cookbook. O'Relly Media, Inc. 2010. 293 c.

Valverde F., Pastor O. Dealing with REST Services in Model-driven Web Engineering Methods // V Jornadas Científico-Técnicas en Servicios Web y SOA, JSWEB. 2009.

Wrzodek C., Drager A., Zell A. KEGG translator: visualizing and converting the KEGG PATHWAY database to various formats // Bioinformatics. 2011. V. 27. No. 16. P. 2314–2315.

DISTRIBUTED RESTful WEB SERVICES FOR RECONSTRUCTION AND ANALYSIS OF GENE NETWORKS

N.L. Podkolodnyy[1], A.V. Semenychev[1], D.A. Rasskazov[1], V.G. Borowsky[1], E.A. Ananko[1], E.V. Ignatieva[1], N.N. Podkolodnaya[1], O.A. Podkolodnaya[1], N.A. Kolchanov[1, 2, 3]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: pnl@bionet.nsc.ru;
[2] Novosibirsk National Research State University, Novosibirsk, Russia;
[3] National Research Centre «Kurchatov Institute», Moscow, Russia

Summary

This paper describes a RESTful Web service-based distributed software system, which focuses on the reconstruction of gene networks by integrating data from heterogeneous data sources, including databases of molecular-genetic interactions, metabolic and signaling pathways, gene networks, etc.

Key words: distributed systems, RESTful-Web services, gene networks, data integration, gene network graph analysis.

УДК 577:004.4:004.94

НОВЫЕ ВОЗМОЖНОСТИ СИСТЕМЫ MGSmodeller

© 2012 г. **Ф.В. Казанцев[1], И.Р. Акбердин[1], Н.Л. Подколодный[1], В.А. Лихошвай[1, 2]**

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

Поступила в редакцию 15 июля 2012 г. Принята к публикации 15 августа 2012 г.

Актуальной проблемой системной биологии является моделирование сложноорганизованных молекулярно-генетических систем и их анализ, что требует разработки специальных подходов, позволяющих рассматривать эти системы как совокупность динамически взаимодействующих подсистем с более простой структурой. В данной работе освещается подход, направленный на ускорение процесса реконструкции и комплексного анализа математических моделей молекулярно-генетических систем с использованием высокопроизводительного кластера.

Ключевые слова: молекулярно-генетические системы, математические модели, MGSmodeller, высокопроизводительные вычисления.

ВВЕДЕНИЕ

Одним из центральных вопросов системной биологии является выявление молекулярно-генетических механизмов функционирования живых систем. Для изучения их динамических характеристик все больше применяются методы математического и компьютерного моделирования, эффективное использование которых невозможно без учета специфики строения молекулярно-генетических систем (МГС). К таким особенностям относятся: линейный характер кодирования информации, при котором важное значение для моделирования наблюдаемой динамики имеют взаимное расположение генов, промоторов, терминальных и других генетических элементов на молекуле ДНК; топология молекулы ДНК в разные периоды клеточного цикла; существование обратимых состояний белковых молекул, меняющих их свойства; явления полиаллельности гена; анизотропия пространственных компартментов живых систем, разделяющая процессы биохимического синтеза; эпигенетическая передача наследственной информации. Эффективным методом решения описанных выше сложностей является модульный подход, при котором последовательно описываются комплексные подсистемы клетки через их более простые (элементарные) составляющие (Ратнер, 1966; Лихошвай и др., 2001; Karr *et al.*, 2012).

Для корректного моделирования динамики функционирования МГС необходимо разрабатывать подходы, объединяющие элементарные подсистемы в рамках одного стандарта, и средства вычислительного эксперимента, позволяющие исследовать модели живых систем большого размера (100–1000 элементарных подсистем) (Karr *et al.*, 2012). Использование высокопроизводительных вычислительных комплексов дает возможность проводить полномасштабный вычислительный эксперимент для исследования такого рода моделей, включая численный анализ особенностей функционирования МГС.

Существующие подходы моделирования направлены на разделение функциональности между специализированными программными системами. Сначала проводят графическую реконструкцию модели МГС в одном инструменте (Hucka *et al.*, 2003; Funahashi *et al.*, 2003; Sauro *et al.*, 2003), затем осуществляется численный анализ модели в других программных

инструментах, таких, как Jws (Olivier, Snoep, 2004), Cellerator (Shapiro *et al.*, 2002), COPASI (Hoops *et al.*, 2006), Matlab. Интеграция систем осуществляется на базе комплекса программ конверторов. Разделение функциональности между программными модулями позволяет распределять задачи и использовать узкоспециализированные решения на каждом из этапов моделирования. В результате разработка новых технологических решений в виде компьютерных систем для создания и анализа математических моделей на базе существующих специализированных программных модулей обеспечивает богатый инструментарий как для численного исследования модели, так и для репрезентации результатов моделирования (Karr *et al.*, 2012).

В данной работе представлено развитие программной среды моделирования молекулярно-генетических систем MGSmodeller (Kazantsev *et al*., 2008). Богатый функционал системы по реконструкции, исследованию и повторному использованию математических моделей МГС был расширен высокопроизводительными методами их сборки и анализа.

МАТЕРИАЛЫ И МЕТОДЫ

Для реконструкции математических моделей МГС используется система MGSmodeller (Kazantsev *et al.*, 2008; http://modelsgroup.bionet.nsc.ru/?page_id=491). Математические модели реконструируются в формате и по правилам стандарта SibML в рамках обобщенного химико-кинетического подхода (Лихошвай и др., 2001). Анализ результатов моделирования производится средствами системы MGSmodeller и программами Matlab, Gnuplot. Модули компиляции и численного исследования реализованы на языке Fortran. Модули аннотации и редактирования языка SibML, а также постобработки результатов реализованы на языке Java. Расчеты в системе MGSmodeller производились с использованием вычислительного кластера ЦКП «Биоинформатика» СО РАН (bioinformatics.bionet.nsc.ru, www2.sscc.ru/HKC-30T/HKC-30T.htm).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

В настоящей работе представлена версия программы MGSmodeller для работы на высокопроизводительном кластере. Как и версия для персонального компьютера (Kazantsev *et al.*, 2008), система предназначена для моделирования МГС в терминах стандарта SibML (Акбердин и др., 2009). Изменению подверглись модули компиляции моделей и модули анализа результатов численного исследования. Переход на высокопроизводительные технологии был связан с необходимостью изучения масштабных комплексных молекулярно-генетических систем (Mironova *et al*., 2012), что, в свою очередь, невозможно было сделать при использовании мощностей персональных компьютеров за приемлемое время.

Постановка задачи компиляции модели

Математические модели в компьютерной среде MGSmodeller представлены в рамках стандарта SibML как совокупность элементарных подсистем молекулярно-генетических систем. Их реконструкция в рамках среды моделирования производится на основе блочного принципа (Лихошвай и др., 2001). Сначала производится декомпозиция исследуемого объекта до уровня элементарных подсистем, которыми могут быть реакции ферментативного синтеза, подсистемы регуляции экспрессии генов, системы сплайсинга, транспорта, трансляции, процессы созревания и модификации белков, деградации макромолекул и др. Далее описываются математические модели каждой подсистемы, из которых формируется база элементарных моделей. На этой основе исследователь конструирует из элементарных моделей, как из строительных блоков, модель исследуемого объекта. Для этого описывается сценарий сборки модели – файл, содержащий заданную структурно-функциональную организацию модели целевого объекта (Лихошвай и др., 2001), в котором указывается система отношений компартментов (структурный уровень организации целевого объекта) и для каждого компартмента указываются подсистемы, которые должны быть включены в него (функциональный уровень организации объекта).

Предложенный подход позволяет эффективно строить модель исходного целевого объекта, а также модели систем с вариацией структурно-функциональной организации, необходимость рассмотрения которых дикту-

ется текущим исследованием и/или задачами изучения влияния перестроек на уровне компартментной организации системы, изучения влияния мутаций, делеций, вставок и других модификаций, протекающих на генетическом уровне; необходимостью изучения альтернативных механизмов функционирования подсистем и многими другими потребностями, которые могут возникнуть в процессе моделирования.

Этап сборки модели осуществляется как процесс «доопределения» моделей элементарных подсистем – атрибутами, которые осуществляют их однозначную привязку к компартментам; атрибутами, несущими дополнительную информацию о функционировании элементарных моделей в составе комплексной модели целевого объекта. Для комплексных моделей, обладающих сложной структурой, эта операция является ресурсоемкой, требующей постоянного сравнения строковых идентификаторов объектов моделирования. При использовании высокопроизводительного кластера данная операция может осуществляться параллельно для разных частей сценария сборки модели, ускоряя процесс в несколько раз (Казанцев и др., 2012). Например, при исследовании распределения фитогормона ауксина, одного из ключевых морфогенов в регуляции роста и развития растений, по двумерному ансамблю клеток (Mironova *et al.*, 2012) время сборки модели было уменьшено с десятка часов до десятка минут средствами рассматриваемого подхода.

Постановка задачи численного исследования

При постановке задачи численного исследования модели в среде MGSmodeller пользователь может задать временны́е точки для вывода промежуточного результата на печать. Для моделирования молекулярно-генетических систем со сложной структурой начальные концентрации и значения параметров модели можно задавать по шаблону. В частности, для многокомпартментных систем возможно одной операцией установить одинаковую концентрацию одноименных веществ во всех компартментах и только для выбранных компартментов задавать индивидуальное значение концентрации, тем самым производя спецификацию параметра локализации (см. табл. 1).

Для интегрирования дифференциальных уравнений при численном исследовании модели в MGSmodeller используется метод Гира. Также реализована возможность экспорта модели в форматы других систем моделирования (Matlab, Mathematica, SBML), что позволяет проводить дополнительный анализ и сравнение результатов моделирования алгоритмами других систем.

В системе моделирования MGSmodeller добавлена возможность численного исследования нескольких моделей (вариантов одной модели) одновременно в параллельном режиме. На этой базе реализована возможность анализа чувствительности математической модели к варьированию параметров с заданным интервалом и шагом изменения (по принципу «параметры изменяются одновременно» или по принципу последовательного варьирования каждого из параметров).

Дополнительные возможности

В результате численного эксперимента для моделей больших размерностей исследователь, как правило, получает большие объемы информации, и возникает проблема их интерпретации, анализа и визуализации. В случае если не хватает возможностей базовых средств визуализации,

Таблица 1

Инструкции к спецификации значений переменных

Объект в SiBML	Семантика
1,5 ‘<I(Pin3), T(protein)>’	Всем переменным, связанным с функционированием белка *Pin3*, присвоить значение 1,5
0,5 ‘<C(c4,05), I(Pin3), T(protein)>	Переменной, соответствующей концентрации белка *Pin3*, функционирующего в компартменте c4,05, присвоить значение 0,5

в рамках системы MGSmodeller результаты моделирования представлены в структурированном виде. Организация атрибутов переменных модели, задающих ассоциацию с контекстом моделирования (см. табл. 2), позволяет проводить постобработку данных сторонними программами, в том числе, используя специализированные инструменты визуального анализа (http://www.gnuplot.info; Cedilnik *et al*., 2006).

Возможности программного модуля были успешно протестированы в работе В.В. Мироновой и соавт. (Mironova *et al.*, 2012). Была показана высокая эффективность работы модуля, выраженная в минимальных затратах программных и временных ресурсов исследователей на этапах создания и анализа модели.

Построение конвейеров обработки данных

Скорость сборки модели и удобные принципы получения ее различных вариантов важны на первых этапах моделирования. Когда установлена структура математической модели, определен комплекс задач и сформулированы гипотезы, которые планируется проверить с помощью разработанной модели, наступает этап численного эксперимента и анализа полученных результатов. Применение высокопроизводительного подхода на этом этапе моделирования дает преимущество – при численном анализе можно создавать цепочки технологических процессов, которые могут выполняться параллельно. И на выходе будет получена уже обработанная сводная информация.

Например, в работе В.В. Мироновой и соавт. (Mironova *et al.*, 2012) была использована следующая последовательность действий в полуавтоматическом режиме: постановка численного эксперимента для варьирования набора параметров – генерация вариантов модели с различными значениями параметров – проведение численных экспериментов – постобработка данных – реорганизация данных – построение различных типов кривых и результирующей поверхности решений для двухмерного массива клеток (рис.). Временные затраты на проведение одного численного эксперимента изменялись от десятка минут до нескольких часов (в зависимости от значения параметров). Было проведено несколько тысяч экспериментов, расчет которых в совокупности занял двое суток.

При исследовании динамики функционирования МГС могут потребоваться методы обработки и визуализации результатов расчетов, отличные от типовых, встроенных в систему. Удобный способ описания результатов расчетов, представленный в статье, существенно уменьшает время создания собственных инструментов и встраивания их в технологическую цепочку процесса обработки данных.

ЗАКЛЮЧЕНИЕ

Исследование молекулярно-генетических механизмов функционирования живых систем с высокой степенью подробности требует рассмотрения и учета особенностей функционирования сотен элементарных подсистем в одно и то же время в различных компартментах системы и т. д. Подготовка данных к этапу анализа в масштабных моделях является отдельной нетривиальной задачей, решение которой заключается в организации модели и данных

Таблица 2

Примеры объектов языка моделирования

Объект в SiBML	Семантика
<C(nucleus), I(CYP79B2), T(gene)>	Переменная, описывающая функционирование гена *CYP79B2*, локализованного в компартменте nucleus
<C(nucleus), I(CYP79B2), T(rna)>	Переменная, описывающая функционирование РНК гена *CYP79B2*, находящейся в компартменте nucleus
<T(protein), I(IAMT1p)>	Переменная, описывающая функционирование белка IAMT1p (в однокомпартментной модели, название компартмента опускается)

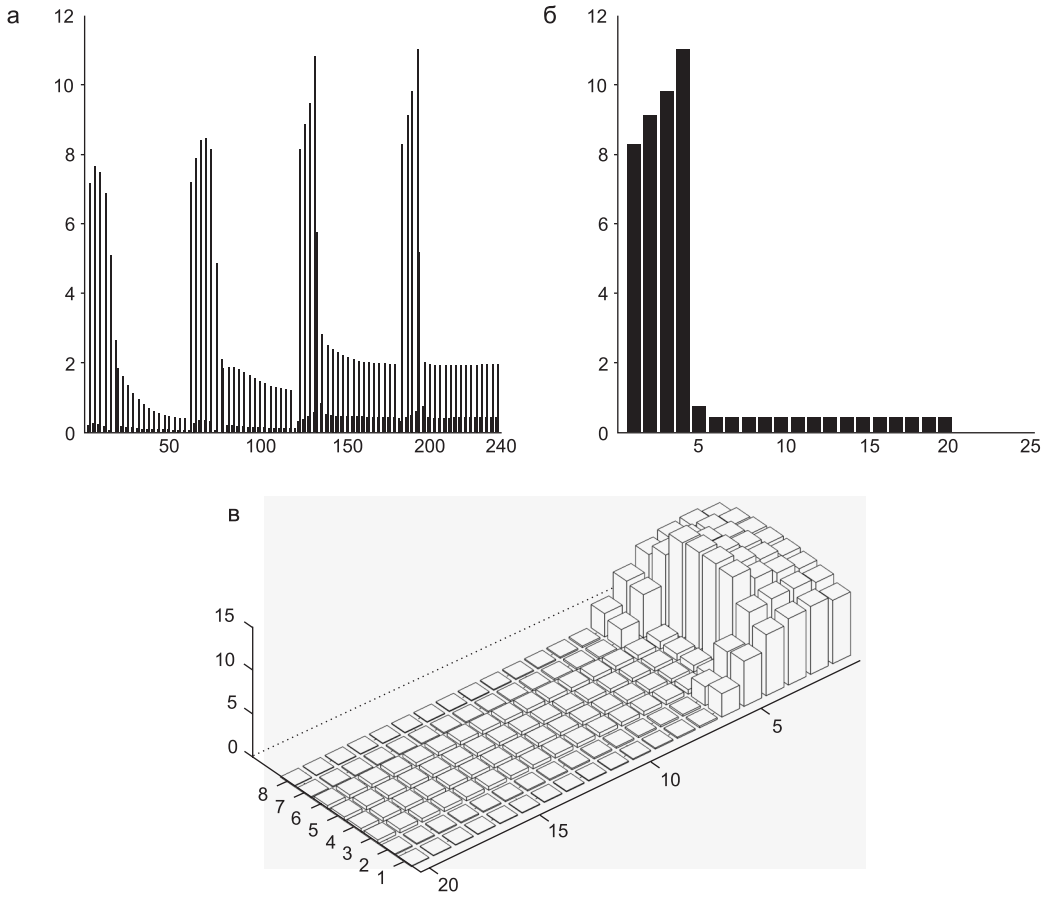

Рис. Примеры сортировки, упорядочения и отображения переменных для ансамбля клеток.

а – вектор всех переменных в конечной точке расчета (по оси X – индекс переменной, по оси Y – ее значение); б – значения переменных, характеризующих концентрацию ауксина в клетках корня вдоль центральной оси (по оси X – индекс клетки, по оси Y – значение переменной в этой клетке); в – интегральная поверхность распределения концентрации фитогормона вдоль осевого среза корня в двухмерном варианте, образованная значениями концентраций ауксина в каждой клетке (по оси X – индекс клетки, по оси Y – индекс клетки, по оси Z – значение переменной в клетке с индексом (X,Y)).

моделирования в форматах, поддерживающих автоматизированный способ их обработки и повторного использования.

Предложен подход к математическому моделированию динамики функционирования биологических систем, который позволяет в процессе реконструкции и исследования моделей сложноорганизованных МГС формировать базы моделей элементарных подсистем для их повторного использования при решении новых задач. Заложенные в программную систему MGSmodeller алгоритмы сокращают время на создание различных вариантов математической модели исследуемого целевого объекта, позволяющих описывать мутации, делеции, вставки и другие молекулярно-генетические модификации, а также проводить проверку гипотез об альтернативных механизмах функционирования подсистем.

Применение высокопроизводительных вычислений позволяет существенно сократить время на проведение анализа комплексной модели и одновременно исследовать воздействие одних и тех же внешних факторов на динамику функционирования различных вариантов структурно-функциональной организации системы.

Анализ сложноорганизованных моделей часто требует применения дополнительных специализированных инструментов. Предложенный подход к организации и описанию структуры атрибутов объектов моделирования и результа-

тов численного исследования облегчает процесс разработки такого инструментария.

Работа выполнена при частичной поддержке РФФИ (гранты № 10-01-00717, № 11-04-01254а, 12-04-31119 мол_а); Минобрнауки (госконтракт П857); Президиума РАН (проекты 6.8 и 30.29); гранта НШ-5278.2012.4; СО РАН (междисциплинарные интеграционные проекты № 80, № 130) и фонда «Династия» (грант для молодых биологов).

ЛИТЕРАТУРА

Акбердин И.Р., Казанцев Ф.В., Омельянчук Н.А., Лихошвай В.А. Математическое моделирование метаболизма ауксина в клетке меристемы побега растения // Информ. вестник ВОГиС. 2009. Т. 13. № 1. С. 170–175.

Казанцев Ф.В., Миронова В.В., Новоселова Е.С. и др. Язык моделирования молекулярно-генетических систем SiBML // Тр. конф. «Параллельные вычислительные технологии (ПаВТ) 2012». Новосибирск, 26–30 марта 2012. Новосибирск: ИВМиМГ СО РАН, 2012. С. 722.

Лихошвай В.А., Матушкин Ю.Г., Ратушный А.В. и др. Обобщенный химико-кинетический метод моделирования генных сетей // Молекуляр. биология. 2001. Т. 3. № 6. С. 1072–1079.

Ратнер В.А. Генетические управляющие системы. Новосибирск: Наука, 1966. 181 с.

Cedilnik A., Geveci B., Moreland K. *et al*. Remote large data visualization in the ParaView framework // Eurographics Parallel Graphics and Visualization. 2006. P. 163–170.

Funahashi A., Morohashi M., Kitano M., Tanimura N. CellDesigner: a process diagram editor for generegulatory and biochemical networks // BIOSILICO. 2003. V. 1. P. 159–162.

Hoops S., Sahle S., Gauges R. *et al*. COPASI – a COmplex PAthway Simulator // Bioinformatics. 2006. V. 22. P. 3067–3074.

Hucka M., Finney A., Sauro H. *et al*. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models // Bioinformatics. 2003. V. 19. P. 524.

Karr J., Sanghvi J., Macklin D. *et al*. A whole-cell computational model predicts phenotype from genotype // Cell. 2012. V. 150. I. 2. P. 248–250.

Kazantsev F.V., Akberdin I.R., Bezmaternykh K.D. *et al*. MGSmodeller – a computer system for reconstruction, calculation and analysis of mathematical models of molecular-genetic system // Proc. of the 6th Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2008). Novosibirsk, 22–28 June 2008. Novosibirsk: Inst. Cytol. Genet., 2008. P. 113.

Mironova V.V., Novoselova E.S., Doroshkov A.V. *et al*. Combined *in silico/in vivo* analysis of mechanisms providing for root apical meristem self-organization and maintenance // Annals Bot. 2012. V. 110. I. 2. P. 349–360. doi:10.1093/aob/mcs069

Olivier B.G., Snoep J.L. Web-based kinetic modelling using JWS Online // Bioinformatics. 2004. V. 20. P. 2143–2144.

Sauro H.M., Hucka M., Finney A. *et al*. Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration // OMICS. 2003. V. 7. Issue. 4. P. 355–372.

Shapiro B.E., Levchenko A., Meyerowitz E.M. *et al*. Cellerator: extending a computer algebra system to include biochemical arrows for signal transduction simulations // Bioinformatics. 2002. V. 19. I. 5. P. 677–678.

NEW FACILITIES OF THE MGSmodeller

F.V. Kazantsev[1], I.R. Akberdin[1], N.L. Podkolodnyy[1], V.A. Likhoshvai[1, 2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia, e-mail: kazfdr@bionet.nsc.ru;
[2] Novosibirsk National Research State University, Novosibirsk, Russia

Summary

Mathematical modeling and analysis of complex molecular-genetic systems (MGS) are the key challenges in the systems biology era. To solve this task the special technologies and programming approaches considering the MGS as an ensemble of dynamic interconnected subsystems with a more simple structure are necessary to be developed. We have presented the approach that is aimed at acceleration of reconstruction of the complex MGS mathematical models and complex analysis using high performance computation techniques.

Key words: molecular-genetic systems, mathematical models, MGSmodeller, high-throughput computation.

УДК 573.2,576.324,004.932.2

МОДЕЛИРОВАНИЕ МОРФОДИНАМИКИ НА РАННИХ СТАДИЯХ ЭМБРИОГЕНЕЗА РАСТЕНИЯ

© 2012 г. **С.В. Николаев[1], Н.А. Колчанов[1, 7, 8], С.К. Голушко[2], Ж.-К. Палаки[3], О. Урбан[4], Е.В. Амелина[2], А.В. Юрченко[5], К.С. Голушко[5], У.С. Зубаирова[1], А.В. Пененко[6], А. Трубюй[4]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: nikolaev@bionet.nsc.ru;
[2] Конструкторско-технологический институт вычислительной техники СО РАН, Новосибирск, Россия;
[3] Национальный институт сельскохозяйственных исследований – Версальский исследовательский центр, Париж, Франция;
[4] Национальный институт сельскохозяйственных исследований – Центр в Жуи-эн-Жозэс, Париж, Франция;
[5] Институт вычислительных технологий СО РАН, Новосибирск, Россия;
[6] Институт вычислительной математики и математической геофизики СО РАН, Новосибирск, Россия;
[7] НИЦ «Курчатовский институт», Москва, Россия;
[8] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

Поступила в редакцию 15 июля 2012 г. Принята к публикации 31 августа 2012 г.

Целью работы было изучение динамики формы (морфодинамики) на ранних стадиях развития зародыша *Arabidopsis thaliana*. В ходе работы была отработана последовательность использования информационных технологий – конвейер процедур – от получения стеков конфокальных снимков и реконструкции по ним трехмерных моделей зародыша до клонального анализа и расчетов механического поведения клеток растущего зародыша. В статье приведены описание конвейера и предварительные результаты.

Ключевые слова: биология развития, эмбриогенез, морфодинамика, морфогенез, зародыш растения, конфокальная микроскопия, анализ изображений, клеточная линия, механика клетки, метод конечных элементов, моделирование.

ВВЕДЕНИЕ

Одной из фундаментальных проблем современной биологии является изучение регуляции формообразования (морфогенеза) в процессе роста организма. Ключевым этапом такого изучения является реконструкция последовательности стадий преобразования формы.

На ранних стадиях эмбриогенеза из одной клетки развивается зародыш, имеющий округлую форму. По мере роста и деления клеток зародыш претерпевает морфологические изменения, приводящие к появлению плоскостей симметрии. В результате к моменту прорастания он приобретает все черты молодого проростка. Клетки объединены в ткани и органы, которые составляют корешок, гипокотиль и семядоли (рис. 1).

Некоторые аспекты морфогенеза зародыша достаточно подробно изучены путем экспертного анализа снимков зародышей на разных стадиях развития. Однако ясно, что для понимания механики процессов развития зародышей растений требуется более детальная картина, которую можно получить с привлечением информационных технологий, таких, как методы автоматизированной обработки

экспериментальных данных, математического и численного моделирования, решения обратных и полуобратных задач.

Постановка конкретных задач при изучении морфодинамики зародыша растения определяет последовательность информационных технологий (конвейер процедур), применяемых на этих этапах. В данной статье описан конвейер процедур для реконструкции клеточных линий и изучения механики деформирования стенок клеток на ранних стадиях морфодинамики зародыша.

Схема конвейера

На рис. 2 приведена схема потока данных между процедурами конвейера. Описание этих процедур составляет основное содержание статьи. Некоторые из них широко известны, описано их применение для решения биологических задач, другие процедуры были либо адаптированы для биологических задач, либо явились оригинальной разработкой. И те и другие приведены в соответствующих частях

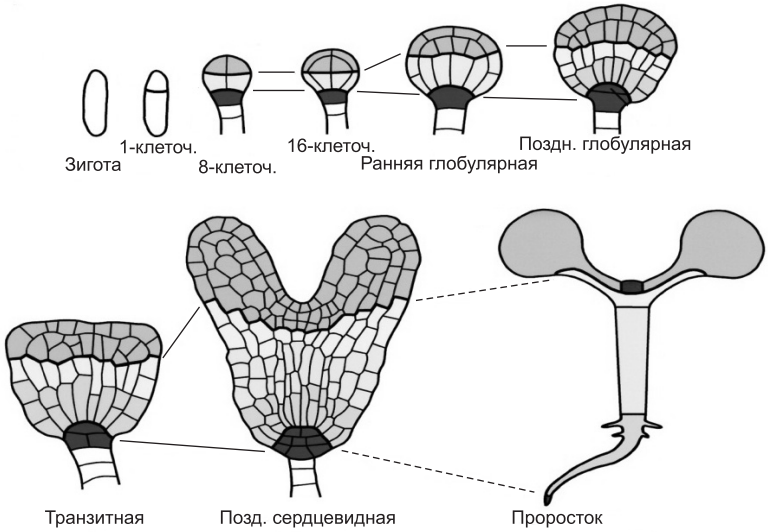

Рис. 1. Схематическое изображение стадий развития зародыша *Arabidopsis thaliana* до проростка (адаптировано из: Laux *et al*., 2004).

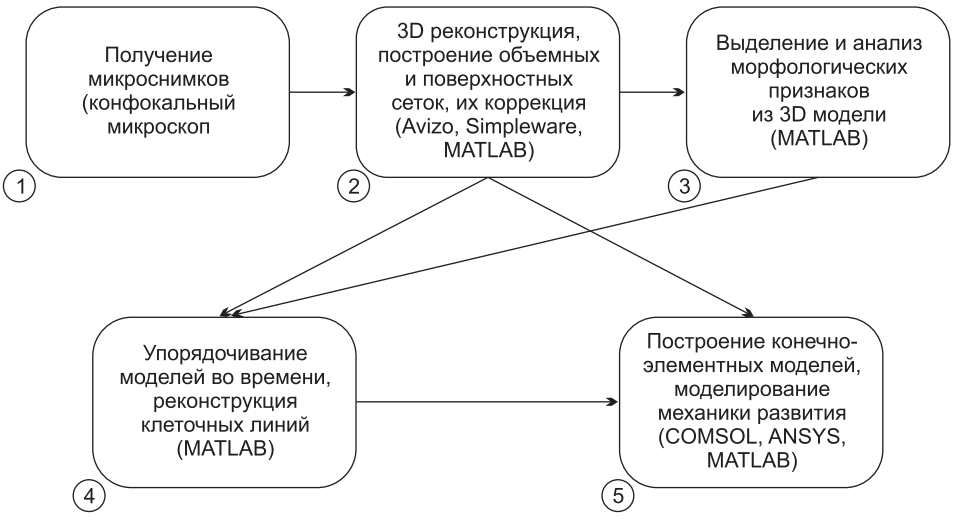

Рис. 2. Схема потока данных между процедурами конвейера.

статьи. Следующий раздел посвящен некоторым деталям формирования изображения в контексте реконструкции клеточного строения ткани и основным шагам пространственной реконструкции, дается представление о клональном дереве и его трассировке. Затем следует раздел, посвященный моделированию механики клеточной стенки, там же в контексте метода конечных элементов рассматривается задача построения сеточного представления геометрии зародыша. Описание экспериментов и методов подготовки образцов и получения снимков не является предметом данной статьи.

Реконструкция клеточного строения растительной ткани

3D реконструкция растительных клеток. Задача 3D реконструкции клеток состоит в построении пространственного образа для каждой клетки в поле изображения. Исходными данными для такого построения является набор изображений последовательных срезов некоторого объема образца. При использовании конфокальной микроскопии эти срезы являются виртуальными – изображение получается сканированием лазерным лучом некоторой фокальной плоскости внутри прозрачного объекта. Сигнал флюоресценции собирается от каждого элементарного объема, вокселя (voxel), и хранится вместе с его координатами. Для наблюдения клетки необходимо, чтобы воксели в пределах одной структуры (например, ядра) были отличимы от вокселей, принадлежащих другой структуре, например, клеточной стенке (Truernit *et al.*, 2008). В зависимости от того, какую субклеточную структуру надо наблюдать, используются различные флюоресцентные метки. Для выделения границ областей, например клеток, на изображении (сегментации изображения) часто используют «алгоритм водораздела» (watershed transformation) как с точками инициации, так и без них (Гонсалес, Вудс, 2006). В качестве точки инициации внутри клетки может быть взят, например, центроид клеточного ядра, если оно детектируется. В настоящее время имеется много программ, реализующих данные алгоритмы сегментации. Стэки сегментированных изображений являются основой для построения пространственной геометрии образца (Pawley, 2006). В данной работе трехмерная реконструкция проводилась с использованием пакета Avizo (http://www.3dvisual.com.au/html/avizo.html).

Набор трехмерных моделей для зародышей на разных стадиях развития дает материал для постановки и решения разнообразных биологических задач, таких, как реконструкция клеточных линий, являющихся основой роста и морфогенеза зародыша, и изучение морфодинамики растущего зародыша.

Клеточные линии и алгоритмы их реконструкции. Под клеточной линией (или клоном) для некоторой клетки будем подразумевать множество потомков данной клетки, появившихся в процессе развития организма. Изучение клеточных линий (клональный анализ) началось с описания Уитманном (Whitmann) паттернов клеточных делений эмбриона пиявки в XIX в. Реконструкция клеточных линий является одной из основных задач при изучении развития организма, в частности морфогенеза, и все чаще применяется при исследовании стволовых клеток и новообразований в организме.

Трассировка (отслеживание) линий дает информацию о числе потомков для данной клетки, их местоположении и статусе дифференцировки. Для такой трассировки в настоящее время имеется несколько экспериментальных подходов, например, в статье K. Kretzschmar с совт. (2012). Так, полное «генеалогическое дерево» для клеток нематоды *Caenorhabditis elegans* было определено с использованием микроскопии Номарского (Sulston *et al*., 1983). Современная цейтраферная микросъемка в нескольких фокальных плоскостях позволила записывать клеточные линии в оцифрованном виде, что стимулировало разработку автоматизированных методик клонального анализа. Автоматическое построение клеточной линии для *C. elegans* было описано в работе Z. Bao с соавт. (2006), а для раннего развития рыбки zebrafish (до стадии 1000 клеток) была реконструирована полная клеточная линия (Olivier *et al*., 2010).

Следует отметить, что для растений реконструкцию клеточных линий облегчает отсутствие миграции клеток и апоптоза (Fernandez *et al*., 2010).

Алгоритмы реконструкции клеточных линий. В соответствии с определением Buckingham с соавт. (2011) целью клонального анализа

является построение полного «4D изображения» клеток *in vivo*. Разработка алгоритмов для реконструкции клеточных линий на основе имеющихся 4D данных является нетривиальной задачей (Olivier *et al*., 2010, Fernandez *et al*., 2010). В общем случае алгоритм включает 3D реконструкцию клеток ткани в разные моменты времени с трассировкой индивидуальных клеток и визуальной верификацией, требующей больших затрат времени. Часто трассировка клеток затруднена из-за роста и деления клеток, их деформирования и движения, что стимулирует совершенствование алгоритмической базы. Например, MARS-ALT-конвейер (Fernandez *et al*., 2010) представляет собой итеративный процесс построения надежной реконструкции клеточной линии. Каждый шаг этого процесса включает: 1) идентификацию «вручную» небольшого набора клеточных линий для каждой пары сегментированных изображений для последовательных моментов времени; 2) оценку деформирующего преобразования для этой пары изображений; 3) оценку нелинейного преобразования вокселей на векторном поле оптических плотностей с учетом погрешностей позиционирования; 4) уточнение поля деформации на основе интенсивности вокселей; 5) построение гипотетической клеточной линии на паре изображений; 6) дополнительное тестирование линий с использованием геометрических и топологических правил. MARS-ALT алгоритм использовался при построении клеточных линий для развития меристемы цветка на протяжении 70 часов.

Особенности реконструкции клеточных линий для зародыша растения. Для ранних стадий эмбрионального развития растения непрерывные наблюдения зародыша внутри семени на клеточном разрешении остаются в настоящее время нерешенной экспериментальной задачей. Для клонального анализа доступны данные о клеточном строении зародышей, фиксированных в некоторые моменты времени в процессе развития. В этой ситуации сложность задачи клонального анализа зависит от вариабельности клеточных линий между зародышами. В случае низкой вариабельности можно попытаться классифицировать коллекцию зародышей по стадиям развития и рассматривать упорядоченную коллекцию в качестве возрастной серии для «обобщенного» зародыша.

Отсутствие миграции клеток позволяет извлечь полезную для клонального анализа информацию из реконструированной сегментированной модели отдельного зародыша. Например, для зародыша арабидопсиса дикого типа паттерн клеточных делений известен и стабилен вплоть до стадии нескольких десятков клеток, и биологи растений могут реконструировать клональное дерево из 3D модели зародыша. Эта априорная информация полезна для формулировки решающих правил для распознавания клональных отношений, которыми пользуются эксперты. Для зародыша, состоящего из сотен клеток, клональный анализ «вручную» является чрезмерно трудоемким и может сопровождаться множеством ошибок. Поэтому автоматизация такой работы на основе решающих правил является актуальной задачей, так как ее решение позволило бы массово реконструировать клеточные линии и проводить сравнения между организмами дикого типа и организмами с разнообразными мутациями.

На рис. 3 приведен пример реконструкции клонального дерева, построенного с использованием алгоритма, разработанного при выполнении данного проекта. Решение о том, появились ли две рядом расположенные клетки в результате деления некоторой материнской клетки, или они принадлежат разным парам сестринских клеток, принимается на основе автоматического анализа деталей строения клеточных стенок. Описание алгоритма является предметом другой статьи и здесь не приводится.

Компьютерное моделирование механического поведения растительных клеток и тканей

Клетка как осмотическая ячейка. Несмотря на сложное строение клетки и наличие механически разнородных структур на мезоскопических масштабах (масштабы субклеточных структур – органелл), механическое поведение клетки в первом приближении определяется клеточной стенкой. В таком приближении клетка представляет собой упругую оболочку, заполненную жидкостью. Считается, что оболочка обладает свойством избирательной проницаемости: одни вещества, в том числе вода, могут свободно проходить через эту

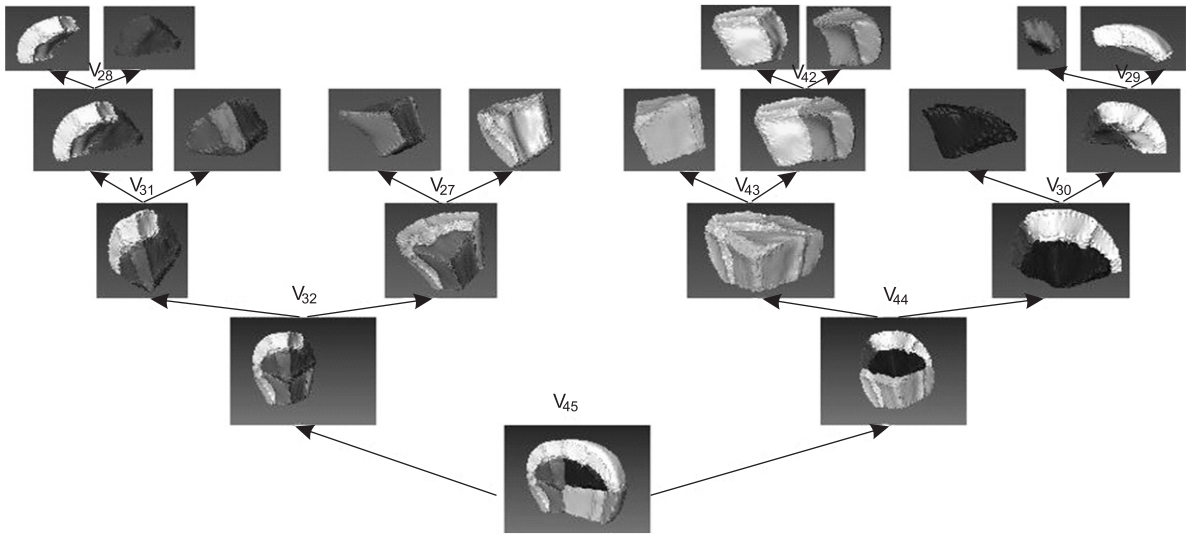

Рис. 3. Пример реконструкции части клонального дерева для зародыша арабидопсиса с использованием алгоритма, основанного на автоматической детекции морфологических особенностей клеточного строения.

оболочку, а для других оболочка является непроницаемой. Таким образом, клетка является осмотической ячейкой: в гипотонической среде она стремится разбухнуть, а в гипертонической среде – сжаться. В каждый момент времени тургорное давление уравновешивается силой упругого натяжения клеточной оболочки. Далее будем считать, что механические свойства оболочки определяются клеточной стенкой, и что в условиях равновесия сила натяжения везде одинакова. Если в клетку поступает вода (при отличной от нуля разности водных потенциалов), то увеличиваются ее объем, площадь стенки и ее упругое натяжение.

Моделирование механического поведения стенки растительной клетки. Материал клеточной стенки состоит как минимум из двух фаз, чем схож с конструкционными композитами. Одну из фаз – волокна целлюлозы, образующие «скелет» стенки, по аналогии с композиционными материалами будем называть волокнами или наполнителем. Другая фаза, представленная в клеточной стенке сильно гидратированной матрицей из молекул гемицеллюлозы и/или пектина, заполняет пространство между волокнами и в терминологии композиционных материалов называется связующим материалом.

Целлюлозные волокна в клеточной стенке зрелых клеток формируют слои с различающейся от слоя к слою преимущественной ориентацией волокон – так называемую вторичную клеточную стенку. В то же время в активно растущей ткани клетки разделены первичной клеточной стенкой, и волокна в матрице такой стенки не образуют какой-либо строго упорядоченной структуры. Таким образом, материал клеточной стенки является неоднородным и анизотропным. Для описания механических свойств первичной клеточной стенки можно использовать осредненные механические характеристики целлюлозного скелета в матрице. Для прогнозирования осредненных механических характеристик и построения физических соотношений, связывающих напряжения и деформации, можно применять различные структурные модели композиционного материала, например, описанные С.К. Голушко и Ю.В. Немировским (2008).

Предполагается, что в результате упругого деформирования под действием тургорного давления происходит увеличение размеров растительной клетки. Затем происходит релаксация механических напряжений клеточной стенки за счет разрыва части связей между молекулами-волокнами целлюлозы, что схоже с пластической деформацией при ослаблении материала. После этого материал клеточной стенки достраивается и упрочняется при дальнейшем увеличении объема клетки, что в целом схоже с поведением гиперупругих материалов. Возникающие циклы «упругое растяжение–ре-

лаксация» можно заменить непрерывным упругим деформированием при увеличивающейся нагрузке, что существенно упрощает математическую и численную модель процесса.

Даже при описанном упрощении модели аналитическое решение задачи об упругом деформировании клеток зародыша возможно только в случаях простой геометрии, т. е. на самых первых стадиях. Для более поздних стадий и, соответственно, сложных форм зародыша необходимо применять численные методы решения. Одним из универсальных методов дискретного представления геометрического объекта для последующего численного анализа является использование неструктурированных сеток, а наиболее развитым методом решения задач механики на неструктурированных сетках – метод конечных элементов.

О построении конечно-элементных моделей многоклеточных структур растительных тканей. Конечно-элементное моделирование включает несколько этапов. Первые два из них – создание геометрической модели и генерация сетки элементов.

В случаях моделирования объектов с относительно простой геометрией (например, зародышей растения на стадии до 4–8 клеток) можно эффективно использовать прямое геометрическое моделирование, например, с помощью CAD-редакторов геометрии. В такой ситуации можно оперировать простыми геометрическими формами, комбинируя их для создания приближенной геометрической модели (рис. 4, а). Другой подход заключается в использовании геометрии, построенной на основе экспериментальных данных по технологии, описанной в предыдущих разделах. В этом случае автоматически реконструируется геометрия и генерируется сетка элементов, что позволяет импортировать соответствующую модель непосредственно в систему компьютерного моделирования и анализа (CAE систему) и использовать ее для расчета (рис. 4, б).

Алгоритмы построения сеток составляют раздел вычислительной геометрии и представлены в многочисленных руководствах, например у Ф. Препарата и М. Шеймос (1989). Качественно построенная сетка может с достаточно высокой точностью представлять геометрию природного объекта, в частности зародыша растения. Эта же сетка может быть использована как основа для построения конечно-элементной модели в системе компьютерного моделирования при решении краевых задач для систем дифференциальных уравнений в частных производных. В настоящее время существует множество пакетов, как коммерческих, так и для свободного использования, предназначенных для построения таких сеток. В настоящей работе использован пакет Simpleware (http://www.simpleware.com). Для обмена данными между пакетами Simpleware и CAE-системами COMSOL и ANSYS использован формат STL (Standard Tessellation Language).

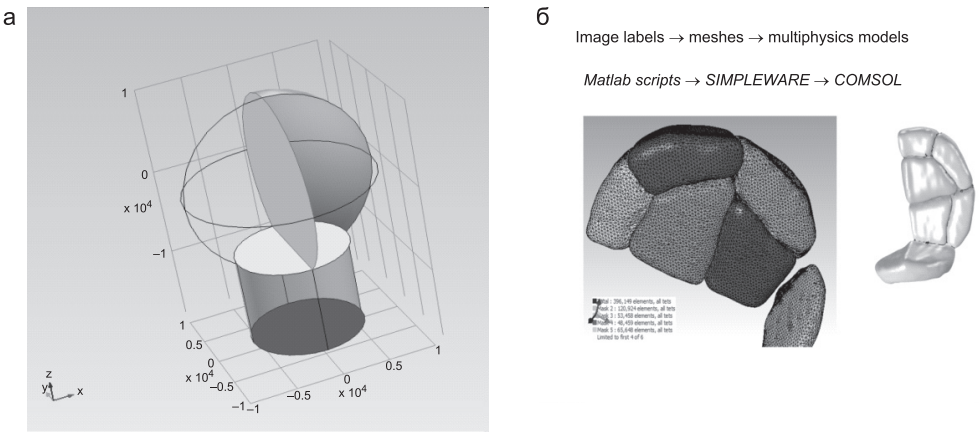

Рис. 4. Примеры построения геометрической модели зародыша с использованием редактора в пакете COMSOL 4.2 (а) на основе 3D реконструкции (б).

Применение компьютерного моделирования для интерпретации особенностей деформирования клеток. Анализ экспериментальных изображений различных стадий развития зародыша растения выявил ряд особенностей деформирования стенок его клеток, нуждающихся в объяснении.

Для изучения возможных причин возникновения изгиба стенки между клетками зародыша и суспензором рассмотрена двухклеточная фаза развития зародыша и исследовано влияние на характер его деформирования различных механических параметров. Для этого в пакете COMSOL построена оболочечная конечно-элементная модель такой системы, и в качестве недеформированного состояния рассматривалась исходная геометрия зародыша. Параметры модели оценены на основании данных из известных источников (Chanliaud *et al*., 2002). В качестве нагрузки рассмотрено избыточное давление в клетках зародыша и суспензоре с разницей всего 2,5 %.

На рис. 5 представлены результаты компьютерного моделирования деформации клеточных стенок зародыша (слева) в сравнении с изображением среза зародыша *A. thaliana* на стадии двух клеток (справа), полученным на конфокальном микроскопе. Проведенные вычислительные эксперименты на оболочечной модели показали, что плоские межклеточные стенки оказываются очень чувствительными к разнице давлений в соседних клетках. Сравнение результатов вычислений с наблюдаемой формой клеточных стенок между клетками зародыша и клеткой суспензора позволяет предположить, что скорости роста клеток зародыша в данном случае превышают скорость роста клетки суспензора, что, в свою очередь, приводит к развитию в этих клетках большего тургорного давления.

Другой характерной деталью, обнаруженной на ряде снимков зародышей, является наличие своеобразных «гребешков» (рис. 6). Возникновение подобного «гребешка» на клеточной стенке наблюдается при численном моделировании деформирования оболочечной модели зародыша при избыточном давлении снаружи клеток. Это позволяет выдвинуть гипотезу о том, что зародыш к моменту съемки подвергался действию гиперосмотической среды (возможно, в процессе подготовки образца).

Выявление особенностей на изображениях, подобных тем, что описаны выше, и их интерпретация актуальны для корректного

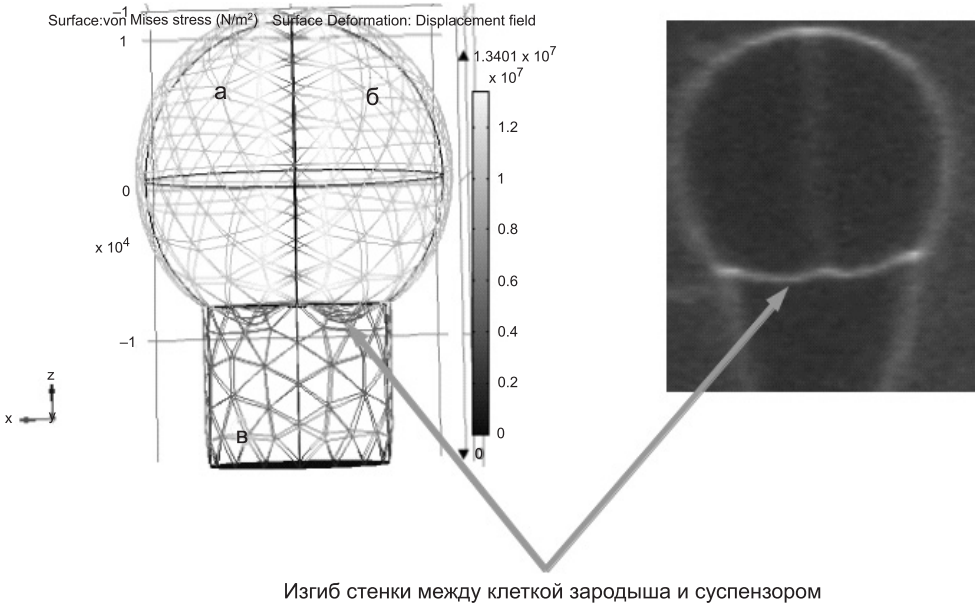

Рис. 5. Сравнение результатов моделирования равновесного состояния стенки в модели зародыша на стадии двух клеток с конфокальным микроизображением.

а, б – клетки зародыша, в – клетка суспензора.

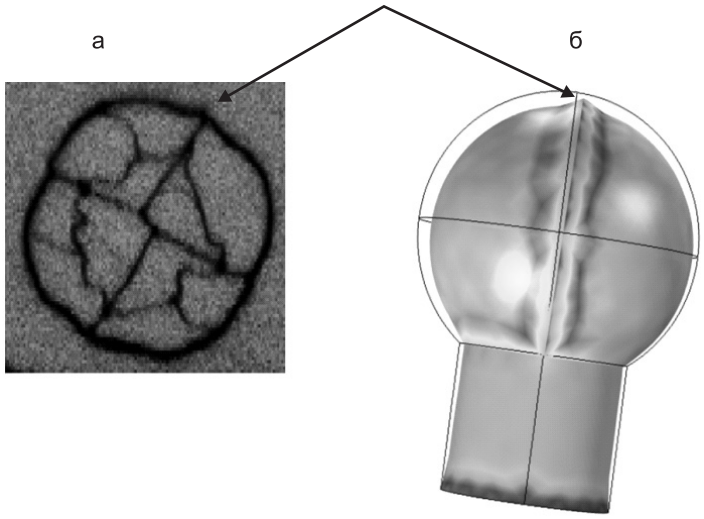

Рис. 6. Наблюдаемые (а) и расчетные (б) «гребешки» на микроизображении и математической модели зародыша.

использования геометрических параметров зародыша при формулировках биологически содержательных гипотез и выводов.

Механика морфодинамики одноклеточного зародыша арабидопсиса. Изучение различных стадий развития зародыша позволяет получить представление о деформации его формы и формы клеток, из которых он состоит, а также их характерных размеров. Так, из экспериментальных снимков видно, что на двух-, четырех- и восьмиклеточной стадиях форма зародыша остается близкой к сферической. При этом форма одноклеточного зародыша сразу после деления зиготы близка к полуэллипсоидальной, практически совпадающей по радиусу с клеткой-суспензором. Предварительные модельные расчеты показали, что деформирование полуэллипсоидальной клетки в сферическую со стенками из однородного изотропного материала под действием избыточного давления может происходить только с существенным увеличением объема исходной клетки. В то же время наблюдения показывают, что округление одноклеточного зародыша может происходить даже при незначительном увеличении исходного объема (примерно на 1/3 от начального).

Как отмечалось, у клетки существуют возможности по упрочнению и ослаблению своих стенок, в том числе локальному. Кроме того, клетка может строить целлюлозный скелет стенки с преобладанием волокон, ориентированных в одном из направлений, в результате чего возникает анизотропия механических свойств стенки на макроуровне, схожая с анизотропией волокнистых композиционных материалов. Было предположено, что изменение механических свойств отдельных участков клеточной стенки зародыша обусловливает наблюдаемые особенности ростовой деформации клетки. С учетом этого поставлена задача определения механических характеристик материала стенки, позволяющих обеспечить переход формы клетки от исходного к деформированному состоянию под действием тургорного давления. Подобная задача ставилась, например, авторами работы (Fayant *et al*., 2010) при моделировании апикального роста пыльцевой трубки. По аналогии в настоящей работе для поиска законов распределения механических характеристик материала оболочка, представляющая зародыш, разбита на слои с помощью плоскостей, перпендикулярных оси вращения, после чего осуществлен поиск требуемых параметров для каждого из слоев. При этом для эффективных механических характеристик материала использованы соотношения из работы В.В. Болотина и Ю.Н. Новичкова (1980), а для решения задачи линейного упругого деформирования применен комплекс программ ANSYS Mechanical™ с трехмерными (объемными) конечными элементами.

На основании данных, полученных с помощью конфокального изображения (рис. 7, а),

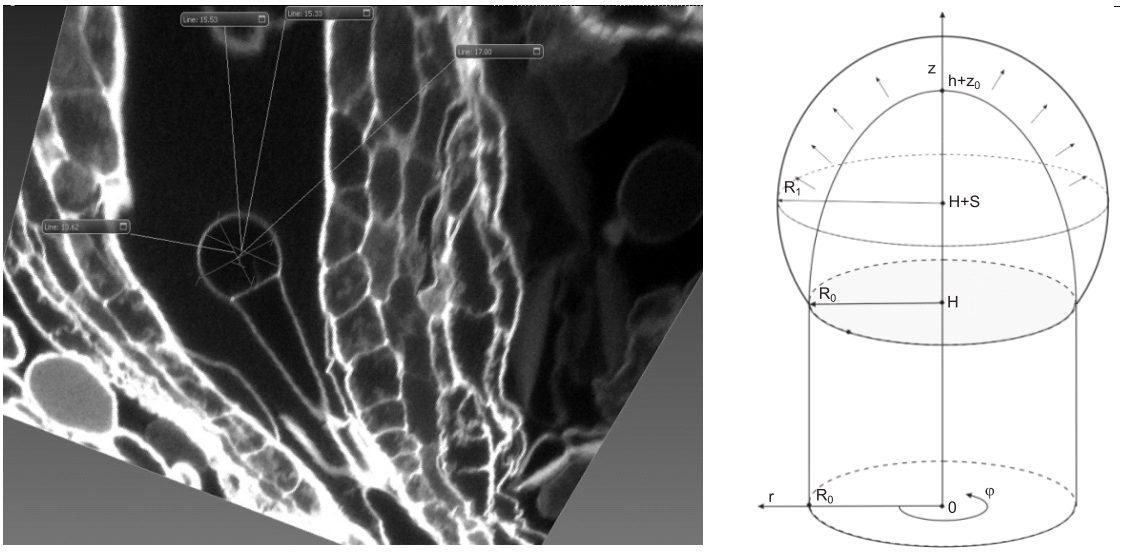

Рис. 7. Данные конфокального изображения зародыша и геометрическая модель.

оценены размеры конечного состояния клетки и построена геометрическая модель одноклеточного зародыша с исходной и конечной формами (рис. 7, б). Модель построена в виде оболочки вращения в цилиндрической системе координат, суспензор задается в виде цилиндрической оболочки, мембрана (стенка между суспензором и зародышем) – в виде круга, а исходная форма клетки зародыша – в виде полуэллипсоида вращения. Форма клетки зародыша после деформирования – часть сфероида. Предполагается, что толщина стенки до деформирования постоянна и на нее действует равномерно распределенное внутреннее давление.

Рассмотрены три варианта обеспечения требуемой формы:

1) оболочка с изотропными слоями, управляемые параметры (параметры проектирования) – модули упругости материалов слоев $E^{[k]}$;

2) анизотропная оболочка с различной степенью упрочнения слоев в направлении касательной к окружности, параметры проектирования – интенсивности армирования $\psi_a^{[k]}$ и модули упругости компонент-фаз материалов слоев;

3) комбинированный вариант оболочки, содержащей как упрочненные в окружном направлении, так и изотропные слои с различными механическими характеристиками.

В таблице приведены примеры законов распределений управляемых параметров и толщин слоев для перечисленных вариантов. В варианте (в) в первом слое – анизотропный материал, аналогичный материалу оболочки (б), а слои 2–7 изотропные, с различными механическими характеристиками. На рис. 8 представлены состояния оболочки до и после деформирования с обозначением расчетных слоев. Величины на графиках приведены в микрометрах. Как следует из расчетов, требуемая конечная форма может

Таблица

Результаты решения задачи формообразования

Номер слоя k	Вариант оболочки					
	а		б		в	
	Высота слоя по z, мкм	$E^{[k]}$	Высота слоя по z, мкм	$\psi_a^{[k]}$	Высота слоя по z, мкм	$E^{[k]} \mid \psi_a^{[k]}$
1	5	10	4	0,20	4	0,2 ($\psi_a^{[k]}$) ψ_a
2	3	6,7	1	0,06	1	8,3
3	2	5,0	2	0,05	1	7,3
4	1	4,0	2	0,03	2	6,7
5	1	3,4	4	0,00	2	5,7
6	1	2,0	–	–	2	5,0
7	–	–	–	–	1	3,2

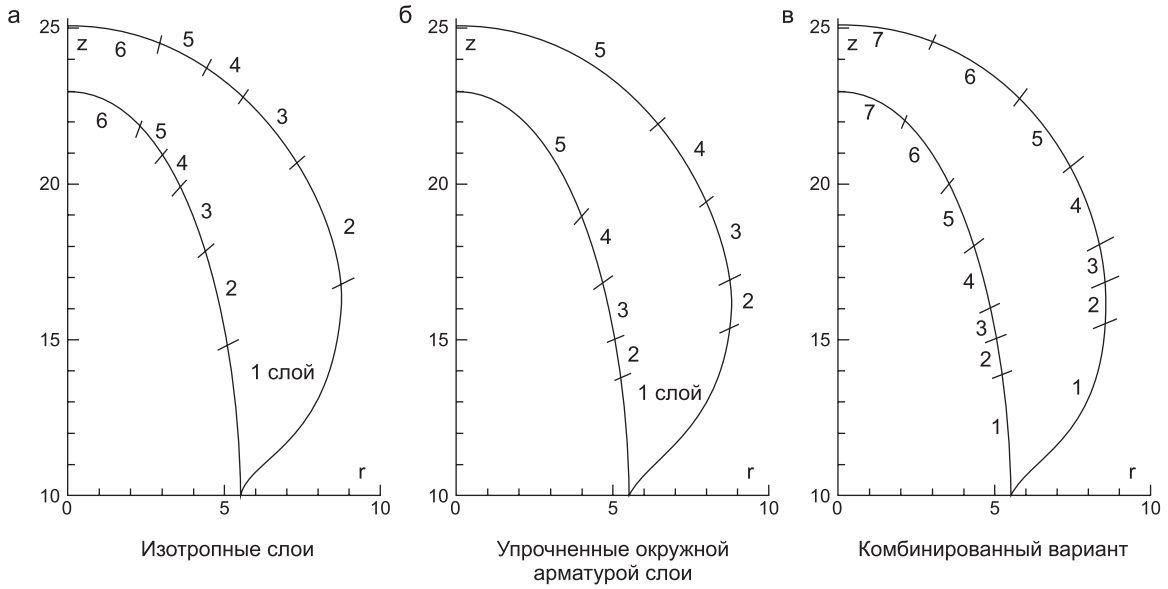

Рис. 8. Различные варианты перехода от начальной к требуемой форме оболочки.

быть достигнута в результате деформирования исходной оболочки с различными начальными свойствами материала стенок клетки зародыша – в общем случае решение поставленной обратной задачи неединственно. Для более точного определения механизмов формообразования необходимы дополнительные данные как о свойствах материала стенок клеток, так и о процессе изменения их формы. Тем не менее во всех рассмотренных вариантах можно найти общие тенденции и закономерности. В частности, отметим, что характеристики жесткости материала по мере приближения к вершине зародыша уменьшаются. Это выражается как в уменьшении значений модулей упругости слоев для изотропного материала стенки клетки, так и в снижении содержания упрочняющих волокон в случае анизотропии.

ЗАКЛЮЧЕНИЕ И ПЕРСПЕКТИВЫ

К настоящему времени накоплен значительный методический опыт в применении различных подходов к изучению живых объектов: от экспериментально-наблюдательных до автоматизированной обработки информации и математического моделирования.

В данной работе представлен вариант конвейера процедур от конфокальной микроскопии до моделирования механического поведения клеток, примененный при изучении развития зародыша *A. thaliana*. Использование неструктурированных сеток для представления геометрии зародыша позволило организовать поток данных из программ построения геометрии в программы для численного моделирования методом конечных элементов. Ключевым моментом такого конвейера является реконструкция клеточного строения зародыша и последовательности клеточных делений при его росте. Результаты построения клеточных линий в перспективе позволят отслеживать траектории некоторых реперных точек объекта при его морфодинамике для выявления последовательности происходящих деформаций, что, в свою очередь, позволит более корректно определять алгоритмы, которые использует клетка для управления процессом морфодинамики.

Работа выполнена при частичной финансовой поддержке гранта РФФИ-НИСИ № 11-04-91397-а.

ЛИТЕРАТУРА

Болотин В.В., Новичков Ю.Н. Механика многослойных конструкций. М.: Машиностроение, 1980.

Голушко С.К., Немировский Ю.В. Прямые и обратные задачи механики упругих композитных пластин и оболочек вращения. М.: Физматлит, 2008.

Гонсалес Р., Вудс Р. Цифровая обработка изображений. М.: Техносфера, 2006.

Препарата Ф., Шеймос М. Вычислительная геометрия: Введение. М.: Мир, 1989.

Bao Z., Murray J.I., Boyle T. *et al.* Automated cell lineage tracing in *Caenorhabditis elegans* // Proc. Natl Acad. Sci. USA. 2006. V. 103. P. 2707–2712.

Buckingham M.E., Meilhac S.M. Tracing cells for tracking cell lineage and clonal behavior // Developm. Cell. 2011. V. 21. P. 394–409.

Chanliaud E., Burrows K.M., Jeronimidis G., Gidley M.J. Mechanical properties of primary plant cell wall analogues // Planta. 2002 V. 215. P. 989–996.

Fayant P., Girlanda O., Chebli Y. *et al*. Finite element model of polar growth in pollen tubes // Plant Cell. 2010. V. 22. P. 2579–2593.

Fernandez R., Das P., Mirabet V. *et al*. Imaging plant growth in 4D: robust tissue reconstruction and lineaging at cell resolution // Nat. Meth. 2010. 7. P. 547–553.

Handbook of Biological Confocal Microscopy / Ed. J.B. Pawley. Springer, N.Y. 2nd ed. 2006. 988 p.

Kretzschmar K., Watt F.M. Lineage Tracing // Cell. 2012. 148. P. 33–45.

Laux T., Würschum T., Breuninger H. Genetic regulation of embryonic pattern formation // Plant Cell. 2004. V. 16. Suppl. 1. P. S190–S202.

Olivier N., Luengo-Oroz M.A., Duloquin L. *et al*. Cell lineage reconstruction of early zebrafish embryos using label-free nonlinear microscopy // Science. 2010. V. 329. P. 967–971.

Truernit E., Bauby H., Dubreucq B. *et al*. High-resolution whole-mount imaging of three-dimensional tissue organization and gene expression enables the study of phloem development and structure in *Arabidopsis* // Plant Cell. 2008. V. 20. P. 1494–1503.

MODELING OF PLANT EMBRYO MORPHODYNAMICS AT EARLY DEVELOPMENTAL STAGES

S.V. Nikolaev[1], N.A. Kolchanov[1, 7, 8], S.K. Golushko[2], J.-C. Palauqui[3], A. Urban[4], E.V. Amelina[2], A.V. Yurchenko[5], K.S. Golushko[5], U.S. Zubairova[1], A.V. Penenko[6], A. Trubuil[4]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia, e-mail: nikolaev@bionet.nsc.ru;
[2] Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia;
[3] Institut National de la Recherche Agronomique – Centre de Versailles-Grignon, Institut Jean-Pierre Bourgin, UMR1318 INRA-AgroParisTech, France;
[4] Institut National de la Recherche Agronomique – Centre de Jouy-en-Josas Unite de Mathematiques et Informatique Appliquees, France;
[5] Institute of Computational Technologies SB RAS, Novosibirsk, Russia;
[6] Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia;
[7] National Research Centre «Kurchatov Institute», Moscow, Russia;
[8] Novosibirsk National Research State University, Novosibirsk, Russia

Summary

Embryo morphodynamics at early developmental stages of *Arabidopsis thaliana* was studied. First, a pipeline was elaborated from confocal microscopy and tissue 3D reconstruction to cell lineage tree reconstruction and numerical simulation of growing embryo mechanics. Tentative results of its use are presented.

Key words: developmental biology, embryogenesis, morphodynamics, morphogenesis, plant embryo, confocal microscopy, image analysis, cell lineage, cell mechanics, finite element method, modeling.

УДК 573.2:57.011

МОДЕЛИРОВАНИЕ РОСТА И РАЗВИТИЯ РАСТИТЕЛЬНЫХ ТКАНЕЙ В ФОРМАЛИЗМЕ L-СИСТЕМ

© 2012 г. **У.С. Зубаирова[1], А.В. Пененко[2], С.В. Николаев[1]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия,
e-mail: ulyanochka@bionet.nsc.ru;
[2] Институт вычислительной математики и математической геофизики СО РАН, Новосибирск, Россия

Поступила в редакцию 15 июля 2012 г. Принята к публикации 31 августа 2012 г.

Даны краткое описание динамических систем с динамической структурой и формализм L-систем для их представления. На примерах продемонстрировано применение L-систем для моделирования роста растительных тканей и регуляции паттернов распределения параметров, характеризующих состояние системы.

Ключевые слова: биология развития, динамические системы с динамической структурой, математическая модель, L-системы.

ВВЕДЕНИЕ

Развивающиеся организмы – динамические системы с динамической структурой

Выяснение механизмов, которые управляют ростом и развитием организмов, является одной из интереснейших задач биологии. Одним из методов изучения этих явлений является построение формализованных моделей явления, которые должны помочь выяснить взаимную согласованность представлений об отдельных процессах и их соответствие экспериментальным данным. Отличительной чертой изучаемых систем в биологии является изменчивость структуры при условии сохранения определенных обобщенных структурных особенностей – структурных инвариантов. В результате наряду с динамикой переменных состояния системы требуется моделировать динамику ее структуры, что ставит задачу математического моделирования *динамических систем с динамической структурой*.

Например, в некотором аспекте состояние каждой клетки в ткани можно охарактеризовать уровнем экспрессии определенного набора генов – эти показатели будут являться переменными состояния каждой клетки и в целом – ткани. Описание в этих переменных функционирующей во времени ткани, изменяющей свое состояние под воздействием внешних и/или внутренних причин, дает нам пример динамической системы. Естественно представить ткань как систему, структура которой определяется набором клеток-подсистем, между которыми имеются определенные связи (например, потоки сигналов между соседними клетками). В результате роста и деления клеток клеточное строение ткани меняется: меняется соседство клеток и, следовательно, изменяются потоки сигналов между ними. С абстрактной точки зрения, это пример динамической структуры системы. Динамическая структура системы вносит дополнительную сложность при моделировании, так как мы вынуждены перестраивать топологию системы в процессе моделирования.

Для моделирования таких систем известно несколько формализмов и их реализаций в виде компьютерных программ, например, различные варианты L-систем (Lindenmayer, 1968; Prusinkiewicz, Lindenmayer, 1990; Prusinkiewicz *et al*., 1993), стохастическая параметризованная грамматика, реализованная в программе Plenum

(http://computableplant.ics.uci.edu/~guy/Plenum.html), MGS – язык для моделирования биологических систем (http://www.ibisc.univ-evry.fr/~mgs/), CompuCell (http://www.compucell3d.org/).

Представление дискретно-непрерывных моделей организмов в формализме L-систем

В данной работе мы приводим примеры моделей с использованием формализма L-систем (сокращение от Lindenmayer systems). L-системы были введены А. Линденмайером для моделирования развития многоклеточных организмов, образующих линейные и ветвящиеся нитевидные структуры – филаменты (Lindenmayer, 1968). В формализме L-систем организм представляется как упорядоченная структура из дискретных единиц, называемых модулями, при этом формализм не накладывает никаких ограничений на природу модулей. Так, в модели низших организмов модули могут представлять клетки, а в модели высших растений – морфофункциональные единицы, такие, как апикальная меристема, междоузлие, листья, цветы. Каждый модуль представляется символом (буква алфавита L-системы), который характеризует его тип. Эволюция системы заключается в дискретном изменении ее структуры: либо изменяется тип подсистемы, либо вместо одной подсистемы возникает несколько. В формализме L-систем такая эволюция системы моделируется переписыванием строки символов по правилам, определенным в L-системе.

Формально L-система – это упорядоченная тройка $G = < V, w, P >$, где V – алфавит системы, w – непустое слово в алфавите V, называемое аксиомой, P – конечное множество правил вида $a \rightarrow u$, сопоставляющих каждому символу a из алфавита V слово u (возможно, пустое) в алфавите V.

Следующий пример (рис. 1) наряду с иллюстрацией работы L-системы демонстрирует одну из биологических интерпретаций данного формализма.

Anabaena catenula. Символы *a* и *b* представляют состояния клеток (их размеры и готовность к делению). Индексы *left* и *right* отражают полярность клетки, определяя позиции, в которых появятся новые дочерние клетки типа *a* и *b*.

Для описания ветвящихся филаментов в алфавит введены новые служебные символы: «[» и «]» для обозначения начала и конца ветви соответственно, «+» и «–» для обозначения стороны, с которой новая ветвь крепится к материнской. Ветвь крепится к символу, за которым непосредственно следует «[». Ветви могут быть многократно вложены друг в друга: на каждой из ветвей может быть несколько подветвей, которые также имеют свои ответвления, и т. д. Сразу заметим, что введенные новые символы не определяют, под каким углом к исходной оси будет располагаться новая ветвь. И конечный результат во многом зависит от того, как мы будем интерпретировать полученную итоговую строку. Можно, например, использовать для интерпретации «черепашью» графику – в этом случае мы будем получать двумерные изображения (рис. 2). Можно ввести другую интерпретацию и сразу после символа начала ветви добавлять значение угла, под которым эта ветвь будет располагаться по отношению к материнской ветви.

Ввиду дискретной природы L-систем, непрерывный рост модуля в промежутках между применениями правил не описывается моделью, т. е. мы получаем картину через определенные временные интервалы. Однако в динамической системе с динамической структурой нас интересуют два процесса, разные по своей природе: изменение признаков модуля (концентрация веществ, размер и т. п.) и изменение структу-

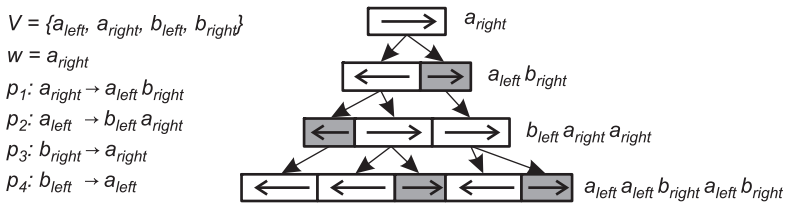

Рис. 1. L-система, моделирующая развитие многоклеточной нитчатой водоросли.

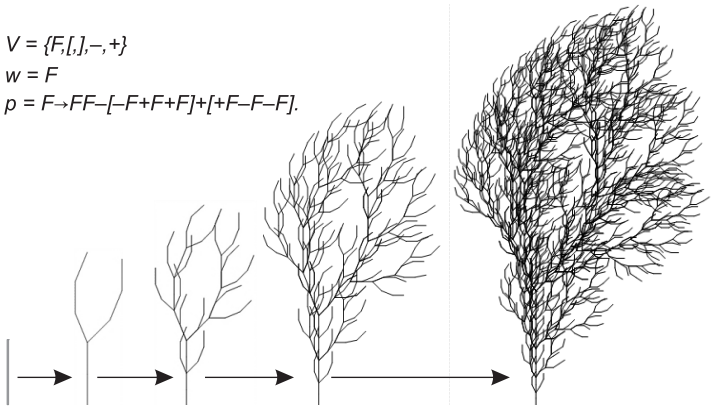

Рис. 2. Пример L-системы, генерирующей ветвящуюся структуру.

На рисунке приведены структуры, полученные на разных шагах работы L-системы.

ры (число и тип модулей). Для моделирования таких систем была разработана модификация, называемая параметризованной L-системой: дополнительно с модулем связан вектор параметров, которые вместе с символом характеризуют состояние модуля. Существенной особенностью параметризованной L-системы является то, что правила переписывания цепочки символов, которые моделируют изменение структуры системы (*динамическая структура*), срабатывают только при выполнении определенных условий на параметры символов (переменные состояния). Параметризованная L-система «работает» на параметризованных словах. Параметризованное слово – это строка, составленная из *модулей* вида $A(a_1, a_2, \ldots, a_n)$, каждый из которых содержит букву и связанные с ней параметры. Буква принадлежит алфавиту V, а значения параметров – некоторому произвольному множеству, в том числе множеству вещественных чисел.

Дифференциальные L-системы (dL-системы) являются сужением параметризованных L-систем, когда параметры являются непрерывными функциями времени. Пока параметр w модуля $A(w)$ остается в области определения допустимых значений D_A, система развивается как непрерывная. Как только значения параметров достигают границы области D_A, правило заменяет модуль $A(w)$ его потомком. Это событие происходит дискретно. Выбор правила может зависеть от того, какой сегмент границы области определения был пересечен. Дифференциальные уравнения описывают в модели непрерывные процессы (например, рост клетки, диффузию веществ в клеточном ансамбле или постепенное удлинение междоузлий). Изменение структуры (например, деление клетки или появление новой ветви) описывается в терминах правил переписывания, которые либо меняют символ (тип модуля), либо стирают символ (например, клетка умирает), либо заменяют один символ на несколько (деление клеток, появление новых фитомеров и т. д.). Правила применяются параллельно, отражая одновременность процессов во всех частях организма. Пример dL-системы будет приведен в следующей главе.

Для спецификации приведенных ниже моделей в формализме L-систем мы использовали пакет *Mathematica*. По своей сути *Mathematica* представляет собой язык программирования высокого уровня, позволяющий реализовывать традиционный процедурный и функциональный стили. Поскольку *Mathematica* имеет встроенную систему переписывания, ее просто использовать для реализации работы L-системы. Правила задаются в виде $a \to b$, здесь символ a заменяется строкой b. Одновременность применения набора правил организуется также с помощью внутренних функций пакета *Mathematica*. Для правил, срабатывание которых зависит от выполнения условий, наложенных на параметры, применяется запись *a:>b/;test* (такое правило срабатывает только в том случае если условие *test* верно). Кроме этого, *Mathematica* позволяет достаточно просто создавать графические образы с применением

различных графических примитивов, что удобно для графической интерпретации текущего состояния.

ПРИМЕНЕНИЯ ФОРМАЛИЗМА L-СИСТЕМ

Модель роста нитчатой водоросли *Anabaena catenula*

Anabaena – один из примеров многоклеточного организма с клеточной специализацией: в одних клетках (*вегетативных*) осуществляется фотосинтез, в то время как в других (*гетероцистах*) – фиксация азота. Продукты, вырабатываемые в разных клетках водоросли, распределяются между всеми ее клетками путем диффузии.

Клетки *Anabaena* при делении остаются соединенными и в результате образуют линейные филаменты. В отсутствие аммиака или нитрата в субстрате отдельные гетероцисты разделены между собой примерно 10 вегетативными клетками разного размера. Филамент растет за счет асимметричного деления вегетативных клеток. Так как при этом гетероцисты отодвигаются друг от друга, новые гетероцисты трансформируются из вегетативных клеток приблизительно посередине между двумя уже существующими. Поэтому расстояние между гетероцистами остается в фиксированных границах на протяжении всего процесса развития филамента. Эта особенность структуры и является моделируемым структурным инвариантом.

Модель развития *Anabaena* представлена следующей dL-системой:

initial string $F_h(l_{max}, c_{max}, right)\ F_v(l_{max}, c_{max}, right)\ F_h(l_{max}, c_{max}, right)$;

$F(l_l, c_l, h_l)\langle \boldsymbol{F_v}(\boldsymbol{l}, \boldsymbol{c}, \boldsymbol{h})\rangle F(l_r, c_r, h_r)$:

if $(l < l_{max})$ **and** $(c > c_{min})$ **solve** $\frac{dl}{dt} = rl, \frac{dc}{dt} = \frac{k(c_l - 2c + c_r) - ci}{l} - vc$;

if $(l \geq l_{max})$ **and** $(c > c_{min})$ **and** $(h = right)$
produce $F_v\big((1-k)\,l_{max}, c, left\big)\,F_v(kl_{max}, c, right)$;

if $(l \geq l_{max})$ **and** $(c > c_{min})$ **and** $(h = left)$
produce $F_v(kl_{max}, c, left)\,F_v\big((1-k)\,l_{max}, c, right\big)$;

if $c \leq c_{min}$ **produce** $F_h(l, c, h)$;

$\boldsymbol{F_h}(\boldsymbol{l}, \boldsymbol{c}, \boldsymbol{h})$: **solve** $\frac{dl}{dt} = r_x(l_{max} - 1), \frac{dc}{dt} = r_c(c_{max} - c)$.

Начальная конфигурация состоит из трех одинаково ориентированных клеток. Далее определено правило для изменения внутренних клеток. Вегетативные клетки F_v и гетероцисты F_h характеризуются длиной l, концентрацией азотсодержащих веществ c и ориентацией $h = \{left, right\}$. Первое правило для вегетативных клеток F_v показывает, что пока длина клетки l меньше максимального значения l_{max} и концентрация вещества c выше порогового значения c_{min}, длина клетки l увеличивается экспоненциально согласно уравнению $\frac{dl}{dt} = rl$, где r – удельная скорость роста, а концентрация веществ изменяется по закону $\frac{dc}{dt} = \frac{k(c_l - 2c + cr) - ci}{l} - vc$, где v – скорость реакции первого порядка. Первое слагаемое в этом уравнении описывает диффузию веществ через клеточные стенки, а второе – экспоненциальный спад концентрации веществ в клетках. Следующие два правила описывают деление вегетативных клеток. Если клетка достигает максимальной длины l_{max}, а концентрация c еще выше уровня c_{min}, клетка делится согласно ориентации на две вегетативные клетки размера kl_{max} и $(1-k)l_{max}$, концентрация вещества c наследуется от их родительской клетки. Если концентрация падает до границы c_{min}, клетка дифференцируется в гетероцисту. Все правила отвечают непрерывному критерию: сохраняются общая длина и концентрация азотсодержащих веществ.

Последняя строка модели характеризует поведение гетероцист. Их длина и концент-

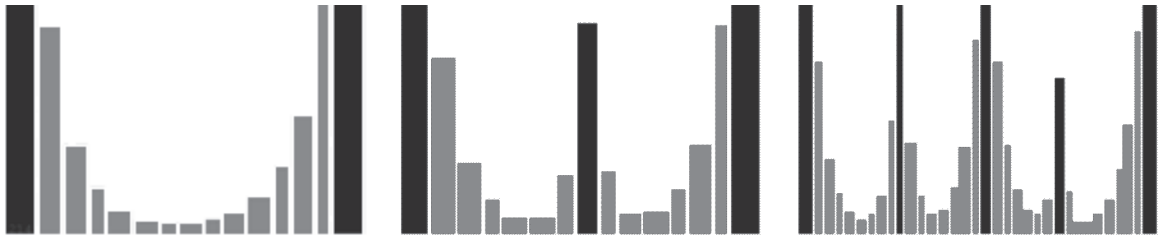

Рис. 3. Результаты моделирования *Anabaena catenula* на трех различных шагах.

Ширина столбца – относительная длина клетки, высота – концентрация вещества *c*. Темные столбцы – гетероцисты, светлые – вегетативные клетки.

рация веществ экспоненциально стремятся к предельным значениям l_{max} и c_{max}. Гетероцисты не подвергаются никаким дальнейшим изменениям. Результаты работы модели приведены на рис. 3.

Представленная модель водоросли *Anabaena catenula* показывает, что простой механизм, основанный на изменении концентрации некоторого вещества (вследствие реакции и диффузии) и простых правил поведения клеток в зависимости от этой концентрации, способен объяснить возникновение пространственно неоднородных клеточный ансамблей, сохраняющих определенный структурный паттерн. Модель хорошо воспроизводит наблюдаемые закономерности в распределении гетероцист в процессе развития водоросли *Anabaena catenula*, а именно: гетероцисты появляются периодично (через 8–10 вегетативных клеток), что подтверждается экспериментальными наблюдениями.

Модель регуляции расположения зон в апикальной меристеме растущего побега растения *Arabidopsis thaliana*

На конце каждой растущей ветви растения расположена апикальная меристема побега (АМП). В АМП выделяют несколько групп клеток (зоны). Так, клетки, расположенные вокруг вертикальной оси меристемы (рис. 4) в радиусе 2–4 клетки в самых верхних 3–4 слоях, синтезируют некоторый белок, называемый CLV3, и принадлежат центральной зоне (ЦЗ). Клетки ЦЗ постоянно делятся с небольшой скоростью и таким образом дают начало всем клеткам побега. Ниже клеток ЦЗ располагаются клетки, экспрессирующие ген *WUS*. Эти клетки относят к организационному центру (ОЦ), толщина которого в вертикальном направлении может составлять 1–3 клетки. Эти группы формируют в АМП определенную стабильную пространственную структуру и оказывают друг на друга регулирующие воздействия. Некоторые дополнительные детали можно найти в работах Николаева с соавт. (2006, 2007, 2010) и в цитируемых в них статьях.

В результате роста и деления происходит смена клеток, составляющих зоны, так что клетки ЦЗ, перемещаясь вниз, становятся клетками ОЦ. В свою очередь клетки ОЦ должны переместиться вниз и стать клетками риб-зоны. В данной работе нас интересовал вопрос о стабильном положении ОЦ на продольной оси АМП относительно верхней точки меристемы и регуляции размера ЦЗ. Для построения одномерной модели мы рассматриваем столбец клеток вокруг центральной оси меристемы (ось *0x* на рис. 4). Соответственно, в начале массива расположены клетки ЦЗ, затем клетки ОЦ и далее клетки, специализирующиеся в клетки сосудистой системы.

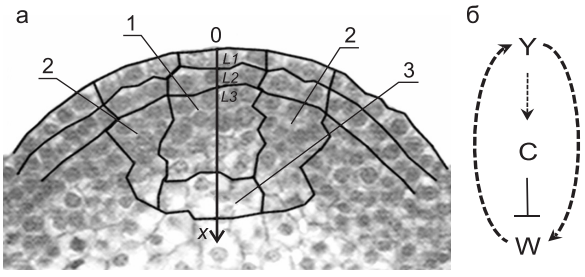

Рис. 4. Продольный срез апикальной меристемы побега *A. thaliana*.

а: 1 – центральная зона, 2 – периферическая зона, 3 – организационный центр; б – взаимная регуляция синтеза веществ *Y*, *C* и *W*. Пунктирные стрелки – активация, сплошная линия – ингибирование.

Для того чтобы изучать эффект деления клеток на структуру АМП, модель была реализована в виде dL-системы. Как отмечено выше, расположение ЦЗ и ОЦ в АМП определяется пространственным распределением концентраций некоторых веществ: *Y*, *C* и *W*. В зависимости от скоростей синтеза этих веществ мы выделяем три типа клеток: S_w – клетки, в которых синтез вещества *W* больше определенного порога, эти клетки в модели представляют ОЦ; S_c – клетки ЦЗ, расположены перед клетками S_w; S_d – клетки, расположенные за клетками S_w. Эти клетки утратили способность делиться и стали специализироваться в клетки сосудистой системы. Клетки типов S_c и S_w при достижении определенного возраста могут делиться и на этом основании объединяются в клетки возобновительной зоны (тип S_r). В итоге при расчете *динамики структуры* в модели различались два типа клеток: S_r и S_d.

В данной модели состояние каждой клетки описывается 4 параметрами: φ – возраст клетки (определяется фазой клеточного цикла), *y*, *c*, *w* – концентрации соответствующих веществ. Таким образом, в нашей модели алфавит L-системы имеет следующий вид:

$$V = \{S_r(\varphi, y, c, w), S_d(\varphi, y, c, w)\}.$$

Динамика переменных состояния каждой клетки описывается следующим образом. Возраст клетки φ увеличивается согласно уравнению:

$$\frac{d\varphi_i}{dt} = \theta\varphi_i,\ i = 1, 2, \ldots, n, \qquad (1)$$

где θ – относительная скорость изменения возраста клетки, в данном случае это постоянная, одинаковая для всех клеток величина, *n* – число клеток.

Параметры *y*, *c*, *w* изменяются благодаря синтезу, диффузии и распаду веществ *Y*, *C* и *W* в клетках. Реакции синтеза веществ происходят со скоростью, зависящей от присутствия других веществ в клетке: скорость синтеза *Y* зависит от концентрации вещества *W* в первой клетке; в зависимости от концентрации *Y* в других клетках может синтезироваться вещество *C*. В тех же клетках в зависимости от концентрации веществ *Y* и *C* может синтезироваться вещество *W* (регуляторные зависимости представлены на рис. 4 справа). Кроме того, все вещества распадаются с определенными скоростями, и между клетками может происходить диффузия веществ *Y* и *W*. В результате динамика концентраций веществ выражается следующей системой уравнений:

$$\begin{aligned}
&\frac{dy_1}{dt} = -a_y y_1 + D_y(y_2 - y_1) + \frac{1}{\tau_y} g(x_1),\ x_1 = h_y + E_{yw} w_1,\\
&\frac{dy_i}{dt} = -a_y y_i + D_y(y_{i-1} - 2y_i + y_{i+1}),\ i = 2, 3, \ldots, n-1,\\
&\frac{dy_n}{dt} = -a_y y_n + D_y(y_{n-1} - y_n),\\
&\frac{dc_i}{dt} = -a_c c_i + \frac{1}{\tau_c} g(U_i),\ U_i = h_c + E_{cy} y_i,\ i = 1, 2, \ldots, n,\\
&\frac{dw_1}{dt} = -a_w w_1 + D_w(w_2 - w_1) + \frac{1}{\tau_w} g(V_1),\ V_1 = h_w + E_{wy} y_1 + E_{wc} c_1,\\
&\frac{dw_i}{dt} = -a_w w_i + D_w(w_{i-1} - 2w_i + w_{i+1}),\ V_i = h_w + E_{wy} y_i + E_{wc} c_i,\ i = 2, \ldots, n-1,\\
&\frac{dw_n}{dt} = -a_w w_n + D_w(w_{n-1} - w_n) + \frac{1}{\tau_w} g(V_n),\ V_n = h_w + E_{wy} y_n + E_{wc} c_n,
\end{aligned} \qquad (2)$$

где y_i, c_i, w_i – концентрации веществ в *i*-й клетке; *a* – коэффициенты распада, *D* – коэффициенты диффузии, E_{ij} – коэффициенты чувствительности регуляции, которые больше нуля, если вещество *j* стимулирует синтез вещества *i*, и меньше нуля, если угнетает; τ_k – коэффициенты, обратные максимальной скорости экспрессии. Параметры h_k, как и E_{ij}, определяют пороговые значения сигмоидной функции $g(x)$.

Аксиома dL-системы задается списком клеток-модулей, снабженных параметрами, значения которых формируются следующим образом. В начальный момент времени возраст клеток φ задается как нормально распределенная слу-

чайная величина со средним значением $\varphi_0 = 1$ и среднеквадратичным отклонением $\sigma = 0{,}1$, с дополнительным условием $\varphi > 0$. Переменные y, c, w в клеточном массиве в начальный момент получают значения стационарного решения системы (1)–(2).

Динамика структуры клеточного массива определяется делением клеток типа S_r в возобновительной зоне. В данной работе мы предполагали, что клетки делятся с некоторой вероятностью в зависимости от их возраста (фазы клеточного цикла, в которой в данный момент времени находится клетка). Как только возраст клетки достигает порогового значения $\varphi = 2\varphi_0$ (в модели $\varphi_0 = 1$), клетка делится на две дочерние, возраст которых $\alpha\varphi$ и $(1-\alpha)\varphi$, где α – нормально распределенная случайная величина со средним значением 0,5 и среднеквадратичным отклонением 0,1, с дополнительным условием $0 < \alpha < 1$. В образовавшихся дочерних клетках значения концентраций веществ Y, C и W наследуются (переписываются) от родительской клетки.

Правила переписывания dL-системы имеют следующий вид:

$$S_r(\varphi, y, c, w) \xrightarrow{\varphi<2\varphi_0,\ c\geq c_0\ w\geq w_0} S_r(\varphi, y, c, w),$$

$$S_r(\varphi, y, c, w) \xrightarrow{\varphi<2\varphi_0,\ c<c_0\ w<w_0} S_d(\varphi, y, c, w),$$

$$S_r(\varphi, y, c, w) \xrightarrow{\varphi\geq 2\varphi_0,\ c\geq c_0\ w\geq w_0} S_r(\alpha\varphi, y, c, w)\ S_r((1-\alpha)\varphi, y, c, w),$$

$$S_r(\varphi, y, c, w) \xrightarrow{\varphi\geq 2\varphi_0,\ c<c_0\ w<w_0} S_d(\alpha\varphi, y, c, w)\ S_d((1-\alpha)\varphi, y, c, w),$$

$$S_d(\varphi, y, c, w) \rightarrow S_d(\varphi, y, c, w).$$

В результате деления клеток возобновительной зоны изменяются структура и размерность динамической системы (1)–(2). В момент времени, когда происходит деление клетки, система перестраивается согласованно с изменением клеточной структуры.

Состояние системы на каждом временном шаге записывается в отдельный список, что позволяет визуализировать результаты (рис. 5).

Для нормального функционирования АМП должна иметь определенную структуру, т. е. «правильное» распределение синтеза (и концентраций) веществ. В результате возмущений, вызываемых делениями клеток, структура АМП может разрушаться. Это значит, что после возмущения она уже не вернется к «правильному» стационарному решению. В данной работе мы считаем, что АМП является живучей, если вероятность разрушения правильной структуры мала на протяжении времени, равного 20 средним клеточным циклам в АМП (например, у *A. thaliana* по времени это в среднем соответствует жизненному циклу растения).

Поскольку в модели диффузия является механизмом распространения сигналов, можно ожидать, что соотношение характерных времен роста и деления клеток, с одной стороны, и диффузии W и Y, с другой стороны, должно играть ключевую роль в устойчивости структуры АМП. В качестве характерного времени диффузионных процессов мы выбрали время достижения определенной концентрации W в первой клетке в результате его диффузии из клеток ОЦ, где происходит его синтез.

В вычислительных экспериментах с моделью было установлено, что уменьшение относитель-

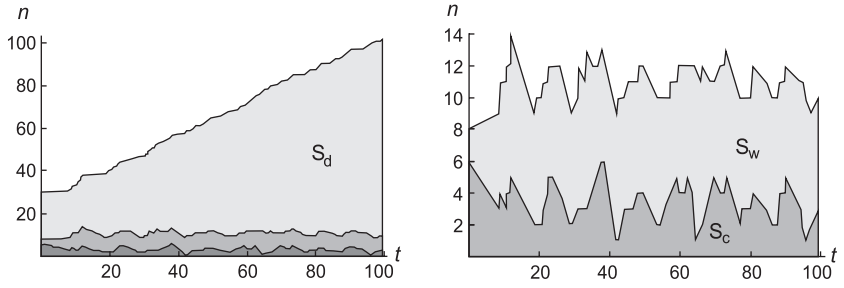

Рис. 5. Результат работы модели. Динамика числа клеток каждого типа в клеточном массиве.

t – время в условных единицах, n – число клеток.

ной длины клеточного цикла приводит к монотонному возрастанию вероятности разрушения структуры возобновительной зоны (рис. 6).

ЗАКЛЮЧЕНИЕ И ПЕРСПЕКТИВЫ

Рассмотренные примеры показывают, что параметризованные L-системы являются удобным формализмом для моделирования роста и развития простых организмов с линейной и ветвящейся структурой. Естественным шагом в продолжение рассмотренной в работе тематики является переход к моделям с более сложной пространственной структурой, например к моделям динамики клеточных пластов. Примеры применения для этого L-систем показывают, что сам формализм накладывает ограничения на множество порождаемых паттернов роста клеточного пласта (Tuza, Lindenmayer, 1992). В качестве альтернативы, у которой этот недостаток отсутствует, интерес представляет модель Хонды (Honda *et al*., 2004) и ее модификации (Merks *et al*., 2011). В этой модели клетки представляются многоугольниками в двумерном случае и многогранниками – в трехмерном, вершины которых перемещаются под воздействием различных сил. Алгоритм деления клеток реализует эмпирическое правило, согласно которому граница между дочерними клетками ориентируется перпендикулярно главной оси эллипсоида момента инерции материнской клетки и делит ее примерно поровну.

Авторы выражают благодарность Н.Л. Подколодному за плодотворное обсуждение работы.

Работа выполнена при частичной финансовой поддержке грантов: РФФИ №11-04-01748-а, интеграционный проект СО РАН №47.

ЛИТЕРАТУРА

Николаев С.В., Колчанов Н.А., Фадеев С.И. и др. Исследование одномерной модели регуляции размеров возобновительной зоны в биологической ткани // Вычисл. технол. 2006. Т. 11. Вып. 2. С. 67–81.

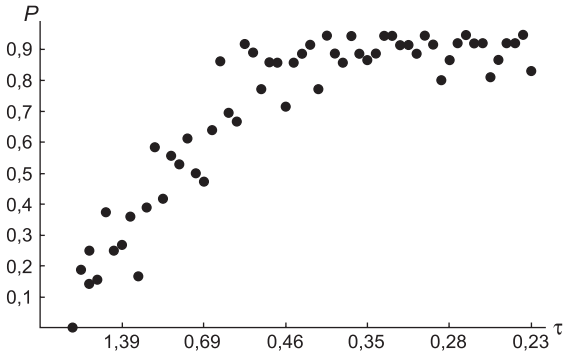

Рис. 6. Вероятность (*P*) разрушения структуры АМП на протяжении времени, равного 20 клеточным циклам, в зависимости от отношения длины клеточного цикла к безразмерному характерному времени диффузии (τ).

Николаев С.В., Пененко А.В., Лавреха В.В. и др. Модельное изучение роли белков CLV1, CLV2, CLV3 и WUS в регуляции структуры апикальной меристемы побега // Онтогенез. 2007. Т. 38. Вып. 6. С. 457–462.

Николаев С.В., Зубаирова У.С., Фадеев С.И. и др. Исследование одномерной модели регуляции размеров возобновительной зоны в биологической ткани с учетом деления клеток // СибЖИМ. 2010. Т. 13. Вып. 4(44). С. 70–82.

Honda H., Tanemura M., Nagai T. A three-dimensional vertex dynamics of model of space-filling polyhedra simulating cell behavior in a cell aggregate // J. Theor. Biol. 2004. V. 226. P. 439–453.

Lindenmayer A. Mathematical models for cellular interaction in development, Parts I and II // J. Theor. Biol. 1968. V. 18. P. 280–315.

Merks R., Guravage M., Inze D., Beemster G. VirtualLeaf: an open-source framework for cell-based modeling of plant tissue growth and development // Plant Physiol. 2011. V. 155(2). P. 656–666.

Prusinkiewicz P., Lindenmayer A. The Algorithmic Beauty of Plants. Springer-Verlag, N.Y., 1990.

Prusinkiewicz P., Hammel M., Mjolsness E. Animation of plant development // Computer Graphics Proceedings, Annual Conference Series. Proc. of SIGGRAPH 93, Anaheim, California (1–6 August, 1993). 1993. P. 351–360.

Tuza Z., Lindenmayer A. Locally Generated Colourings of Hexagonal Cell Division Patterns: Application to Retinal Cell Differentiation // Lindenmayer Systems: Impacts on Theoretical Computer Science, Computer Graphics, and Developmental Biology / Eds G. Rozenberg, A. Salomaa. Berlin: Springer-Verlag, 1992. P. 333–350.

MODELING OF PLANT TISSUE GROWTH AND DEVELOPMENT WITH L-SYSTEMS

U.S. Zubairova[1], A.V. Penenko[2], S.V. Nikolaev[1]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: ulyanochka@bionet.nsc.ru;
[2] Institute of Computational Mathematics and Mathematical Geophysics, Novosibirsk, Russia

Summary

An introduction to modeling of dynamical systems possessing dynamical structures with L-systems is given. Application of L systems is illustrated by models of plant tissue growth and control of state variable distribution in the growing tissue.

Key words: developmental biology, dynamical systems with dynamical structure, mathematical model, L-systems.

УДК 575;004.94;579.23+578.81

ВЫСОКОПРОИЗВОДИТЕЛЬНОЕ МОДЕЛИРОВАНИЕ ЭВОЛЮЦИИ ПРОКАРИОТИЧЕСКИХ СООБЩЕСТВ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММНОГО КОМПЛЕКСА «ГАПЛОИДНЫЙ ЭВОЛЮЦИОННЫЙ КОНСТРУКТОР»

© 2012 г. **З.С. Мустафин[1, 2], Ю.Г. Матушкин[1, 2], С.А. Лашин[1, 2]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: Zidane-7@yandex.ru;
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

Поступила в редакцию 15 июля 2012 г. Принята к публикации 31 августа 2012 г.

В работе приведены результаты разработки высокопроизводительной версии программного комплекса «Гаплоидный эволюционный конструктор» (http://evol-constructor.bionet.nsc.ru), предназначенного для моделирования функционирования и эволюции прокариотических сообществ. Разработана параллельная версия программы, предназначенная для работы на высокопроизводительных кластерах с поддержкой MPI. Оказалось, что общее ускорение параллельной версии программного комплекса почти линейно зависит от числа используемых процессоров, и время расчета сложных моделей сообществ на кластере ЦКП «Биоинформатика» СО РАН уменьшилось с десятков часов до нескольких минут.

Ключевые слова: микробное сообщество, оптимизация, параллельное программирование, моделирование, эволюция.

ВВЕДЕНИЕ

Моделирование эволюции и функционирования прокариотических сообществ является актуальной задачей современной системной биологии как с фундаментальной, так и с практической точек зрения. Прокариоты способны катализировать огромное количество разнообразных биохимических реакций и потому являются участниками большинства природных процессов (Заварзин, 2003). Многие виды прокариот используются человеком в технологических процессах. Математическое и компьютерное моделирование поведения и эволюции прокариотических сообществ в тех или иных условиях можно использовать для нужд современной биологии и медицины (Wang, Post, 2012). Данная работа посвящена развитию методов моделирования эволюции и функционирования сообществ одноклеточных гаплоидных организмов и программного комплекса «Гаплоидный эволюционный конструктор» (ГЭК – доступен по адресу http://evol-constructor.bionet.nsc.ru).

В компьютерной модели ГЭК рассматриваются следующие уровни биологической организации: геномный, метаболический, популяционный и экоценотический (Lashin *et al*., 2012). Также в ГЭК реализована возможность моделирования функционирования генных сетей с учетом популяционных и экоценотических факторов. Впервые применен подход описания «обобщенных геномов популяций» с помощью техники генетических спектров, что позволяет значительно сократить время расчета модели с сохранением точности, сопоставимой с точностью классических индивидуально ориентированных моделей. Кроме того, ГЭК позволяет разрабатывать множество разных подмоделей для одного слоя, фактически создавая библиотеки подмоделей. При этом поскольку интерфейсы взаимодействия между подмоделями

разных слоев (в частности входные и выходные данные) четко специфицированы, при построении общей модели появляется возможность комбинирования различных сочетаний подмоделей разных слоев. Это позволяет исследовать различные аспекты эволюционного процесса в рамках одного программного средства.

ГЭК позволяет моделировать мутации, горизонтальный перенос и потерю генов, фиксацию генетических изменений. Горизонтальный перенос, а также потеря генетического материала изменяют уникальный набор метаболических реакций, характерных для данной популяции клеток («вида»), т. е. фактически структуру клеточного метаболизма, что моделирует появление новых штаммов/«видов» клеток и лежит в основе развития биоразнообразия, моделируемого ГЭК. К числу других возможностей ГЭК относятся возможности моделирования фаговой инфекции (Лашин и др., 2011) и функционирования генных сетей (Lashin, Matushkin, 2012). Интеграция методики моделирования генных сетей открывает широкие методические перспективы для исследования эволюции генных сетей с учетом надгенетических и надорганизменных уровней биологической организации, таких, как популяционный и экоценотический.

В работе приведены результаты разработки и программной реализации высокопроизводительного алгоритма моделирования популяционных процессов в рамках ГЭК: описаны эффективный алгоритм расчета изменения численности прокариотической популяции и параллельная реализация алгоритма с использованием технологии MPI (Корнеев, 2002); проведены тестирование и оценка производительности алгоритма на высокопроизводительном кластере.

Алгоритм расчета изменения численности популяции

Анализ однопроцессорной версии ГЭК при помощи профилировщика Intel Parallel Amplifier (http://software.intel.com/ru-ru/articles/intel-parallel-studio-home) показал, что при моделировании сообществ с высоким генетическим разнообразием (10^6–10^8 уникальных аллельных комбинаций) практически все время выполнения уходит на вычисление единственной функции – изменения численности популяции независимо от типа трофической стратегии (рис. 1).

На рис. 1 показано, что с ростом количества аллельных комбинаций время выполнения программы практически полностью концентрируется на функции изменения численности популяции. Опишем эту функцию более формально.

Основным объектом, с которым работает эта функция, является ОГП (обобщенный геном популяции). ОГП в ГЭК – это многомерное распределение частот аллелей для всех генов, присутствующих у особей популяции. Наличие гена в клетках популяции в ГЭК подразумевает наличие в метаболизме этих клеток процесса синтеза или утилизации соответствующего субстрата; аллель как вариант гена определяет конкретное значение константы скорости соответствующего процесса. Заметим, что в рамках ГЭК каждый признак однозначно определяется одним геном, и ген рассматривается как единица наследования.

На рис. 2 показан ОГП, содержащий 4 гена: с тремя, одним, четырьмя, двумя и возможными аллельными вариантами соответственно. Для расчета прироста численности популяции с учетом различной приспособленности особей, несущих разные аллельные комбинации, необходимо рассмотреть все возможные такие комбинации и учесть в каждой из них изменение размера субпопуляции с помощью заданной пользователем функции роста популяции, так называемой трофической стратегии (Лашин и др., 2009), а затем посчитать итоговый размер популяции и концентрации аллельных вариантов.

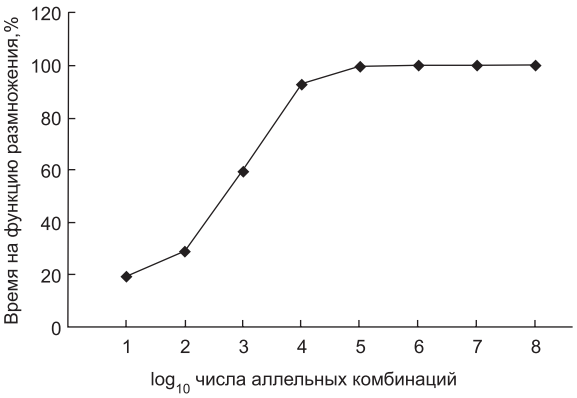

Рис. 1. Время выполнения функции изменения численности популяции относительно времени работы программы.

```
0(0.5)  1(0.2)  2(0.3)          – распределение аллелей для гена 1
3(1)                           – распределение аллелей для гена 2
4(0.1)  5(0.1)  6(0.4)  7(0.4) – распределение аллелей для гена 3
8(0.9)  9(0.1)                 – распределение аллелей для гена 4
```

Рис. 2. Пример представления обобщенного генома популяции в ГЭК.

Идея предлагаемого в статье алгоритма, реализующего изменение численности популяции, заключается в переборе всех возможных аллельных комбинаций (всего их n) в одном цикле длины n. Комбинации записываются с помощью специального массива, состоящего из индексов каждого аллельного варианта комбинации. В качестве примера рассмотрим табл. 1.

Все аллельные варианты всех генов записываются в один массив (полужирным текстом выделены текущие аллельные варианты каждого гена). С помощью массива индексов (в стартовой комбинации он имеет вид (0, 3, 4, 8)) выписывается первая комбинация, и индекс текущего аллельного варианта в последнем гене увеличивается на 1. Если при этом в последнем гене получен последний аллельный вариант, то значение выбранного аллельного варианта в предыдущем гене увеличивается на 1, а в последнем гене выбирается первый аллельный вариант. Таким образом, с помощью массива индексов осуществляется полный перебор всех аллельных комбинаций в популяции.

Достоинство такого подхода состоит в том, что полученный цикл может быть разбит на любое число параллельных процессов (протестировано до 900 процессов). Каждому процессу необходимо «знать» массивы значений и концентраций аллельных вариантов (так называемые «развертки» значений и концентраций), а также стартовую и конечную позиции своего фрагмента цикла. По окончании вычислений в каждом процессе полученные данные суммируются в корневой процесс с помощью функции массового суммирования MPI_Reduce, происходит присваивание посчитанных результатов исходному объекту и алгоритм завершает работу (рис. 3). Все пересылки реализуются с помощью функций MPI_Bcast и MPI_Reduce, все процессы выполняют приблизительно одинаковый объем работы (несущественные отличия возникают, если число комбинаций не кратно числу процессов).

Таблица 1

Возможные аллельные комбинации

Аллельные варианты всех генов	Полученные комбинации
(**0**, 1, 2, **3**, **4**, 5, 6, 7, **8**, 9)	(0, 3, 4, 8)
(**0**, 1, 2, **3**, **4**, 5, 6, 7, 8, **9**)	(0, 3, 4, 9)
(**0**, 1, 2, **3**, 4, **5**, 6, 7, **8**, 9)	(0, 3, 5, 8)
…	…
(0, 1, **2**, **3**, 4, 5, 6, 7, 8, **9**)	(2, 3, 7, 9)

Тестирование алгоритма

В рамках работы было проведено тестирование алгоритма на 6-ядерных процессорах X5670 2.93 GHz (Westmere) кластера НКС 30-Т (http://bioinformatics.bionet.nsc.ru/). Алгоритм был верифицирован на тестовом наборе сценариев ГЭК (Лашин и др., 2009; Lashin *et al*., 2012). Затем были составлены специальные нагрузочные тесты (10^6–10^8 аллельных комбинаций). Результаты тестирования алгоритма приведены в табл. 2.

В ряде вычислительных экспериментов было показано, что эффективность распараллеливания иногда превышает значение 1. Это объясняется тем, что на различных узлах кластера могут быть получены различные значения времени выполнения программы. Расчет параллельной версии на узлах, отличных от узлов, на которых рассчитывалась последовательная версия, может дать увеличение или уменьшение эффективности. Среднеквадратичное отклонение для каждой выборки результатов показано в таблице в столбце «дисперсия», с округлением вверх до секунд. Столбец «ускорение» показывает эффективность распараллеливания для различного числа процессов. Получено практически линейное ускорение на протяжении всего тестирования.

Таким образом, время выполнения программы на современных процессорах сократилось с 8 ч до 2 мин, при этом требуется всего 12 уз-

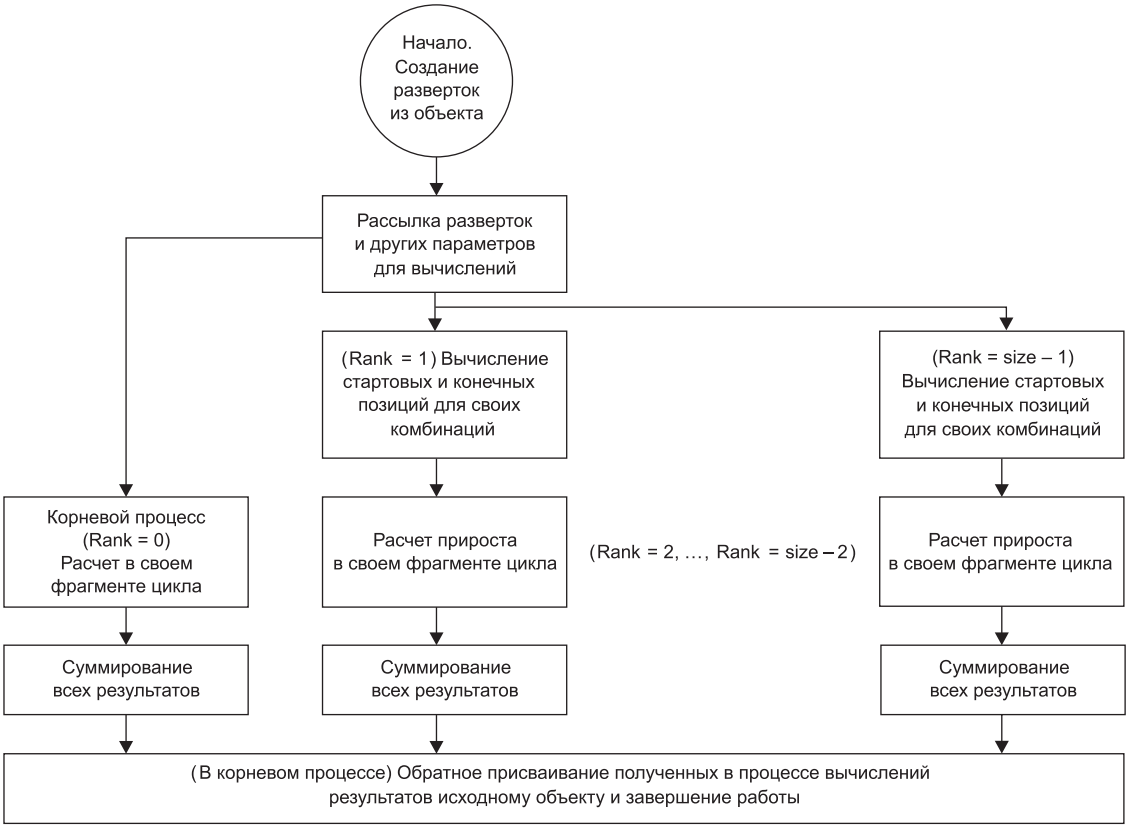

Рис. 3. Схема распараллеливания алгоритма.

Таблица 2

Результаты тестирования на X5670

Количество процессов	Время работы ч:м:с	Эффективность распараллеливания	Ускорение	Дисперсия, с
1	8:02:26	1	1	249
2	4:10:24	0,9633	1,9266	65
4	2:05:20	0,9623	3,8492	188
8	1:02:03	0,9718	7,7744	4
16	0:31:06	0,9695	15,512	3
24	0:20:48	0,9664	23,1936	3
36	0:13:46	0,9734	35,0424	1
64	0:07:52	0,9582	61,3248	2
96	0:05:13	0,9633	92,4768	1
144	0:03:32	0,9481	136,5264	1
264	0:02:03	0,8914	235,3296	2

лов вычислительного комплекса, и на каждый параллельный процесс необходимо всего 2 Мб оперативной памяти.

ЗАКЛЮЧЕНИЕ

Нами была проведена оптимизация алгоритма расчета изменения численности прокариотической популяции в программном комплексе ГЭК, которая позволила проводить расчеты на параллельных высокопроизводительных вычислительных кластерах. Чем сложнее структура моделируемого сообщества, чем больше в этом сообществе генетическое разнообразие, тем большую долю в выполнении программы занимает выполнение функции расчета изменения численности популяции, соответственно, на сложных моделях выигрыш от оптимизации является максимальным. Показано, что ускорение в зависимости от числа использованных процессоров близко к линейному и достигает максимума при моделировании сообществ с большим генетическим разнообразием. Мы считаем, что именно такие сообщества представляют наибольший интерес для исследования, и надеемся, что высокопроизводительная версия ГЭК, представленная в данной статье, позволит пользователю увеличить сложность и разнообразие моделируемых биологических ситуаций, что будет способствовать развитию эволюционной биологии.

БЛАГОДАРНОСТИ

Работа была поддержана следующими грантами: РФФИ 12-07-00671, Междисциплинарные интеграционные проекты СО РАН №№ 47, 87, Научная школа-5278.2012.4; Программа РАН № 28.

ЛИТЕРАТУРА

Заварзин Г.А. Лекции по природоведческой микробиологии. М.: Наука, 2003. С. 348.

Корнеев В.Д. Параллельное программирование в MPI. 2-е изд. испр. Новосибирск: ИВМиМГ СО РАН, 2002. 215 с.

Лашин С.А., Суслов В.В., Матушкин Ю.Г. Моделирование эволюции трофически замкнутых сообществ с компенсаторным и некомпенсаторным метаболизмом // Информ. вестник ВОГиС. 2009. Т. 13. № 1. С. 150–158.

Лашин С.А., Матушкин Ю.Г., Суслов В.В., Колчанов Н.А. Эволюционные тренды в системах «Прокариотическое сообщество» и «Прокариотическое сообщество–фаг» // Генетика. 2011. Т. 47. № 12. С. 1676–1685.

Lashin S.A., Matushkin Yu.G. Haploid evolutionary constructor: new features and further challenges // In Silico. Biol. 2012. V. 11. No. 3. P. 125–135.

Lashin S.A., Matushkin Yu.G., Suslov V.V., Kolchanov N.A. Computer modeling of genome complexity variation trends in prokaryotic communities under varying habitat conditions // Ecol. Modelling. 2012. V. 224. No. 1. P. 124–129.

Wang G., Post W.M. A theoretical reassessment of microbial maintenance and implications for microbial ecology modeling // FEMS Microbiol. Ecol. 2012. Sep;81(3):610-7. doi: 10.1111/j.1574-6941.2012.01389.x. Epub 2012 Apr 30. (ссылка в пабмед: http://www.ncbi.nlm.nih.gov/pubmed/22500928).

HIGH-THROUGHPUT SIMULATIONS OF PROKARYOTIC COMMUNITY EVOLUTION WITH HAPLOID EVOLUTIONARY CONSTRUCTOR

Z.S. Mustafin[1, 2], Yu. G. Matushkin[1, 2], S.A. Lashin[1, 2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia, e-mail: Zidane-7@yandex.ru;
[2] Novosibirsk National Research State University, Novosibirsk, Russia

Summary

The results of the development of a high-throughput version of the software package Haploid Evolutionary Constructor (HEC), available at http://evol-constructor.bionet.nsc.ru, are presented. The software is used to simulate the functioning and evolution of prokaryotic communities. A parallel version of the software package was created using the MPI technology. The test was performed on a cluster of the Bioinformatics shared access center. The acceleration obtained was almost linear. The simulation time of complex bacterial communities was reduced from dozens of hours to several minutes.

Key words: microbial communities, optimization, parallel computing, modeling, evolution.

УДК 575;004.94;579.23+578.81

РАЗРАБОТКА ПРОСТРАНСТВЕННО РАСПРЕДЕЛЕННОЙ МОДЕЛИ ЭВОЛЮЦИИ ПРОКАРИОТИЧЕСКИХ СООБЩЕСТВ

© 2012 г. **С.А. Лашин[1, 2], Е.А. Мамонтова[1, 2], Ю.Г. Матушкин[1, 2]**

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

Поступила в редакцию 15 июля 2012 г. Принята к публикации 31 августа 2012 г.

В статье описаны методика моделирования функционирования и эволюции прокариотических сообществ и программный комплекс «Гаплоидный эволюционный конструктор» (http://evol-constructor.bionet.nsc.ru) для построения моделей с пространственным распределением по одной координате (1D). Приведено сравнение 0D (с полным перемешиванием) и 1D моделей прокариотического сообщества вида «отравитель–жертва». Показано влияние пространственного распределения субстратов и прокариотических клеток на стабильность сообщества.

Ключевые слова: моделирование эволюции, микробное сообщество, пространственное распределение.

ВВЕДЕНИЕ

Математическое и компьютерное моделирование является важным инструментом теоретического исследования эволюции биологических сообществ. В данной работе рассматриваются методика моделирования эволюции популяций одноклеточных гаплоидных организмов (клеток) и ее реализация в виде программного комплекса «Гаплоидный эволюционный конструктор» (ГЭК, http://evol-constructor.bionet.nsc.ru) (Lashin, Matushkin, 2012). ГЭК позволяет моделировать функционирование сетей популяций, трофически связанных между собой субстрат-продуктными отношениями, в среде с протоком. В модели учтены различные уровни биологической организации: генетический (мутации, горизонтальный перенос генов), метаболический (утилизация, синтез, транспорт метаболитов), популяционный (размножение, конкуренция, естественный отбор), экоценотический (проток в среде, обмен метаболитами). Проблема вычислительной сложности в ГЭК решена путем описания популяции как единого объекта: популяция определяется как сообщество генетически идентичных или почти идентичных (с точностью до аллелизма) одноклеточных гаплоидных асексуальных организмов (клеток). Каждый ген определяет эффективность одного из трех типов метаболических процессов: утилизации субстратов, синтеза продуктов и секреции продуктов.

Расширение методики моделирования и программы ГЭК заключается в переходе от модели с полным перемешиванием субстратов и клеток в гомогенной среде (0D) к модели с пространственным распределением субстратов и клеток по одной координате (1D).

Традиционно пространственное распределение данных в системе описывали посредством нелинейных дифференциальных уравнений в частных производных. Базовая модель системы, содержащей источник энергии, с учетом химических реакций и диффузии рассматривает систему уравнений

$$\begin{cases} \dfrac{\partial x}{\partial t} = P(x, y) + D_x \dfrac{\partial^2 x}{\partial r^2} \\ \dfrac{\partial y}{\partial t} = Q(x, y) + D_y \dfrac{\partial^2 y}{\partial r^2} \end{cases}, \qquad (1)$$

где r – пространственная координата, $D_x\partial^2 x/\partial r^2$ и $D_y\partial^2 y/\partial r^2$ описывают диффузию веществ x и y вдоль этой координаты. Данные уравнения,

называемые уравнениями «реакция–диффузия», были предложены А. Тьюрингом в его работе (1952 г.), ставшей классической. Диффузия нелинейно связанных компонент x и y в данной системе приводит не к выравниванию, а периодическому во времени, неравномерному в пространстве распределению (Ризниченко, 2003).

Дальнейшее развитие пространственно распределенных систем привело к возникновению новых техник моделирования, основанных на методах теории вероятностей, дискретной математики и др. К ним относят, например, методы Монте-Карло (Haviland, Lavin, 1962; Bird, 1965; Хлопков, 2006), клеточные автоматы (Бандман, 2006).

В ранее разработанных пространственно распределенных моделях популяционной биологии (Разумовский, 1981; Раутиан, Жерихин, 1997; Левченко, 2003) основной упор делается на миграцию особей, тесно связанную с видообразованием и сукцессией. В то же время эти модели характеризуют популяции высокоразвитых организмов и не описывают случай бактериальных сообществ, в которых процессы диффузии играют бóльшую роль, нежели самопроизвольные миграции.

Особенности микроорганизмов – высокое отношение поверхности к объему, как следствие – высокая интенсивность обмена с окружающей средой, быстрые темпы размножения и прироста биомассы – отражены в классической модели проточной культуры микроорганизмов, разработанной Моно (Monod, 1942) и Гербертом (Herbert, 1958), однако пространственное распределение в ней не учтено: предполагается непрерывное перемешивание субстрата и клеток.

Таким образом, добавление пространственного распределения к числу факторов, учитываемых ГЭК, делает эту методику более гибкой и универсальной, расширяет спектр решаемых биологических задач.

Описание алгоритма

С помощью ГЭК моделируются функционирование и эволюция сообщества, состоящего из прокариотических клеток, живущих в единой проточной среде фиксированного объема с идеальным перемешиванием организмов и субстратов в ней. Организмы могут потреблять (утилизировать) субстраты и производить (синтезировать), после чего секретировать в окружающую среду продукты, которые могут в свою очередь потребляться другими организмами. Численность популяций за одно поколение изменяется в зависимости от количества потребленных субстратов, размера популяции, скорости протока в окружающей среде, коэффициента смертности популяции и ее «естественного прироста» (такая зависимость называется «трофической стратегией») (рис. 1).

Созданные в рамках этой работы программы позволяют исследовать расширенную модель,

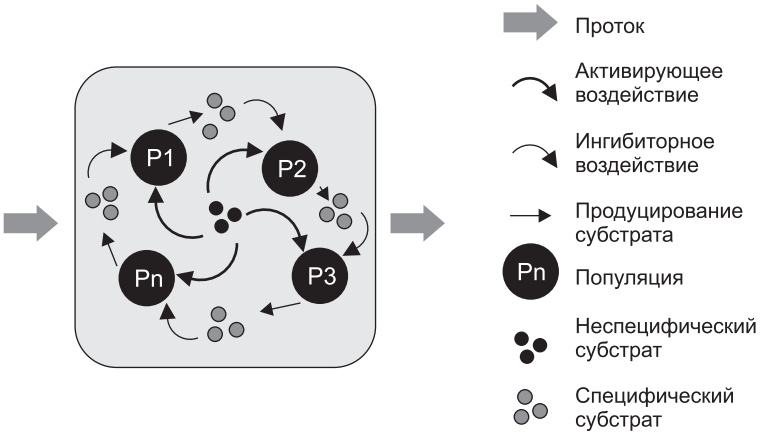

Рис. 1. Схема основных процессов взаимодействия популяций друг с другом и со средой в случае идеального перемешивания.

Неспецифический субстрат необходим для существования всем популяциям, специфические субстраты, производимые одной популяцией, могут оказывать на другие популяции как активирующее, так и ингибирующее действие.

в которой среда рассматривается не как единый объем, а как упорядоченный конечный набор равных объемов – «точечных сред» с идеальным перемешиванием в каждой из них и общим сквозным протоком. Для удобства будем считать, что нумерация сред возрастает по направлению протока (рис. 2).

Итерационный шаг эволюционного процесса в расширенной модели частично повторяет соответствующий шаг для модели с идеальным перемешиванием: для каждой из n точечных сред моделируются процессы утилизации субстрата организмами, размножение популяций, секреция продуктов в среду, мутации и горизонтальный перенос генов (последний этап имеет стохастический характер). Простой этап моделирования протока в среде заменяется в расширенной модели составным этапом, включающим моделирование взаимодействия соседних точечных сред (диффузию субстратов и миграции организмов) и моделирование протока (рис. 3).

Итак, перераспределение субстратов в среде моделируется двумя процессами: диффузией и переносом с помощью протока.

Пусть $N_i^j(t)$ – концентрация i-го неспецифического субстрата в j-й точечной среде в момент времени t; $S_i^j(t)$ – концентрация i-го специфического субстрата в j-й точечной среде в момент времени t. Тогда формулы перераспределения неспецифического субстрата на шаге $t+1$ принимают вид:

$$\begin{cases} N_i^j(t+1) = N_i^j(t) + k_{flow}\left(N_{flow,i} - N_i^j(t)\right) + D\left(N_i^{j+1}(t) - 2N_i^j(t)\right) - Cons_i^j(P) \text{ при } j=1, \\ N_i^j(t+1) = N_i^j(t) + k_{flow}\left(N_i^{j-1}(t) - N_i^j(t)\right) + D\left(N_i^{j-1}(t) - 2N_i^j(t)\right) - Cons_i^j(P) \text{ при } j=n, \\ N_i^j(t+1) = N_i^j(t) + k_{flow}\left(N_i^{j-1}(t) - N_i^j(t)\right) + D\left(N_i^{j-1}(t) + N_i^{j+1}(t) - 2N_i^j(t)\right) - Cons_i^j(P) \text{ при } 1<j<n. \end{cases} \quad (2)$$

k_{flow} – скорость протока (доля объема притока в и оттока из точечной среды), D – коэффициент диффузии i-го субстрата, $Cons_i^j(P)$ – слагаемое, описывающее потребление субстрата клетками сообщества.

Формулы для специфических субстратов аналогичны, за исключением добавления слагаемого $Synt_i^j$, описывающего совокупный синтез i-го специфического субстрата клетками сообщества, а также отсутствия действия притока в первой точечной среде. Потребление и синтез субстратов рассмотрены более подробно в наших ранних работах (Lashin *et al*., 2010; Lashin, Matushkin, 2012).

Перемещение организмов (клеток) в отличие от перемещения субстратов характеризуется миграцией с некоторой собственной скоростью. Моделирование перераспределения организмов в среде состоит из последовательного вычисления доли популяции, мигрирующей против потока, затем вдоль потока. Расчет долей популяции, мигрирующей против потока и вдоль протока ($Ratio_{left}$ и $Ratio_{right}$ соответственно), описывается формулой

$$Ratio_{left,right} = U \pm k_{flow}, \quad (3)$$

где U – доля мигрирующих клеток популяции («рассеяние»).

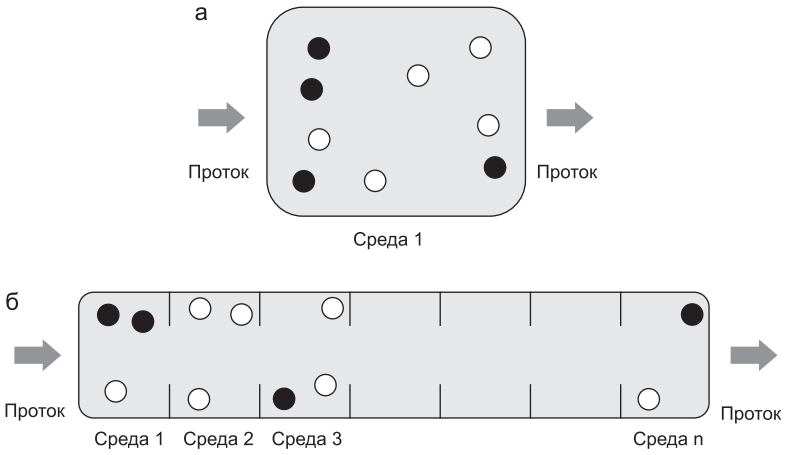

Рис. 2. Физическое представление моделей ГЭК.

а – схема модели с полным перемешиванием (0D); б – схема пространственно распределенной модели (1D).

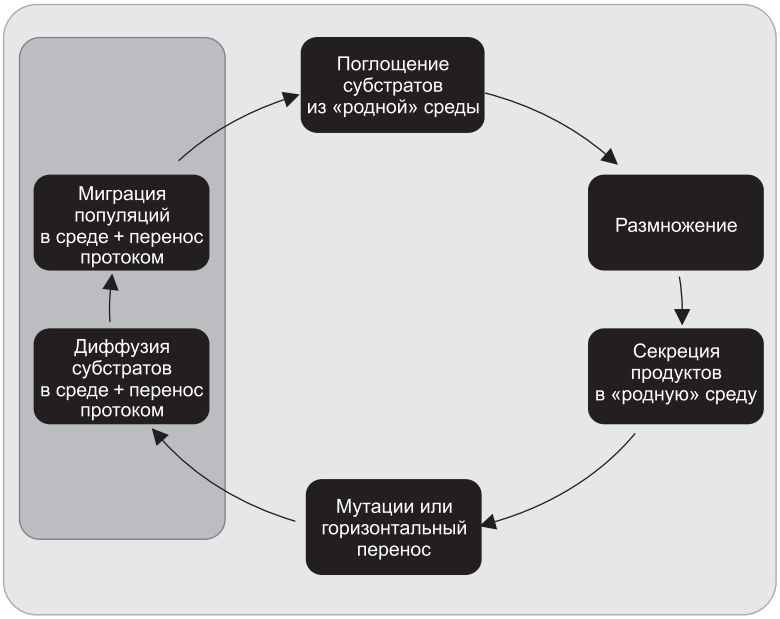

Рис. 3. Схема одного итерационного шага эволюционного процесса для расширенной модели методики ГЭК.

Темная область отмечает шаги, отсутствующие в модели 0D.

Далее производится ряд проверок:

– если $Ratio_{left} < 0$, перемещения против потока не происходит, скорости популяции недостаточно, чтобы преодолеть действие протока;

– если $Ratio_{left} > 1/3$, его значение уменьшается до 1/3. Такая оценка сверху продиктована следующими соображениями: в отсутствие протока наибольшая возможная доля миграции составляет по одной трети популяции по каждому из двух направлений (одна треть остается в родительской среде). В остальных случаях доля миграции вдоль протока предполагается большей, чем доля миграции против него;

– для первой точечной среды устанавливается $Ratio_{left} = 0$, против протока из общей среды миграции – нет;

– если $Ratio_{left} + Ratio_{right} > 1$, значение $Ratio_{right}$ уменьшается до $1 - Ratio_{left}$ во избежание миграции субпопуляции, превосходящей размером исходную популяцию;

– для последней точечной среды $Ratio_{right} = k_{flow}$ миграция за пределы общей среды не предусмотрена, доля переноса «вправо» – это доля популяции, вымываемая протоком.

Таким образом, существующая математическая модель функционирования и эволюции бактериальных сообществ была расширена путем пространственного распределения по одной координате, описаны соответствующие методы перераспределения субстратов и организмов в среде.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Интеграция расширенной методики в программный комплекс ГЭК позволила построить компьютерные модели и провести верификацию методики на результатах их функционирования.

В первую очередь были построены базовые модели, отражающие действие протока, диффузионные процессы для субстратов и миграции организмов в среде. В начальный момент времени рассматриваемый объект (субстрат/популяция) помещался в первую из десяти моделируемых точечных сред. Результаты моделирования представлены на графиках (рис. 4): ярко выражены диффузионные процессы для субстратов, сопровождаемые накоплением неспецифического субстрата и вымыванием специфического. Для популяции наблюдаются прирост численности в условиях достаточного количества специфического субстрата и гибель организмов, вызванная его невосполняемостью.

Была также промоделирована коэволюция системы «отравитель–жертва». Данная система представлена двумя популяциями, существующими в одной среде. Популяция P_1 («отравитель») производит специфический субстрат S_2, оказывающий ингибиторное воздействие на P_2; P_2 («жертва») производит специфический субстрат S_1, жизненно необходимый для P_1 (рис. 5, а). В терминах модели ГЭК трофическая страте-

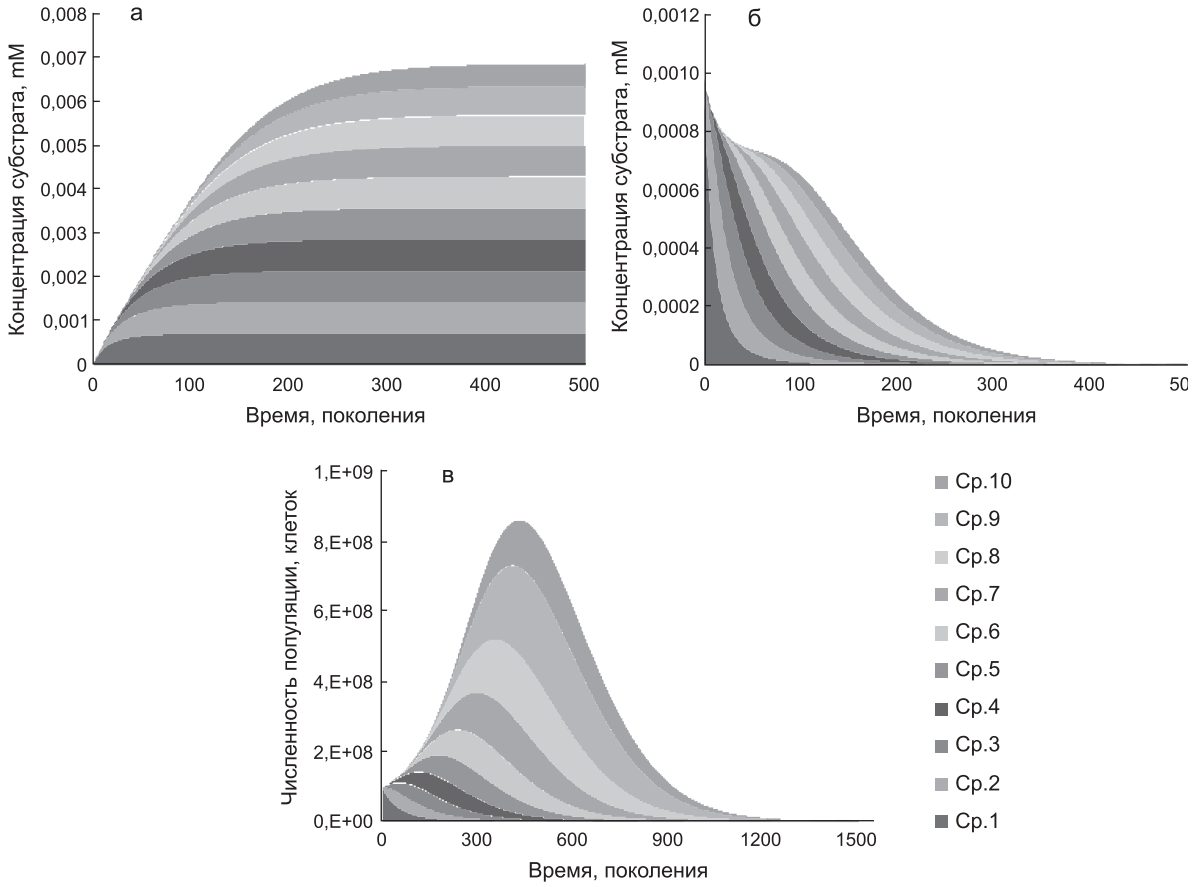

Рис. 4. Распространение субстратов в среде без популяций.

а – неспецифический субстрат, б – специфический субстрат; в – распространение популяции в среде с равномерным начальным распределением субстратов.

гия отравителя компенсаторная (недостаток неспецифического субстрата может частично компенсироваться специфическими субстратами), соответствующая закону компенсации факторов (Rübel, 1927); трофическая стратегия жертвы – ингибиторная (специфические субстраты оказывают на популяцию негативное влияние).

Компенсаторная и ингибиторная трофические стратегии описываются формулами:

$$\Delta P = \sqrt{r_1 n_1(P) \cdot c_1 s_1(P)} - k_{flow} \cdot P - k_{death} \cdot P^2, \quad (4)$$

$$\Delta P = P \cdot (r_1 n_1 - c_2 s_2 - k_{flow}) - k_{death} \cdot P^2, \quad (5)$$

где r_1 – значение константы эффективности утилизации неспецифического субстрата N_1; n_1 – количество съеденного неспецифического субстрата N_1; c_1 – значение константы эффективности утилизации специфического субстрата S_1; s_1 – количество съеденного неспецифического субстрата S_1; c_2 – значение константы эффективности утилизации специфического субстрата S_2; s_2 – количество съеденного неспецифического субстрата S_2.

В зависимости от значений констант эффективности утилизации специфического субстрата и их отношений возможны три основных сценария развития системы: гибель сообщества (рис. 5, б), установление стационарного состояния численностей популяций (рис. 5, в) или незатухающие колебания в системе (рис. 5, г).

Близость к источнику вещества является в данной модели важным фактором устойчивости системы. Например, помещение популяций в начальный момент в третью среду делает сообщество более уязвимым: стационарное состояние возможно лишь при крайне малой чувствительности жертвы к ингибитору, в остальных случаях сообщество вымирает. При помещении популяций в шестую среду сообщество гибнет даже при крайних значениях констант c_1 и c_2.

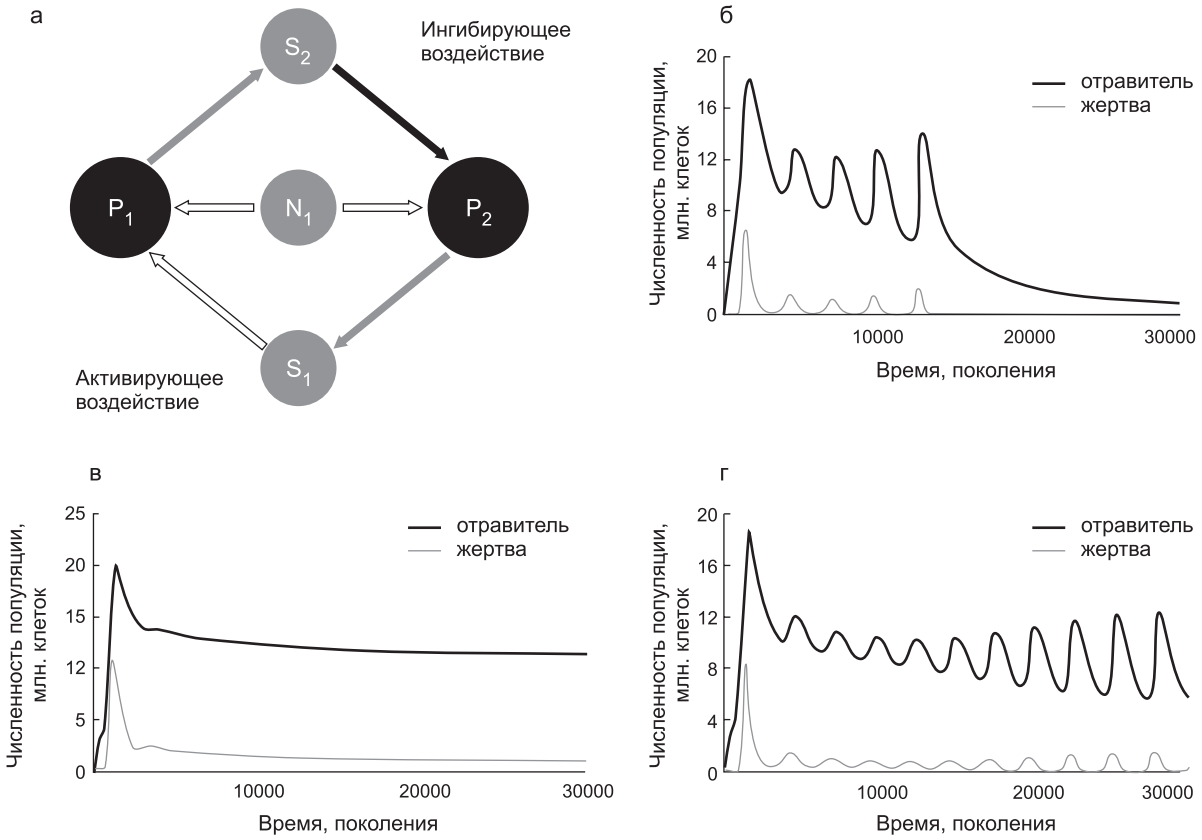

Рис. 5. Система «отравитель–жертва».

а – схема трофического взаимодействия популяций с различными трофическими стратегиями: P_1 «травит» P_2, P_2 «кормит» P_1. Возможные сценарии развития системы: б – гибель сообщества ($c_1 = 6$, $c_2 = 2$), в – выход на стационарное состояние ($c_1 = 5$, $c_2 = 1$), г – незатухающие колебания в системе ($c_1 = 5$, $c_2 = 2$).

Начальное размещение жертвы и отравителя в различные точечные среды также влияет на устойчивость системы. Уже в случае разбиения на две точечные среды стационарное состояние возможно лишь при $c_1 = 1$, остальные случаи приводят систему к гибели. Отметим также, что для двух точечных сред не имеет существенного значения, кто был размещен в первой среде, отравитель или жертва.

При разбиении на большее число точечных сред при тех же условиях и размещении популяций в крайние среды жертва достаточно быстро гибнет независимо от чувствительности к ингибитору, отравитель же может просуществовать некоторое время при больших значениях константы c_1, но это не спасает сообщество от гибели. Однако в таком случае начальное размещение жертвы и отравителя оказывает некоторое влияние на динамику численности последнего: отравитель получает преимущество, если он помещен изначально в последнюю среду, жертва – в первую.

Наконец, было рассмотрено влияние мутаций, произошедших в первой точечной среде, на поведение системы в целом. На основе компьютерных моделей ГЭК были получены следующие результаты. Мутации, изменяющие ген утилизации субстрата S_1 отравителем, способны приводить к следующим изменениям:

1) выводить систему из стационарного состояния, приводя к незатухающим колебаниям;

2) выводить систему из стационарного состояния, приводя ее к гибели.

Мутации, изменяющие ген утилизации S_2 жертвой, способны привести к следующим изменениям:

1) приводить систему с колебаниями в стационарное состояние;

2) приводить систему, обреченную на гибель, к стационарному состоянию (рис. 6, а, б);

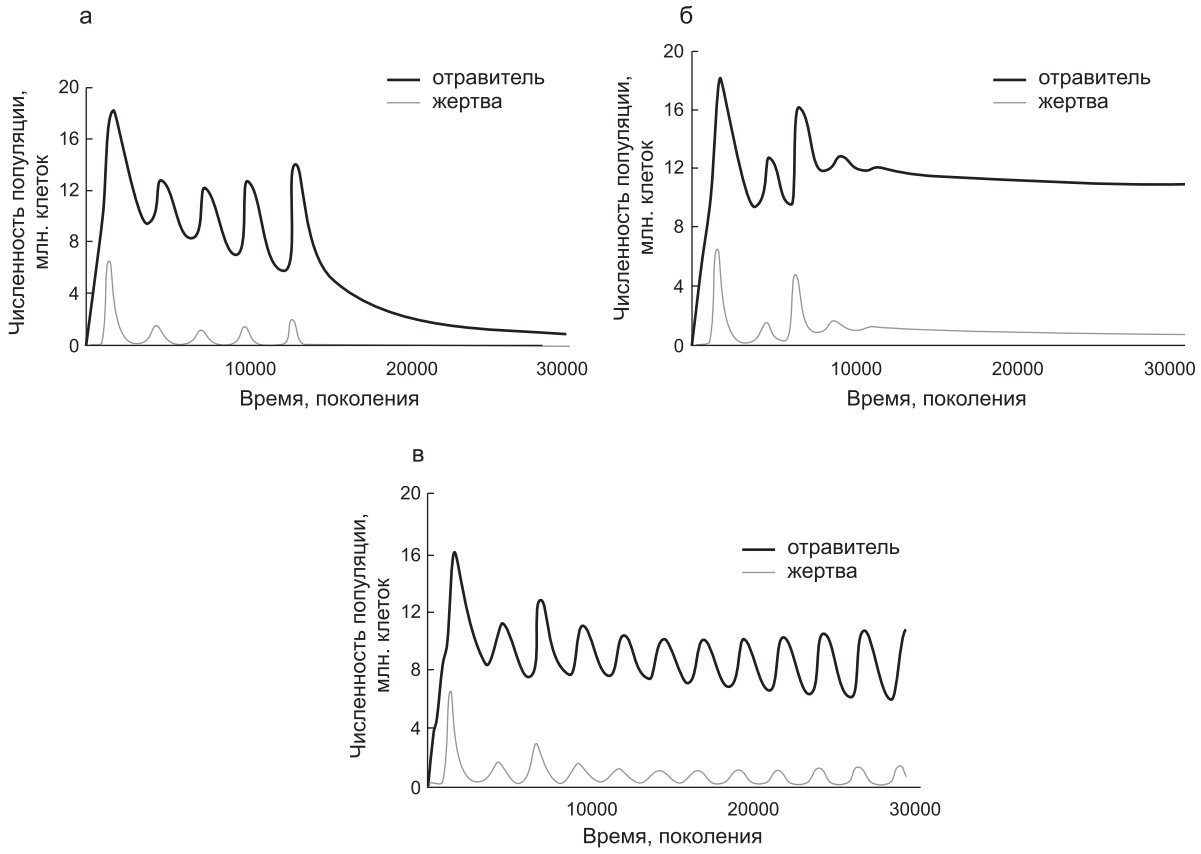

Рис. 6. Моделирование динамики системы «отравитель–жертва» при $c_1 = 6$, $c_2 = 2$.

а – без мутаций; б – под действием мутации, произошедшей на 5000-м поколении, изменившей значение c_2 на 1; в – под действием мутации, произошедшей на 5000-м поколении, изменившей значение c_2 на 1,5.

3) приводить систему, обреченную на гибель, к незатухающим колебаниям (рис. 6, а, в).

При этом характер динамики популяций во всех точечных средах повторяет общий характер изменений в первой точечной среде с некоторым сглаживанием колебаний; сглаженность графика увеличивается с увеличением номера среды.

Тем самым показано, что система является крайне нестабильной, немаловажную роль играет изначальное размещение популяций в среде: нахождение в одной точечной среде более выигрышно для популяций, чем попадание в разные среды, положительную роль играет близость к источнику вещества. Рассмотрены мутации, изменяющие гены утилизации специфического субстрата у жертвы или отравителя, и их распространение в среде. Показано, что мутация, изменившая значение одного гена в одной точечной среде, способна изменить сценарий развития системы «отравитель–жертва» и характер динамики численности сообщества.

БЛАГОДАРНОСТИ

Работа была поддержана грантами: РФФИ 10-04-01310; Научная школа-5278.2012.4; Программа РАН № 28; Междисциплинарный интеграционный проект СО РАН № 87.

ЛИТЕРАТУРА

Бандман О.Л. Клеточно-автоматные модели пространственной динамики // Системная информатика. 2006. Вып. 10. С. 59–113.

Левченко В.Ф. Эволюция биосферы до и после появления человека. СПб: Ин-т эволюционной физиологии и биохимии РАН, 2003. 164 с.

Разумовский С.М. Закономерности динамики биоценозов. М.: Наука, 1981. 231 с.

Раутиан А.С., Жерихин В.В. Модели филоценогенеза и уроки экологических кризисов геологического прошлого // Журн. общ. биологии. 1997. Т. 58. Вып. 4. С. 20–47.

Ризниченко Г.Ю. Математические модели в биофизике и экологии. М.; Ижевск: Институт компьютерных исследований, 2003. 184 с.

Хлопков Ю.И. Статистическое моделирование в вычислительной аэродинамике. М.: ООО «Азбука-2000», 2006. 158 с.

Bird G.A. Shock-wave structure in rigid sphere gas // Rarefied Gas Dynamics. 1965. V. 1. P. 216–222.

Haviland J.K., Lavin M.D. Application of the Monte-Karlo method to heat hauster in rarefied of gases // Phys. Fluids. 1962. V. 5. No. 11. P. 1399–1405.

Herbert D. Recent progress in microbiology. Oxford: Blackwell, 1958. P. 381–416.

Lashin S.A., Matushkin Yu.G. Haploid evolutionary constructor: new features and further challenges // In Silico. Biol. 2012. V. 11. No. 3. P. 125–135.

Lashin S.A., Suslov V.V., Matushkin Yu.G. Comparative modeling of coevolution in communities of unicellular organisms: adaptability and biodiversity // J. Bioinform. and Computat. Biol. 2010. V. 8. No. 3. P. 627–643.

Monod J. Recherches sur la croissanse des cultures bacteriennes. P.: Hermann, 1942.

Rübel A.E. Ecology, plant geography, and geobotany: their history and aim // Bot. Gaz. 1927. V. 84. No. 4. P. 428–439.

SPATIALLY DISTRIBUTED MODELING OF PROKARYOTIC COMMUNITY EVOLUTION

S.A. Lashin[1, 2], E.A. Mamontova[2], Yu.G. Matushkin[1, 2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;
[2] Novosibirsk National Research State University, Novosibirsk, Russia,
e-mail: lashin@bionet.nsc.ru

Summary

This paper describes the development of an approach to the simulation of prokaryotic community activity and evolution and the software package «Haploid evolutionary constructor» (http://evol-constructor.bionet.nsc.ru). The initial model with ideal mixing (0D) is expanded to a spatially distributed model (1D). The 0D and 1D poisoner–prey prokaryotic community models are compared. It is shown that the community stability is influenced by the spatial distribution of substrates and prokaryotic cells.

Key words: evolution modeling, bacterial community, spatial distribution, diffusion.

УДК 004.65

ИНФОРМАЦИОННЫЙ ПОРТАЛ «БИОТЕХНОЛОГИЯ РАСТЕНИЙ» – ИНТЕРНЕТ-РЕСУРС ДЛЯ ПОДДЕРЖКИ ЭКСПЕРИМЕНТОВ В ОБЛАСТИ ГЕННОЙ ИНЖЕНЕРИИ РАСТЕНИЙ, ГЕНЕТИКИ И СЕЛЕКЦИИ ПШЕНИЦЫ

© 2012 г. **А.В. Кочетов, О.Г. Смирнова, С.М. Ибрагимова, Д.А. Рассказов, Д.А. Афонников, М.А. Генаев, А.В. Дорошков, Т.А. Пшеничникова, А.В. Симонов, Е.В. Морозова**

Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения РАН, Новосибирск, Россия, e-mail: ak@bionet.nsc.ru

Поступила в редакцию 15 июля 2012 г. Принята к публикации 31 августа 2012 г.

Разработан Интернет-портал для информационной поддержки проведения НИР в области биотехнологии растений, содержащий специализированные модули (базы данных – БД) и программные компоненты), включая БД внешних информационных ресурсов, БД промоторов для трансгенеза растений, БД трансляционных энхансеров для трансгенеза растений, БД WheatPGE для поддержки экспериментов в области генной инженерии растений и селекционно-генетических экспериментов на пшенице. Модульная структура позволяет использовать этот ресурс в качестве платформы, к которой могут добавляться новые информационные и программные компоненты. Интернет-ресурс доступен на сайте ИЦиГ СО РАН (http://bioagrotech.bionet.nsc.ru/).

Ключевые слова: информационный ресурс, база данных, промотор, трансляционный энхансер, генная инженерия, пшеница, фенотипирование, селекция.

ВВЕДЕНИЕ

Биотехнология относится к приоритетным направлениям развития современной науки. В РФ для координации НИОКР в этой области была образована технологическая платформа (ТП) «Биоиндустрия и биоресурсы – БиоТех2030», которая является формой реализации института частно-государственного партнерства и инструментом осуществления научно-технической и инновационной политики в области биотехнологий (http://www.biotech2030.ru/). В сферу компетенции ТП входит широкий круг задач, связанных с биоэнергетикой, источниками возобновляемого сырья, разработкой технологий биопродукции новых материалов, биокатализа, агробиотехнологий и т. д. Для решения этих задач должны применяться современные методы общей и молекулярной биологии, генетики, генной, хромосомной и клеточной инженерии. Эффективное проведение НИР и ОКР на современном уровне в настолько широком спектре задач невозможно без адекватного информационного обеспечения.

Анализ подходов к информационной поддержке НИОКР в генной инженерии растений, селекционной генетике и биотехнологии

Генная инженерия растений используется для получения генетически модифицированных организмов (ГМО), обладающих новыми промышленно ценными свойствами. Генетические модификации позволяют получать штаммы бактерий и грибов – продуцентов ферментов, аминокислот, биологически активных веществ; также широко используются трансгенные растения. С помощью ГМО развиваются методы получения биотоплива (Ageitos *et al*., 2011; Wiley *et al*., 2011), разрабатываются технологии оптимизации трансгенеза (Abdeev *et al*., 2009)

и наработки биопрепаратов в трансгенных растений (molecular pharming) (Komarova *et al*., 2010; Hassan *et al*., 2011), наработки вторичных метаболитов оптимизированными культурами тканей (Bulgakov *et al*., 2011).

В агробиологии увеличение стрессоустойчивости и устойчивости растений к фитопатогенам и, как считают, использование методов системной биологии может привести к быстрому прогрессу (Pritchard, Birch, 2011). Получение стрессоустойчивых растений за счет комбинации классической селекции и генной инженерии (molecular breeding) позволит расширить площади их выращивания и решить продовольственную проблему (Varshney *et al*., 2011). Обсуждаются перспективы использования ГМ растений в сельском хозяйстве – рассматривается возможность 40–80 % увеличения продуктивности при их использовании (Skryabin, 2010).

Одним из перспективных подходов, лежащих в основе получения новых сортов в генетике и селекции растений, считается картирование локусов, отвечающих за хозяйственно ценные признаки (Kumar *et al*., 2010). Современный селекционно-генетический эксперимент использует данные о тысячах и десятках тысяч растений (Ajjawi *et al*., 2010; Brachi *et al*., 2010). Очевидно, что для выборок такого размера традиционные способы определения большинства фенотипических характеристик малоэффективны. Для повышения эффективности решения указанных выше задач в последнее время в мире все более интенсивно используются информационные и телекоммуникационные технологии.

Подход для поддержки генетических коллекций, основанный на информационной поддержке селекционно-генетических экспериментов у растений, предложен в базе данных Germinate (Lee *et al*., 2005). Интересным проектом является база данных Sol Genomic Network (http://solgenomics.net/), которая содержит информацию о фенотипе, генотипе, полногеномных данных и генных и метаболических сетях для растений семейства пасленовых. Эта база данных позволяет производить поиск фенотипа, результатов анализа количественных признаков (QTL), списка маркеров, генов, метаболических сетей. Она тесно интегрирована с геномными данными.

Следует отметить, что большинство этих систем используют поддержку хранения информации о фенотипе в виде изображений. Кроме того, методы анализа изображений используются при анализе фенотипа растения все более интенсивно (Eberius, Lima-Guerra, 2009). Наиболее перспективный подход – создание систем для поддержки лабораторных экспериментов. Высокопроизводительное фенотипирование, эффективный сбор, хранение большого объема данных, их интеграция с геномными данными позволили создать прорывную технологию анализа взаимосвязи между генотипом и фенотипом у *Arabidopsis thaliana* (Lu *et al*., 2008). Однако данная система не позволяет учитывать влияние окружающей среды на развитие фенотипа растения. Другим интересным проектом является система PHENOME для сбора, хранения и анализа данных о фенотипе у томата (Vankadavath, 2009).

Одним из подходов, позволяющих существенно ускорить фенотипирование, является использование анализа цифровых изображений. Например, они были успешно применены для оценки биомассы растения (Hartmann *et al*., 2011; Golzarian *et al*., 2011), анализа морфологии и развития корня у риса (Iyer-Pascuzzi *et al*., 2010), анализа морфологии опушения листа (Kaminuma *et al*., 2008).

Использование мобильных устройств позволяет существенно повысить эффективность решения задач в области селекционно-генетических экспериментов, особенно для полевых наблюдений. Например, в информационной системе PHENOME (Vankadavath, 2009) для сбора информации о признаках растений используются карманные компьютеры (PDA). Это позволяет собирать большое количество данных в полевых условиях (результаты измерений анатомических признаков растений, их плодов, устойчивости к заболеваниям). Затем эта информация заносится в центральную базу данных.

В мире активно развивается специализированная информационная поддержка биотехнологических исследований. Помимо специфических (узконаправленных) разработок, существуют научно-исследовательские институты (государственные и частные), оказывающие услуги по информационной поддержке и

проработке проектов (Biotechnology Information Institute, http://www.bioinfo.com/; Information System for Biotechnology, http://www.isb.vt.edu/ (правительство США); Bioinformatics Information System Network, http://www.btisnet.gov.in/index.asp (правительство Индии); ArgosBiotech, http://www.argosbiotech.de/ – компания, обеспечивающая информационный портал и маркетинг для бизнеса и научного сообщества в области биотехнологий.

Однако специализированные ресурсы для поддержки биотехнологических экспериментов в открытом доступе отсутствуют. В качестве примера можно рассмотреть планирование НИР с применением методов генной инженерии растений. Получение технологически эффективных ГМО требует грамотного планирования, которое может включать следующие этапы (рассматривается ситуация создания биопродуцента технологически значимого белка или вторичного метаболита):

Выбор организма для создания биопродуцента. Выбор организма в ряде случаев задан изначально, если речь идет о специфическом вторичном метаболите, характерном для определенного организма. Однако в некоторых случаях требуется создание продуцента белка (фермента или фармакологического препарата), для чего могут быть использованы различные подходы. Разные биопродуценты имеют свои преимущества и недостатки, среди которых следует выделить степень близости продукта к природному варианту (для белков важны посттрансляционные модификации), отсутствие токсичных примесей, сложность выделения и очистки, количественные характеристики синтеза и себестоимость продукции. Например, белки человека можно производить с помощью культур соответствующих клеток, и полученный продукт будет практически идентичен натуральному. Однако стоимость такой продукции будет высокой вследствие требований к стерильности, отсутствию в культуре вирусов или прионов, а также из-за низкого выхода. С другой стороны, культуры микроорганизмов могут давать высокий уровень биопродукции, однако полученный белок может характеризоваться конформацией, отличной от природного варианта, в частности, это касается посттрансляционных модификаций. Выбор организма-биопродуцента в данном случае зависит от экспертной оценки его особенностей, определяющих преимущества и недостатки в рамках решения конкретной технологической задачи.

Выбор генов-мишеней. При создании ГМО глубина собственно модификаций может варьировать и зависит от поставленной задачи. Существующие технологии позволяют как усиливать экспрессию определенного белка (за счет внесения трансгена в геном), так и снижать или выключать экспрессию (например, с помощью использования нокаутных штаммов, генетического сайленсинга или РНК-интерференции). Процесс выбора прост в тех случаях, когда он задан изначально и планируется получение ГМО, которые несут один трансген, например, при биопродукции фармакологически значимого (чужеродного) белка или применяемого в биотехнологическом производстве фермента. Однако может ставиться более сложная задача получения вторичного метаболита либо ГМО должен характеризоваться дополнительными параметрами (например присутствием дополнительных специфических белков, участвующих в процессинге основного продукта). В качестве примера можно привести разработки, в которых для повышения выхода определенного метаболита у ГМО выключают метаболические цепи, конкурирующие за субстрат или интермедиаты. В этих случаях необходимы моделирование биохимических контуров и расчет параметров, обеспечивающих оптимальный уровень синтеза.

Выбор векторной системы. Выбор векторной системы обычно определяется спецификой поставленной технологической задачи и организмом-реципиентом генетической конструкции, с помощью которого будет реализовываться проект. Этот выбор весьма разнообразен (вирусные векторы, плазмиды разных типов, интегрирующиеся в геном конструкты и т. п.).

Дизайн генетической конструкции. Этот этап включает выбор адекватного промотора, при необходимости – подбор энхансера трансляции и поли(А)-сигнала, оптимизацию кодонного состава. Также необходимо удостовериться в отсутствии ложных сигналов экспрессии: поскольку ДНК трансгена часто относится к организму другой таксономической принадлежности, она характеризуется нуклеотидным составом, отличным от геномной ДНК организ-

ма-хозяина (например, ДНК млекопитающих обогащена G+C в сравнении с двудольными растениями). Это, в свою очередь, может привести к образованию комбинаций нуклеотидов, которые будут распознаваться в клетках организма-хозяина как сигналы экспрессии.

Выбор метода трансгенеза. В большинстве случаев метод трансгенеза определяется организмом, использованным для получения ГМО, и особенностями векторной системы. Однако в некоторых случаях существует выбор: так, трансгенные растения можно получать с помощью агробактериальной трансформации, бомбардировки частицами с сорбированной на них ДНК, трансфекции протопластов с помощью электропорации или ПЭГ и т. п. Каждый из методов имеет преимущества и недостатки: например, получение трансгенных растений с помощью агробактериальной трансформации позволяет получать ГМО с одной инсерцией чужеродной ДНК (это важно для предотвращения генетического сайленсинга), но обычно связано с необходимостью регенерации растений из каллусов (успех которой связан с особенностями генотипа и часто требует трудоемкого подбора условий для индукции морфогенеза). С другой стороны, бомбардировка микрочастицами с сорбированной на них ДНК (particle bombardment) часто позволяет регенерировать трансгенные растения с меньшими сложностями, но при этом в геном обычно встраивается большое число копий ДНК трансгена.

Выбор метода культивирования ГМО. Этот этап связан с особенностями как ГМО, так и технологического процесса. Например, трансгенные растения можно выращивать в теплице, но также можно и культивировать *in vitro* в виде культуры каллусов или корней (например, при трансгенезе с помощью *Agrobacterium rhizogenes*).

Выбор системы очистки продукта. Этот этап практически полностью определяется особенностями собственно продукта.

Таким образом, как можно видеть, процедура планирования НИР представляет собой сложный процесс, требующий работы с большим объемом постоянно обновляющихся литературных данных. Эффективность такой работы может быть существенно увеличена при наличии соответствующих информационных ресурсов (экспертных систем). В открытом доступе нет ресурсов, позволяющих решать такие задачи в комплексе.

В статье представлен информационный ресурс, содержащий специализированные модули для решения следующих задач:

– дизайн генетической конструкции для получения трансгенных растений (выбор промотора и трансляционного энхансера);

– сбор, хранение и анализ данных о фенотипических параметрах растений пшеницы в селекционного-генетическом эксперименте. Интернет-портал «Биотехнология растений» представляет собой информационный ресурс модульного типа, т. е. к этой платформе могут добавляться новые компоненты, предназначенные для решения других задач.

Структура информационного портала «Биотехнология растений»

Информационный портал «Биотехнология растений» (ИП БР) доступен по адресу http://bioagrotech.bionet.nsc.ru/ и включает:

1) базу данных промоторов (БДП), содержащую 289 учетных записей;

2) базу данных трансляционных энхансеров (БДТЭ), содержащую 58 учетных записей;

3) базу внешних информационных ресурсов (БВИР), содержащую 15 записей;

4) базу данных WheatPGE (БWPGE), содержащую 30 информационных полей для описания фенотипических признаков растения пшеницы и 5 информационных полей для описания мест произрастания растений пшеницы;

5) модуль интерфейса внешнего уровня (МИВУ), обеспечивающий через Интернет доступ к информационным ресурсам портала и внешних Web-источников, навигацию по ИП БР;

6) модуль интерфейса для базы данных промоторов (МИБДП), обеспечивающий взаимодействие пользователя с БДП;

7) модуль интерфейса для базы данных трансляционных энхансеров (МИБТЭ), обеспечивающий взаимодействие пользователя с БДТЭ;

8) модуль интерфейса для базы внешних информационных ресурсов (МИБВИР), обеспечивающий взаимодействие пользователя с БВИР;

9) модуль интерфейса для базы WheatPGE (МИБWPGE), обеспечивающий взаимодействие пользователя с БWPGE.

Модуль интерфейса верхнего уровня (МИВУ) приведен на рис. 1. Он обеспечивает доступ к БД внешних информационных ресурсов, к специализированным БД промоторов и трансляционных энхансеров для экспериментов с трансгенными растениями, а также для информационной поддержки селекционно-генетических экспериментов на пшенице.

Базы данных БДВИР, БДП и БДТЭ разработаны на платформе Sequence Retrieval Systems 6.1, которая развернута на сервере баз данных под управлением Red Hat Enterprise Linux 5.7.

Информационно-поисковая система SRS позволяет осуществлять автоматическую индексацию поисковых полей, что дает возможность пользователю применять различные комбинации запросов для гибкого поиска.

База данных БВИР предоставляет возможность для поиска внешнего информационного ресурса, который может быть использован для решения конкретной задачи, стоящей перед пользователем. Формат этой неспециализированной БД максимально прост: одна таблица включает четыре поля (идентификатор, адрес, комментарий и ключевые слова). Поиск по ключевым словам позволяет отобрать потенциально подходящие варианты ресурсов, после чего дополнительная информация может быть получена из комментария, поле «address» содержит ссылку непосредственно на описываемый ресурс. Пример поиска по полю «ключевые слова» (запрос «вектор») приведен на рис. 2.

На рис. 3 в качестве примера показан скриншот результатов выполнения этого запроса к базе данных БВИР – результат поиска по данному запросу.

База данных промоторов (БДП), детально описанная в работах Смирновой с соавт. (Smirnova *et al*., 2012; Смирнова и др., 2012а), содержит три типа взаимосвязанных таблиц: описание гена, промоторных участков, нуклеотидных последовательностей промоторов. Особенность этой БД заключается в том, что в ней содержится информация о транскрипционной активности делеционных вариантов промоторов, определенная в экспериментах с трансгенными растениями. Такие эксперименты используются для выявления структуры промоторов генов растений и их транскрипционного контроля. БДП позволяет применить эти данные для планирования опытов в области биотехнологии и генной инженерии растений. Текущий выпуск базы содержит информацию

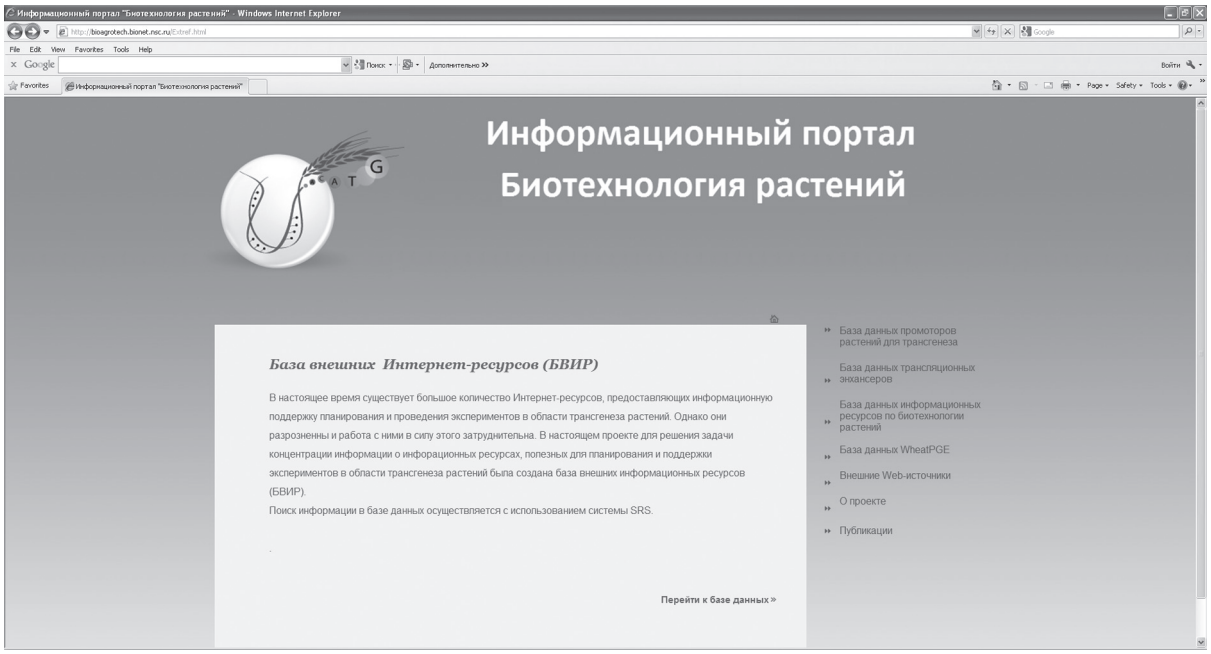

Рис. 1. Модуль интерфейса верхнего уровня (МИВУ) информационного портала «Биотехнология растений». Страница доступа к БВИР.

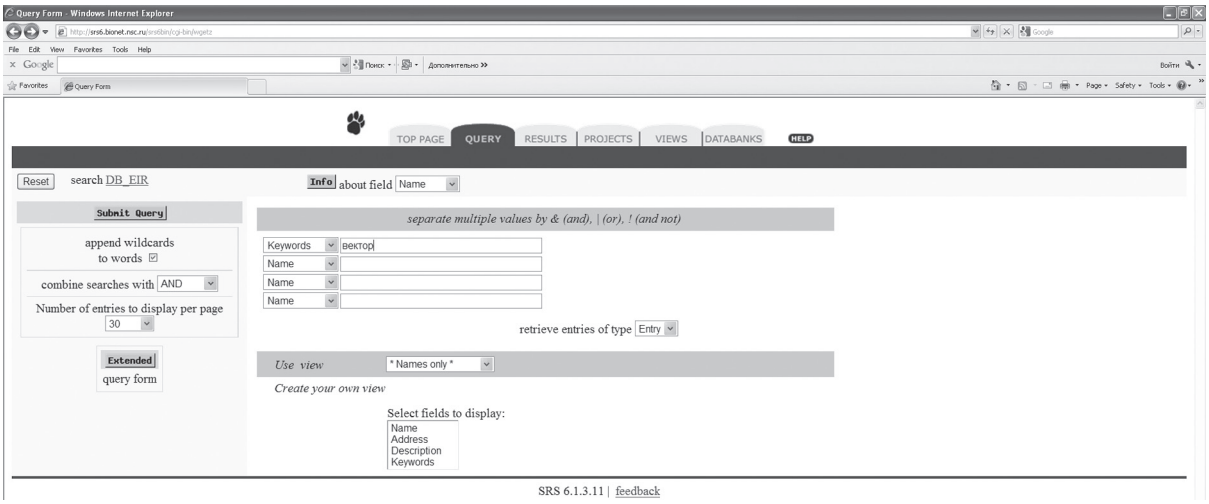

Рис. 2. Страница формы для составления запроса к БВИР (типовая форма системы SRS). В качестве примера использован поисковый запрос «вектор».

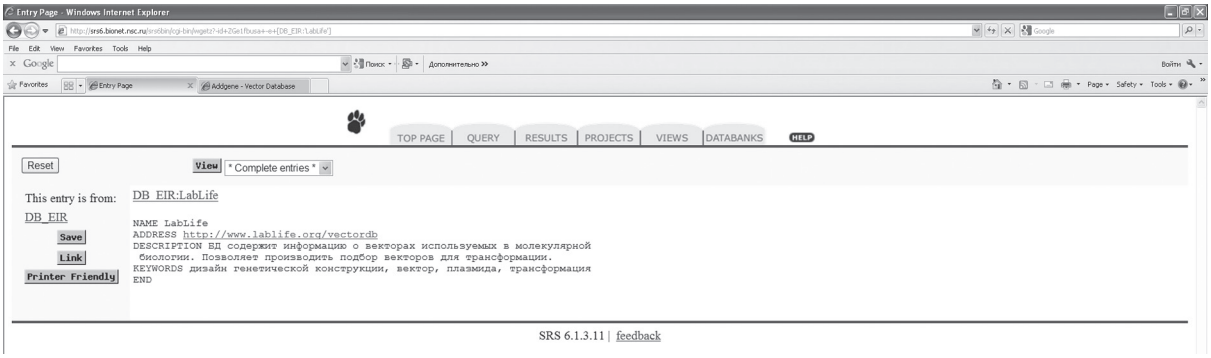

Рис. 3. Страница вывода результата поиска в БВИР: пример записи БВИР, в ключевых словах которой содержится слово «вектор». Приведены краткое описание (DESCRIPTION) и ссылка на информационный ресурс (ADDRESS).

о 289 промоторах, 289 нуклеотидных последовательностях и 158 генах. Представленные промоторы принадлежат 27 видам растений. Их активность описана более чем в 40 различных органах и тканях трансгенных растений. Список видов трансгенных растений, использованных для изучения активности промоторов, включает 33 наименования. Описано действие более 40 видов регуляторов на активность промоторов.

Типичные примеры запросов для БД TGP:

– найти промоторы, работающие в конкретном виде растений (поиск в поле Target species);

– найти промоторы, на которые влияет конкретный регулятор (поиск в поле Regulator);

– найти промоторы, работающие в конкретном виде растений и на которые влияет конкретный регулятор (поиск в полях Target species и Keywords и/или Regulator);

– найти промоторы, выделенные из определенного вида растений (поиск в поле PromoterID);

– найти промоторы, на которые влияют несколько различных регуляторов;

– найти промоторы, активные в определенном органе или ткани (поиск в полях STAGE_ORGAN_TISSUE и/или COMMENT);

– найти промоторы, которые активны в определенном органе или ткани (поиск в полях REGULATOR и STAGE_ORGAN_TISSUE или COMMENT).

База данных БДТЭ, детально описанная в работе Смирновой с соавт. (2012б), содержит структурированную информацию о локализованных в мРНК регуляторных сигналах, кото-

рые контролируют экспрессию генов на посттранскрипционном уровне. Эта информация полезна для планирования генно-инженерных экспериментов, поскольку трансляционные энхансеры нельзя заменить другими функциональными элементами в структуре генетической конструкции.

База WheatPGE (Генаев, 2011, 2012; Генаев и др., 2012а, б) ориентирована на изучение взаимоотношений фенотип–генотип–окружающая среда у пшеницы и предназначена для обеспечения проведения высокопроизводительного фенотипирования в ходе селекционно-генетических экспериментов.

Логическая модель данных включает таблицу растения, связанную с 4 блоками информации – генотипом, фенотипом, местом произрастания и экспериментом. Всего в текущей версии БД содержит 32 таблицы и 55 отношений между ними.

Генотип растения описывается 9 таблицами, включающими информацию о сорте растения или линии. Генотип связан с рядом таблиц, описывающих генетические маркеры. Такая привязка позволяет документировать эксперименты на пшенице, которые направлены на выявление мест локализации генов, контролирующих фенотипические признаки пшеницы, на хромосомных картах.

Фенотип растения описывается 15 таблицами: таблица, описывающая базовые признаки растения (длина стебля, число колосьев, урожайность); таблицы, описывающие структуру урожая (колосья); таблицы описывающие характеристики листьев; таблицы, описывающие опушение листа; таблица, описывающая длительность стадий развития растения.

Место произрастания описано 2 таблицами, основная информация в которых содержит название места произрастания, широту, долготу, тип климата, климатические характеристики (среднегодовую температуру, среднегодовую влажность, средние температуры января и июля).

Блок информации, связанный с проведением эксперимента, содержит 2 таблицы, описывающие событие и их список для растения. Событие содержит поля названия, типа и значения.

Функции интеграции базы данных и различных методов массового фенотипирования растений выполняет модуль интерфейса, обеспечивающий взаимодействие пользователя с БWPGE. Нами использована методология разработки программного обеспечения MVC, поддержку которой обеспечивает Catalyst – свободный кроссплатформенный программный каркас для создания Web-приложений, написанных на языке Perl.

МИБWPGE обеспечивает работу с базой через мобильные устройства (планшетные компьютеры и смартфоны), для которых доступ в Интернет сейчас возможен практически из любой точки страны. Это обеспечивает ввод данных в БWPGE в полевых условиях, что позволяет существенно ускорить процесс фенотипирования растений. Для удобства взаимодействия с базой данных с мобильных устройств для идентификации растений мы используем систему QR кодов, которые являются матричными штрих-кодами и могут быть сканированы камерой мобильного устройства. QR-код присваивается в базе каждому растению, может быть распечатан на плотной бумаге и прикреплен к его стеблю. В дальнейшем при измерении параметров растения в процессе эксперимента достаточно считать этот код, открыть в браузере ссылку для этого растения и занести параметры в базу.

Решение задач в области биотехнологий с использованием ресурсов информационного портала «Биотехнология растений»

Рассмотрим несколько примеров решения задач, связанных с устойчивостью растений пшеницы (*Triticum aestivum*) к засухе, с использованием ресурсов информационного портала «Биотехнология растений». Одним из возможных решений может являться создание трансгенных растений, которые несут целевые гены под управлением промоторов, активизирующихся в ответ на засушливые условия среды. Для поиска таких промоторов можно использовать базу данных промоторов (БДП). Для этого требуется отобрать гены, которые активизируются у пшеницы в условиях засухи, т. е. содержат ключевое слово «drought-induced» и не индуцируются, например, в условиях солевого стресса. Для этого в базу данных можно внести запрос по таблице генов на странице

«Results» в поле расширенного запроса «(([tgp_gene-Species:wheat*] & [tgp_gene-Keywords:drought-induced*]) ! ([tgp_gene-Keywords:salt-induced*]))» (рис. 4, а) и нажать кнопку «Expression». В результате запроса будут получены 4 записи, одна из которых (TGP_GENE:Ta:Ltp1) приведена на рис. 4, б.

Последовательности этих промоторов можно найти по ссылкам, приведенным в записи (SEQUENCE_ID Ta:Ltp1_P1S).

Обеспечить высокий уровень наработки целевого белка в условиях засухи можно не только за счет регуляции транскрипции, но и усилив уровень трансляции его мРНК. Для этого можно подобрать фрагменты вставок в последовательность гена, которые бы усиливали процесс трансляции его мРНК. Такие фрагменты можно найти с помощью базы БДТЭ. Для этого необходимо на странице запроса по таблице объектов БДТЭ (рис. 5, а) ввести название организма («Triticum aestivum», пшеница) и нужный тип локализации энхансера, например, «5′UTR» (5′ нетранслируемый район).

По данному запросу в БДТЭ получено 10 записей. Пример одной из записей, TRANSIG_OBJ:ART5ENH03, приведен на рис. 5, б. В записи указаны последовательность энхансера («tagatatccgcgctt»), а также его краткое описание.

Отметим, что одним из важных фенотипических признаков у пшеницы, который проявляется в связи с ответом на стресс в условиях засухи, является опушение листьев. Опушение покровов растения яровой мягкой пшеницы влияет на их влагоудерживающую способность (Лихенко, 2007). Сильное, «войлочное»,

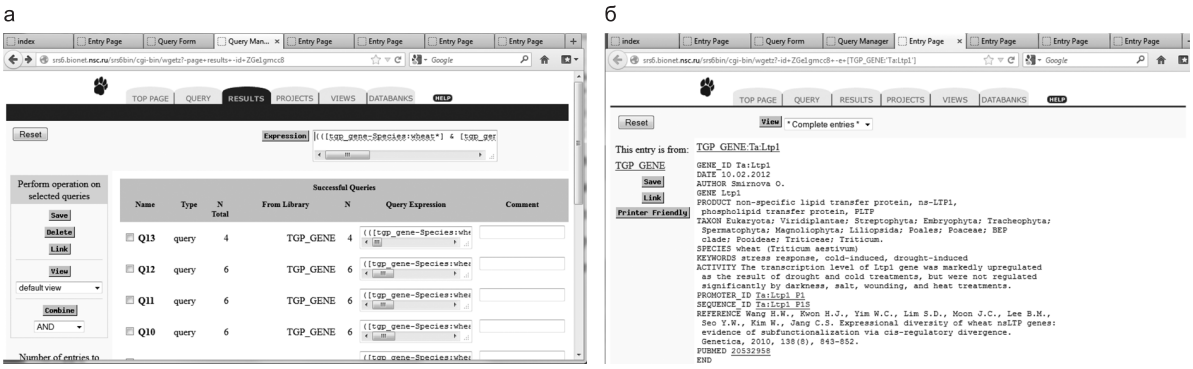

Рис. 4. Поиск в БДП промоторов генов, имеющих повышенную экспрессию в ответ на засушливые условия среды.

а – ввод запроса; б – пример записи TGP_GENE:Ta:Ltp1 для гена, имеющего высокий уровень экспрессии в условиях засухи.

Рис. 5. Поиск в БДТЭ трансляционных энхансеров, расположенных в 5′-нетранслируемом районе генов пшеницы.

а – ввод запроса; б – пример записи TRANSIG_OBJ:ART5ENH03 для трансляционного энхансера.

«мохнатое» опушение характерно для ряда засухоустойчивых сортов, относящихся к степной экологической группе; для сортов, произрастающих во влажном климате, напротив, характерно очень слабое опушение (Крупнов, Цапайкин, 1990). Поэтому анализ наследования количественных характеристик опушения листа у пшеницы поможет в идентификации генов, контролирующих этот признак. Это сделает возможным целенаправленное создание линий и сортов пшеницы с повышенной плотностью опушения, которое может обеспечить дополнительную устойчивость растений к засухе. Для проведения подобных работ можно использовать систему BWPGE в составе ИП БР. С ее помощью можно исследовать количественные характеристики опушения листа пшеницы у родительских сортов и линий, а также

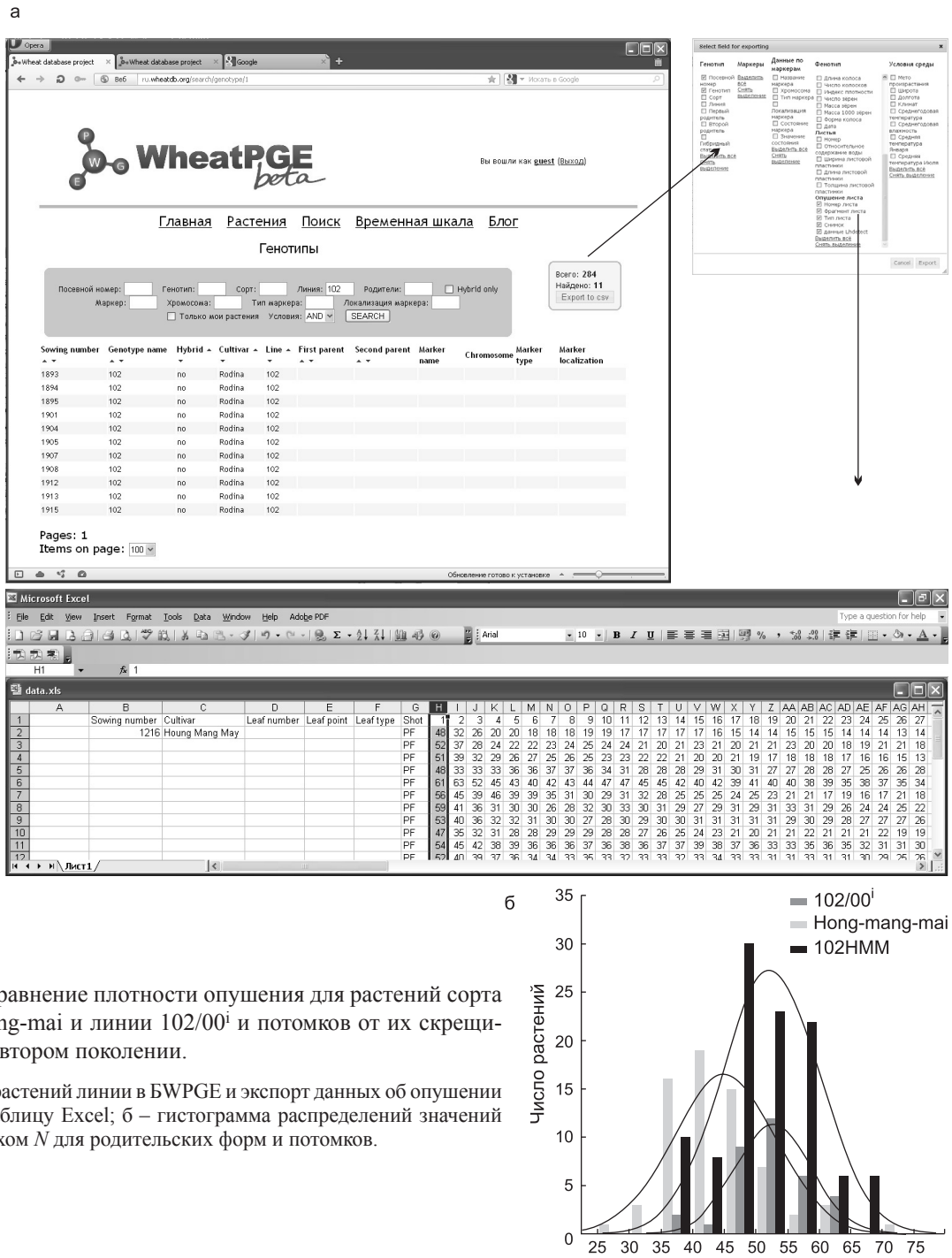

Рис. 6. Сравнение плотности опушения для растений сорта Hong-mang-mai и линии 102/00^{i} и потомков от их скрещивания во втором поколении.

а – поиск растений линии в BWPGE и экспорт данных об опушении листа в таблицу Excel; б – гистограмма распределений значений числа трихом *N* для родительских форм и потомков.

у потомков от их скрещивания. Оценка числа трихом на листовой пластинке для растений проводится при помощи технологии высокопроизводительного фенотипирования на основе анализа изображений (Genaev *et al*., 2012).

В качестве примера применения БWPGE для решения подобных задач рассмотрим сравнение числа трихом на поверхности листовой пластинки у родительских форм и потомков от их скрещивания в поколении F_2 на примере сорта Hong-mang-mai и линии 102/00^i. Для этого необходимо провести поиск растений указанных сортов в БWPGE. В результате выполнения этой операции отображается список записей растений (рис. 6, а). Для этих растений информацию об опушении листа необходимо экспортировать в формате CSV и загрузить в таблицу Excel. Аналогичную процедуру необходимо провести и для потомков от скрещивания Hong-mang-mai и линии 102/00^i во втором поколении.

Из сопоставления распределений числа трихом у родительских растений и потомков видно (рис. 6, б), что опушение растений во втором поколении от скрещивания сортов Hong-mang-mai с линией 102/00^i более интенсивно по сравнению с родительскими генотипами. Таким образом, разработанный нами подход к анализу можно использовать для создания гибридов пшеницы, имеющих повышенную плотность опушения, что может обеспечить им устойчивость в стрессовых условиях среды, в частности при засухе.

ЗАКЛЮЧЕНИЕ

Разработан информационный Интернет-портал «Биотехнология растений», который в настоящее время может использоваться для решения ряда задач в областях генной инженерии растений и планирования селекционно-генетических экспериментов на пшенице. Модульная структура ресурса позволяет рассматривать его в качестве прототипа платформы, на которой могут интегрироваться новые специализированные модули, направленные на решение конкретных задач и востребованные специалистами в области биотехнологии.

Работа поддержана грантом Министерства образования и науки РФ в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 гг.» (07.514.11.4052).

ЛИТЕРАТУРА

Генаев, М.А., Дорошков А.В., Морозова Е.В. и др. Компьютерная система WheatPGE для анализа взаимосвязи фенотип–генотип–окружающая среда у пшеницы // Вавилов. журн. генет. и селекции. 2011. Т. 15. С. 784–793.

Генаев М.А., Дорошков А.В., Пшеничникова Т.А. и др. Информационная поддержка селекционно-генетического эксперимента у пшеницы в системе WheatPGE // Матем. биология и биоинформатика. 2012. Т. 7. № 2. С. 410–424.

Крупнов В.А., Цапайкин А.П. Опушение листьев пшеницы: генетические и экологические аспекты // С.-х. биология. Сер. «биология растений». 1990. № 1. С. 51–57.

Лихенко И.Е. О взаимосвязи опушения органов растений яровой мягкой пшеницы с хозяйственно и биологически ценными признаками в условиях Западной Сибири // Растениеводство и селекция. 2007. № 6. С. 25–31.

Смирнова О.Г., Рассказов Д.А., Афонников Д.А., Кочетов А.В. TGP – база данных промоторов для трансгенеза растений // Матем. биология и биоинформатика. 2012а. Т. 7. № 2. С. 444–460.

Смирнова О.Г., Рассказов Д.А., Кочетов А.В. Информационная поддержка экспериментов по трансгенезу растений: база данных трансляционных энхансеров // Вавилов. журн. генет. и селекции. 2012б. Т. 16. № 4/1. С. 766–773.

Abdeev R.M., Abdeeva I.A., Bruskin S.S. *et al*. Bacterial thermostable beta-glucanases as a tool for plant functional genomics // Gene. 2009. V. 436. P. 81–89.

Ageitos J.M., Vallejo J.A., Veiga-Crespo P., Villa T.G. Oily yeasts as oleaginous cell factories // Appl. Microbiol. Biotechnol. 2011. V. 90. P. 1219–1227.

Ajjawi I., Lu Y., Savage L.J. *et al*. Large-scale reverse genetics in Arabidopsis: case studies from the Chloroplast 2010 Project // Plant Physiol. 2010. V. 152. P. 529–540.

Brachi B., Faure N., Horton M. *et al*. Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature // PLoS Genet. 2010. V. 6. P. e1000940.

Bulgakov V.P., Inyushkina Y.V., Fedoreyev S.A. Rosmarinic acid and its derivatives: biotechnology and applications // Crit. Rev. Biotechnol. 2011.

Eberius M., Lima-Guerra J. High-Throughput Plant Phenotyping – Data Acquisition, Transformation, and Analysis // Bioinformatics: Tools and Applications / Ed. D. Edwards *et al*. Springer Science+Business Media, LLC, 2009. P. 259–278.

Genaev M.A., Doroshkov A.V., Pshenichnikova T.A. *et al*. Extraction of quantitative characteristics describing wheat leaf pubescence with a novel image processing technique // Planta. 2012. In press.

Golzarian M.R., Frick R.A., Rajendran K. *et al*. Accurate inference of shoot biomass from high-throughput images of cereal plants // Plant Methods. 2011. V. 7. 2.

Hartmann A., Czauderna T., Hoffmann R. *et al*. HTPheno: an image analysis pipeline for high-throughput plant phenotyping // BMC Bioinformatics. 2011. V. 12. P. 148.

Hassan S.W., Waheed M.T., Lössl A.G. New areas of plant-made pharmaceuticals // Expert Rev. Vaccines. 2011. V. 10. P. 151–153.

Iyer-Pascuzzi A.S., Symonova O., Mileyko Y. *et al*. Imaging and analysis platform for automatic phenotyping and trait ranking of plant root systems // Plant Physiol. 2010. V. 152. P. 1148–1157.

Kaminuma E., Yoshizumi T., Wada T. *et al*. Quantitative analysis of heterogeneous spatial distribution of Arabidopsis leaf trichomes using micro X-ray computed tomography // Plant J. 2008. V. 56. P. 470–482.

Komarova T.V., Baschieri S., Donini M. *et al*. Transient expression systems for plant-derived biopharmaceuticals // Expert Rev. Vaccines. 2010. V. 9. P. 859–876.

Kumar G.R., Sakthivel K., Sundaram R.M. *et al*. Allele mining in crops: prospects and potentials // Biotechnol. Adv. 2010. V. 28. P. 451–461.

Lee J.M., Davenport G.F., Marshall D. *et al*. GERMINATE: a generic database for integrating genotypic and phenotypic information for plant genetic resource collections // Plant Physiol. 2005. V. 139. P. 619–631.

Lu Y., Savage L.J., Ajjawi I. *et al*. New connections across pathways and cellular processes: industrialized mutant screening reveals novel associations between diverse phenotypes in *Arabidopsis* // Plant Physiol. 2008 V. 146. P. 1482–1500.

Pritchard L., Birch P. A systems biology perspective on plant-microbe interactions: biochemical and structural targets of pathogen effectors // Plant Sci. 2011. V. 180. P. 584–603.

Skryabin K. Do Russia and Eastern Europe need GM plants? // N. Biotechnol. 2010. V. 27. P. 593–595.

Smirnova O.G., Ibragimova S.M., Kochetov A.V. Simple database to select promoters for plant transgenesis // Transgenic Res. 2012. 21. P. 429–437.

Vankadavath R.N., Hussain A.J., Bodanapu R. *et al.* Computer aided data acquisition tool for high-throughput phenotyping of plant populations // Plant Methods. 2009. V. 5. 18.

Varshney R.K., Bansal K.C., Aggarwal P.K. *et al*. Agricultural biotechnology for crop improvement in a variable climate: hope or hype? // Trends Plant Sci. 2011. V. 16. P. 363–371.

Wiley P.E., Campbell J.E., McKuin B. Production of biodiesel and biogas from algae: a review of process train options // Water Environ. Res. 2011. V. 83. P. 326–338.

INFORMATIONAL PORTAL «PLANT BIOTECHNOLOGY» – INTERNET RESOURCE TO SUPPORT EXPERIMENTS IN PLANT GENE ENGINEERING, GENETICS AND WHEAT BREEDING

A.V. Kochetov, O.G. Smirnova, S.M. Ibragimova, D.A. Rasskazov, D.A. Afonnikov, M.A. Genaev, A.V. Doroshkov, T.A. Pshenichnikova, A.V. Simonov, E.V. Morozova

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: ak@bionet.nsc.ru

Summary

New Internet-resource to support the research in plant biotechnology is presented. This Internet portal contains specialized modules (databases and software) and allows users to combine these modules to solve various tasks as well as it permits the further resource development by addition of new modules. Currently the resource contains the database of external informational sources, the database on promoters for plant transgenesis, the database on translational enhancers for plant transgenesis, and the database WheatPGE to support the experiments in a wheat breeding. The resource is available at ICG www-site (http://bioagrotech.bionet.nsc.ru/).

Key words: informational resource, database, promoter, translational enhancers, genetic engineering, bread wheat, phenotyping, breeding.

УДК 004.75

BioInfoWF – СИСТЕМА АВТОМАТИЧЕСКОЙ ГЕНЕРАЦИИ WEB-ИНТЕРФЕЙСОВ И WEB-СЕРВИСОВ ДЛЯ БИОИНФОРМАЦИОННЫХ ИССЛЕДОВАНИЙ

© 2012 г. **М.А. Генаев[1], Е.Г. Комышев[2], К.В. Гунбин[1], Д.А. Афонников[1, 2]**

[1] Учреждение Российской академии наук Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: mag@bionet.nsc.ru;
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

Поступила в редакцию 5 июля 2012 г. Принята к публикации 25 июля 2012 г.

В настоящей работе представлена система BioInfoWF для автоматической генерации Web-интерфейсов и Web-сервисов для вычислительных модулей в области биоинформатики. Для каждого вычислительного модуля, используемого в этой системе, вводится метаописание на языке XML. На основе метаописаний BioInfoWF автоматически генерируют Web-интерфейсы и Web-сервисы, которые в дальнейшем могут использоваться как в различных биоинформационных системах, так и непосредственно в самой системе BioInfoWF. Вычислительные модули в нашей системе могут объединяться в конвейеры, для которых автоматически генерируется пользовательский Web-интерфейс. Разработанный нами инструмент существенно упрощает разработку и публикацию модулей анализа биоинформатических данных в сети, что обеспечивает их доступность для сообществ биологов и биоинформатиков. Система BioInfoWF распространяется под свободной лицензией GNU GPL. Дистрибутив и пользовательская документация системы BioInfoWF доступны на сайте http://bioinfowf.bionet.nsc.ru.

Ключевые слова: биоинформатика, интеграция данных, конвейерная обработка данных, метаописание данных, Web-интерфейс, Web-сервис.

ВВЕДЕНИЕ

Методы биоинформатики все шире применяются в самых различных областях биологии: при анализе данных высокопроизводительного секвенирования (Pop, Salzberg, 2008), в биомедицине (Fernald *et al*., 2011), генетике (Moore *et al*., 2010), изучении молекулярной эволюции (Sánchez *et al*., 2011), при анализе фенотипических признаков организмов (Hartmann *et al*., 2011). Для решения конкретной биологической задачи рутинными процедурами являются обращение к базам данных, использование большого числа программ, объединенных в цепочки, визуализация полученных результатов на разных этапах обработки данных. Успех решения зависит от эффективной обработки огромного количества информации, представленной в разных форматах; применения большого числа математических моделей, которые реализованы разными вычислительными программами и алгоритмами. Все это обусловливает важность использования в биоинформатике конвейерных систем организации обработки данных (Deelman *et al*., 2009).

При разработке подобных систем актуальной задачей является обеспечение возможности повторного использования разработанных конвейеров, их доступности другим пользователям, в том числе и в виде отдельных модулей. Такие возможности предоставляет Интернет, в частности, организация работы конвейеров на основе Web-сервисов. Web-сервис – это программная система, идентифицируемая строкой универсального идентификатора ресурса (universal resource identifier, URI), его общедоступные интерфейсы определены на языке XML и основаны на базе открытых стандартов

и протоколов. Web-сервис является единицей модульности при использовании сервис-ориентированной архитектуры приложения и обеспечивает взаимодействие программных систем независимо от платформы.

При большом разнообразии имеющихся Web-сервисов встает задача их структуризации. В настоящее время разработано несколько популярных систем каталогизации и конструирования конвейеров в области биоинформатики, которые используют Web-сервисы. Проект BioCatalogue (Bhagat *et al.*, 2010) является реестром биологических Web-сервисов. BioCatalogue предоставляет общий интерфейс для регистрации, просмотра и аннотирования Web-служб в сообществе наук о жизни. При аннотации сервисов учитываются их технический тип, биоинформатическая категория, поставщик этого сервиса, различные пользовательские теги, форматы входных/выходных данных. Сервисы BioCatalogue также являются предметом постоянного мониторинга, позволяющего идентифицировать проблемы с обслуживанием и изменением сервиса. Это позволяет осуществлять фильтрацию недоступных или ненадежных ресурсов. Сервисы в BioCatalogue доступны не только в виде программного интерфейса, но и с помощью Web-интерфейса на основе технологий «Web 2.0» (Bhagat *et al*., 2010).

При решении конкретной задачи в области биоинформатики, как правило, используется набор методов, каждый из которых реализован в виде отдельного приложения или Web-сервиса. В настоящее время разработано несколько популярных систем в области биоинформатики, позволяющих организовывать такие методы в конвейеры. Схема конвейера в этом случае является планом эксперимента для решения конкретной биологической задачи. Одной из популярных систем, предоставляющих конвейерную обработку данных, является Galaxy (Goecks *et al*., 2010). Пользователи, не имеющие опыта программирования, могут легко задать параметры, запуская вычислительные модули и конвейеры. При проведении вычислений Galaxy фиксирует промежуточную информацию, для того чтобы любой пользователь мог повторить и подробно проанализировать результаты численного анализа. В системе Galaxy пользователи могут совместно использовать и публиковать конвейеры через Web и создавать Web-документы, описывающие протоколы и смысл исследований.

Другой популярной системой подобного типа является Taverna (Oinn *et al*., 2004). Это графическая среда для управления и запуска конвейеров, реализованная на языке Java в виде приложения с графическим интерфейсом (GUI). Taverna обеспечивает удобство интеграции программ, баз данных и Web-сервисов, доступных в Интернет. Это позволяет биоинформатикам конструировать конвейеры для решения задач в различных областях, таких, как секвенирование и аннотация геномов. Распределенная Web-сервисная архитектура системы позволяет использовать много различных провайдеров ресурсов для проведения анализа. Схему готового конвейера можно сохранить, отредактировать и запустить повторно, причем не только через GUI интерфейс, но и как консольное приложение.

Популярность конвейерной обработки данных и накопление огромного числа различных готовых конвейеров, каждый из которых нацелен на решение узкой биологической задачи, привели к появлению таких ресурсов, как myExperiment (Goble *et al*., 2010). myExperiment – среда для совместной работы, где ученые могут публиковать свои планы эксперимента, конвейеры, делиться ими с другими пользователями. Конвейеры, объекты (программы и данные из баз), их связки (так называемые пакеты) можно искать, сортировать и классифицировать как фотографии и видео в Интернете. Однако в отличие от Facebook или MySpace myExperiment полностью ориентирован на потребности исследователей. myExperiment обеспечивает удобный доступ к знаниям (информации, научным данным) для следующего поколения ученых, предоставляя пул научных методов, формируя сообщества и определяя протокол общения внутри этого сообщества. Все это приводит к уменьшению времени выхода на биологический эксперимент.

Широта решаемых задач и популярность работы в системах, подобных Taverna и Galaxy, напрямую зависят от доступности Web-сервисов и разнообразия задач, которые они способны выполнять. В настоящей работе мы представляем систему BioInfoWF для автоматической генерации Web-сервисов и Web-интерфейсов. Предложено формальное метаописание вычислительных модулей на XML. На основе таких

метаописаний система BioInfoWF автоматически генерирует Web-интерфейсы и Web-сервисы, которые в дальнейшем могут использоваться как в различных биоинформационных системах (Biocatalogue, Taverna, myExperiment), так и непосредственно в самой системе BioInfoWF. Вычислительные модули в нашей системе могут объединяться в конвейеры, для которых автоматически генерируется пользовательский Web-интерфейс. Разработанный нами инструмент существенно упрощает разработку и публикацию модулей анализа биоинформатических данных в сети и их доступность в сообществах биоинформатиков и биологов, таких, как Biocatalogue, myExperiment и др.

Таким образом, BioInfoWF позволяет решать задачи конвейерной обработки данных, обеспечивает генерацию удобного пользовательского интерфейса «Web 2.0» для запуска вычислительных модулей и конвейеров, а также дает возможность автоматически создавать Web-сервисы.

МАТЕРИАЛЫ И МЕТОДЫ

Архитектура системы

Предлагаемая система включает (рис. 1):

1) менеджер запуска вычислительных модулей и готовых конвейеров BioInfoWF;

2) репозитории, в которых хранятся: а) описания вычислительных модулей и схем конвейеров, которые являются наборами XML файлов; б) вычислительные модули, базы данных и программы визуализации данных; в) пользовательские данные;

3) системы генерации Web-интерфейсов и Web-сервисов.

Описание конвейера и его исполнение

В системе BioInfoWF конвейер представляет собой набор вычислительных модулей, связанных топологией выполнения задачи. Модули представляют собой программы, запускаемые в консольном режиме в среде Linux. Управляющие параметры (названия входных и выходных файлов, параметры алгоритмов) передаются в расчетные модули через командную строку или переменные окружения. В ходе выполнения задачи выходные данные одного модуля могут подаваться на вход другому модулю.

Разработанная нами система позволяет интегрировать любые вычислительные модули, организованные подобным образом. При этом данные могут находиться как на локальной машине, так и на удаленной. Схема интеграции (порядок выполнения процедур) и метаописания вычислительных модулей определяются на языке XML.

Рис. 1. Основные структурные элементы системы BioInfoWF.

Серверная часть, реализованная на языке Perl, выполняет запуск конвейера и отслеживает статус выполнения каждого вычислительного модуля. На вход приложению подается описание схемы конвейера и вычислительных модулей. Приложение запускает конвейер, создавая файл с отчетом. В отчете указывается статус выполнения каждого узла в конвейере. Серверная часть поддерживает параллельный запуск узлов конвейера, при этом максимальное количество потоков определяется в конфигурационном файле приложения.

Язык описания вычислительных модулей схем конвейеров

Для формального описания схемы конвейера и вычислительных модулей, из которых он состоит, нами был разработан язык на основе XML. Описание конвейера представлено двумя файлами. Первый описывает вычислительные модули, второй задает топологию конвейера (рис. 2).

Описания вычислительных модулей содержатся в хранилище и состоят из следующих разделов:

– *Входные файлы*. Описание того, какие входные файлы подаются на вход модулю; для каждого файла указывается его идентификатор, даются описание, формат файла.

– *Выходные файлы*. Описание выходных файлов вычислительного модуля.

– *Параметры и опции*. Описание параметров и опций для вычислительного модуля. Для каждого параметра задаются идентификатор, описание, тип параметра (например, строка, число или бинарное значение), значение по умолчанию, внешний вид поля запроса значения для параметра на странице WWW-браузера.

– *Правила генерации командной строки*. Описание набора входных параметров вычислительного модуля и списка входных и выходных файлов. На основе этой информации менеджер запуска генерирует командную строку для запуска вычислительного модуля.

– *Правила поведения пользовательского интерфейса*. Опциональный раздел, который указывает, каким образом Web-интерфейс будет динамически реагировать на действия пользователей.

Пример описания в формате XML команды kill, которая прекращает выполнение какого-либо вычислительного процесса, приведен на рис. 3. Программа на вход принимает идентификационный номер процесса, который надо завершить. Вторая строчка файла (рис. 3) описывает название вычислительного модуля и путь, где располагается исполняемый файл модуля. В нашем случае kill – это команда окружения bash, поэтому указание полного пути не требуется. Секция output (строки 7–14) описывает выходные файлы, в этом примере описываются два файла с идентификаторами *stdout* и *stderr*, которые мы в дальнейшем проассоциируем со стандартными потоками вывода 1 и 2 соответственно. Аналогичным образом описывается секция input для входных файлов. В нашем примере входных файлов нет,

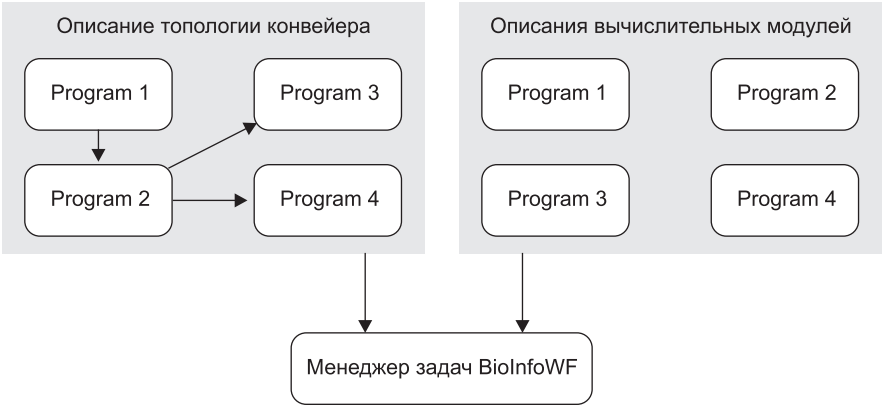

Рис. 2. Описание любого конвейера включает описание вычислительных модулей, вовлеченных в конвейер, и связей между этими модулями.

поэтому секция отсутствует. В секции options описывается единственная опция, которая будет передавать идентификатор id процесса команде kill. Опция имеет тип *int* и представление *text*, которое будет соответствовать *<input type=«text» />* при генерации Web-интерфейса. Значение по умолчанию для опции не задано. Секция cmdline описывает правила генерации командной строки. На входе мы имеем две хеш-таблицы, *$options* и *$files*. Ключами в этих хеш-таблицах служат id из секций *input*, *output* и *cmdline*. На выходе необходимо сформировать переменную *$cmd*, которая содержала бы готовую командную строку для вычислительного модуля.

Генерация пользовательского Web-интерфейса

Пользовательский интерфейс генерируется автоматически на основе XML описаний вычислительных модулей и схемы конвейера. Пользователю предлагается работать с уже готовыми схемами. В текущей версии клиентской части возможна работа только с последовательными конвейерами.

Схема генерации Web-страниц представлена на рис. 4. Менеджер запуска задач BioInfoWF получает на вход описание схемы конвейера и вычислительных модулей в формате XML и с использованием библиотеки Perl HTML::Template генерирует файл в формате html. Автоматизация достигается за счет того, что в описании каждого входного параметра указывается тип элемента HTML для его визуального представления (выпадающее меню, радиокнопка, текстовое поле и т. п.). Пользователь может управлять поведением конвейера, изменяя его схему и входные данные для вычислительных модулей. Для визуализации и редактирования входных и выходных данных имеются возможности подключения внешних программ, реализованных, как правило, в виде Java Applet приложений.

Реакция интерфейса на действие пользователя при его работе с HTML-страницами достигается за счет внедрения в описание каждого модуля правил поведения, реализованных с помощью библиотеки jQuery. Динамическое изменение интерфейса удобно использовать, когда существуют зависимости между параметрами исполняемых модулей. Например, в модуле, описывающем задачу множественного выравнивания белковых последовательностей программой Mafft, выбор матрицы сравнения аминокислот обусловливает ряд дополнительных опций, которые зависят от ее типа. При выборе матрицы на Web-странице отображаются только опции, связанные типом выбранной матрицы.

```
1  <programs>
2  <program name="Kill" exe="kill" cluster="off">
3  <description>
4    kill - terminate a process
5  </description>
6
7  <output>
8  <file id="stdout" type="text" name="STDOUT" description="Standard output" />
9  <file id="stderr" type="text" name="STDERR" description="Standard error" />
10 </output>
11
12 <options>
13 <option id="PID" name="PID" description="PID" view="text" type="int" default="" />
14 </options>
15
16 <cmdline>
17   $cmd = " $options{PID} "
18   "1>\"$files{stdout}\" ".
19   "2>\"$files{stderr}\"";
20 </cmdline>
21
22 </program>
23 </programs>
```

Рис. 3. Пример XML описания вычислительного модуля.

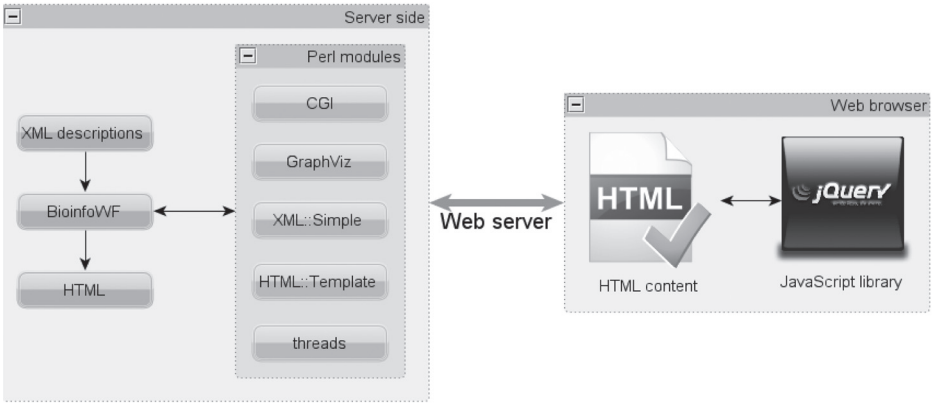

Рис. 4. Схема процесса генерации Web-интерфейса в системе BioInfoWF.

Предложенный подход позволяет автоматически создавать Web-интерфейсы для любых вычислительных модулей, описанных в системе.

В текущей версии интерфейса доступны следующие базовые опции управления конвейером.

– Установка входных файлов, параметров и опций для каждого вычислительного модуля в конвейере.

– Старт с произвольного узла и остановка на произвольном узле в конвейере.

– Отслеживание статуса выполнения каждого вычислительного модуля в конвейере.

– Просмотр входных/выходных файлов для каждого этапа расчета в конвейере.

Привязка форматов входных/выходных файлов к различным приложениям для их визуализации.

Генерация Web-сервисов

Генерация Web-сервисов производится автоматически приложением Java, модулем генерации Web-сервисов (рис. 5). Интерфейс генерируется автоматически на основе метаописания вычислительных модулей, содержащего всю необходимую для этого информацию. Для генерации необходимо запустить модуль WebServicesGen в консольном режиме, передав в параметрах файл с метаописанием вычислительных модулей и название вычислительного модуля, после чего Web-сервис будет сгенерирован и размещен на сервере. Таким образом, в системе BioInfoWF реализована быстрая автоматическая генерация Web-сервисов для новых вычислительных модулей.

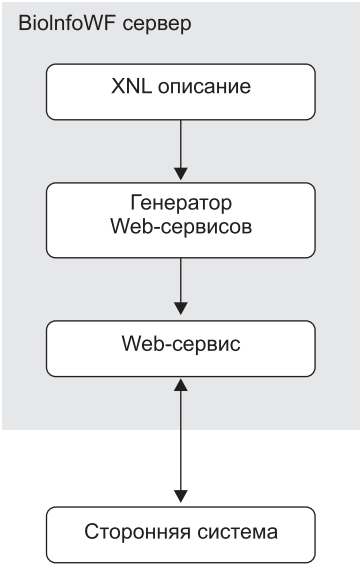

Рис. 5. Схема генерации Web-сервисов и взаимодействия их со сторонней системой.

Задача сгенерированных Web-сервисов – предоставление доступа к вычислительным модулям системы BioInfoWF. При этом модуль генерации Web-сервисов производит Java-сервлет, аналог клиентского апплета, исполняющегося на сервере. На каждый вычислительный модуль генерируется отдельный сервлет, который в свою очередь реализует как REST, так и WSDL/SOAP Web-сервис как синхронного, так и асинхронного типа на основе протокола HTTP.

REST – подход к архитектуре построения Web-приложений – позволяет использовать простые HTTP запросы, содержащие параметры в заголовках, а передаваемые данные – в теле HTTP запроса. Таким образом, простой REST

Web-сервис возможно использовать в системах, не поддерживающих более сложный SOAP протокол. Для использования REST Web-сервиса вычислительного модуля необходимо послать HTTP запрос PUT на адрес *http://BioInfoWF.server/ServiceName/* (где *BioInfoWF.server* – адрес сервера BioInfoWF в сети, а *ServiceName* – название соответствующего вычислительного модуля системы BioInfoWF), в заголовке которого содержатся параметры вычислительного модуля, а в теле – данные. Ответ на этот запрос будет содержать посчитанные вычислительным модулем данные. Для получения краткой справочной информации о вычислительном модуле необходимо послать GET запрос на тот же адрес.

Язык описания Web-сервисов WSDL вместе с протоколом обмена структурированными сообщениями в распределенной вычислительной среде SOAP позволяют выполнять более гибкие запросы и предоставлять формальную информацию о Web-сервисе. Системе, в которой реализована поддержка SOAP/WSDL Web-сервисов, достаточно предоставить WSDL описание, указав URI, по которому оно находится, после чего эта система может взаимодействовать с Web-сервисом.

Асинхронный тип Web-сервиса полезен при вычислениях, требующих временных затрат, когда поддержание постоянного HTTP соединения на запрос–ответ проблематично. В этом случае для загрузки входных данных при инициализации вычисления, их выгрузке и проверке состояния вычисления используются отдельные HTTP запросы и ответы.

Java-сервлет, реализующий эти Web-сервисы совместно, упаковывается в war-файл (jar архив для сервлетов), который может быть развернут на любом контейнере сервлетов. Для контейнера сервлетов Catalina пакета Tomcat достаточно поместить этот файл в папку Webapps/.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Система BioInfoWF распространяется под свободной лицензией GNU GPL. Дистрибутив и пользовательская документация системы доступны на официальном сайте http://bioinfowf.bionet.nsc.ru. Приведем пример использования системы BioInfoWF для решения задач молекулярной эволюции генов и белков.

Решение этой задачи обычно заключается в последовательном выполнении набора операций с нуклеотидными или аминокислотными последовательностями (извлечение гомологов из банка данных, выравнивание, реконструкция филогении, оценка режима эволюции). Ранее были созданы несколько вариантов специализированных конвейерных систем для решения задач молекулярной эволюции и филогении (Dereeper *et al*., 2008; Sбnchez *et al*., 2011). При помощи системы BioInfoWF нами был создан пакет анализа молекулярной эволюции SAMEM (Гунбин и др., 2011; Gunbin *et al*., 2012). SAMEM состоит из двух основных конвейеров, анализа эволюции генов (I) и анализа эволюции белков (II) и двух дополнительных конвейеров, собирающих выборки генов и белков (III) и производящих их первичный анализ (IV). Уникальной особенностью конвейера I является возможность исследования данных ранее предложенными методами оценки K_R/K_C (Zhang, 2000; Smith, 2003), а также оригинальным, впервые предложенным, методом оценки K_R/K_C (Gunbin *et al*., 2012). При анализе данных конвейером II существенной особенностью является оригинальный метод анализа скоростей фиксации различных типов аминокислотных замен на ветвях филогенетического дерева (Гунбин и др., 2011; Gunbin *et al*., 2012). Метод основан на Марковском моделировании эволюции (Pupko *et al*., 2002) и непараметрическом перестановочном тесте (Gunbin *et al*., 2011), осуществляющем сравнение числа ожидаемых и наблюдаемых аминокислотных замен, и позволяет анализировать режимы молекулярной эволюции на глубоких ветвях филогенетического дерева. Подход, реализованный в системе SAMEM, также позволяет 1) использовать все известные свойства аминокислот и 2) проводить статистическое соотнесение изменения этих свойств с признаками фенотипа.

Вычислительные модули пакета SAMEM могут быть доступны как в качестве пользовательского Web-интерфейса, так и в качестве Web-сервисов. Примеры Web-интерфейса SAMEM и использование модулей SAMEM в виде Web-сервисов в системе Taverna приведены на рис. 6.

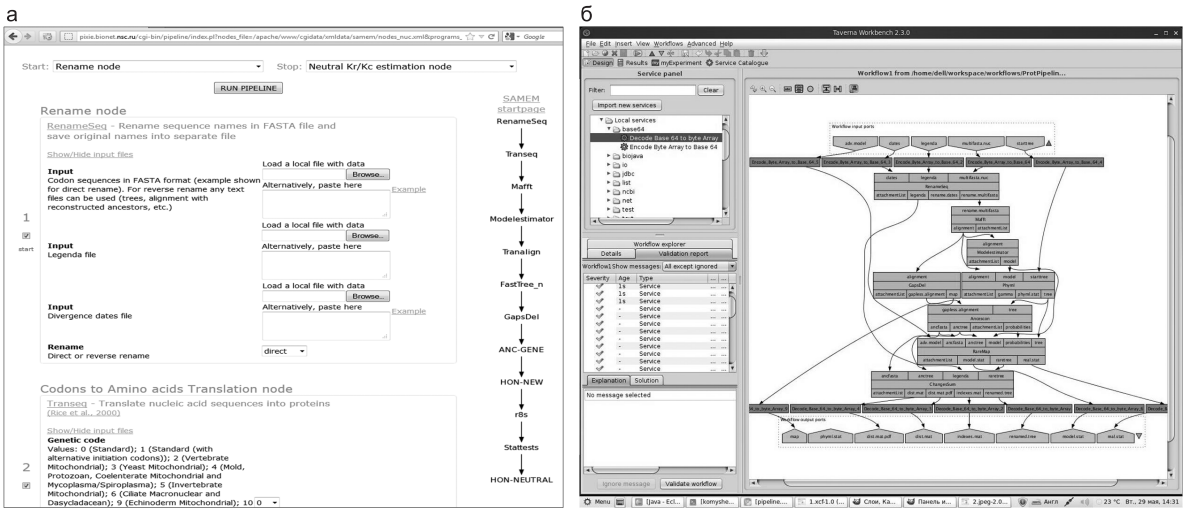

Рис. 6. Интерфейс пользователя системы SAMEM, реализованной на платформе BioInfoWF.

а – реализация конвейера в виде Web-интерфейса; б – реализация этого конвейера в системе Taverna с использованием Web-сервисов.

ЗАКЛЮЧЕНИЕ

В настоящей работе предложена система BioInfoWF, которая позволяет генерировать Web-сервисы для решения биоинформатических задач, организовывать их работу в виде конвейера, автоматически генерировать пользовательский Web-интерфейс. Удобство системы заключается в быстром подключении вычислительных модулей на основе исполняемых приложений и Web-сервисов. Использование системы было продемонстрировано на примере создания конвейеров решения задач по анализу молекулярной эволюции генов и белков – SAMEM.

Работа поддержана грантом Министерства образования и науки РФ в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 гг.» (07.514.11.4023).

ЛИТЕРАТУРА

Гунбин К.В., Генаев М.А., Турнаев И.И., Афонников Д.А. Компьютерная система анализа режимов молекулярной эволюции генов и белков: анализ эволюции циклинов В // Вестн. Томского гос. ун-та. Биология. 2011. 4. С. 175–189.

Bhagat J., Tanoh F., Nzuobontane E. *et al*. BioCatalogue: a universal catalogue of Web services for the life sciences // Nucl. Acids Res. 2010. V. 38. P. 689–694.

Deelman E., Gannon D., Shields M., Taylor I. Workflows and e-Science: An overview of workflow system features and capabilities // Future Generation Computer Systems. 2009. V. 25. P. 528–540.

Dereeper A., Guignon V., Blanc G. *et al*. Phylogeny.fr: robust phylogenetic analysis for the non-specialist // Nucl. Acids Res. 2008. V. 36. P. 465–469.

Fernald G.H., Capriotti E., Daneshjou R. *et al*. Bioinformatics challenges for personalized medicine // Bioinformatics. 2011. V. 27. P. 1741–1748.

Goble C.A., Bhagat J., Aleksejevs S. *et al*. myExperiment: a repository and social network for the sharing of bioinformatics workflows // Nucl. Acids Res. 2010. V. 38. P. 677–682.

Goecks J., Nekrutenko A., Taylor J., Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences // Genome Biol. 2010. 11:R86.

Gunbin K.V., Suslov V.V., Genaev M.A., Afonnikov D.A. Computer system for analysis of molecular evolution modes (SAMEM): analysis of molecular evolution modes at deep inner branches of the phylogenetic tree // In Silico Biol. 2012. In press.

Gunbin K.V., Suslov V.V., Turnaev I.I. *et al*. Molecular evolution of cyclin proteins in animals and fungi // BMC Evol. Biol. 2011. V. 11. P. 224.

Hartmann A., Czauderna T., Hoffmann R. *et al*. HTPheno: an image analysis pipeline for high-throughput plant phenotyping // BMC Bioinformatics. 2011. V. 12. P. 148.

Moore J.H., Asselbergs F.W., Williams S.M. Bioinformatics challenges for genome-wide association studies //

Bioinformatics. 2010. V. 26. P. 445–455.
Oinn T., Addis M., Ferris J. *et al*. Taverna: a tool for the composition and enactment of bioinformatics workflows // Bioinformatics. 2004. V. 20. P. 3045–3054.
Pop M., Salzberg S.L. Bioinformatics challenges of new sequencing technology // Trends Genet. 2008. V. 24. P. 142–149.
Pupko T., Pe'er I., Hasegawa M. *et al*. A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families // Bioinformatics. 2002. V. 18. P. 1116–1123.
Sánchez R., Serra F., Tárraga J. *et al*. Phylemon 2.0: a suite of Web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing // Nucl. Acids Res. 2011. V. 39. P. 470–474.
Smith N.G. Are radical and conservative substitution rates useful statistics in molecular evolution? // J. Mol. Evol. 2003. V. 57. P. 467–478.
Zhang J. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes // J. Mol. Evol. 2000. V. 50. No. 1. P. 56–68.

BioInfoWF – WEB SERVICES AND WEB INTERFACES GENERATOR FOR BIOINFORMATICS ANALYSIS

M.A. Genaev[1], E.G. Komyshev[2], K.V. Gunbin[1], D.A. Afonnikov[1,2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: mag@bionet.nsc.ru;
[2] Novosibirsk National Research State University, Novosibirsk, Russia

Summary

The BioinfoWF (Bioinformatics WorkFlow) system for automated generation of Web interface and Web services for bioinformatics programs. Each program module used in the system has metadescription in XML. The metadescriptions are used for automated generation of Web interface and Web services that can be used further in bioinformatics workflows. Computational modules can be organized in workflows. The tool we have developed significantly simplify the design and publication of modules for bioinformatics data analysis via the internet and their availability for scientific communities. The developed system makes is distributed under GNU GPL. The Source codes and documentation for BioinfoWF are available at http://bioinfowf.bionet.nsc.ru.

Key words: bioinformatics, data integration, workflow data processing, data metadescription, Web interface, Web service.

УДК 579.842.11:57.042:53.047:53.05.

СИСТЕМА ДЕТЕКЦИИ БИОАНАЛИТИЧЕСКОГО КОМПЛЕКСА НОВОГО ПОКОЛЕНИЯ

© 2012 г. **Е.В. Сысоев[3], А.К. Поташников[3], Ю.В. Обидин[3], Т.Н. Горячковская[1], В.С. Базин[3], В.М. Попик[2], С.Е. Пельтек[1], Н.А. Колчанов[1, 4, 5]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия;
[2] Институт ядерной физики СО РАН, Новосибирск, Россия;
[3] Конструкторско-технологический институт научного приборостроения СО РАН, Новосибирск, Россия, e-mail: potash@tdisie.nsc.ru;
[4] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия;
[5] НИЦ «Курчатовский институт», Москва, Россия

Поступила в редакцию 15 июля 2012 г. Принята к публикации 1 августа 2012 г.

Разработана и создана система детекции для определения наличия антител/антигенов в биологических жидкостях. Система предназначена для регистрации кинетики свободной иммунодиффузии флюоресцентных нанокомплексов в каналах микрофлюидного модуля биоаналитического комплекса нового поколения. Система состоит из четырех параллельных каналов возбуждения, системы регистрации флюоресценции и программы обработки изображения. Система позволяет регистрировать наличие антител/антигенов в биологических жидкостях в концентрациях менее чем 0,1 мкг/мл.

Ключевые слова: биоаналитические комплексы, биобезопасность, микрофлюидная система.

ВВЕДЕНИЕ

Большинство исследований в биологии и химии в течение всей их истории проводились макроскопическими методами – в колбах, пробирках, чашках Петри. Прогресс микроэлектроники всего за несколько десятков лет привел к впечатляющей миниатюризации вычислительных приборов: от огромных приборов, занимающих целые комнаты, до миниатюрных микросхем. В связи с этим у многих исследователей возникла идея ввести сходные технологии в химию и биологию, что и стало началом микрофлюидных технологий. Ограничивающим фактором в системе микрометровых размеров являются возможности человека эффективно манипулировать биологическими объектами.

В настоящее время в мировой науке происходит технологическая революция, основанная на использовании микрофлюидных систем, направленная на переход к малым и сверхмалым размерам устройств для изучения функции биологических макромолекул, геномов, клеток, клеточных структур, а также для фармакологии, клинической диагностики, биохимических исследований, аналитической и индустриальной химии и др.

Главной характеристикой микрофлюидных систем является возможность оперировать микро- и нанообъемами анализируемых жидкостей. Основной целью данной работы являются разработка и создание системы детекции для биоаналитического комплекса нового поколения, основанного на соединении технологии микрофлюидных систем и процесса свободной иммунодиффузии флюоресцентных нанокомплексов антиген/антитело для снижения стоимости клинических анализов крови за счет уменьшения объемов дорогостоящих реагентов,

биоопасных отходов и уменьшения издержек при массовом производстве.

В биоаналитическом комплексе в каналах микрофлюидной системы происходит реакция свободной иммунодиффузии, регистрация которой осуществляется путем измерения интенсивности флюоресцентного сигнала с помощью цифровой камеры и последующей цифровой обработки полученных изображений (Пельтек и др., в печати).

МАТЕРИАЛЫ И МЕТОДЫ

1. Структура системы детекции

Структура системы детекции (СД) сигнала интенсивности флюоресценции представлена на рис. 1. В ее состав входят:

– твердотельный лазер на основе лазерного диода;

– оптическая система ОС-1, обеспечивающая формирование светового пучка и фильтрацию спонтанного излучения лазера;

– оптическая система ОС-2, представляющая собой делитель лазерного луча на четыре параллельных пучка;

– оптическая система ОС-3, состоящая из объектива и набора светофильтров;

– цифровая камера;

– компьютер с программным обеспечением.

Система предназначена для возбуждения флюоресценции в микрофлюидном модуле и количественной оценки ее величины как функции времени. ОС-2 создана на основе светоделительных кубиков с целью придания возможности адаптации к конкретному микрофлюидному модулю.

2. Оптическая схема системы детекции

Оптическая схема СД биоаналитического комплекса приведена на рис. 2.

Лазер 1 является источником света для возбуждения флюоресцентного сигнала флюоресцентных нанокомплексов «антиген–антитело» в каналах микрофлюидного модуля. Основными требованиями, предъявляемыми к лазеру, являются узкая спектральная полоса возбуждения и достаточно высокая мощность излучения в заданном спектральном диапазоне. Лазер может работать в непрерывном и импульсном режимах и имеет длину волны излучения 473 нм.

Интерференционный светофильтр 2 с узкой полосой пропускания, центр которой совпадает с основной длиной волны лазера, служит для устранения спонтанного излучения лазера в полосе частот спектра люминесценции. Объектив 3 обеспечивает формирование пучка лазера с требуемым размером поперечного сечения (50 мкм) в рабочей плоскости микрофлюидного модуля. Поворотная призма 4 служит для юстировки лазерного пучка.

Светоделительные кубики предназначены для формирования четырех лазерных пучков для возбуждения флюоресцентных нанокомплексов параллельно в четырех точках детекции микрофлюидного модуля. Светоделительные кубики установлены таким образом, чтобы плоскость светоделительных граней была параллельна попадающему в кубики лазерному пучку. Светоделительные грани первого и второго светоделительных кубиков повернуты на 90° относительно друг друга вокруг оси падающего лазерного пучка. Такое расположение светоделительных кубиков позволяет получить

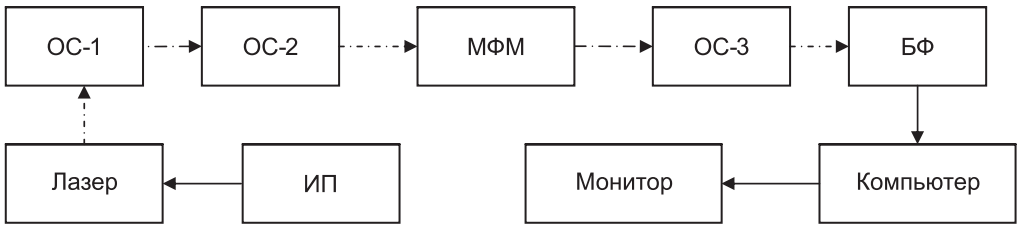

Рис. 1. Структура системы детекции.

ОС-1 – оптическая система 1, ОС-2 – оптическая система 2, ММФ – модуль микрофлюидный; ОС-3 – оптическая система 3; БФ – блок фотоприемников (видеокамера); ИП – источник питания.

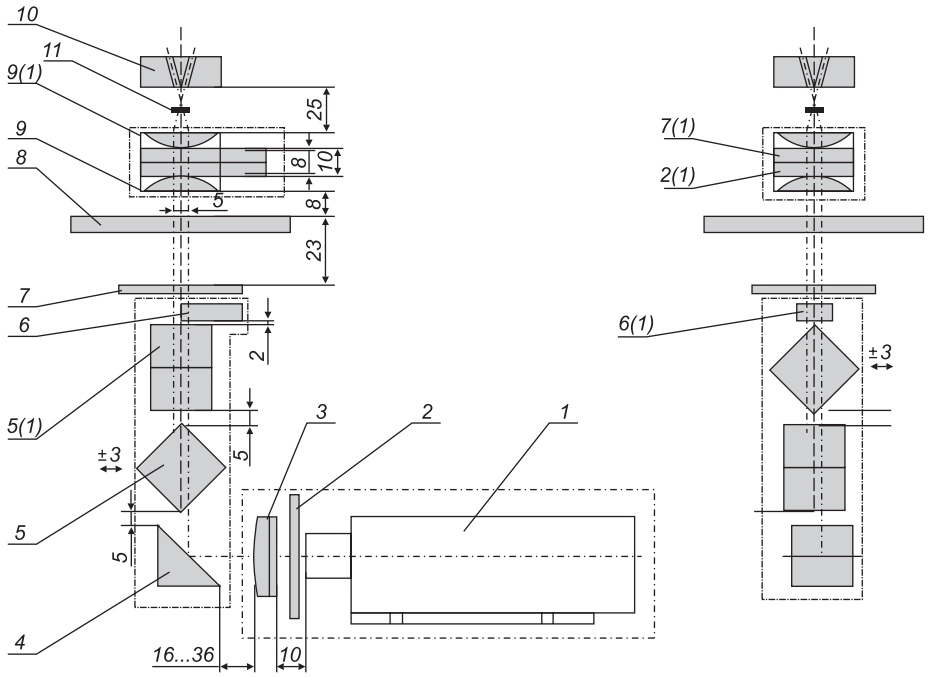

Рис. 2. Оптическая схема системы детекции.

1 –лазер; 2 – интерференционный фильтр, отрезающий спонтанное излучение; 3 – формирующий объектив; 4 – поворотная призма; 5 – два светоделительных кубика; 6 – две юстировочные пластинки; 7 – поляризационный светофильтр; 8 – микрофлюидный модуль; 9 – фокусирующий объектив; 10 – блок фотоприемный; 11 – поляризационный светофильтр; 12 – интерференционный светофильтр.

из одного лазерного пучка четыре, идущих параллельно друг другу.

Поляризационный светофильтр 7 обеспечивает узкую линию возбуждения флюоресценции нанокомплексов. Плоскопараллельные стеклянные пластинки 6 позволяют выполнить тонкую подстройку координат точек детекции в микрофлюидном модуле. Объектив 9 предназначен для фокусировки излучения люминесценции на фотоприемном блоке 10. Конструктивно в объектив 9 встроены поляризационный фильтр 7(1) и узкополосный интерференционный режекторный фильтр 2(1).

Спектр отфильтрованного лазерного излучения приведен на рис. 3.

Работа оптической схемы. Пучок света, выходящий из лазера 1, проходит через интерференционный светофильтр 2**,** объектив 3 и после призмы 4 светоделительными кубиками 5 преобразуется в четыре идущих параллельно друг другу лазерных пучка. Далее два из четырех пучков проходят через две юстировочные пластинки 6. После этого все четыре пучка проходят поляризационный фильтр 7. Таким образом, на микрофлюидный модуль попадают четыре идущих параллельно друг другу лазерных пучка, сфокусированных в строго определенных точках детекции в каналах микрофлюидного модуля. Свет лазерных пучков и флюоресценции попадает в объектив 9**.** Излучение лазера подавляется фильтрами, а флюоресцентное излучение проходит на выход объектива и фокусируется в заднем фокусе объектива, в плоскости которого установлена видеокамера.

На рис. 4 и 5 приведены спектры возбуждения флюоресцентных нанокомплексов с различными типами квантовых точек.

Разнесение точек фокусировки излучения лазера и флюоресценции вдоль оптической оси позволяет уменьшить плотность мощности лазерных пучков в задней фокальной плоскости объектива 9 в 10^2–10^3 раз. В совокупности лазерное излучение подавляется в 10^{14}–10^{15} раз. Если мощность лазера составляет 0,04 Вт, то в задней фокальной плоскости объектива 9 она составляет менее $10^{-16} \div 10^{-17}$ Вт на каждый из четырех лазерных пучков.

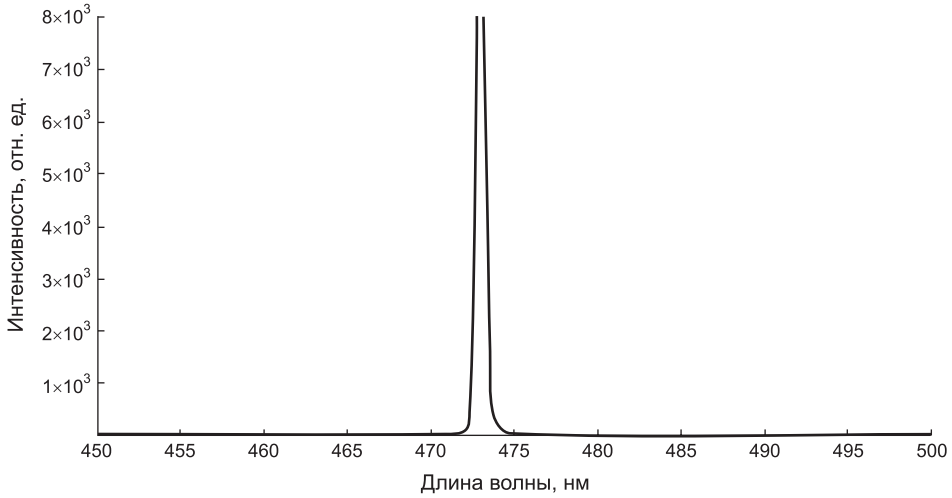

Рис. 3. Спектр фильтрованного излучения лазера.

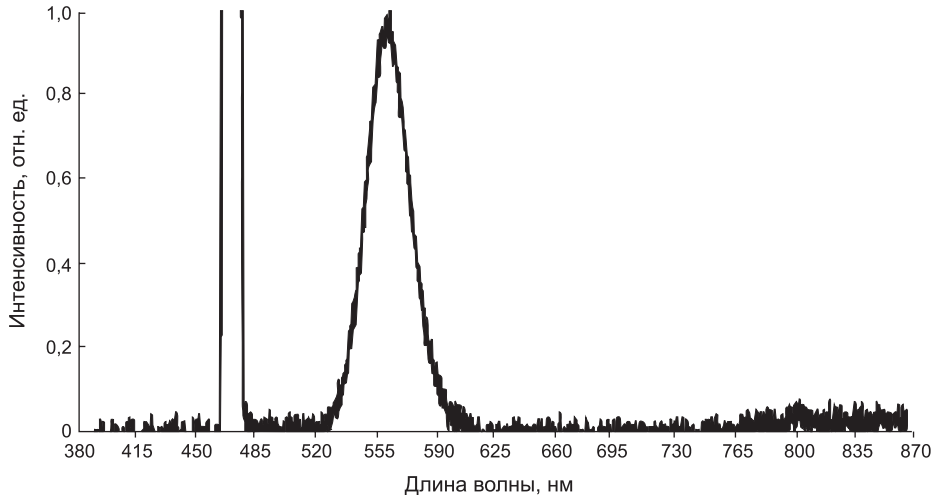

Рис. 4. Спектры возбуждения и флюоресценции нанокомплексов с квантовыми точками Qtracker 565.

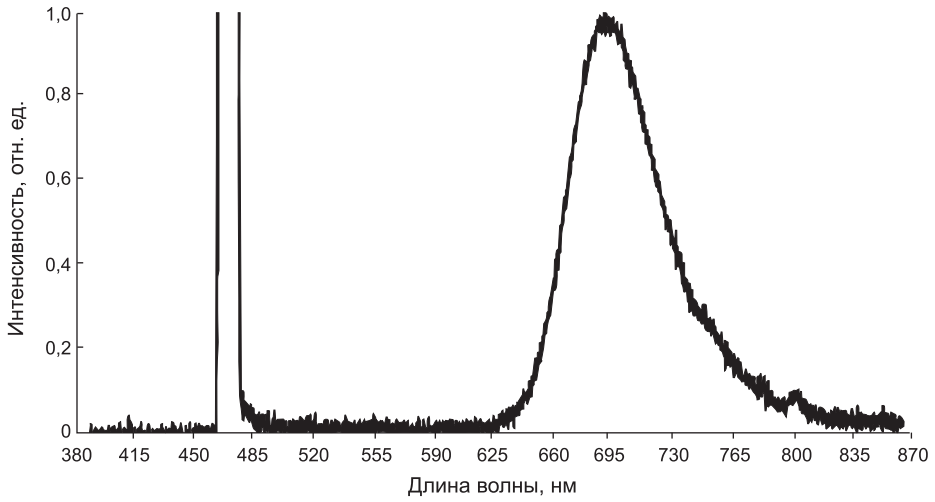

Рис. 5. Спектры возбуждения и флюоресценции нанокомплексов с квантовыми точками Qtracker 705.

3. Фотоприемник

В качестве фотоприемника используется цифровая видеокамера Видеоскан-415Ц-2001 с CCD-матрицей размером 1/2” (6,5×4,83 мм). Разрешение камеры – 782×582 пикселей, размер пикселя 8,3×8,3 мкм. Динамический диапазон – 1000, имеется возможность 12-разрядного представления элементов изображений. Программируемое время экспозиции – от 3,5 мкс до 10 мин. Видеокамера в непрерывном режиме или в режиме принудительного запуска регистрирует изображения каналов микрофлюидного модуля.

Обработка изображений цифровой камеры

Для изображений люминесцирующих объектов характерны яркостные перепады по отношению к темному фону. Для идеального изображения уровень фона соответствует уровню черного в сигнале. В зависимости от интенсивности свечения флюоресцирующие объекты могут перекрывать все яркостные градации динамического диапазона. Для слабосветящихся объектов минимальный уровень сигнала соответствует уровню собственных шумов фотоприемника и маскируется им.

Методы внутрикадровой цифровой обработки (Прэтт, 1982) для таких изображений оказываются неэффективными. Так, например, операции сглаживания вместе с некоторым подавлением шумов приводят к «размытию» яркостных перепадов, а медианная фильтрация устраняет только одиночные шумовые выбросы.

Изображения фотолюминесцирующих объектов статичны, поэтому для улучшения их качества можно применять межкадровую обработку. Весьма эффективен в этом случае метод цифрового шумоподавления (Цифровое телевидение …, 1980), позволяющий уменьшить шумы в изображении за счет усредняемых (накапливаемых) кадров изображения. Без цифрового шумоподавления невозможно получение качественного изображения флюоресцирующего объекта, поэтому данную операцию следует отнести к разряду обязательных функций.

Изображение может быть дополнительно улучшено за счет контрастирования методом преобразования яркостей, сущность которого заключается в перераспределении значений

Рис. 6. Гистограмма изображения флюоресцирующего объекта.

яркостей исходного изображения в пределах заданного динамического диапазона. Для определения границ диапазона яркостей исходного изображения обычно используется гистограмма их распределения. Для интенсивно флюоресцирующих объектов гистограмма имеет два ярко выраженных пика, соответствующих наиболее часто встречающимся значениям фона и сигнала (рис. 6).

Яркостная коррекция с целью удаления фона необходима всегда при дальнейшем автоматизированном количественном анализе и весьма полезна для визуального контроля с целью улучшения субъективного восприятия изображения (Корнышев, Тимофеев, 2007).

Система детекции работает под управлением программы «BioChip».

Программа имеет встроенные средства для калибровки и поверки характеристик цветопередачи. Результаты калибровки сохраняются в файле параметров и автоматически корректируют передаточную функцию камеры.

Главное окно программы представлено на рис. 7.

Программно реализованы два режима выделения зоны флюоресценции: пороговая обработка и заранее задаваемая зона произвольного диаметра. Программа определяет следующие оценки зарегистрированной флюоресценции:

− количество светящихся точек изображения;

− суммарную яркость флюоресцирующих точек;

− среднюю длину волны флюоресценции;

− насыщенность цвета.

Кроме того, программа строит гистограмму результирующего изображения, показывает

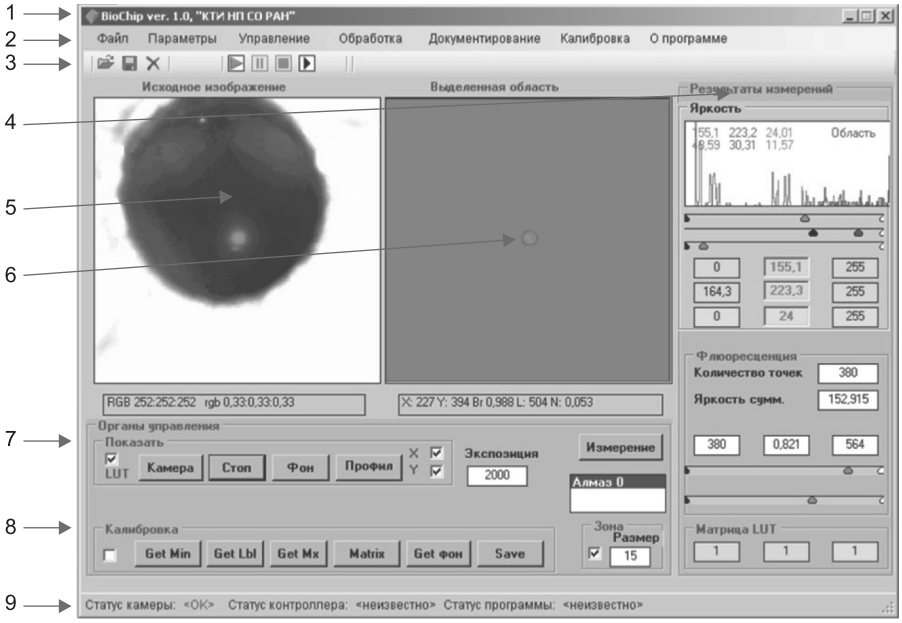

Рис. 7. Главное окно программы BioChip.

1 – заголовок окна, 2 – главное меню, 3 – панель инструментов, 4 – группа элементов для отображения результатов измерений, 5 – окно исходного изображения, 6 – окно обработанного изображения, 7 – группа органов управления, 8 – группа средств калибровки, 9 – строка состояния.

среднее значение яркости по трем цветовым координатам (R, G, B), отображает установленные пороги по яркости, насыщенности и длине волны, которые используются при выделении точек флюоресценции.

На рис. 7 видны два окна изображений. В первом (левом) представлено исходное изображение, регистрируемое камерой. Используя набор задаваемых оператором параметров, программа автоматически выделяет из всего изображения область детекции, в которой осуществляется реакция свободной иммунодиффузии. В правом окне показаны результаты обработки.

Результаты содержат: гистограмму яркости, цифровые индикаторы значений яркости по каждой из цветовых координат (R, G, B), индикатор количества выделенных точек изображения, индикатор суммарной яркости всех точек, а также индикаторы значения насыщенности и длины волны доминирующего цвета.

Алгоритм работы программы «BioChip»

Процедура пороговой обработки изображения выделяет точки детекции. При пороговом выделении точек флюоресценции программа для каждой из точек изображения последовательно выполняет 6 операций:

1. Вычисление квадрата расстояния от центра кадра (x_0, y_0) до обрабатываемой точки (x_i, y_i) $R^2 = (x_i - x_0)^2 + (y_i - y_0)^2$ и сравнение его с заданным значением, что позволяет исключить из рассмотрения точки, расположенные за пределами заданного радиуса от центра кадра.

2. Вычитание из яркости каждой точки, представленной значениями цветовых составляющих (R_i, G_i, B_i), значения яркости фона в этой же точке: $R_i = R_i - R_{cp}$; $G_i = G_i - G_{i\varphi} + G_{cp}$; $B_i = B_i - B_{i\varphi} + B_{cp}$. К полученному значению добавляется средний уровень фона для сохранения уровня яркости. Вычитание фона позволяет более точно вычислять интегральные характеристики точек флюоресценции.

3. Определение суммарной яркости точки ($Bright = R_i + G_i + B_i$) и сравнение ее с порогом. Операция выбраковывает яркие точки, не связанные с точками детекции, и позволяет исключать дефекты и блики в изображении.

4. Сравнение яркости по каждому из цветов с нижним и верхним порогами. Пороги могут быть заданы оператором на этапе настройки.

Это позволяет исключать дефекты и блики в разных цветах.

5. Вычисление насыщенности цвета для каждой выделенной точки изображения:

$$N = \sqrt{(R_i - R_s)^2 + (G_i - G_s)^2 + (B_i - B_s)^2}$$

и сравнение ее с порогами. Здесь (R_s, G_s, B_s) – координаты источника освещения (по умолчанию вычисляются из фона).

6. Определение доминирующей длины волны, при использовании значений точек локуса, заданных таблично. При обращении к таблице необходимо знать текущие цветовые координаты (R_i, G_i, B_i) и координаты источника (R_s, G_s, B_s).

На основе полученных локальных характеристик отдельных флюоресцирующих точек (яркость, насыщенность, длина волны) находятся среднее значение и среднеквадратическое отклонение (СКО). Высокий уровень СКО свидетельствует о возможной недостоверности оценок. Средние значения характеристик представляются оператору на экране монитора.

Любую из 6 указанных выше операций пороговой обработки можно исключить, используя меню или опции программы.

Результаты экспериментальной проверки системы детекции

Для получения количественной оценки чувствительности канала регистрации был ис-

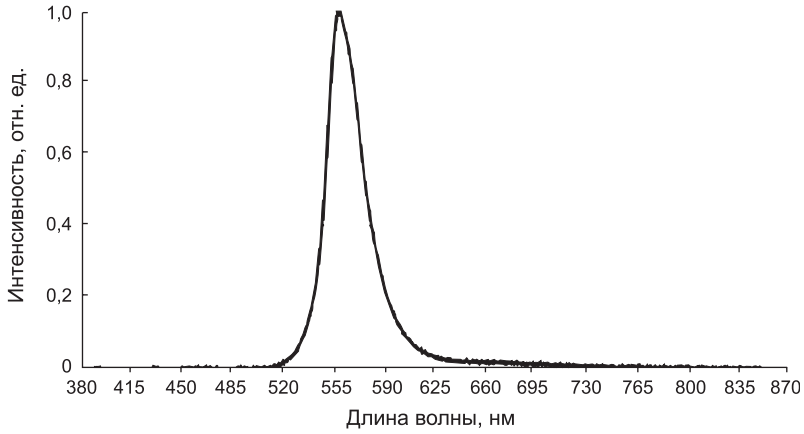

Рис. 8. Спектр излучения имитатора.

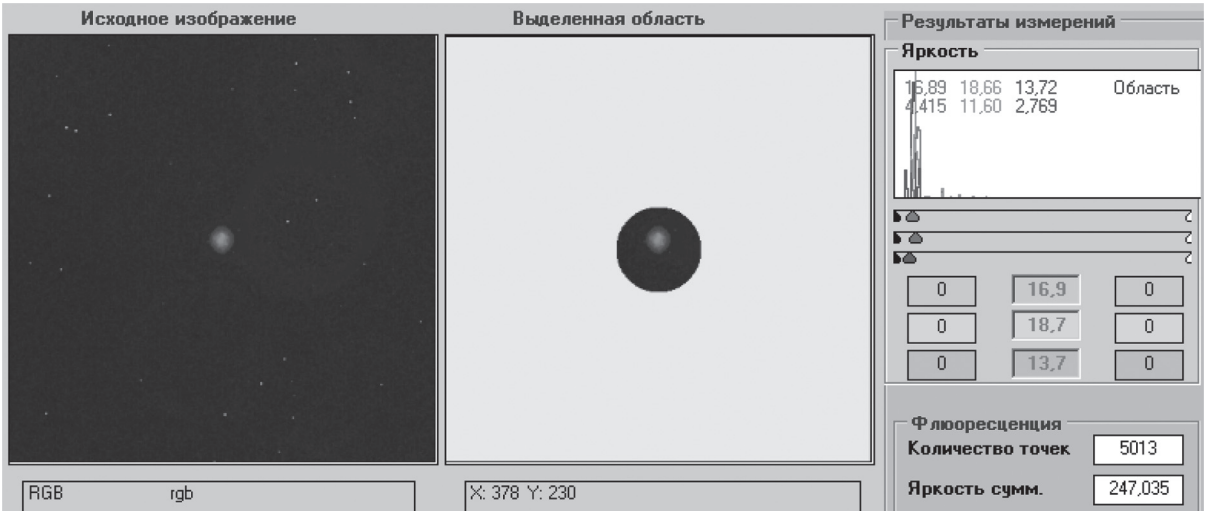

Рис. 9. Изображение светящейся точки с 2 фильтрами.

Ослабление 6400, экспозиция 20 с.

пользован специально изготовленный имитатор источника флюоресценции, спектр которого близок к спектру флюоресценции квантовых точек (рис. 4, 5).

Имитатор представляет собой светодиод с калиброванной диафрагмой диаметром 50 мкм. Он обеспечивает световую мощность $8 \cdot 10^{-9}$ Вт. В связи с низкой мощностью излучения она измерялась методом сравнения и интегрирования спектральных яркостей в заданном диапазоне длин волн. Для получения спектральных характеристик использовался сертифицированный спектрофотометр «Колибри-2» компании «ВМК-Оптоэлектроника». Спектр имитатора приведен на рис. 8.

Светимость имитатора соответствовала освещенности в плоскости изображений ~ 0,1 лк. Такая освещенность примерно в 10 раз превышает предельную чувствительность человеческого глаза. Эталон снимался через 1 и 2 нейтральных светофильтра с пропусканием 1 : 80. Полученные изображения приведены на рис. 9.

ЗАКЛЮЧЕНИЕ

Разработана и апробирована система детекции для функционирования в составе биоаналитического комплекса нового поколения, основанного на соединении метода свободной иммунодиффузии и микрофлюидных технологий. Чувствительность системы составляет $1{,}6 \cdot 10^{-6}$ лк, что позволяет определять наличие антител/антигенов в биологических жидкостях в концентрациях менее чем 0,1 мкг/мл.

ЛИТЕРАТУРА

Пельтек С.Е., Горячковская Т.Н., Банникова С.В. и др. Исследование реакции свободной иммунодиффузии в каналах микрофлюидного модуля // Автометрия. (В печати).

Прэтт У. Цифровая обработка изображений. М.: Мир, 1982. Кн. 1. 312 с.

Цифровое телевидение / Под ред. М.И. Кривошеева. М.: Связь, 1980. 259 с.

Корнышев Н.П., Тимофеев А.В. Компьютерное моделирование телевизионных систем визуализации люминесцирующих малоконтрастных объектов // Вопр. радиоэлектроники. Сер. Техника телевидения. 2007. Вып. 1. С. 43–47.

DETECTION SYSTEM OF A NEW-GENERATION BIOANALYTICAL DEVICE

E.V. Sysoev[3], A.K. Potashnikov[3], Y.V. Obidin[3], T.N. Goryachkovskaya[1], V.S. Bazin[3], V.M. Popik[2], S.E. Peltek[1], N.A. Kolchanov[1, 4, 5]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia;
[2] Budker Institute of Nuclear Physics SB RAS, Novosibirsk, Russia;
[3] Technological Design Institute of Scientific Instrument Engineering SB RAS, Novosibirsk, Russia, e-mail: potash@tdisie.nsc.ru;
[4] Novosibirsk National Research State University, Novosibirsk, Russia;
[5] National Research Centre «Kurchatov Institute», Moscow, Russia

Summary

We designed and manufactured a system for detection of antibodies/antigenes in biological fluids. This system records the kinetics of free immunodiffusion of fluorescent nanocomplexes inside the channels of a microfluid module of a new-generation bioanalytical device. The system consists of four parallel excitation channels, a system for fluorescence detection, and an image-processing program. Antibody/antigen concentrations below 0,1 μg/ml can be detected in biological fluids.

Key words: bioanalytical devices, biosafety, microfluid system.