ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

ОСНОВАН В 1997 г.

Том 18
4/2
Декабрь 2014

VAVILOV JOURNAL OF GENETICS AND BREEDING

FOUNDED IN 1997

Vol. 18
4/2
December 2014

«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. 363-4977*5415

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 18
4/2
Декабрь 2014

Содержание

VAVILOV JOURNAL
OF GENETICS AND BREEDING
FOUNDED IN 1997

Vol. 18
4/2
December 2014

Content

ПРЕДИСЛОВИЕ

Развитие новых высокопроизводительных экспериментальных технологий в области молекулярной биологии и генетики привело к возможности генерации беспрецедентно огромных объемов данных, описывающих особенности работы клетки на молекулярном уровне. В связи с этим критически возрастает роль таких научных направлений, как биоинформатика и системная компьютерная биология, обеспечивающих возможность автоматического конвейерного анализа и интерпретацию получаемых экспериментальных данных, моделирования биологических систем и процессов.

Компьютерная системная биология является быстро растущей междисциплинарной областью научных исследований, объединяющей биоинформатику, информатику, математическое моделирование, микробиологию, молекулярную биологию и генетику, биостатистику и другие области знаний. Для развития современной биоинформатики и компьютерной системной биологии характерна интеграция теоретических, экспериментальных и компьютерных подходов при проведении комплексных исследований.

В настоящем выпуске журнала, сформированном по материалам Международной конференции по биоинформатике регуляции и структуры геномов и системной биологии (BGRS/SB-2014) (http://conf.nsc.ru/BGRSSB2014/), представлены результаты исследований, проводимых в СО РАН по различным направлениям современной биоинформатики и компьютерной системной биологии, а также генетической и метаболической инженерии и биотехнологии, в том числе: биомедицинские исследования с использованием полногеномного анализа; анализ взаимосвязей функциональных и геномных характеристик, аллелей генов с появлением заболеваний; компьютерный анализ и функциональная аннотация сайтов связывания транскрипционных факторов; исследование особенностей экспрессии генов, контролирующих тонус сосудов в почках крыс НИСАГ со стресс-зависимой артериальной гипертензией; анализ структурных и динамических особенностей мутантов белка SOD1, ассоциированных с боковым амиотрофическим склерозом; разработка алгоритма восстановления аминокислотной последовательности циклических пептидов из масс-спектров; применение методов компьютерного анализа графов для поиска структурно-функциональных закономерностей организации генной сети циркадного ритма; исследование структуры и эволюции сетей научного соавторства; моделирование процессов морфогенеза растений; разработка Web-сервисов для оценки эффективности элонгации трансляции генов; комплексные экспериментально-биоинформационные исследования в области генетической и метаболической инженерии и биотехнологии, ориентированные на обработку технологии осахаривания биомассы мискантуса при помощи коммерческих ферментных препаратов; получение этанола из растительной биомассы, теоретические и практические аспекты проблемы биологического окисления углеводородов микроорганизмами; изучение конформационных состояний ДНК в газовой фазе, позволяющее расширить знания о закономерностях компактизации ДНК в естественных и искусственных условиях.

Н.А. Колчанов

Приглашенные редакторы: **Н.Л. Подколодный, Ю.Л. Орлов**

УДК:475.15.275

ПОЛНОГЕНОМНЫЙ АНАЛИЗ ПУЛИРОВАННЫХ ВЫБОРОК ДНК КОГОРТ ЧЕЛОВЕКА

© 2014 г. **В.Н. Бабенко[1, 2], В.Н. Максимов[1, 3], Е.В. Кулакова[1], Н.С. Сафронова[1, 2], М.И. Воевода[1–3], Е.И. Рогаев[1, 4]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: bob@bionet.nsc.ru;
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия;
[3] Институт терапии СО РАМН, Новосибирск, Россия;
[4] Институт общей генетики РАН, Москва, Россия

Поступила в редакцию 10 октября 2014 г. Принята к публикации 27 октября 2014 г.

Цель работы – разработка, валидация и применение подходов приоритетного скрининга точковых полиморфизмов в эксперименте полногеномного аллелотипирования 16 когорт человека, составленных из больных и здоровых индивидов, по схеме «случай – контроль» на основе пулированных выборок. Генотипирование произведено на платформе Illumina Omni1S, имеющей 1,2 млн маркеров на планшете, выборки в среднем составили 200 человек. При сравнении частот аллелей между когортами патологических и контрольных выборок выявляли наборы полиморфизмов, достоверно отличающиеся друг от друга по частоте. Установлено, что выборки больных индивидов показывают систематические, устойчивые отклонения от здоровых по числу достоверно отличающихся полиморфизмов, дисперсия интенсивности аллелей между повторами одной когорты меньше таковой при случайном выборе пар разных когорт.

Ключевые слова: аллелотипирование, пулированная выборка, полногеномный анализ, частота аллеля, точковый полиморфизм.

ВВЕДЕНИЕ

Пулированная, или смешанная, ДНК представляет собой смесь образцов геномной ДНК нескольких индивидов. Она может служить основой для выявления изменения частот точковых полиморфизмов в данных по сравнительному анализу выборок больных и здоровых индивидов, а именно: при достоверном различии частоты точкового полиморфизма между нормой и патологией мы классифицируем данный полиморфизм как кандидатный для дальнейшего рассмотрения.

Применяя технологию анализа пулированной выборки, заметим, что генотипирование невозможно для пулированного материала, поскольку индивидуальные генотипы в смешанном образце не различимы. Поэтому цель такого анализа – оценка частот аллелей по интенсивности сигнала каждого из аллелей в выборке (аллелотипирование). Для этого вполне подходит методология генотипирования на чипе. В настоящее время аллелотипирование пулированных выборок на генотипирующих чипах Illumina широко распространено (Wilkening *et al.*, 2007; Iliadis *et al.*, 2012; Berglund *et al.*, 2013; Gai *et al.*, 2013; Ozerov *et al.*, 2013; Teumer *et al.*, 2013) и доказало свою эффективность в вопросе выявления кандидатных полиморфизмов по схеме «случай – контроль».

С другой стороны, фирма Illumina прогрессивно увеличивает плотность точковых полиморфизмов на планшете, вводя технологически новые платформы и используя новые полиморфизмы, выявленные в проекте «1000 геномов» (1000GP; http://www.1000genomes.org). В частности, массив, использованный в нашей рабо-

те, содержит 1,2 млн маркеров (http://support.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_human_omni1s.pdf). Такое число полиморфизмов является представительной выборкой относительно состояния dbSNP к настоящему времени, что предполагает извлечение новой информации из ранее не известных полиморфных сайтов.

В настоящем исследовании метод пулированного аллелотипирования применен к ряду когорт индивидов Новосибирской области для выявления кандидатных полиморфизмов. Шестнадцать изученных выборок представляли образцы крови больных и здоровых когорт индивидов проекта HAPEE, а также образцы, собранные в Институте терапии РАМН г. Новосибирска (Malyutina *et al.*, 2002; Nicholson *et al.*, 2008; Bobak *et al.*, 2009; Pajak *et al.*, 2013).

В указанных работах дан анализ биометрических, патофизиологических и биохимических медицинских показателей. В предлагаемой работе впервые проведен анализ полногеномного аллелотипирования этих данных, проанализированы их качество и статистические свойства для оценки их пригодности в целях дальнейшего рассмотрения.

МАТЕРИАЛЫ И МЕТОДЫ

Описание выборок

Основная часть выборок собрана в рамках проекта HAPIEE (Health, Alohol and Psychosocial factors In Eastern Europe; http://www.ucl.ac.uk/easteurope/hapiee_open.htm) в г. Новосибирске. Методы построения выборок вкупе с анализом морфометрических, биохимических и психосоциальных данных по данному проекту неоднократно опубликованы (Malyutina *et al.*, 2002; Nicholson *et al.*, 2008; Bobak *et al.*, 2009; Vikhireva *et al.*, 2010; Pajak *et al.*, 2013). Анализ проведен на 16 когортах (табл. 1), краткая аннотация выборок и способ выбора представлены в приложении.

Геномную ДНК выделяли из венозной крови методом фенол-хлороформной экстракции. Концентрацию исходных образцов оценивали на спектрофотометре Nanodrop 1000 (ThermoFisher

Таблица 1

Генотипированные когорты

Когорта	Кол-во человек	Кол-во репликаций
Мальчики	200	2
Девочки	200	2
Мужчины	200	2
Женщины	200	2
Инфаркт миокарда (HAPIEE), мужчины	200	2
Инфаркт миокарда (Куимов), мужчины	200	2
Внезапная смерть, мужчины	200	2
Артериальная гипертензия с метаболическим синдромом		
мужчины	200	2
женщины	200	2
Артериальная гипертензия без метаболического синдрома		
мужчины	200	2
женщины	200	2
Долгожители, женщины	130	2
Инсульт инцидентный ишемический, мужчины	60	2
Глаукома, мужчины и женщины	149	2
Пилоты, мужчины	202	2
Синдром Вольфа – Паркинсона – Уайта, мужчины и женщины	81	2

Scientific Inc., USA). Затем оцененные концентрации ДНК были выравнены до концентрации 50 нг/мкл. Качество ДНК-образцов проверено с помощью ПЦР. После этого 10 мкл каждого образца были взяты для пулирования. Для анализа на чипе отобрано 1 мкг пулированной ДНК с концентрацией 50 нг/мкл. Две реплики для каждой когорты были собраны и пулированы (объединены) независимым образом. Препарат приготовлен в соответствии с протоколом изготовителя (www.illumina.com).

Генотипирование проводили в центре «Биоинженерия» РАН (http://www.biengi.ac.ru/) на платформе Illumina Omni1S-8v1H12 (www.illumina.com), содержащей 8 планшетов по 1,2 млн маркеров. Были использованы 4 чипа, таким образом, проанализировано 32 образца пулированной ДНК. Полученные предварительные данные были процессированы с помощью программы Illumina GenomeStudio 2011.1 (http://support.illumina.com/downloads.ilmn) и размещены в СУБД Mysql v.6.3. (www.mysql.com).

Из 32 экспериментов были извлечены выборки полиморфизмов с частотами в диапазоне [0,01 … 0,99], средняя численность таких полиморфизмов на планшет составила 830 тыс. Затем были скомпилированы целевые выборки 16 когорт, в которых в качестве частоты аллеля взята средняя двух реплик. Средняя численность этих полиморфизмов на когорту 560 тысяч. (Обе частоты не ниже 0,01.) После этого произведено попарное сравнение выборок когорт по частотам полиморфизмов.

Достоверность попарных различий частот оценивали с помощью критерия χ^2 для четырехпольной таблицы, составленной как

$(n_{11} = N_1 p_1, n_{12} = N_1(1-p_1), n_{21} = N_2 p_2, n_{22} = N_2(1-p_2))$:

$$\chi^2 = \frac{N(n_{11}n_{22} - n_{21}n_{12})^2}{(n_{11}+n_{12})(n_{11}+n_{21})(n_{12}+n_{22})(n_{21}+n_{22})}$$

где p_1, p_2 – частоты B аллелей полиморфизма в выборке 1 и 2, N_1, N_2 – объемы выборок 1 и 2.

В качестве внешних контрольных частот взяты частоты соответствующих полиморфизмов выборки полиморфизмов 360 европейских представителей проекта «1000 геномов» в популяциях GBR, TSI, FIN, CEU (www.1000genomes.org). Для внутреннего контроля взяты когорты девочек, мальчиков, мужчин и женщин. Данные об аннотированных полиморфизмах в базах данных OMIM, GENEREVIEWs взяты из dbSNP v.137 с указанием соответствующей аннотации в системе фильтров dbSNP (http://www.ncbi.nlm.nih.gov/snp/limits), насчитывающих 18 291 и 34 373 полиморфизмов соответственно. Таблица о полиморфизмах, найденных в проектах полногеномного ассоциативного анализа (Genome Wide Association Studies, GWAS), взята на сайте базы данных Университета Калифорнии Санта-Круз (www.genome.ucsc.edu).

Оценка LD парного сцепления по полиморфизмам осуществлена программой plink (Purcel *et al.*, 2007) на европейской популяции (см. выше) с параметром $r^2 > 0{,}8$. Половые хромосомы в случае мужчин обработаны по отдельному сценарию.

РЕЗУЛЬТАТЫ

Валидация исходных данных

Внутри- и межпопуляционная дисперсия интенсивности аллелей

Результатом генотипирования по отдельной точке на платформе Illumina является значение двумерного вектора (X, Y) (www.illumina.com/genotyping.ilmn). Значения вектора (N, 0) и (0, N) соответствуют гомозиготным генотипам, а (N/2, N/2) – гетерозиготе, где N – значение сигнала на красной или зеленой «бусинах», соответствующих аллелям. В общем случае сигнал, называемый интенсивностью, имеет целые значения (N1, N2). В целом на анализируемой платформе Illumina значения интенсивности N менялось от 0 до 62 000. Значение X было, как правило, выше значения Y, видимо, в силу выбора b аллеля на платформе в сторону минимального.

Мы применили дисперсионный F-тест для анализа внутри- и межпопуляционных различий для оценки конкордантности реплик по отношению к случайным данным (пары разных когорт). В качестве внутрипопуляционных данных взяты соответствующие двумерные Raw координаты выявленных гетерозиготных генотипов каждой из когорт по каждому маркеру. Для межпопуляционных различий были случайно выбраны 16 пар реплик без совпадения по когорте. По каждой паре выборок вычислено

суммарное отличие всех общих полиморфизмов отдельно по осям *x* и *y* (табл. 2). В результате мы получили достоверное различие ($p < 0,015$) по абсциссе (*x*) и недостоверное, хотя систематическое, отличие ($p < 0,12$) по ординате (*y*) между указанными парами, подтверждающее систематическое отличие в меньшую сторону внутрипопуляционных различий (табл. 2).

Другим критерием валидации было попарное различие мужчин и женщин одной когорты, где были представлены и мужчины, и женщины, и случайный выбор мужчин и женщин разных когорт. F-критерий показал достоверное отличие от случайного ($p < 0,014$) и здесь.

Попарное сравнение когорт по частотам аллелей

После компилирования усредненных значений частот аллелей мы провели попарное сравнение между когортами. Для этого определили матрицу попарных различий как число достоверно различающихся полиморфизмов (см. Материалы и методы) относительно контрольной (мужчины или женщины) когорты. В качестве иллюстрации на рис. 1 даны числа статистически отличных по частоте (СОПЧ) полиморфизмов относительно соответствующих контрольных групп (мужчины и женщины) для 14 когорт. Видно, что значительное число СОПЧ-полиморфизмов относительно контрольных когорт наблюдается (в порядке убывания) у когорт внезапной смерти (ВС), глаукомы и синдрома Вольфа – Паркинсона – Уайта (Wolff – Parkinson – White, WPW), при этом когорта девочек также характеризуется высокими отличиями от женской и мужской контрольных когорт по числу СОПЧ-полиморфизмов. Характерную особенность составляет систематическое, хотя и небольшое, превышение числа различий СОПЧ-полиморфизмов при сравнении с контрольной когортой женщин практически у всех когорт относительно сравнения с контрольной когортой мужчин (столбик «женщины» в гистограмме систематически выше), что, видимо, обусловлено наличием двух Х-хромосом при оценке полиморфизма у женщин. Малое число СОПЧ-полиморфизмов в когорте инсульт инцидентный ишемический обусловлено сравнительно малым размером выборки (см. табл. 1). Рисунок 1 показывает систематические отличия, по крайней мере по количеству найденных кандидатных СНП, при сравнении с независимыми выборками контроля.

Для визуализации распределения 16 когорт использована матрица попарных отличий по числу СОПЧ-полиморфизмов для построения графика многомерного шкалирования (МШ) (рис. 2). Видно, что контрольные когорты оказались в левой половине графика (девочки, мальчики, мужчины, женщины). Положение когорты инцидентного инсульта, также попавшей в левую половину, можно считать незначимой вследствие значительно меньшего объема выборки и, следовательно, минимального числа СОПЧ-полиморфизмов (см. рис. 1), делающей данную выборку «аутлайером». Следует заметить, что девочки и женщины расположены на графике достаточно далеко, в то время как мужчины и мальчики расположены относитель-

Таблица 2

F-тест по сравнению внутри- и межпопуляционных различий сигнала Raw по осям *x* и *y*

Показатель	Raw *x* различия		Raw *y* различия	
	наблюдаемые	ожидаемые	наблюдаемые	ожидаемые
Среднее	651,8	1 102,3	634,4	803,1
Дисперсия	44 737,6	143 406,7	63 672,7	11 8123,2
Наблюдения	16	16	16	16
Df	15	15	15	15
F	0,3119		0,539	
P (F ≤ f) одностороннее	0,015		0,12	
F критическое одностороннее	0,41		0,41	

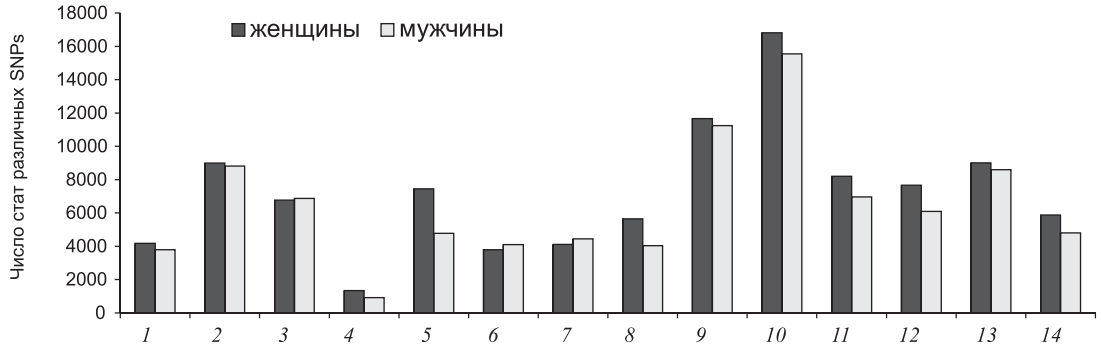


Рис. 1. Количество статистически достоверных отличий полиморфизмов относительно контрольной выборки мужчин и женщин в 14 когортах.

1 – мальчики; *2* – девочки; *3* – долгожители, женщины; *4* – инсульт инцидентный ишемический, мужчины; *5* – артериальная гипертензия (АГ) с метаболическим синдромом (МС), мужчины; *6* – АГ с МС, женщины; *7* – АГ без МС, женщины; *8* – АГ без МС, мужчины; *9* – глаукома, мужчины и женщины; *10* – внезапная смерть, мужчины; *11* – инфаркт миокарда (Куимов), мужчины; *12* – инфаркт миокарда (HAPIEE), мужчины; *13* – синдром Вольфа – Паркинсона – Уайта, мужчины и женщины; *14* – пилоты, мужчины.

но близко, что, по всей видимости, свидетельствует о смещении выборки девочек, вероятно, в силу некоторого отличия концентрации ДНК. Наибольшее отличие от контрольных групп по числу СОПЧ-полиморфизмов, как и на рис. 1, имеют: а) ВС, мужчины, б) синдром WPW и в) глаукома.

Рисунки 1 и 2 иллюстрируют результат выполнения первой фазы выявления кандидатных полиморфизмов визуализацией распределения СОПЧ-полиморфизмов; можно сделать выводы о том, что в среднем несколько тысяч полиморфизмов в каждой выборке достоверно отличаются от контроля. Различия количества СОПЧ-полиморфизмов при смене контрольной существенны. Мы решили также проверить, насколько часто они повторены в обеих контрольных выборках.

Для пяти патологий с наиболее выраженным числом отличий (см. рис. 1) мы провели валидацию оценки числа общих при разных контрольных выборках (мужчины, женщины) СОПЧ-полиморфизмов (табл. 3). Видно, что в среднем 25–38 % отличий систематически повторены относительно обоих контролей (мужчин и женщин). Ожидаемая вероятность случайного совпадения одного маркера при выборе 9 тыс. из 500 тыс. маркеров ($p = 0{,}018$) равна $p = 0{,}0003$. Систематическое отклонение оценено как число совпадающих у 14 когорт СОПЧ-полиморфизмов по отношению к двум контрольным когортам (мужчин и женщин), таких полиморфизмов найдено не было. При числе достоверных различий между мужчинами и женщинами 6 700 из 500 000, и, следовательно, 30 % (около 3 000) совпадений, данной вероятностью можно пренебречь.

ОБСУЖДЕНИЕ

В работах по анализу данных при пулированном способе нанесения материала ставятся две задачи: 1) валидация и адекватная оценка частот аллелей по данным эксперимента и 2) поиск наиболее вероятных кандидатных полиморфизмов, в нашем случае по схеме «случай – контроль». При валидации полиморфизмов одним из вариантов является проверка гаплотипического конформизма, использованная, в частности, Gaj и соавт. (2012). При такой валидации одним из критериев к кандидатным полиморфизмам было отклонение от контроля не менее двух сцепленных полиморфизмов, включая кандидатный. В этой работе контрольные гаплотипы и частоты восстановлены по выборке нативной (польской) популяции, на основе индивидуального полногеномного генотипирования (секвенирования). К сожалению, индивидуальных полногеномных данных по российской (сибирской) популяции достаточного объема у нас не было, поэтому ис-

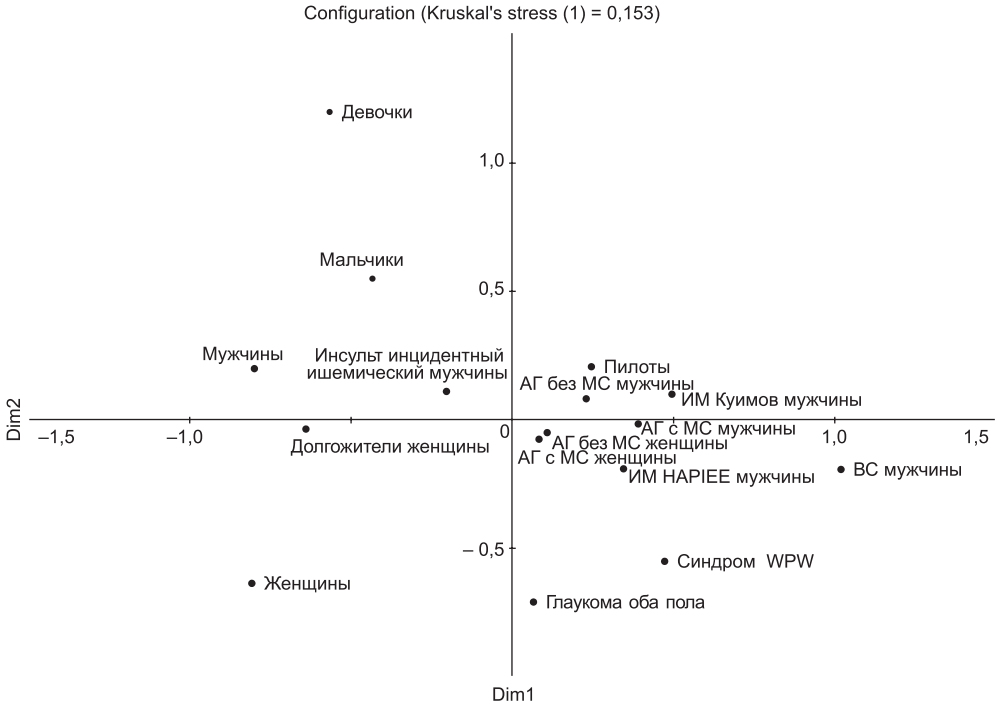


Рис. 2. График многомерного шкалирования матрицы попарных различий частот аллелей когорт.

Таблица 3

Распределение СОПЧ-полиморфизмов по выборкам пяти патологических когорт относительно контрольной выборки женщин (КВЖ) и мужчин (КВМ)

Кол-во СОПЧ-полиморфизмов	Всего	Общие, в КВЖ и КВМ*
Внезапная смерть		
КВЖ	16 820	5 473 (33 %)
КВМ	15 550	
Глаукома		
КВЖ	11 672	3 564 (32 %)
КВМ	11 244	
Инфаркт миокарда (HAPIEE)		
КВЖ	7 663	1 519 (25 %)
КВМ	6 089	
Синдром WPW		
КВЖ	9 009	3 254 (38 %)
КВМ	8 596	

* В процентах по отношению к предыдущему столбцу.

пользована выборка европейцев для выявления сцепленных по r^2 полиморфизмов, содержащих кандидатные. Мы идентифицировали их с помощью программы plink (Purcell *et al.*, 2007) по европейской популяции из проекта «1000 геномов» (Abecasis, Auton, 2012). Установлено, что примерно 10 % кандидатных полиморфизмов значимо сцеплены между собой ($r^2 > 0,8$), при этом их частоты имеют достоверно меньшую дисперсию, чем при отсутствии сцепления, что подтверждает состоятельность оценки частот для пулированных образцов на чипе. Рассматривая

обогащение различных геномных сегментов полиморфизмами из GWAS, стоит отметить, что, как и ожидалось, наибольшее количество ассоциированных с признаком полиморфизмов расположено в экзонных областях (плотность превышена в 2,6 раза по сравнению с контролем). При анализе взаимодействия SNP-eQTL или влияния полиморфизма на экспрессию генов стоит отметить, что она практически полностью соответствует распределению GWAS-полиморфизмов: наибольшую роль играют экзонные состояния, и потом все остальные.

В ряде случаев мы наблюдали совместный, противоположный эффект меж- (внутригенных) полиморфизмов на несколько соседних генов, что, по-видимому, обусловлено модулированием хроматинового состояния доменов ДНК. Доля СОПЧ-полиморфизмов, имеющих известную eQTL-ассоциацию по внешним данным (2,1 %), недостоверно превышает долю таковых в общем пуле полиморфизмов (1,1 %). Тем не менее мы планируем использовать эту информацию при дальнейшей селекции кандидатных полиморфизмов.

ЗАКЛЮЧЕНИЕ

Стратегия выявления кандидатных полиморфизмов по пулированным данным при наличии множества когорт показывает систематичность отличий когорт больных индивидов от когорт здоровых по числу СОПЧ-полиморфизмов. Безусловно, этот вывод основан на обобщенных характеристиках когорт и условном показателе соотношений СОПЧ-полиморфизмов. Он не исключает как возможные систематические смещения по точному соотношению ДНК пулированных индивидов, вариации концентрации ДНК в различных выборках, так и возможное выпадение качества генотипирования на отдельных полиморфизмах. Работа по оценке указанных факторов продолжается.

БЛАГОДАРНОСТИ

Работа поддержана бюджетным проектом ИЦиГ СО РАН VI.61.1.2, грантом Правительства Российской Федерации 14.B25.31.0033 и грантом РФФИ 14-04-01906.Вычисления производились в суперкомпьютерном центре ИВМиМГ СО РАН (www.sscc.ru).

Авторы благодарны Елене Пивоваровой и Татьяне Колесниковой, а также рецензентам за ценные замечания.

ЛИТЕРАТУРА

Abecasis G.R., Auton A. *et al.* An integrated map of genetic variation from 1,092 human genomes. 1000 Genomes Project Consortium // Nature. 2012. V. 491. No. 7422. P. 56–65.

Berglund E.C., Lindqvist C.M., Hayat S. *et al.* Accurate detection of subclonal single nucleotide variants in whole genome amplified and pooled cancer samples using HaloPlex target enrichment // BMC Genomics. 2013. No. 14. P. 856.

Bobak M., Richards M., Malyutina S. *et al.* Association between year of birth and cognitive functions in Russia and the Czech Republic: cross-sectional results of the HAPIEE study // Neuroepidemiology. 2009. V. 33. No. 3. P. 231–239.

Gaj P., Maryan N., Hennig E.E. *et al.* Pooled sample-based GWAS: a cost-effective alternative for identifying colorectal and prostate cancer risk variants in the Polish population // PLoS One. 2012. V. 7. No. 4. P. e35307.

Iliadis A., Anastassiou D., Wang X. Fast and accurate haplotype frequency estimation for large haplotype vectors from pooled DNA data // BMC Genet. 2012. V. 13. P. 94.

Malyutina S., Bobak M., Kurilovitch S. *et al.* Relation between heavy and binge drinking and all-cause and cardiovascular mortality in Novosibirsk, Russia: a prospective cohort study // Lancet. 2002. V. 360. No. 9344. P. 1448–1454.

Nicholson A., Pikhart H., Pajak A. *et al.* Socio-economic status over the life-course and depressive symptoms in men and women in Eastern Europe // J. Affect. Disord. 2008. V. 105. No. (1–3). P. 125–136.

Ozerov M., Vasemägi A., Wennevik V. *et al.* Cost-effective genome-wide estimation of allele frequencies from pooled DNA in Atlantic salmon (Salmo salar L.) // BMC Genomics. 2013. V. 14. P. 12.

Pajak A., Szafraniec K., Kubinova R. *et al.* Binge drinking and blood pressure: cross-sectional results of the HAPIEE study // PLoS One. 2013. V. 8. No. 6. P. e65856.

Purcell S., Neale B., Todd-Brown K. *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis // American J. Human Genetics. 2007. V. 81. No. 3. P. 559–575.

Teumer A., Ernst F.D., Wiechert A. *et al.* Comparison of genotyping using pooled DNA samples (allelotyping) and individual genotyping using the affymetrix genome-wide human SNP array 6.0 // BMC Genomics. 2013. V. 14. P. 506.

Vikhireva O., Pikhart H., Pajak A. *et al.* Non-fatal injuries in three Central and Eastern European urban population samples: the HAPIEE study // Eur. J. Public Health. 2010. V. 20. No. 6. P. 695–701.

Wilkening S., Chen B., Wirtenberger M. *et al.* Allelotyping of pooled DNA with 250 K SNP microarrays // BMC Genomics. 2007. V. 8. P.77.

ПРИЛОЖЕНИЕ

Краткая аннотация выборок

Мальчики Девочки	14–17 лет, учащиеся 9–11 классов общеобразовательных школ Октябрьского района (обследовано 10 % от всех учащихся этих классов в этом районе)
Мужчины Женщины	45–69 лет (HAPIEE)
Инфаркт миокарда (HAPIEE), мужчины	Группа больных инфарктом миокарда была сформированы на основе популяционной выборки 45–69-летних жителей Октябрьского и Кировского районов г. Новосибирска (9 400 человек), которая была собрана НИИ терапии СО РАМН в ходе работы по международному проекту HAPIEE (Health, Alcohol and Psychosocial factors In Eastern Europe). Программа исследования включала: измерение артериального давления, антропометрия (рост, вес, объем талии, бедер), социально-демографические характеристики, опрос о курении, потреблении алкоголя (частота и типичная доза), уровне физической активности, оценку липидного профиля (общий холестерин; триглицериды, холестерин липопротеидов высокой плотности), опрос на выявление стенокардии напряжения (Rose), ЭКГ покоя в 12 отведениях
Инфаркт миокарда (Куимов), мужчины	В исследование включены больные ОКС, поступившие в блок интенсивной терапии городской клинической больницы № 1 Новосибирска с 01 апреля 2009 по 30 марта 2010 г. (средний возраст 59,1 ± 6,1 года), в том числе с ОКС с подъемом сегмента ST 180 человек (117 мужчин) и 58 человек с ОКС без подъема сегмента ST. Средний возраст мужчин 56,2 ± 5,2 года. Диагноз ОКС установлен по совокупности критериев, разработанных Европейским обществом кардиологов и Американской коллегией кардиологов (2000), включающих: а) типичный болевой приступ, б) изменения ЭКГ в двух и более последовательных отведениях (высокоамплитудный Т, отрицательный Т, подъем сегмента ST, патологический Q, депрессия сегмента ST, наличие QR), в) динамические изменения в уровне ферментов (КФК, КК-МВ, ТнТ, ТнI)
Внезапная смерть, мужчины	Набор аутопсийного материала проводился у мужчин, умерших внезапно в возрасте 25–64 лет (жителей Октябрьского района Новосибирска), подвергнутых судебно-медицинскому исследованию. Средний возраст умерших составил 53,6 ± 7,9 года. При секционном исследовании производили забор образцов ткани печени или миокарда в количестве 5–10 г. Мужчины, умершие внезапно (без морфологических изменений характерных для инфаркта миокарда, дилатационной и гипертрофической кардиомиопатий и др.). С учетом ограниченной информации о времени развития фатального события в исследуемую группу включены случаи смерти, развившейся в течение 1 ч или при отсутствии свидетелей смерти в течение не более 24 ч и расцененных по данным аутопсии как смерть сердечного генеза
АГ с МС, мужчины	45–69 лет HAPIEE
АГ без МС, мужчины	45–69 лет HAPIEE
АГ с МС, женщины	45–69 лет HAPIEE
АГ без МС, женщины	45–69 лет HAPIEE
Геронты, женщины	90 лет и старше
Инсульт инцидентный ишемический, мужчины	Инсульт в HAPIEE 45–69 лет (инсульт, случившийся после прохождения скрининга). Данные предоставлены регистром инсульта
Глаукома, мужчины и женщины (ПОУГ)	79 женщин, 70 мужчин, 41–85 лет, средний возраст 67,2 года
Пилоты, мужчины	Действующие пилоты гражданской авиации, 35–55 лет

Синдром Вольфа – Паркинсона – Уайта, мужчины и женщины	Группа представлена пациентами с синдромом WPW в количестве 81 человек, среднего возраст 32,0 ± 1,6 года, из которых 43 были мужчинами (53,0 %) и 38 женщинами (47,0 %). Они были выбраны из группы пациентов с пароксизмальными реципрокными атриовентрикулярными тахикардиями, пролеченных в отделении сердечно-сосудистой хирургии ГУЗ дорожной клинической больницы на ст. Новосибирск-главный, отделениях нарушений ритма Областного кардиодиспансера и НИИ патологии кровообращения им. Е.Н. Мешалкина г. Новосибирска в течение 2000–2003 гг. Критериями включения были: отсутствие ИБС, врожденных и приобретенных пороков сердца, перенесенных острых мио- и перикардитов различной этиологии в течение не менее одного месяца до обследования, нарушения функции щитовидной железы, онкологических заболеваний и беременности

GENOME-WIDE SNP ALLELOTYPING OF HUMAN COHORTS BY POOLED DNA SAMPLES

V.N. Babenko[1], V.N. Maximov[1, 3], E.V. Kulakova, E.S. Safronova[1, 2], M.I. Voevoda[1–3], E.I. Rogaev[1, 4]

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

Summary

The research concerns the task of identification of contrasted single nucleotide polymorphisms (SNPs) obtained in genome-wide pooled allelotyping of 16 human cohorts, comprising healthy and ill persons, by the nested case–control approach. The genotyping platform was the Illumina Omni1S chip with 1.2 million markers. The mean pooled sample size was about 200 individuals. The candidate selection was based on statistical comparison of allele frequencies in a "case–control" study. Samples of ill patients show significant deviations from healthy persons in the numbers of significantly differing polymorphisms. The variance of allele frequencies among repeats in a single cohort was less than that in random choice of pairs from different cohorts.

Key words: genome-wide genotyping, pooled sample, allele frequency, SNP.

УДК 575.8.57.017.73

АНАЛИЗ ВЗАИМОДЕЙСТВИЯ АЛЛЕЛЕЙ ГЕНОВ ЛИПИДНОГО ОБМЕНА ПРИ ДИСЛИПИДЕМИИ

© 2014 г. **И.В. Николаев, Р.В. Мулюкова, Л.Р. Каюмова, Е.В. Воробьева, В.Ю. Горбунова**

Башкирский государственный педагогический университет им. М. Акмуллы, Уфа, Россия, e-mail: obg_bspu@mail.ru

Поступила в редакцию 8 сентября 2014 г. Принята к публикации 15 сентября 2014 г.

В клинической практике часто используют показатели липидного обмена. К его нарушениям (дислипидемии) относят повышение уровня общего холестерина (ОХС), триглицеридов (ТГ), а также изменение ряда других показателей, являющихся результатом нарушения синтеза, транспорта и расщепления липопротеинов. Клиническая значимость метаболических нарушений, объединенных рамками дислипидемии, ассоциируется, в первую очередь, с высоким риском развития сердечно-сосудистых заболеваний, сахарного диабета второго типа и ожирения. Исследована связь *SNP* следующих генов: *G-2548A* в промоторной области гена лептина (*LEP*), *A223G* в 4-м экзоне гена рецептора лептина (*LEPR*), *T495G* в 8-м интроне гена липопротеинлипазы (*LPL*), *C34G* в 8-м экзоне гена ядерного рецептора (*PPARG*), с нарушениями липидного обмена, и показан их кумулятивный эффект в развитии дислипидемии.

Ключевые слова: дислипидемия, лептин, рецептор лептина, липопротеинлипаза, ядерный рецептор.

ВВЕДЕНИЕ

Дислипидемия (гиперлипидемия) – аномально повышенный уровень липидов (липопротеинов) и/или нарушение их соотношения. Выделяют два типа дислипидемий: первичные, генетически обусловленные, и вторичные, приобретенные в результате приема каких-либо лекарственных препаратов либо вследствие болезни. Гиперлипидемия считается важным фактором риска развития сердечно-сосудистых заболеваний, в основном, из-за ее связи со значительным влиянием холестерина на развитие атеросклероза (Ребров, Гайдукова, 2010).

Липиды попадают в организм, главным образом, в форме триглицеридов жирных кислот. В кишечнике под действием ферментов поджелудочной железы они подвергаются гидролизу, продукты которого всасываются клетками стенки кишечника. Здесь из них вновь синтезируются нейтральные жиры, которые через лимфатическую систему поступают в кровь и либо транспортируются в печень, либо отлагаются в жировой ткани (Марри и др., 1993). В контроль и реализацию процесса метаболизма липидов вовлечено большое число генов и их продуктов. В нашем исследовании были изучены локусы генов четырех наиболее важных из них: генов лептина и его рецептора *LEP* и *LEPR*, липопротеинлипазы *LPL* и γ-рецептора, активируемого пролифератором пероксисом *PPARG*.

Ранее было показано, что SNP G-2548A (rs7799039; 7:128238730), локализованный в промоторе гена лептина *LEP*, при замене гуанина на аденин может приводить к уменьшению концентрации лептина и неспособности жировой ткани секретировать этот гормон (Мельниченко, 2001; Wang *et al.*, 2006). По современным представлениям, лептин стимулирует окисление свободных жирных кислот в митохондриях (Schulze, Kratzsch, 2005). Он изменяет метаболизм жиров и глюкозы, а также регулирует нейроэндокринную функцию. Лептин может либо оказывать прямое влияние, либо активировать лептиновые рецепторы в гипоталамусе, которые изменяют экспрессию

нейропетидов и приводят к снижению аппетита, повышению расхода энергии за счет изменения тонуса симпатической системы и обмена веществ в периферических органах и тканях (Schwartz, Seeley, 1997; Мельниченко, 2001; Mantzoros, 2004). SNP A223G (rs1137101, 1:65592830), расположенный в 4-м экзоне гена *LEPR*, может приводить к нарушению рецепции лептина (Park *et al.*, 2006), что увеличивает риск развития ожирения, сахарного диабета второго типа и сердечно-сосудистых заболеваний (Duarte *et al.*, 2007; Constantin *et al.*, 2010).

В нарушении липидного обмена важную роль также играет липопротеинлипаза (LPL) – многофункциональный белок и ключевой фермент метаболизма липидов. Она является основным компонентом триглицерид-насыщенных хиломикронов и липопротеинов очень низкой плотности и играет важную роль в формировании липопротеинов высокой плотности. SNP T495G (rs320, 8:19961566), локализованный в 8-м интроне гена *LPL*, может повышать концентрацию триглицеридов и холестерина (Heinzmann *et al.*, 1991; Ma *et al.*, 2003).

Помимо гидролиза триглицеридов плазмы до диглицеридов липопротеинлипаза также участвует во взаимодействии липопротеинов с ядерными рецепторами (Hayden, Henderson, 1999), в том числе с γ-рецептором, активируемым пролифератором пероксисом (*PPARG*), который определяет дифференциацию адипоцитов и регулирует функционирование генов, связанных с: аккумуляцией жира (синтеза триглицеридов), дифференцировкой адипоцитов и миобластов, чувствительностью к инсулину, активностью остеобластов и остеокластов (Semple *et al.*, 2006). SNP C34G (rs1801282, 3:12351626), находящийся в 8-м экзоне гена *PPARG*, ассоциирован со снижением его транскрипционной активности (Montagner *et al.*, 2011), развитием метаболического синдрома (Jeninga *et al.*, 2009) и сахарного диабета второго типа (Laakso, 2004).

Цель настоящего исследования – комплексный анализ взаимодействия аллелей генов, влияющих на метаболизм липидов: *LEP*, *LEPR*, *LPL*) и *PPARG*, продукты которых являются биологически активными веществами, специфичными для жировой ткани. Были поставлены следующие задачи: сравнение частот генотипов и аллелей полиморфных локусов типированных генов в группах с низким и высоким уровнем ОХС и ТГ в сыворотке крови, а также при стратификации индивидов по полу и возрасту; определение и оценка сочетаний аллелей типированных локусов, способствующих или препятствующих высокому уровню ОХС и ТГ в сыворотке крови; определение потенциального влияния мутаций типированных локусов генов на внутриклеточные сигнальные каскады, *in silico* оценка влияния рассмотренных мутаций на физико-химические свойства продуктов соответствующих генов. Для анализа связи SNP G-2548A в гене *LEP*, *A223G* в гене *LEPR*, T495G в гене *LPL* и C34G в гене *PPARG* с нарушениями липидного обмена в популяции людей, проживающих в Республике Башкортостан, использованы молекулярные, биохимические, статистические и биоинформатические методы.

МАТЕРИАЛЫ И МЕТОДЫ

В работе использованы образцы ДНК 457 (241 мужчина и 216 женщин) практически здоровых лиц, проживающих в Республике Башкортостан. Средний возраст испытуемых составил 23,64 ± 6,87 года (от 18 до 63 лет). Забор крови для выделения ДНК производили после медицинского осмотра и анкетирования испытуемых на предмет наличия хронических заболеваний, с их письменного согласия.

Выборка была поделена на три группы: с высоким, низким и нормальным уровнем ОХС и ТГ в сыворотке крови, на основе рекомендаций Всероссийского научного общества кардиологов (Диагностика и коррекция…, 2009), с модификациями (табл. 1). Работа проведена в Центре молекулярно-генетических исследований Башкирского государственного педагогического университета им. М. Акмуллы.

ДНК выделяли методом фенольно-хлороформной экстракции по Мэтью (Mathew, 1984), полимеразную цепную реакцию синтеза ДНК проводили по Мюллису (Mullis *et al.*, 1985), анализ полиморфизма длин рестрикционных фрагментов – по Лэйнгдалу (Langdahl *et al.*, 1998), электрофорез в 7%-м полиакриламидном геле – по Маниатису (Маниатис и др., 1984). Результаты электрофореза ДНК визуализировали в ультрафиолетовом свете трансиллюминатора

Таблица 1

Градация показателей общего холестерина и триглицеридов в сыворотке крови человека (по рекомендациям ВНОК, 2009 г. с модификациями)

Показатель	Уровень (концентрация), ммоль/л (*n*)		
	низкий	высокий	в пределах физиологической нормы
Общий холестерин	до 3,8 (137)	больше 5,2 (63)	3,8–5,2 (257)
$X \pm m$	3,33 ± 0,43	5,98 ± 0,63	4,45 ± 0,39
Триглицериды	до 0,5 (31)	больше 1,7 (62)	0,5–1,7 (364)
$X \pm m$	0,42 ± 0,09	2,20 ± 0,56	1,05 ± 0,32

*З*десь и далее: *n* – количество человек, *X* – среднее значение в группе, *m* – ошибка среднего арифметического.

Vilber Lourmart TFX-20M. Материалом для проведения биохимических анализов послужила сыворотка венозной крови, без следов гемолиза, очищенная от эритроцитов и соответствующая специальным требованиям. Уровень ОХС и ТГ определяли ферментным методом с использованием реактивов фирмы Cormay (Германия) на анализаторе «Флюорат-02-АБЛФ-Т».

Проведен как качественный (сравнение частот генотипов и аллелей в изученных группах), так и количественный (однофакторный дисперсионный) анализ связи уровня ОХС и ТГ в сыворотке крови с мутациями в типированных полиморфных локусах генов. Для моделирования воздействия дислипидемии первого типа на человека при частотном анализе сравнивали группы индивидов с низкими и высокими показателями ОХС и ТГ в сыворотке крови. Испытуемые были стратифицированы по гендерному признаку и возрасту. При разделении по возрасту выделено две группы: группа 1 – 18–30 лет (385 человек) и группа 2 – 31–63 года (72 человека).

Для оценки количественных различий в выделенных группах был использован *t*-критерий Стьюдента для независимых выборок. В группах индивидов, различающихся по гендерному признаку, сравнивали частоту аллелей и генотипов типированных полиморфных локусов генов.

Стратификацию индивидов по национальному признаку не проводили, так как определение межэтнических различий не входило в задачи исследования. Для более точной оценки влияния каждого из типированных локусов на уровень ОХС и ТГ в сыворотке крови проведен однофакторный дисперсионный анализ (ANOVA) без разделения индивидов на группы. Статистические расчеты проводились в программах Microsoft Exel 2003 (Microsoft Corp., 2002) и Statistica 6.1 (Statsoft Inc, 2007). При попарном сравнении частот генотипов и аллелей в двух различных группах использовали двусторонний критерий Фишера (F). Статистически достоверными считали различия частот аллелей и генотипов при значении $p \leq 0{,}05$.

Межгенное взаимодействие локусов типированных генов оценивали с помощью программы Multifactor Dimensionality Reduction 2.0 (MDR 2.0), основанной на методе логистической регрессии (Moore *et al.*, 2006). При проведении этого вида расчетов использованы данные индивидов с низким и высоким уровнем ОХС и ТГ.

Биоинформатический анализ нуклеотидных последовательностей полиморфных локусов проводили с применением программы поиска сайтов связывания транскрипционных факторов TFSCAN (Rice *et al.*, 2000; http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::tfscan). Информацию об общегеномной локализации SNP получали из dbSNP, интегрированной в состав базы данных GeneBank (http://ncbi.nlm.nih.gov/SNP). Локализацию сайтов связывания транскрипционных факторов, затрагивающих SNP, приводили в соответствие с их общегеномной локализацией на основе данных Консорциума описания генома человека, доступного в базе данных GeneBank.

Нуклеотидные и аминокислотные последовательности анализировали в двух вариантах: для мутантных и нормальных аллелей изученных генов. Информацию об аминокислотной последовательности белка, а также структурных и функциональных доменах, входящих в его состав, получали из базы данных UniProt (http://uniprot.org). Для анализа использовали аминокислотные последовательности канонических изоформ белков. Оценку физико-химических свойств белковых структур проводили в программе ProtParam (Gasteiger *et al.*, 2005; http://web.expasy.org/protparam).

Критериями изменения физико-химических свойств белков считали изменение молекулярной массы, изоэлектрической точки, алифатического индекса, являющегося одним из показателей термостабильности (Ikai, 1980), и индекса нестабильности белка, оценивающего стабильность белка *in vitro*, который у стабильных белков не должен превышать 40 (Guruprasad *et al.*, 1990). При поиске гомологов доменов белков, затронутых мутацией, использовали BLAST-поиск по базе данных Protein Data Bank, с алгоритмом PSI-BLAST (матрица BLOSUM62), реализованным в базе данных GeneBank.

Для оценки конформационных изменений доменов белков под действием изученных мутаций полиморфных локусов генов проводили моделирование пространственных белковых структур при помощи программного комплекса Schrödinger Suite 2013 (Schrödinger Inc., 2013). Файлы-шаблоны для моделирования с данными о кристаллизованных белковых структурах получали в базе данных Protein Data Bank (http://pdb.org). Изменения конформации белковых структур под действием мутаций анализировали в программе Vadar 1.8 (Willard *et al.*, 2003; http://vadar.wishartlab.com/index.html). Критериями изменения конформации считали изменение общего объема домена и его доступной площади.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

При сравнении групп индивидов, различающихся по гендерному признаку, у женщин выявлено небольшое увеличение средних значений уровня общего холестерина в сыворотке крови по сравнению с мужчинами (4,43 против 4,23 ммоль/л) и уменьшение средних значений уровня ТГ (1,14 против 1,17 ммоль/л), однако статистически достоверных значений оно не достигало (табл. 2).

При разделении индивидов с учетом возраста в старшей возрастной группе (31–63 года) выявлено увеличение среднего уровня ОХС (4,60 против 4,28 ммоль/л), однако оно не достигало статистически значимых значений (табл. 2). В этой же возрастной группе отмечено достоверное увеличение среднего уровня ТГ в сыворотке крови в 1,15 раза (1,30 против 1,13 ммоль/л, $p = 0{,}019$). Полученные данные согласуются с гипотезой М.А. Даренской и др. (2006) о возрастном повышении уровня липидов в сыворотке крови и изменении ее реологических свойств.

Таблица 2

Общая характеристика исследуемой выборки

Группы	Общий холестерин, ммоль/л	Триглицериды, ммоль/л
Женщины, $n = 216$	4,43 ± 0,89	1,14 ± 0,46
Мужчины, $n = 241$	4,23 ± 0,97	1,17 ± 0,65
p (χ^2)	> 0,05	> 0,05
Группа 1 (средняя, 18–30 лет), $n = 385$	4,28 ± 0,93	1,13 ± 0,57
Группа 2 (старшая, 31–63 года), $n = 72$	4,60 ± 0,95	1,30 ± 0,54
p (χ^2)	> 0,05	0,019

Анализ ассоциаций исследуемых ДНК-локусов с уровнем общего холестерина и триглицеридов

Сравнительный анализ распределения частот генотипов и аллелей генов *PPARG, LPL, LEP* и *LEPR* показал статистически значимые различия между группами индивидов, разделенными по гендерному признаку. Частота генотипа *LPL*G/*G* в группе женщин была достоверно выше, чем в группе мужчин (0,34 против 0,22, $p = 0{,}008$, $\chi^2 = 7{,}37$). Частота аллеля *LPL*G* у женщин также была достоверно выше (0,59 против 0,53, $p = 0{,}03$, $\chi^2 = 4{,}3$). Этот факт можно объяснить различиями в метаболизме липидов у женщин и мужчин (Després *et al.*, 2000).

PPARG. При анализе распределения частот генотипов и аллелей полиморфного варианта гена *PPARG* с учетом показателя общего холестерина выявлены достоверные различия (табл. 3). В группе лиц с низким уровнем ОХС обнаружено повышение аллеля *PPARG*C* по отношению к группе лиц, имеющих высокий уровень ОХС (0,90 против 0,78, $p = 0{,}009$, $\chi^2 = 6{,}97$), а также понижение аллеля *PPARG*G* (0,10 против 0,22, $p = 0{,}009$, $\chi^2 = 6{,}97$), что может указывать на вовлеченность изученного полиморфного варианта C34G (rs1801282) гена *PPARG* в регуляцию метаболизма липидов.

Как известно, снижение активности PPARγ у носителей аллеля *PPARG*G* подавляет липолиз в адипоцитах, что снижает уровень циркулирующих свободных ЖК и увеличивает утилизацию мышцами глюкозы (Boden, 1997). Кроме того, необходимо отметить тот факт, что исследуемый полиморфный вариант гена *PPARG* снижает транскрипционную активность некоторых генов-мишеней, в том числе гена фактора некроза опухолей α, лептина, резистина, адипонектина (Meirhaeghe *et al.*, 2005). При проведении анализа распределения частот генотипов и аллелей рассматриваемого ДНК-локуса с учетом уровня триглицеридов не выявлено достоверно значимых различий.

LPL. Сравнительный анализ распределения частот генотипов и аллелей полиморфного варианта гена *LPL* с учетом уровня ОХС показал статистически значимые различия (табл. 3). Частота генотипа *LPL*T/*T* в группе индивидов с низким уровнем ОХС в сыворотке крови была достоверно выше, чем в группе с высоким

Таблица 3

Результаты анализа ассоциаций исследуемых ДНК-локусов с показателями общего холестерина и триглицеридов в сыворотке крови (приведены статистически значимые данные)

Генотип/аллель	Уровень ОХС, ммоль/л (*n*)		*p* (χ^2)
	низкий, менее 3,8	высокий, более 5,2	
*PPARG*C*	0,90 ± 0,02 (135)	0,10 ± 0,02 (12)	0,009 (6,97)
*PPARG*G*	0,78 ± 0,02 (62)	0,22 ± 0,02 (34)	
*LPL*T/*T*	0,18 ± 0,03 (24)	0,05 ± 0,03 (3)	0,03 (4,97)
*LPL*T/*G*	0,66 ± 0,04 (91)	0,49 ± 0,06 (31)	0,03 (4,68)
*LPL*G/*G*	0,16 ± 0,03 (22)	0,46 ± 0,06 (29)	0,0005 (18,86)
*LPL*G*	0,49 ± 0,03 (35)	0,71 ± 0,04 (25)	0,0007 (15,13)
*LPL*T*	0,51 ± 0,03 (59)	0,29 ± 0,04 (26)	
	Уровень ТГ, ммоль/л		
	низкий, менее 0,5	высокий, более 1,7	
*LPL*G/*G*	0,19 ± 0,04 (6)	0,44 ± 0,06 (27)	0,04 (4,28)
*LPL*G*	0,52 ± 0,06 (32)	0,69 ± 0,04 (86)	0,03 (4,87)
*LPL*T*	0,48 ± 0,06 (30)	0,31 ± 0,04 (38)	

уровнем (0,18 против 0,05, $p = 0,03$, $\chi^2 = 4,97$). Гетерозиготный генотип *LPL*T/*C* также наблюдался с достоверно большей частотой (0,66 против 0,49, $p = 0,03$, $\chi^2 = 4,68$). Частота генотипа *LPL*G/*G* выше в группе лиц с высоким уровнем ОХС в сыворотке крови (0,46 против 0,16, $p = 0,0005$, $\chi^2 = 18,86$). Частота аллеля *LPL*G* достоверно выше в группе с высокими значениями ОХС в сыворотке крови (0,71 против 0,49, $p = 0,0007$, $\chi^2 = 15,13$).

При проведении анализа распределения частот генотипов и аллелей описываемого полиморфного варианта гена *LPL* с учетом уровня триглицеридов выявлены достоверные различия (табл. 3). В группе лиц с высоким уровнем ТГ отмечено повышение генотипа *LPL*G/*G* (0,44 против 0,19 в группе лиц с низким уровнем ТГ, $p = 0,04$, $\chi^2 = 4,28$) и аллеля *LPL*G* (0,69 против 0,52, $p = 0,03$, $\chi^2 = 4,87$). Как полагают Ма и соавт. (Ма *et al.*, 2003), наличие генотипа *LPL*T/*T* – один из факторов, определяющих физиологически нормальную концентрацию триглицеридов.

LEP и LEPR. Анализ сравнения частот генотипов и аллелей анализируемых ДНК-локусов *LEP* и *LEPR* в группах, разделенных с учетом показателей ОХС и ТГ, не выявил достоверных различий. Поэтому было решено провести количественную оценку влияния аллелей и генотипов изученных генов на уровень ОХС и ТГ в сыворотке крови.

Однофакторный дисперсионный анализ

Проведенный однофакторный дисперсионный анализ (ANOVA) позволил количественно оценить влияние аллелей и генотипов типированных полиморфных локусов генов на уровень рассматриваемых показателей. В результате однофакторного дисперсионного анализа выявлено статистически значимое влияние генотипа *LPL*G/*G* на высокий уровень общего холестерина ($F = 24,16$, $p = 0,0001$) и триглицеридов ($F = 11,97$, $p = 0,001$). Как показано ANOVA, на низкий уровень триглицеридов влияет наличие в генотипе аллеля *LPL*T* ($F = 4,77$, $p = 0,03$). Полученные данные согласуются с исследованиями зарубежных ученых, которые показали ассоциацию мутантного аллеля *LPL*G* с первичной гиперглицеринемией у европейцев и японцев (Kathiresan *et al.*, 2008), а также у белых американцев с высокой концентрацией ОХС в сыворотке крови (Cooper *et al.*, 2007).

При исследовании взаимосвязи полиморфного варианта с показателем триглицеридов выявлены статистически более высокие значения ТГ у обладателей генотипа *LEPR*G/*G* ($F = 8,13$, $p = 0,005$). Вероятно, уровень ТГ в сыворотке крови увеличивается вследствие того, что этот мутантный аллель обусловливает нарушение рецепции лептина, в результате чего у человека плохо работает чувство насыщения (Sun *et al.*, 2010). В целом, данные, полученные в результате сравнительного анализа частот генотипов и аллелей, согласуются с результатами однофакторного дисперсионного анализа, что еще раз подтверждает важную роль описываемых ДНК-локусов в метаболизме липидов, в частности общего холестерина и триглицеридов.

Анализ межгенных взаимодействий

С помощью программы MDR 2.0 была проведена оценка характера взаимодействия полиморфных локусов: G-2548A гена *LEP*, A223G гена *LEPR*, T495G гена *LPL*, C34G гена *PPARG* с высокими показателями ОХС и ТГ. Были выбраны оптимальные, статистически значимые, четырехфакторные модели взаимодействия аллелей генов липидного обмена (*LEP, LEPR, LPL, PPARG*). Отобраны сочетания генотипов типированных локусов генов, приводящие к высокому уровню ОХС (> 5,2 ммоль/л) и ТГ (> 1,7 ммоль/л) в сыворотке крови, а также препятствующие развитию этого признака во всей рассмотренной выборке индивидов (табл. 4).

Анализ результатов работы программы MDR 2.0 показал, что наибольшее влияние на появление высокого уровня ОХС и ТГ оказывают аллели генов лептина *LEP* и липопротеинлипазы *LPL*. Сочетания генотипов, вовлеченных в регуляцию уровня ОХС, различались между собой мутантным аллелем гена липопротеинлипазы *LPL*G* – фермента, осуществляющего функцию синтеза липопротеинов высокой плотности из липопротеинов низкой плотности только при нормальной концентрации этого фермента в сыворотке крови (Ма *et al.*, 2003). Анализ графического изображения модели взаимодействия локусов (рисунок, *а*), определяющих высокий

Таблица 4

Характеристика моделей межгенных взаимодействий исследуемых полиморфных ДНК-локусов

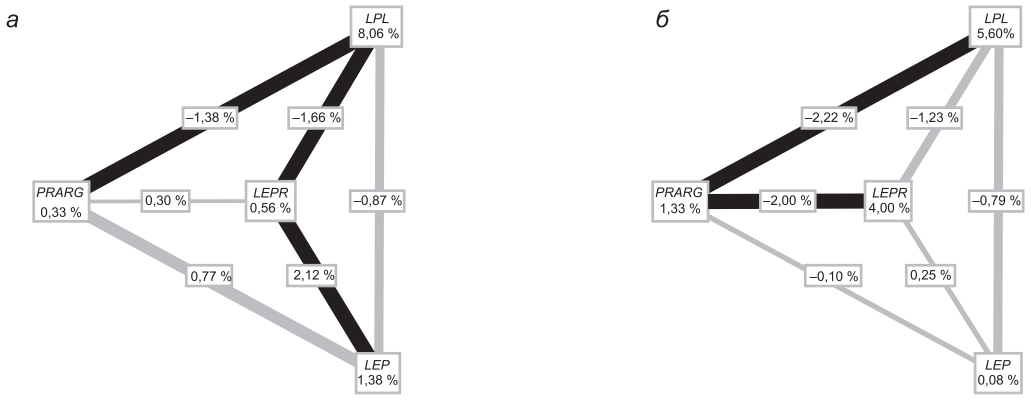
Признак	Tr. Bal. Acc	Ts. Bal. Acc	Se	Sp	CV Cons	Общий p (χ^2)	OR	Влияние
Высокий ОХС	*LEP*A/*G, LEPR*A/*G, PPARG*C/*C, LPL*G/*G*							+
	*LEP*A/*G, LEPR*A/*G, PPARG*C/*C, LPL*G/*T*							–
	0,77	0,63	0,78	0,76	10/10	< 0,0001 (51,43)	11,03 (5,42–22,47)	
Высокий ТГ	*LEP*A/*G, LEPR*A/*G, PPARG*C/*C, LPL*G/*T*							+
	*LEP*G/*G, LEPR*A/*G, PPARG*C/*C, LPL*G/*T*							–
	0,76	0,50	0,69	0,81	10/10	< 0,0001 (20,73)	9,43 (3,33–26,73)	

Примечание. Tr. Bal. Acc – тренировочная сбалансированная точность; Ts. Bal. Acc – тестируемая сбалансированная точность; Se – чувствительность модели, Sp – специфичность; CV Cons – повторяемость пересчетов; OR – отношение шансов; +/– – способствующее/препятствующее развитию признака сочетание генотипов.

уровень ОХС, показал наличие дублирующего эффекта, т. е. эффекта полимерии между локусами генов *LEPR*, *PPARG*, *LPL*, *LEP*, и эффекта синергизма, т. е. комплементарности между локусами генов *LEP*, *LEPR* и *PPARG*.

Сочетания генотипов, вовлеченных в регуляцию уровня ТГ в сыворотке крови, различаются по наличию мутантного аллеля гена *LEP*A,* который присутствует в сочетании, способствующем высокому уровню ТГ. Анализ графического изображения модели взаимодействия локусов (рисунок, *б*), определяющих высокий уровень ТГ, показал наличие связи между локусами генов *LEPR*, *PPARG*, *LPL* и *LEP* по типу полимерии. Выявлено, что сочетание локусов генов, определяющих высокий уровень ОХС в сыворотке крови, отличается от такового сочетания, определяющего высокий уровень ТГ, присутствием мутантного аллеля гена *LPL*G*. При этом генотип, принятый программой MDR 2.0 за препятствующий высокому уровню ОХС в исследованной выборке, способствует высокому уровню ТГ в сыворотке крови. Очень часто проявление гиперхолистеринемии и триглицеридемии сопряжены между собой. Вероятно, возможен кумулятивный эффект действия мутантных аллелей генов *LPL* и *LEP*, что может привести к манифестации первичной дислипидемии. В исследованных генах изучены сайты связывания транскрипционных факторов и показано, что в результате мутации в гене *LEPR* (SNP A223G) формируется слабой силы потенциальный сайт связывания рецептора эстрогена первого типа ESR1 (GGTCA, 1:65592830-65592834 последовательности генома человека). Некоторые исследования подтверждают вовлеченность этого транскрипционного фактора в регуляцию транскрипции рецептора лептина, развитие дислепидемии и метаболического синдрома (Barros, Gustafsson, 2011; Faulds *et al.*, 2012).

Анализ локуса гена γ-рецептора, активируемого пролифераторами пероксисом *PPARG*, несущего SNP C34G, выявил, что в случае мутации данного локуса исчезает средней силы потенциальный сайт связывания транскрипционного фактора CREB1 (TGACG, 3:12351622-12351626 последовательности генома человека). CREB1 вовлечен в регуляцию работы большого числа генов. Он своеобразный переключатель многих молекулярных процессов внутри клетки (Bartsch *et al.*, 1998). Возможно, отсутствие данного сайта связывания CREB1 приводит к описанному типу манифестации SNP C34G гена



Графическое изображение результатов анализа взаимодействий между генами при отклонении от физиологической нормы уровня общего холестерина (*а*) и триглицеридов (*б*). Толщина линий графа демонстрирует силу взаимодействий локусов, цвет линий графа – характер взаимодействия локусов, проценты на гранях и вершинах графа – уровень энтропии оптимального взаимодействия элементов системы.

PPARG. Этот факт подтвержден в модельном эксперименте, поставленном на мышах, в ходе которого показано, что у мышей с дефицитом CREB1 наблюдались гиперэкспрессия *PPARG* и нарушения в работе печени (Herzig *et al.*, 2003). Для локусов генов *LEP* и *LPL* изменений профиля сайтов связывания транскрипционных факторов выявлено не было. При дальнейшем проведении исследования ввиду большей информативности подробно рассматривали два локуса: *LEPR* и *PPARG*.

Анализ информации, представленной в базе данных UniProt, выявил, что SNP A223G, локализованный в гене *LEPR*, затрагивает экстрацеллюлярную цепь белка, осуществляющую рецепторную функцию, а SNP C34G, локализованный в гене *PPARG*, изменяет гидрофобные свойства цепи белка. Поиск по базе данных Protein Data Bank установил, что домены белков, затронутые изучаемыми мутациями, в ней не представлены. Для моделирования пространственных структур изучаемых доменов белковых продуктов генов *LEPR* и *PPARG* брали аминокислотные последовательности белковых структур канонических изоформ белков (P48357-1, P37231-1), депонированных в базе данных UniProt: для LEPR – позиции белковой последовательности 98–248 аминокислоты, для PPARG – позиции 1–51 аминокислоты.

Для анализируемых доменов рассмотренных белков в базе данных Protein Data Bank были найдены кристаллизованные белковые структуры, пригодные в качестве шаблонов при моделировании. Для LEPR – 3V6O (комплекс рецептора лептина LEPR с моноклональными антителами мыши). Цепь A этой структуры обладает гомологией 37 % с анализируемым нами доменом, что выше порогового значения в 30 % и позволяет достоверно судить о действии рассматриваемой мутации на белковый домен (Pipel, Lancet, 1999). Для PPARG – 1DP4 (A-рецептор предсердного натрийуретического пептида), гомология составила 34 %. Изменение химических свойств белков LEPR и PPARG и конформации их доменов под действием SNP A223G и C34G изучали при помощи программ Vadar 1.8 и ProtParam (табл. 5). Полученные данные позволили установить, что мутантные аллели генов *LEPR* и *PPARG* ведут к конформационным изменениям доменов белков за счет замены аминокислот, что изменяет их физико-химические свойства. Мутантный аллель гена *LEPR*G* меняет аминокислоту 223 с глицина на аргинин, что увеличивает молекулярную массу белка и изменяет его электростатические свойства. Объем экстрацеллюлярного домена белка, затронутого мутацией, увеличивается, при одновременном увеличении его доступной площади, что, возможно, способно оказывать некоторое влияние на рецепцию лептина (Carrillo-Vázquez *et al.*, 2013; Verkerke *et al.*, 2014).

Мутантный аллель гена *PPARG*C*, обусловливая замену аминокислоты 12 белкового продукта с пролина на аланин, приводит к небольшому уменьшению молекулярной массы белка при уменьшении объема и площади домена,

Таблица 5

Изменение свойств и структуры белков LEPR и PPARG под влиянием SNP A223G и C34G

Белковая структура / аллель гена	Молекулярная масса (Да)	Доступная площадь (A^2)	Общий объем (A^3)	Теоретическая pI (pH)	Индекс нестабильности	Алифатический индекс
LEPR/**A*	75219,7	5970,3	6162,7	7,66	45,45	86,12
LEPR/**G*	75247,8	6002,9	6221,0	7,85	45,45	86,12
PPARG/**C*	57620,1	4845,4	6061,6	5,61	50,20	86,32
PPARG/**G*	57594,1	4601,2	5969,1	5,61	49,86	86,51

pI – изоэлектрическая точка белка

затронутого мутацией. Индекс нестабильности белка, характеризующий его нестабильность *in vitro*, уменьшается, увеличивается термостабильность белка, изменяя его гидрофобность. Вероятно, это может влиять на функционирование белка и его способность регулировать дифференциацию адипоцитов и энергетический метаболизм, что повышает риск развития дислипидемии (Ahluwalia *et al.*, 2002).

ЗАКЛЮЧЕНИЕ

Проведенное исследование выявило особенности влияния мутаций в рассмотренных локусах генов *LEP*, *LEPR*, *PPARG* и *LPL* на метаболизм липидов и его нарушения у индивидов, проживающих в Республике Башкортостан. Отмечены достоверные различия в распределении частот генотипов и аллелей рассмотренных генов в группах индивидов, различающихся по уровню ОХС и ТГ. Методом однофакторного дисперсионного анализа показано количественное влияние генотипов и аллелей типированных генов на уровень ОХС и ТГ в сыворотке крови. Выявлены достоверные гендерные различия в распределении частот генотипов и аллелей, показано достоверное увеличение концентрации ТГ в сыворотке крови у более взрослых индивидов. Анализ межгенных взаимодействий локусов изученных генов позволил установить сочетания генотипов изученных локусов генов, способствующие повышению уровня ОХС и ТГ и приводящие к нарушениям липидного обмена. Методами биоинформатики на примере рецептора лептина LEPR и γ-рецептора PPARG изучено возможное конформационное изменение доменов белков, затронутых мутациями полиморфных локусов соответствующих генов, в результате которых изменяются их объем и доступная площадь. Установлено изменение физико-химических свойств белковых структур, обусловленных конформационными изменениями доменов белков LEPR и PPARG.

Нарушения в генах, кодирующих эти белковые структуры, как и в других рассмотренных, определяют недостаточный уровень выработки соответствующих продуктов, в результате чего не может быть обеспечено нормальное физиологическое соотношение компонентов липидного обмена, что может приводить к гипертрофии жировой ткани и в итоге к увеличению массы тела. На примере локусов генов *LEPR* и *PPARG* показано влияние изученных мутаций на потенциальные сайты связывания транскрипционных факторов, которые могут быть задействованы в регуляторных сигнальных каскадах организма человека, способствуя развитию такого комплексного нарушения, как первичная дислипидемия.

ЛИТЕРАТУРА

Даренская М.А., Колесникова Л.И., Бардымова Т.П. и др. Закономерности изменений показателей процесса пероксидации липидов у практически здоровых в различные периоды становления репродуктивной системы // Бюл. ВСНЦ СО РАМН. 2006. № 1 (47). С. 119–122.

Диагностика и коррекция нарушений липидного обмена с целью профилактики и лечения атеросклероза. Рекомендации экспертов Всероссийского научного

общества кардиологов (четвертый пересмотр). М., 2009. 19 с.

Маниатис Т., Фрич Э., Сэмбрук Дж. Молекулярное клонирование. М.: Мир, 1984. С. 220–228.

Марри Р., Греннер Д., Мейес П., Родуэлл В. Б. Биохимия человека. М.: Мир, 1993. 384 с.

Мельниченко Г.А. Ожирение в практике эндокринолога // Русский медицинский журнал. 2001. Т. 9. Вып. 2. С. 61–74.

Ребров А.П., Гайдукова И.З. Особенности дислипидемии при псориатическом артрите: взаимосвязь с атеросклерозом, факторами сердечно-сосудистого риска и системным воспалением // Саратовский научно-медицинский журнал. 2010. Т. 6. Вып. 3. С. 51–55.

Ahluwalia M., Evans M., Morris K. *et al.* // The influence of the Pro12Ala mutation of the PPAR-gamma receptor gene on metabolic and clinical characteristics in treatment-naпve patients with type 2 diabetes // Diabetes Obes. Metab. 2002. No. 4 (6). P. 376–378.

Barros R.P., Gustafsson J.A. Estrogen receptors and the metabolic network // Cell Metab. 2011. No. 14 (3). P. 289–299.

Bartsch D., Casadio A., Karl K.A., Serodio P., Kandel E.R. CREB1 encodes a nuclear activator, a repressor, and a cytoplasmic modulator that form a regulatory unit critical for long-term facilitation // Cell. 1998. No. 95. P. 211–223.

Boden G. Role of fatty acids in the pathogenesis of insulin resistance and NIDDM // Diabetes. 1997. No. 46. P. 3–10.

Carrillo-Vázquez J.P., Chimal-Vega B., Zamora-Lъpez B. Stru-ctural consequences of the polymorphism Q223R in the human leptin receptor: A molecular dynamics study // Am. J. Agric. Biol. Sciences. 2013. No. 8 (3). P. 239–248.

Constantin A., Costache G., Sima A. *et al.* Leptin G-2548A and leptin receptor Q223R gene polymorphisms are not associated with obesity in Romanian subjects // Biochem. Biophys. Res. Commun. 2010. No. 391 (1). P. 282–286.

Cooper A., Spirin V., Schmidt S. *et al.* Common single-nucleotide polymorphisms act in concert to affect plasma levels of high-density lipoprotein cholesterol // Am. J. Hum. Genet. 2007. No. 81. P. 1298–1303.

Després J.P., Couillard C., Gagnon J. *et al.* Race, visceral adipose tissue, plasma lipids, and lipoprotein lipase activity in men and women: the health, risk factors, exercise training, and genetics (heritage) family study // Arterioscler. Thromb. Vasc. Biol. 2000. No. 20. P. 1932–1938.

Duarte S., Francischetti E., Genelhu V. *et al.* LEPR p.Q223R, beta3-AR p.W64R and LEP c.-2548G>A gene variants in obese Brazilian subjects // Genet. Mol. Res. 2007. No. 6 (4). P. 1035–1043.

Faulds M.H., Zhao C. *et al.* The diversity of sex steroid action: regulation of metabolism by estrogen signaling // Journal Endocrinology. 2012. No. 212. P. 3–12.

Jeninga E.H., Gurnell M., Kalkhoven E. Functional implications of genetic variation in human PPARg // Trends in Endocrinology and Metabolism. 2009. V. 20. No. 8. P. 380–387.

Gasteiger E., Hoogland C., Gattiker A. *et al.* Protein identification and analysis tools on the ExPASy server. The Proteomics Protocols Handbook. T.: Humana Press, 2005. P. 571–607.

Guruprasad K., Reddy B.V., Pandit M.W. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence // Protein Eng. 1990. No. 4. P. 155–161.

Hayden M.R., Henderson H. The molecular biology and genetics of human lipoprotein lipase. Lipoproteins in Health and Disorder. L., 1999. P. 132–137.

Heinzmann C., Kirchgessner T., Lusis A. DNA polymorphism haplotypes of the human lipoprotein lipase gene // Hum. Genet. 1991. No. 86. P. 578–584.

Herzig S., Hedrick S., Morantte I. *et al.* CREB controls hepatic lipid metabolism through nuclear hormone receptor PPAR-γ // Nature. 2003. No. 426. P. 190–193.

Ikai A.J. Thermostability and aliphatic index of globular proteins // J. Biochem. 1980. No. 88. P. 1895–1898.

Kathiresan S., Melander O., Anevski D. *et al.* Polymorphisms associated with cholesterol and risk of cardiovascular events // New Eng. J. Med. 2008. No. 358. P. 1240–1249.

Laakso M. Gene variants, insulin resistance, and dyslipidemia // Curr. Opin. Lipidol. 2004. V. 2. No. 15. P. 115–120.

Langdahl B.L., Ralston S.H., Grant S.F., Eriksen E.F. An Spl binding site polymorphism in the *COLIA1* gene predicts osteoporotic fractures in both men and women // J. Bone Miner Res. 1998. No. 13 (9). P. 1384–1389.

Ma Y.Q., Thomas G.N., Ng M. The lipoprotein lipase gene HindIII polymorphism is associated with lipid levels in early-onset type 2 diabetic patients. // Metabolism. 2003. No. 52 (3). P. 338–343.

Mantzoros C.S. Leptin and the hypothalamus: neuroendocrine control of food intake // Mol. Psychiatry. 2004. V. 4. P. 8–12.

Mathew C.C. The isolation of high molecular weight eukaryotic DNA // Methods molecular biology. N.Y., 1984. V. 2. P. 31–34.

Meirhaeghe A., Cottel D., Amouyel P., Dallongeville J. Association between peroxisome proliferator-activated receptor γ haplotypes and the metabolic syndrome in French men and women // Diabetes. 2005. No. 54. P. 3043–3048.

Moore J.H. *et al.* A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility // J. Theoretical Biology. 2006. No. 241. P. 252–261.

Montagner A., Rando G., Degueurce G. *et al.* New insights into the role of PPARs // Prostaglandins, Leukot, Essent. Fatty Acids. 2011. V. 85. No. 5. P. 235–243.

Mullis K.B., Saiki R.K., Scharf S. *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. 487–491.

Park K.S., Shin H.D., Park B.L. *et al.* Polymorphisms in the leptin receptor (LEPR) – putative association with obesity and T2DM // Genet. 2006. No. 51. P. 85–91.

Pipel Y., Lancet D. The variable and conserved interfaces of modeled olfactory receptor proteins // Protein Science. 1999. No. 8. P. 969–977.

Rice P., Longden I., Bleasby A. The European Molecular Biology Open Software Suite // Trends in Genetics. No. 16 (6). 2000. P. 276–277.

Schulze P., Kratzsch J. Leptin as a new diagnostic tool in chronic heart failure // Clin.Chim. Acta. 2005. V. 362. P. 1–11.

Schwartz M.W., Seeley R.J. Seminars in medicine of the Beth Israel Deaconess medical center: Neuroendocrine responses to starvation and weight loss // New Engl. J. Med. 1997. V. 336. P. 1803–1811.

Semple R.K., Chatterjee V.K., O'Rahilly S. PPAR gamma and human metabolic disease // J. Clin. Invest. 2006. V. 116. No. 3. P. 581–589.

Sun Q., Cornelis M.C., Kraft P. *et al.* Genome-wide association study identifies polymorphisms in LEPR as determinants of plasma soluble leptin receptor levels // Hum. Mol. Genet. 2010. No. 19 (9). P. 1846–1855.

Verkerke H., Naylor C., Zabeau L. *et al.* Kinetics of leptin binding to the Q223R leptin receptor // PLoS One. 2014. No. 9 (4). P. 43–48.

Wang T.N., Huang M.C., Chang W. *et al.* G-2548A polymorphism of the leptin gene is correlated with extreme obesity in Taiwanese aborigines // Obesity. 2006. No. 14 (2). P. 183–187.

Willard L., Ranjan A., Zhang H. *et al.* VADAR: a web server for quantitative evaluation of protein structure quality // Nucleic Acids Res. 2003. No. 31 (13). P. 3316–3319.

ANALYSIS OF THE INTERACTION OF LIPID METABOLISM ALLELES IN DYSLIPIDEMIA

I.V. Nikolaev, R.V. Mulyukova, L.R. Kayumova, E.V. Vorobieva, V.Yu. Gorbunova

Akmulla Bashkir State Pedagogical University, Ufa, Russia,
e-mail: obg_bspu@mail.ru

Summary

Parameters of lipid metabolism in the spectrum of blood serum are the most commonly used indicators in clinical practice. Their disturbances (dyslipidemia) are: elevated levels of total cholesterol and triglycerides, as well as changes of other parameters resulting from aberrations in lipoprotein synthesis, transport and cleavage. The clinical significance of metabolic disorders covered by the term *dyslipidemia* is associated primarily with high risks of cardiovascular diseases, diabetes mellitus type 2, and obesity. Associations of certain SNPs (G-2548A in the promoter region of the leptin gene (*LEP*), A223G in exon 4 of the leptin receptor gene (*LEPR*), T495G in intron 8 in the lipoprotein lipase gene (*LPL*), and C34G in exon 8 of a nuclear receptor (*PPARG*)) with disturbed lipid metabolism have been investigated, and their cumulative contribution to the development of dyslipidemia is demonstrated.

Key words: dyslipidemia, single-nucleotide polymorphism, leptin, leptin receptor, lipoprotein lipase, nuclear receptor.

УДК 575.162

ГЕНЫ, КОНТРОЛИРУЮЩИЕ ПИЩЕВОЕ ПОВЕДЕНИЕ И МАССУ ТЕЛА ЧЕЛОВЕКА, И ИХ ФУНКЦИОНАЛЬНЫЕ И ГЕНОМНЫЕ ХАРАКТЕРИСТИКИ

© 2014 г. **Е.В. Игнатьева[1–3], Д.А. Афонников[1–3], Е.И. Рогаев[2], Н.А. Колчанов [1, 3]**

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

Поступила в редакцию 29 сентября 2014 г. Принята к публикации 27 октября 2014 г.

С целью систематизации информации о генах, участвующих в регуляции массы тела и пищевого поведения, была сформирована компиляция, включающая 424 гена, полученных (а) по данным из экспериментальных и обзорных статей, (б) из базы OMIM, (в) по данным мета-анализа экспериментов по полногеномному поиску ассоциаций. Четыре гена из компиляции (*BDNF, MC4R, PCSK1, POMC*) подтверждены всеми тремя источниками данных и рассматриваются как наиболее значимые в системе регуляции массы тела (приоритет 1). Выявлены группы, включающие 3 и 29 генов, подтвержденных двумя из трех источников данных (приоритет 2). Идентифицированы метаболические и сигнальные пути, участвующие в регуляции массы тела, которые можно считать потенциальными мишенями для фармакологических воздействий. Обнаружены районы хромосом человека, содержащие близкорасположенные гены из компиляции, содержащие в числе других гены, внесенные в компиляцию только по данным мета-анализа экспериментов по полногеномному поиску ассоциаций (*ETV5, MIR148A, NFE2L3, TMEM160),* что может помочь интерпретировать функции этих генов. К числу двенадцати генов из компиляции, наименее толерантных к мутациям, отнесены гены *LRP1, LRP5, RAI1, FASN, LYST, RPTOR, DGKD, LRP1B, NCOA1, ADCY3*. Компиляция может быть полезна как источник информации о генах-кандидатах, значимых для оценки риска развития ожирения и разработки фармакологических подходов к коррекции избыточной массы тела.

Ключевые слова: пищевое поведение, регуляция массы тела, локализация в геноме, толерантность к мутациям.

ВВЕДЕНИЕ

Регуляция массы тела – сложный фенотипический признак, который контролируется как генетическими факторами, так и факторами среды. Генетические факторы, провоцирующие ожирение, достаточно сложны. На долю моногенных форм заболевания приходится всего 5 % от всех наблюдаемых случаев в человеческих популяциях. Наиболее изученными локусами, мутации которых связаны с моногенными формами патологии, являются гены, кодирующие лептин (*LEP*) и его рецептор (*LEPR*), рецептор меланокортина типа 4 (*MC4R*), проопиомеланокортин (*POMC*), пробелок конвертаза субтилизин/кексин типа 1 (*PCSK1*), целенаправленный гомолог-1 гена дрозофилы (*SIM1*) (Blakemore *et al*., 2010; Zegers *et al*., 2012).

Обнаружение новых генов, нарушение которых связано с повышенной массой тела, позволяет выявлять биохимические и сигнальные пути, а также механизмы их регуляции, контролирующие рассматриваемый фенотипический признак. Такие гены могут рассматриваться как гены-кандидаты, используемые

для оценки риска развития патологии, а также как потенциальные мишени для фармакологических воздействий. Выявление генов, ассоциированных с патологиями, осуществляется с использованием различных стратегий, включая семейный анализ, исследование генов-кандидатов, а также полногеномный анализ ассоциаций (GWAS) (Herrera *et al*., 2011). Однако каждый из подходов имеет свои ограничения: 1) семейный анализ позволяет выявлять только варианты с высокой пенетрантностью; 2) результаты, полученные в ассоциативных исследованиях генов-кандидатов на разных популяционных выборках, не всегда подтверждают друг друга; 3) данные экспериментов GWAS с трудом поддаются интерпретации ввиду слабой изученности функциональной роли многих генов. Таким образом, несмотря на обилие информации, полученной различными экспериментальными методами, в том числе с использованием полногеномных подходов, данные о генетических предпосылках развития ожирения еще не достаточно полны.

Основой для формирования списка генов, потенциально задействованных в развитии ожирения, может быть рассмотрение физиологических систем, контролирующих массу тела, включая базовый метаболизм, регулируемый нервной, эндокринной и иммунной системами.

Важнейшей системой организма, от которой зависит масса тела, является система регуляции пищевого поведения, функционирующая при участии белковых продуктов генов, экспрессируемых как в мозге (Olszewski *et al*., 2008), так и в периферических органах и тканях: желудке, кишечнике, поджелудочной железе, жировой ткани. Центральное звено системы составляют нейроны аркуатных ядер гипоталамуса, секретирующие нейропептид Y (NPY) и агутиподобный белок (AgRP), а также альфа-меланоцитстимулирующий гормон (α-MSH), который образуется из проопиомеланокортина (POMC) под действием прогормон-конвертаз (PCSK1 и PCSK2) (Yeo, Heisler, 2012). Активность нейронов аркуатных ядер гипоталамуса контролируется гормонами (лептином, инсулином, грелином, полипептидом YY (PYY), глюкокортикоидами, адренокортикотропином, кортикотропин-релизинг гормоном), нейромедиаторными системами мозга (серотонергическая, дофаминергическая, адреналиновая, ГАМК-ергическая), а также нейротрофическими факторами BDNF и др. (Yeo, Heisler, 2012; Maniam *et al*., 2012).

Неотъемлемый этап исследований генетических основ предрасположенности к заболеваниям – теоретическая оценка потенциального влияния конкретных нуклеотидных замен на уровень экспрессии гена или функцию белка (Ponomarenko *et al*., 2002; Choi *et al*., 2012) либо возможной роли отдельных генов в развитии патологического процесса (Masoudi-Nejad *et al*., 2012; Smedley *et al*., 2014). Petrovski с соавт. (2013) была предложена мера Residual Variation Intolerance Score (RVIS), которая характеризует толерантность гена к мутациям. Отрицательный показатель RVIS свидетельствует о том, что ген находится под давлением стабилизирующего отбора, а положительный указывает на то, что стабилизирующий отбор ослабевает и, наоборот, возможен движущий либо балансирующий отбор. При сравнении выборки генов из базы OMIM со всеми остальными генами человека была выявлена достоверная корреляция между связью гена с заболеванием и пониженным (относительно среднего значения, рассчитанного для полногеномной выборки генов) значением RVIS. Таким образом, мера RVIS представляется нам удобным критерием оценки возможной роли генов в развитии патологии.

Целями нашей работы было формирование компиляции генов, участвующих в регуляции пищевого поведения и массы тела, определение их функциональных и геномных характеристик, а также выявление генов с пониженной толерантностью к мутациям на основе показателя RVIS, что указывает на значимость таких генов в развитии патологии. Была сформирована компиляция, включающая 424 гена, четыре из которых (*BDNF, MC4R, PCSK1, POMC)* имели наивысший приоритет. Выявлены метаболические и сигнальные пути, участвующие в регуляции массы тела. Обнаружены участки генома человека, содержащие близкорасположенные гены из компиляции. С использованием показателя RVIS установлено, что содержащиеся в компиляции гены *LRP1, LRP5, RAI1, FASN, LYST, RPTOR, DGKD, LRP1B, NCOA1, ADCY3, ZNF608, INSR* наименее толерантны к мутациям, и это указывает на повышенную значимость

мутантных вариантов этих генов в развитии патологий, в частности ожирения.

МАТЕРИАЛЫ И МЕТОДЫ

Гены были внесены в компиляцию на основании трех информационных источников (табл. 1). Первым источником, именуемым в дальнейшем *Публикации*, были научные статьи (включая как экспериментальные, так и обзорные), в которых охарактеризована роль генов в регуляции пищевого поведения у человека либо у близких видов (мыши и крысы). Если в статье была приведена информация о гене млекопитающего, в компиляцию вносился гомологичный ген человека с соответствующей ссылкой на источник данных.

Вторым источником была база OMIM. Поиск по базе OMIM проводился по ключевым словам «obesity», либо «hyperphagia», либо «anorexia». В результате поиска из базы OMIM получен список из 333 генов, именуемый в дальнейшем *OMIM*. Список генов *OMIM* включал гены двух категорий. К первой, именуемой в дальнейшем *OMIM-allelic variant*, отнесены 73 гена, имеющие хотя бы одно из вышеперечисленных ключевых слов в поле «allelic variants». Ко второй категории, именуемой *OMIM-all text*, отнесены 260 генов, имеющие перечисленные выше ключевые слова только в поле «all text». Поскольку для генов, отнесенных к категории *OMIM-all text*, информация об аллельных вариантах, ассоциированных с изменением пищевого поведения, отсутствовала, гены этой группы можно рассматривать как участвующие в регуляции пищевого поведения и массы тела и выявленные в базе OMIM методом Text Mining.

Третий источник составили статьи, посвященные результатам мета-анализа данных экспериментов по полногеномному поиску ассоциаций (GWAS). В компиляцию были включены гены человека, для которых в результате мета-анализа была показана ассоциация с повышенным индексом массы тела с достоверностью $p < 5{,}0 \times 10^{-8}$ (полногеномный уровень). В этот список, именуемый в дальнейшем *GWAS-мета-анализ*, были включены 48 генов человека. При этом 39 генов из 48 были выявлены на основе мета-анализа данных GWAS, полученных с использованием популяционных выборок населения европеоидного происхождения. Кроме того, в проанализированных нами статьях были

Таблица 1

Информационные источники, на основе которых была сформирована компиляция генов пищевого поведения и регуляции массы тела

Информационный источник	Краткое обозначение информационного источника	Кол-во генов	Кол-во публикаций либо поисковый запрос
Научные публикации, о функциональной причастности генов к регуляции пищевого поведения	*Публикации*	83	33 (11 обзорных статей, 22 статьи, описывающие данные экспериментов)
База данных OMIM	*OMIM-allelic variants*	73	‘Search: ‘hyperphagia’ OR ‘obesity’ OR ‘anorexia’ (Records with: gene map locus; Prefixes: +, *; Search in: allelic variants)
	OMIM-all text	260	‘Search: ‘hyperphagia’ OR ‘obesity’ OR ‘anorexia’ (Records with: gene map locus; Prefixes: +, *; Search in: all text)[1]
Научные публикации, описывающие, результаты мета-анализа данных GWAS	*GWAS-мета-анализ*	48	6

[1] Гены, полученные в результате запроса по полю allelic variants, были исключены из списка OMIM-all text.

представлены данные анализа популяционных выборок африканского, восточно-азиатского, северо-американского и австралийского происхождения.

Функциональную аннотацию генов и выявление обогащенных терминов Gene Ontology и метаболических путей осуществляли с помощью Интернет-инструмента DAVID (Huang da *et al*., 2009). Расположение генов на хромосомах исследовали на основе данных таблицы gene2refseq из базы EntrezGene. Для всех генов генома человека были определены координаты их центров, вычисленные как среднее арифметическое между позициями начала и конца каждого гена на хромосоме. Близкорасположенные гены выявляли, используя позиции центров генов. Проводили поиск геномных районов, содержащих центральные позиции четырех генов в пределах 2 Мб.

Чтобы оценить толерантность генов к мутациям, использовали скор RVIS. Значения RVIS для 16 956 генов человека (Dataset S2) были экстрагированы из публикации Petrovski с соавт. (2013). В этом исследовании RVIS определялся на основе соотношения количества замен с частотой встречаемости минорного аллеля выше 0,1 (MAF > 0,1) к общему количеству замен в кодирующих частях генов. Значение RVIS, равное нулю, соответствует среднему значению этой величины для всех генов из выборки, проанализированной авторами, т. е. практически среднему значению по геному. У генов с положительным RVIS количество замен с MAF > 0,1 превышает ожидаемое, а у генов, имеющих отрицательный RVIS, обнаружено пониженное (по сравнению со среднегеномным) количество замен с MAF > 0,1 (Petrovski *et al*., 2013).

РЕЗУЛЬТАТЫ

Информационное содержание компиляции

С использованием трех информационных источников *Публикации,* OMIM, *GWAS-мета-анализ* (см. раздел Материалы и методы и табл. 1) сформирована компиляция объемом 424 гена (рис. 1). 83 гена были описаны в публикациях, 333 гена внесены в компиляцию на основе запросов к базе OMIM, и 48 генов внесены по данным информационного источника *GWAS-мета-анализ*. Списки генов, полученных из трех информационных источников, частично

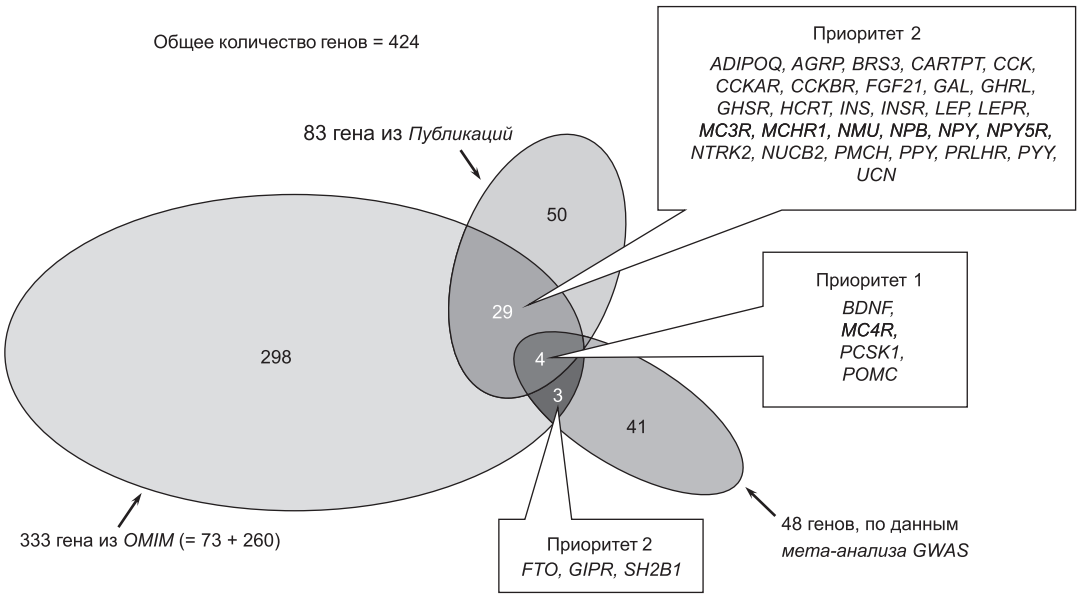


Рис. 1. Количество генов, внесенных в компиляцию на основе информационных источников *Публикации*, *OMIM* и *GWAS-мета-анализ* (представление в форме диаграммы Венна). Гены, выявленные по данным всех трех источников (приоритет 1) либо по данным из двух источников (приоритет 2), обозначены на выносках, исходящих из соответствующих областей диаграммы.

перекрывались (рис. 1). Обнаружено, что четыре гена (*BDNF, MC4R, PCSK, POMC*) имеют свидетельства о причастности к регуляции пищевого поведения либо массы тела из всех трех информационных источников (*Публикации, OMIM, GWAS-мета-анализ*). Следовательно, эти гены можно рассматривать как наиболее значимые в системе регуляции массы тела (на рис. 1 обозначены как Приоритет 1).

Выявлены также два подмножества генов, принадлежащие двум из трех списков. Три гена (*FTO*, *GIPR*, *SH2B1*) содержались на пересечении списков *OMIM* и *GWAS-мета-анализ*. Еще 29 генов отмечены на пересечении списков *Публикации* и *OMIM*. На рис. 1 эти два подмножества генов обозначены как Приоритет 2.

Функциональная аннотация генов

С использованием системы DAVID был охарактеризован набор клеточных функций белков, кодируемых генами из компиляции. Около четверти генов (24,8 %) кодируют рецепторы клеточной поверхности; 23,9 % генов кодируют белки, проаннотированные GO термином *receptor binding activity*, т. е. сигнальные молекулы (гормоны, нейропептиды и т. д.); 15 % генов кодируют транскрипционные регуляторы, и, наконец, 7 % генов кодируют белки с киназной активностью (рис. 2). Около четверти генов из компиляции составили гены с очень гетерогенными функциями, эта группа на рис. 2 имеет обозначение Other.

С использованием Интернет-инструмента DAVID были выявлены метаболические и сигнальные пути, неслучайно часто ($p < 10^{-2}$) представленные, т. е. перепредставленные, в аннотации генов из компиляции. Уровень перепредставленности выше двух имели 15 путей из базы KEGG, 7 путей из REACTOME и 8 путей из BIOCARTA (рис. 3).

Локализация генов в геноме

Расположение генов из компиляции в геноме человека представлено на рис. 4. Обнаружено одиннадцать геномных районов протяженностью 2 Мб, включающих центральные позиции четырех близкорасположенных генов. Четыре из одиннадцати выявленных геномных района длиной 2 Мб располагались отдельно друг от друга на хромосомах 7, 11, 19, 20.

Кроме того, в трех ситуациях (хромосомы 3, 17, 19), выявленные нами районы протяженностью 2Мб перекрывались между собой. На хромосоме 3 на перекрывании таких участков содержится пятерка близкорасположенных генов (*IGF2BP2, ETV5, HRG, KNG1, ADIPOQ*). На хромосоме 17 найдена семерка генов (*HCRT, STAT3, PPY, PYY, SLC4A1, CRHR1, MAPT*). На хромосоме 19 имелась пятерка генов (*TMEM160, FGF21, GYS1, LHB, CPT1C*).

Таким образом, с использованием этого критерия выявлено семь групп близкорасположенных генов (их локализация выделена на рис. 4 овалами).

Выявление генов, наименее устойчивых к мутациям, на основе величины RVIS

При ранжировании в соответствии со значениями RVIS были выбраны 12 генов с наиболее низким скором (табл. 2). Данные гены обладали значением RVIS < –2,13. Это, согласно S. Petrovski с соавт. (2013), означает, что эти гены попадают в число 1,49 % наименее толерантных к мутациям генов генома человека.

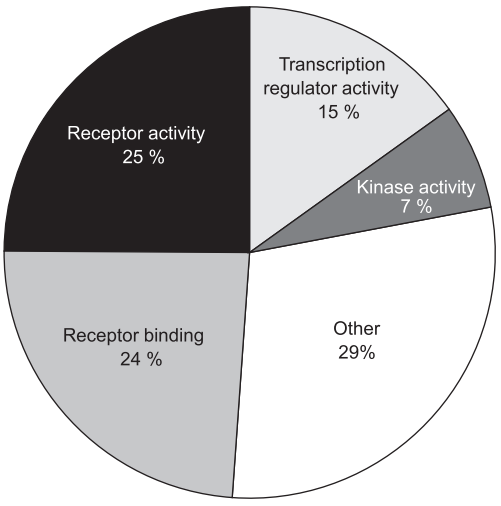


Рис. 2. Распределение генов из компиляции по основным категориям базы Gene Ontology (подраздел *Molecular functions*). Приведены GO термины и доли генов, имеющих в аннотации данные GO термины, от общего количества генов в компиляции.

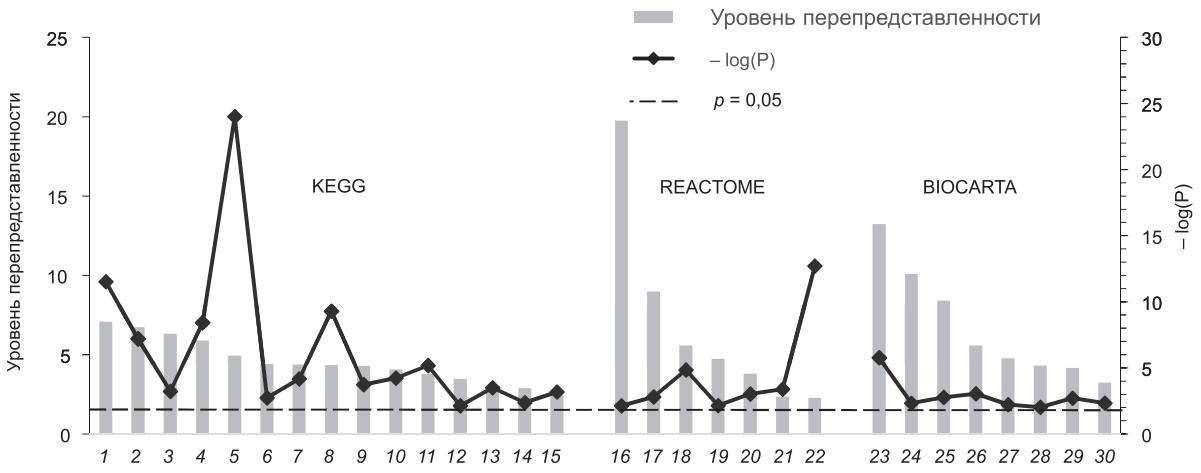


Рис. 3. Метаболические и сигнальные пути из баз KEGG, REACTOME и BIOCARTA, перепредставленные в аннотации генов из компиляции более чем в два раза и уровнем достоверности $p \leq 10^{-2}$.

1 – Adipocytokine signaling pathway; *2* – Type II diabetes mellitus; *3* – Maturity onset diabetes of the young; *4* – PPAR signaling pathway; *5* – Neuroactive ligand-receptor interaction; *6* – Aldosterone-regulated sodium reabsorption; *7* – NOD-like receptor signaling pathway; *8* – Insulin signaling pathway; *9* – Acute myeloid leukemia; *10* – Pancreatic cancer; *11* – Toll-like receptor signaling pathway; *12* – mTOR signaling pathway; *13* – Neurotrophin signaling pathway; *14* – Progesterone-mediated oocyte maturation; *15* – Calcium signaling pathway; *16* – Mitochondrial Uncoupling; *17* – Proteins Signal attenuation; *18* – Signaling by Insulin receptor; *19* – Regulation of beta-cell development; 20 – Hormone biosynthesis; *21* – Signalling by NGF; *22* – Signaling by GPCR; *23* – Visceral Fat Deposits and the Metabolic Syndrome; *24* – Reversal of Insulin Resistance by Leptin; *25* – Role of PPAR-gamma Coactivators in Obesity and Thermogenesis; *26* – IL-6 signaling pathway; *27* – Insulin Signaling Pathway; *28* – Role of ERBB2 in Signal Transduction and Oncology; *29* – Signal transduction through IL1R; *30* – Mechanism of Gene Regulation by Peroxisome Proliferators via PPARa(alpha).

Таблица 2

Двенадцать генов из компиляции, имеющих наиболее низкое значение RVIS, что указывает на низкую толерантность к мутациям в кодирующей части

Символ гена	Название гена	Источник данных	RVIS
LRP1	low density lipoprotein receptor-related protein 1	*OMIM-all text*	–7,28
LRP5	low density lipoprotein receptor-related protein 5	*OMIM-all text*	–3,72
RAI1	retinoic acid induced 1	*OMIM-all text*	–3,68
FASN	fatty acid synthase	*OMIM-all text*	–3,39
LYST	lysosomal trafficking regulator	*OMIM-all text*	–3,04
RPTOR	regulatory associated protein of MTOR, complex 1	*GWAS-мета-анализ*	–2,58
DGKD	diacylglycerol kinase, delta 130kDa	*OMIM-all text*	–2,34
LRP1B	low density lipoprotein receptor-related protein 1B	*GWAS-мета-анализ*	–2,29
NCOA1	nuclear receptor coactivator 1	*OMIM-all text*	–2,19
ADCY3	adenylate cyclase 3	*GWAS-мета-анализ*	–2,17
ZNF608	zinc finger protein 608	*GWAS-мета-анализ*	–2,16
INSR	insulin receptor	*OMIM-allelic variants*, *Публикации*	–2,14

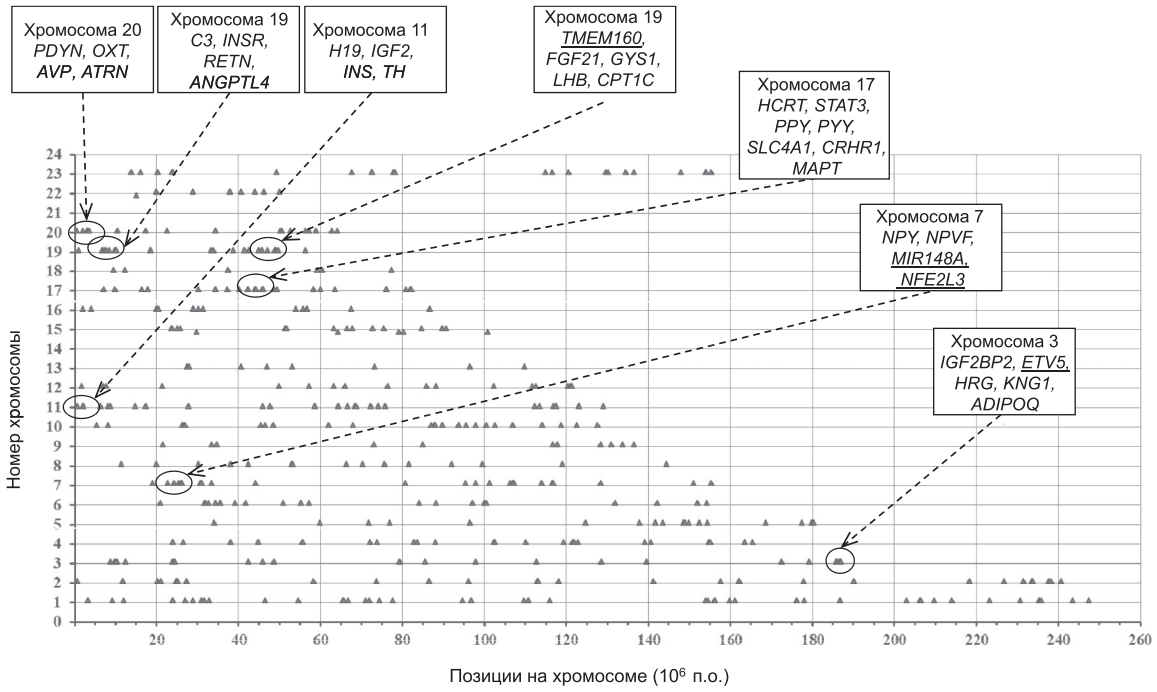


Рис. 4. Локализация генов из компиляции на хромосомах человека. Гены обозначены треугольниками так, что их позиции по оси 0X соответствуют координатам центров генов на хромосоме, а по оси 0Y – номеру хромосомы. Овалами выделены либо четыре отдельно расположенные района хромосом протяженностью $\leq 2 \times 10^6$ п. о., включающие центральные позиции четырех генов из компиляции (хромосомы 7, 11, 19, 20), либо три участка (хромосомы 3, 17, 19), где такие районы длиной $\leq 2 \times 10^6$ п. о. перекрываются, что позволяет выявить пять, семь и еще пять близкорасположенных генов соответственно. В прямоугольных выносках приведены названия генов. Гены, внесенные в компиляцию только на основании информационного источника *GWAS-мета-анализ*, подчеркнуты.

ЗАКЛЮЧЕНИЕ

В данной работе представлена компиляция генов, регулирующих пищевое поведение и массу тела, сформированная с использованием трех информационных источников (см. табл. 1). Выявлены четыре гена (*MC4R, PCSK, POMC, BDNF*) с высоким приоритетом, которые содержались во всех трех информационных источниках (см. рис. 1). Белки, кодируемые генами *MC4R, PCSK, POMC*, экспрессируются в аркуатных ядрах гипоталамуса и выполняют ключевую функцию в системе передачи сигнала насыщения. Ген *BDNF* кодирует нейротрофический фактор мозга, активация которого в вентромедиальных ядрах гипоталамуса снижает аппетит (Yeo, Heisler, 2012).

Значительную долю генов в компиляции составляли гены, кодирующие рецепторы клеточной поверхности, и белки, способные связываться с рецепторами, а также белки с киназной активностью (см. рис. 2). Это наблюдение отражает тот факт, что пищевое поведение контролируется обширной системой нейронов, взаимодействующих между собой посредством сигнальных веществ (нейропептидов, гормонов и т. п.), которые связываются с рецепторами на поверхности клетки и передают сигналы в цитоплазму посредством каскадов сигнальной трансдукции (Olszewski *et al*., 2008). В компиляции выявлена также значительная доля генов с регуляторными функциями, а именно, транскрипционных регуляторов (63 гена, т. е. 15 % выборки). Известно, что каждый транскрипционный регулятор контролирует активность большой кассеты генов (Merkulova *et al*., 2013), поэтому данная группа генов перспективна для выбора потенциальных мишеней фармакологического воздействия, управляя активностью которых можно препятствовать развитию патологии.

Гены из компиляции задействованы в большом разнообразии метаболических и сигнальных путей, что отражает сложную природу механизмов регуляции массы тела (см. рис. 3). Мы рассматриваем выявленные пути как возможные точки приложения эффектов фармакологических препаратов, которые могут быть разработаны в будущем с целью коррекции массы тела.

При исследовании распределения генов в геноме обнаружены участки, содержащие близкорасположенные гены (рис. 4). Можно предполагать, что в некоторых из описанных нами случаев близкое расположение генов подразумевает наличие координированной регуляции экспрессии. Координированная регуляция может обеспечиваться за счет формирования соответствующей 3D структуры хроматина, функционирования так называемых транскрипционных фабрик, а также районов хроматина с барьерной функцией (Razin *et al*., 2011; Wang *et al*., 2012). Проверка этой гипотезы требует более детального анализа с привлечением дополнительных экспериментальных данных.

В пределах трех из выявленных нами районов близкорасположенных генов находились гены (*ETV5, MIR148A, NFE2L3, TMEM160*), внесенные в компиляцию только на основе информационного источника *GWAS-мета-анализ* (подчеркнутые символы генов на рис. 4). Данное наблюдение поможет сформировать гипотезы относительно механизмов функционирования генов, обнаруженных в экспериментах GWAS, роль которых в регуляции массы тела охарактеризована пока недостаточно.

Исследование другой геномной характеристики, скора RVIS (Petrovski *et al*., 2013), позволило выявить гены с наименьшей толерантностью к мутациям (см. табл. 2), что указывает на повышенный риск развития патологий, в частности ожирения, в случае обнаружения мутаций в этих генах. Среди двенадцати генов с самым низким скором (RVIS < –2,13) обнаружены два гена, кодирующие белки с транскрипционными регуляторными функциями (*RAI1*, *ZNF608*). Наше наблюдение создает мотивацию к дальнейшим исследованиям генов-мишеней транскрипционных факторов RAI1 и ZNF608 на основе технологий Chip-Seq и т. п.

Выполненные в настоящей работе систематизация данных о генах, регулирующих пищевое поведение и массу тела, а также их анализ позволят расширить знания о механизмах регуляции этих фенотипических признаков. Результаты работы могут быть полезны при разработке более эффективных подходов персонализированной медицины на всех этапах, включая профилактику (генотипирование с целью оценки риска патологии), диагностику (выявление генов, несущих повреждающие мутации), а также выбор средств фармакологического воздействия.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке РНФ (проект № 14-24-00123).

ЛИТЕРАТУРА

Blakemore A.I., Froguel P. Investigation of Mendelian forms of obesity holds out the prospect of personalized medicine // Ann. N. Y. Acad. Sci. 2010. V. 1214. P. 180–189.

Choi Y., Sims G.E., Murphy S. *et al.* Predicting the functional effect of amino acid substitutions and indels // PLoS One. 2012. V. 7. No. 10. P. e46688.

Herrera B.M., Keildson S., Lindgren C.M. Genetics and epigenetics of obesity // Maturitas. 2011. V. 69. No. 1. P. 41–49.

Huang da W. *et al.* Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources // Nat. Protoc. 2009. V. 4. No. 1. P. 44–57.

Maniam J., Morris M.J. The link between stress and feeding behavior // Neuropharmacology. 2012. V. 63. No. 1. P. 97–110.

Masoudi-Nejad A., Meshkin A., Haji-Eghrari B., Bidkhori G. Candidate gene prioritization // Mol. Genet. Genomics. 2012. V. 287. No. 9. P. 679–698.

Merkulova T.I., Ananko E.A., Ignat'eva E.V., Kolchanov N.A. Regulatory transcription codes in eukaryotic genomes // Genetika. 2013. V. 49. No. 1. P. 37–54.

Olszewski P.K., Cedernaes J., Olsson F. *et al.* Analysis of the network of feeding neuroregulators using the Allen Brain Atlas // Neurosci. Biobehav. Rev. 2008. V. 32. No. 5. P. 945–956.

Oshchepkov D.Y., Vityaev E.E., Grigorovich D.A., Ignatieva E.V., Khlebodarova T.M. SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition // Nucleic Acids Res. 2004. V. 32. P. W208–W212.

Petrovski S. Wang Q., Heinzen E.L., Allen A.S., Goldstein D.B. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes // PLoS Genet. 2013. V. 9. No. 8. P. e1003709.

Ponomarenko J.V., Merkulova T.I., Vasiliev G.V. *et al.* rSNP_Guide, a database system for analysis of transcription factor binding to target sequences: application to SNPs and site-directed mutations // Nucleic Acids Res. 2001. V. 29. No. 1. P. 312–316.

Razin S.V., Gavrilov A.A., Pichugin A. *et al.* Transcription factories in the context of the nuclear and genome organization // Nucleic Acids Res. 2011. V. 39. No. 21. P. 9085–9092.

Smedley D., Köhler S., Czeschik J.C. *et al.* Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases // Bioinformatics. 2014. V. 30. pii: btu508 ah.

Wang J., Lunyak V.V., Jordan I.K. Genome-wide prediction and analysis of human chromatin boundary elements // Nucleic Acids Res. 2012. V. 40. No. 2. P. 51–529.

Yeo G.S., Heisler L.K. Unraveling the brain regulation of appetite: lessons from genetics // Nat. Neurosci. 2012. V. 15. No. 10. P. 1343–1349.

Zegers D., Van Hul W., Van Gaal L.F., Beckers S. Monogenic and complex forms of obesity: insights from genetics reveal the leptin-melanocortin signaling pathway as a common player // Crit. Rev. Eukaryot. Gene Expr. 2012. V. 22. No. 4. P. 325–343.

HUMAN GENES CONTROLLING FEEDING BEHAVIOR OR BODY MASS AND THEIR FUNCTIONAL AND GENOMIC CHARACTERISTICS: A REVIEW

E.V. Ignatieva[1–3], D.A. Afonnikov[1–3], E.I. Rogaev[2], N.A. Kolchanov[1, 3]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: eignat@bionet.nsc.ru;
[2] Center for Brain Neurobiology and Neurogenetics, Novosibirsk, Russia;
[3] Novosibirsk National Research State University, Novosibirsk, Russia

Summary

The goals of this study were to create a compilation of genes controlling human body weight and feeding behavior and to summarize functional and genomic information on these genes. Information on 424 human genes was obtained from scientific publications, OMIM and meta-analysis of GWAS data. Four genes (*BDNF, MC4R, PCSK1*, and *POMC*) were confirmed by all three data sources; thus, these genes have the highest priority (No. 1). Genes of other two groups (3 and 29 genes) were confirmed by two of three data sources; thus having priority No. 2. Pathways important for body mass regulation were revealed, and they may be candidate pharmacological targets for obesity treatment. Regions of human chromosomes containing closely located genes from the compilation were revealed. Some groups of closely located genes included genes (*ETV5, MIR148A, NFE2L3*, and *TMEM160)* confirmed by GWAS meta-analysis only. This finding may be helpful in the identification of their functions. Use of Residual Variation Intolerance Score (RVIS) revealed genes with decreased tolerance to functional genetic variation: *LRP1, LRP5, RAI1, FASN, LYST, RPTOR, DGKD, LRP1B, NCOA1*, and *ADCY3*. The compilation can be used in genotyping for pathology risk estimation and for designing new pharmacological approaches for treatment of human obesity.

Key words: feeding behavior, genomic location, regulation of body mass, tolerance to functional genetic variation.

УДК 577.214:577.218:004.651

ВЛИЯНИЕ ФЛАНКИРУЮЩИХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ НА ТОЧНОСТЬ РАСПОЗНАВАНИЯ САЙТОВ СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ

© 2014 г. **Т.М. Хлебодарова[1], Д.Ю. Ощепков[1], В.Г. Левицкий[1, 2], О.А. Подколодная[1], Е.В. Игнатьева[1], Е.А. Ананько[1], И.Л. Степаненко[1], Н.А. Колчанов[1, 2]**

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

Поступила в редакцию 25 сентября 2014 г. Принята к публикации 1 октября 2014 г.

Развитие *in vitro* технологий привело к появлению новых экспериментальных данных о связывании белков с ДНК, которые накапливаются в базах данных и используются при исследовании механизмов регуляции экспрессии генов и разработке компьютерных методов распознавания сайтов связывания в геномах про- и эукариот. Однако пока не ясно, насколько *in vitro* селектированные последовательности отражают истинную структуру природных сайтов связывания транскрипционных факторов (ТФ). С использованием Кульбака – Лейблера критерия расстояний проведено сравнение сходства частотных матриц сайтов связывания ТФ, построенных на основе выборок искусственно селектированных последовательностей и природных сайтов. Показано, что для 80 % ТФ (из числа исследованных) наблюдается высокое сходство коровых последовательностей природных и искусственных сайтов. Для 20 % ТФ их *in vitro* селектированные последовательности имеют в коровой структуре сайта более широкий спектр допустимых значимых нуклеотидов, не встречающихся среди природных сайтов. Методом весовых матриц проведена оценка оптимальной длины последовательностей ДНК, включающих природные сайты связывания, при которой удается достичь максимальной точности их распознавания. Обнаружено, что примерно для 80 % ТФ (из исследованных) оптимальная для распознавания длина сайта связывания значительно превышает длину коровой последовательности и длину *in vitro* селектированных сайтов. Выявленные особенности *in vitro* селектированных сайтов связывания ТФ накладывают определенные ограничения на их использование при разработке компьютерных методов распознавания потенциальных сайтов в геномных последовательностях.

Ключевые слова: транскрипционные факторы, сайты связывания, частотные и весовые матрицы, *in vitro* селектированные последовательности.

ВВЕДЕНИЕ

Ключевым звеном тонкой регуляции экспрессии генов является структура регуляторных последовательностей промоторов генов, определяющая спектр возможных воздействий со стороны регуляторов транскрипции. Поэтому не удивительно, что поток информации по изучению механизмов регуляции транскрипции и структуры промоторов не ослабевает. В связи с этим в последнее десятилетие всё большее внимание уделяется созданию и развитию баз данных по регуляции транскрипции как у эу-, так и у прокариот (Wingender *et al.*, 2001; Lescot *et al.*, 2002; Praz *et al.*, 2002; Matys *et al.*, 2003; 2006; Kolchanov *et al.*, 2002; 2008; Munch *et al.*, 2003; Zhao *et al.*, 2005; Liu *et al.*, 2008; Grote *et al.*, 2009 и др.).

Развитие новых технологий создает большие возможности для накопления и анализа

информации о структуре регуляторных районов генов. Так, с развитием *in vitro* технологий, в частности SELEX (Systematic Evolution of Ligands by EXponential enrichment), SAAB (Selected And Amplified Binding site imprint assay), REPSA (Restriction Endonuclease Protection Selection and Amplification), CASTing (Cyclical Amplification and Selection of Targets) и других более поздних модификаций методов, например SELEX SAGE, SELEX-seq и др., используемых для селекции сайтов связывания транскрипционных факторов (Blackwell, Weintraub, 1990; Pollock, Treisman, 1990; Wright *et al.*, 1991; Hardenbol *et al.*, 1997; Roulet *et al.*, 2002; обзоры: Djordjevic, 2007; Wang *et al.*, 2011), появилось много информации о структуре сайтов связывания для различных ТФ как про-, так и эукариот.

Подобного рода данные необходимы для изучения механизмов функционирования ТФ, построения методов распознавания сайтов связывания ТФ и районов, регулирующих транскрипцию генов, а также для функциональной аннотации геномов. Созданы специализированные базы данных, предназначенные для накопления и систематизации информации об искусственно селектированных сайтах связывания ТФ (Ponomarenko *et al.*, 2000; Sandelin *et al.*, 2004; Bryne *et al.*, 2008; Newburger, Bulyk, 2009; Portales-Casamar *et al.*, 2010; Chen *et al.*, 2011 и др.). Накопление подобной информации идет также и в базе TRANSFAC, одной из наиболее известных баз по регуляции транскрипции (Matys *et al.*, 2003, 2006).

Однако вопрос о том, насколько *in vitro* селектированные последовательности отражают истинную структуру природных сайтов связывания ТФ и каковы возможности их использования для создания компьютерных методов поиска природных сайтов в геномах различных видов организмов, остается открытым. Информация об этом противоречива и неоднозначна (Robison *et al.*, 1998; Shultzaberger, Schneider, 1999; Roulet *et al.*, 2000; Ehret *et al.*, 2001).

Чтобы ответить на этот вопрос, необходим сравнительный анализ большого количества данных, полученных из разных источников. Имея значительный объем информации о структуре природных сайтов связывания ТФ в базе TRRD (Kolchanov *et al.*, 2002, 2008), мы пошли по пути объединения этих данных с данными, полученными с помощью *in vitro* технологий, и создали базу данных ArtSite (Khlebodarova *et al.*, 2006).

В этой базе накапливаются частотные матрицы, описывающие структуру как природных, так и *in vitro* селектированных сайтов связывания ТФ эу- и прокариот. (txt-файл базы может быть получен по запросу у авторов.) Матрицы получены на основе выравнивания последовательностей этих сайтов относительно наиболее консервативных нуклеотидов. В настоящее время база данных ArtSite содержит более 630 матриц, которые описывают структуру сайтов связывания более чем 300 ТФ. Из них более 100 матриц построено на основе природных, функциональных сайтов, которые описывают структуру сайтов связывания для 134 транскрипционных факторов.

Такое большое количество данных позволяет сопоставить результаты распознавания сайтов связывания ТФ, полученных методами, построенными на основе выборок природных сайтов, и искусственно селектированных последовательностей. Ранее мы сравнили структуры коровых последовательностей природных и искусственных сайтов связывания для пяти ТФ, имеющих разные ДНК-связывающие домены и, соответственно, разные типы связывания с ДНК (USF, SP1, YY1, RXR/RAR и E2F1/DP1) (Khlebodarova *et al.*, 2006). Мы получили очень высокий уровень сходства матриц, взятых из разных источников.

Эти данные позволили нам предположить, что, по крайней мере, для исследованных факторов существует возможность использования выборок *in vitro* селектированных последовательностей для распознавания потенциальных природных сайтов ТФ в геномах различных организмов. Для проверки этого предположения мы решили провести более полное сравнение матриц, построенных с использованием последовательностей сайтов, выявленных на основе селекции *in vitro*, и природных сайтов. Ресурс базы ArtSite позволил провести подобное сравнение для 35 ТФ. Кроме того, мы поставили задачу оценить оптимальную длину последовательностей ДНК природных сайтов связывания ТФ, при которой удается достичь максимальной точности их распознавания.

МАТЕРИАЛ И МЕТОДЫ

Для анализа сходства природных и *in vitro* селектированных сайтов связывания транскрипционных факторов были использованы последовательности сайтов, аннотированные в базе ArtSite, и частотные матрицы, созданные на их основе (Khlebodarova *et al.*, 2006). Использованы выборки сайтов связывания только тех транскрипционных факторов, для которых присутствовали данные и по природным, и по *in vitro* селектированным сайтам.

Для измерения сходства частотных матриц сайтов связывания ТФ был использован критерий, основанный на расстоянии Кульбака – Лейблера (Kullback – Leibler), описанный Aerts с соавт. (Aerts *et al.*, 2003). Расстояние Кульбака – Лейблера – широко известный статистический метод сравнения и оценки различия распределений. В применении к частотным матрицам он позволяет сравнивать выборки сайтов связывания транскрипционных факторов и оценивать степень их различия. Согласно критерию, значения расстояний менее и равные 0,2 определяют высокий уровень сходства матриц, от 0,2 до 0,3 – средний, более 0,3 – слабый.

Для оценки оптимальной длины сайтов связывания транскрипционных факторов, при которой достигается максимальная точность распознавания их потенциальных сайтов в геномных последовательностях, использован метод оптимизации весовых матриц (Levitsky *et al.*, 2007).При распознавании методом весовых матриц ошибка первого рода (недопредсказание сайтов из выборки обучения) была зафиксирована на уровне 50 %, а ошибка второго рода (перепредсказание случайных последовательностей, полученных из выборки обучения путем перемешивания) минимизировалась.

Точность распознавания оценена стандартным методом независимого скользящего контроля (jackknife test). Максимальная длина последовательности природного сайта, для которой проведена оценка точности распознавания, была равна 50 нуклеотидам. Нами отобрано 29 выборок сайтов связывания транскрипционных факторов, для которых оказалось возможным произвести расчеты точности распознавания для длин матриц вплоть до 50 п. н. (табл. 1).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Анализ сходства матриц, полученных на основе природных и *in vitro* селектированных сайтов

Анализ сходства матриц, полученных на основе природных и *in vitro* селектированных сайтов связывания для 28 выборок сайтов связывания ТФ (ССТФ) эукариот с использованием критерия расстояний Кульбака – Лейблера приведен в табл. 2. Согласно полученным оценкам, расстояние менее 0,2, свидетельствующее о высоком уровне сходства природных и искусственных матриц ССТФ, показано для 80 % матриц из числа исследованных. Для шести ТФ (~20 %), а именно: C/EBPα, C/EBPβ, PEA3, PDX1, MYOD и SREBP1 (SRE тип сайтов) – это расстояние превышало 0,2, и только для одного ТФ, EGR1, оно было больше 0,3, что указывало на средний и слабый уровень сходства природных и искусственных матриц соответствующих ТФ (табл. 2). Для PEA3 средний уровень сходства (0,25) искусственных и природных матриц можно было бы объяснить видовыми различиями ТФ, так как в первом случае это были сайты связывания ТФ PEA3 *Brachydanio rerio* (рыбы), а во втором – человека и мыши.

Что касается остальных ТФ, то во всех случаях матрицы построены на основе сайтов связывания ТФ млекопитающих, причем различия в структуре сайтов C/EBPβ, EGR1 и MYOD были выражены сильнее, чем для PEA3 (табл. 2). Эти различия нельзя объяснить особенностями какого-либо ДНК-связывающего домена соответствующих ТФ, так как в рассматриваемых случаях типы доменов были разные: C/EBP – bZIP, EGR1 – Zinc finger, MYOD – bHLH, PDX1 – Homeo домен, PEA3 – Ets домен. Более того, матрицы сайтов связывания ТФ, имеющих в своей структуре те же типы ДНК-связывающих доменов, что и перечисленные выше ТФ, например CREB (bZIP), GATA (zinc finger), MYOG (bHLH) и ELK1 (Ets домен), различались незначительно (табл. 2).

Одним из возможных объяснений выявленных различий может быть то, что для некоторых ТФ формирование тонких механизмов регуляции их генов-мишеней привело к отбору сайтов с узким диапазоном аффинности, что, несомненно, отразилось на структуре сайтов.

Таблица 1

Количество и длина природных и *in vitro* селектированных последовательностей сайтов связывания транскрипционных факторов в матрицах, использованных для сравнения

Транскрипционный фактор	Кол-во последовательностей сайтов ТФ, использованных для построения матрицы		Длина последовательностей сайтов ТФ, использованных для сравнения, п. н.	
	природные сайты	*in vitro* селекция	природные сайты (max†/min‡)	*in vitro* селекция
AP2	43	185	50/9	20
AHR/ARNT	16	24	50/7	13
CEBPA	48	81	50/12	16
CEBPB	48	99	50/12	16
c-MYB	16	28	50/9	14
MYC/MAX	22	26	50/9	26
EGR1	23	55	50/9	21
ELK1	13	18	50/12	26
ETS1	59	15	50/10	10
GATA1	45	25	50/8	10
GATA2	27	53	50/7	20
GATA3	12	67	50/7	20
HMG1Y	14	15	50/10	20
HSF1	31	41	50/12	27
HSF2	13	33	50/14	27
MEF2	10	104	50/10	40
MYOD	13	24	50/7	10
MYOG	19	44	50/10	14
PDX1	10	30	50/8	10
PEA3	10	36	50/9	26
PPAR/RXR	39	72	50/14	27
USF	52	31	50/10	20
SOX5	19	23	50/7	26
SOX9	9	73	50/9	26
PU.1	23	57	50/13	16
SREBP1*	38	30	50/8	16
SREBP1**	15	7	50/10	16
SRF	13	46	50/10	26
YY1	22	55	50/10	15

Примечание. * – SRE тип сайта; ** – E-box тип сайта; † – максимальная длина последовательности сайта, для которой проведена оценка точности распознавания; ‡ – соответствует длине, при которой различие Кульбака – Лейблера матриц минимально.

Таблица 2

Оценка сходства и точности распознавания природных и *in vitro* селектированных сайтов связывания транскрипционных факторов эукариот

Транскрипционный фактор	Длина коровой последовательности сайтов, п. н. ‡	Оптимальная для распознавания длина сайта, п. н.	Расстояние Кульбака – Лейблера	Увеличение точности распознавания при опт. длине сайтов †
AP2	9	49	0,12	5,06
AHR/ARNT	7	9	0,18	2,14
C/EBPα	12	14	**0,23**	2,17
C/EBPβ	12	13	**0,27**	**1,16**
c-MYB	9	14	0,15	4,30
MYC/MAX	9	13	0,20	7,10
EGR1	9	49	**0,32**	33,0
ELK1	12	44	0,16	18,8
ETS1	10	49	0,17	18,2
GATA1	8	46	0,08	17,6
GATA2	7	49	0,19	5,06
GATA3	7	**11**	0,10	2,57
HMG1Y	10	19	0,17	**1,24**
HSF1	12	47	0,19	25,5
HSF2	14	45	0,16	23,3
MYOD	7	16	**0,26**	8,82
MYOG	10	32	0,18	3,08
PDX1	8	10	**0,21**	2,16
PEA3	9	36	**0,25**	16,3
PPAR	14	44	0,17	8,84
USF	10	31	0,13	3,26
SOX5	7	48	0,16	**162**
SOX9	9	30	0,18	18,3
PU1	13	28	0,14	27,4
SREBP1*	8	16	**0,22**	9,19
SREBP1**	10	17	0,09	2,79
SRF	10	38	0,18	32,6
YY1	10	25	0,18	2,44

Примечание. * – SRE тип сайта; ** – E-box тип сайта; ‡ – соответствует длине, при которой различие Кульбака – Лейблера матриц минимально; † – соответствует значению отношения точности распознавания сайта при оптимальной длине к точности его распознавания при учете только кора.

При искусственной селекции сайты отбираются в зависимости от условий эксперимента, которые могут быть настроены на отбор как высоко-, так и низкоаффинных сайтов, что может не соответствовать сформированной в результате эволюции структуре сайта. В этом смысле показательна ситуация с ТФ SREBP1, для которого исследователи выявили два типа сайтов, значительно различающихся по структуре. Именно те сайты, через которые осуществляется специфическая регуляция транскрипции генов в зависимости от уровня холестерина в клетке (SRE тип), сильнее отличаются от искусственных сайтов (уровень сходства 0,22), нежели те, которые участвуют в широком спектре регуляторных событий (E-box тип). Для этого типа сайтов показан высокий уровень сходства (0,09) с *in vitro* селектированными последовательностями (табл. 2). В целом, эти данные свидетельствуют о достаточно высоком уровне сходства природных и искусственных сайтов связывания ТФ эукариот.

Хотелось бы отдельно рассмотреть те случаи, когда для селекции последовательностей *in*

vitro использовали ТФ, различные по видовому происхождению. В базе имеются такие данные для пяти факторов: ETS1, c-Myb, GATA1, GATA2 и GATA3. Для сравнения использовали матрицы природных сайтов, которые построены на основе сайтов связывания человека, мыши и цыпленка. Как видно из табл. 3, для трех типов матриц (c-Myb, GATA2 и GATA3) уровень сходства высок и не зависит от видового происхождения фактора, использованного для селекции. Что же касается ETS1 и GATA1, то, согласно нашим оценкам, корректность результата поиска будет зависеть от видового происхождения ТФ, использованного для селекции сайтов. Особенно это значимо для GATA1. Анализ матриц показал, что для обоих ТФ в популяции природных сайтов, независимо от их видового происхождения, присутствует практически один тип кора, gata для GATA1 (рис. 1, *а*) и ggaa для ETS1. В селекционном эксперименте выявлены два значимых типа кора, и частота выявления второго кора, gatt для GATA1 (рис. 1, *б*, *в*) и ggat для ETS1, зависит от видового происхождения фактора. До 40 % последовательностей содержат второй тип кора при использовании для селекции сайтов GATA1 мыши (рис. 1, *в*) или ETS1 цыпленка.

Как видно из табл. 3, именно появление второго значимого нуклеотида в коровой последовательности сайта связывания сильно отражается на сходстве природных и *in vitro* селектированных матриц (рис. 1, пара сравниваемых матриц *а*/*б*, расстояние 0,0793; пара *а*/*в*, расстояние 0,3511), поскольку разделение сайтов в этой матрице только по этому признаку существенно увеличивает уровень сходства для gata-сайтов (рис. 1, пара *а*/*г*, расстояние 0,1618) и снижает таковой для gatt-сайтов (рис. 1, пара *а*/*д*, расстояние 0,4750). Это означает, что использование последних для распознавания потенциальных сайтов в геноме приведет к обнаружению большого числа не характерных для *in vivo* сайтов. Насколько широко распространено это явление, мы пока сказать не можем. Однако его наличие ставит под сомнение корректность создания широкомасштабных баз потенциальных сайтов

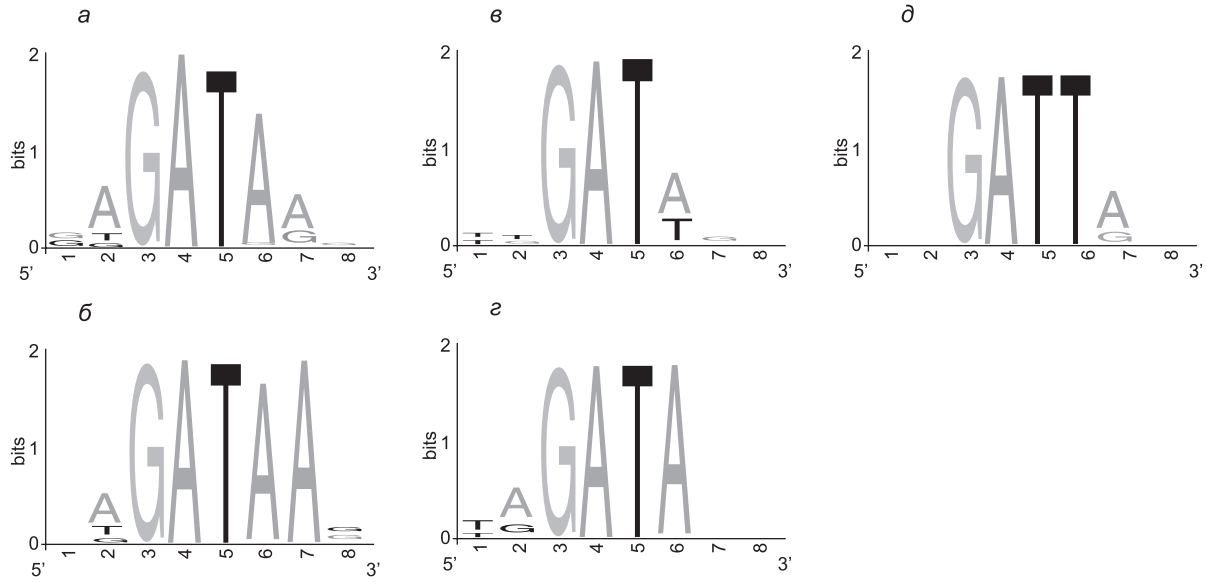


Рис. 1. Графическое изображение структуры сайта GATA1, построенное с помощью программы WebLogo (http://weblogo.berkeley.edu/):

а – матрица природных сайтов ТФ GATA1, полученная на основе последовательностей из генов человека (25), мыши (10), крысы (6) и цыпленка (6); *б* и *в* – матрицы искусственных сайтов, полученные в результате селекции с GATA1 цыпленка и мыши соответственно; *г* и *д* – матрицы искусственных сайтов с gata и gatt кором соответственно, полученные в результате селекции с GATA1 мыши. Цифры в скобках указывают количество проанализированных последовательностей природных сайтов из геномов соответствующих организмов. Ось ординат – степень консервативности нуклеотида в последовательности сайта (показана высотой буквы), ось абсцисс – позиция нуклеотида в матрице от 5′ конца сайта.

Таблица 3

Расстояние Кульбака – Лейблера между матрицами, построенными на основе природных и *in vitro* селектированных сайтов, полученных в результате селекции с ТФ различного видового происхождения

Тип матрицы	Видовое происхождение ТФ	Расстояние Кульбака – Лейблера	Длина сайтов в матрице, п. н.	Число сайтов связывания ТФ в сравниваемых парах матриц	
				in vitro селекция	природные
GATA1	Цыпленок	0,0793	8	25	47
GATA1	Мышь	**0,3511**	7	69	47
GATA1 (gata-core)	Мышь	0,1618	8	42	47
GATA1 (gatt-core)	Мышь	**0,4750**	8	24	47
GATA2	Цыпленок	0,0988	7	49	14
GATA2	Человек	0,1864	7	53	14
GATA3	Цыпленок	0,0996	7	67	12
GATA3	Человек	0,0980	7	63	12
ETS1	Цыпленок	**0,2251**	10	59	40
ETS1	Мышь	0,1731	10	15	40
c-Myb	Мышь	0,1510	9	28	16
c-Myb	Цыпленок	0,1335	11	49	16

связывания ТФ (Marinescu *et al.*, 2005) без предварительного изучения структуры их сайтов связывания.

Оценка оптимальной длины сайтов связывания транскрипционных факторов, при которой достигается максимальная точность распознавания их потенциальных сайтов

Так как длина фланговых последовательностей сайтов связывания ТФ влияет на точность их распознавания (Levitsky *et al.*, 2007), мы решили оценить оптимальную для распознавания длину сайтов связывания тех ТФ, для которых существуют выборки искусственных сайтов. Для этого мы использовали метод весовых матриц (Levitsky *et al.*, 2007) и выборки природных сайтов с увеличенными фланговыми последовательностями. Нами были построены методы динуклеотидных весовых матриц, критерием отбора длины которых стала оценка точности предсказания с помощью jack-knife теста. Не для всех 35 выборок сайтов связывания ТФ удалось набрать достаточное количество необходимых последовательностей, поэтому в данном исследовании использовано только 28 выборок. Для этих типов сайтов (табл. 2) приведены результаты анализа сходства матриц (расстояние Кульбака – Лейблера), построенных на основе коровых последовательностей природных и *in vitro* селектированных сайтов, а также изменения точности распознавания природных сайтов (по отношению ошибок второго рода) при увеличении длины их последовательности до оптимальной, т. е. такой, когда ошибка второго рода (перепредсказание) была минимальной.

Как следует из данных (табл. 2), оптимальная для распознавания длина сайтов связывания для всех исследованных ТФ превышала длину коровой последовательности сайта и только у шести ТФ (AHR, C/EBPα, C/EBPβ, MYC/MAX, GATA3 и PDX1) это превышение было незначительным (на 1–4 нуклеотида). При этом точность распознавания, изменение которой оценивали по отношению ошибок перепредсказания, увеличилась в два раза и более для большинства исследованных ТФ. А в случае ТФ SOX5 точность возросла на 2 порядка.

Исключение составили только ТФ C/EBPβ и HMG1Y. Однако и для этих ТФ увеличение длины флангов их сайтов связывания при построении матрицы улучшило точности их распознавания и снизило уровень перепред-

сказания почти на 20 % (табл. 2). Что касается столь высокого значения увеличения точности распознавания для ТФ SOX5, то анализ выборки сайтов связывания этого ТФ показал, что, в отличие от выборок других ССТФ, она содержит большое количество высокогомологичных сайтов, что, как показано нами, может приводить к завышению оценки точности распознавания. Удаление таких сайтов из выборки существенно (в три раза) снизило точность распознавания сайта.

Полученные данные позволяют предположить, что фланговые последовательности сайтов связывания практически всех исследованных ТФ эукариот содержат дополнительную информацию, необходимую для более точного распознавания их потенциальных сайтов. Например, согласно данным, представленным в табл. 2, длина коровой последовательности сайтов связывания ТФ SRF, построенной на основе выборки искусственно селектированных последовательностей, равна 10 пн. На рис. 2 она соответствует десятибуквенному мотиву CCA(T/A)ATAAGG, представленному в позициях 17–26. Оказалось, что оптимальная для распознавания длина сайта, полученная на основе анализа выборки геномных последовательностей сайтов связывания этого фактора равна 38 п. н.

Как видно на рис. 2, эта последовательность за пределами консенсуса как в 5′-, так и 3′-фланкирующих участках обогащена короткими консервативными кластерами нуклеотидов (позиции 2–4, 9, 11, 12, 15, 29, 32–35, 39). Отметим, что, хотя во всех случаях уровень консерватизма нуклеотидов оказался более низким, чем в коровой части сайта, точность распознавания сайтов связывания ТФ SRF на основе расширенной матрицы длиной 38 нуклеотидов оказалась в 32 раза выше, чем при распознавании на основе матрицы, соответствующей коровой последовательности длиной 10 нуклеотидов (см. табл. 2).

Консервативные участки в выравнивании отображены в виде стопки букв, причем высота каждой буквы пропорциональна частоте ее встречаемости в данной позиции, а общая высота стопки пропорциональна консерватизму последовательности в данной позиции, выраженной в битах (ось ординат). Номера по оси абсцисс соответствуют номеру позиции в выравнивании. Коровая часть сайта соответствует позициям 17–26.

Отметим, что рост точности распознавания при оптимизации матриц происходит благодаря двум факторам: (1) привлечению динуклеотидной, а не мононуклеотидной статистики для построения матриц и (2) собственно наращива-

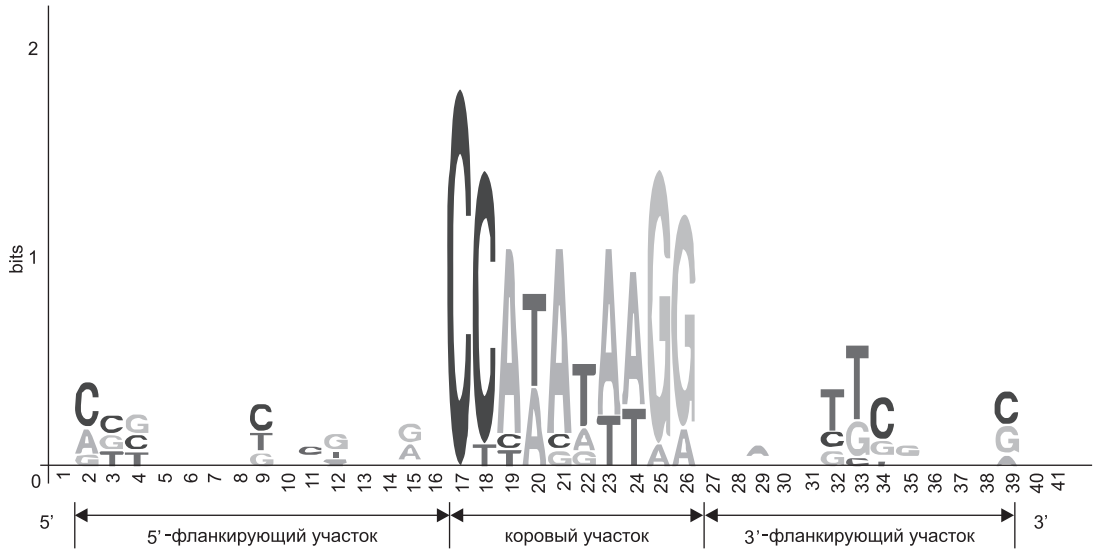


Рис. 2. Представление консервативных нуклеотидов в пределах коровой части сайта связывания фактора SRF и его фланкирующих участках, построенное с помощью программы WebLogo (http://weblogo.berkeley.edu/).

нию длины матрицы за счет менее консервативных флангов. Преимущество динуклеотидных матриц по сравнению с мононуклеотидными подтверждено как давними работами, так и современными исследованиями (Zhang, Marr, 1993; Gershenson *et al.*, 2005; Siddharthan, 2010; Nahdi, Ioshikhes, 2012). Нами ранее было показано, что привлечение фланкирующих последовательностей существенно повышает точность методов распознавания (Levitsky *et al.*, 2007; 9 типов ССТФ). Недавнее применение использованной в настоящей работе технологии учета консервативного контекста в последовательностях, фланкирующих кор сайта, к последовательностям, извлеченным из эксперимента по массовому секвенированию сайтов связывания ТФ, показывает, что длина динуклеотидной матрицы, обеспечивающей наивысшую точность, составляет около 30 нт (Kulakovskiy *et al.*, 2013; ССТФ FoxA). Хотя фланкирующие последовательности коровых районов сайтов существенно менее консервативны по сравнению с кором (рис. 2), их привлечение позволяет правильно распознавать неверные предсказания, выявляемые с помощью мононуклеотидных матриц короткой длины (10 нуклеотидов и менее).

Таким образом, проведенное сравнение структуры сайтов связывания ТФ, полученных из разных источников, позволило сделать заключение: несмотря на то что примерно для 80 % ТФ (из исследованных) показано высокое сходство коровых последовательностей природных и искусственных сайтов, использование выборок *in vitro* селектированных сайтов связывания ТФ для разработки компьютерных методов распознавания в геномных последовательностях их потенциальных сайтов не вполне корректно. Ограниченная длина селектированных сайтов (табл. 2) резко снижает точность их распознавания и приводит к выявлению значительного числа неверно предсказанных сайтов. Более того, для 20 % исследованных ТФ их *in vitro* селектированные СС имеют более широкий спектр допустимых значимых нуклеотидов в коровой структуре сайта (см. рис. 1, табл. 2). Их использование при распознавании также может привести к увеличению числа неверно предсказанных сайтов (росту ошибки перепредсказания) за счет последовательностей, вообще не встречающихся среди природных сайтов. Причина последнего, вероятно, кроется в том, что для природных сайтов связывания ряда ТФ характерна более высокая частота встречаемости среднеаффинных сайтов, нежели среди выявленных *in vitro* (Khlebodarova *et al.*, 2006), что представляет результат отбора природных сайтов скорее по функциональному критерию, чем по критерию аффинности к белку.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке гранта РНФ № 14-24-00123.

Авторы выражают благодарность Н.Л. Подколодному за ценные замечания при обсуждении результатов работы.

ЛИТЕРАТУРА

Aerts S., Van Loo P., Thijs G. *et al.* Computational detection of cis -regulatory modules // Bioinformatics. 2003. V. 19. Suppl 2. P. ii5–14.

Blackwell T.K., Weintraub H. Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection // Science. 1990. V. 250. P. 1104–1110.

Bryne J.C., Valen E., Tang M.H. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update // Nucl. Acids Res. 2008. V. 36. P. D102–D106.

Chen L., Wu G., Ji H. hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data // Bioinformatics. 2011. V. 27. P. 1447–1448.

Djordjevic M. SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways // Biomol. Eng. 2007. V. 24. P. 179–189.

Ehret G.B., Reichenbach P., Schindler U. *et al.* DNA binding specificity of different STAT proteins. Comparison of *in vitro* specificity with natural target sites // J. Biol. Chem. 2001. V. 276. P. 6675–6688.

Gershenzon N.I., Stormo G.D., Ioshikhes I.P. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites // Nucleic Acids Res. 2005. V. 33. P. 2290–2301.

Grote A., Klein J., Retter I. *et al.* PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes // Nucl. Acids Res. 2009. V. 37. P. D61–D65.

Hardenbol P., Wang J., Van Dyke M. Identification of preferred hTBP DNA binding sites by the combinatorial method REPSA // Nucl. Acids Res. 1997. V. 25. P. 3339–3344.

Khlebodarova T.M., Podkolodnaya O.A., Oshchepkov D.Y. *et al.* ARTSITE DATABASE: Structures of natural and *in vitro* selected transcription factor binding sites //

Bioinformatics of Genome Regulation and Structure II. Ed. By N. Kolchanov and R. Hofestaedt, Springer Science+Business Media, Inc., 2006. P. 55–65.

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. 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 // Intelligent Data Analysis, 2008. V. 12. No. 5. P. 443–461.

Kulakovskiy I., Levitsky V., Oshchepkov D. *et al.* From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites // J. Bioinform. Comput. Biol. 2013. V. 11. P. 1340004.

Lescot M., Dehais P., Thijs G. *et al.* PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences // Nucl. Acids Res. 2002. V. 30. P. 325–327.

Levitsky V.G., Ignatieva E.V., Ananko E.A. *et al.* Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions // BMC Bioinformatics. 2007. V. 8. P. 481.

Liu X., Yu X., Zack D.J. *et al.* TiGER: a database for tissue-specific gene expression and regulation // BMC Bioinformatics. 2008. V. 9. P. 271. doi: 10.1186/1471-2105-9-271.

Matys V., Fricke E., Geffers R. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles // Nucl. Acids Res. 2003. V. 31. P. 374–378.

Matys V., Kel-Margoulis O.V., Fricke E. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes // Nucl. Acids Res. 2006. V. 34. P. D108–D110.

Munch R., Hiller K., Barg H. *et al.* PRODORIC: prokaryotic database of gene regulation. Nuc. Acids Res. 2003. V. 31. P. 266–269.

Nandi S., Ioshikhes I. Optimizing the GATA-3 position weight matrix to improve the identification of novel binding sites // BMC Genomics. 2012. V. 13. P. 416.

Newburger D.E., Bulyk M.L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions // Nucl. Acids Res. 2009. V. 37. P. D77–D82.

Pollock R., Treisman R. A sensitive method for the determination of protein-DNA binding specificities // Nucl. Acids Res. 1990. V. 18. P. 6197–6204.

Ponomarenko J.V., Orlova G.V., Ponomarenko M.P. *et al.* SELEX_DB: a database on *in vitro* selected oligomers adapted for recognizing natural sites and for analyzing both SNPs and site-directed mutagenesis data // Nucl. Acids Res. 2000. V. 28. P. 205–208.

Portales-Casamar E., Thongjuea S., Kwon AT. *et al.* JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles // Nucl. Acids Res. 2010. V. 38. P. D105–D110.

Praz V., Perier R., Bonnard C., Bucher P. The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data // Nucl. Acids Res. 2002. V. 30. P. 322–324.

Robison K., McGuire A.M., Church G.M. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome // J. Mol. Biol. 1998. V. 284. P. 241–254.

Roulet E., Bucher P., Schneider R. *et al.* Experimental analysis and computer prediction of CTF/NFI transcription factor DNA binding sites // J. Mol. Biol. 2000. V. 297. P. 833–848.

Roulet E., Busso S., Camargo A.A. *et al.* High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites // Nat. Biotechnol. 2002. V. 20. P. 831–835.

Sandelin A., Alkema W. Engstrom P., Wasserman W.W., Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles // Nucl. Acids Res. 2004. V. 32. P. D91–94.

Shultzaberger R.K., Schneider T.D. Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX // Nucl. Acids Res. 1999. V. 27. P. 882–887.

Siddharthan R. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix // PLoS One. 2010. V. 5. P. e9722.

Wang J., Lu J., Gu G., Liu Y. *In vitro* DNA-binding profile of transcription factors: methods and new insights // J. Endocrinol. 2011. V. 210. P. 15–27.

Wingender E., Chen X., Fricke E. *et al.* The TRANSFAC system on gene expression regulation // Nucl. Acids Res. 2001. V. 29. P. 281–283.

Wright W.E., Binder M., Funk W. Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site // Mol. Cell. Biol. 1991. V. 11. P. 4104–4110.

Zhang M.Q., Marr T.G. A weight array method for splicing signal analysis // Comput. Appl. Biosci. 1993. V. 9. P. 499–509.

Zhao F., Xuan Z., Liu L., Zhang M.Q. TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies // Nucl. Acids Res. 2005. V. 33. P. D103–D107.

EFFECT OF FLANKING SEQUENCES ON THE ACCURACY OF THE RECOGNITION OF TRANSCRIPTION FACTOR BINDING SITES

T.M. Khlebodarova[1], D.Yu. Oshchepkov[1], V.G. Levitsky[1, 2], O.A. Podkolodnaya[1], E.V. Ignatieva[1], E.A. Ananko[1], I.L. Stepanenko[1], N.A. Kolchanov[1, 2]

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

Summary

The development of *in vitro* methods produced new experimental information on protein binding to DNA, which is accumulated in databases and used in studies of mechanisms regulating gene expression and in the development of computer-assisted methods of binding site recognition in pro- and eukaryotic genomes. However, it is still questionable to what extent sequences selected in vitro reflect the actual structures of natural transcription factor (TF) binding sites. The Kullback – Leibler divergence was applied to the comparison of frequency matrices of TF binding sites constructed on samples of artificially selected sequences and natural sites. Core sequences of natural and artificial sites showed high similarity for 80 % of all TFs studied. For 20 % of TFs, binding site sequences selected in vitro had a broader range of permissible significant nucleotides not found in natural sites. The optimum lengths of DNA sequences including natural binding sites, at which they are recognized most accurately, were estimated by the weight matrix method. For approximately 80 % of the TFs studied, the optimum binding site length notably exceeded the lengths of the core sequences, as well as the lengths of *in vitro* selected sites. The detected features of *in vitro* selected TF binding sites impose constraints on their use in the development of computer-assisted methods of the recognition of candidate sites in genomic sequences.

Key words: transcription factors, binding sites, frequency and weight matrices, *in vitro* selected sequences.

УДК 573.2

КОМПЬЮТЕРНЫЙ АНАЛИЗ И ФУНКЦИОНАЛЬНАЯ АННОТАЦИЯ САЙТОВ СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ AP2/ERF В ГЕНОМЕ *ARABIDOPSIS THALIANA* L.

© 2014 г. **О.А. Черных[1], В.Г. Левицкий[1, 2], Н.А. Омельянчук[1], В.В. Миронова[1, 2]**

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

Поступила в редакцию 12 сентября 2014 г. Принята к публикации 1 октября 2014 г.

У растений этилен участвует как в регуляции процессов развития, так и в ответе на стрессовые воздействия. Сигнал с рецепторов этилена активирует гены одного из самых больших семейств транскрипционных факторов, APETALA2/ETHYLENE RESPONSE FACTORs (ERFs). Сайты связывания ERF транскрипционных факторов содержат специфический GCCGCC-мотив и называются GCC-боксами. В настоящей работе две компьютерные программы для предсказания сайтов связывания транскрипционных факторов (oPWM и SiteGA) применены для анализа последовательностей экспериментально подтвержденных GCC-боксов. Распознаны GCC-боксы и исследовано их распределение в геноме *Arabidopsis thaliana* L. Проведены функциональная аннотация распознанных GCC-боксов и анализ их роли в ответе на фитогормон этилен.

Ключевые слова: этилен, транскрипционные факторы, сайты связывания, арабидопсис.

ВВЕДЕНИЕ

Фитогормон этилен является единственным газообразным гормоном растений. Он участвует в контроле большого числа важнейших процессов в жизни растительного организма, начиная от роста и развития до ответов на стрессы биотического и абиотического происхождения (Shakeel *et al.*, 2013). За последние годы были идентифицированы многие ключевые компоненты сигнального пути от рецепторов этилена до связывания транскрипционных факторов (ТФ) с промоторами генов-мишеней (Fujimoto *et al.*, 2000).

Действующая модель пути передачи этиленового сигнала включает в себя следующие семейства генов: *ETR1-like*/*ETR2-like*, *CTR1, EIN2, EIN3/EIL1* и *ERF* (Ethylene Responsive Elements) (Shakeel *et al.*, 2013). Передача сигнала этилена осуществляется по механизму негативной регуляции, т. е. в отсутствие этилена все компоненты пути находятся в активированном состоянии, блокируя реакцию на этилен на транскрипционном уровне. Рецепторы этилена представлены двумя подсемействами (ETR1-like и ETR2-like) и расположены на мембране эндоплазматического ретикулума (Bleecker, Kende, 2000). Рецепторы находятся в комплексе с серин/треониновой протеинкиназой CTR1. При связывании этилена CTR1 становится неактивной, что приводит к дефосфорилированию и протеолитическому расщеплению следующего компонента пути EIN2 (Ju, Chang, 2012; Qiao *et al.*, 2012). EIN2 служит главным позитивным регулятором, от которого сигнал этилена передается к ТФ EIN3/EIL1 и ERF, локализованным в ядре.

Транскрипционные факторы активируются каскадно, т. е. один за другим. EIN3/EIL1 активирует ТФ ERF, который, в свою очередь, связывается со специфической последовательностью, называемой GCC-боксом и имеющей консенсус GCCGCC (Ohme-Takagi, Shinshi, 1995; Solano *et al.*, 1998). Консенсус GCCGCC

консервативен для многих видов растений и широко используется для распознавания ERF мишеней (Shinshi *et al.*, 1995; Fujimoto *et al.*, 2000; Choudhury *et al.*, 2008).

К настоящему моменту в промоторах отдельных генов различных видов растений исследованы GCC-боксы, например, в работах Solano с соавт. (1998) и Nakano с соавт. (2006) изучено влияние нуклеотидных замен в их последовательности. GCC-боксы находят в промоторах некоторых связанных с ответом на патогенные воздействия генов, многие из которых являются этилен-зависимыми (Stepanova, Ecker, 2000). Тем не менее только для небольшого числа потенциальных генов-мишеней экспериментально показана функциональность GCC-боксов в их промоторах в реакции на этилен.

В нашей работе с помощью компьютерных программ oPWM и SiteGA, разработанных для предсказания сайтов связывания ТФ, проведен поиск GCC-боксов в геноме *Arabidopsis thaliana* и изучены особенности их распределения. Мы также проанализировали связь между наличием в промоторах GCC-боксов и этилен-чувствительностью генов, используя данные полногеномных микрочип-экспериментов, в которых было исследовано изменение экспрессии генов в ответ на воздействие этилена.

МАТЕРИАЛЫ И МЕТОДЫ

Создание выборок GCC-боксов

На основе данных из публикаций нами были составлены обучающая выборка (1) и позитивная выборка (2). Обучающая выборка (1) состоит из 24 последовательностей, найденных в промоторах генов семи видов растений (табл. 1), и содержит нуклеотидные последовательности GCC-боксов, подтвержденных в экспериментах, исследовавших этилен-чувствительность генов. Позитивная выборка (2) составлена на основе данных из публикаций и содержит промоторы 54 генов *A. thaliana*, экспрессировавшихся в ответ на воздействие этилена. Отбор генов произведен по нескольким условиям: эффект этилена на экспрессию генов проверен методом количественного ПЦР (qRT-PCR), в публикации имеются точные данные по увеличению или уменьшению экспрессии генов под действием этилена. Экспрессия генов должна быть установлена в течение 3–24 ч после обработки этиленом. Гены позитивной выборки (2) не повторяют состав обучающей выборки (1).

Распознавание потенциальных GCC-боксов

Выборка 24 экспериментально подтвержденных GGC-боксов собрана по литературным данным. Выравнивание сайтов было произведено с помощью разработанного нами подхода на основе генетического алгоритма (Mironova *et al.*, 2014). Для распознавания использованы разработанные нами ранее методы oPWM и SiteGA (Levitsky *et al.*, 2007).

Метод oPWM – модификация канонического метода весовых матриц, это позволяет рассматривать динуклеотидный контекст и в полной мере привлекать фланкирующие районы. Отличительная черта метода SiteGA состоит в учете зависимостей между удаленными позициями сайтов, что не предусмотрено в методе весовых матриц.

Лого, моно- и динуклеотидное лого на рис. 1, *а*, *б* представляют консервативные контекстные характеристики GCC-боксов, выявляемые с помощью метода oPWM. Согласно методике SiteGA, для распознавания GCC-боксов нами было отобрано $N = 60$ локально-позиционированных динуклеотидов (ЛПД), каждый из которых представлен позициями начала и конца участка и типом динуклеотида.

Для расчета наиболее значимых зависимостей между ЛПД мы провели следующий анализ. Для $60 \times (60 - 1)/2$ ЛПД были отобраны корреляции пар, значимость которых задана условиями $p < 0{,}05$, $p < 0{,}01$ или $p < 0{,}001$. Затем для каждого из этих условий рассчитана плотность ЛПД (рис. 1, *в*). Для одной корреляции вклад каждой позиции учитывался как $0{,}5/L_1$ и $0{,}5/L_2$ (L_1 и L_2 означают длины участков двух ЛПД). В данной работе проанализированы сайты, распознанные обоими методами.

Анализ данных микрочип-эксперимента

Выборки данных микрочип-экспериментов взяты из базы NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/). Для GCC-боксов всего было проанализировано шесть экспериментов

Таблица 1

Обучающая выборка нуклеотидных последовательностей экспериментально подтвержденных GCC-боксов

Ген	Сайт	Вид	Статья
HLS1	TTAACGCAGACATAGCCGCCATTTTCAACTTCTCACTC	*Arabidopsis thaliana*	Fujimoto *et al.*, 2000
PDF1.2a	ATTCAGATTAACCAGCCGCCCATGTGAACGATGTAGCA		Zarei *et al.*, 2011
Chn48	ATAAAAAGGTAAGAGCCGCCACATAATATATGTAACCT	*Nicotiana tabacum*	Shinshi *et al.*, 1995
ACS3	CTATTACATAGTAAGCCGCCCACCGTATCTCAAAATAG		Zhang *et al.*, 2009
Cel5	GTCACATTTTTATCGTCGCCGTGAATTGTGGTATAGTA	*Lycopersicon esculentum*	Tournier *et al.*, 2003
prb-1b	CAAGTATGACTAATGGCGGCTCTTATCTCACGTGATG	*Nicotiana tabacum*	Sessa *et al.*, 1995
PR-5	GGCCTTTACATTTAGCCGCCTAGCTCTATCTTTACCAA	*Nicotiana sylvestris*	Sato *et al.*, 1996
gln2	GCCTCCTCATTAGAGCCGCCACTAAAATAAGACCGATC	*Nicotiana tabacum*	Grimmig *et al.*, 2003
ATHCHIB	TTGATCACGAACCCGCCGCCTCATATTCATAATTAAAG	*Arabidopsis thaliana*	Samac *et al.*, 1990
CH5B	TTCACGCTTGGGAAGCCGCCGGGGTGGGCCCGCAGAAA	*Phaseolus vulgaris*	Solano *et al.*, 1998
CHN50	GGATGAAGCTAAAAGCCGCCAGTCTCACTAAGAAAAAT	*Nicotiana tabacum*	Nakano *et al.*, 2006
Osmotin	TCTATGTGCGAAAAGCCGCCATACTCCTATATAAACCA		
RD29B	AGAAACAAATGTATGTCGGCCAACAAGTTAATTTGGGT	*Arabidopsis thaliana*	Cheng *et al.*, 2013
ELI3-2	CGGATTATGTCAACACCGCCATGGAACGGCTTGCAAAG		
GEA6	GAGAGAAGAATTACACCGACGATTTCACCATGAAGAGA		
LEA4	TATCTTGTTCTCTCGCCGACCAAGACTTGCTATAAATA		
COR15B	GAAAAAAAAAGCAGGTCGGCCATGAAATTGTGGCTACA		
COR47-DRE1	TCTTATTTCTTGAAGTCGGTAGATGAATATCATGATAT		
HSP101-DRE1	CTTTAATTTATACAAGCCTCCTTTTTTGTACATCTATTT		
HSP70-DRE	CTGAATTTTGACTTGCCGACTCCCCTGCTTGCTACTTT		
PDF1.2b	AGTCAGATTAACCAGCCGCCCATGCAAAGCCAAAGCAG	*Arabidopsis thaliana*	Wang *et al.*, 2013
AT2G37130	ACTTTCTTAATTATGGCGGCTGTAAATACATGTACAAT		
PMT1	TATTATATCGAGTTGCGCCCTCCACTCCTCGTGTCCAA	*Nicotiana tabacum*	Sears *et al.*, 2014
AtHAK5	TAAAAGTTTCAACAGCCGGCAATACGTGTTTGAGACGC	*Arabidopsis thaliana*	Son *et al.*, 2012
ACS	AACACGTCATTGTTGCCGCCAACACTGAAGCTTCCTAT	*Musa acuminata*	Choudhury *et al.*, 2008
ACO	GAGACCGATGGAAGCAGCCAAACTTGGTCCCCGATC		
BERF1	TCCTCCATCACTGTGCCGCCCGTGCTCGCCTCTCCCGG	*Hordeum vulgare*	Osnato *et al.*, 2010

GCC-бокс подчеркнут. Выборка использована для применения методов распознавания SiteGA и oPWM.

(табл. 2). Из данных микрочип-экспериментов были взяты значения экспрессии генов при контрольной обработке растения воздухом и обработке этиленом. При наличии нескольких реплик эксперимента было рассчитано среднее значение экспрессии. Для расчета значимости был проведен t-тест. Ген считался значимо регулируемым этиленом, если (1) отношение Ф уровней экспрессии обработанного и контрольного образцов было не менее/более заданного порогового значения (Ф = 1,5/0,667 для активации/репрессии) и (2) отличие средних уровней экспрессии для обработанных и контрольных реплик было значимым по t-тесту ($p < 0,05$). Долю генов, содержащих предсказанные сайты и увеличивших или уменьшивших свою экспрессию под воздействием этилена, сравнили с долей всех генов в геномной выборке, имею-

щих такую же реакцию на этилен (табл. 3). Был проанализирован район [–300; +1] от старта инициации транскрипции (коровый промотор). Всего в анализ вошло 21 098 генов, для которых были аннотированы старты транскрипции (база ENSEMBL) и имелись данные экспериментов с микрочипами. Ассоциация между наличием GCC-бокса в коровом промоторе и этилен-зависимой экспрессией признавалась значимой, если была выявлена, как минимум, в двух экспериментах. Этот порог был выбран по биномиальному распределению (см. ниже).

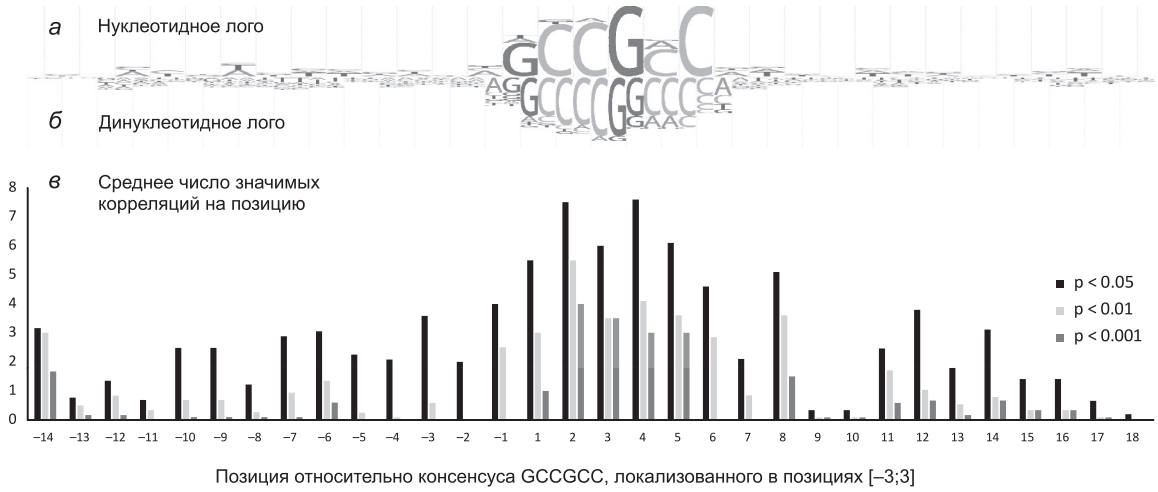


Рис. 1. Контекстные особенности GCC-сайтов, в которых функциональность GCC-боксов подтверждена экспериментально, выявленные методом oPWM (Levitsky *et al.*, 2007), осуществленным для моно- (*а*) и динуклеотидной (*б*) позиционно-весовой матрицы (PWM). Лого для PWMs рассчитаны в соответствии с Kulakovskiy и соавт. (2010) (*в*). Метод SiteGA (Levitsky *et al.*, 2007) выявил значимые корреляции между частотами локально-позиционированных динуклеотидов. По оси ординат показано среднее число значимых корреляций на позицию относительно консенсуса GCCGCC (ось абсцисс).

Таблица 2

Эксперименты с микрочипами, использованные для анализа ассоциаций наличия в промоторных районах генов GCC-боксов с этилен-чувствительной экспрессией этих генов

GEO ID	Условия обработки	Ткань	Ссылка
GSE14247	10 ppm этилен 4 ч	Трехнедельные растения	Qiao *et al.*, 2012
GSE39384	10 μM АЦК 1 ч	Семидневные проростки	Goda *et al.*, 2008
GSE39384	10 μM АЦК 30 мин		
GSE39384	10 μM АЦК 3 ч		
GDS3505	10 ppm этилен 4 ч	Трехдневные проростки	Stepanova *et al.*, 2007
GSE5174	10 ppm 4 ч	Этиолированные растения, выращенные в темноте	Olmedo *et al.*, 2006

Данные из базы NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/). Во всех экспериментах использованы данные по дикому фенотипу. Для проведения анализа взяты данные по изменению экспрессии генов при обработке растения воздухом (контроль) и этиленом. АЦК – 1-амино-циклопропан-1-карбоновая кислота – предшественник этилена.

Таблица 3

Схема для статистического анализа данных микрочип экспериментов с использованием углового преобразования Фишера для сравнения долей

Число генов	Наличие предсказанного GGC-бокса	
	да	нет
Увеличивших/уменьшивших уровень экспрессии	*a*	*b*
Не изменивших уровень экспрессии	*c*	*d*

Чтобы подтвердить статистическую значимость различий, посчитанных для выборок генов, мы использовали отношения *a/c* и *b/d* (табл. 3) между: 1) числом этилен-зависимых генов, значимо увеличивших или уменьшивших свою экспрессию в микрочиповом эксперименте (*a* и *b*) и 2) числом генов, не изменивших свою экспрессию в полногеномной выборке (*c* и *d*). Первая пропорция p_1 была рассчитана для выборки генов, в которых присутствовал потенциальный GCC-бокс. Вторая пропорция p_2 была рассчитана для генов, не содержащих этилен-чувствительные элементы.

Поскольку математическое ожидание значений *a* и *c* в некоторых случаях не превышало 10, для оценки значимости нами использован тест Фишера для сравнения долей (угловое преобразование Фишера для сравнения долей):

$$p_1 = 2 * arcsin\left(\sqrt{\frac{a}{a+c}}\right), p_2 = 2 * arcsin\left(\sqrt{\frac{b}{b+d}}\right).$$

Так как расчет отношений проводился для шести микрочип-экспериментов, для учета множественного сравнения мы применили биномиальное распределение, а именно рассчитали минимально необходимое количество микрочип-экспериментов для оценки найденной ассоциации как значимой ($p < 0{,}05$).

В результате расчет биномиального распределения

$$P(k) = \sum_{k}^{6} \frac{6!}{(6-k)!\,k!} 0.05^k 0.95^{6-k}$$

при $k = 3$ дает $P(3) = 0{,}0083 < 0{,}01$, тогда как $P(2) = 0{,}0394 < 0{,}05$. Следовательно, при $k = 2$, $k = 3$ имеем статистическую значимость, т. е. значимость изменения экспрессии в двух или более микрочип-экспериментах из шести можно рассматривать как неслучайное событие. Функциональная аннотация GCC-боксов проведена в системе AgriGO (Du *et al.*, 2010) с использованием инструмента SEA (Singular enrichment analysis).

РЕЗУЛЬТАТЫ

Распознавание GCC-боксов

Для распознавания GCC-боксов в геноме *A. thaliana* нами были созданы обучающая выборка (1), содержащая нуклеотидные последовательности экспериментально подтвержденных GCC-боксов, и позитивная выборка (2) этилен-чувствительных генов. Для полногеномного распознавания GCC-боксов мы применили две компьютерные программы, oPWM и SiteGA, основанные соответственно на методе весовых матриц и анализе локально-позиционированных динуклеотидов (Levitsky *et al.*, 2007). С помощью обеих программ исследованы нуклеотидные последовательности из обучающей выборки (1) (см. табл. 1). В результате анализа найдены особенности нуклеотидного контекста как в самом GCC-боксе, так и на его флангах (рис. 1).

Для распознавания был применен объединенный метод SiteGA&PWM, предполагающий распознавание потенциального сайта одновременно SiteGA и oPWM. С использованием позитивной выборки (2) были выбраны пороги обоих методов (0.934 SiteGA, 0.687 oPWM). При этом условии нами распознаются 15 из 27 сайтов обучающей выборки (1) (ошибка 1-го рода = 44 %). Как ошибку 2-го рода можно использовать вероятность распознавания сайтов для полного генома $1{,}9 \times 10^{-5}$.

Распределение GCC-боксов в геноме *A. thaliana*

Найденные контекстные особенности (рис. 1) были использованы для распознавания GCC-боксов во всем геноме *A. thaliana*. Всего GCC-боксы были распознаны в промоторах [–2000; +1] 941 гена, что составляет 3,5 % генома *A. thaliana*. Для анализа особенностей распределения GCC-боксов мы рассчитали их относительную плотность распределения в различных районах генома: 5′- и 3′НТР, транскриптах, экзонах, интронах и промоторных районах (рис. 2, *а*).

Неожиданным оказалось, что повышенная плотность GCC-боксов найдена не только в промоторных районах, но и в экзонах, большую часть которых составляют кодирующие. Наибольшая плотность GCC-боксов обнаружена в экзонах, высокая плотность также была характерна для 5′НТР. Для исследования функциональной важности найденных особенностей мы проанализировали плотность распределения GCC-боксов в различных областях (5′НТР, экзонах, интронах, промоторных районах и 3′НТР) генов из позитивной выборки (2), составленной из нуклеотидных последовательностей 54 генов, для которых реакция на этилен была показана экспериментально.

Для этих генов, как и в целом геноме, повышенная плотность GCC-боксов была найдена в экзонах, но в отличие от полногеномного распределения в выборке этилен-чувствительных генов значимое обогащение было найдено еще и в промоторных областях [–1000; +1]. GCC-боксы в 5′НТР генов позитивной выборки не найдены. Далее мы проанализировали более

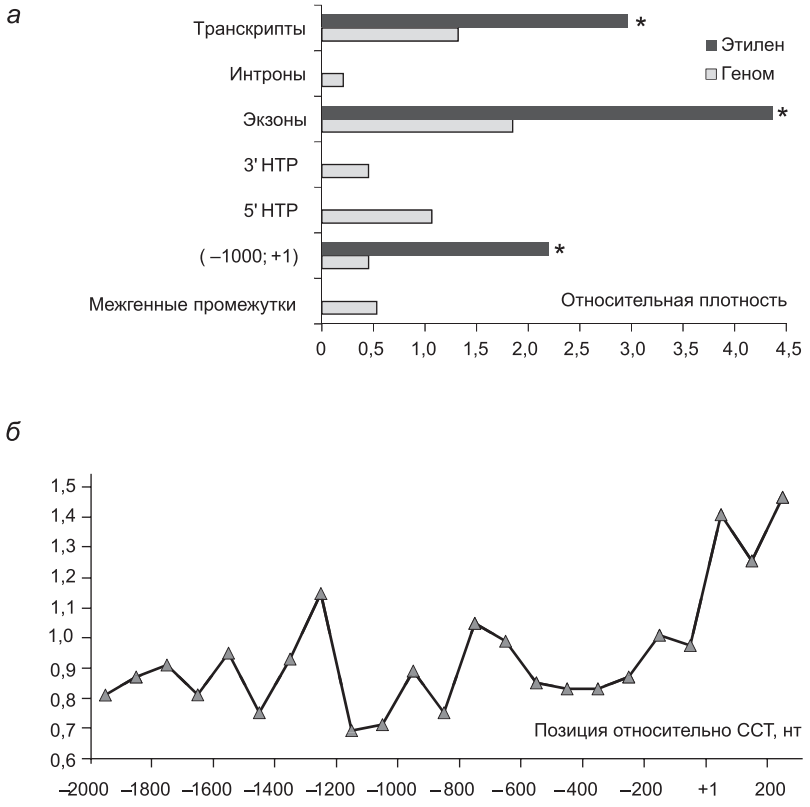


Рис. 2. Распределение GCC-боксов в геноме *A. thaliana*: *а* – плотность распределения GCC-боксов в разных областях генома для полного генома и позитивной выборки. Каждая плотность нормализована на среднюю плотность мотива в полном геноме. * отмечена значимость различий между плотностью в полном геноме и позитивной выборке по *t*-тесту (p <0,05); *б* – относительная плотность распределения GCC-боксов вдоль регуляторных областей [–2000; +200] в полном геноме.

детально плотность распределения GCC-боксов вдоль регуляторных областей генов *A. thaliana*. GCC-боксы оказались локализованы вдоль промоторов неравномерно – плотность распределения была выше вокруг старта инициации транскрипции (рис. 2, *б*). Из проведенного анализа можно сделать следующие выводы. Значимое обогащение предсказанных методами PWM и SiteGA сайтов связывания ERF факторов в регуляторных районах этилен-чувствительных генов свидетельствует об адекватности распознавания GCC-боксов комбинированием этих двух методов.

Распознанные в настоящей работе GCC-боксы оказались неравномерно распределены в геноме *A. thaliana*, их повышенная плотность в кодирующих участках генома отмечена не только при анализе генома в целом, но и для выборки этилен-чувствительных генов, у которых плотность распределения GCC-боксов в экзонах оказалась при этом даже значимо выше, чем в целом по геному (рис. 2, *а*).

Обогащение GCC-бокса в экзонах может быть объяснено высокой встречаемостью тринуклеотида GCC в кодирующих участках, так как он является кодоном для одной из самых распространенных аминокислот – аланина. Альтернативное объяснение этого результата: GCC-бокс может быть дуоном – кодоном, который кроме своего прямого назначения составляет часть сайта связывания ТФ (Stergachis *et al.*, 2013). Однако нами было показано, что доля генов, изменивших свою экспрессию в ответ на этилен и несущих GCC-боксы в экзонах, не превышает доли генов, изменивших свою экспрессию в ответ на этилен в целом по геному. Таким образом, функциональность GCC-боксов в экзонах генов не подтверждена.

Функциональная аннотация GCC-боксов

Функциональная аннотация GCC-боксов состояла из двух задач: (1) функциональная аннотация генов, содержащих в промоторах GCC-боксы, по геномной онтологии (ГО) и (2) поиск ассоциаций между наличием GCC-бокса в промоторе и этилен-чувствительностью генов. В анализе был использован список генов с предсказанными GCC-боксами в районах [–500; +1]. Выбор данного участка обоснован особенностями распределения GCC-бокса в 5′-регуляторных областях этилен-чувствительных генов и в среднем по геному, а именно: плотность распределения GCC-боксов в среднем по геному выше вокруг старта инициации транскрипции (рис. 2, *б*), при этом в 5′НТР этилен-чувствительных генов сайты не были обнаружены (рис. 2, *а*).

Функциональная аннотация генов, содержащих GCC-боксы

Для исследования процессов, связанных с экспрессией генов, содержащих потенциальные GCC-боксы, мы провели функциональную аннотацию генов, в промоторах которых они были распознаны. Число генов с GCC-боксами только в прямой ориентации – 2 029. Число генов, содержащих GCC-боксы в обратной ориентации, – 1 941. Число генов с GCC-боксами в обеих цепях – 3 971. Аннотация проводилась в системе AgriGO (Du *et al.*, 2010). В результате выявлен ряд терминов, значимо обогащенных для трех выборок генов, содержащих GCC-боксы в: (1) прямой (GCC+); (2) обратной (GCC–); или (3) любой ориентации (GCC+/–) относительной цепи транскрипции.

На рис. 3 представлена блок-схема терминов ГО, относящихся к функциям генов, содержащих GCC-боксы. Значимо обогащенные термины обозначены светло- и темно-серым ($p < 0{,}05$ и $p < 10^{-4}$ по тесту Бенджамини соответственно). Термины ГО, значимо обогащенные для генов GCC+ и GCC–, отличались. GCC+ оказались ассоциированы с метаболизмом азотных соединений, а GCC– с организацией клеточных компонентов. Однако, когда мы объединили гены, содержащие GCC-боксы в обеих ориентациях, были найдены дополнительные значимо обогащенные термины, в том числе связанные с постэмбриональным развитием ($p < 0{,}007$ по тесту Бенджамини).

Анализ взаимосвязи между наличием в промоторе GCC-бокса и этилен-чувствительностью гена

Для проверки функциональности потенциального GCC-бокса в промоторе и его связи с этилен-зависимой экспрессией генов нами

был проведен анализ данных шести микрочип-экспериментов, в которых было исследовано воздействие этилена на проростки *A. thaliana* (табл. 2). Для отдельного полногеномного эксперимента сравнили доли генов, увеличивших или уменьшивших свою экспрессию под воздействием этилена в среднем по геному с таковой для выборки генов, содержащих предсказанные GCC-боксы (табл. 4). Связь между наличием потенциальных GCC-боксов

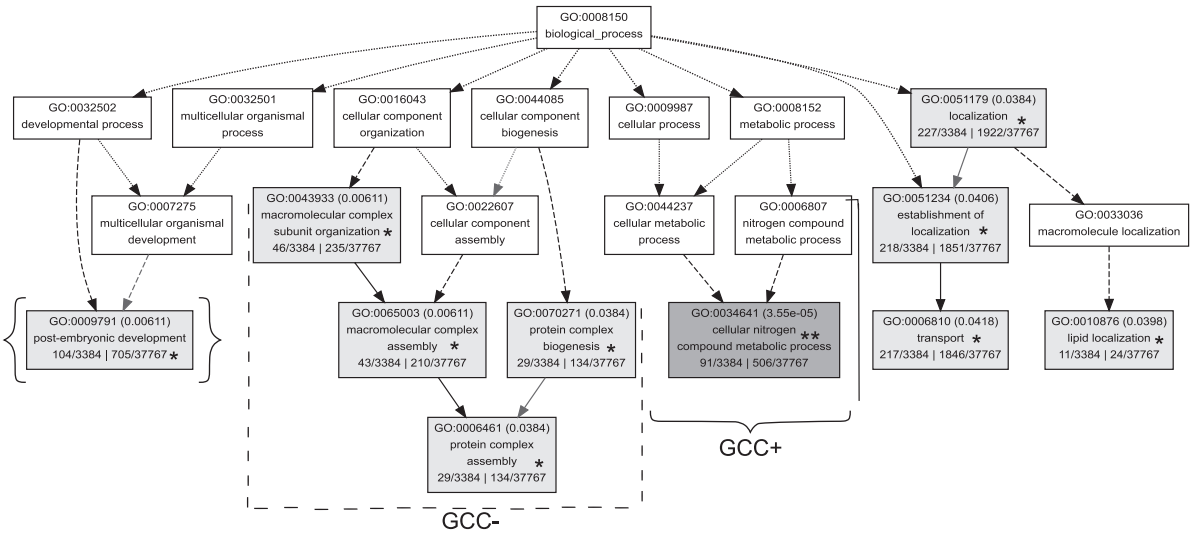


Рис. 3. Блок-схема терминов GO (Gene Ontology), значимо обогащенных для генов, содержащих GCC-боксы.

Фигурными скобками отмечены термины, обогащенные для GCC+ (сайт в прямой ориентации), штриховой линией – термины для GCC– (сайт в обратной ориентации). В двойных фигурных скобках – термины, связанные с развитием. Значимость обогащения по тесту Бенджамини: $p < 0{,}05$* (светло-серые блоки), $p < 10^{-4}$** (темно-серые). В блоках указано число распознанных генов (*x, y*) в формате *x*/3 384, *y*/37 767, где 3 384 – число аннотированных генов, содержащих GCC-бокс из 3 971 гена *A. thaliana* и поданных нами на анализ в систему AgriGO, а 37 767 – число генов, составляющих, по последним данным, полный геном *A. thaliana*.

Таблица 4

Ассоциация наличия GCC-боксов в промоторных районах генов ([–300; +1] относительно сайтов старта транскрипции) с этилен-зависимой активацией экспрессии генов

GEO ID	Цепь	Число генов с GCC-боксами, увеличившими экспрессию (*a*)	Число генов с GCC-боксами (*c*)	Отношения *a/c* | *b/d*, %	*p*-value
GSE14247	+/–	141	631	22,3 | 14,3	$9{,}0 \times 10^{-8}$
GSE14247	+	83	372	22,3 | 14,3	$3{,}1 \times 10^{-5}$
GSE14247	–	59	267	22,1 | 14,3	$4{,}6 \times 10^{-4}$
GDS3505	+/–	17	631	2,7 | 1,7	0,04
GSE3938 3 ч	–	7	267	2,6 | 0,9	$1{,}4 \times 10^{-2}$

Представлены значимые результаты для прямой (+), обратной (–) и обеих (+/–) цепей, полученные при анализе данных шести микрочип-экспериментов (см. табл. 2): *a* – число генов, содержащих в промоторе хотя бы один GCC-бокс и увеличивших свою экспрессию в ответ на воздействие этилена; *c* – число генов, содержащих GCC-бокс в промоторе, среди всех проанализированных генов генома (D = 21 098). Также указаны отношения *a/c* и *b/d*, с помощью которых были высчитаны значимости (*p*-value). Объяснение расчетов с величинами *a* – *d* см. в табл. 3.

и этилен-зависимой экспрессией признавали значимой, если значимые различия в экспрессии генов обнаруживались более чем в двух микрочип-экспериментах, что соответствует $p < 0,05$ по биномиальному распределению.

В результате проведенного анализа показано, что наличие GCC-боксов в промоторном районе [–300; +1] значимо ассоциировано с активацией экспрессии генов в ответ на воздействие этилена на растения (табл. 4). Случаев подавления экспрессии для генов, содержащих GCC-боксы, не выявлено. Мы также проанализировали, влияет ли ориентация GCC-бокса в промоторе гена на его этилен-чувствительность. Активация генов в ответ на воздействие этилена оказалась характерной для обратной ориентации (GCC–). Оба результата были выявлены в двух из шести микрочип-экспериментов (значимость $p < 0,05$ по биномиальному критерию). Для прямой ориентации (GCC+) была найдена одна ассоциация с активацией экспрессии в ответ на воздействие этилена, что по биномиальному распределению не значимо. Для сравнительной оценки функциональной значимости GCC-боксов в различных их ориентациях необходимо привлечь результаты дополнительных микрочип-экспериментов по оценке изменения экспрессии генов в ответ на воздействие этилена. Известно, что семейство ТФ ERF, специфически связывающихся с GCC-боксами, очень велико (125 генов у *A. thaliana*) (Nakano *et al.*, 2006) и лишь некоторые представители этого семейства изменяют свою экспрессию в ответ на этилен. Выявленные ассоциации свидетельствуют в пользу адекватности примененных нами методов распознавания GCC-боксов, так как с их использованием получены значимые результаты как по функциональной аннотации, так и по анализу данных микрочип-экспериментов.

ЗАКЛЮЧЕНИЕ

Нами проведен биоинформатический анализ последовательностей GCC-боксов, которые являются сайтами связывания ТФ семейства ERF. Наличие GCC-бокса в промоторе гена исследователи ассоциируют с их этилен-чувствительностью (Fujimoto *et al.*, 2000; Stepanova *et al.*, 2007). Первые GCC-боксы были найдены в генах, значимо изменяющих свою экспрессию в ответ на этилен и связанных с ответом на стрессовые воздействия (Sessa *et al.,* 1995; Sato *et al.*, 1996). Однако, так как ранее полногеномный анализ GCC-боксов не проводился, достоверность этой ассоциации была неизвестна. Нами впервые осуществлен анализ распределения GCC-боксов в геноме *A. thaliana*. В результате показано, что GCC-боксы действительно обогащены в проксимальных районах этилен-чувствительных генов и ассоциированы с активирующим эффектом этилена. Установлено, что этилен не только увеличивает, но и уменьшает экспрессию ряда генов (Hess *et al.*, 2011; Cheng *et al.*, 2013; Mase *et al.*, 2013). Вероятно, на характер изменения экспрессии генов влияет локализация GCC-боксов в прямой или обратной цепи относительно старта инициации транскрипции. Также возможно влияние других ТФ в пути передачи этиленового сигнала, например, EIN3/EIL1 (Solano *et al.*, 1998; Alonso *et al.*, 2003), сайты связывания которых изучены недостаточно. Детальное прояснение этого вопроса и роли ориентации сайтов относительно старта инициации транскрипции для проксимального района является целью наших ближайших исследований.

Предсказанные нами GCC-боксы оказались значимо обогащены не только в промоторах, но и в кодирующих частях этилен-чувствительных генов. Это может свидетельствовать либо о том, что GCC-бокс может быть частью дуона (Stergachis *et al.*, 2013), либо в ряде случаев о том, что регулируемые ERF-гены могут экспрессироваться с альтернативных стартов транскрипции, а значит, сайт располагается в 5′-области для некоторых изоформ, кодируемых геном. Однако нами было показано, что предсказанные сайты в экзонах не имеют функционального значения. Изучение этого вопроса требует более детального анализа с привлечением дополнительных экспериментальных данных.

Также нами выявлены некоторые закономерности влияния ориентации GCC-боксов относительно старта инициации транскрипции на функцию гена. Группы генов, имеющих GCC-боксы в обратной ориентации, были значимо ассоциированы с увеличением экспрессии в ответ на этилен. Для GCC-боксов в прямой ориентации нами не было показано значимой

связи с этилен-чувствительностью. Интересно, что гены, содержащие GCC-боксы в определенной ориентации, оказались специфически обогащены другими терминами генной онтологии помимо постэмбрионального развития. Это свидетельствует о том, что GCC-боксы могут быть функциональны для ТФ ERF, которые не участвуют в генной сети передачи этилена, но участвуют в других важных для метаболизма клетки процессах.

БЛАГОДАРНОСТИ

Выражаем благодарность И.В. Мироновой за помощь в составлении обучающей выборки и И.В. Медведевой за экстракцию данных из базы ENSEMBL. Работа поддержана грантом РФФИ-12-04-33112-мол-а-вед и бюджетным проектом ИЦиГ СО РАН VI.61.1.2.

ЛИТЕРАТУРА

Alonso J.M., Stepanova A., Solano R. *et al.* Five Components of the Ethylene-Response Pathway Identified in a Screen for Weak Ethylene-Insensitive Mutants in Arabidopsis // Proc. Natl AS USA. 2003. V. 100. No. 5. P. 2992–2997.

Bleecker A.B., Kende H. Ethylene: A Gaseous Signal Molecule in Plants // Ann. Rev. Cell Developmental Biology. 2000. V. 16. No. 1. P. 1–18.

Cheng M., Liao P., Kuo W., Lin T. The Arabidopsis ETHYLENE RESPONSE FACTOR1 Regulates Abiotic Stress-Responsive Gene Expression by Binding to Different Cis-Acting Elements in Response to Different Stress Signals // Plant Physiology. 2013. V. 162. No. 3. P. 1566–1582.

Choudhury S.R., Roy S. *et al.* Characterization of Differential Ripening Pattern in Association with Ethylene Biosynthesis in the Fruits of Five Naturally Occurring Banana Cultivars and Detection of a GCC-Box-Specific DNA-Binding Protein // Plant Cell Reports. 2008. V. 27. No. 7. P. 1235–1249.

Du Z., Zhou X., Ling Y. *et al.* agriGO: A GO Analysis Toolkit for the Agricultural Community // Nucleic Acids Research. 2010. V. 38. P. 64–70.

Fujimoto S., Ohta M. *et al.* Arabidopsis Ethylene-Responsive Element Binding Factors Act as Transcriptional Activators or Repressors of GCC Box-Mediated Gene Expression // The Plant Cell. 2000. V. 12. No. 3. P. 393–404.

Goda H., Sasaki E., Akiyama K. *et al.* The AtGenExpress Hormone and Chemical Treatment Data Set: Experimental Design, Data Evaluation, Model Data Analysis and Data Access // Plant J.: For Cell Molecular Biology. 2008. V. 55. No. 3. P. 526–542.

Grimmig B., Gonzalez-Perez M. *et al.* Ozone-Induced Gene Expression Occurs via Ethylene-Dependent and Independent Signalling // Plant Molecular Biology. 2003. V. 51. No. 4. P. 599–607.

Hess N., Klode M. *et al.* The Hypoxia Responsive Transcription Factor Genes ERF71/HRE2 and ERF73/HRE1 of Arabidopsis Are Differentially Regulated by Ethylene // Physiologia Plantarum. 2011. V. 143. No. 1. P. 41–49.

Ju C., Chang C. Advances in Ethylene Signalling: Protein Complexes at the Endoplasmic Reticulum Membrane // AoB Plants. 2012. No. 1.

Kulakovskiy I.V., Boeva V.A., Favorov A.V., Makeev V.J. Deep and Wide Digging for Binding Motifs in ChIP-Seq Data // Bioinformatics. 2010. V. 26. No. 20. P. 2622–2623.

Levitsky V.G., Ignatieva E.V., Ananko E.A. *et al.* Effective Transcription Factor Binding Site Prediction Using a Combination of Optimization, a Genetic Algorithm and Discriminant Analysis to Capture Distant Interactions // BMC Bioinformatics. 2007. V. 8. No. 1. P. 481.

Mase K., Ishihama N., Mori H. *et al.* Ethylene-Responsive AP2/ERF Transcription Factor MACD1 Participates in Phytotoxin-Triggered Programmed Cell Death // Molecular Plant-Microbe Interactions. 2013. V. 26. No. 8. P. 868–879.

Mironova V., Omelyanchuck N., Levitsky V. Computational analysis of Auxin Responsive Elements in *Arabidopsis thaliana* Genome // BMC Genomics review. 2014. In print.

Nakano T., Suzuki K., Fujimura T., Shinshi H. Genome-Wide Analysis of the ERF Gene Family in Arabidopsis and Rice // Plant Physiology. 2006. V. 140. No. 2. P. 411–432.

Ohme-Takagi M., Shinshi H. Ethylene-Inducible DNA Binding Proteins That Interact with an Ethylene-Responsive Element // Plant Cell. 1995. V. 7. No. 2. P. 173–182.

Olmedo G., Guo H., Gregory B. *et al.* ETHYLENE-INSENSITIVE5 Encodes a 5′→3′ Exoribonuclease Required for Regulation of the EIN3-Targeting F-Box Proteins EBF1/2 // Proc. Natl AS USA. 2006. V. 103. No. 36. P. 13286–13293.

Osnato M., Stile M.R., Wang Y. *et al.* Cross Talk between the KNOX and Ethylene Pathways Is Mediated by Intron-Binding Transcription Factors in Barley // Plant Physiology. 2010. V. 154. No. 4. P. 1616–1632.

Qiao H., Shen Z., Huang S. *et al.* Processing and Subcellular Trafficking of ER-Tethered EIN2 Control Response to Ethylene Gas // Science. 2012. V. 338. No. 6105. P. 390–393.

Samac D. A., Hironaka C.M., Yallaly P.E., Shah D.M. Isolation and Characterization of the Genes Encoding Basic and Acidic Chitinase in Arabidopsis Thaliana // Plant Physiology. 1990. V. 93. No. 3. P. 907–914.

Sato F., Kitajima S. *et al.* Ethylene-Induced Gene Expression of Osmotin-like Protein, a Neutral Isoform of Tobacco PR-5, Is Mediated by the AGCCGCC Cis-Sequence // Plant Cell Physiology. 1996. V. 37. No. 3. P. 249–255.

Sears M., Zhang H., Rushton P. *et al.* NtERF32: A Non-NIC2 Locus AP2/ERF Transcription Factor Required in Jasmonate-Inducible Nicotine Biosynthesis in Tobacco // Plant Molecular Biology. 2014. V. 84. No. 1-2. P. 49–66.

Sessa G., Meller Y., Fluhr R. A GCC Element and a G-Box Motif Participate in Ethylene-Induced Expression of the PRB-1b Gene // Plant Molecular Biology. 1995. V. 28. No. 1. P. 145–153.

Shakeel S.N., Wang X., Binder B.M., Schaller G.E. Mechanisms of Signal Transduction by Ethylene: Overlapping

and Non-Overlapping Signalling Roles in a Receptor Family // AoB Plants. 2013. V. 5. No. 1.

Shinshi H., Usami S., Ohme-Takagi M. Identific ation of an Ethylene-Responsive Region in the Promoter of a Tobacco Class I Chitinase Gene // Plant Molecular Biology. 1995. V. 27. No. 5. P. 923–932.

Solano R., Stepanova A., Chao Q., Ecker J.R. Nuclear Events in Ethylene Signaling: A Transcriptional Cascade Mediated by ETHYLENE-INSENSITIVE3 and ETHYLENE-RESPONSE-FACTOR1 // Genes Development. 1998. V. 12. No. 23. P. 3703–3714.

Son G.H., Wan J., Kim H. *et al.* Ethylene-Responsive Element-Binding Factor 5, ERF5, Is Involved in Chitin-Induced Innate Immunity Response // Molecular Plant-Microbe Interactions. 2012. V. 25. No. 1. P. 48–60.

Stepanova A.N., Ecker J.R. Ethylene Signaling: From Mutants to Molecules // Current Opinion Plant Biology. 2000. V. 3. No. 5. P. 353–360.

Stepanova A.N., Yun J., Likhacheva A.V., Alonso J.M. Multilevel Interactions between Ethylene and Auxin in Arabidopsis Roots // Plant Cell. 2007. V. 19. No. 7. P. 2169–2185.

Stergachis A., Haugen E. *et al.* Exonic Transcription Factor Binding Directs Codon Choice and Affects Protein Evolution // Science. 2013. V. 342. No. 6164. P. 1367–1372.

Tournier B., Sanchez-Ballesta M. *et al.* New Members of the Tomato ERF Family Show Specific Expression Pattern and Diverse DNA-Binding Capacity to the GCC Box Element // FEBS Letters. 2003. V. 550. No. 1-3. P. 149–154.

Wang P., Du Y., Zhao X. *et al.* The MPK6-ERF6-ROS-Responsive Cis-Acting Element7/GCC Box Complex Modulates Oxidative Gene Transcription and the Oxidative Response in Arabidopsis // Plant Physiology. 2013. V. 161. No. 3. P. 1392–1408.

Zarei A., Körbes A.P., Younessi P. *et al.* Two GCC Boxes and AP2/ERF-Domain Transcription Factor ORA59 in Jasmonate/ethylene-Mediated Activation of the PDF1.2 Promoter in Arabidopsis // Plant Molecular Biology. 2011. V. 75. No. 4-5. P. 321–231.

Zhang Z., Zhang H., Quan R., Wang X.C., Huang R. Transcriptional Regulation of the Ethylene Response Factor LeERF2 in the Expression of Ethylene Biosynthesis Genes Controls Ethylene Production in Tomato and Tobacco // Plant Physiology. 2009. V. 150. No. 1. P. 365–377.

COMPUTATIONAL ANALYSIS AND FUNCTIONAL ANNOTATION OF AP2/ERF TRANSCRIPTION FACTOR BINDING SITES IN *ARABIDOPSIS THALIANA* L. GENOME

O.A. Chernykh[1], V.G. Levitsky[1, 2], N.A. Omelyanchuk[1], V.V. Mironova[1, 2]

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

Summary

The plant hormone ethylene regulates both developmental processes and various stress responses in plants. Ethylene perception in plants is followed by activation of some transcription factors from the large family of APETALA2/ETHYLENE response factors (ERFs). ERF TF binding sites contain a specific GCCGCC motif, called GCC-box. In this study, we applied TF binding site recognition tools oPWM and SiteGA for sequence analysis of experimentally proven GCC-boxes. We carried out GCC box recognition and tested its distribution in the *Arabidopsis thaliana* L. genome. Functional annotation and microarray data analysis of the genes possessing predicted GCC-boxes elucidated their role in ethylene response.

Key words: ethylene, transcription factor, binding site, *Arabidopsis thaliana*.

УДК 519.95

ВЫБОР ИНФОРМАТИВНЫХ ПРИЗНАКОВ ДЛЯ ДИАГНОСТИКИ ЗАБОЛЕВАНИЙ ПО ГЕНЕТИЧЕСКИМ ДАННЫМ

© 2014 г. **Н.Г. Загоруйко, О.А. Кутненко, И.А. Борисова, В.В. Дюбанов, Д.А. Леванов, О.А. Зырянов**

Федеральное государственное бюджетное учреждение науки Институт математики им. С.Л. Соболева Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: zag@mail.ru

Поступила в редакцию 28 сентября 2014 г. Принята к публикации 24 октября 2014 г.

В связи с появлением и активным использованием ДНК-микрочипов при решении различных задач в медицине, биоинформатике и молекулярной биологии усилилась потребность в алгоритмах Data Mining, способных обрабатывать задачи, в которых число анализируемых объектов на порядки меньше числа описывающих признаков. Однако большинство из существующих ныне алгоритмов изначально не предназначено для решения подобных сложных, плохо обусловленных задач. Нами разработан подход, основанный на идее конкурентного сходства, который позволяет разрабатывать алгоритмы, лучше приспособленные для этих целей. Одним из таких алгоритмов является предложенный нами алгоритм FRiS-GRAD, который одновременно решает задачу распознавания и задачу выбора системы информативных признаков. Эффективность его работы проиллюстрирована на различных медицинских задачах в сравнении с наиболее популярными алгоритмами выбора информативных признаков и распознавания.

Ключевые слова: экспрессия генов, функция конкурентного сходства, выбор информативных признаков, распознавание.

ВВЕДЕНИЕ

В настоящее время увеличивается количество публикаций с данными об экспрессии генов у пациентов − носителей различных заболеваний. Данные имеют вид таблиц из M объектов (пациентов) и N признаков (экспрессии генов). Один из основных видов анализа таких данных состоит в выборе подмножества признаков, по которым можно было бы делать диагностику заболеваний. Эту задачу можно было бы решить, оценивая информативность каждого из N генов в отдельности и выбирая заданное количество $n < N$ генов с наибольшей индивидуальной информативностью.

Но между генами имеются зависимости, учет которых заставляет выбирать такие гены, которые дополняли бы друг друга и образовывали подмножества генов с максимальной коллективной информативностью. Точное решение этой задачи методом полного перебора всех сочетаний из N по n генов в общем случае получить нельзя. Предлагаются различные эвристические методы выбора информативных характеристик. В работе Jeffery с соавт. (2006) описаны десять наиболее популярных методов выбора признаков. Их типичным недостатком является предположение о том, что признаки независимы. Кроме того, без всякого обоснования выбрано заданное количество n генов.

В данной работе описан алгоритм FRiS-GRAD для выбора информативных признаков, который учитывает взаимные зависимости между признаками и автоматически определяет их оптимальное количество n. Алгоритм основан на использовании тернарной меры сходства между объектами в виде FRiS-функции (Zagoruiko *et al.*, 2008). Это позволяет сделать прозрачным способ построения решающих правил, оценить количественно компактность образов и информативность признаков. Приведены примеры решения реальных генетических задач.

Что такое FRiS-функция?

Главным элементом всех методов анализа данных является мера сходства между объектами или признаками. Считать, что для оценки сходства между объектами *a* и *b* достаточно знать расстояние *r*(*a, b*) между ними, неправильно. Можно ли при *r*(*a, b*) = 5 считать, что объект *a* похож на объект *b* настолько, чтобы их можно было включить в один класс? А Москва от Санкт-Петербурга далеко или близко? Ответить на такие вопросы можно, только зная ответ на вопрос «По сравнению с чем?»

Следовательно, надо знать не только расстояние *r*(*a, b*), но и расстояние *r*(*a, c*) до объекта *c*, который является ближайшим к *a* конкурентом объекту *b*. Сходство объекта *a* с объектом *b* в конкуренции с объектом *c* оценивается по формуле

$$F(a,b \mid c) = \frac{r(a,c) - r(a,b)}{r(a,c) + r(a,b)}. \quad (1)$$

Если объекты *a* и *b* совпадают, то их сходство равно 1. Если расстояния от *a* до *b* и *c* одинаковы, то сходство равно 0. Если же *a* совпадает с *c*, то сходство *a* с *b* равно –1.

Как построены решающие правила?

Используется решающее правило прецедентного типа. Из M_i объектов a_i обучающей выборки каждого *i*-го образа выбирают типичные объекты («столпы»). Столпом назначается такой объект a_i, сходство с которым всех остальных объектов a_j данного образа в конкуренции с ближайшими объектами b_j чужого образа максимально (Zagoruiko *et al.*, 2008). Вокруг каждого выбранного столпа формируется кластер. В него входят объекты, сходство которых *F* со столпом выше порога, например, $F > 0$. Если какие-то объекты не вошли в кластеры (оказались незащищенными), то среди них выбирается следующий столп. Им становится объект любого из *K* образов, сходство с которым остальных незащищенных объектов этого образа в конкуренции с любым ближайшим объектом чужого образа максимально. Такая процедура последовательного увеличения столпов продолжается, пока все *M* объектов не окажутся включенными в кластеры.

В итоге формируется список из *k* столпов с указанием количества объектов, которые входят в их кластеры. Решающее правило для распознавания принадлежности контрольного объекта *z* к одному из *K* образов состоит в следующем. Вычисляются расстояния от *z* до всех *k* столпов. Выбираются два самых близких столпа, принадлежащие разным образам. Объект *z* считается принадлежащим тому образу, на столп которого он похож больше всего. Величина FRiS-функции показывает надежность принятого решения.

Как оценить компактность?

Компактность *i*-го кластера C_i равна сумме сходств всех M_i объектов a_j кластера со своим столпом s_i в конкуренции с ближайшими столпами s_v других образов (Загоруйко и др., 2010):

$$C_i = \sum_{j=1}^{M_i} F(a_j, s_i \mid s_v). \quad (2)$$

Если для описания обучающей выборки *K* образов потребовалось использовать *k* столпов, то компактность *C* описания *M* объектов выборки равна

$$C = \frac{1}{M} \sum_{i=1}^{M} C_i. \quad (3)$$

Алгоритм выбора столпов

1.	Для всех a_i, $a_j \in M_i$
1.1.	вычислить $S_i = \sum F(a_i, a_j \mid b_j)$.
2.	Объект $a_i = \text{argmax}\{S_i/i = 1, ..., M_i\}$ назначается столпом s_i.
3.	Повторить пункты 1 и 2 *K* раз.
4.	Сформировать *K* кластеров.
5.	Если все объекты входят в кластеры, то конец.
6.	Если вне кластеров есть $M' < M$ объектов, для них повторить пункты 1–5.

Уклонение от переобучения

С ростом числа столпов сумма компактностей C_i кластеров монотонно увеличивается. Но мощность M_i очередных кластеров обычно меньше предыдущих. Наступает такой момент, когда появляются кластеры, состоящие из одного или нескольких объектов. Наличие таких кластеров свидетельствует о наступлении стадии переобучения. Для обнаружения этого момента перехода от обучения к переобучению используется функция Q качества описания K образов, которая тем больше, чем больше компактность C и чем меньше количество столпов k:

$$Q = C\,\frac{K}{k} \qquad (4)$$

Наличие штрафа за превышение количества столпов k над количеством образов K приводит к тому, что функция Q сначала растет, затем начинает снижаться. Точка перегиба функции $Q = f(k)$ указывает на момент, когда процесс наращивания числа столпов нужно остановить. Объекты, которые к этому моменту не вошли ни в один кластер, не отражают основные закономерности распределения образов и из дальнейшего использования исключаются (цензурируются). Эксперименты на большом числе модельных задач показали, что цензурируются обычно 12–15 % обучающей выборки. Ошибка распознавания контрольной выборки уменьшается в результате цензурирования в 1,5–2,0 раза (Загоруйко, 2013).

Как выбрать признаки?

Исходные данные часто содержат признаки, которые не несут полезную информацию для решения конкретной задачи. Нужно выбрать такое подмножество признаков (в нашем случае – генов), которые необходимы и достаточны для диагностики заданного заболевания.

Известны «жадные» алгоритмы выбора признаков Addition (Ad), когда на каждом шаге к имеющимся признакам добавляется самый полезный, и Deletion (Del), когда из имеющихся признаков исключается самый бесполезный признак. Оба этих алгоритма локально оптимальны. Чтобы уклониться от попадания в локальный оптимум, используют комбинированную процедуру AdDel, в которой чередуются этапы наращивания подсистемы на $n1$ признаков с процедурой сокращения подсистемы на $n2$ признаков, $n2 < n1$.

В процессе увеличения размерности подсистемы такой процедурой «два шага вперед – один назад» информативность подсистемы растет, затем рост останавливается и начинается уменьшение информативности. Точка перегиба функции информативности указывает на оптимальное количество n признаков.

Можно добавлять и исключать не отдельные признаки, а гранулы, состоящие из двух или трех признаков. Самые информативные пары и тройки признаков можно находить методом полного перебора.

На этом основан алгоритм выбора информативных признаков FRiS-GRAD (гранулированный AdDel), который мы использовали при решении разных задач, в том числе задач с генетическими данными (Загоруйко, 2013).

Решение задачи диагностики лейкемии

Особенность генетических задач заключается в том, что количество признаков (генов) велико: тысячи, десятки тысяч. Это на два – три

Алгоритм выбора признаков

1. Для всех $j = 1, \ldots, N$ признаков вычислить компактность C_j.
2. Признак $x_j = \text{argmax} C_j\{C_j/j = 1, ..., N\}$ внести в подсистему.
3. Повторить пункты 1 и 2 $n1$ раз.
4. Оценить компактность C'' подсистемы.
5. Признак x_j', без которого получается $C''_{\max}$, исключить из подсистемы.
6. Повторить пункты 4 и 5 $n2$ раза, $n2 < n1$.
7. Для признаков, не входящих в подсистему, повторить пункты 2–6.
8. Если на i-м и $(i + 1)$-м шагах $C''_{i+1} < C''_i$, то конец.

порядка больше количества объектов (пациентов). Одна из задач состояла в выборе подмножества генов, по экспрессии которых можно было бы отличать друг от друга пациентов с двумя типами лейкемии – ALL и AML (Guyon *et al.*, 2002). Обучающая выборка содержала 38 объектов, тестовая – 34 объекта. Исходное количество признаков (генов) $N = 7\,129$.

Результаты решения этой задачи, описанные в работе (Guyon *et al.*, 2002), таковы. Информативное подмножество признаков выбиралось методом RFE (разновидностью алгоритма Deletion), решающие правила основаны на методе опорных векторов SVM (Vapnik, 1998). Были найдены наилучшие подсистемы, размерность которых кратна степени числа 2: 4 096, 2 048, …, 4, 2 и 1. По двум лучшим признакам, которые можно выбрать по результатам обучения, правильно распознано 30 объектов из 34, по четырем лучшим признакам – 31, по 128 признакам – 33 объекта (табл. 1).

Нами на тех же данных получены следующие результаты. Информативное подмножество признаков выбрано с помощью алгоритма FRiS-GRAD. Информативность признаков оценена по критерию FRiS-компактности. Из 7 129 признаков выбрано 18 признаков, из которых программа FRiS-Stolp построила 30 вариантов решающих правил. В состав каждого правила входит с разными весами от трех до шести признаков. Первые 10 правил показаны в табл. 2.

Первые 27 правил из 30 дают результат 34 из 34. Различия между приведенными результатами могут зависеть как от метода выбора признаков, так и от типа решающих правил. Для сравнения решающих правил SVM и FRiS был проведен такой эксперимент.

В подпространстве двух признаков (генов 803 и 4846), выбранных методом RFE, по правилу SVM получено 30 правильных ответов, а FRiS-методом – 33.

По лучшему одному гену (4846), выбранному методом RFE, результат SVM равен 27, а результат FRiS равен 30. А по лучшему гену (2461), выбранному алгоритмом GRAD, метод FRiS дает 32 правильных ответа (табл. 3).

Отсюда можно сделать вывод, что как метод выбора признаков, так и решающие правила, основанные на FRiS-функции, обладают высокими конкурентными качествами.

Таблица 1

Результаты обучения и контроля при выборе признаков методом RFE и решающем правиле SVM (обучающая выборка 38 объектов, тестовая выборка 34 объекта)

Число признаков	Критерий выбора	Распознано правильно
7 129	0,85	29
4 096	0,71	24
2 048	0,85	29
1 024	0,94	32
512	0,88	30
256	0,94	32
128	0,97	33
64	0,94	32
32	0,97	33
16	1,00	34
8	1,00	34
4	0,91	31
2	0,88	30
1	0,79	27

Таблица 2

Выбор признаков методом FRiS-GRAD, решающие правила FRiS-Stolp

FRiS	Решающие правила	P
0,72656	537/1,1833/1,2641/2,4049/2	34
0,71373	1454/1,2641/1,4049/1	34
0,71208	2641/1,3264/1,4049/1	34
0,71077	435/1,2641/2,4049/2,6800/1	34
0,70993	2266/1,2641/2,4049/2	34
0,70973	2266/1,2641/2,2724/1,4049/2	34
0,70711	2266/1,2641/2,3264/1,4049/2	34
0,70574	2641/2,3264/1,4049/2,4446/1	34
0,70532	435/1,2641/2,2895/1,4049/2	34
0,70243	2641/2,2724/1,3862/1,4049/2	34

Таблица 3

Результаты распознавания двумя решающими правилами SVM и FRiS-Stolp по двум лучшим признакам, выбранным методом RFE

Метод	Best features	SVM	FRiS-Stolp
RFE	803,4846	30 (88 %)	33 (97 %)
	4846	27 (79 %)	30 (88 %)

Таблица 4

Результаты сравнения FRiS-методов с лучшими результатами, полученными сорока наиболее известными методами

Задача	ALL1	Leuk	Prost	DLBCL	Colon	ALL4	Myel	ALL3	ALL2
Признаки	12 625	7 129	12 625	7 129	2 000	12 625	12 625	12 625	12 625
Объекты $m1/m2$	95/33	47/25	50/53	58/19	22/40	26/67	36/137	65/35	24/91
Рекорды из 40	100,00	95,85	90,19	94,30	88,60	82,06	82,90	59,58	78,23
FRiS	100,00	100,00	96,3	96,9	95,6	88,2	84,8	87,6	85,6
Рейтинг FRiS	1	1	1	1	1	1	1	1	1

Сравнение с наиболее известными методами выбора признаков

В работе Jeffery с соавт. (2006) проведено сравнение десяти наиболее известных методов выбора признаков на основе результатов решения девяти задач диагностики по генетическим данным. Для каждой выбранной системы признаков строились решающие правила четырех наиболее известных типов. Для каждой из девяти задач было получено сорок различных решений.

Мы выбрали лучшие из них (рекорды) и сравнили их с результатами, полученными комбинацией алгоритма выбора признаков FRiS-GRAD с алгоритмом построения решающего правила FRiS-Stolp (табл. 4). В таблице показаны имена задач, размерность признакового пространства N, количества объектов первого ($m1$) и второго ($m2$) образов и две строки результатов. В последней строке показано место, занятое результатами решения всех девяти задач FRiS-методами.

Для каждой задачи по результату, полученному каждым методом, можно указать его рейтинг: лучший результат занимает первое место, худший – десятое. Если просуммировать места, занятые методом на всех задачах, то можно определить его общий рейтинг. Результаты таких подсчетов представлены в табл. 5, в последней строке которой показана сумма рейтинговых мест, занятых FRiS-методом. Такой же анализ был проведен и по четырем использованным решающим правилам. Его результаты показаны в табл. 6, в которой, как и в табл. 5, чем меньше сумма рейтинговых мест, тем лучше.

Таблица 5

Сумма рейтинговых мест, занятых методами выбора признаков

Метод выбора признаков	Рейтинг
Fold change	47
Between group analysis	43
Analysis of variance (ANOVA)	43
Significance analysis of microarrays	42
Rank products	42
Welch t-statistic	39
Template matching	38
Area under the ROC curve	37
MaxT	37
Empirical Bayes t-statistic	32
FRiS-GRAD	9

Таблица 6

Сумма рейтинговых мест, полученных решающими правилами

Решающее правило	Рейтинг
Between group analysis (BGA)	35
K-nearest neighbours (kNN)	32
Naive bayes classification (NBC)	25
Support vector machines (SVM)	19
FRiS-Stolp	9

ЗАКЛЮЧЕНИЕ

По приведенным результатам можно сделать вывод о высокой эффективности сочетания алгоритмов выбора признаков FRiS-GRAD и построения решающих правил FRiS-Stolp для решения сложных задач диагностики по генетическим данным.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке РФФИ по грантам 11-01-00156 и 14-01-00039 и интеграционным проектам СО РАН № 54 и 87.

ЛИТЕРАТУРА

Загоруйко Н.Г. Когнитивный анализ данных. Новосибирск: Академическое издательство ГЕО, 2013. 186 с.

Загоруйко Н.Г., Борисова И.А., Дюбанов В.В., Кутненко О.А. Количественная мера компактности и сходства в конкурентном пространстве // Сибирский журнал индустриальной математики. Новосибирск, 2010. Т. 13. № 1 (41). С. 59–71.

Guyon I., Weston J., Barnhill S., Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines // Machine Learnin. 2002. V. 46 (1–3). P. 389–422.

Jeffery I., Higgins D., Culhane A. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data // BMC Bioinformatics. 2006. V. 7. P. 359.

Vapnik V.N. Statistical Learning Theory. Wiley-Interscience, 1998.

Zagoruiko N.G., Borisova I.A., Dyubanov V.V., Kutnenko O.A. Methods of Recognition Based on the Function of Rival Similarity // Pattern Recognition Image Analisys. 2008. V. 18. No. 1. P. 1–6.

FEATURE SELECTION IN THE TASK OF MEDICAL DIAGNOSTICS ON MICROARRAY DATA

N. G. Zagoruiko, O. A. Kutnenko, I. A. Borisova, V. V. Dyubanov, D.A. Levanov, O.A. Zyranov

Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia,
e-mail: zag@math.nsc.ru

Summary

In tasks of modern biology, the numbers of attributes often exceed the numbers of objects by orders of magnitude. For the solution of such tasks, a Data Mining method based on using a new measure of similarity between objects in the form of the Function of Rival Similarity (FRiS) is offered. On this basis, methods of quantitative estimation of compactness of patterns, construction of decision rules, and feature selection are developed. All these techniques are implemented in the FRiS-GRAD algorithm. The high efficiency of the algorithm is illustrated by results of solving the task of disease recognition on a microarray dataset.

Key words: gene expression, function of rival similarity, feature selection, pattern recognition.

УДК 577.217.53:577.322.52:004.738

ELOE – ВЕБ-ПРИЛОЖЕНИЕ ДЛЯ ОЦЕНКИ ЭФФЕКТИВНОСТИ ЭЛОНГАЦИИ ТРАНСЛЯЦИИ ГЕНОВ

© 2014 г. **В.С. Соколов[1], Б.С. Зураев[1, 2], С.А. Лашин[1, 2], Ю.Г. Матушкин[1, 2]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия,
e-mail: sokovlad1@bionet.nsc.ru;
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

Поступила в редакцию 18 сентября 2014 г. Принята к публикации 8 октября 2014 г.

Многие современные исследования изучают важную характеристику гена – эффективность его экспрессии. Как известно, она определяется на уровнях транскрипции, трансляции, посттрансляционной модификации и др. В работе представлена программа EloE (Elongation Efficiency), сортирующая гены организма в порядке уменьшения их предполагаемой скорости элонгации трансляции на основе анализа их нуклеотидных последовательностей. Полученные теоретические данные достоверно коррелируют с доступными экспериментальными данными по экспрессии генов различных организмов, например *S. cerevisiae* и *H. pylori*. Также программа выявляет предпочтительные кодоны в геноме организма и строит распределение стабильности потенциальных вторичных структур в районах 5′- и 3′-концов мРНК. Программа может быть использована для предварительной оценки уровня экспрессии генов исследуемого организма, экспериментальные данные для которого еще не доступны. Результаты работы EloE могут быть переданы в сторонние программные инструменты, которые моделируют искусственные генетические конструкции для генно-инженерных экспериментов.

Ключевые слова: индекс эффективности элонгации, веб-приложение, эффективность трансляции, вторичные структуры.

ВВЕДЕНИЕ

Эффективность экспрессии генов организма определяется на многих уровнях, важнейшими из которых являются транскрипция, трансляция, посттрансляционная модификация. Изучение факторов, регулирующих трансляцию, – актуальная задача современной биологии. Результаты ее решения могут быть использованы, например, при создании генно-инженерных конструкций и принесут большую пользу в таких областях, как медицина и сельское хозяйство.

В работах (Thanaraj, Argos, 1996; Lopinski *et al.*, 2000; Takyar *et al.*, 2005; Eck, Stephan, 2008) показано, что у большинства прокариотических, а также многих эукариотических видов уровень экспрессии генов зависит от их кодонного состава и от наличия и стабильности вторичных структур в мРНК. Эти факторы влияют на скорость движения рибосомного комплекса по мРНК в процессе трансляции и тем самым – на скорость синтеза белка. На данный момент разработано множество различных индексов для выявления предпочтительных кодонов (Ikemura, 1981; Bennetzen, Hall, 1982; McLachlan *et al.*, 1984). Индекс адаптации кодонов (CAI) – один из первых в этом ряду (Sharp, Li, 1986). Также существует большое количество программ для оценки насыщенности мРНК вторичными структурами (Zuker *et al.*, 1999; Hofacker, 2003; Zuker, 2003).

В статье Н.В. Владимирова с соавт. (Vladimirov *et al.*, 2007) описано пять типов эволюционной оптимизации первичной структуры генов, основанных на факторах, влияющих на эффективность экспрессии генов на уровне трансляции. К этим факторам относятся: час-

тоты кодонов в гене, наличие и распределение вторичных структур в мРНК и стабильность этих структур. Разные комбинации названных параметров формируют пять типов эволюционной оптимизации, которые учитывают:

1) только кодонный состав мРНК;
2) только количество локальных инвертированных повторов в мРНК;
3) только энергетическую стабильность потенциальных шпилек в мРНК;
4) кодонный состав и количество локальных инвертированных повторов в мРНК;
5) кодонный состав и энергетическую стабильность потенциальных шпилек в мРНК.

Индекс эффективности элонгации (ИЭЭ, EEI – Elongation Efficiency Index), предложенный в статье В.А. Лихошвая и Ю.Г. Матушкина (2000), позволяет классифицировать большинство прокариот и некоторых эукариот (в основном одноклеточных, например дрожжей) по этим пяти типам оптимизации первичной структуры генов. Данный индекс оценивает предполагаемую эффективность прохождения стадии элонгации трансляции для каждого гена организма. Поскольку элонгация является одной из наиболее энерго- и времязатратных стадий трансляции, на основе индекса ИЭЭ можно сделать предположения об эффективности трансляции в целом (Там же).

Наша работа посвящена исследованию связи контекстных характеристик генов с их эффективностью трансляции. В представленной программе для исследования эффективности элонгации трансляции был выбран именно ИЭЭ (EEI) (Лихошвай, Матушкин, 2000; Likhoshvai, Matushkin, 2002), поскольку он позволяет учитывать в расчетах как кодонный состав гена, так и его насыщенность локальными инвертированными повторами (потенциальными шпильками в составе вторичной структуры мРНК).

Данный индекс позволяет ранжировать по эффективности элонгации трансляции гены даже тех организмов, для которых другие индексы, основанные на учете частот использования кодонов, не работают (Лихошвай, Матушкин, 2000). Также ранее было показано, что ИЭЭ коррелирует с другими параметрами, оценивающими эффективность экспрессии генов, в частности с плотностью нуклеосомной упаковки в промоторном районе генов дрожжей (Vladimirov *et al.*, 2007; Матушкин и др., 2013). Для массового анализа геномов различных организмов была необходимость в создании общего программного интерфейса с возможностью изменения параметров расчетов, доступного в сети интернет и позволяющего производить анализ сразу нескольких (до нескольких тысяч) геномов за один запуск. Такая задача была решена в форме специального веб-приложения.

РЕЗУЛЬТАТЫ

Для классификации видов по пяти типам эволюционной оптимизации первичной структуры их генов и оценки их эффективности элонгации трансляции создано web-приложение EloE, доступное по адресу http://www-bionet.sscc.ru:7780/EloE. Вид интерфейса представлен на рис. 1.

Исходные данные составляют файлы с аннотированной нуклеотидной последовательностью полного генома в формате gbk (данные могут быть взяты в базе GenBank ftp://ftp.ncbi.nih.gov/genbank/genomes). Для произведения расчетов требуется создание zip-архива с аннотированными геномами (gbk) исследуемых организмов. Геном каждого организма должен располагаться в архиве в отдельной папке. Архив загружается в программу с помощью кнопки Upload. Все результаты, в том числе список генов организма, отсортированных по индексу ИЭЭ, сохраняются в файлы и могут быть загружены после окончания расчетов (кнопка Download results).

Основные файлы с результатами располагаются для каждого организма в отдельной папке Organism_name:

1) organism_name_all.txt – файл со всеми пятью типами индекса ИЭЭ, рассчитанными для всех генов организма, учитываемых в расчетах;
2) organism_name_eeiN.txt (N = {1, 2, 3, 4, 5}) – файл только с тем типом индекса ИЭЭ, который работает в данном организме;
3) organism_name_genes_and_flanks.txt – файл с подробной информацией по каждому гену и его нуклеотидной последовательностью с флангами;
4) organism_name_number_eei.txt – файл с указанием номера гена, его позиции в опероне (только для прокариот) и значения ИЭЭ;

5) organism_name_ribpos.txt – файл с расположением генов рибосомных белков в списках генов организма, отсортированных по каждому из пяти типов ИЭЭ в порядке увеличения (рис. 2 и 3).

Общие результаты по всем анализируемым геномам собраны в одном файле organism_index.txt. Данные во всех файлах разделены знаком табуляции.

Интерфейс программы позволяет изменять параметры расчетов: размеры фланкирующих районов генов, длины инвертированных повторов в мРНК и расстояние между мономерами этих повторов. В начале/конце первого/последнего кодирующего экзона обычно расположены специфические кодоны, характерные именно для сайтов начала/конца трансляции. Поэтому их учет может негативно повлиять на расчеты ИЭЭ. В программе можно указать количество кодонов, которые не будут учтены в расчетах, или поставить галочку Use auto calculation of flanks' length. Тогда программа сама определит оптимальное количество неучитываемых кодонов. Также можно заказать дополнительные выходные файлы.

Одной из возможностей программы является генерация файла organism_name_lciij_profile_out.xlsx (при установленной галочке Calculate Local Complementarity Index for individual nucleotides). В нем хранятся значения для построения профилей индексов локальной комплементарности (ИЛК, LCI – Local Complementarity Index), которые строятся в web-приложении. Индекс локальной комплементарности имеет смысл среднего количества локальных совершенных инвертированных повторов определенной длины в мРНК. Такие повторы потенциально могут образовывать шпильки в составе вторичной структуры мРНК и замедлять движение рибосомного комплекса в процессе элонгации трансляции (Lopinski *et al.*, 2000; Takyar *et al.*, 2005). Индекс локальной комплементарности индивидуального нуклеотида показывает среднюю стабильность шпилек, в образовании которых может принимать участие данный нуклеотид. Индекс рассчитывается в районах старт-кодона (± 600 нуклеотидов) и стоп-кодона (± 600 нуклеотидов) трансляции. Для этого вместе с последовательностью гена из файла gbk экстрагируются фланкирующие районы длиной 600 нуклеотидов. Изменение длины экстрагируемых флангов не влияет на расчеты индексов ИЭЭ.

Для каждого организма EloE строит график с позициями генов рибосомных белков для каждого типа ИЭЭ (рис. 2 и 3). Гены на графике отсор-

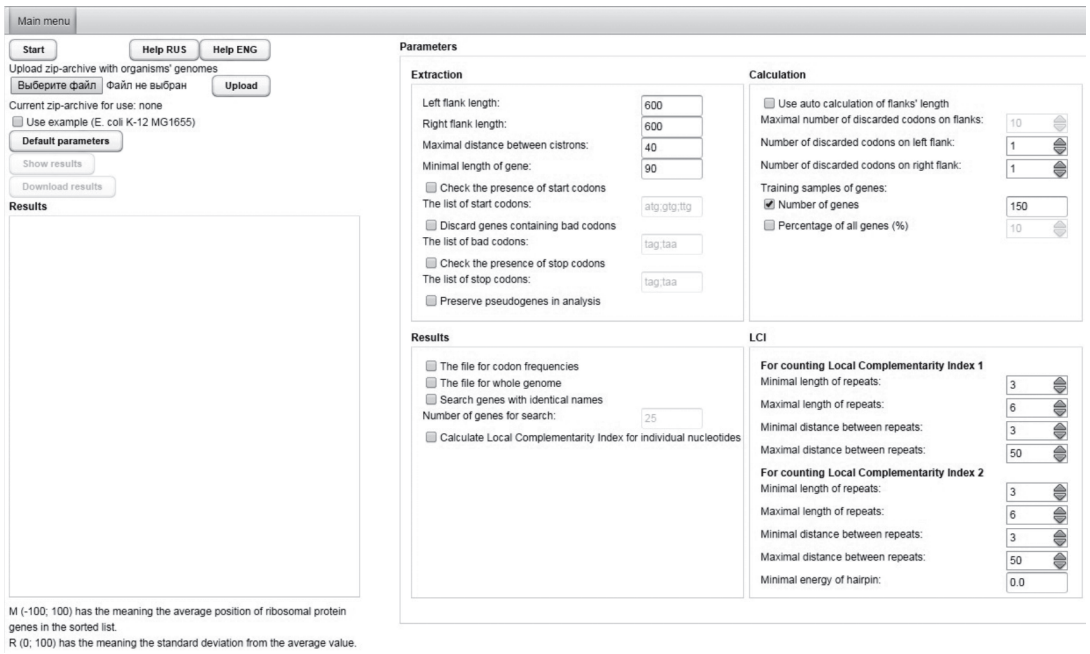


Рис. 1. Интерфейс web-приложения EloE.

тированы в порядке увеличения ИЭЭ. Таким образом, наилучший тип ИЭЭ для организма – это такой тип, для которого гены рибосомных белков расположены правее и плотнее (рис. 2 и 3). Как видно из рис. 2, в *E. coli* лучше всего работает первый тип ИЭЭ, т. е. эффективность элонгации в большей степени зависит от частот кодонов в гене. У *Mycoplasma fermentans* JER (рис. 3) лучше работает второй тип ИЭЭ, т. е. эффективность элонгации в основном определяется количеством инвертированных повторов в гене.

К особенностям программы EloE относятся:

1) возможность обработки более одного генома (до нескольких тысяч) за один запуск;

2) расчет дополнительных параметров и их визуализация, например ИЛК индивидуальных нуклеотидов (LCI) (рис. 4).

Приведенный на рис. 4 профиль ИЛК индивидуальных нуклеотидов отображает среднюю по всем генам организма стабильность потенциальных вторичных структур в районе старт- и стоп-кодона трансляции. Спад профиля в районе старт-кодона (рис. 4, *а*) говорит о потенциально меньшей стабильности шпильки в этом районе мРНК. С другой стороны, пик профиля в районе стоп-кодона (рис. 4, *б*) указывает на повышенную стабильность шпильки.

Выходные файлы содержат такие параметры генов, как: индекс ИЭЭ, индекс ИЛК, GC-состав, длина, позиция в опероне (для прокариот) и др. Дополнительной функцией программы является построение усредненных профилей стабильности вторичных структур в районах 5′- и 3′-концов мРНК всех генов организма. Для прокариот можно выбирать, по каким генам строить профиль: по всем или только по первым, средним, последним или единственным цистронам в операх.

При помощи программы EloE было проведено исследование геномов 62 штаммов *Mycoplasma* (Sokolov *et al.*, 2014). У группы микоплазм (*C.M. haemolamae*, *C.M. haemominutum*,

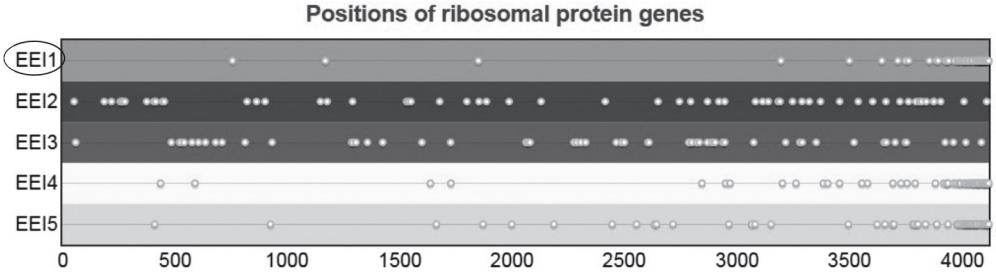


Рис. 2. Распределение генов рибосомных белков (точки) в списке генов *E. coli* K-12 MG1655, расположенных слева направо в порядке увеличения EEI1-5. Наилучший тип индекса (EEI1) выделен кружком – гены рибосомных белков расположены правее и плотнее.

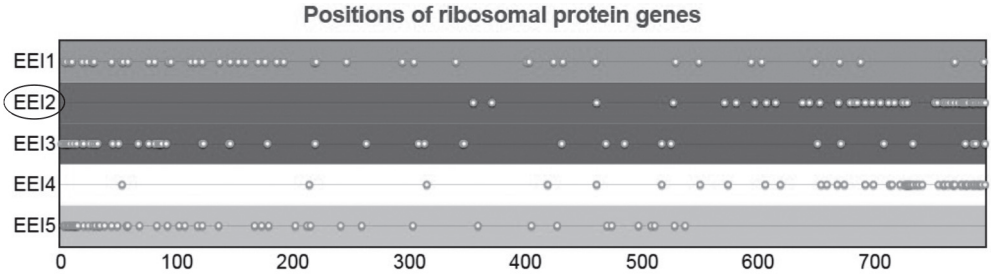


Рис. 3. Распределение генов рибосомных белков (точки) в списке генов *Mycoplasma fermentans* JER, расположенных слева направо в порядке увеличения EEI1-5. Наилучший тип индекса (EEI2) выделен кружком – гены рибосомных белков расположены правее и плотнее.

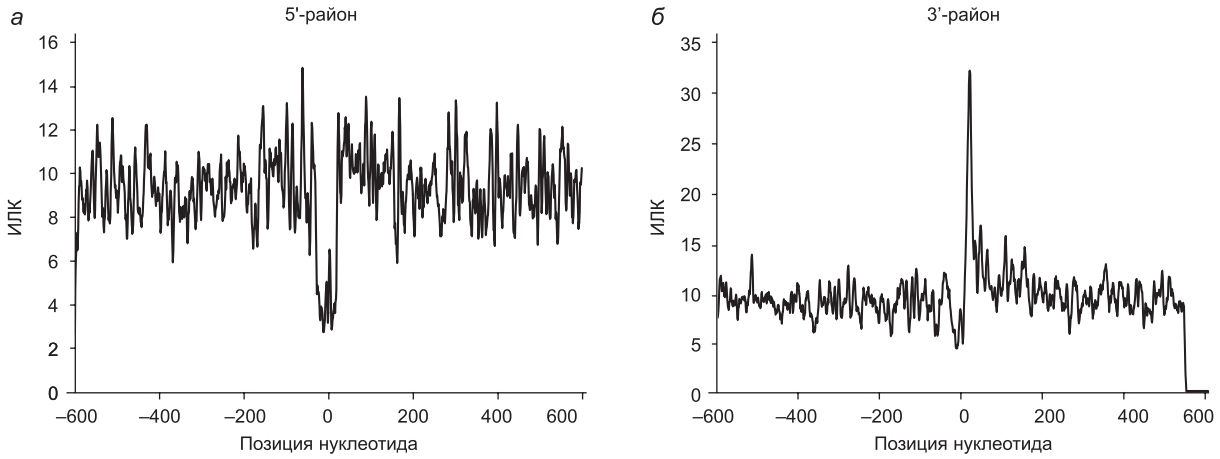


Рис. 4. Визуализация профиля среднего значения ИЛК индивидуальных нуклеотидов по всем генам *Mycoplasma fermentans* JER. Ноль на оси абсцисс на графике *а* – старт-кодон трансляции, на графике *б* – стоп-кодон.

M. haemocanis, *M. haemofelis*, *M. pneumoniae*, *M. suis*) впервые выявлено пониженное содержание в генах локальных инвертированных повторов по сравнению с другими микоплазмами. Также при построении профилей распределения локальных инвертированных повторов в районах старт- и стоп-кодонов трансляции у *M. haemofelis* обнаружен не характерный для остальных микоплазм пик в районе старт-кодона.

ЗАКЛЮЧЕНИЕ

Эффективность экспрессии гена – одна из его главнейших характеристик. Программа EloE позволяет ранжировать гены организма по вычисляемой эффективности одной из важнейших стадий трансляции – элонгации. Учет одновременно кодонного состава и локальных инвертированных повторов в мРНК позволяет программе EloE анализировать более широкий класс организмов, для которых учета только данных по частотам использования кодонов недостаточно. Дополнительные результаты, такие как график распределения стабильности вторичных структур вблизи флангов генов, позволяют более детально исследовать особенности нуклеотидных последовательностей.

Эти данные могут быть полезны во многих областях современных исследований. Особенно это важно при изучении организмов, для которых еще не получены экспериментальные данные по экспрессии их генов. Программа находится в открытом доступе по адресу http://www-bionet.sscc.ru:7780/EloE.

БЛАГОДАРНОСТИ

Работа выполнена при частичной поддержке гранта РФФИ № 13-04-00620, государственного контракта № 1/223-114 и бюджетного проекта № VI.61.1.2.

ЛИТЕРАТУРА

Лихошвай В.А., Матушкин Ю.Г. Предсказание эффективности экспрессии генов по их нуклеотидному составу // Молекулярная биология. 2000. Т. 34. № 3. С. 406–412.

Матушкин Ю.Г. и др. Эффективность элонгации генов дрожжей кореллирует с плотностью нуклеосомной упаковки в 5′-нетранслируемом районе // Математическая биология и биоинформатика. 2013. Т. 8. № 1. С. 248–257.

Bennetzen J.L., Hall B.D. Codon selection in Yeast // J. Biol. Chem. 1982. V. 257. No. 6. P. 3026–3031.

Eck S., Stephan W. Determining the relationship of gene expression and global mRNA stability in *Drosophila melanogaster* and *Escherichia coli* using linear models // Gene. 2008. V. 424. No. 1. P. 102–107.

Hofacker I.L. Vienna RNA secondary structure server // Nucleic acids research. 2003. V. 31. No. 13. P. 3429–3431.

Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli system // J. Molecular Biology. 1981. V. 151. No. 3. P. 389–409.

Likhoshvai V.A., Matushkin Y.G. Differentiation of single-cell organisms according to elongation stages crucial for gene expression efficacy // FEBS letters. 2002. V. 516. No. 1. P. 87–92.

Lopinski J.D., Dinman J.D., Bruenn J.A. Kinetics of ribosomal pausing during programmed–1 translational frameshifting // Mol. Cell. Biol. 2000. V. 20. No. 4. P. 1095–1103.

McLachlan A.D., Staden R., Boswell D.R. A method for measuring the non-random bias of a codon usage table // Nucleic acids research. 1984. V. 12. No. 24. P. 9567–9575.

Sharp P.M., Li W.H. An evolutionary perspective on synonymous codon usage in unicellular organisms // Journal molecular evolution. 1986. V. 24. No. 1-2. P. 28–38.

Sokolov V.S., Likhoshvai V.A., Matushkin Y.G. Gene expression and secondary mRNA structures in different Mycoplasma species // Russian Journal Genetics: Applied Research. 2014. V. 4. No. 3. P. 208–217.

Takyar S., Hickerson R.P., Noller H.F. mRNA helicase activity of the ribosome // Cell. 2005. V. 120. No. 1. P. 49–58.

Thanaraj T.A., Argos P. Ribosome-mediated translational pause and protein domain organization // Protein Science. 1996. V. 5. No. 8. P. 1594–1612.

Vladimirov N.V., Likhoshvai V.A., Matushkin Y.G. Correlation of codon biases and potential secondary structures with mRNA translation efficiency in unicellular organisms // Molecular Biology. 2007. V. 41. No. 5. P. 843–850.

Zuker M. Mfold web server for nucleic acid folding and hybridization prediction // Nucleic acids research. 2003. V. 31. No. 13. P. 3406–3415.

Zuker M., Mathews D.H., Turner D.H. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide // RNA biochemistry and biotechnology. Springer Netherlands, 1999. P. 11–43.

ELOE: A WEB APPLICATION FOR ESTIMATION OF GENE TRANSLATION ELONGATION EFFICIENCY

V.S. Sokolov[1], B.S. Zuraev[1, 2], S.A. Lashin[1, 2], Yu.G. Matushkin[1, 2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: sokovlad1@bionet.nsc.ru;
[2] Novosibirsk National Research State University, Novosibirsk, Russia

Summary

Expression efficiency is one of major characteristics of genes considered in a number of modern investigations. It is known that gene expression efficiency in an organism is regulated at many stages: transcription, translation, posttranslational protein modification, and others. In this study, a special EloE (Elongation Efficiency) web application is described. It sorts genes in an organism in the order of decreasing theoretical rate of the elongation stage of translation deduced from their nucleotide sequences. The predictions done in this way show a significant correlation with available experimental data on gene expression in various organisms, for instance, *S. cerevisiae* and *H. pylori*. In addition, the program identifies preferential codons in a genome and defines the distribution of stability of potential secondary structures in 5′ and 3′ regions of mRNA. EloE can be useful in preliminary estimation of translation elongation efficiency of genes in organisms for which experimental data are not available yet. Some results can be used, for instance, in other programs modeling artificial genetic constructs in gene engineering experiments.

Key words: elongation efficiency index; web application; translation efficiency; secondary structures.

УДК 575.117.2:612.821.33:616.12-008.333.1

СНИЖЕННЫЙ УРОВЕНЬ ЭКСПРЕССИИ ГЕНОВ, КОНТРОЛИРУЮЩИХ ТОНУС СОСУДОВ В ПОЧКАХ КРЫС НИСАГ СО СТРЕСС-ЗАВИСИМОЙ АРТЕРИАЛЬНОЙ ГИПЕРТЕНЗИЕЙ

© 2014 г. **О.Е. Редина[1], Л.О. Климов[1], Н.И. Ершов[1], Т.О. Абрамова[1], Л.Н. Иванова[1, 2], А.Л. Маркель[1, 2]**

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

Поступила в редакцию 16 сентября 2014 г. Принята к публикации 3 октября 2014 г.

Сравнивали транскрипционную активность генов в почках гипертензивных крыс НИСАГ и нормотензивных крыс WAG для определения генов-кандидатов стресс-зависимой артериальной гипертензии, которые достоверно экспрессируются в почках только одной из двух сравниваемых линий. Анализ экспрессии генов проведен на микрочипах Illumina (USA). При анализе транскрипционной активности генов в корковом веществе почек выявлено три гена (*Klk1, Klk1c10* и *Kng1*), имеющих отношение к функционированию калликреин-кининовой системы. Все три гена достоверно экспрессировались в почках нормотензивных крыс, в почках гипертензивных крыс их экспрессия не детектировалась. Снижение уровня экспрессии этих генов и гена *Gucy1a3* у гипертензивных крыс позволяет предполагать ослабление функции калликреин-кининовой системы у крыс НИСАГ, что может нарушать гемоциркуляцию в почечных тельцах и способствовать развитию гипертензии. В мозговом веществе почек функциональная аннотация генов, достоверно экспрессирующихся только у одной из двух сравниваемых линий, показала различия в экспрессии генов регуляции иммунного ответа.

Ключевые слова: крысы НИСАГ, транскрипционная активность генов, микрочипы, артериальная индуцируемая стрессом гипертензия, эмоциональный стресс.

ВВЕДЕНИЕ

Гипертоническая болезнь (ГБ) – одно из самых распространенных заболеваний, характеризуется стойким повышением артериального давления (АД). Причины ГБ, или эссенциальной гипертонии, в отличие от вторичных гипертензий, по-прежнему остаются неизвестными, хотя основные механизмы регуляции АД у человека и млекопитающих хорошо изучены (Guyton, 1990; Cowley, 1992; Lifton *et al.*, 2001). Гипертоническая болезнь характеризуется наличием выраженной генетической компоненты (до 50 % изменчивости) (Havlik *et al.*, 1979; Levy *et al.*, 2000). Анализ генов-кандидатов, как и полногеномные исследования на больших популяциях людей, показали сложную полигенную детерминацию заболевания (Hirschhorn, 2005; Levy *et al.*, 2009; Newton-Cheh *et al.*, 2009; Wang *et al.*, 2009). Развитие ГБ обусловлено действием большого числа генов, каждый из которых вносит лишь небольшой вклад в патогенез заболевания. В итоге ГБ формируется в результате взаимодействия многих средовых и генетических факторов (Mullins *et al.*, 2006), при этом многие гены изменяют уровень своей экспрессии (Lynn *et al.*, 2009).

Для изучения физиологических и молекулярно-генетических механизмов развития артериальной гипертензии в условиях эмоционального

стресса нами получена экспериментальная модель стресс-зависимой артериальной гипертензии – линия крыс с наследуемой индуцируемой стрессом артериальной гипертензией (НИСАГ, или ISIAH) (Markel, 1992). При создании линии НИСАГ селекцию проводили на повышение АД при действии мягкого эмоционального стресса, вызванного получасовой рестрикцией крысы в тесной проволочной клетке. К настоящему времени получено более 30 поколений инбридинга.

Показано, что линия НИСАГ высокоинбредная (Адаричев и др., 1996), характеризуется повышенным АД в покое (175,0 ± 3,5 мм рт. ст. у самцов и 165,0 ± 3,0 у самок) и его значительным повышением в условиях мягкого эмоционального стресса (до 195,0 ± 2,4 мм рт. ст. у самцов и 174,0 ± 3,2 мм рт. ст. у самок). Крысы НИСАГ имеют специфические для ГБ морфологические изменения органов, в том числе изменения морфологии почек, гипертрофию левого желудочка сердца (Markel *et al.*, 1999; Шмерлинг и др., 2001; Филюшина и др., 2013). У крыс НИСАГ изменен уровень катехоламинов в надпочечниках и плазме крови (Маркель и др., 2006; Markel *et al.*, 2007).

Ранее установлено, что у крыс НИСАГ по сравнению с контрольными крысами WAG (Wistar Albino Glaxo) изменена экспрессия некоторых генов, регулирующих гипоталамо-гипофизарно-надпочечниковую систему (Хворостова и др., 2002, 2003; Markel *et al.*, 2007), а также ряда генов, контролирующих состояние симпатической нервной системы, сосудистого тонуса и водно-солевого баланса (Пыльник и др., 2011; Федосеева и др., 2011; Абрамова и др., 2013).

Важную роль при развитии ГБ играет функциональное состояние почек. Они контролируют баланс натрия, объем циркулирующей крови и внеклеточной жидкости, что является ключевым механизмом регуляции АД (Mullins *et al.*, 2006). Хронические заболевания почек представляют серьезный фактор риска для развития заболеваний сердечно-сосудистой системы, включая ГБ, инфаркт миокарда, мозговой инсульт (Korstanje, DiPetrillo, 2004). Для изучения молекулярно-генетических механизмов регуляции физиологических и патофизиологических процессов всё более широко используют сравнительный анализ уровня экспрессии мРНК генов на микроматрицах (Bareyre, Schwab, 2003; Park, Prolla, 2005; Viemann *et al.*, 2005; Mukherjee *et al.*, 2006). Для оценки дифференциальной транскрипционной активности генов почки у гипертензивных крыс НИСАГ и контрольных крыс WAG проведен сравнительный анализ экспрессии генов на микроматрицах Illumina (USA). Функциональный анализ дифференциально экспрессирующихся генов в почках гипертензивных крыс НИСАГ и нормотензивных крыс WAG показал, что эти линии отличаются по экспрессии генов, имеющих отношение к регуляции ответа на стресс, регуляции ионного транспорта, функции иммунной системы (Redina *et al.*, 2014). В ранее проведенном анализе были рассмотрены дифференциально экспрессирующиеся гены с достоверно детектируемой экспрессией у крыс обеих линий ($p < 0,01$) (Redina *et al.*, 2014). Однако можно предполагать, что некоторые гены, определяющие различия в функциональном состоянии почек гипертензивных и контрольных крыс, могут достоверно экспрессироваться только в одной из линий. Цель настоящей работы – выявление генов-кандидатов стресс-зависимой артериальной гипертензии, которые достоверно экспрессируются в почках только одной из двух сравниваемых линий.

МАТЕРИАЛЫ И МЕТОДЫ

Животные

В эксперименте использовали крыс гипертензивной линии НИСАГ и нормотензивной линии WAG. В каждой группе было по три животных-самца в возрасте 6–7 мес. Систолическое АД измеряли непрямым методом на хвосте (tail-cuff method), оно составило 173,67 ± 1,86 мм рт. ст. у крыс НИСАГ и 124,67 ± 2,67 мм рт. ст. у крыс WAG. Крысы содержались в стандартных условиях в виварии ИЦиГ СО РАН, воду и сбалансированный корм получали без ограничения. Эксперименты выполнены в соответствии с международными Правилами проведения работ с использованием экспериментальных животных. Анализ экспрессии генов проведен раздельно в двух отделах почки – в корковом и мозговом веществе. Для получения образцов тканей крыс декапитировали, быстро выделяли

почку, на поперечном срезе разделяли корковое и мозговое вещество. Сразу после выделения образцы тканей (50 мг) гомогенизировали в 1 мл тризола (TRIzol reagent, Invitrogen Life Technologies, USA) и хранили при –70 °С до выделения РНК.

Анализ микрочипов

Образцы тканей посылали в специализированную компанию ЗАО «Геноаналитика» (г. Москва, Россия), где проводили выделение РНК и технологическую часть эксперимента. Для реакции амплификации использовали 400 нг РНК и TotalPrep RNA Labeling Kit с Biotinylated-UTP (Ambion, Austin, TX). Гибридизацию осуществляли на микрочипах RatRef-12 Expression BeadChip (Illumina, Inc., California, USA), включающих 22 524 пробы для 22 228 генов крысы. Последовательности проб были выбраны из базы данных National Center for Biotechnology Information RefSeq database (Release 16; Illumina, San Diego, CA, USA). Гибридизацию, отмывку и окрашивание флуоресцентным реагентом Cy3-стрептавидином проводили в соответствии с рекомендациями Illumina Gene Expression Direct Hybridization Manual. Гибридизацию всех сравниваемых образцов одной ткани проводили на одном стекле. Результаты гибридизации на микрочипах сканировали с помощью Illumina BeadArray reader.

Статистическую обработку результатов гибридизации, включая log2-трансформацию и нормализацию методом квантилей, проводили с помощью программного пакета R/Bioconductor: beadarray (Dunning *et al.*, 2007). Дифференциальную экспрессию генов анализировали с помощью программного пакета R/Bioconductor: limma (Smyth, 2004) с применением эмпирического байесовского подхода и поправки Бенжамини – Хохберга. Для отбора генов, экспрессирующихся в почках только одной из сравниваемых линий, использовали сортировку по параметру ‘detection’ *p*-value, который должен быть меньше 0,01 для всех образцов одной линии (достоверная детекция) и больше 0,1 для всех образцов другой линии (недостоверная детекция). Достоверными считали различия при скорректированном уровне значимости $p \leq 0{,}05$. Для функциональной аннотации дифференциально экспрессирующихся генов применяли Web-инструмент DAVID 2008 (http://david.abcc.ncifcrf.gov) с использованием стандартных значений уровней значимости обогащения терминами GO (Gene Ontology) ≤ 0,1 (Huang *et al.*, 2009a, b).

РЕЗУЛЬТАТЫ

Список генов, достоверно экспрессирующихся в почках только одной из двух сравниваемых линий, представлен в табл. 1. У НИСАГ было репрессировано 20 генов, и восемь генов репрессировано у WAG в корковом веществе почки. В мозговом веществе почки 9 генов репрессировано у НИСАГ и три гена – у WAG.

Среди генов, достоверно экспрессирующихся в почках только одной из двух сравниваемых линий, шесть генов были общими в корковом и мозговом веществе почек. Все 6 генов (*Ankra2, Ctla2a, Gucy1a3, Loc498449, Rpl30, RT1-A2*) были репрессированы у крыс НИСАГ.

Сравнение генов, достоверно экспрессирующихся в почках только одной из двух анализируемых линий крыс, со списком генов, аннотированных в базе данных RGD как гены, имеющие отношение к развитию ГБ, показало, что только ген *Klk1* входит в этот список. Кроме того, гены *Klk1* и *Kng1* аннотированы в базе данных RGD как имеющие отношение к заболеваниям почек – нефросклерозу (*Klk1* и *Kng1)* и почечной недостаточности (*Klk1*).

Среди генов, представленных в табл. 1, четыре гена в корковом веществе почек связаны с уровнем напряжения сосудистой стенки, возникающего в ответ на движение крови (shear stress) (Ekstrand *et al.*, 2010). Проведение функциональной аннотации генов, достоверно экспрессирующихся в почках только одной из двух сравниваемых линий, с помощью Web-инструмента DAVID позволило выявить несколько биологических процессов, характеризующих различия функции почек у гипер- и нормотензивных крыс (табл. 2).

Анализ коркового вещества почек выявил два гена (*Gucy1a3* и *Kng1*), достоверно экспрессирующихся только у нормотензивных крыс и участвующих в контроле вазодилятации и диаметра кровеносных сосудов. Кроме того, функциональная аннотация показала, что ген

Таблица 1

Гены, экспрессирующиеся в почках только у одной из сравниваемых линий крыс НИСАГ и WAG

Chr.	Acc.#	Символ гена	Название гена	Экспрессия	
				НИСАГ	WAG
			kidney_cortex		
X	NM_207595.1	*Ankra2*	ankyrin repeat, family A (RFXANK-like), 2	Нет	Да
9	XM_001063205.1	*Bnip3-ps1*	BCL2/adenovirus E1B interacting protein 3, pseudogene 1	Нет	Да
19	NM_019293.1	*Car5a*	carbonic anhydrase 5a, mitochondrial	Нет	Да
17	XM_001065725.1	*Ctla2a*	cytotoxic T lymphocyte-associated protein 2 alpha	Нет	Да
2	XM_579393.1	*Gucy1a3*	guanylate cyclase 1, soluble, alpha 3	Нет	Да
19	XM_238042.4	*Hhip*	Hedgehog-interacting protein	Нет	Да
1	NM_001005382.1	*Klk1** #	kallikrein 1	Нет	Да
1	XM_001080455.1	*Klk1c10*	kallikrein 1-related peptidase C10	Нет	Да
11	NM_012696.2	*Kng1*#	kininogen 1	Нет	Да
17	XM_346949.2	*LOC361229*	hypothetical LOC361229	Нет	Да
12	XM_344072.2	*LOC363865*	similar to tumor protein, translationally-controlled 1	Нет	Да
15	XM_573708.1	*LOC498449*	similar to Ubiquitin-conjugating enzyme E2 E1 (Ubiquitin-protein ligase E1)	Нет	Да
16	XM_001071886.1	*LOC689753*	similar to K06A9.1b	Нет	Да
X	XR_007560.1	*RGD1560706*	similar to LRRGT00057	Нет	Да
2	XM_001058612.1	*RGD1564247*Δ	similar to SUMO/sentrin specific protease 5	Нет	Да
17	XM_573998.2	*Rnf182*	ring finger protein 182	Нет	Да
7	NM_022699.2	*Rpl30*	ribosomal protein L30	Нет	Да
20	NM_001008829.1	*RT1-A2*	RT1 class Ia, locus A2	Нет	Да
20	NR_002149.1	*Sfta2*	surfactant associated 2	Нет	Да
3	NM_053372.1	*Slpi*	secretory leukocyte peptidase inhibitor	Нет	Да
4	NM_173136.1	*Akr1b8*Δ	aldo-keto reductase family 1, member B8	Да	Нет
1	NM_001025767.1	*Blnk*	B-cell linker	Да	Нет
1	XM_574627.2	*Fam111a*Δ	family with sequence similarity 111, member A	Да	Нет
7	XM_216959.2	*LOC300024*	similar to Ly6-B antigen gene	Да	Нет
19	XM_226326.3	*LOC307731*	similar to L-lactate dehydrogenase A chain (LDH-A) (LDH muscle subunit) (LDH-M)	Да	Нет
20	XM_574750.1	*LOC365566*	similar to Ubiquitin-conjugating enzyme E2S	Да	Нет
17	XR_006738.1	*LOC689842*Δ	similar to Nucleolar GTP-binding protein 1 (Chronic renal failure gene protein) (GTP-binding protein NGB)	Да	Нет
1	NM_022715.2	*Mvp*	major vault protein	Да	Нет
			kidney_medulla		
X	NM_207595.1	*Ankra2*	ankyrin repeat, family A (RFXANK-like), 2	Нет	Да
16	NM_053770.1	*Argbp2*	Arg/Abl-interacting protein ArgBP2	Нет	Да
17	XM_001065725.1	*Ctla2a*	cytotoxic T lymphocyte-associated protein 2 alpha	Нет	Да
2	XM_579393.1	*Gucy1a3*	guanylate cyclase 1, soluble, alpha 3	Нет	Да
15	XM_573708.1	*LOC498449*	similar to Ubiquitin-conjugating enzyme E2 E1 (Ubiquitin-protein ligase E1)	Нет	Да
1	XM_214751.3	*Mrpl18*	mitochondrial ribosomal protein L18	Нет	Да
7	NM_022699.2	*Rpl30*	ribosomal protein L30	Нет	Да
20	NM_001008829.1	*RT1-A2*	RT1 class Ia, locus A2	Нет	Да
20	XM_001055146.1	*Spock2*	sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 2	Нет	Да

Окончание таблицы 1

Chr.	Acc.#	Символ гена	Название гена	Экспрессия	
				НИСАГ	WAG
1	NM_133539.1	*Mrpl17*	mitochondrial ribosomal protein L17	Да	Нет
20	NM_053299.1	*Ubd*	ubiquitin D	Да	Нет

* Гены аннотированы в базе данных RGD как имеющий отношение к развитию гипертонии.
# Гены аннотированы в базе данных RGD как имеющие отношение к развитию заболеваний почек.
Δ Гены, изменяющие уровень экспрессии в гладкомышечных клетках в ответ на воздействие fluid shear stress, т. е. силы, действующей на объект в результате движения жидкости по жесткой поверхности (Ekstrand *et al.*, 2010).

Kng1 участвует в регуляции клеточной адгезии, ответа на стресс и в регуляции биологических процессов, включая иммунные. В мозговом веществе почек функциональная аннотация генов, достоверно экспрессирующихся в почках одной из двух сравниваемых линий, показала различия в экспрессии генов регуляции иммунного ответа у гипер- и нормотензивных крыс.

ОБСУЖДЕНИЕ

В настоящей работе выявлены гены, достоверно экспрессирующиеся в почках только одной линии – гипертензивных крыс НИСАГ или нормотензивных крыс WAG. Транскрипционная активность большинства таких генов не детектирована в почках гипертензивных крыс – 71,4 % в корковом и 75,0 % в мозговом веществе почки. Снижение уровня экспрессии многих генов (67 % из 505 проанализированных генов) было показано при старении почек, что ассоциировано с проявлением таких процессов, как гломерулосклероз, атрофия почечных канальцев, фиброзные изменения в мелких артериях (Melk *et al.*, 2005).

Сравнительные электронно-микроскопические исследования у взрослых (6 мес.) крыс НИСАГ и Wistar показали гипертрофию почечных телец в гипертензивной почке крыс НИСАГ, сопровождающуюся множественными структурными изменениями. Комплекс этих изменений указывал на увеличение функциональной нагрузки на фильтрационный барьер и начальные стадии гломерулярного (Шмерлинг и др., 2001) и реномедуллярного склероза (Филюшина и др., 2013) в почках крыс НИСАГ. Снижение уровня транскрипционной активности ряда генов может быть связано с функциональными нарушениями в почках крыс НИСАГ. Снижение транскрипционной активности большого числа генов было показано также при исследовании животных с ГБ, индуцированной внешними воздействиями, например при инъекции ангиотензина-II (Yuan *et al.*, 2003; Makhanova *et al.*, 2010) или солевой нагрузке (Hopcroft *et al.*, 2010). В эксперименте с солевой нагрузкой снижение экспрессии многих генов связывали с адаптацией организма, направленной против развития гипертонии (Там же).

В отличие от моделей индуцируемой гипертонии, в генетических моделях, таких как крысы НИСАГ, артериальная гипертензия является результатом селекции, в процессе которой мог произойти отбор определенных полиморфизмов, приводящих к изменению (повышению или понижению) уровня экспрессии генов и могущих отражаться на изменении синтеза функциональных белков, вызывая наблюдаемые у крыс НИСАГ отклонения в фенотипе. Однако отбор по определенному признаку в процессе селекции, как правило, влечет за собой и приобретение ряда признаков, контролируемых генами, тесно сцепленными с теми, по которым идет отбор. Исходя из этого нельзя исключить, что выявленные различия в экспрессии ряда генов, представленных в данной работе (см. табл. 1), могут быть обусловлены случайными генными вариациями, не связанными непосредственно с регуляцией уровня артериального давления. Функции многих генов, описанных в настоящей работе как дифференциально экспрессирующиеся в почках гипер- и нормотензивных крыс, не известны. Однако анализ генов с известной функцией позволил выявить ряд важных особенностей в функционировании генов гипертензивной почки. В настоящей работе при анализе

Таблица 2

Функциональная аннотация генов, достоверно экспрессирующихся в почках только одной из двух сравниваемых линий НИСАГ и WAG

Термины GO	Число генов	*p*-value	Символ гена	Название гена
Корковое вещество почек				
Regulation of response to stress	3	1,9E-2	*RT1-A2* *Klk1* *Kng1*	RT1 class I, CE14; RT1 class I, CE16; RT1 class Ia, locus A2; RT1 class Ib, locus Cl; RT1 class Ia, locus A1; RT1 class I, A3 Kallikrein 1 Kininogen 1
Vasodilation, regulation of blood vessel size	2	2,5E-2	*Gucy1a3* *Kng1*	Guanylate cyclase 1, soluble, alpha 3 Kininogen 1
Negative regulation of cell adhesion	2	2,8E-2	*Kng1* *LOC689842*	Kininogen 1 Similar to Nucleolar GTP-binding protein 1 (Chronic renal failure gene protein) (GTP-binding protein NGB); similar to G protein-binding protein CRFG; GTP binding protein 4; similar to isopentenyl diphosphate delta-isomerase type 2
Negative regulation of biological process	5	4,0E-2	*Hhip* *RT1-A2* *Gucy1a3* *Kng1* *LOC689842*	Hedgehog-interacting protein RT1 class I, CE14; RT1 class I, CE16; RT1 class Ia, locus A2; RT1 class Ib, locus Cl; RT1 class Ia, locus A1; RT1 class I, A3 Guanylate cyclase 1, soluble, alpha 3 Kininogen 1 Similar to Nucleolar GTP-binding protein 1 (Chronic renal failure gene protein) (GTP-binding protein NGB); similar to G protein-binding protein CRFG; GTP binding protein 4; similar to isopentenyl diphosphate delta-isomerase type 2
Negative regulation of immune system process	2	7,1E-2	*RT1-A2* *Kng1*	RT1 class I, CE14; RT1 class I, CE16; RT1 class Ia, locus A2; RT1 class Ib, locus Cl; RT1 class Ia, locus A1; RT1 class I, A3 Kininogen 1
Мозговое вещество почек				
Negative regulation of immune response	2	1,8E-2	*RT1-A2* *A2m*	RT1 class I, CE14; RT1 class I, CE16; RT1 class Ia, locus A2; RT1 class Ib, locus Cl; RT1 class Ia, locus A1; RT1 class I, A3 Alpha-2-macroglobulin

транскрипционной активности генов в корковом веществе почек обнаружено три гена (*Klk1, Klk1c10* и *Kng1*), имеющих отношение к функционированию калликреин-кининовой системы (ККС) организма. Все три гена достоверно экспрессировались в почках нормотензивных крыс. В почках гипертензивных крыс их экспрессия не детектировалась. Калликреин-кининовая система является ключевой протеолитической системой, участвующей в регуляции широкого спектра физиологических функций и развитии многих патологических состояний (Елисеева, 2001). Она ингибирует апоптоз, воспалительные процессы, развитие гипертрофии и фиброза и стимулирует ангио- и нейрогенез в сердце, почках, мозге и кровеносных сосудах (Chao, Chao, 2005). В базе данных RGD ген *Klk1* аннотирован как имеющий отношение к развитию ГБ и таким заболеваниям почек, как фиброз и почечная недостаточность. Сниженный уровень

тканевого калликреина отмечают у человека и модельных животных с гипертонией, а также с заболеваниями сердечно-сосудистой системы и почек (Chao, Chao, 2005; Iwai *et al.*, 2005).

У крыс со спонтанной гипертензией (SHR) было показано снижение АД при воздействии калликреина (Wang *et al.*, 1995). Усиление функции ККС в результате введения калликреина крысам с соль-чувствительной гипертонией линии SS (Dahl salt-sensitive rats) в течение длительного времени ослабляло у них повреждение почек (Uehara *et al.*, 1997). Установлено, что при этом происходят обратное развитие повреждений канальцевого аппарата и инволюция склеротических нарушений гломерулярного аппарата почки (Chao *et al.,* 1998).

Кининоген активирует пролиферацию эн-дотелиальных клеток (Pérez *et al.*, 2006) и фибробластов (Aravena *et al.*, 2005) и подавляет пролиферацию лимфоцитов (Acuna-Castillo *et al.*, 2005). Давно показано, что злокачественная ГБ ассоциирована с низким уровнем кининогена в плазме крови (Almeida *et al.*, 1981). Под действием калликреинов из кининогенов выщепляются биологически активные пептиды-кинины, например брадикинин, обладающий сосудорасширяющим действием. Брадикинин служит мощным стимулятором высвобождения эндотелий-зависимых расслабляющих факторов, таких как оксид азота (NO), эндотелий-зависимый фактор гиперполяризации и простациклин (Bönner *et al.*, 1990).

Gucy1a3 (guanylate cyclase soluble subunit α3) кодирует альфа-субъединицу растворимой гуанилатциклазы. Активируя последнюю, NO увеличивает образование циклического гуанозинмонофосфата (цГМФ) в гладкомышечных клетках, что приводит к расслаблению гладких мышц сосудов, их расширению и снижению АД (Rapoport *et al.*, 1983).

Ранее при изучении структурных особенностей капилляров почечных клубочков крыс НИСАГ выявлены снижение количества капилляров и их неравномерное распределение в клубочках. Было сделано заключение, что найденные изменения гломерулярных капилляров и примыкающих к ним подо- и мезангиоцитов свидетельствуют о нарушении функции гломерулярного фильтра и соответствуют морфологическим признакам гломерулосклероза (Лазарев и др., 2002). Можно предположить, что снижение уровня экспрессии генов *Klk1, Kng1* и *Gucy1a3* в корковом веществе почек крыс НИСАГ может нарушать процессы циркуляции крови в почечных тельцах и участвовать в процессе развития гипертензивного состояния.

В мозговом веществе почек среди генов с достоверно различающейся экспрессией у крыс НИСАГ и WAG найдены гены (*RT1-A2, A2m*), относящиеся к регуляции иммунного ответа. В настоящее время активно изучается и показана важная роль воспалительных процессов в развитии эссенциальной ГБ (Androulakis *et al.*, 2011). Воспаление и оксидативный стресс могут участвовать в процессах ремоделирования и повреждения сосудов при ГБ (Touyz, 2004). Ген *RT1-A2* локализован на хромосоме 20 в локусе АД Bp195, описанном при картировании соль-зависимой гипертонии (Moreno *et al.*, 2003).

Ген *A2m* локализован на хромосоме 4 и также попадает в локусы АД, описанные в исследованиях при изучении соль-зависимой ГБ (Schork *et al.*, 1995; Garrett *et al.*, 2002). Ген *A2m* кодирует белок, являющийся главным ингибитором металлопротеиназ. Изменения баланса металлопротеиназ и их ингибиторов могут быть связаны с ремоделированием сосудов при экспериментальной ГБ, обусловленной сужением почечной артерии одной из почек (two kidney-one clip hypertension - 2K-1C) (Castro *et al.*, 2010). Как показано нами, у крыс НИСАГ локус, достоверно ассоциированный с уровнем АД в покое и после воздействия эмоционального стресса, находится на хромосоме 1 и пересекается с локусом относительного веса селезенки, что указывает на возможную связь развития стресс-зависимой гипертензии у крыс НИСАГ с изменением генетического контроля функции селезенки (Редина и др., 2014).

Известно, что при ГБ наблюдается эндотелиальная дисфункция (Ghiadoni *et al.*, 2012). Экспрессия генов и белков в эндотелиальных (Chiu *et al.*, 2009) и гладкомышечных клетках (Ekstrand *et al*., 2010) может изменяться при воздействии, оказываемом движущейся кровью на стенки сосудов (shear stress). При таком воздействии в гладкомышечных клетках изменяется экспрессия многих генов, ассоциированных с ответом на оксидативный стресс и с гипоксией (Ekstrand *et al*., 2010).

В настоящей работе в корковом веществе почек найдено четыре гена, которые изменяют уровень экспрессии в гладкомышечных клетках при увеличении стресса «трения крови» (shear stress) (см. табл. 1). Эти гены могут быть рассмотрены как гены-кандидаты для дальнейших исследований функции гладкомышечных и эндотелиальных клеток сосудов в почках гипертензивных крыс НИСАГ.

ЗАКЛЮЧЕНИЕ

Исследования детерминации генетического контроля заболеваний и физиологических признаков, проведенные на модельных животных, могут иметь прямое отношение к изучению механизмов патологии человека (Korstanje, DiPetrillo, 2004). Сходство генетического контроля некоторых заболеваний человека и животных подтверждается сходством списков генов, аннотированных в базе данных RGD как гены, ответственные за развитие целого ряда заболеваний, в том числе и артериальной гипертонии (http://rgd.mcw.edu).

Мы предполагаем, что в дальнейшей работе среди выявленных нами генов с дифференциальной экспрессией в почках крыс НИСАГ и WAG будут определены дополнительные гены-кандидаты стресс-зависимой гипертензии, а также найдены полиморфизмы, потенциально связанные с процессами ремоделирования стенки сосудов и старения почек. Изучение генетической базы гипертензивных состояний у экспериментальных животных расширяет наши знания о причинах и механизмах развития ГБ человека.

БЛАГОДАРНОСТИ

Работа поддержана грантом РФФИ № 13-04-01492 и бюджетным проектом VI.53.2.4.

Авторы выражают благодарность компании «ЗАО Геноаналитика» за проведение технологической части анализа микрочипов.

ЛИТЕРАТУРА

Абрамова Т.О., Редина О.Е., Смоленская С.Э., Маркель А.Л. Повышенный уровень экспрессии мРНК гена Ephx2 в почках гипертензивных крыс линии НИСАГ (ISIAH) // Молекул. биология. 2013. Т. 47. № 6. С. 942–948.

Адаричев В.А., Корохов Н.П., Остапчук В. и др. Характеристика линий крыс с нормотензивным и гипертензивным статусом методом геномного фингерпринтинга // Генетика. 1996. Т. 32. С. 1669–1677.

Елисеева Е. Ангиотензин-превращающий фермент, его физиологическая роль // Вопросы медицинской химии. 2001. Т. 47. № 1. С. 43–54.

Лазарев В.А., Филюшина Е.Е., Бузуева И.И. и др. Структурные особенности капилляров почечных клубочков крыс гипертензивной линии НИСАГ // Бюл. СО РАМН. 2002. № 1. С. 89–92.

Маркель А.Л., Калашникова Е.В., Горякин С.В. и др. Характеристика функциональной активности симпатоадреналовой системы у гипертензивных крыс линии НИСАГ // Бюл. эксперим. биол. мед. 2006. Т. 141. № 3. С. 244–247.

Пыльник Т.О., Плетнева Л.С., Редина О.Е. и др. Влияние эмоционального стресса на экспрессию мРНК гена альфа-ENaC в почке гипертензивных крыс линии НИСАГ // Доклады Академии наук. 2011. Т. 439. № 4. С. 563–565.

Редина О.Е., Смоленская С.Э., Абрамова Т.О., Маркель А.Л. Генетические локусы, контролирующие вес селезенки и уровень артериального давления у крыс НИСАГ со стресс-зависимой артериальной гипертензией // Молекулярная биология. 2014. Т. 48. № 3. С. 407–415.

Федосеева Л.А., Рязанова М.А., Антонов Е.В. и др. Экспрессия генов рениновой системы почки и сердца у гипертензивных крыс линии НИСАГ // Биомедицинская химия. 2011. Т. 57. № 4. С. 410–419.

Филюшина Е.Е., Шмерлинг М.Д., Бузуева И.И. и др. Структурные особенности реномедуллярных интерстициальных клеток крыс гипертензивной линии НИСАГ // Бюл. эксперимент. биологии и медицины. 2013. Т. 155. № 3. С. 391–396.

Хворостова В., Горякин С.В., Петрова Г.В. и др. Характеристика гипоталамо-гипофизарно-надпочечниковой системы у гипертензивных крыс линии НИСАГ // Российск. физиол. журнал им. И.М. Сеченова. 2002. Т. 88. № 11. С. 1423–1432.

Хворостова В., Калашникова Е.В., Черкасова О.П. и др. Особенности экспрессии гена глюкокортикоидного рецептора у гипертензивных крыс линии НИСАГ // Российск. физиол. журнал им. И.М. Сеченова. 2003. Т. 89. № 12. С. 1523–1528.

Шмерлинг М.Д., Филюшина Е.Е., Лазарев В.А. и др. Ультраструктурные особенности почечных телец у крыс с наследственной индуцированной стрессом артериальной гипертензией // Морфология. 2001. Т. 120. № 6. С. 70–74.

Acuna-Castillo C., Aravena M., Leiva-Salcedo E. *et al.* T-kininogen, a cystatin-like molecule, inhibits ERK-dependent lymphocyte proliferation // Mech. Ageing Dev. 2005. V. 126. No. 12. P. 1284–1291.

Almeida F.A., Stella R.C., Voos A. *et al*. Malignant hypertension: a syndrome associated with low plasma kininogen and kinin potentiating factor // Hypertension. 1981. V. 3. No. 6. Pt. 2. P. II-46–49.

Androulakis E., Tousoulis D. *et al.* Inflammation in hypertension: current therapeutic approaches // Curr. Pharm. Des. 2011. V. 17. No. 37. P. 4121–4131.

Aravena M., Pérez C., Pérez V. *et al.* T-kininogen can either induce or inhibit proliferation in Balb/c 3T3 fibroblasts, depending on the route of administration // Mech. Ageing Dev. 2005. V. 126. No. 3. P. 399–406.

Bareyre F.M., Schwab M.E. Inflammation, degeneration and regeneration in the injured spinal cord: insights from DNA microarrays // Trends Neurosci. 2003. V. 26. No. 10. P. 555–563.

Bönner G., Preis S., Schunk U. *et al.* Hemodynamic effects of bradykinin on systemic and pulmonary circulation in healthy and hypertensive humans // J. Cardiovasc. Pharmacol. 1990. V. 15. Suppl. 6. P. S46–56.

Castro M.M., Rizzi E., Prado C. *et al.* Imbalance between matrix metalloproteinases and tissue inhibitor of metalloproteinases in hypertensive vascular remodeling // Matrix Biol. 2010. V. 29. No. 3. P. 194–201.

Chao J., Chao L. Kallikrein-kinin in stroke, cardiovascular and renal disease // Exp. Physiol. 2005. V. 90. No. 3. P. 291–298.

Chao J., Zhang J.J., Lin K.F., Chao L. Adenovirus-mediated kallikrein gene delivery reverses salt-induced renal injury in Dahl salt-sensitive rats // Kidney Int. 1998. V. 54. No. 4. P. 1250–1260.

Chiu J.J., Usami S., Chien S. Vascular endothelial responses to altered shear stress: pathologic implications for atherosclerosis // Ann. Med. 2009. V. 41. No. 1. P. 19–28.

Cowley A.W. Long-term control of arterial blood pressure // Physiol. Rev. 1992. V. 72. No. 1. P. 231–300.

Dunning M.J., Smith M.L., Ritchie M.E., Tavare' S. beadarray: R classes and methods for Illumina bead-based data // Bioinformatics. 2007. V. 23. No. 16. P. 2183–2184.

Ekstrand J., Razuvaev A. *et al*. Tissue factor pathway inhibitor-2 is induced by fluid shear stress in vascular smooth muscle cells and affects cell proliferation and survival // J. Vasc. Surg. 2010. V. 52. No. 1. P. 167–175.

Garrett M., Joe B. *et al*. Identification of blood pressure quantitative trait loci that differentiate two hypertensive strains // J. Hypert. 2002. V. 20. No. 12. P. 2399–2406.

Ghiadoni L., Taddei S., Virdis A. Hypertension and endothelial dysfunction: therapeutic approach // Curr. Vasc. Pharmacol. 2012. V. 10. No. 1. P. 42–60.

Guyton A.C. Long-term arterial pressure control: an analysis from animal experiments and computer and graphic models // Am. J. Physiol. 1990. V. 259. No. 5. Pt 2. P. R865–877.

Havlik R. *et al*. Blood pressure aggregation in families // Am. J. Epidemiol. 1979. V. 110. No. 3. P. 304–312.

Hirschhorn J.N. Genetic approaches to studying common diseases and complex traits // Pediatr. Res. 2005. V. 57. No. 5. Pt. 2. P. 74R–77R.

Hopcroft L.E., McBride M.W., Harris K.J. *et al*. Predictive response-relevant clustering of expression data provides insights into disease processes // Nucleic Acids Res. 2010. V. 38. No. 20. P. 6831–6840.

Huang D.W., Sherman B.T., Lempicki R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists // Nucleic Acids Res. 2009a. V. 37. No. 1. P. 1–13.

Huang D.W., Sherman B.T., Lempicki R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources // Nat. Protoc. 2009b. V. 4. No. 1. P. 44–57.

Iwai N., Yasui N., Naraba H. *et al*. Klk1 as one of the genes contributing to hypertension in Dahl salt-sensitive rat // Hypertension. 2005. V. 45. No. 5. P. 947–953.

Korstanje R., DiPetrillo K. Unraveling the genetics of chronic kidney disease using animal models // Am. J. Physiol. Renal. Physiol. 2004. V. 287. No. 3. P. F347–352.

Levy D., DeStefano A.L., Larson M.G. *et al*. Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the framingham heart study // Hypertension. 2000. V. 36. No. 4. P. 477–483.

Levy D., Ehret G.B., Rice K. *et al.* Genome-wide association study of blood pressure and hypertension // Nat. Genet. 2009. V. 41. No. 6. P. 677–687.

Lifton R.P., Gharavi A.G., Geller D.S. Molecular mechanisms of human hypertension // Cell. 2001. V. 104. No. 4. P. 545–556.

Lynn K., Li L., Lin Y. *et al*. A neural network model for constructing endophenotypes of common complex diseases: an application to male young-onset hypertension microarray data // Bioinformatics. 2009. V. 25. No. 8. P. 981–988.

Makhanova N.A., Crowley S.D., Griffiths R.C., Coffman T.M. Gene expression profiles linked to AT1 angiotensin receptors in the kidney // Physiol. Genomics. 2010. V. 42A. No. 3. P. 211–218.

Markel A.L. Development of a new strain of rats with inherited stress-induced arterial hypertension // Genetic hypertension. Paris: Colloque INSERM, 1992. V. 218. P. 405–407.

Markel A.L., Maslova L.N., Shishkina G.T. *et al*. Developmental influences on blood pressure regulation in ISIAH rats // Development of the hypertensive phenotype: basic and clinical studies. Amsterdam; Lausanne; NewYork; Oxford; Shannon; Singapore; Tokyo: Elsevier, 1999. V. 19. P. 493–526.

Markel A.L., Redina O.E., Gilinsky M.A. *et al*. Neuroendocrine profiling in inherited stress-induced arterial hypertension rat strain with stress-sensitive arterial hypertension // J. Endocrinol. 2007. V. 195. No. 3. P. 439–450.

Melk A., Mansfield E.S., Hsieh S.C. *et al*. Transcriptional analysis of the molecular basis of human kidney aging using cDNA microarray profiling // Kidney Int. 2005. V. 68. No. 6. P. 2667–2679.

Moreno C., Dumas P., Kaldunski M. *et al*. Genomic map of cardiovascular phenotypes of hypertension in female Dahl S rats // Physiol. Genomics. 2003. V. 15. No. 3. P. 243–257.

Mukherjee S., Belbin T., Spray D. *et al*. Microarray technology in the investigation of diseases of myocardium with special reference to infection // Front Biosci. 2006. V. 11. P. 1802–1813.

Mullins L.J., Bailey M.A., Mullins J.J. Hypertension, kidney, and transgenics: a fresh perspective // Physiol. Rev. 2006. V. 86. No. 2. P. 709–746.

Newton-Cheh C., Johnson T., Gateva V. *et al.* Genome-wide association study identifies eight loci associated with blood pressure // Nat. Genet. 2009. V. 41. No. 6. P. 666–676.

Park S.K., Prolla T.A. Gene expression profiling studies of aging in cardiac and skeletal muscles // Cardiovasc. Res. 2005. V. 66. No. 2. P. 205–212.

Pérez V., Leiva-Salcedo E. *et al*. T-kininogen induces endothelial cell proliferation // Mech. Ageing Dev. 2006. V. 127. No. 3. P. 282–289.

Rapoport R.M., Draznin M.B., Murad F. Endothelium-dependent relaxation in rat aorta may be mediated through cyclic GMP-dependent protein phosphorylation // Nature. 1983. V. 306. No. 5939. P. 174–176.

Redina O.E., Smolenskaya S.E., Abramova T.O. *et al*. Differential transcriptional activity of kidney genes in hypertensive ISIAH and normotensive WAG rats // Clinical Experimental Hypertension. 2014. DOI: 10.310 9/10641963.2014.954711.

Schork N.J., Krieger J., Trolliet M. *et al*. A biometrical genome search in rats reveals the multigenic basis of blood pressure variation // Genome Res. 1995. V. 5. No. 2. P. 164–172.

Smyth G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments // Stat. Appl. Genet. Mol. Biol. 2004. V. 3. No. 1. P. Article 3.

Touyz R.M. Reactive oxygen species, vascular oxidative stress, and redox signaling in hypertension: what is the clinical significance? // Hypertension. 2004. V. 44. No. 3. P. 248–252.

Uehara Y., Hirawa N., Numabe A. *et al*. Long-term infusion of kallikrein attenuates renal injury in Dahl salt-sensitive rats // Am. J. Hypertens. 1997. V. 10. No. 5. Pt. 2. P. 83S–88S.

Viemann D., Schulze-Osthoff K., Roth J. Potentials and pitfalls of DNA array analysis of the endothelial stress response // Biochim. Biophys. Acta. 2005. V. 1746. No. 2. P. 73–84.

Wang C., Chao L., Chao J. Direct gene delivery of human tissue kallikrein reduces blood pressure in spontaneously hypertensive rats // J. Clin. Invest. 1995. V. 95. No. 4. P. 1710–1716.

Wang Y., O'Connell J.R., McArdle P.F. *et al*. From the Cover: Whole-genome association study identifies STK39 as a hypertension susceptibility gene // Proc. Natl. Acad. Sci. USA. 2009. V. 106. No. 1. P. 226–231.

Yuan B., Liang M., Yang Z. *et al*. Gene expression reveals vulnerability to oxidative stress and interstitial fibrosis of renal outer medulla to nonhypertensive elevations of ANG II // Am. J. Physiol. Regul. Integr. Comp. Physiol. 2003. V. 284. No. 5. P. R1219–1230.

THE DOWNREGULATION OF GENES CONTROLLING VASCULAR TONE IN KIDNEYS OF ISIAH RATS WITH STRESS-INDUCED ARTERIAL HYPERTENSION

O.E. Redina[1], L.O. Klimov[1], N.I. Ershov[1], T.O. Abramova[1], L.N. Ivanova[1, 2], A.L. Markel[1, 2]

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

Summary

The transcriptional activity of genes was studied in kidneys of hypertensive ISIAH and normotensive WAG rats in order to detect genes significantly expressed in kidneys of just one of the analyzed strains. Gene profiling was performed on the Illumina RatRef-12 Expression BeadChip microarray platform. The expression of three genes (*Klk1, Klk1c10*, and *Kng1*) related to the kallikrein-kinin system was significant in the WAG renal cortex but was not detected in hypertensive kidneys. The downregulation of these three genes and *Gucy1a3* in ISIAH renal cortex suggests the weakened function of the kallikrein-kinin system in hypertensive kidneys, which may cause blood circulation disturbances in renal glomeruli and mediate the development of hypertension in ISIAH rats. The functional annotation of the genes significantly expressed in renal medulla of just one of the compared rat strains revealed the genes involved in immune response regulation.

Key words: ISIAH rats, transcriptional activity of genes, microarrays, stress-induced arterial hypertension, emotional stress.

УДК 575.112:577.322.2:004.94

СТРУКТУРНЫЕ И ДИНАМИЧЕСКИЕ ОСОБЕННОСТИ МУТАНТОВ БЕЛКА SOD1, АССОЦИИРОВАННЫХ С БОКОВЫМ АМИОТРОФИЧЕСКИМ СКЛЕРОЗОМ

© 2014 г. **Н.А. Алемасов, Н.В. Иванисенко, В.А. Иванисенко**

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 24 сентября 2014 г. Принята к публикации 21 октября 2014 г.

Одной из причин гибели нейронов головного и спинного мозга при заболевании боковым амиотрофическим склерозом является образование внутриклеточных белковых агрегатов, вызванных мутациями в гене *SOD1*. Ранее показано, что продолжительность жизни пациентов отрицательно коррелирует с термостабильностью мутантных форм белка SOD1, носителями которых они были. В настоящей работе сделано предположение, что усиливать агрегацию мутантов SOD1 может не только дестабилизация структуры белка за счет разрушения водородных связей, но также возникновение новых водородных связей, стабилизирующих патогенную конформацию мутанта. Методом молекулярной динамики оценено время существования водородных связей в белке. Установлено, что корреляция этой оценки с продолжительностью жизни пациентов ($R = 0,89$, $p < 0,00001$) оказалась существенно сильнее корреляции, полученной ранее на основе анализа термостабильности мутантов.

Ключевые слова: SOD1, нейродегенеративные заболевания, молекулярная динамика, водородные связи, предсказание, время жизни.

ВВЕДЕНИЕ

Боковой амиотрофический склероз (БАС) – нейродегенеративное заболевание, при котором поражаются двигательные нейроны (моторная кора головного мозга, передние рога спинного мозга и ядра черепно-мозговых нервов). Поражение двигательных нейронов вызывает прогрессирующий паралич и атрофию мышц, приводящую в итоге к смерти пациентов из-за нарушения функции легких (Haidet-Phillips *et al.*, 2011; Alavi *et al.*, 2013). Частота встречаемости заболевания составляет один – два случая на 100 тыс. (Brown, 1997; Alavi *et al.*, 2013). По разным данным, от 1 до 23 % случаев БАС имеют наследственную природу (Eisen *et al.*, 2008; Alavi *et al.*, 2013). Наиболее распространенными наследственными причинами развития БАС считаются мутации в гене *SOD1*, кодирующем фермент супероксиддисмутазу-1 (Bosco *et al.*, 2010; Ivanova *et al.*, 2014). На данный момент известно свыше 170 мутаций этого гена, вызывающих различные формы БАС (http://alsod.iop.kcl.ac.uk/).

На молекулярном уровне мутации приводят к различным изменениям в структуре белка SOD1. Одним из следствий мутации белков супероксиддисмутазы-1 (SOD1) является образование внутриклеточных агрегатов (Bruijn *et al.*, 1998; Deng *et al.*, 2006). Предложено большое количество объяснений агрегации SOD1 (Ross, Poirier, 2004).

Byström с соавт. (2010) показано, что продолжительность жизни пациентов со дня первой диагностики заболевания отрицательно коррелирует ($R = 0,78$) с потерей термостабильности мутантов, носителями которых они были. Однако найденная закономерность хорошо работала только для ограниченного круга мутаций, в то время как эффекты большого количества мутаций, связанных с изменением заряда аминокислотных остатков, не подчинялись построенной зависимости.

Большинство работ (Stathopulos *et al.*, 2003; Sato *et al.*, 2005; Chiti, Dobson, 2009), посвященных проблеме агрегации мутантных белков, основано на экспериментальных данных по изменению их термодинамической стабильности и упускает анализ отдельных факторов, включающих сети водородных связей, солевые мостики и другие физико-химические, структурные и конформационные характеристики белков.

В настоящей работе сделано предположение, что некоторые мутации могут оказывать пространственно-распределенный эффект на физико-химические и структурные характеристики, определяющие подверженность белка агрегации. При этом происходят локальные разнонаправленные изменения характеристик различных участков пространственной структуры белка, компенсирующие друг друга в масштабе всей структуры. Такой интегральный показатель, как термостабильность, не позволяет выявить закономерности агрегации белков. В частности, патогенные мутации SOD1, влияющие на сеть водородных связей путем разрушения одних и возникновения других связей, могут повышать вероятность перехода из нативной конформации SOD1 в ее патогенную форму.

Для проверки этой гипотезы методом молекулярной динамики (МД) (Alder, Wainwright, 1959) оценена стабильность водородных связей, которая рассчитана как суммарное время существования водородной связи за период моделирования. Нами рассмотрены следующие мутации белка SOD1 человека: Ала4Вал, Цис6Ала, Вал7Глу, Гли12Арг, Гли37Арг, Лей38Вал, Гли41Сер, Гис46Арг, Гис48Глн, Асп76Вал, Асн86Лиз, Ала89Вал, Асп90Ала, Гли93Арг, Вал94Ала, Глу100Гли, Асп101Асн, Цис111Сер, Гли114Ала, Вал118Лей, Асп124Вал, Асп125Гис, Гли127Арг. Показано, что разница во временах существования водородных связей для мутантов и белка дикого типа имеет корреляцию ($R = 0{,}89$, $p < 0{,}00001$) с продолжительностью жизни более сильную, чем корреляция, основанная на термостабильности, выявленная в работе Byström с соавт. (2010).

Построена регрессионная модель, предсказывающая продолжительность жизни пациентов по структурным характеристикам мутантов SOD1. С использованием подхода случайных перестановок показано статистически значимое отличие продолжительности жизни пациентов, предсказанной с помощью регрессии, от продолжительности, ожидаемой по случайным причинам (*S*).

МЕТОДЫ

Протокол молекулярной динамики

В работе для моделирования в рамках метода молекулярной динамики был применен программный комплекс AMBER 12 (Salomon-Ferrer *et al.*, 2013). В качестве аппаратной платформы выступал гибридный высокопроизводительный кластер ЦКП «Биоинформатика» (http://bioinformatics.bionet.nsc.ru/), который содержит ускорители NVIDIA Tesla M2090. При моделировании МД применен следующий набор параметров: шаг интегрирования – 2 фс; радиус отсечения взаимодействий – 10 Å ; температура – 300 K; время моделирования – 50 нс. В качестве модели воды использована TIP3P. Кубическая ячейка имела сторону 12 Å и состояла из более чем 20 000 молекул воды, растворенного в ней белка (гомодимер, 154 аминокислотных остатка, PDBID: 2C9V) из 4 376 атомов. В зависимости от заряда мутантного белка в раствор добавляли около 36 ионов Na^+ и 30 ионов Cl^-, что соответствовало концентрации 0,137 моль. Для обеспечения достоверности результатов моделирования каждая траектория повторена 5 раз. В итоге для 24 форм белка получено 120 траекторий, общей протяженностью более 6 мкс.

Белок для нормального функционирования в клетке требует включения в свой состав ионов меди и цинка. Известно, что количество атомов меди ~0,2/димер, а цинка ~1,5/димер (Ayers *et al.*, 2014), это говорит о том, что моделирование может осуществляться без учета иона меди.

Данные по времени жизни пациентов

Данные по временам жизни пациентов, носителей известных мутаций, взяты из базы данных ALS mutation database (Yoshida *et al.*, 2010) и исследования Wang с соавт. (2008). Данные по мутациям Ала89Вал (Sato *et al.*, 2005) и Гли127Арг (Holmøy *et al.*, 2010) были найдены в

отдельных работах. Всего использованы сведения о 18 заменах: Ала4Вал, Вал7Глу, Гли12Арг, Гли37Арг, Лей38Вал, Гли41Сер, Гис46Арг, Гис48Глн, Асп76Вал, Асн86Лиз, Ала89Вал, Асп90Ала, Гли93Арг, Глу100Гли, Асп101Асн, Гли114Ала, Асп125Гис, Гли127Арг. Регрессии построены на основе 16 мутаций, без информации об Ала89Вал и Гли127Арг, поскольку по каждой из этих двух мутаций есть сведения только об одном пациенте-носителе.

Мутация Вал94Ала была промоделирована для проверки гипотезы: замена валина на аланин в области белка, доступной растворителю, не должна влиять на подверженность его агрегации и, следовательно, на возникновение БАС. Время жизни пациентов, носителей мутации Вал94Ала, в литературных источниках не найдено.

Структурные характеристики мутантных форм белка и его дикого типа

Для построения модели, связывающей время жизни пациента, несущего определенную мутацию, и структурные характеристики белка, была создана таблица, состоящая из N строк и M столбцов, где N – количество аминокислотных остатков белка SOD1 ($N = 306$), M – количество исследуемых мутаций ($M = 23$).

В ячейках таблицы содержалось относительное усредненное по пяти траекториям время существования внутримолекулярных водородных связей (НВ), контактирующих с остатком $n \in [1; N]$ из мутанта $m \in [1; M]$. Величина НВ для остатка n рассчитана путем суммирования времен существования водородной связи в траектории МД для всех атомов остатка вне зависимости от того, являются ли они донорами или акцепторами данной связи. Далее проводилась нормировка НВ на всю длину траектории. В таблицу заносилась разность НВ мутантов и белка дикого типа. На следующем шаге были отобраны аминокислотные остатки, которые удовлетворяли следующему критерию: коэффициент корреляции Пирсона между строками таблицы, содержащими НВ, и соответствующими временами жизни пациентов с мутациями из обучающей выборки должен превосходить пороговое значение R = 0,5. Величина порога выбрана таким образом, что можно сказать: большая часть НВ мутантов для выбранной строки коррелирует со временами жизни пациентов. Для построения регрессионного уравнения, связывающего время жизни пациентов и структурные характеристики мутантных SOD1, носителями которых они были, рассчитывали суммарный вектор из выбранных по порогу строк.

Сравнение построенной модели со случайными моделями

В качестве метода статистической проверки построенной модели на корректность был выбран метод рандомизации, использованный на двух этапах: «shuffle» и «permutation». На этапе «shuffle» применен метод, аналогичный методу bootstrap (Efron, 1979), но без возвращения выбранных случайным образом значений.

Все имеющиеся литературные данные были разбиты на обучающую и тестовую выборку в пропорции 2:1. Таким образом, из 16 мутаций с известными литературными данными для построения регрессии использовано 11 мутаций, для проверки предсказания – пять мутаций.

1. Этап «shuffle» заключался в таком построении обучающей выборки, чтобы из имеющихся 16 мутаций выбрать случайным образом две трети, то есть 11. Тестовая выборка составлена из оставшихся пяти мутаций. Столбцы таблицы с НВ перестанавливали в соответствии с совершенной выборкой мутаций.

2. Этап «permutation» представлял собой случайное перемешивание (permutation) элементов вектора из 16 элементов, соответствующих литературным данным о времени жизни пациентов с данными мутациями. При этом таблица со значениями НВ не изменялась. Пропорции для обучающей и тестовой выборки аналогичны предыдущему этапу: 11/5.

Для каждой выборки из «shuffle» и «permutation» находился свой набор строк таблицы с параметрами НВ, коррелирующими с соответствующей строкой времен жизни пациентов.

После для обоих этапов происходило построение распределений величины S – среднего квадрата разности между вектором длины 5, содержащим предсказанные времена жизни, и литературными данными для соответствующих пяти мутаций из тестовой выборки. Было

проведено 10^5 выборок (shuffle) и столько же перемешиваний (permutation). Полученные распределения величины *S* сравнивали с применением непараметрических критериев χ^2, Краскела – Уоллиса (Kruskal, Wallis, 1952) и Колмогорова – Смирнова (Kolmogorov, 1933; Smirnov, 1948). Данный подход использован для оценки степени случайности предсказаний: в случае если распределение *S*, полученное на этапе «shuffle», не отличимо в смысле упомянутых выше критериев от *S* на этапе «permutation», то это свидетельствует в пользу случайности предсказаний.

Построение регрессии

В случае достоверного отличия распределений *S*, полученных на предыдущем шаге для построения финальной регрессии и последующего предсказания времени жизни пациентов, использован весь набор известных литературных данных, состоящий из 16 мутаций. Таким образом, предсказания были сделаны для оставшейся доли мутаций, для которых нет литературных данных о времени жизни пациентов: Цис6Ала, Ала89Вал, Вал94Ала, Цис111Сер, Вал118Лей, Асп124Вал, Гли127Арг.

РЕЗУЛЬТАТЫ

На первом шаге анализа осуществлено моделирование молекулярной динамики 24 димеров SOD1 для расчета сетей водородных связей, образованных в структуре анализируемых белков и оценки их характеристик. В качестве основной характеристики водородной связи рассчитывалась ее стабильность как суммарное время существования в течение моделирования МД, нормированное на всю ее длительность. Для проверки качества оценки стабильности было установлено, что выбранная длительность моделирования в 50 нс достаточна для выхода этого показателя на постоянный уровень.

Перед построением финальной регрессии методом случайного перемешивания была проанализирована достоверность построенной модели, используемой для предсказания времени жизни пациентов (см. Материалы и методы). Показано статистически значимое отличие наблюдаемой величины *S* от ожидаемой по случайным причинам (см. таблицу), отсюда следует, что результаты предсказаний модели взяты из генеральной совокупности, достоверно отличной от случайной. Из рис. 1 видно, что форма кривых для обоих распределений различна, и при этом график для распределения *S* на этапе «permutation» расположен правее графика *S* для «shuffle». Величина этого сдвига косвенно отражена в значениях критериев сравнения. Таким образом, предсказания модели также в среднем лучше случайных предсказаний.

Таблица HB содержит количество строк на порядок больше числа столбцов. Следовательно, при построении на ее основе регрессии может возникнуть проблема переобучения. Настоящий шаг в том числе дает возможность проверить, имеет ли место проблема: если бы переобучение состоялось, то распределения *S* для обоих этапов были бы не отличимы друг от друга и регрессия оказалась бы недостоверной. Как видно (таблица, рис. 1), этого не произошло.

После успешного анализа достоверности модели была построена финальная регрессия. Коэффициент линейной корреляции Пирсона между изменением стабильности водородных связей, вызванным мутациями в белке SOD1, и временем жизни пациентов-носителей этих мутаций составил $R = 0{,}8877$, $p = 0{,}0000045$. Построенная модель линейной регрессии (время

Значимость различия между ожидаемым по случайным причинам и наблюдаемым распределением величины *S*

Критерий	Критическое значение критерия	Уровень значимости отличий
χ^2	4 117	$p < 0{,}000001$
Краскела – Уоллиса	796	$p < 4 \times 10^{-175}$
Колмогорова – Смирнова	0,2064	$p < 5 \times 10^{-186}$

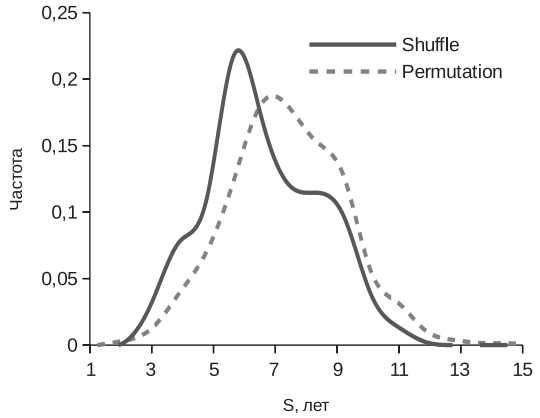


Рис. 1. Распределения наблюдаемого (сплошная линия) и ожидаемого по случайным причинам (прерывистая) значений *S* при проверке достоверности модели. Графики получены сглаживанием сплайн-кривой между каждой из 32 точек гистограммы.

жизни = 7,04 × *HB* + 4,19) позволила предсказать времена жизни для двух пациентов-носителей мутаций Ала89Вал и Гли127Арг (рис. 2).

В работе Holmøy с соавт. (2010) приведены сведения о времени жизни пациента, носителя мутации Гли127Арг, составившем 7 мес. Как видно из рисунка, предсказания хорошо согласуются с данным фактом. Также имелась информация об одном пациенте с мутацией Ала89Вал (Sato *et al.*, 2005). Заболевание у этого пациента длилось около 6 лет, и на момент публикации этих авторов пациент был жив. Согласно предсказаниям модели, мутация Ала89Вал занимает промежуточное положение по степени тяжести заболевания, что согласуется с описанными выше фактами. Предсказания также были сделаны еще для пяти мутаций (Цис6Ала, Вал94Ала, Цис111Сер, Вал118Лей, Асп124Вал), для которых нет литературных данных о времени жизни пациентов. Оказалось, что все эти мутации распределились от 1,8 до 12,7 года на шкале времени жизни пациентов. Рисунок 3 демонстрирует чувствительные к мутациям аминокислотные остатки, обнаруженные на основе анализа модели. На этапе построения финальной регрессии выявлено 11 аминокислотных остатков, для которых значение HB коррелирует со временем жизни пациентов.

ОБСУЖДЕНИЕ

В настоящей работе мы предположили, что как стабилизация, так и дестабилизация структуры белка влияют на увеличение вероятности нахождения его в метастабильном патогенном состоянии (Ross, Poirier, 2004) и, следовательно, на подверженность белка агрегации (Bruijn *et al.*, 1998). В частности, структура может стабилизироваться за счет появления новых водородных связей в структуре мутанта, отсутствующих в белке дикого типа. Вновь возникшие водородные связи, в свою очередь, могут стабилизировать белок в патогенной конформации, увеличивая его способность образовывать агрегаты.

На поверхности свободной энергии в пространстве конформационных состояний белка такие мутации могут понижать локальный минимум энергии, соответствующий патогенной конформации белка. Разрушение водородных связей в результате мутаций может снижать потенциальный барьер на поверхности свободной энергии между нативным и патогенным состояниями белка, что повышает вероятность перехода между этими двумя состояниями.

Анализ изменения стабильности индивидуальных водородных связей в результате мутаций позволил нам рассчитать динамику и перестройку сети водородных связей, характеристики которой имели высокую корреляцию с продолжительностью жизни пациентов. Ранее показано, что в процессе динамики белков происходит постоянный разрыв первичных водородных связей и создание альтернативных, что предотвращает резкое увеличение конформационной энтропии при повышении температуры и, следовательно, поддерживает стабильность белка (Khechinashvili *et al.*, 2006). Воздействие таких дестабилизирующих факторов по-разному влияет на водородные связи, образованные между сближенными и удаленными друг от друга в первичной структуре аминокислотными остатками (Nisius, Grzesiek, 2012).

Согласно нашим расчетам (рис. 3), в частности, мутации в белке SOD1 оказывают стабилизирующее влияние на водородные связи, образованные аминокислотным остатком Гис120, а для водородных связей остатка Сер142 этот эффект является дестабилизирующим. Извест-

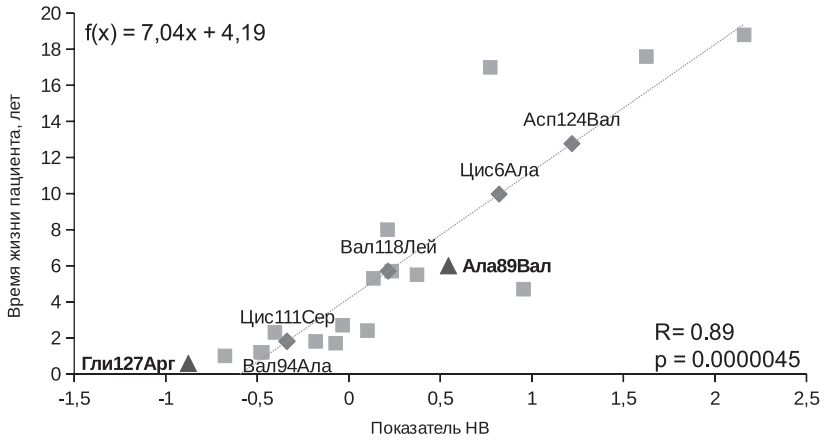


Рис. 2. Регрессионная зависимость между показателем НВ в мутантах белка SOD1 и временем жизни пациентов-носителей этих мутаций. Квадратами (■) показаны мутантные белки, входящие в обучающую выборку. Треугольниками (▲) обозначены предсказания времени жизни для двух пациентов-носителей мутаций Ала89Вал и Гли127Арг. Ромбами (♦) отмечены предсказания для мутаций Цис6Ала, Вал94Ала, Цис111Сер, Вал118Лей, Асп124Вал.

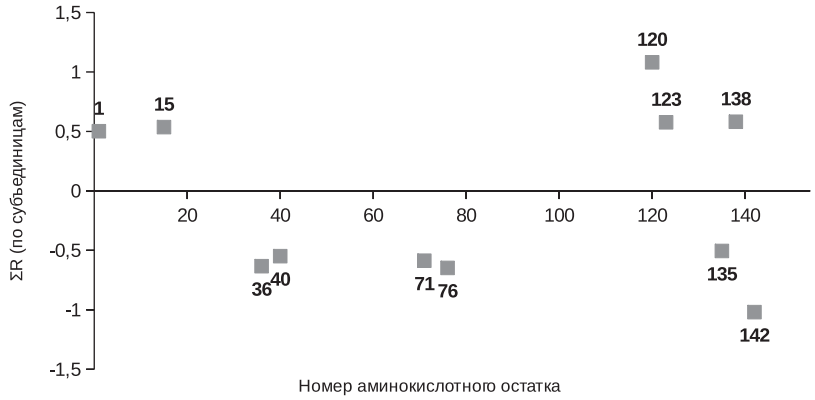


Рис. 3. Профиль корреляций изменения стабильности индивидуальных водородных связей со временем жизни пациентов. По оси абсцисс отложены номера аминокислотных остатков в структуре мономера, образующих водородные связи. По оси ординат отложена сумма коэффициентов корреляции для аминокислотных остатков по обеим субъединицам белка.

но, что Гис120 непосредственно участвует в связывании иона меди, необходимого для ферментативной активности SOD1 (Nagano, 2012). Примером отрицательной корреляции между стабильностью водородных связей в мутантных белках SOD1 и продолжительностью жизни пациентов, носителей этих мутаций, может служить сеть водородных связей, образуемых с участием остатка Гис71. Этот остаток связывает ион Zn, что стабилизирует структуру SOD1 (Arnesano *et al.*, 2004; Ding, Dokholyan, 2008).

Известно, что остатки 38–40 стабилизируют β-бочонок (Deng *et al.*, 1995). При этом Лей38 имеет ван-дер-ваальсовый контакт с Гис43, который, в свою очередь, связан с каталитически важным остатком Арг143 (Muneeswaran *et al.*, 2014). В работе Wright с соавт. (2013) выделены два района (1–30 и 90–120), имеющие особое значение для агрегации мутантных белков SOD1. Оказалось, что в этих районах стабильность водородных связей коррелирует с продолжительностью жизни пациентов (рис. 3).

ЗАКЛЮЧЕНИЕ

Таким образом, нами выдвинута гипотеза о влиянии как стабилизации, так и дестабилизации структуры белка на увеличение веро-

ятности нахождения белка в промежуточном, метастабильном состоянии и, следовательно, на подверженность его агрегации. Показано, что в локальной дестабилизации белка чаще участвуют те аминокислотные остатки, которые связаны с выполняемой ими той или иной функцией. Выявлена важная роль водородных связей в увеличении вероятности локальной дестабилизации его структуры. В дальнейшем планируется расширить список исследуемых мутаций белка SOD1.

БЛАГОДАРНОСТИ

Работа поддержана междисциплинарными интеграционными проектами СО РАН № 130, 39, 47, а также проектом фундаментальных исследований СО РАН VI.61.1.2.

ЛИТЕРАТУРА

Alavi A., Nafissi S., Rohani M. *et al.* Genetic analysis and SOD1 mutation screening in Iranian amyotrophic lateral sclerosis patients // Neurobiol. Aging. 2013. V. 34. No. 5. P. 1516.e1–1516.e8.

Alder B.J., Wainwright T.E. Studies in Molecular Dynamics. I. General Method // J. Chem. Phys. 1959. V. 31. No. 2. P. 459.

Arnesano F., Banci L., Bertini I. *et al.* The unusually stable quaternary structure of human Cu, Zn-superoxide dismutase 1 is controlled by both metal occupancy and disulfide status // J. Biol. Chem. 2004. V. 279. No. 46. P. 47998–48003.

Ayers J., Lelie H., Workman A. *et al.* Distinctive features of the D101N and D101G variants of superoxide dismutase 1; two mutations that produce rapidly progressing motor neuron disease // J. Neurochem. 2014. V. 128. No. 2. P. 305–314.

Bosco D.A., Morfini G., Karabacak N.M. *et al.* Wild-type and mutant SOD1 share an aberrant conformation and a common pathogenic pathway in ALS // Nat. Neurosci. 2010. V. 13. No. 11. P. 1396–1403.

Brown R.H. Amyotrophic lateral sclerosis. Insights from genetics // Arch. Neurol. 1997. V. 54. No. 10. P. 1246–1250.

Bruijn L.I., Houseweart M.K., Kato S. *et al.* Aggregation and motor neuron toxicity of an ALS-linked SOD1 mutant independent from wild-type SOD1 // Science. 1998. V. 281. No. 5384. P. 1851–1854.

Byström R., Andersen P.M., Grцbner G., Oliveberg M. SOD1 mutations targeting surface hydrogen bonds promote amyotrophic lateral sclerosis without reducing apo-state stability // J. Biol. Chem. 2010. V. 285. No. 25. P. 19544–19552.

Chiti F., Dobson C.M. Amyloid formation by globular proteins under native conditions // Nat. Chem. Biol. 2009. V. 5. No. 1. P. 15–22.

Deng H.X., Tainer J.A., Mitsumoto H. *et al.* Two novel SOD1 mutations in patients with familial amyotrophic lateral sclerosis // Hum. Mol. Genet. 1995. V. 4. No. 6. P. 1113–1116.

Deng H.X., Shi Y., Furukawa Y. *et al.* Conversion to the amyotrophic lateral sclerosis phenotype is associated with intermolecular linked insoluble aggregates of SOD1 in mitochondria // Proc. Natl. Acad. Sci. U. S. A. 2006. V. 103. No. 18. P. 7142–7147.

Ding F., Dokholyan N. V. Dynamical roles of metal ions and the disulfide bond in Cu, Zn superoxide dismutase folding and aggregation // Proc. Natl. Acad. Sci. U. S. A. 2008. V. 105. No. 50. P. 19696–19701.

Efron B. Bootstrap methods: another look at the jackknife // Ann. Stat. 1979. V. 7. No. 1. P. 1–26.

Eisen A., Mezei M.M., Stewart H.G. *et al.* SOD1 gene mutations in ALS patients from British Columbia, Canada: clinical features, neurophysiology and ethical issues in management // Amyotroph. Lateral Scler. 2008. V. 9. No. 2. P. 108–119.

Haidet-Phillips A.M., Hester M.E., Miranda C.J. *et al.* Astrocytes from familial and sporadic ALS patients are toxic to motor neurons // Nat. Biotechnol. 2011. V. 29. No. 9. P. 824–828.

Holmøy T., Wilson J.A., von der Lippe C. *et al.* G127R: A novel SOD1 mutation associated with rapidly evolving ALS and severe pain syndrome // Amyotroph. Lateral Scler. 2010. V. 11. No. 5. P. 478–480.

Ivanova M.I., Sievers S.A., Guenther E.L. *et al.* Aggregation-triggering segments of SOD1 fibril formation support a common pathway for familial and sporadic ALS // Proc. Natl. Acad. Sci. U. S. A. 2014. V. 111. No. 1. P. 197–201.

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.

Kolmogorov A. Sulla determinazione empirica di una legge di distribuzione // G. dell'Istituto Ital. degli Attuari. 1933. V. 4. P. 1–11.

Kruskal W.H., Wallis W.A. Use of Ranks in One-Criterion Variance Analysis // J. Am. Stat. Assoc. 1952. V. 47. No. 260. P. 583–621.

Muneeswaran G., Kartheeswaran S., Muthukumar K. *et al.* Comparative structural and conformational studies on H43R and W32F mutants of copper-zinc superoxide dismutase by molecular dynamics simulation // Biophys. Chem. 2014. V. 185. P. 70–78.

Nagano S. Oxidative Modifications of Cu, Zn-Superoxide Dismutase (SOD1)–The Relevance to Amyotrophic Lateral Sclerosis (ALS) // Amyotrophic Lateral Sclerosis. InTech. 2012. P. 301–312.

Nisius L., Grzesiek S. Key stabilizing elements of protein structure identified through pressure and temperature perturbation of its hydrogen bond network // Nat. Chem. 2012. V. 4. No. 9. P. 711–717.

Ross C.A., Poirier M.A. Protein aggregation and neurodegenerative disease // Nat. Med. 2004. V. 10 Suppl. P. S10– S17.

Salomon-Ferrer R., Case D.A., Walker R.C. An overview of the Amber biomolecular simulation package // Wiley Interdiscip. Rev. Comput. Mol. Sci. 2013. V. 3. No. 2. P. 198–210.

Sato T., Nakanishi T., Yamamoto Y. *et al.* Rapid disease progression correlates with instability of mutant SOD1 in familial ALS // Neurology. 2005. V. 65. No. 12. P. 1954–1957.

Smirnov N. Table for estimating the goodness of fit of empirical distributions // Ann. Math. Stat. 1948. V. 19. P. 279–281.

Stathopulos P.B., Rumfeldt J.A.O., Scholz G.A. *et al.* Cu/Zn superoxide dismutase mutants associated with amyotrophic lateral sclerosis show enhanced formation of aggregates in vitro // Proc. Natl. Acad. Sci. U. S. A. 2003. V. 100. No. 12. P. 7021–7026.

Wang Q., Johnson J.L., Agar N.Y.R., Agar J. N. *et al.* Protein aggregation and protein instability govern familial amyotrophic lateral sclerosis patient survival // PLoS Biol. 2008. V. 6. No. 7. P. e170.

Wright G.S.A., Antonyuk S.V., Kershaw N.M. *et al.* Ligand binding and aggregation of pathogenic SOD1 // Nat. Commun. 2013. V. 4. P. 1758.

Yoshida M., Takahashi Y., Koike A. *et al.* A mutation database for amyotrophic lateral sclerosis // Hum. Mutat. 2010. V. 31. No. 9. P. 1003–1010.

STRUCTURAL AND DYNAMIC PROPERTIES OF MUTANTS OF THE SOD1 PROTEIN ASSOCIATED WITH AMYOTROPHIC LATERAL SCLEROSIS

N.A. Alemasov, N.V. Ivanisenko, V.A. Ivanisenko

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

Summary

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease, which affects motor neurons in the brain and spinal cord and leads to patients' death. One of the causes of motor neuron degeneration and death is the formation of intracellular protein aggregates formed by a mutant SOD1 protein. Recently, it has been shown that the survival time of ALS patients with specific mutation in SOD1 gene inversely correlates with thermodynamic stability of the SOD1 mutant protein. In the present paper, we hypothesize that mutant SOD1 aggregation can be facilitated by not only destabilization due to hydrogen bonds disruption but also by formation of new hydrogen bonds, which can stabilize intermediate "pathogenic" conformations of the mutant SOD1 protein. Molecular dynamics simulations were conducted to estimate frequencies of hydrogen bond occurrence in the protein structure. It was shown that the regression model based on frequencies of hydrogen bond occurrence significantly better correlated with patients' survival time ($R = 0.89$, $p < 0.00001$) than the estimation based on thermodynamic stability analysis of mutant SOD1 proteins.

Key words: SOD1, neurodegenerative diseases, molecular dynamics, hydrogen bonds, prediction, survival time.

УДК 61:575; 658.011.56

ЦИРКАДНЫЕ ЧАСЫ МЛЕКОПИТАЮЩИХ: ГЕННАЯ СЕТЬ И КОМПЬЮТЕРНЫЙ АНАЛИЗ

© 2014 г. **О.А. Подколодная[1], Н.Н. Подколодная[1, 2], Н.Л. Подколодный[1, 3]**

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

Поступила в редакцию 25 сентября 2014 г. Принята к публикации 21 октября 2014 г.

В работе представлены результаты реконструкции и анализа генной сети циркадных часов млекопитающих. Применение методов теории графов позволило провести анализ структуры генной сети и выделить центральную компоненту регуляции циркадного ритма, которая включает базовые регуляторные контуры, проходящие через ключевой элемент циркадных часов – белок Clock/Bmal1. Использование кластерного анализа позволило выявить подсистемы, имеющие четкую биологическую интерпретацию и участвующие в функционировании циркадных часов путем взаимодействия с центральной компонентой. Такая структурная модель, включающая центральную компоненту и взаимодействующие с ней функциональные подсистемы, может быть основой для построения математической модели динамики генной сети регуляции циркадного ритма.

Ключевые слова: циркадные часы, генные сети, методы анализа графов.

ВВЕДЕНИЕ

Циркадные часы являются универсальным адаптивным механизмом эукариот. Они обеспечивают согласованное протекание процессов в живых организмах на всех уровнях от молекулярно-генетических до поведенческих. Общепринято мнение, что основой функционирования этого механизма служат молекулярно-генетические осцилляторы, присутствующие практически в каждой клетке организма (Albrecht, 2012). У млекопитающих главный водитель циркадного ритма, сформированный нейронами супрахиазматических ядер гипоталамуса (СХЯ), синхронизует ритм сети периферических циркадных осцилляторов, расположенных как в мозге (вне СХЯ), так и в периферических тканях (Там же). Принципиальная структура клеточных циркадных осцилляторов одинакова в нейронах СХЯ и клетках периферических тканей (Yagita *et al.*, 2001). Универсальным регуляторным модулем циркадных часов млекопитающих является отрицательная обратная связь, обеспечивающая ритмическую регуляцию транскрипции генов, кодирующих белки Per1-3 и Cry1-2 (Reppert, Weaver, 2001). Коротко она может быть описана следующим образом: гетеродимерный транскрипционный фактор (ТФ) Clock/Bmal1 активирует транскрипцию генов *Per* и *Cry*, а белковый гетеродимерный комплекс Per/Cry за счет белок-белковых взаимодействий подавляет активность ТФ Clock/Bmal1. Это, в свою очередь, приводит к замыканию регуляторного контура за счет снижения уровня транскрипции генов *Per* и *Cry* и соответствующему восстановлению активности Clock/Bmal1. В ходе этого цикла ритмически изменяется уровень гетеродимеров Per/Cry и активность ТФ Clock/Bmal1. Кроме данной основной петли в циркадных часах млекопи-

тающих описаны и другие транскрипционно-трансляционные петли как с отрицательными, так и с положительными обратными связями, которые обеспечивают стабильность функционирования циркадных часов (Albrecht, 2012).

Помимо этого, функционирование циркадных часов регулируется различными механизмами, включая дополнительные обратные связи с регуляторами транскрипции, посттрансляционную модификацию белков (циклы фосфорилирования/дефосфорилирования и ацетилирования/деацетилирования, полиADP-рибозилирование, протеосомная деградация белков), стабилизацию, деградацию, транспорт RNA и белков и др. (Virshup *et al.*, 2007; Asher, Schible, 2011; Albrecht, 2012; Morf *et al.*, 2012; Подколодная, 2014). Можно отметить характерную особенность молекулярно-генетического механизма циркадных часов – наличие обратных связей их компонентов различных уровней с центральными компонентами часов Clock/Bmal1 и Per(1-3)/Cry(1-2) и между собой и избыточность компонент часового механизма.

Оба фактора существенно затрудняют исследование механизмов и построение более детальных математических моделей, учитывающих тканеспецифические особенности регуляции циркадных часов (Ripperger, Brown, 2010). Первым и одним из важнейших этапов построения моделей молекулярно-генетических систем является реконструкция и структурный анализ генной сети, описывающей основные молекулярно-генетические события и их взаимосвязи в рассматриваемой системе. Цель работы – реконструкция генной сети циркадного осциллятора млекопитающих и анализ ее структуры с помощью методов теории графов.

МАТЕРИАЛЫ И МЕТОДЫ

Методы реконструкции генных сетей

На первом этапе работы генная сеть Circadian Rhythm была реконструирована с использованием системы GeneNet (Ananko *et al.*, 2005). Система позволяет накапливать и систематизировать данные о генах, РНК, белках, малых молекулах, реакциях и регуляторных событиях, экстрагированных из научных публикаций и баз данных, используя специализированный редактор генных сетей. Экстрагированная на основе ручной аннотации информация представляется в текстовом и графическом виде. Для дальнейшей работы по реконструкции генной сети на основе первичных знаний, накопленных в базе данных GeneNet, о молекулярно-генетических взаимодействиях при регуляции циркадного ритма млекопитающих разработаны специальные программные средства, позволяющие:

1. Экстрагировать информацию из XML-представления генной сети в системе GeneNet в стандартные форматы gml, graphml и sbml.
2. Осуществлять реконструкцию генной сети для определенного вида организма на основе использования информации по близким видам организмов из базы данных GeneNet.
3. Объединять различные генные сети, представленные в базе данных GeneNet.
4. Осуществлять проверку целостности генной сети, восстановление недостающей информации из других баз данных и удаление дублирующей информации, включая описание одних и тех же молекулярных событий с различной степенью детальности.
5. Осуществлять проверку связности графа генной сети и выделять связные компоненты графа.

Методы анализа графов генных сетей

Для анализа структуры генной сети использованы библиотека программ Networkx 1.9 (http://networkx.github.io/), системы Cytoscape v.3.1.1 (http://www.cytoscape.org/) и Gephi v.0.8.2-beta (http://gephi.github.io/), а также собственные программы, написанные на языке Python 2.7, в рамках которых осуществлялись загрузка и преобразование графа генной сети, подключение внешних библиотек, вызов методов анализа и интеграция результатов обработки.

Для выявления центральных или наиболее важных вершин в генной сети применяли методы расчета различных показателей центральности вершин (индексы структурной важности вершин графа) (Koschützki, Schreiber, 2008), реализованные в библиотеке программ Networkx 1.9, включая:

- центральность по степени вершин (degree centrality) – наиболее распространенный и простой показатель, который соответствует

степени вершины, т. е. числу дуг, связанных с вершиной, нормированной на число вершин графа. Для ориентированного графа генной сети вводятся раздельные показатели для входных и выходных степеней вершины. Вершины с максимальным количеством дуг, или хабы, обычно соответствуют белкам – регуляторным молекулам;

- центральность вершины графа по близости (closeness centrality) – обратно пропорциональна сумме кратчайших расстояний от этой вершины до других вершин в графе. Этот показатель определяет важность конкретной вершины графа для быстрой передачи информации в графе;
- центральность вершины по посредничеству (betweenness centrality) – взвешенная сумма всех кратчайших путей между вершинами в графе, проходящих через данную вершину. Этот показатель определяет важность вершины графа с точки зрения ее влияния на пути передачи информации между вершинами графа;
- эксцентриситет вершин – максимальное расстояние от вершины до других вершин графа. Вершины с минимальным эксцентриситетом находятся в центре графа.

Важным методом анализа ориентированных графов является поиск сильносвязных компонент и регуляторных контуров в генной сети, который осуществлен нами с использованием алгоритма R. Tarjan (1973), реализованного в пакете программ Networkx 1.9. Сильносвязной компонентой графа называется подграф, в котором существует путь между любыми вершинами подграфа. Таким образом, вершины любого регуляторного контура в графе генных сетей входят в максимальную сильносвязную компоненту этого графа.

Кластеризация вершин и выявление структурных модулей генной сети регуляции циркадного ритма выполнялись с помощью итерационного метода Louvain, который обеспечивает оптимальное разбиение графа на структурные модули, или кластеры (Blondel *et al.*, 2008). Для генерации случайных графов применены методы, реализованные в библиотеке программ Networkx 1.9. Для визуализации графа генных сетей использована система Gephi и Cytoscape.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Реконструкция генной сети Circadian rhythm

Генная сеть циркадного осциллятора реконструирована на основе анализа и систематизации данных, представленных в 279 научных публикациях, а также информации из баз данных SWISS-PROT и Entrez Gene. Вносились данные, полученные при исследовании клеточных линий или тканей и органов мыши, крысы или человека. Реконструированный фрагмент генной сети регуляции циркадного ритма включает 54 гена, 122 белка, 16 небелковых субстанций, 436 реакций и регуляторных событий. Для анализа генной сети построен ее граф (рис. 1).

В табл. 1 представлены основные характеристики реконструированной генной сети Circadian rhythm. Размер графа определяется числом вершин (676) и дуг (936). Диаметр графа (19) соответствует максимальному расстоянию между его вершинами. Средняя степень вершин графа зависит от представления графа генной сети и детальности описания механизмов регуляции. В систему GeneNet в отличие от других вариантов описания генных сетей (KEGG, PathwayDB, REACTOME и другие базы данных) включено описание регуляторных событий, которые имеют меньшее значение степени вершин. Поэтому наблюдаемое значение средней степени вершин графа (2,74) нельзя сразу сравнивать с показателями генных сетей, полученных из других источников. Для сравнения необходимо предварительно генные сети привести к единому представлению.

Поэтому для выявления особенностей графа генной сети Circadian rhythm проведено сравнение его характеристик с характеристиками моделей случайных графов Эрдеша – Реньи (Erdős – Rényi model) и малых миров (Watts – Strogatz model), которые часто используют для структурного моделирования биологических сетей (Erdős, Rényi, 1959; Watts, Strogatz, 1998; Newman, 2003; Barabasi, Oltvai, 2004). На рис. 2 представлено распределение степеней вершин графа генной сети циркадного ритма, а также моделей случайных графов: модель Реньи с тем же количеством узлов и дуг и модель малых миров со средней степенью вершин, как в генной сети. Средняя степень вершин у всех графов

Рис. 1. Граф генной сети Circadian rhythm. Размер вершин графа пропорционален степени вершины. Наибольшие размеры вершин соответствуют белкам Clock/Bmal1, per1/cry1, per2/cry2, Sirt1.

Таблица 1

Характеристика графа генной сети Circadian rhythm

Показатель	Значение
Кол-во вершин	676
Кол-во дуг	936
Средняя степень вершин	2,74
Диаметр графа	19
Коэффициент ассоциативности вершин по степени	– 0,156
Средний коэффициент кластеризации	0,0072

практически совпадает. Можно видеть, что распределение степеней вершин в графе генной сети Circadian rhythm отличается от моделей случайных графов. В графе генной сети присутствуют специфичные для нее неоднородности, обусловленные различными типами вершин: генов, РНК, белков, реакций и регуляторных событий. В частности, граф генной сети имеет вершины со степенью более 10, вероятность появления которых в моделях случайных графов очень мала ($p < 0{,}0005$ для модели случайных графов Реньи и $p < 10^{-6}$ для модели случайных графов малых миров) (рис. 2). Максимальные степени вершин характерны для четырех компонент графа, составляющих основу циркадного осциллятора – транскрипционного фактора

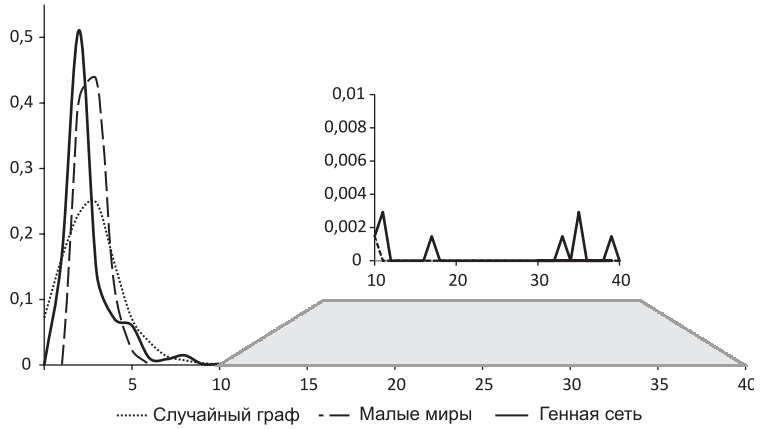


Рис. 2. Распределение степеней вершин графа генной сети Circadian rhythm и моделей случайных графов (ось абсцисс – степени вершин, ось ординат – частота распределения числа вершин графа с данной степенью). Сплошная линия – для генной сети, пунктирная – для случайного графа с тем же количеством вершин и дуг (модель случайных графов Реньи), штриховая линия – для модели случайных графов малых миров со средней степенью вершин, как в генной сети.

Clock/Bmal1 (значение степени вершины 39), регуляторов его активности гетеродимерных комплексов Per1/Cry1 (35) и Per2/Cry2 (35) и NAD-зависимой деацетилазы Sirt1 (33). Большое количество связей также у активной формы коактиватора транскрипции Pgc1a (17), ТФ Rev-erb alpha (11), генов *Arntl* (11) и *Nr1d1*(10). В табл. 2 представлены характеристики некоторых вершин графа генной сети Circadian rhythm, соответствующих регуляторным белкам и белковым комплексам.

Показателями, определяющими важность вершин графа генной сети для выполнения различного типа функций, являются степень вершины и различные типы центральности вершин. Следует отметить, что для вершин с максимальными значениями степени значения центральности различного типа также максимальны. Вершины, имеющие большое число дуг, как правило ТФ, связывают регуляторными путями других участников регуляции циркадных часов между собой, так как находятся на перекрестках путей.

Минимальный эксцентриситет (eccentricity) в реконструированной генной сети Circadian rhythm имеют вершины графа, соответствующие белкам Clock/Bmal1, Err alpha, Dec1, Dec2, Pgc1a 177p538p, Per1/Cry1, Per2/Cry2 и Sirt1. Таким образом, по результатам анализа можно сделать выводы, что центральными вершинами в генной сети Circadian rhythm являются белки Clock/Bmal1, Per1/Cry1, Per2/Cry2, Sirt1 и Pgc1a 177p538p, которые играют важную роль в передаче регуляторных сигналов в генной сети циркадного осциллятора.

Кластерный анализ

На следующем этапе анализа выявляли структурные модули, или кластеры, в графе генной сети Circadian rhythm. Близкие друг другу вершины объединены в единый кластер. Каждая вершина может входить только в один кластер. В результате анализа выявлено 18 кластеров с числом вершин более трех, которые были упорядочены в соответствии с числом входящих в них вершин. В табл. 3 представлена биологическая интерпретация некоторых наиболее интересных примеров кластеров.

Самый большой кластер объединяет 28 вершин, формирующих пути превращения (образование комплексов, циклы фосфорилирования и дефосфорилирования) основных негативных регуляторов циркадного осциллятора – белков Per(1-2) и Cry(1-2) и киназ Ck1e и Ck1d, которые вносят основной вклад в определение величины его периода. Двадцать вершин кластера 3 образуют путь регуляции транскрипции гена *Per1*, ТФ Crebp, GR, Dec1 и Dec2 и посттранскрипционную регуляцию экспрессии его

Таблица 2

Характеристика некоторых вершин в генной сети Circadian rhythm

Вершина	Степень вершины	Центральность по			Эксцентриситет
		степени	близости	посредничеству	
Clock/Bmal1	39	0,058	0,239	0,452	11
Per1/Cry1	35	0,0519	0,20	0,094	11
Per2/Cry2	35	0,0519	0,20	0,090	11
Sirt1	33	0,049	0,18	0,186	12
Pgc1a 177p538p	17	0,025	0,18	0,11	11
Rev-erb alpha	11	0,016	0,16	0,06	14
Arntl	11	0,016	0,19	0,06	14
Nr1d1	10	0,015	0,19	0,06	14
Per1	8	0,012	0,20	0,075	13
Ror alpha	8	0,012	0,18	0,03	12
Prox1	8	0,012	0,17	0,01	12
Per2	8	0,012	0,20	0,083	12
Cyp7	8	0,012	0,1	0,03	14
Bhlhe40	8	0,012	0,18	0,04	14

Таблица 3

Примеры биологической интерпретации некоторых кластеров графа генной сети Circadian rhythm

Номер кластера	Кол-во вершин	Биологическая интерпретация
1	28	Образование комплексов Per1/2 и Cry1/2 с киназами CK1e и CK1d, а также циклы фосфорилирования и дефосфорилирования белков Per и киназ
3	20	Экспрессии гена *Per1*
5	11	Регуляция уровня ТФ Clock/Bmal1за счет образование комплексов белков Clock и Bmal1 с Dec1, Dec2 и Id2
7	11	Убиквитинирование Per1 и Per2
8	10	Экспрессия гена ТФ Dbp и регуляция этим ТФ транскрипции генов *Top1* и *Cyp3A4*
11	9	Экспрессия и регуляция активности белка Pgc1a (ген *Ppargc1a*)
12	9	Убиквитинирование Cry1 и Cry2
13	8	ТФ Clock/Bmal1 и компоненты, подавляющие его активность различными путями
15	7	Путь синтеза NAD^+ и регуляция им активности фермента Parp1
16	7	Образование TF Ppar gamma/Rxr alpha: регуляция транскрипции гена *Pparg* фактором Cebpa, который транскрибируется с гена *Cebpa*
17	7	Регуляция экспрессии гена *Ldlr* с участием ТФ Hes1 и Hes6
18	7	Убиквитинирование Cry1

RNA-связывающим белком Lark. Кластеры 18 (7 вершин), 12 (9 вершин) и 7 (11 вершин) представляют пути убиквитинирования белков Per1, Per2, Cry1 Cry2. Кластер 13 (8 вершин) объединяет белки, подавляющие активность ТФ Clock/Bmal1, используя различные механизмы. Кластер 11 объединяет 9 вершин – компонент пути экспрессии гена *Ppargc1a*, кодирующего коактиватор транскрипции Pgc1a и регуляцию активности этого белка киназой AMPK, активность которого регулируется, в свою очередь, соотношением ATP/AMP.

Другим примером может служить кластер 9, компоненты которого представляют собой путь синтеза NAD^+ и регуляцию активности фермента Parp1 соотношением NAD^+/NAM. Таким образом, можно констатировать, что проведенная нами кластеризация вершин графа генной сети Circadian rhythm позволила выявить биологически интерпретируемые группы объектов.

Мотивы

Как уже указано выше, особенность генной сети Circadian rhythm составляет большое количество обратных связей между ее элементами. Мы называем регуляторным мотивом подграф, созданный элементами генной сети, образующими регуляторный контур с обратной связью. В отличие от известного понятия «неслучайные структурные мотивы» в теории графов (Newman, 2003; Kim *et al.*, 2011), регуляторный мотив необязательно является часто повторяющейся структурой в графе, однако важно, что он несет определенную функциональную нагрузку. Для выявления таких мотивов использованы методы поиска регуляторных контуров в графе генной сети. Рассмотрим некоторые из регуляторных мотивов данной генной сети (рис. 3).

Самый короткий регуляторный мотив – основная регуляторная петля циркадного осциллятора, включающая ТФ Clock/Bmal1 гены *Per* и *Cry* и гетеродимер Per/Cry. Здесь отрицательная обратная связь реализуется за счет белок-белковых взаимодействий гетеродимеров Clock/Bmal1 и Per/Cry (рис. 3, *а*).

Более длинные мотивы (рис. 3, *б–ж*) образованы петлями как с отрицательными (рис. 3, *б*, *в*), так и с положительными обратными связями (рис. 3, *е*, *ж*). Отрицательные обратные связи в мотивах *б–г* реализуются за счет подавления транскрипции гена *Arntl*, кодирующего субъединицу Bmal1 гетеродимера Clock/Bmal1, ТФ Rev-erb alpha, Err alpha, Klf10, в то время как в мотиве *д* за счет посттранскрипционной модуляции экспрессии Bmal1, осуществляемой miR-142-3p (гены *Rev-erba*, *Erra*, *Klf10* и *Mir142* являются мишенями ТФ Clock/Bmal1).

В мотивах *е*, *ж* положительная обратная связь осуществляется за счет активации транскрипции гена *Arntl* ТФ Ror alpha и Ppar alpha, гены которых, в свою очередь, также являются мишенями ТФ Clock/Bmal1. Как видим, центральный положительный регулятор данной генной сети ТФ Clock/Bmal1 участвует во всех

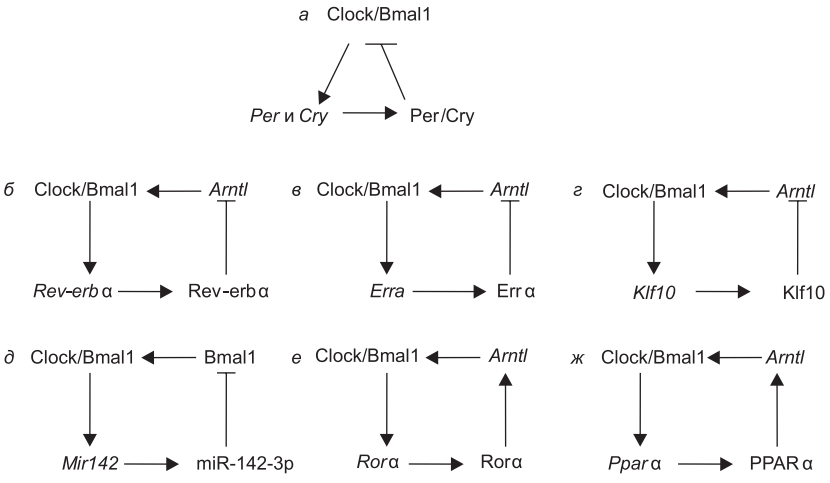


Рис. 3. Примеры регуляторных мотивов в генной сети Circadian Rhythm (комментарии см. в тексте).

этих мотивах, а также имеет максимальную степень вершин в графе, максимальный уровень центральности графа и минимальный эксцентриситет. Можно предположить, что центральная компонента регуляции циркадного ритма должна включать ТФ Clock/Bmal1, который обеспечивает взаимодействие с другими участниками процесса регуляции циркадного ритма. Такое взаимодействие обеспечено большим числом дублирующих путей, из которых наибольший интерес представляют пути с минимальным временем их реализации, которые и определяют главную регуляторную компоненту. В связи с этим для дальнейшего анализа структуры генной сети Circadian rhythm и выявления центральной компоненты регуляции циркадного ритма мы избрали следующую стратегию:

1. Выделение классов реакций или регуляторных событий с различным временем реализации, например модификация, метаболические реакции, транспорт, процессы транскрипции или трансляции и др. Определение весов дуг в графе для реакций или регуляторных событий различных классов. В настоящей работе мы выделяли только два уровня класса реакций и событий: а) медленные (вес = 1): транспорт, процессы транскрипции или трансляции; б) быстрые (вес = 0,01 << 1,0): модификации белков, метаболические реакции и т. д. Использование весов дуг в графе позволяет считать, что время передачи сигнала между объектами в генной сети пропорционально взвешенному расстоянию между соответствующими вершинами в графе генной сети (сумма весов дуг пути между вершинами).

2. Поиск всех кратчайших путей, имеющих минимальное взвешенное расстояние от Clock/bmal1 до регуляторных элементов сети: ТФ, корегуляторов транскрипции и миРНК, представленных в генной сети, считая, что они являются источниками регуляторных сигналов и регуляторных контуров (Clock/Bmal1; Atf4; Atf-5; Bmal1; Cebpa; Clock; Creb; Dbp; Dec1; Dec2; E4bp4; Err alpha; Foxo1; GR; Hes1; Hes6; Hif-1; Hlf; Hltf; Hnf4 alpha; Id2; Klf10; MyoD; Nrf1; Ppar alpha; Ppar alpha/Rxr alpha; Ppar gamma/Rxr alpha; Pparg; Prox1; Rar alpha; RelB; Rev-erb alpha; Ror alpha; Ror gamma; Rxr alpha; Shp; Smad3; Sirt1; Top1; Per1; Per2; Cry1; Cry2; miR-142-3p; miR-419).

3. Поиск всех кратчайших путей, имеющих минимальное взвешенное расстояние от найденных мишеней ТФ Clock/Bmal1 до Clock/Bmal1 (обратные связи).

4. Объединение всех полученных путей в граф и выделение в нем максимальной сильносвязной компоненты. Эта компонента будет объединять все кратчайшие регуляторные контуры (в смысле времени регуляции), связывающие Clock/Bmal1 с другими регуляторами циркадного ритма.

Такая процедура позволила выявить сильносвязную компоненту, соответствующую графу регуляторных контуров Clock/Bmal1, включающую 163 вершины и 188 дуг, объединенных в 27 регуляторных контуров (рис. 4).

Из 27 обратных связей 10 оказались отрицательными, 17 положительными. Выявленные нами регуляторные контуры, в конечном итоге, модулируют уровень экспрессии гена *Arntl,* а также и активность ТФ Clock/Bmal1. Эта модуляция может осуществляться через короткие контуры, как в очевидных случаях, показанных нами на рис. 3, и более сложными путями. Примером может быть положительная обратная связь между экспрессией гена *Per2* и уровнем активности ТФ Clock/Bmal1, которая, на первый взгляд, противоречит описанной выше структуре циркадного осциллятора. Путь реализации событий этой положительной обратной связи, представленных в графе, приведен на рис. 5. Рассмотрим более подробно процессы, обеспечивающие эту обратную связь:

1. Экспрессия гена *Per2* активируется ТФ Clock/Bmal1, ATF4.

2. Наработанный в результате белок Per2 образует комплекс Per2/Cry2/CK1d.

3. Гетеродимер Per2/Cry2 проникает в ядро и подавляет активность ТФ Clock/Bmal1.

4. В результате этого снижается экспрессия гена *Mir142,* служащего мишенью ТФ Clock/Bmal1.

5. miR-142-3p, продукт гена *Mir-142*, является негативным регулятором экспрессии Bmal1, поэтому снижение его уровня приведет к увеличению уровня Bmal1 и увеличению активности ТФ Clock/Bmal1.

В данном случае контур с положительной обратной связью формируется за счет взаимодействия двух коротких контуров с отрицательной

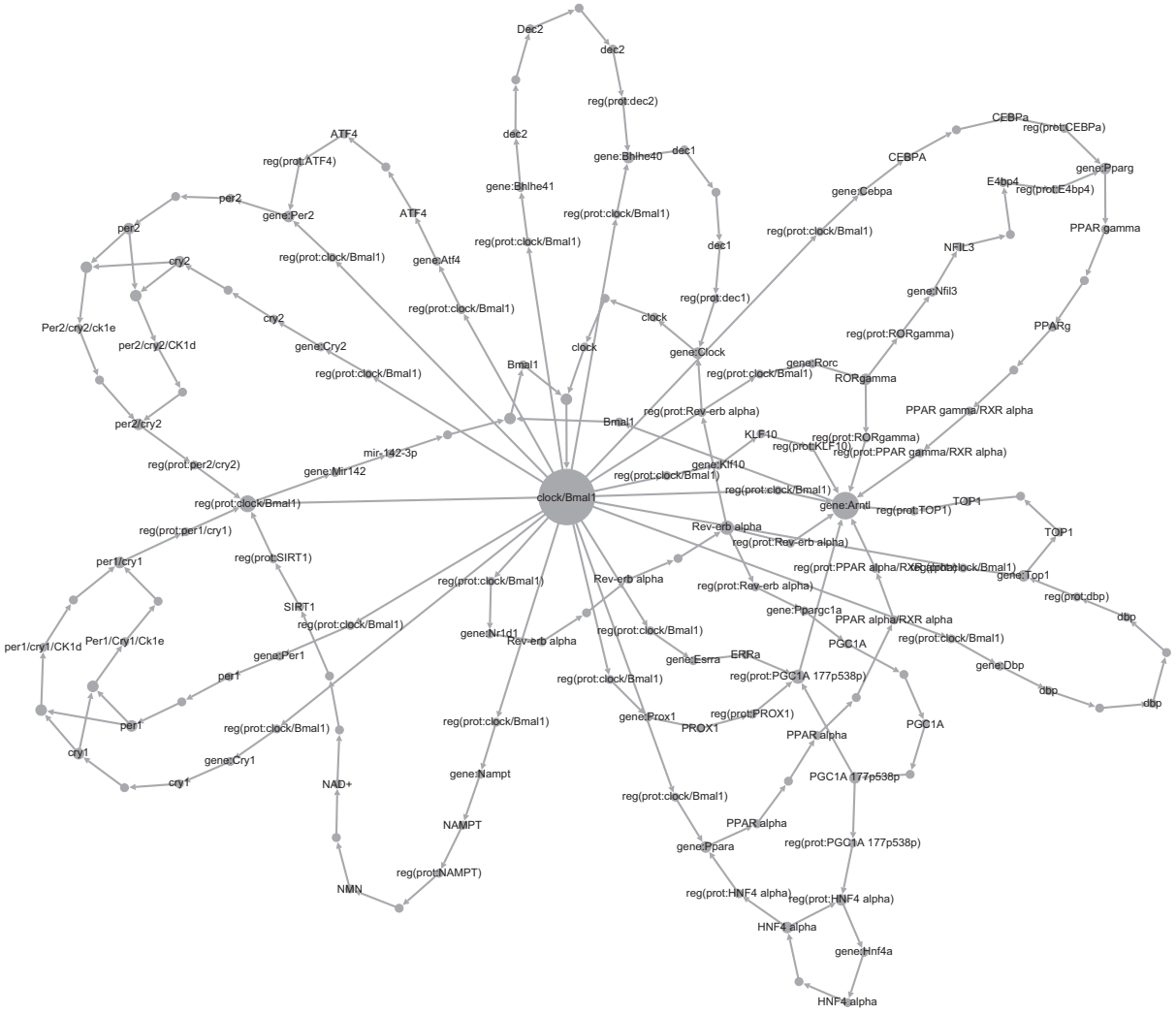


Рис. 4. Сильносвязная компонента графа генной сети Circadian Rhythm, образованная регуляторными контурами, связывающими Clock/Bmal1 с другими регуляторными элементами генной сети.

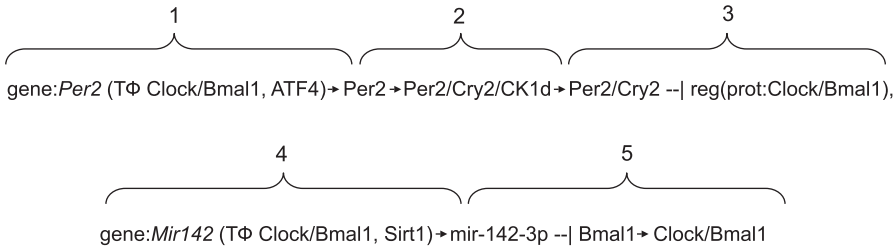


Рис. 5. Пример положительной обратной связи сильносвязной компоненты графа генной сети Circadian Rhythm.

обратной связью (см. рис. 3, *а, д*). Этот пример показывает продуктивность использованного нами подхода, позволяющего наряду с описанными выше петлями обратных связей выявить дополнительные более сложные регуляторные контуры, которые могут способствовать устойчивости работы циркадного осциллятора. Отметим, что представленные здесь регуляторные контуры и участвующие в них гены, вероятно, можно рассматривать как комплекс

компонент ядра циркадного осциллятора, в то время как группы генов, полученные нами в результате кластерного анализа вершин графа генной сети, можно представить как элементы, выполняющие отдельные функции в рамках механизма циркадных часов. Так, белки и гены, объединенные в кластерах 1, 7, 12 и 18 (табл. 3) в сумме обеспечивают пути превращения белков Per 1-2 и Cry 1-2, определяющие динамику гетеродимеров Per(1-2)/Cry(1-2), от которой зависит величина периода циркадного осциллятора. Компоненты кластеров 5 и 13 отвечают за снижение активности ТФ Clock/Bmal1 за счет белок-белковых взаимодействий. Кластер 15 – синтез NAD^+ и регуляция им активности фермента Parp1, одного из модификаторов Clock/Bmal1. Ранее отмечено, что среди прямых мишеней ТФ Clock/Bmal1 гены ТФ и корегуляторов транскрипции представляют значительную группу. И, хотя не все они образуют петли обратных связей с ядром осциллятора, наличие таких генов предполагает возможность вовлечения в циркадную ритмику экспрессии больших массивов генов за счет кассетной регуляции их транскрипции этими факторами.

ЗАКЛЮЧЕНИЕ

Представленные в работе подходы к реконструкции и анализу генных сетей на основе методов теории графов обеспечивают возможность реконструкции и анализа сложно организованных генных сетей, включая выявление их модульной организации. Применение этих подходов позволило реконструировать расширенный вариант генной сети циркадного осциллятора млекопитающих, провести анализ структуры генной сети и выделить центральную компоненту циркадного осциллятора, которая включает базовые регуляторные контуры, проходящие через ключевой элемент циркадных часов – белок Clock/Bmal1.

Использование кластерного анализа позволило выявить подсистемы, имеющие четкую биологическую интерпретацию и участвующие в функционировании циркадных часов путем взаимодействия с центральной компонентой. Такая структурная модель, включающая как центральную компоненту, так и взаимодействующие с ней функциональные подсистемы, может стать основой для построения расширенной математической модели динамики генной сети циркадного осциллятора.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке гранта РНФ № 14-24-00123.

ЛИТЕРАТУРА

Подколодная О.А. Молекулярно-генетические аспекты взаимодействия циркадных часов и метаболизма энергетических субстратов млекопитающих // Генетика. 2014. Т. 50. № 2. С. 125–137.

Albrecht U. Timing to Perfection: The Biology of Central and Peripheral Circadian Clocks // Neuron. 2012. V. 74. P. 246–260.

Ananko E.A., Podkolodny N.L., Stepanenko I.L., Podkolodnaya O.A., Rasskazov D.A., Miginsky D.S., Likhoshvai V.A., Ratushny A.V., Podkolodnaya N.N., Kolchanov NA. GeneNet in 2005 // Nucleic Acids Res. 2005. V. 33 (Database issue). P. D425–D427.

Asher G., Schibler U. Crosstalk between components of circadian and metabolic cycles in mammals // Cell Metab. 2011. V. 13. P. 125–137.

Barabasi A.-L., Oltvai Z.N. Network biology: understanding the cell's functional organization // Nat. Rev. Genet. 2004. V. 5. P. 101–113.

Blondel V.D. *et al.* Fast unfolding of communities in large networks // J. Stat. Mechanics. 2008. V. 10008. P. 1–12.

Erdős P., Rényi A. On random graphs I // Publ. Math. Debrecen. 1959. V. 6. P. 290–297.

Kim W., Li M. *et al.* Biological network motif detection and evaluation // BMC Systems Biology. 2011. V. 5. P. S5.

Koschützki D., Schreiber F. Centrality analysis methods for biological networks and their Application to gene regulatory networks // Gen. Regul. Syst. Bio. 2008. V. 2. P. 193–201.

Morf J., Rey G., Schneider K. *et al.* Cold-inducible RNA-binding protein modulates circadian gene expression posttranscriptionally // Science. 2012. V. 338. P. 379–383.

Newman M.E.J. The structure and function of complex networks // SIAM Reviews. 2003. V. 45. P. 167–256.

Reppert M., Weaver D. Molecular analysis of mammalian circadian rhythms // Ann. Rev. Phys. 2001. V. 63. P. 647–676.

Ripperger J.A., Brown S.A. Transcriptional regulation of circadian clock // Protein Reviews 2010. V. 12. P. 37–78.

Tarjan R. Enumeration of the elementary circuits of a directed graph // SIAM J. Computing. 1973. V. 2. P. 211–216.

Virshup D.M., Eide E.J., Forger D. *et al.* Reversible protein phosphorylation regulates circadian rhythms // Cold Spring Harb Symp Quant Biol. 2007. V. 72. P. 413–420.

Watts D.J., Strogatz S.H. Collective dynamics of 'small-world' networks // Nature. 1998. V. 393 (6684). P. 440–442.

Yagita K., Tamanini F., van Der Horst G.T., Okamura H. Molecular mechanisms of the biological clock in cultured fibroblasts // Science. 2001. V. 292. P. 278–281.

THE MAMMALIAN CIRCADIAN CLOCK: GENE REGULATORY NETWORK AND THEIR COMPUTER ANALYSIS

O.A. Podkolodnaya[1], N.N. Podkolodnaya[1, 2], N.L. Podkolodnyy[1, 3]

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

Summary

This paper presents the results of the reconstruction and analysis of gene regulatory network of the circadian clock in mammals. Application of graph theory methods makes it possible to analyze the structure of the gene network and identify the central component of circadian clock regulation, which includes the basic regulatory circuits passing through the key element of the circadian clock, the Clock/Bmal1 protein. Cluster analysis has revealed subsystems with clear biological interpretation, which are involved in the functioning of the circadian clock by interacting with the central component. This structural model, which includes the central component and functional subsystems that interact with the central component, can provide grounds for the construction of a mathematical model of the dynamics of the gene network regulating the circadian rhythm.

Key words: circadian clock, gene network, graph analysis methods.

УДК 577.214.626+316.452

ИССЛЕДОВАНИЕ СТРУКТУРЫ И ЭВОЛЮЦИИ СЕТЕЙ НАУЧНОГО СОАВТОРСТВА НА ОСНОВЕ АНАЛИЗА НОВОСИБИРСКИХ ПУБЛИКАЦИЙ В ОБЛАСТИ БИОЛОГИИ И МЕДИЦИНЫ

© 2014 г. **И.И. Титов[1, 2], А.А. Блинов[2]**

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

Поступила в редакцию 9 октября 2014 г. Принята к публикации 22 октября 2014 г.

Из-за разнообразия взаимодействий сообщества живой материи, от бактериальных колоний до человеческих сообществ, имманентно более сложны, нежели ансамбли частиц в неживой природе. Одним из проявлений внутри- и межгрупповых взаимодействий в социуме являются сети соавторства научных публикаций. В нашей работе рассмотрена такая сеть для новосибирского научного сообщества в области биологии и медицины. Используя базу данных PubMed, мы построили сеть и рассчитали ее статистические характеристики. Распределение организаций по научной активности оказывается распределением с толстым хвостом и подчиняется так называемому закону Парето: 83 % публикаций и 75 % авторов принадлежат примерно 20 % самых активным организациям. Сравнение сетей последних показывает, что сети вузов обладают более выраженным ядром, нежели сети научно-исследовательских институтов. Проведен анализ «демографической» структуры ныне активных авторов. Показано, что значительную долю составляют авторы с коротким публикационным стажем, а дефицит авторов наблюдается среди впервые опубликованных в 1991–1997 гг. В целом, динамика сети оказывается нестационарной с сохранением тенденции к повышению активности.

Ключевые слова: сеть соавторства, структура сети, эволюция сети, статистический анализ.

ВВЕДЕНИЕ

В наше время едва ли не самым важным условием производства научного знания является взаимодействие ученых. При непосредственном влиянии научных сотрудников друг на друга это взаимодействие реализуется в виде обмена мнениями, разделения труда и т. д., увеличивая продуктивность (Lee, Bozeman, 2005) и цитируемость (Sooryamoorthy, 2009; Gazni, Didegah, 2011), и фиксируется в виде совместного авторства публикации – явного продукта научного сотрудничества.

Данные о соавторстве позволяют формально построить упрощенную социальную сеть научных работников, в которой ее члены распространяют, интерпретируют и производят новое знание посредством социальных взаимодействий. Исследования последних 15 лет показали, что сети соавторства в разных областях науки организованы сходным образом: они кластеризованы и обладают малым диаметром (Newman, 2004).

Эти свойства не могут быть получены при случайно-равномерном размещении ребер и характерны также для других социальных сетей, таких как сети киноактеров и директоров компаний (Newman, 2003).

Особенно ярко социальные контакты научных работников должны проявляться в сообществах компактных, междисциплинарных и географически изолированных научных центров, например в Новосибирском научном

центре (ННЦ). Последнюю четверть века новосибирская, как и вся российская наука, не находилась в стабильных условиях, а пережила трансформацию общественного строя. Поэтому исследование структуры и эволюции сети соавторства Новосибирского научного центра представляет особый интерес.

МАТЕРИАЛЫ И МЕТОДЫ

Данные о научных публикациях были выделены из базы данных PubMed (состояние на август 2014 г.) фильтрацией по названию города. Всего была найдена 5 571 публикация. Отметим, что в базе PubMed присутствует лишь часть научных публикаций, а выбор PubMed из нескольких аналогичных баз данных обусловлен субъективно ее большей однородностью. Затем был составлен список организаций Новосибирска, которые присутствуют в аффилиации авторов публикаций, и добавлен ГНЦ «Вектор», который находится в п. г. т. Кольцово в непосредственной близости от Новосибирска.

Далее этот список был сокращен автоматической идентификацией вариантов названий одной и той же организации. На последнем этапе были вручную объединены те организации, которые меняли названия с сохранением юридического лица (Институт химической биологии и фундаментальной медицины Сибирского отделения Российской академии наук и др.). Окончательный список состоял из 62 организаций города Новосибирска, сотрудники которых имеют публикации в базе PubMed.

Для каждой публикации каждый из ее авторов отнесен к одной из его организаций. Если автор имел публикации с разными аффилиациями, он оказывался независимо включенным в соответствующие сети.

Далее для каждой организации список авторов сокращался автоматической идентификацией различия англоязычных написаний фамилий, а также отождествлением авторов без второго инициала. Окончательный список состоял из 8 162 авторов.

Статистический анализ сетей соавторства проведен при помощи пакета NetInference (Титов и др., 2013). Эволюция сети восстановлена при помощи определения соответствия между кластерами на соседних временных срезах.

РЕЗУЛЬТАТЫ

Сначала охарактеризуем публикационную активность в целом. Во-первых, все новосибирские институты СО РАН имеют публикации в базе PubMed, несмотря на ее специализацию в области биологии и медицины, – вероятно, вследствие присущей ННЦ интенсивности междисциплинарных контактов. Во-вторых, большое число организаций (часть из них не являются ни научными, ни учебными) представлены малым числом публикаций. В то же время более половины (51 %) публикаций принадлежат трем наиболее активным организациям. Поэтому неудивительно, что распределение числа организаций по числу публикаций выглядит как распределение с толстым хвостом (рис. 1). Интересно, что активность организаций подчиняется так называемому закону Парето: 12 (19,4 %) организациям соответствует 83 % статей и 75 % авторов. В дальнейшем мы сфокусировали свое внимание на рассмотрении именно этих 12 наиболее активных организаций: нескольких институтов СО РАН и СО РАМН, ГНЦ «Вектор» и двух вузов – НГУ и НГМУ.

Теперь обратимся к изменению публикационной активности новосибирской медико-биологической науки во времени. В целом можно отметить рост активности, однако очень неоднородный (рис. 2) – с отчетливыми пиками в 1991, 2003 и 2014 гг. Эта картина характерна как для менее активных, так и для большин-ства самых активных организаций, но наиболее ярко

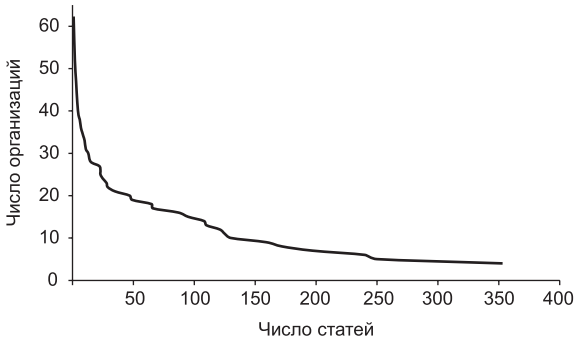


Рис. 1. Кумулятивное распределение числа новосибирских организаций по числу научных публикаций, аннотированных в базе PubMed. Функция $y(x)$ на графике показывает число организаций, которые имеют не менее x публикаций.

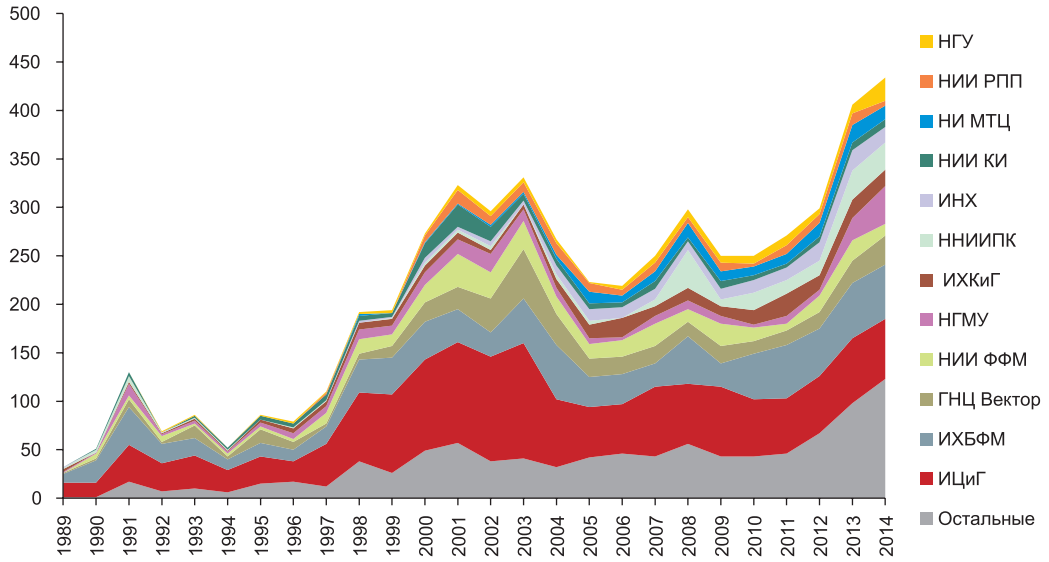


Рис. 2. Динамика числа публикаций в базе PubMed у 12 наиболее активных организаций Новосибирска. По оси абсцисс отложены года публикаций. Толщина полосы соответствующего цвета показывает число публикаций организации в заданный год. Таким образом, верхняя кривая отражает динамику числа публикаций всех новосибирских организаций. Организации представлены снизу вверх в порядке убывания полного числа публикаций в базе PubMed. Остальные 50 организаций выделены в группу «Остальные» (нижняя полоса). Напомним, что за 2014 г. данные неполные.

выражена у трех лидеров (ИЦиГ, ИХБФМ и ГНЦ «Вектор», рис. 2). Как показано ниже, яма 1990-х гг. и рост последних лет сопряжены с изменением числа научных работников, которые публикуются впервые. Исключение из общей тенденции составляет НГУ, число публикаций которого оставалось невысоким до тех пор, пока не выросло резко в 2014 г., вероятно, вследствие образования новых научных лабораторий.

В целом, оценка вкладов тех или иных факторов изменения научной активности требует отдельного исследования. Среди возможных факторов отметим изменение численности научных работников в результате внутренней и внешней эмиграции. В частности, снижение активности, наблюдаемое в середине 2000-х гг., могло быть вызвано прекращением новосибирской аффилиации эмигрантов.

Может ли нестационарный характер активности научных организаций быть связан с необычным поведением статистических характеристик авторов? Здесь под стационарностью процесса мы понимаем постоянство во времени его параметров, в том числе параметров роста. Для ответа на вопрос перейдем к статистическому описанию индивидуальной публикационной активности, в частности, проследим изменение активности авторов во времени. Сначала мы определили общее количество статей у каждого автора. В целом, распределение авторов по числу публикаций удовлетворяет степенному закону (данные не показаны), известному уже почти 90 лет (Lotka, 1926).

Далее для каждого автора, который имел публикацию в 2013–2014 гг., мы определили дату его первого появления в базе PubMed и построили распределение числа авторов по времени, которое прошло с момента первого появления в базе данных (рис. 3). Сходные распределения в виде пирамиды строят в обычных переменных для демографического анализа возрастной структуры населения. Молодыми учеными будем называть тех авторов, у которых публикационный стаж короче пяти лет. На полученной кривой можно выделить три участка: с начальным быстрым снижением, плато и яма (рис. 3).

Первый из участков, из сотрудников с публикационным стажем менее пяти лет, содержит две трети (66,4 %) из ныне активных научных работников. Этому участку можно сопоставить рост активности последних лет, где трехкратное

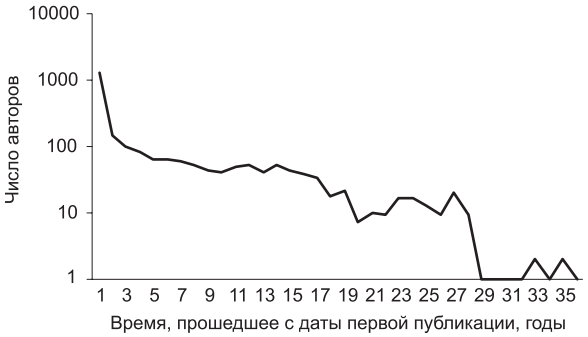


Рис. 3. Распределение числа авторов, которые опубликовали свои работы в 2013–2014 гг., по времени, прошедшему с даты их первой публикации (в полулогарифмических координатах).

увеличение числа авторов сопряжено с увеличением темпа публикаций более чем в полтора раза (см. рис. 2). Можно ли объяснить увеличение темпа публикаций ростом числа молодых ученых? Обе тенденции роста частично связаны друг с другом напрямую и частично вызваны действием третьего фактора.

Этот вывод обосновывается следующей простой оценкой. Для последнего четырехлетнего интервала времени мы определили доли публикаций, в которых присутствуют исключительно молодые или опытные научные работники.

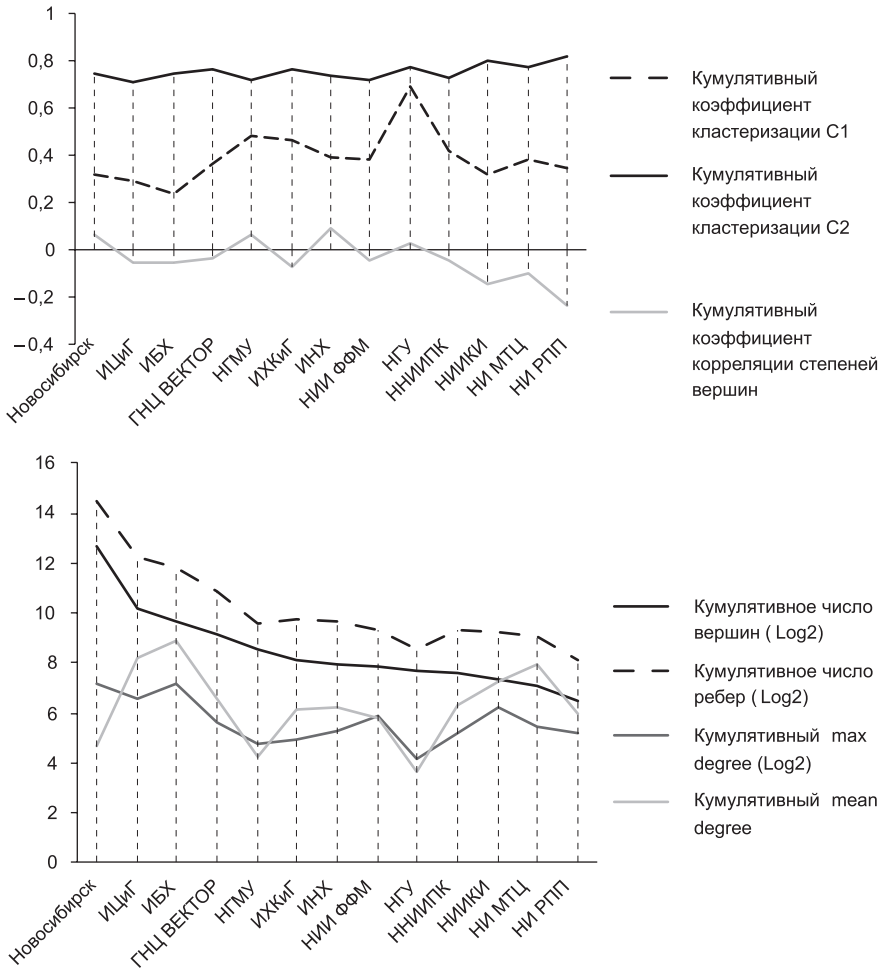


Рис. 4. Статистические характеристики сетей соавторства Новосибирска и 12 наиболее активных организаций. Некоторые характеристики стохастических сетей зависят от числа вершин в них, поэтому организации, в отличие от рис. 2, приведены в порядке уменьшения числа вершин в их сетях. Числа вершин и ребер, максимальная степень вершины показаны в полулогарифмических координатах. Для удобства общей шкалы по оси ординат приведены значения логарифмов этих характеристик.

Эти доли оказались равны 8 и 35 % соответственно, при двукратном перевесе численности первых. Тогда отношение продуктивности опытных и молодых научных работников можно оценить как $\frac{2 \cdot 35}{8} \approx 9$ раз. По той же логике продуктивность молодых научных работников при участии в работе опытных увеличивается в $\frac{100-35-8}{8} \approx 9$ раз. И наоборот, участие молодых научных работников в работе опытных повышает продуктивность работы последних в $\frac{100-35-8}{35} \approx 1{,}6$ раза.

Следует подчеркнуть крайнюю примитивность приведенных оценок, поскольку в них пренебрегается такими существенными обстоятельствами, как разное качество публикаций, индивидуальность авторов, неаддитивность и неравенство реальных вкладов соавторов, вариации численности и состава отдельных коллективов и обеих групп в целом, организация труда, сложившаяся структура научных коллективов и т. д.

Учет вышеназванных обстоятельств и накопленный к настоящему моменту объем публикаций позволят создавать точные наукометрические модели для прогноза эффективности работы научных организаций.

На втором участке число сотрудников слабо зависит от времени первой публикации вплоть до ямы между 17 и 22 годами. Это падение соответствует авторам, первые публикации которых появились в 1991–1997 гг., и может быть одним из факторов снижения общей активности, наблюдаемого в это же время (см. рис. 2).

От описания индивидуальной активности перейдем к рассмотрению взаимодействий авторов. Для этого мы построили сети соавторства, где каждому автору для каждой аффилиации соответствует своя вершина графа, а совместной публикации соответствует ребро между вершинами. Вес как вершины, так и ребра этого графа зависит от числа публикаций.

Взвешивание ребер и вершин графа может использоваться для количественной оценки социального влияния в сети, что было реализовано в нашей предыдущей работе для идентификации научных коллективов (Титов и др., 2013). Здесь же мы рассматриваем статистику построенных сетей соавторства (рис. 4).

Разнообразие значений характеристик сетей и, в частности, немонотонный характер их зависимости от числа вершин свидетельствуют об индивидуальности архитектуры рассмотренных сетей, поскольку их нельзя получить друг из друга масштабным преобразованием. Из рис. 4 видно, что во всех 12 сетях и сети новосибирского сообщества наблюдается высокая степень кластеризации.

Этот факт отмечен ранее для сетей соавторства разных областей знаний (Newman, 2004). При этом из общей выборки выделяются оба вуза: НГМУ и НГУ. Для них характерны высокие значения глобальных коэффициентов кластеризации, низкая плотность и ассортативность (взаимное притяжение вершин с близким числом входящих ребер).

Все эти свойства свидетельствуют об иной организации сетей вузов в сравнении с научными институтами: для первых характерно наличие единственного ярко выраженного ядра сети в сравнении со вторыми, более однородными и децентрализованными.

ЗАКЛЮЧЕНИЕ

Исследование сложных систем удобно сводить к изучению свойств сетей, которые моделируют эти системы. Для таких сетей часто характерны особая архитектура, богатая динамика и необычная эволюция. В этой работе мы исследуем сеть соавторства ННЦ, построенную по данным публикаций из базы PubMed. Мы показываем, что активность медико-биологического сообщества Новосибирска на протяжении последней четверти века, в целом демонстрируя значительный рост и обладая организацией типичного научного сообщества, испытавала драматические пики и провалы. Нестационарность эволюции сети проявляется во временных характеристиках авторов, в то время как структурная статистика авторов соответствует хорошо известному распределению активности.

БЛАГОДАРНОСТИ

Работа поддержана интеграционным проектом СО РАН № 21 и проектом фундаментальных исследований СО РАН VI.61.1.2.

ЛИТЕРАТУРА

Титов И.И., Блинов А.А., Рудниченко К.А., Крутов П.В., Казанцев А.Л., Куликов А.И. NETINFERENCE: набор программ для анализа структуры и динамики сетей // Вавиловский журнал генетики и селекции. 2013. Т. 17. № 4/1. С. 615–619.

Gazni A., Didegah F. Investigating different types of research collaboration and citation impact: a case study of Harvard University's publications // Scientometrics. 2011. V. 87. No. 2. P. 251–265.

Lee S., Bozeman B. The impact of research collaboration on scientific productivity // Social Studies of Science. 2005. V. 35. No. 5. P. 673–702.

Lotka A.J. The frequency distribution of scientific productivity // J. Wash. Acad. Sci. 1926. V. 16. No. 12. P. 317–324.

Newman M.E.J. The structure and functions of complex networks // SIAM review. 2003. V. 45. No. 2. P. 167–256.

Newman M.E.J. Coauthorship networks and patterns of scientific collaboration // PNAS. 2004. V. 101. No. S. 1. P. 5200–5205.

Sooryamoorthy R. Do types of collaboration change citation? Collaboration and citation patterns of South African science publications // Scientometrics. 2009. V. 81. No. 1. P. 177–193.

EXPLORING THE STRUCTURE AND EVOLUTION OF THE NOVOSIBIRSK BIOMEDICAL CO-AUTHORSHIP NETWORK

I.I. Titov[1, 2], A.A. Blinov[2]

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

Summary

The interaction diversity within the communities of living matter, from bacterial colonies to human societies, makes them inherently more complex than ensembles of particles in inanimate nature. Co-authorship networks are a particular case of intra- and inter-group social interactions. In this paper, we analyze the Novosibirsk biomedical scientific community as an example of such a network. Using the PubMed database, we have built a community network and calculated its statistics. The distribution of organizations by scientific activity has a fat tail and obeys the Pareto principle: 83% of publications and 75% of authors belong to the 20% of the most active organizations. A comparison of their networks shows that networks of the universities have a more pronounced core rather than those of research institutions. We have plotted the "demographic" structure of currently active authors and found out two facts: (1) an abundance of authors with short "publication experience" and (2) a deficit of authors whose first publication is dated back to 1991-1997. In general, the network dynamics is non-steady, and the activity tends to increase.

Keywords: co-authorship network, network structure, network evolution, statistical analysis.

УДК 573.22:57.011

L-СИСТЕМА ДЛЯ МОДЕЛИРОВАНИЯ ПЛОСКИХ ОДНОМЕРНО РАСТУЩИХ РАСТИТЕЛЬНЫХ ТКАНЕЙ

© 2014 г. **У.С. Зубаирова[1], С.К. Голушко[2], А.В. Пененко[3], 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] Федеральное государственное бюджетное учреждение науки Конструкторско-технологический институт вычислительной техники Сибирского отделения Российской академии наук, Новосибирск, Россия;
[3] Федеральное государственное бюджетное учреждение науки Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия

Поступила в редакцию 7 октября 2014 г. Принята к публикации 22 октября 2014 г.

Разработана система, которая позволяет моделировать симпластный рост ткани линейного листа и в вычислительных экспериментах генерировать паттерны клеточной структуры ткани. Для этого мы модифицировали формализм дифференциальных L-систем и в этом формализме описали динамическую модель симпластного роста с учетом биомеханики. В качестве примера приведены результаты вычислительного эксперимента для клеток, которые росли симпластно в прямоугольном листе, и клеток, которые росли свободно. Распределения размеров клеток при свободном и симпластном росте значимо различаются.

Ключевые слова: линейный лист, симпластный рост, динамические системы с динамической структурой, математическая модель, L-системы.

ВВЕДЕНИЕ

Эпидермальные слои клеток растений, как и эпителии животных, являются удобной модельной системой для изучения процессов формирования клеточных паттернов (Ryu *et al.*, 2013). Клетки в таких слоях делятся перпендикулярно поверхности слоя, поэтому мы можем легко наблюдать пространственное расположение клеток в ткани друг относительно друга, их размеры, а также процессы формирования таких структур в развитии. Для растительной ткани характерен симпластный рост (Priestley, 1930), при котором клетки не сдвигаются друг относительно друга, поэтому деление клеток – единственное, что влияет на топологию ткани.

Для изучения процессов роста растительной ткани используют модели, в которых морфодинамика обычно представлена движением вершин многоугольников в потенциале обобщенных сил, включающих механические напряжения границ клеток, их объемов и других параметров (Honda *et al.*, 2004; Hamant *et al.*, 2008; Merks *et al.*, 2011). В работах с использованием таких моделей, например, установлено, что механические напряжения в ткани влияют на потоки ауксинов (Hamant *et al.*, 2008; Nakamura *et al.*, 2012). Также в ряде исследований показано: диффузионные потоки морфогенов могут так регулировать рост и пролиферацию клеток, что при определенном соотношении параметров поддерживаются пространственные паттерны концентраций морфогенов, необходимые для функционирования ткани (Николаев и др., 2010; Chickarmane *et al.*, 2012; Yadav *et al.*, 2013).

Однако при изучении вопросов формирования пространственных паттернов, на наш взгляд, важно использовать модель механики роста с ясной биофизической интерпретацией. Желательно также, чтобы геометрия роста ткани и возникающих механических напря-

жений в ней была простой. Это позволило бы изучать принципиальные стороны процесса, не замаскированные эффектами, возникающими от сложной геометрии. В данной работе мы поставили вопрос о том, как будут изменяться размеры клеток в составе плоской одномерно растущей растительной ткани, если у каждой ее клетки есть механизм регуляции роста с механикой изменения объема за счет поступления воды (Ortega, 2010) и с упругопластическим поведением материала клеточной стенки. Для изучения этого вопроса мы создали систему для моделирования плоской одномерно растущей растительной ткани. Такая простая геометрия характерна, например, для эпидермиса линейного листа (Williams, 1974) и эпидермиса корня (Dolan, 1996). Для того чтобы моделировать процессы роста ткани, состоящей из разных типов делящихся и неделящихся клеток, в результате чего изменяется клеточная структура ткани, целесообразно использовать формализм, предназначенный для описания динамических систем с динамической структурой. Одним из таких формализмов является формализм L-систем (Lindenmayer, 1968) и его варианты (Prusinkiewicz, Lindenmayer, 1990). В данной работе мы модифицировали формализм дифференциальных L-систем (Prusinkiewicz *et al.*, 1993), а именно, учитывая специфику роста рассматриваемых тканей, разработали вариант склеенных одномерных дифференциальных L-систем. В этом формализме мы описали динамическую модель симпластного роста с учетом механики клеток и реализовали модель в пакете Mathematica 9.

1. Модель симпластного роста клеток в составе плоской одномерно растущей растительной ткани

1.1. Клеточная структура ткани

В качестве объекта для моделирования мы выбрали эпидермис с простой геометрией роста, а именно эпидермис линейного листа. Длина пластинки линейного листа во много раз превосходит ее ширину, относительно одинаковую на всем протяжении, такие листья характерны для злаков. Клеточная структура эпидермиса листа злаков представляет собой почти параллельные продольные ряды клеток в направлении от основания до кончика листа, которые формируются в процессе роста листа из меристематического слоя клеток, расположенного в основании листа (Williams, 1974). В результате поверхность линейного листа в модели можно представить как кирпичную кладку (рис. 1) из прямоугольных клеток, уложенных в продольные ряды, в которых все клетки имеют одинаковую ширину и разную длину вследствие разных скоростей роста клеток, при этом «модельный лист» тоже имеет форму прямоугольника. Ввиду такой простой топологии, несмотря на то что нас интересует поверхность листа, мы можем моделировать ткань не как двумерную, а как несколько одномерных цепочек, которые склеены между собой (симпластный рост).

Поскольку эпидермис линейного листа является плоской структурой, мы рассматривали его двумерную модель, пренебрегая изменением толщины клеток в процессе роста. Клетки в эпидермисе листа образуют параллельные продольные ряды, из чего можно заключить, что в ширину клетки растут согласованно и при росте не возникает дополнительных сил между соседними клетками в одном таком ряду. Поэтому в нашей модели мы пренебрегли изменением ширины клеток в процессе их роста. Таким образом, мы предполагаем, что все клетки имеют форму параллелепипеда с одинаковыми толщиной и шириной r, мкм, и площадью основания S_c, мкм^2, и растут только в длину l.

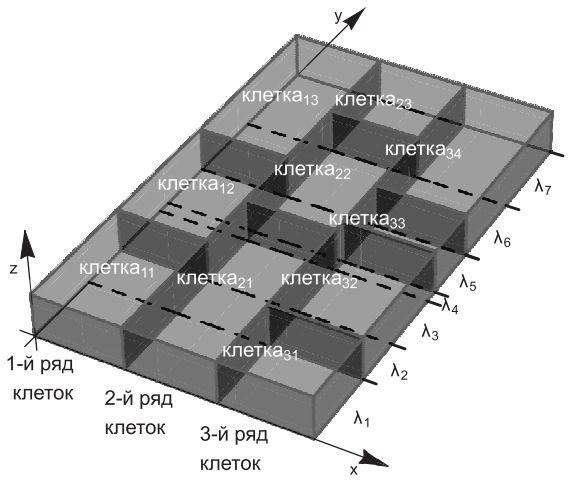


Рис. 1. Разбиение модели листовой пластинки на фрагменты (обозначены штрихами).

Итак, рассмотрим прямоугольный пласт из прямоугольных клеток, которые растут только в направлении 0*y*. Занумеруем продольные ряды клеток в направлении 0*x* индексом *n* (*n* = 1, ..., *N*), а в каждом продольном ряду в направлении *0y* – индексом *m* (*m* = 1, ..., M_n). Для моделирования симпластного роста клеток в ткани каждую клетку разобьем на фрагменты следующим образом: границу каждой клетки, параллельную 0*x,* продолжим через весь клеточный пласт, в результате в направлении 0*y* лист будет разбит на фрагменты, которые занумеруем индексом *k* (*k* = 1, ..., *K*), а длину фрагмента обозначим через λ_k. На рисунке 1 показан прямоугольный лист, состоящий из трех клеточных продольных рядов. Заметим, что клетки включают в себя разные фрагменты, так, например, первая клетка в первом продольном ряду состоит из фрагментов λ_1, λ_2, первая клетка во втором продольном ряду состоит из фрагментов λ_1, λ_2, λ_3, первая клетка в третьем продольном ряду состоит только из фрагмента λ_1.

1.2. Модель механики симпластного роста клеток в составе ткани

Существенной чертой нашей модели является неоднородный рост клеточной стенки в пределах одной клетки, возникающий в результате склеивания клеток, которые расположены в параллельный рядах и растут с разными скоростями. Реальная скорость роста в составе ткани длины *m*-й клетки в *n*-м продольном ряду есть сумма скоростей роста принадлежащих ей фрагментов:

$$\frac{dl_{nm}}{dt} = \sum_{\lambda_k \in l_{nm}} \frac{d\lambda_k}{dt}.$$

С механической точки зрения, фрагмент λ_k представляет собой склеенные фрагменты клеток, на каждый из которых действует своя сила, определяемая тургорным давлением в соответствующей клетке. Следовательно, на весь фрагмент действует сумма этих сил. Если для упрощения модели предположим, что параметры, определяющие механическое поведение, одинаковы для всех клеток, то суммарные значения сечений для всего фрагмента будут равны соответствующим значениям для одной клетки, умноженным на количество клеток. Из этого следует, что изменение длины общего для всех клеток фрагмента определяется формулой:

$$\frac{d\lambda_k}{dt} = \frac{\lambda_k}{N} \sum_{n=1}^{N} \left(\frac{dl_{nm}}{l_{nm}\, dt}\right)_f \ (\forall n\, m\colon \lambda_k \in l_{nm}),$$

где λ_k – *k*-й фрагмент, l_{nm} – длина *m*-й клетки в *n*-м ряду, $\left(\frac{dl_{nm}}{l_{nm}\, dt}\right)_f$ – удельная скорость изменения длины клетки в данный момент времени, если бы у нее не было механических связей с соседними клетками. Эта скорость определяется из модели роста одиночных клеток.

Мы рассматриваем клетку в форме параллелепипеда, растущего только в длину. Для описания механики изменения объема каждой отдельной клетки мы использовали модель Ортеги (Ortega, 2010), в которой рост клетки имеет две составляющих: биосинтез сухой биомассы и увеличение объема за счет поступления воды в клетку. Движущей силой для потока воды является водный потенциал клетки относительно окружающей среды, который равен разности между внутриклеточным осмотическим и тургорным давлением.

В качестве единицы измерения сухой биомассы клетки в модели мы ввели изоосмотическую длину клетки (l_i), рост которой описан функцией от времени (аппроксимация по экспериментальной кривой роста дрожжевой клетки (Bryan *et al.*, 2010)):

$$l_i(t) = l_i(t_0) e^{\upsilon (t - t_0)^2},$$

где нормировочный коэффициент υ и показатель степени подобраны из условия, что изоосмотическая длина клетки вырастает от 5 до 10 мкм примерно за 25 ч.

Скорость изменения релаксированной длины клетки в нашей модели определяется формулой:

$$\frac{dl_r}{dt} = \begin{cases} 0, \text{если } \frac{S_w}{S_c} E\, \varepsilon \le 3 \text{ бар} \\ \eta\, l_i\, \varepsilon^2, \text{если } \frac{S_w}{S_c} E\, \varepsilon > 3 \text{ бар}, \end{cases}$$

где η – коэффициент с размерностью обратного времени, $\varepsilon = (l - l_r)/l_r$. Строя такую функцию, мы предполагали, что клетка может выделять на строительство своей стенки часть имеющихся ресурсов (l_i) в зависимости от деформации по-

следней (ε). Таким образом, переменными состояния модели являются реальная (l), релаксированная (l_r) и изоосмотическая (l_i) длины клетки. В этих переменных уравнение для изменения размеров клетки записывается следующим образом:

$$\frac{dl}{l\,dt} = \frac{2\,L\,(S_c + r\,l)}{l}\left(\frac{\gamma(l_i - l)}{l} - \frac{S_w}{S_c}E\,\frac{l_r - l}{l_r}\right),$$

где S_w – площадь сечения клеточной стенки, S_c – площадь сечения клетки перпендикулярно ее длине, r – ширина клетки, L – коэффициент проводимости клеточных барьеров для воды, E – модуль Юнга материала клеточной стенки. В данной работе мы ограничиваемся гипотезой, что клетка в ткани растет автономно, т. е. единственное ограничение – это механические взаимодействия между клетками вследствие того, что они склеены между собой. В таком случае мы можем предположить, что рост изоосмотической и релаксированной длин клеток, когда они растут в составе ткани, регулируется клеткой так же, как при росте отдельной клетки.

2. Представление модели симпластного роста ткани в формализме склеенных L-систем и ее программная реализация

2.1. Формализм L-систем

Процесс роста и деления клеток в одном продольном ряду удобно моделировать с использованием формализма L-систем (Николаев и др., 2010). L-системы (Lindenmayer systems) – теоретические модели развития на основе формальных языков, предложенные А. Линденмайером (Lindenmayer, 1968) для описания роста одномерных клеточных ансамблей. В формализме L-систем организм представлен как упорядоченная структура из дискретных единиц, называемых модулями. Каждый модуль обозначен символом (буквой 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. Самым простым классом L-систем являются L-системы, которые представляют собой упорядоченную тройку, состоящую из алфавита, непустого слова, называемого аксиомой, и конечного множества правил переписывания (Prusinkiewicz, Lindenmayer, 1990). Этот формализм расширяют параметризованные L-системы, работающие на параметризованных словах, с каждой буквой которых связан вектор параметров. Одну из разновидностей параметризованных L-систем составляют дифференциальные L-системы (dL-системы) (Prusinkiewicz *et al.*, 1993). В таких L-системах вместо дискретных шагов вывода введено непрерывное течение времени, а изменение параметров подчинено системе дифференциальных уравнений. Подробнее формализм L-систем рассмотрен нами ранее (Зубаирова и др., 2012).

2.2. Склеенные одномерные дифференциальные L-системы

С учетом специфики геометрии (параллельные продольные ряды клеток) для моделирования ткани из растущих и делящихся клеток в данной работе мы модифицировали формализм dL-систем. Динамику клеточной структуры линейного листа, состоящего из N параллельных продольных рядов клеток, мы моделировали с помощью N одномерных dL-систем, каждая из которых моделирует динамику структуры отдельного продольного ряда.

Алфавит каждой dL-системы состоит из одной буквы, снабженной следующими переменными состояния: l_i, l_r, l – изоосмотическая, релаксированная и реальная длины клетки, l_{0i} – начальная изоосмотическая длина клетки, t_0 – момент времени, когда клетка появилась. Для того чтобы обеспечить согласованную работу dL-систем, моделирующих динамику клеточной структуры отдельных продольных рядов клеток, мы ввели еще одну одномерную dL-систему, которая моделирует динамику структуры фрагментов клеток. Алфавит этой dL-системы также состоит из одной буквы, а параметрами являются длина фрагмента λ_k и вектор длины N, i-й компонентой которого является номер клетки, содержащей данный фрагмент λ_k, в i-м продольном клеточном ряду. Динамика переменных состояния l_i, l_r, l и λ_k оп-

ределяется общей системой дифференциальных уравнений, составляющих модель симпластного роста клеток в составе ткани:

$$l_{i_{nm}}(t) = l_{i_{nm}}(t_0)e^{\nu(t-t_0)^2},$$

$$\frac{dl_{r_{nm}}}{dt} = \begin{cases} 0, \text{если } \frac{S_w}{S_c}E\,\varepsilon \le 3 \text{ бар} \\ \eta\, l_{i_{nm}}\,\varepsilon^2, \text{если } \frac{S_w}{S_c}E\,\varepsilon > 3 \text{ бар} \end{cases},$$

$$\frac{d\lambda_k}{dt} = \frac{\lambda_k}{N}\sum_{n=1}^{N}\frac{2\,L\,(S_c + r\,l_{nm})}{l_{nm}}\left(\frac{\gamma(l_{i_{nm}} - l_{nm})}{l_{nm}} - \right.$$

$$\left. - \frac{S_w}{S_c}E\,\frac{l_{r_{nm}} - l_{nm}}{l_{r_{nm}}}\right)(\forall n\, m\colon \lambda_k \in l_{nm}),$$

$$\frac{dl_{nm}}{dt} = \sum_{\lambda_k \in l_{nm}}\frac{d\lambda_k}{dt},$$

где k – номер фрагмента клетки, n – номер продольного ряда клеток, а m – номер клетки в нем, остальные коэффициенты и параметры см. в разделе 1.2. Значения параметров: l_0 = 5 мкм, r = 2 мкм, S_c = 4 мкм², S_w = 0,8 мкм² (Williams, 1974), L = 0,2 (мкм · ч · бар)$^{-1}$ (адаптирован из условия согласования модели и соответствует величине, приведенной П.С. Нобелем (Nobel, 2005)), γ = 750 бар (Там же), E = 375 бар (Gibson, 2012), υ = 1/900 ч$^{-1}$ (подобран из условия, что изоосмотическая длина клетки вырастает от 5 до 10 мкм за 25 ч – аппроксимация по экспериментальной кривой роста дрожжевой клетки из статьи (Bryan *et al.*, 2010)), η = 0,05 ч$^{-1}$, σ_α = 0,05 (Jönsson *et al.*, 2005; Smith *et al.*, 2006).

Правило переписывания общее для всех N + + 1 dL-систем и срабатывает в момент деления какой-либо клетки, когда ее изоосмотическая длина достигает критического значения. В момент деления клетки ее параметры переписываются следующим образом: (1) все длины делятся в отношении α:(1 − α) (в данной работе обсуждаются результаты, полученные в случае, когда коэффициент υ считается нормально распределенной случайной величиной со средним значением μ_α и среднеквадратическим отклонением σ_α, с дополнительным условием 0,1 < α < 0,9); (2) начальные изоосмотические длины дочерних клеток l_{0i} получают значения $\alpha\, l_i$ и (1 – α)l_i; (3) параметры t_0 получают значения момента времени деления клетки; (4) в соответствии с тем, как разделилась реальная длина клетки, находят фрагмент λ_k и коэффициент его деления; (5) вектор номеров клеток разделившегося фрагмента λ_k переписывается в соответствии с новыми номерами клеток. Такой механизм роста ткани, когда клетки растут с постоянной скоростью, а затем, когда их размер достигает определенного порога, асимметрично делятся, неоднократно применялся в моделях (Jönsson *et al.*, 2005; Smith *et al.*, 2006; Николаев и др., 2010). Аксиому dL-системы составляет ряд инициальных клеток (ориентирован в направлении оси 0x), которые начинают расти в направлении, перпендикулярном оси инициального слоя, и, пролиферируя, формируют параллельные продольные ряды клеток, так что каждый продольный ряд клеток растет из своей инициали. Если клетки в одном продольном ряду растут независимо от клеток в соседних продольных рядах, имеем случай свободно растущих клеток. Если клетки в соседних продольных рядах растут, оказывая механическое воздействие на соседей вследствие того, что их стенки склеены между собой, – случай симпластного роста. Начальные значения для изоосмотической, релаксированной и реальной длин клетки равны l_0. Модель реализована в пакете Mathematica 9, на рис. 2 приведена блок-схема программы. Струк-

Рис. 2. Блок-схема программы, реализующей симпластный рост ткани.

тура данных представляет два списка. Первый описывает двумерную ткань, его элементами являются продольные ряды клеток, которые, в свою очередь, также представлены списками клеток. Второй содержит фрагменты клеток. Состояние системы на каждом временном шаге записывается в отдельный список, который отражает все изменения, происходившие со структурой листа во время вычислительного эксперимента, и позволяет визуализировать результаты.

2.3. Результаты вычислительных экспериментов

Разработанная программа позволяет моделировать симпластный рост ткани линейного листа и в вычислительных экспериментах генерировать паттерны клеточной структуры. На рис. 3 мы видим распределение клеток в модельном образце ткани, полученном в вычислительном эксперименте. Показан пример визуализации размеров клеток и их возраста, т. е. их состояния в фазе клеточного цикла.

На рис. 4 представлен результат вычисления наблюдаемой и изоосмотической длин клеток в ходе их свободного роста (черные точки) в составе ткани (серые). Из графика ясно, что простая модель управления ростом релаксированной длины клетки при симпластном росте приводит к отклонению реальной длины от изоосмотической, в то время как рассеяние точек для свободного роста формирует биссектрису угла. Это можно интерпретировать следующим образом. Ниже биссектрисы расположены клетки, сжатые по сравнению с их реальной длиной при свободном росте, а выше ее – растянутые. Поэтому регуляторный механизм роста клеточной стенки хорошо работает при свободном росте клетки, однако при ее симпластном росте в составе ткани реальная длина клетки и ее динамика отличаются от таковых для свободной клетки. Это происходит из-за механического влияния склеенных с ней клеток из других продольных рядов, и входной сигнал для механизма регуляции получает «неправильное» значение, что приводит к неадекватной регуляции. Поэтому распределения наблюдаемых длин клеток при симпластном и свободным росте значительно различаются.

ЗАКЛЮЧЕНИЕ

В работе представлена система, которая позволяет моделировать симпластный рост ткани линейного листа и в вычислительных экспериментах генерировать паттерны клеточной структуры ткани. В вычислительном алгоритме реализована математическая модель механики одномерного симпластного роста растительных тканей. Для описания динамики клеточной структуры ткани мы модифицировали формализм дифференциальных L-систем. Имплементация вычислительного алгоритма в пакете Mathematica 9 позволяет естественным образом применить стиль функционального программирования при моделировании динамики клеточной структуры ткани в процессе ее роста. В приведенных результатах моделирования роста линейного листа показано, что симпластный рост предъявляет совершенно другие требования к регуляции роста расти-

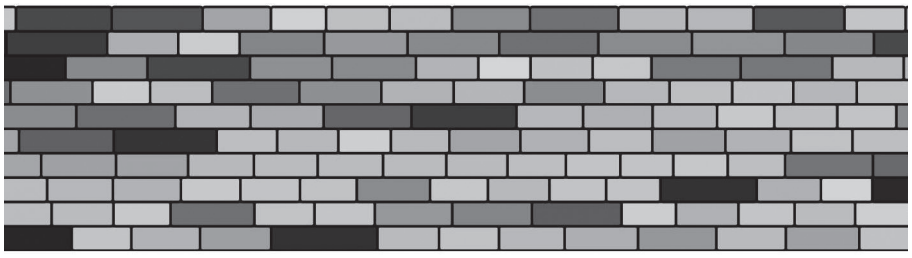


Рис. 3. Фрагмент клеточной структуры линейного листа, полученный в результате вычислительного эксперимента. Оттенкам серого соответствует стадия клеточного цикла клетки: светлые клетки – только поделившиеся, темные – готовые к делению.

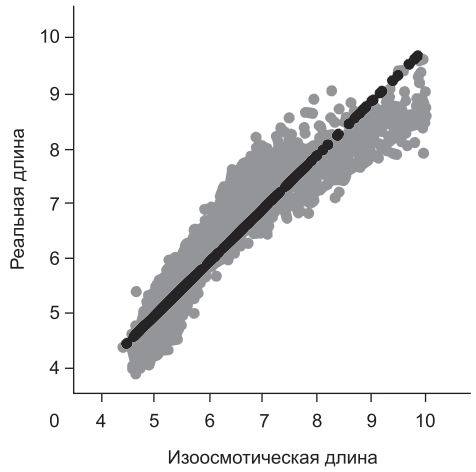


Рис. 4. Рассеяние точек, представляющих клетки, в координатных осях (изоосмотическая/реальная длина). Серые точки обозначают клетки, растущие симпластно, черные – растущие свободно.

тельной клетки по сравнению со свободным ростом. Помимо модуля расчета механики роста программа содержит модули для расчетов переноса морфогенов по ткани и управления дифференциальной активностью генов. Кроме того система моделирования предусматривает возможность изменять функции роста клеток и механизмы регуляции. Все это позволяет легко модифицировать модель роста ткани и планировать соответствующие вычислительные эксперименты для изучения особенностей регуляции и режимов функционирования клеток при их росте в составе ткани.

БЛАГОДАРНОСТИ

Авторы выражают благодарность Д.А. Афонникову и Н.Л. Подколодному за ценные советы при обсуждении рукописи. Работа выполнена при частичной финансовой поддержке гранта РНФ № 14-14-00734 «Изучение молекулярных механизмов развития органов растений методами системной биологии».

ЛИТЕРАТУРА

Зубаирова У.С., Пененко А.В., Николаев С.В. Моделирование роста и развития растительных тканей в формализме L-систем // Вавиловский журнал генетики и селекции. 2012. Т. 16. № 4/1. С. 816–824.

Николаев С.В., Зубаирова У.С., Фадеев С.И., Мйолснесс Э., Колчанов Н.А. Исследование одномерной модели регуляции размеров возобновительной зоны в биологической ткани с учетом деления клеток // СибЖИМ. 2010. Т. 13. Вып. 4 (44). С. 70–82.

Bryan A.K., Goranov A., Amon A., Manalisa S.R. Measurement of mass, density, and volume during the cell cycle of yeast // PNAS. 2010. V. 107. P. 999–1004.

Chickarmane V.S., Gordon S.P., Tarr P.T. *et al.* Cytokinin signaling as a positional cue for patterning the apical-basal axis of the growing Arabidopsis shoot meristem // Proc. Natl Acad. Sci. USA. 2012. V. 109 (10). P. 4002–4007.

Dolan L. Pattern in the Root Epidermis: An Interplay of Diffusible Signals and Cellular Geometry // Annals Botany 1996. V. 77. P. 547–553.

Gibson L.J. The hierarchical structure and mechanics of plant materials // J. Royal Society Interface. 2012. published online.

Hamant O., Heisler M., Jönsson H. *et al.* Developmental patterning by mechanical signals in Arabidopsis // Science. 2008. Dec. 12. V. 322 (5908). P. 1650–1655.

Honda H., Tanemura M., Nagai T. A three-dimensional vertex dynamics cell model of space-filling polyhedra simulating cell behavior in a cell aggregate // Journal Theoretical Biology. 2004. V. 226 (4). P. 439–453.

Jönsson H., Heisler M. G., Shapiro B. E. *et al.* An auxin-driven polarized transport model for phyllotaxis // PNAS. 2005. V. 103. P. 1633–1638.

Lindenmayer A. Mathematical models for cellular interaction in development // J. Theor. Biology. 1968. V. 18. P. 280–315.

Merks R., Guravage M., Inze D., Beemster G. Virtual Leaf: An Open-Source Framework for Cell-Based Modeling of Plant Tissue Growth and Development // Plant Physiol. 2011. V. 155 (2). P. 656–666.

Nakamura M., Kiefer C.S., Grebe M. Planar polarity, tissue polarity and planar morphogenesis in plants // Curr. Opin Plant Biol. 2012. V. 15 (6). P. 593–600.

Nobel P.S. Physicochemical and Environmental Plant Physiology. Amsterdam: Elsevier Academic Press, 2005.

Ortega J.K. Plant Cell Growth in Tissue // Plant Physiology. 2010. V. 154. P. 1244–1253.

Priestley J. Studies in the physiology of cambial activity // New Physiology. 1930. V. 29. P. 96–140.

Prusinkiewicz P., Lindenmayer A. The algorithmic beauty of plants. NewYork: Springer, 1990.

Prusinkiewicz P., Hammel M., Mjolsness E. Animation of plant development // Proc. SIGGRAPH 93. Anaheim, California. Ann. Conference Series. 1993. P. 351−360.

Ryu K.H., Zheng X., Huang L., Schiefelbein J. Computational modeling of epidermal cell fate determination systems // Current Opinion Plant Biology. 2013. V. 16 (1). P. 5–10.

Smith R., Guyomarc'h S. *et al.* A plausible model of phyllotaxis // Proc. Natl Acad. Sci. 2006. V. 103. P. 1301–1306.

Williams R.F. The Shoot Apex and Leaf Growth: A Study in Quantitative Biology. London; New York: Cambridge University Press, 1974.

Yadav R.K., Perales M., Gruel J. *et al.* Plant stem cell maintenance involves direct transcriptional repression of differentiation program // Mol. Syst. Biol. 2013. V. 9. P. 654.

AN L-SYSTEM FOR MODELING OF UNIDIMENSIONALLY GROWING FLAT PLANT TISSUES

U.S. Zubairova[1], S.K. Golushko[2], A.V. Penenko[3], S.V. Nikolaev[1]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: ulyanochka@bionet.nsc.ru;
[2] Design and Technology Institute of Digital Techniques SB RAS, Novosibirsk, Russia;
[3] Institute of Computational Mathematics and Mathematical Geophysics SB RAS,
Novosibirsk, Russia

Summary

In this work, a mathematical model and its implementation are proposed for computational simulation of one-dimensional symplastic growth of tissues. We modified the formal grammar of differential L-systems, and in this grammar, we described a dynamic model of symplastic growth with regard to its biomechanics. The results of the simulation of linear leaf blade growth are compared with those for a free-growing cell population. It is shown that in the model proposed symplastic growth causes a greater deviation of the actual cell length from its isosmotic length than in freely growing cells.

Key words: linear leaf, symplastic growth, dynamical systems with dynamic structure, mathematical model, L-systems.

УДК 577.175.12: 581.43:581.824.22

О РАСПРЕДЕЛЕНИЯХ КОНЦЕНТРАЦИЙ АУКСИНА В КЛЕТКАХ ГОРИЗОНТАЛЬНОГО СЛОЯ КОРНЯ

© 2014 г. **Е.С. Новоселова[1], В.В. Миронова[1, 2], Т.М. Хлебодарова[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: likho@bionet.nsc.ru;
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

Поступила в редакцию 29 августа 2014 г. Принята к публикации 22 октября 2014 г.

У растений различают ди-, три-, тетра-, пента- или полиархную структуру центрального цилиндра корня. Вид симметрии отражает характерное взаимное расположение пучков сосудистых тканей флоэмы и ксилемы на поперечном срезе корня. Механизмы формирования различных типов симметрий в структуре центрального цилиндра остаются недостаточно исследованными. Предполагается, что процесс дифференцировки запускается и контролируется фитогормоном ауксином, который выступает в роли морфогена (Sachs, 1969). В работе представлена модель, описывающая транспорт ауксина через одноклеточный слой клеток поперечного среза корня. Изучены стационарные распределения концентраций гормона ауксина, которые могут устанавливаться в поперечном слое клеток. Показано, что нелинейные процессы регуляции транспорта ауксина способны обеспечить существование неравномерных распределений его концентраций, несущих целевую морфогенетическую информацию о диархной структуре цилиндра корня. Однако целевые морфогенетические поля всегда сосуществуют с равномерными распределениями, в которых морфогенетическая информация отсутствует. Полученные результаты свидетельствуют в пользу гипотезы о том, что одним из механизмов формирования целевого распределения концентрации ауксина в клетках поперечного слоя корня может быть неравномерный поток ауксина соответствующей конфигурации из побега в корень.

Ключевые слова: моделирование, морфоген, ауксин, корень, дифференцировка сосудистых тканей растений, флоэма, ксилема.

ВВЕДЕНИЕ

В растениях транспорт воды, питательных и минеральных веществ осуществляется по специализированным проводящим (сосудистым) тканям, флоэме и ксилеме. В меристематической зоне корня эти ткани слабо специализированы и называются протофлоэмой и протоксилемой. Протоксилема и протофлоэма – это два типа непрерывных рядов клеток, активно транспортирующих гормон ауксин вдоль корня в направлении от побега к его кончику.

Гормон ауксин является морфогеном (Sachs, 1969; Benková *et al*., 2009), и его неравномерное распределение в тканях задает позиционную информацию, необходимую для специализации тканей (Petrášek, Friml, 2009). В формирование этого распределения включены биосинтез ауксина и его транспорт, пассивный (диффузия) и активный (полярный). Полярный транспорт осуществляется белками семейства PIN и AUX/LAX, расположенными асимметрично на базальной и апикальной сторонах мембран клеток соответственно (Swarup *et al.*, 2001; Benková *et al*., 2003; Petrášek, Friml, 2009). Первые осуществляют отток ауксина из клетки (Benková *et al*., 2003), а вторые его приток внутрь клетки (Swarup *et al*., 2001). Их экспрессия регулируется ауксином (Vieten *et al*., 2005). На поперечном срезе центрального цилиндра корня относительное расположение тяжей ксилемы и флоэмы

формирует характерный сосудистый рисунок, который в зависимости от количества лучей ксилемы обладает ди-, три-, тетра-, пента- или полиархным типом симметрии (рис. 1).

Механизмы формирования симметрий в расположении ксилемы и флоэмы в горизонтальном слое корня растений исследованы слабо. В ряде теоретических работ (Mitchison *et al*., 1980, 1981; Merks *et al*., 2007; Bayer *et al*., 2009) исследованы механизмы специализации тяжей клеток при формировании проводящей ткани растения, однако механизмы относительного позиционирования параллельных сосудов в ткани в них не рассматривались.

В то же время существуют данные, которые позволяют предположить, что неравномерное распределение ауксина в инициалях протоксилемы и протофлоэмы может быть частью механизма формирования характерного сосудистого рисунка на поперечном срезе корня. Так, некоторые экспериментальные данные указывают на то, что содержание ауксина повышено в инициалях протоксилемы, но не в других клетках поперечного среза меристематической зоны корня арабидопсиса (Bishopp *et al*., 2011a). В работе Ibañes с соавт. (2009) показана роль асимметричной локализации белков PIN1, осуществляющих активный транспорт ауксина из клетки, в формировании полиархной структуры побега, а в работе Muraro и его коллег (2014) продемонстрирована возможность формирования диархной структуры центрального цилиндра корня под управлением минимальной генной сети, контролируемой ауксином, цитокинином и микроРНК.

Ранее нами была разработана модель ауксин-регулируемой дифференцировки клеток сосудистой системы корня (Novoselova *et al*., 2013), которая является развитием наших теоретических исследований механизмов распределения ауксина в корне растений (Лихошвай *и др*., 2007; Mironova *et al*., 2010; 2012). Полученные нами результаты позволили связать различные косвенные данные о распределениях ауксина в единую систему и объяснить закономерности его распределения в протососудистых тканях корня, которые определяли дифференцировку клеток меристематической зоны корня в направлении ксилемы или флоэмы вдоль оси центрального цилиндра корня. Естественно было предположить, что архитектура строения центрального цилиндра корня, с характерной

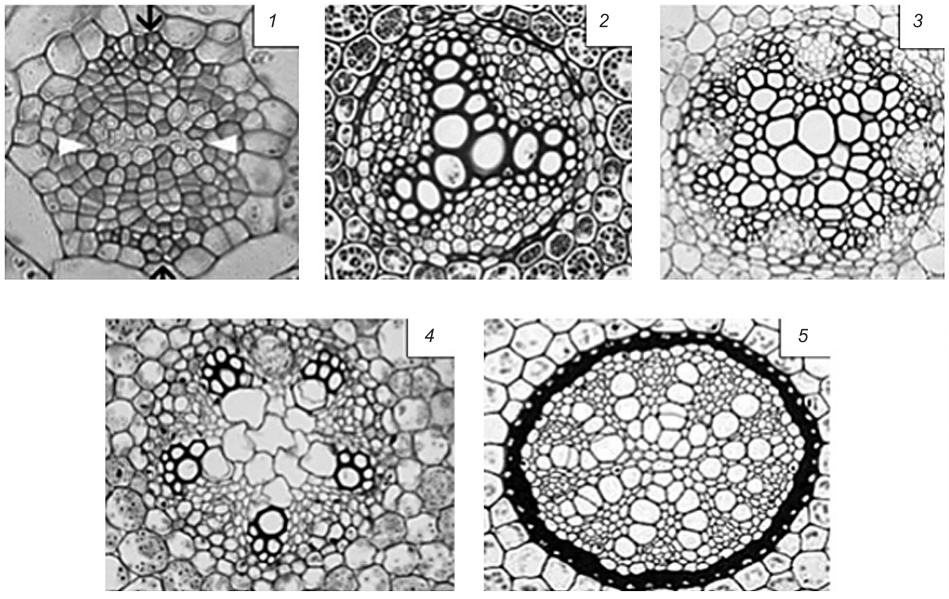


Рис. 1. Разные типы симметрий в расположениях групп первичной ксилемы на поперечных срезах корня различных растений.

1 – диархный, *2* – триархный, *3* – тетрархный, *4* – пентархный и *5* – полиархный.
Фотографии взяты с сайтов http://botit.botany.wisc.edu/Resources/Botany/Root/Ranunculus/Mature/ и http://www.bio.miami.edu/dana/226/226F09_9print.html и находятся в свободном доступе.

локализацией флоэмы и ксилемы на поперечном срезе, является следствием тех же процессов, которые определяют дифференцировку этих тканей вдоль оси центрального цилиндра. Из анализа математических моделей транспорта ауксина в инициальных клетках протофлоэмы и протоксилемы корня (Novoselova *et al*., 2013) также следовало, что в инициали протоксилемы поступает больший поток ауксина из вышерасположенной сосудистой ткани, чем в клетки протофлоэмы.

В целом, эти данные позволяли предположить, что возникновение различных видов симметрий в структуре корня может быть следствием неравномерного распределения в горизонтальном слое инициальных сосудистых клеток корня ауксина, поступающего из верхней части проростка и предопределяющего развитие клеток в направлении ксилемы или флоэмы. Причем те клетки, в которых накапливается большее количество ауксина, предетерминируются по ксилемному пути, а те, в которые поступает меньше ауксина, – по флоэмному. Но за счет каких процессов это достигается? Существующие экспериментальные данные не позволяют однозначно ответить на этот вопрос и указывают на возможность существования разных механизмов, участвующих в формировании симметрий (Ibañes *et al*., 2009; Bishopp *et al*., 2011a, b; Muraro *et al*., 2011, 2014).

Например, у двудольных растений одной из возможностей формирования симметрий может быть структурирование потока ауксина в проростке, на уровне формирования семядолей. Косвенно об этом свидетельствует тот факт, что ксилемные элементы в корне дифференцируются, как правило, в плоскости формирования семядолей (Bauby *et al*., 2007) и нарушение их формирования дезорганизует структуру центрального цилиндра корня (Help *et al.*, 2011).

Но нельзя исключить, что неравномерное распределение ауксина может достигаться за счет механизма, который формируется внутри плоского слоя клеток корня за счет нелинейного взаимодействия процессов вертикального активного транспорта ауксина с процессами его поперечной диффузии. Для проверки этой гипотезы нами была разработана математическая модель распределения ауксина в поперечном одноклеточном слое клеток, расположенном в центральном цилиндре корня на уровне меристематической зоны, где происходит предетерминация клеток сосудистой системы.

Результаты анализа показывают, что нелинейные процессы транспорта, протекающие внутри моделируемого ансамбля, обеспечивают существование в нем неравномерных распределений концентраций ауксина, несущих целевую морфогенетическую информацию, например, о диархной структуре цилиндра корня. Однако целевые морфогенетические поля всегда сосуществуют с равномерными распределениями, в которых морфогенетическая информация отсутствует. Сделан вывод, что для формирования необходимого морфогенетического поля в плоском поперечном ансамбле клеток необходимо использовать внешние факторы. Одним из них может стать неравномерность потока ауксина, которая формируется в побеге и затем поступает в плоский ансамбль поперечного среза корня.

МОДЕЛЬ

Модель описывает процессы поступления ауксина из побега в плоский слой клеток, имитирующий одноклеточный слой, лежащий в поперечном срезе корня, и перераспределение ауксина в этом слое за счет протекания в клетках процессов диффузии, активного транспорта и деградации/диссипации. Считаем, что клетки являются неспециализированными, т. е. в них протекают одинаковые процессы. Каждую клетку в поперечном срезе в первом приближении рассматриваем в виде многогранника. Каждая боковая грань (участок грани) клетки либо контактирует с боковой гранью (участком грани) другой клетки, либо является внешней границей. Также различаем верхнюю грань стенки клетки, обращенную к побегу, и нижнюю, обращенную к кончику корня. Для ранней стадии развития растения считаем, что в клетках слоя нет синтеза ауксина *de novo*, т. е. в каждую клетку слоя ауксин поступает только из побега через верхнюю грань. В каждой клетке ауксин может безвозвратно деградировать/диссипировать, также ауксин может уходить из нее через боковые внутренние грани в другие клетки. Ауксин может активно транспортироваться из клетки через нижнюю грань в нижние слои корня. В поперечном направлении активный

транспорт не рассматриваем. Модель имеет следующий вид:

$$\frac{dx_i}{dt} = \alpha_i + k_{pt}\sum_{\substack{j=1\\ j\neq i}}^{n} \omega_{i,j}\left(x_j - x_i\right) - \left(k_{at}P\left(x_i\right) + k_d\right)x_i,$$

$$\omega_{i,j} = \begin{cases} 1, j \in N_i \\ 0, j \notin N_i \end{cases}, i = 1,...,n. \qquad (1)$$

Переменные системы (1) имеют следующий смысл: n – общее количество клеток в слое, x_i – концентрация ауксина в i-й клетке, N_i – множество номеров клеток, обменивающихся с i-й клеткой в результате пассивной диффузии, $P(x)$ – функция, описывающая скорость активного транспорта ауксина из клетки в направлении кончика корня. Считаем, что функция определена в неотрицательной области значений переменной, неотрицательная, гладкая, ограничена сверху и достаточно простая, в том смысле, что у системы (1) имеется конечное число стационаров.

В модели заданы следующие параметры: α_i – скорость притока ауксина из побега в i-ю клетку; k_{pt} – константа скорости пассивного транспорта ауксина; k_{at} – константа скорости активного транспорта ауксина; k_d – константа скорости деградации/диссипации ауксина. Концентрации и время в модели (1) измеряются в условных единицах.

Используемые в работе термины

Транспортный контакт (контакт) между клетками – наличие диффузии между клетками.

Равномерный поток ауксина из побега в корень – в каждую клетку поперечного среза корня из побега поток поступает с одной и той же скоростью:

$$(\alpha_i = \alpha,\ i = 1,\ \ldots,\ n). \qquad (2)$$

Равномерный стационар – концентрации ауксина в клетках равны:

$$(x_i = x,\ i = 1,\ \ldots,\ n). \qquad (3)$$

РЕЗУЛЬТАТЫ

Исследование стационаров модели (1) при неравномерном потоке ауксина из побега в корень

Раздел посвящен исследованию стационаров модели (1) и механизмов их формирования, так как мы полагаем, что неравномерные стационары могут выступать в качестве морфогенетического поля, под управлением которого на ранних этапах развития растения происходит специализация клеток цилиндра корня.

Сначала отметим, что поток ауксина из побега может выступать в качестве эффективного фактора формирования в поперечном слое клеток морфогенетического поля распределения концентраций ауксина.

Действительно, пусть в системе (1) необходимо обеспечить существование стационара (x_1, ..., x_n). Подставим значения внутриклеточных концентраций ауксина x_i в правую часть системы (1) и приравняем ее к нулю. Тогда искомые значения интенсивностей потоков из побега в индивидуальные клетки, которые обеспечивают существование данного стационара, равны $\alpha_i = (k_{at}P\,(x_i) + k_d)x_i$, $i = 1, ..., n$. Более того, всегда можно обеспечить устойчивость данного стационара. Для этого достаточно взять функцию P, удовлетворяющую неравенству:

$$\max_{i=1,\ldots,n}\left[P\left(x_i\right) + P'\left(x_i\right)x_i\right] > -\frac{k_d}{k_{at}}.$$

Таким образом, у растения существует простой механизм реализации в поперечном срезе стационара целевой конфигурации. Основан он на внешнем первоначальном формировании морфогенетической информации в побеге и в последующем ее переносе в корень. Этот путь не противоречит биологической сути изучаемого процесса: растение в определенный момент развития вполне способно перейти от равномерного потока ауксина из побега в корень к неравномерному. Поэтому следует признать, что неравномерность потока ауксина из побега в корень является естественным фактором реализации конкретного целевого распределения концентраций ауксина в клетках поперечного слоя.

Возможность его реализации в развитии растений подкреплена экспериментальными

данными, согласно которым дифференцировка проводящих тканей в зародыше начинается уже после формирования семядолей – основного источника ауксина (Capron *et al.*, 2009).

Тем не менее, не оспаривая очевидных преимуществ такого механизма формирования морфогенетических полей, нельзя исключить возможность существования внутренних механизмов, которые позволяют решать проблему формирования необходимых распределений внутриклеточных концентраций ауксина, не прибегая к внешним источникам, т. е. при равномерном потоке ауксина из побега в корень, который не несет в себе позиционной информации.

Исследование стационаров модели (1) при равномерном потоке ауксина из побега в корень

Формально проблема поиска стационаров системы (1) сводится к определению неотрицательных решений системы алгебраических уравнений, которая получается путем приравнивания правой части системы (1) к нулю. Сразу отметим очевидное наблюдение, что если поток ауксина из побега в корень равномерный, то система (1) имеет как минимум один равномерный стационар. В то же время легко построить пример системы (1), в которой наряду с равномерным стационаром будут существовать и неравномерные. В этом можно убедиться уже для $n = 2$. Модель (1) принимает следующий вид:

$$\begin{cases} \dfrac{dx_1}{dt} = \alpha - k_{pt}\left[x_1 - x_2\right] - \left(k_{at}P(x_1) + k_d\right)x_1, \\ \dfrac{dx_2}{dt} = \alpha - k_{pt}\left[x_2 - x_1\right] - \left(k_{at}P(x_2) + k_d\right)x_2. \end{cases} \quad (4)$$

Зададим значения $0 < F_1 < F_2 < 1$ и вычислим

$$0 < w_1 = \frac{k_{at}F_1 + k_d}{k_{pt}} < w_2 = \frac{k_{at}F_2 + k_d}{k_{pt}}.$$

Найдем решение системы

$$\begin{pmatrix} -1 - w_1 & 1 \\ 1 & -1 - w_2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -\mathrm{A} \\ -\mathrm{A} \end{pmatrix},$$

$$\mathrm{A} = \frac{\alpha}{k_{pt}} : x_1 = \frac{\mathrm{A}(2 + w_2)}{w_1 w_2 + w_1 + w_2},$$

$$x_2 = \frac{\mathrm{A}(2 + w_1)}{w_1 w_2 + w_1 + w_2},$$

Построим функцию Хилла $P(x) = \dfrac{1}{1 + \left(\dfrac{x}{K_I}\right)^{h_I}}$,

проходящую через точки (x_1, F_1), (x_2, F_2).

Для этого решим уравнения:

$$\left.\begin{aligned} \frac{1}{1 + \left(\frac{x_1}{K_I}\right)^{h_I}} = P(x_1) \Rightarrow \left(\frac{x_1}{K_I}\right)^{h_I} = \frac{1 - P(x_1)}{P(x_1)} \\ \frac{1}{1 + \left(\frac{x_2}{K_I}\right)^{h_I}} = P(x_2) \Rightarrow \left(\frac{x_2}{K_I}\right)^{h_I} = \frac{1 - P(x_2)}{P(x_2)} \end{aligned}\right\} \Rightarrow$$

$$\Rightarrow h_I = \frac{\ln\dfrac{\left(1 - P(x_1)\right)P(x_2)}{\left(1 - P(x_2)\right)P(x_1)}}{\ln\left(\dfrac{x_1}{x_2}\right)}, K_I = \frac{x_1}{\sqrt[h_I]{\dfrac{1 - P(x_1)}{P(x_1)}}}.$$

Численный пример множественности стационаров для двумерной системы (4) приведен на рис. 2.

На основе двумерной системы (4) легко строятся системы более высокой размерности, имеющие стационары, которые могут выступать в качестве морфогенетического поля для формирования диархной структуры цилиндра корня.

В качестве примера приведем модель, которая описывает слой, состоящий из четырех клеток, соединенных между собой транспортными связями по схеме: первая и четвертая клетки обмениваются ауксином в поперечном направлении со второй и третьей клеткой, вторая и третья клетки также обмениваются ауксином. Других транспортных контактов клетки не имеют.

Модель (1) для такой системы клеток имеет следующий вид:

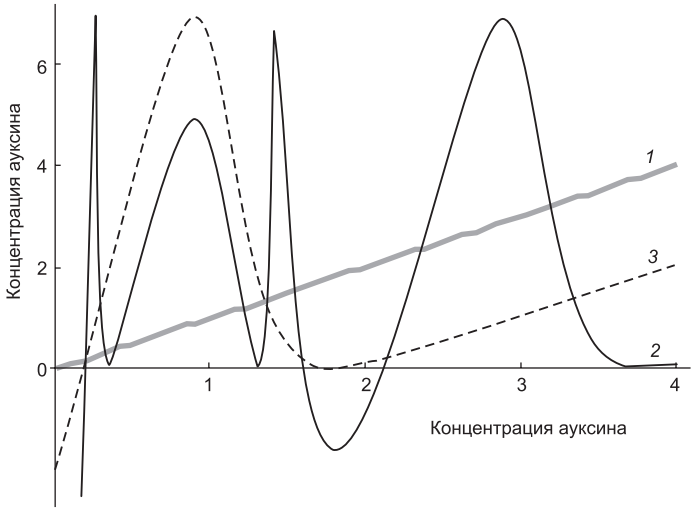


Рис. 2. Множественность стационаров, рассчитанная по модели (4).

Линия *1* – изменение параметра, задающего концентрацию ауксина в 1-й клетке, линии *2* и *3* – рассчитанные концентрации ауксина в 1-й и 2-й клетке. Пересечения кривых *1* и *2* соответствуют стационарам. На графике их показано восемь. Первый и пятый (слева направо) – равномерные стационары (в этих точках кривые *1*, *2*, *3* пересекаются одновременно), остальные – неравномерные. Девятый стационар лежит в интервале (444,440, … 444,445) и на графике не представлен в силу удаленности. Этот стационар равномерный. Расчеты проведены со следующими значениями параметров: $\alpha = 2{,}0$, $k_{pt} = 1{,}0$, $k_{at} = 10{,}0$, $k_d = 0{,}0045$, $h_1 = 9{,}225$, $K_1 = 1{,}122$.

$$\begin{cases} \dfrac{dx_1}{dt} = \alpha - k_{pt}\left[2x_1 - x_2 - x_3\right] - \left(k_{at}P(x_1) + k_d\right)x_1, \\ \dfrac{dx_2}{dt} = \alpha - k_{pt}\left[3x_2 - x_1 - x_3 - x_4\right] - \left(k_{at}P(x_2) + k_d\right)x_2, \\ \dfrac{dx_3}{dt} = \alpha - k_{pt}\left[3x_3 - x_1 - x_2 - x_4\right] - \left(k_{at}P(x_3) + k_d\right)x_3, \\ \dfrac{dx_4}{dt} = \alpha - k_{pt}\left[2x_4 - x_2 - x_3\right] - \left(k_{at}P(x_4) + k_d\right)x_4, \end{cases}$$

и в ней при $\alpha = 2{,}0$, $k_{pt} = 0{,}5$, $k_{at} = 10{,}0$, $k_d = 0{,}0045$, $h_1 = 9{,}225$, $K_1 = 1{,}122$ существует стационар диархной конфигурации: $x_1 = x_4 < x_2 = x_3$.

Таким образом, при равномерном потоке ауксина из побега в ансамбль клеток нелинейность активного транспорта обеспечивает существование в них неравномерных распределений.

Причем не составляет труда указать условия, при которых конфигурация неравномерного распределения концентраций ауксина в клетках будет нести позиционную информацию о диархной структуре цилиндра корня. Однако любой неравномерный стационар, в том числе и целевой, никогда не существует в клеточном ансамбле в единственном экземпляре. Всегда в системе присутствует как минимум дополнительный равномерный стационар. Этот результат означает, что для формирования в процессе развития растения стационара нужной (целевой) конфигурации в поперечном срезе корня на фоне равномерного потока ауксина из побега, необходимо обеспечить условия, при которых все остальные стационары будут неустойчивыми, т. е. возникает необходимость исследовать равномерные стационары системы (1) на устойчивость. Решение вопроса дано в следующем разделе.

О равномерных стационарах модели (1)

Стационар назовем минимальным, если он содержит минимальное значение компонента. Будем считать, что нелинейная функция активного транспорта достаточно проста, в том смысле, что у системы (1) имеется конечное число стационаров, тогда минимальный стационар всегда существует. Более того минимальный стационар системы (1) является равномерным и устойчивым.

Доказательство равномерности минимального стационара

Пусть $\alpha_1 \le ... \le \alpha_n$ – минимальный стационар системы (1). Принятый порядок возрастания значений переменных не ограничивает общности рассуждения, так как в случае необходимости мы можем произвести перенумерацию переменных системы (1).

Предположим, что данный стационар является неравномерным. Тогда

$$\alpha + k_{pt}\left[\sum_{j\in N_1} a_j - |N_1| a_1\right] - \left(k_{at}P(a_1) + k_d\right)a_1 = 0 \Rightarrow$$

$$\Rightarrow \alpha - \left(k_{at}P(a_1) + k_d\right)a_1 \le 0 .$$

Если в неравенстве выполняется равенство, то имеем

$\alpha_1 = ... = \alpha_J \le ... \le \alpha_n$, $J = \max(j \in N_1)$.

Тогда имеем $\alpha - (k_{at} P(a_J) + k_d)\, a_J \le 0$. Если снова выполняется равенство, то увеличиваем номер до $J = \max(j \in N_2)$. Очевидно, что найдется такой номер J, для которого уже выполняется строгое неравенство. В противном случае стационар равномерный, что противоречит предположению. Отсюда имеем, что если стационар не является равномерным, то для всех минимальных значений его компонентов, в том числе и для первой переменной, имеем строгое неравенство.

Так как при $a_1 = 0$ имеем $\alpha - (k_{at} P(a_1) + k_d)\, a_1 = a > 0$, то в силу непрерывности функции P существует $0 < \alpha_0 < \alpha_1 \Rightarrow \alpha - (k_{at} P(a_0) + k_d)\, a_0 = 0$. Получили противоречие.

Этим мы доказали, что минимальный стационар системы (1) является равномерным.

Доказательство устойчивости минимального стационара

Все компоненты равномерной точки покоя вычисляются из уравнения $\alpha = (k_{at} P(a) + k_d)\, a$.

Обозначим решение через a_0. Так как точка покоя $(a_0, \ldots a_0)$ минимальная, то

$$\forall a_l < a_0 \Rightarrow \alpha > \left(k_{at}P(a_l) + k_d\right)a_l \text{ и}$$

$$\forall a_0 < a_r < a_0 + \varepsilon \Rightarrow \alpha < \left(k_{at}P(a_r) + k_d\right)a_r .$$

$$\frac{\left(k_{at}P(a_r) + k_d\right)a_r - \left(k_{at}P(a_l) + k_d\right)a_l}{a_r - a_l} > 0 \Rightarrow$$

$$\Rightarrow \left.\frac{d\left[\left(k_{at}P(a) + k_d\right)a\right]}{da}\right|_{a=a_0} \ge 0 .$$

Если выполняется строгое неравенство, то стационар является устойчивым.

Пусть $\left.\dfrac{d\left[\left(k_{at}P(a) + k_d\right)a\right]}{da}\right|_{a=a_0} = 0 .$

В бесконечно малой окрестности минимального стационара можем записать

$$\frac{d(a)}{dt} = -\sum_{k=k_{\min}}^{\infty} \left.\frac{d^k\left[\left(k_{at}P(a) + k_d\right)a\right]}{k!\,da^k}\right|_{a=a_0} \left(a - a_0\right)^k, i = 1, ..., n ,$$

где $k_{\min}$ равно минимальному значению члена разложения функции в ряд Тейлора, который не равен нулю. Так как слева функция a строго убывает, то

$$\left.\frac{d^{k_{\min}}\left[\left(k_{at}P(a) + k_d\right)a\right]}{k_{\min}!\,da^{k_{\min}}}\right|_{a=a_0} (-1)^{k_{\min}+1} < 0 .$$

В результате получаем, что независимо от четности $k_{\min}$ минимальное стационарное решение устойчиво.

Таким образом, мы доказали, что у системы (1) всегда существует равномерный устойчивый минимальный стационар. Поэтому даже если в ней же существует устойчивый неравномерный стационар, то он не является единственным устойчивым стационаром.

Более того, достаточно очевидно, что примерно равные начальные внутриклеточные концентрации ауксина, скорее, будут лежать в бассейне притяжения равномерного стационара, чем в бассейне притяжения неравномерного стационара, поэтому при равномерном потоке ауксина из побега в корень в качестве итогового распределения, вероятнее всего, будет формироваться именно равномерный стационар, а не другие.

Отсюда вытекает, что система (1) не имеет внутренних механизмов отбора целевого неравномерного стационара из имеющихся стационаров и эта проблема может решаться на постоянной основе только за счет использования внешних факторов.

ОБСУЖДЕНИЕ

Исходя из собственных исследований (Novoselova *et al*., 2013) и ряда экспериментальных данных (Bauby *et al*., 2007; Bishop *et al*., 2011a; Help *et al.*, 2011), мы высказали предположение, что характерное расположение тяжей сосудистых тканей флоэмы и ксилемы на поперечном срезе корня определяется морфогенетическим полем, в качестве которого выступает неравномерное распределение ауксина в клетках поперечного слоя корня. Наиболее просто его формировать через неравномерное поступление ауксина из проростка в ранний корень. Однако также нельзя исключить, что морфогенетическое поле может формироваться за счет внутренних связей и процессов, протекающих в клетках слоя. Для проверки состоятельности этой гипотезы была разработана математическая модель, описывающая поток ауксина, синтезированного в побеге, в горизонтальный слой недифференцированных клеток меристематической зоны корня и его транспорт в нижележащие слои корня.

Анализ модели показал, что даже при равномерном потоке ауксина из побега в корень можно так подобрать функцию активного транспорта и параметры модели (1), что в системе будет присутствовать неравномерное стационарное распределение концентраций ауксина. Тем не менее основной вывод работы состоит в том, что даже если в клеточном ансамбле существует целевой неравномерный стационар, то проблема его фиксации не может быть решена за счет внутренних механизмов, так как в системе всегда дополнительно присутствует равномерное устойчивое минимальное распределение.

Этот результат приводит к выводу, что на определенном этапе развития в клеточном ансамбле должны быть либо модифицированы существующие между клетками связи, либо должны инициироваться новые процессы, которые могут выступить в качестве дополнительного морфогенетического фактора. Такими факторами могут выступать активный транспорт ауксина в поперечных направлениях, асимметричный синтез ауксина *de novo* в клетках слоя, неравномерный поток ауксина из побега и т. д. Например, в работе Muraro с соавт. (2014) показана возможность влияния распределения PIN-транспортеров в центральном цилиндре корня на стабильность распределения ауксина по диархному типу. Однако ничего не известно о механизмах формирования этих потоков в нужной конфигурации.

Асимметричность синтеза ауксина в клетках горизонтального слоя может быть связана с возникновением на определенном этапе в данном слое меристематических клеток, в которых происходит основной синтез ауксина *de novo* (Ljung *et al.*, 2005). Однако, чтобы сформировать меристематические клетки, необходимо уже иметь градиент концентрации ауксина. В результате получаем противоречие: чтобы формировать градиент концентрации ауксина, требуется наличие этого градиента.

Поэтому наиболее вероятным фактором формирования морфогенетического поля в поперечном слое клеток корня, по нашему мнению, являются не структурные отношения между клетками слоя и не процессы, протекающие в них, а подходящая конфигурация неравномерного потока ауксина из побега в корень. Это предположение поддержано рядом исследований, свидетельствующих о том, что асимметрия центрального цилиндра корня закладывается еще в зародыше растения, после начала формирования семядолей (Capron *et al*., 2009). Ксилемные элементы в корне арабидопсиса дифференцируются, как правило, в плоскости формирования семядолей (Bauby *et al*., 2007), предопределяя диархную структуру центрального цилиндра корня на поперечном срезе.

Тот факт, что у мутантов арабидопсиса со слитыми семядолями строение центрального цилиндра корня дезорганизовано (Help *et al*., 2011), предполагает наличие связи между процессами формирования семядолей и дифференцировкой клеток сосудистых тканей, которая реализуется через ауксин как морфоген (Benková *et al*., 2009). В корне взрослого растения симметрия поддерживается нисходящим по флоэме потоком ауксина, однако «правильная» структура проводящих пучков формируется *de novo* во всех боковых и придаточных корнях. Является ли входящий поток ауксина в развивающиеся корни также асимметричным – вопрос для дальнейшего экспериментального и теоретического исследования.

Формирование распределений ауксина более сложных, чем диархное, представляет отдельный интерес. Однако, в целом, существующих экспериментальных данных недостаточно для выдвижения конкретных гипотез о механизмах формирования сложных симметрий в структуре поперечного среза корня различных растений. Можно только полагать, что в формировании сложных неравномерных распределений ауксина могут играть роль и другие гормоны, в частности цитокинин. Его преимущественный транспорт из побега в корень по сосудистым пучкам флоэмы и индуцирующее влияние на активность некоторых PIN-белков (Bishopp *et al*., 2011a, b; Muraro *et al*., 2011, 2014) могут создавать дополнительные условия для перераспределения ауксина в горизонтальном слое клеток центрального цилиндра и накопления ауксина в клетках, предопределяя их развитие в направлении ксилемы. Это направление предполагает развитие более сложных моделей гормон-зависимого механизма формирования сосудистого пучка корня и будет предметом дальнейших исследований.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке РФФИ (проект 13-01-00344), СО РАН (интеграционный проект № 80), фонда «Династия» (грант для молодых биологов) и бюджетного проекта (VI.61.1.2).

ЛИТЕРАТУРА

Лихошвай В.А., Омельянчук Н.А., Миронова В.В. и др. Математическая модель распределения ауксина в корне растения // Онтогенез. 2007. Т. 38. С. 446–456.

Bauby H., Divol F., Truernit E. *et al*. Protophloem differentiation in early *Arabidopsis thaliana* development // Plant Cell Physiol. 2007. V. 48. P. 97–109.

Bayer E.M., Smith R.S., Mandel T. *et al*. Integration of transport-based models for phyllotaxis and midvein formation // Genes Dev. 2009. V. 23. P. 373–384.

Benková E., Ivanchenko M.G., Friml J. *et al*. A morphogenetic trigger: is there an emerging concept in plant developmental biology? // Trends Plant Sci. 2009. V. 14. P. 189–193.

Benková E., Michniewicz M., Sauer M. *et al*. Local, efflux-dependent auxin gradients as a common module for plant organ formation // Cell. 2003. V. 115. P. 591–602.

Bishopp A., Lehesranta S., Vatén A. *et al*. Phloem-transported cytokinin regulates polar auxin transport and maintains vascular pattern in the root meristem // Curr. Biol. 2011a. V. 21. P. 927–932.

Bishopp A., Help H., El-Showk S. *et al*. A mutually inhibitory interaction between auxin and cytokinin specifies vascular pattern in roots // Curr. Biol. 2011b. V. 21. P. 917–926.

Capron A., Chatfield S., Provart N., Berleth T. Embryogenesis: pattern formation from a single cell // Arabidopsis Book. 2009. V. 7. P. e0126.

Help H., Mähönen A.P., Helariutta Y., Bishopp A. Bisymmetry in the embryonic root is dependent on cotyledon number and position // Plant Signal Behav. 2011. V. 6. P. 1837–1840.

Ibañes M., Fàbregas N., Chory J., Caño-Delgado A.I. Brassinosteroid signaling and auxin transport are required to establish the periodic pattern of Arabidopsis shoot vascular bundles // Proc. Natl. Acad. Sci. USA. 2009. V. 106. P. 13630–13635.

Ljung K., Hull A.K., Celenza J. *et al*. Sites and regulation of auxin biosynthesis in Arabidopsis roots // Plant Cell. 2005. V. 17. P. 1090–1104.

Mironova V.V., Omelyanchuk N.A., Novoselova E.S. *et al*. Combined *in silico*/*in vivo* analysis of mechanisms providing for root apical meristem self-organization and maintenance // Ann. Bot. 2012. V. 110. P. 349–360.

Mironova V.V., Omelyanchuk N.A., Yosiphon G. *et al.* A plausible mechanism for auxin patterning along the developing root // BMC Syst Biol. 2010. V. 98. P. 98.

Mitchison G.J. A model for vein formation in higher plants // Proc. R. Soc. Lond. B. 1980. V. 207. P. 79–109.

Mitchison G.J., Hanke D.E., Sheldrake A.R. The polar transport of auxin and vein patterns in plants // Phil. Trans. R. Soc. Lond. B. 1981. V. 295. P. 461–471.

Muraro D., Wilson M., Bennett M.J. Root development: cytokinin transport matters, too! // Curr. Biol. 2011. V. 21. P. R423–425.

Muraro D., Mellor N., Pound M.P. *et al*. Integration of hormonal signaling networks and mobile microRNAs is required for vascular patterning in Arabidopsis roots // Proc. Natl Acad. Sci. USA. 2014. V. 111, P. 857–862.

Novoselova E.S., Mironova V.V., Omelyanchuk N.A. *et al*. Mathematical modeling of auxin transport in protoxylem and protophloem of *Arabidopsis thaliana* root tips // J. Bioinform. Comput. Biol. 2013. V. 11. 1340010.

Petrášek J., Friml J. Auxin transport routes in plant development // Development. 2009. V. 136. No. 16. P. 2675–2688.

Swarup R., Friml J., Marchant A. *et al*. Localization of the auxin permease AUX1 suggests two functionally distinct hormone transport pathways operate in the Arabidopsis root apex // Genes Dev. 2001. V. 15. P. 2648–2653.

Vieten A., Vanneste S., Wisniewska J. *et al*. Functional redundancy of PIN proteins is accompanied by auxin-dependent cross-regulation of PIN expression // Development. 2005. V. 132. P. 4521–4531.

AUXIN DISTRIBUTION IN A TRANSVERSE ROOT SECTION

E.S. Novoselova[1], V.V. Mironova[1, 2], T.M. Khlebodarova[1], V.A. Likhoshvai[1, 2]

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

Summary

Plants differ in the types of the root central cylinder: diarch, triarch, tetrarch, pentarch, or polyarch. The type of the symmetry is the reflection of the relative positions of xylem and phloem bundles in a cross section of the root. The mechanisms forming different types of symmetries in the central cylinder remain poorly understood. It is assumed that vasculature differentiation is triggered and controlled by plant hormone auxin (Sachs, 1969). We have developed a model that describes auxin flow through a cell layer, imitating a cross section of the vascular cylinder in a root. We have studied the stationary distributions of auxin in the cell layer depending on the model parameters. It is shown that the nonlinear processes of auxin transport regulation are responsible for the formation of asymmetric auxin distributions, which may be interpreted as the positional information for development of the diarch structure of the vascular cylinder. However, these distributions always coexist with uniform stationary distributions, not providing positional information. It is hypothesized that the most likely factor in the formation of the final auxin distribution in a root section is an appropriate geometry of the auxin flow from the shoot to the root.

Key words: mathematical modeling, morphogen, auxin, root, vascular tissue differentiation, phloem, xylem.

УДК 573.2, 57.017.6

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ РЕГУЛЯЦИИ ФИТОГОРМОНАМИ ФОРМИРОВАНИЯ МЕРИСТЕМАТИЧЕСКОЙ ЗОНЫ КОРНЯ

© 2014 г. **В.В. Лавреха[1], Н.А. Омельянчук[1], В.В. Миронова[1, 2]**

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

Поступила в редакцию 15 сентября 2014 г. Принята к публикации 28 октября 2014 г.

Расположенная в кончике корня апикальная меристема растения – один из удобных объектов исследования организации ниши стволовых клеток. В апикальной меристеме корня митотически слабо активные клетки покоящегося центра соседствуют с активно делящимися клетками, которые теряют эту способность на определенном расстоянии от покоящегося центра. Известно, что важную роль в регуляции формирования такой структуры играют фитогормоны ауксин и цитокинин, однако конкретные механизмы поддержания ее в динамике пока неизвестны. В работе предложена математическая модель, которая обобщает экспериментальные данные о распределении ауксина и цитокинина вдоль продольной оси корня и их роли в регуляции клеточного цикла. Минимальный механизм регуляции клеточного цикла ауксином и цитокинином, лежащий в основе модели, позволил продемонстрировать *in silico* самоорганизацию меристематической зоны корня в градиентах концентраций этих веществ.

Ключевые слова: *Arabidopsis thaliana*, математическое моделирование, ауксин, цитокинин, клеточный цикл.

ВВЕДЕНИЕ

Вдоль вертикальной оси от кончика корень растений подразделяется последовательно на колумеллу, меристематическую зону (МЗ), зоны элонгации и дифференцировки (Dolan *et al.*, 1993) (рис. 1, *а*). В МЗ по продольной оси снизу вверх расположены пролиферационный и переходный (транзитный) домены (Ivanov, Dubrovsky, 2013). Внизу пролиферационного домена находится ниша стволовых клеток (НСК), а далее расположены митотически-активные (транзитно-амплифицирующиеся, ТА) клетки. НСК состоит из стволовых клеток, окружающих митотически слабо активный покоящийся центр (ПЦ), у *Arabidopsis thaliana* L. состоящий из четырех клеток (Dolan *et al.*, 1993). В переходном домене ТА клетки начинают дифференцироваться, теряют способность к делению и выходят из клеточного цикла. Ауксин и цитокинин играют антагонистические роли в регуляции деления клеток растений: ауксин вызывает деление клеток в МЗ, в то время как цитокинин способствует началу дифференцировки клеток в переходном домене (Dello Ioio *et al.*, 2008). Увеличение концентрации цитокинина в тканях за счет обработки экзогенным гормоном или усиления эндогенного синтеза приводит к ингибированию роста корней и уменьшению размеров пролиферационного домена в МЗ, в то время как снижение эндогенного уровня цитокинина имеет противоположный эффект (Kuderova *et al.*, 2008).

В корне градиент ауксина является основополагающим фактором в установлении местоположения НСК и поддержании ее размеров (Sabatini *et al.*, 1999). Распределение концентрации ауксина имеет максимум в ПЦ и колумелле (рис. 1, *б*), который формируется через активное перераспределение ауксина PIN транспортерами (Grieneisen *et al.*, 2007). Ауксин, регулируя транскрипцию генов *PIN* (на

рис. 1, *г* представлен паттерн экспрессии гена *PIN1*), стабильность и поляризацию белков PIN, контролирует становление своего градиента (Mironova *et al.*, 2010, 2012). Основываясь на данных об активности цитокинин-чувствительного репортера TCS (Zürcher *et al.*, 2013), можно предполагать повышенную концентрацию цитокинина в клетках кончика корня и переходном домене (рис. 1, *в*).

Приблизиться к изучению связи между распределением морфогенов в кончике корня и их ролью в регуляции роста и деления клеток удалось совсем недавно, с развитием методов математического моделирования. На данный момент опубликовано три работы, в которых с помощью математических моделей проанализированы механизмы гормональной регуляции клеточной динамики в корне (Grieneisen *et al.*, 2007; Mironova *et al.*, 2010; Barrio *et al.*, 2013).

В работе Grieneisen с соавт. (2007) основным регулятором роста и деления клеток назван ауксин. Регуляция осуществляется по следующим правилам: (1) клетка не растет больше определенного размера; чтобы клетка могла поделиться, (2) ее размер должен быть больше минимального; (3) уровень ауксина должен быть выше порогового. Клеточный цикл в модели состоит из двух фаз: T1 – фазы медленного роста и T2 – фазы быстрого роста, время прохождения клеткой которых зависит от концентрации ауксина.

В работе Barrio с соавт. (2013) регуляторы клеточного цикла в меристеме корня представлены CYCD, CYCA и CYCB циклинами, колебания концентрации которых определяют переходы между фазами G1/S и G2/M. Авторы предположили, что период клеточного цикла обратно пропорционален концентрации ауксина в клетке. Модель Лотки – Вольтерры была выбрана авторами в качестве модели осцил-

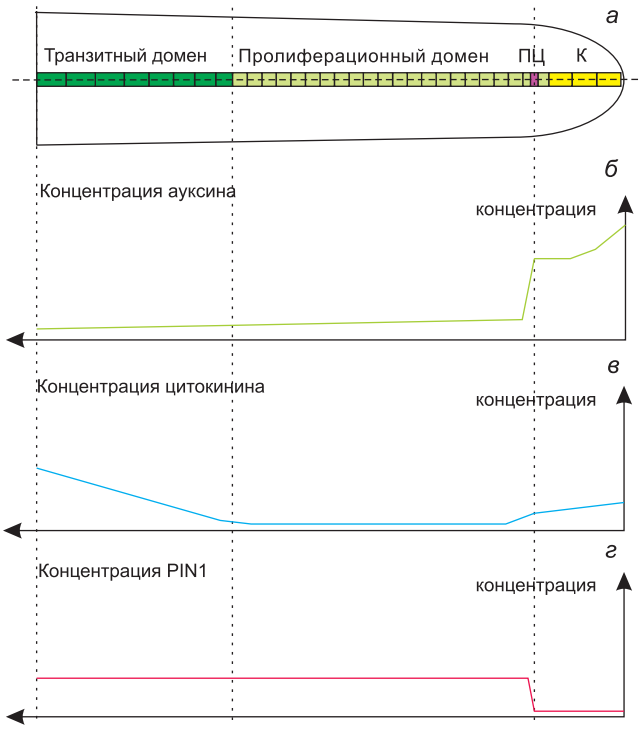


Рис. 1. Структура МЗ и распределение концентраций ауксина, цитокинина и белка-транспортера ауксина PIN1 в соответствии с зонированием: *а* – в кончике корня выделяют следующие подзоны сверху вниз (на рисунке слева направо): транзитный домен, пролиферационный домен с ПЦ в его составе и К (колумелу); *б* – профиль распределения ауксина в корне, восстановленный по активности ауксин-чувствительного репортера DR5 (Sabatini *et al.*, 1999); *в* – профиль распределения цитокинина в корне, восстановленный по активности цитокинин-чувствительного репортера TCS (Zürcher *et al.*, 2013); *г* – паттерн экспрессии *PIN1* (Dello Ioio *et al.*, 2008).

ляций CYCD и CYCB/CYCA в зависимости от концентрации ауксина (Lotka, Dublin, 1925, Вольтерра, 1976).

В опубликованной нами ранее работе (Mironova *et al.*, 2010) гормональная регуляция клеточного цикла осуществлялась с учетом дополнительного морфогена – фактора деления (*Division Factor*). Мы предположили, что скорость деления клетки нелинейно зависит от концентрации фактора деления: низкая (или нулевая) реализуется при недостатке или избытке фактора деления, а высокая – при средних значениях его концентрации. Распределение фактора деления, в свою очередь, регулируется ауксином. В данной работе математическая модель (Mironova *et al.*, 2010) находит дальнейшее развитие. Мы заменили гипотетический фактор деления на реальные гормоны растений, ауксин и цитокинин, и таким образом получили минимальный необходимый механизм формирования структуры кончика корня (колумеллы, пролиферационного и транзитного доменов) в зависимости от градиентов ауксина и цитокинина. Непротиворечивость и достаточность предложенного механизма формирования меристемы вдоль продольной оси корня были протестированы с помощью математического моделирования.

МАТЕРИАЛЫ И МЕТОДЫ

1. Основные допущения модели

Ниже перечислим основные допущения модели и их биологическое обоснование.

1. В ПЦ и колумелле имеется максимум концентрации ауксина (Sabatini *et al.*, 1999). Формирование градиента концентрации ауксина в кончике корня в настоящей модели описано по механизму отраженной волны (рис. 1, *а*, *б*), аналогично (Mironova *et al.*, 2010).

2. В работе Мироновой с соавт. (Mironova *et al.*, 2010) единственным источником ауксина в зоне моделирования является его приток из побега, которого недостаточно для сохранения максимума концентрации ауксина при увеличении клеточного ансамбля больше 100 клеток. Дальнейшее развитие корня требует учета синтеза ауксина *de novo*, что соответствует экспериментальным данным Bhalerao с соавт. (2002).

3. Согласно данным о цитокинин-чувствительном репортере TCS (Zürcher *et al.*, 2013), повышенная концентрация цитокинина наблюдается в колумелле и переходном домене меристемы (рис. 1, *в*). Формирование градиента концентрации цитокинина в корне ранее нигде не исследовано, и в данной работе мы впервые предлагаем его механизм:

а) изменение концентрации цитокинина в клетках зависит от процессов синтеза, диффузии и деградации;

б) так как в клетках с высоким уровнем экспрессии репортера TCS (Zürcher *et al.*, 2013) наблюдается повышенная концентрация ауксина (рис. 1), в модели мы описали ауксин-чувствительный синтез цитокинина (рис. 2, *б*);

в) другим источником цитокинина, рассмотренным в модели, служит поток цитокинина из побега, который идет по флоэме корня в его кончик (Bishopp *et al.*, 2011).

4. В модели мы рассмотрели упрощенную схему клеточного цикла, состоящего из двух фаз G1 и G2, и двух наиболее важных контрольных точек G1/S и G2/M. Согласно экспериментальным данным, для клеток, расположенных в различных подзонах МЗ, характерно преимущественное нахождение в определенной фазе клеточного цикла (Breuer *et al.*, 2014; рис. 3, *а*, *б*). Клетки колумеллы преимущественно находятся в фазе G1. Клетки пролиферационного домена активно делятся, поэтому могут быть в фазах G1 или G2. В переходном домене клетки все еще способны перейти контрольную точку G1/S, но не способны к делению и находятся в фазе G2.

5. Известно, что и ауксин и цитокинин необходимы для прохождения клеточного цикла. Зная информацию о распределении фаз клеточного цикла (рис. 3, *б*) относительно градиентов концентрации ауксина и цитокинина (рис. 1, *б*, *в*), мы предложили упрощенную модель регуляции клеточного цикла в меристеме корня растений (рис. 2, *в*). Мы предположили, что вероятность прохождения контрольной точки G1/S зависит от двух параметров: размера клетки и концентрации ауксина в ней. Принимая во внимание данные о влиянии цитокинина на размер пролиферационного домена, полученные в работе (Dello Ioio *et al.*, 2008), мы предположили, что вероятность прохождения

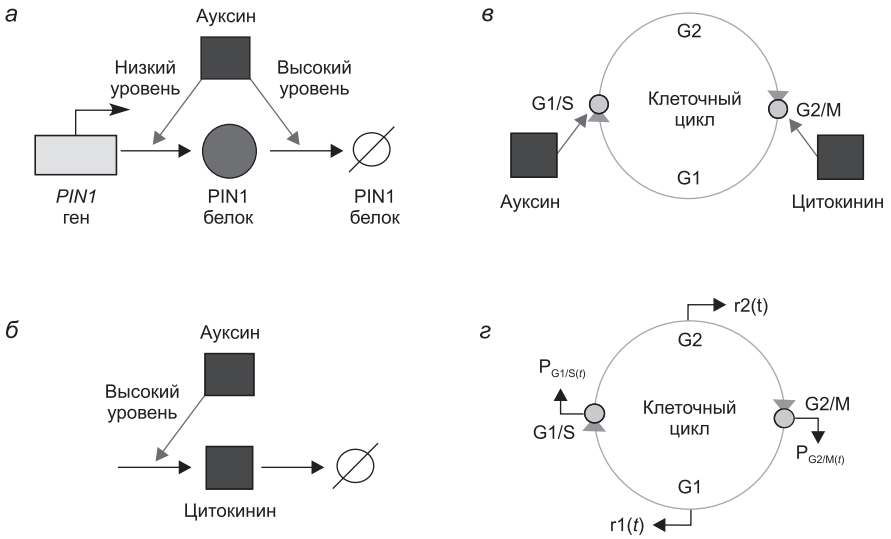


Рис. 2. Схема гормональной регуляции динамических процессов, рассмотренных в модели: *а* – влияние ауксина на синтез и деградацию белка PIN1 (Mironova *et al.*, 2010); *б* – влияние ауксина на динамику концентрации цитокинина в настоящей модели; *в* – схематичное представление гормональной регуляции клеточного цикла в модели; *г* – функции скоростей роста ($r_1(t)$, $r_2(t)$) и вероятности прохождения контрольных точек клеточного цикла ($P_{G1/S}(t)$, $P_{G2/M}(t)$), рассмотренные в модели.

G2/M зависит от концентрации цитокинина в клетке.

6. Известно, что при выходе из МЗ клетки начинают расти более интенсивно (рис. 3, *в*), что было учтено в модели: скорость роста клеток в фазе G2 в три раза больше, чем в фазе G1. Вне зависимости от фазы клеточного цикла, существует ограничение на предельный размер клетки (Sablowski, Dornelas, 2014), что было учтено в математической модели.

2. Переменные в математической модели

В модели рассмотрен ансамбль клеток, расположенных вдоль продольной оси корня, количество клеток N может меняться в течение численного расчета. Самая дистальная клетка, находящаяся на кончике корня, имеет номер 1, последняя клетка в ансамбле N находится в переходном домене. Клетка с номером i имеет длину $2r_i$, (r_i – радиус клетки), координаты центра x_i, фазу клеточного цикла G_i, концентрации активных веществ a_i, $PIN1_i$ и c_i. В настоящей модели изменения концентрации ауксина (a_i) и белка PIN1 ($PIN1_i$) описаны в соответствии с предыдущей версией модели (Mironova *et al.*, 2010). Гипотетический *Division Factor* исключен из модели, но введен гормон цитокинин (c_i). Рассмотрим описание модели более подробно.

Транспортер ауксина PIN1

Как и в модели Мироновой с соавт. (Mironova *et al.*, 2010), концентрация белка PIN1 зависит от концентрации ауксина в клетке (см. рис. 2, *а*) и описана кусочно-линейными функциями:

$$f_{s,pin1}(a_i) = \begin{cases} 0, if\ a_i < T_a^{sp}, \\ 1, else \end{cases}; \quad (1)$$

$$f_{d,pin1}(a_i) = \begin{cases} 9, if\ a_i\ \ T_a^{dp}, \\ 0{,}01, else \end{cases}; \quad (2)$$

$$\frac{dPIN1_i}{dt} = K_{s,pin1} f_{s,pin1}(a_i) - \\ - K_{d,pin1}\left(1 + f_{d,pin1}(a_i)\right) PIN1_i, \quad (3)$$

где a_i, $PIN1_i$ – концентрации ауксина и белка PIN1 в клетке i; $K_{s,\,pin1}$ и $K_{d,\,pin1}$ – коэффициенты синтеза и деградации PIN1; $f_{s,\,pin1}(a_i)$ и $f_{d,\,pin1}(a_i)$ – кусочно-линейные функции управления скоростями синтеза и деградации PIN1 соответственно, в зависимости от концентрации ауксина в

клетке; T_a^{sp}, T_a^{dp} – пороговые значения концентрации ауксина для переключения функций $f_{s,\,pin1}(a_i)$ и $f_{d,\,pin1}(a_i)$.

Ауксин

Изменение концентрации ауксина в настоящей модели описано аналогично (Mironova *et al.*, 2010): рассмотрены процессы диффузии, деградации и активного транспорта. Ауксин поступает из побега через сосудистую систему корня в клетку *N* с интенсивностью α*(t)*, которая растет линейно по времени *t* (4): $\alpha(t) = \alpha_0 + kt$.

Нерегулируемый (пассивный) транспорт ауксина между клетками осуществляется за счет диффузии с коэффициентом D_a. Скорость деградации ауксина в клетке описана линейно с коэффициентом $K_{d,\,a}$. Концентрация ауксина в клетке уменьшается за счет ее удлинения, что учтено функцией скорости разбавления $f_d(r_i)$: $f_d(r_i) = \frac{r_i'}{r_i}$.

Ауксин активно переносится из клетки с номером *i* в клетку с номером *i* – 1 со скоростью $f_t(PIN1_i)$, определяемой концентрацией PIN1 транспортера в клетке *i*:

$f_t(PIN1_i) = K_{at,a} * PIN1_i$.

Все эти процессы подробно описаны в модели (Mironova *et al.*, 2010). В настоящую модель введено дополнение – нерегулируемый синтез ауксина в каждой клетке с интенсивностью σ_0. Общая система уравнений, описывающая изменения концентрации ауксина в клетках, представлена ниже:

$$\frac{da_1}{dt} = -D_a(a_1 - a_2) + \sigma_0 - \\ - K_{d,a}a_1 + f_t(PIN1_2)a_2 - f_d(r_1)a_1 \quad (4)$$

$$\frac{da_i}{dt} = D_a(a_{i+1} + a_{i-1} - 2a_i) + \sigma_0 - \\ - K_{d,a}a_i + f_t(PIN1_{i+1})a_{i+1} - \\ - f_t(PIN1_i)a_i - f_d(r_i)a_i \quad (5)$$

$$\frac{da_N}{dt} = \alpha(t) + D_a(a_{N-1} - a_N) + \sigma_0 \\ - K_{d,a}a_N - f_t(PIN1_N)a_N - f_d(r_N)a_N, \quad (6)$$

где *N* – клетка, ближайшая к побегу, a_i и PIN_i – концентрации ауксина и белка PIN1 в клетке *i*; r_i – размер *i*-й клетки; α(*t*) – поток ауксина из побега; α_0 – базальный уровень интенсивности потока ауксина, поступающего в *N*-ю клетку, *k* – константа прироста интенсивности потока ауксина по времени; D_a и $K_{d,\,a}$ – коэффициенты диффузии и деградации ауксина соответственно; $K_{at,\,a}$ – константа скорости активного транспорта; σ_0 – константа нерегулируемого синтеза ауксина в клетке; $f_d(r_i)$ – функция скорости разбавления концентрации вещества в клетке *i*, представляющая отношение скорости прироста радиуса клетки r'_i к ее текущему радиусу r_i; r'_i – скорость прироста радиуса клетки за время *dt*.

Цитокинин

Изменение концентрации цитокинина описано процессами диффузии, деградации и синтеза. Один из источников цитокинина – его поток в клетку *N* с интенсивностью β(*t*), которая растет линейно по времени *t*: $\beta(t) = \beta_0 + lt$.

Нерегулируемый (пассивный) транспорт цитокинина между клетками осуществляется за счет диффузии с коэффициентом D_c. Скорость деградации цитокинина в клетке описана линейной функцией с коэффициентом $K_{d,\,c}$. Уменьшение концентрации цитокинина в клетке за счет ее удлинения учтено с использованием функции скорости разбавления $f_d(r_i)$ (см. Ауксин). Регулируемый синтез цитокинина описан кусочно-линейной функцией $f_{s,\,c}(a_i)$. При концентрации ауксина ниже пороговой (T_a^{sc}) рассматривается базальная (невысокая) скорость синтеза цитокинина, при концентрации ауксина выше T_a^{sc} скорость синтеза цитокинина увеличивается на порядок (рис. 2, *б*):

$$f_{s,c}(a_i) = \begin{cases} 1, \; if \; a_i > T_a^{sc} \\ 0{,}01, else \end{cases}$$

Изменение концентрации цитокинина в клетках описано уравнениями:

$$\frac{dc_1}{dt} = -D_c(c_1 - c_2) + K_{s,c}f_{s,c}(a_1) - \\ - K_{d,c}c_1 - f_d(r_1)c_1 \quad (7)$$

$$\frac{dc_i}{dt} = D_c(c_{i+1} + c_{i-1} - 2c_i) + K_{s,c}f_{s,c}(a_i) - \\ - K_{d,c}c_i - f_d(r_i)c_i \quad (8)$$

$$\frac{dc_N}{dt} = \beta(t) + D_c(c_{N-1} - c_N) + K_{s,c}f_{s,c}(a_N) -$$

$$- K_{d,c}c_N - f_d(r_N)c_N, \quad (9)$$

где a_i, c_i – концентрации ауксина и цитокинина в клетке i; r_i – размер i-й клетки; D_c и $K_{d,c}$ – коэффициенты диффузии и деградации цитокинина в клетке соответственно; $f_{s,c}(a_i)$ – кусочно-линейная функция скорости синтеза цитокинина, где T_a^{sc} – пороговое значение концентрации ауксина для регуляции скорости синтеза цитокинина, $K_{s,c}$ – константа скорости синтеза цитокинина, $\beta(t)$ – скорость потока цитокинина из сосудистой системы корня с параметрами β_0 и l.

Моделирование роста и деления клеток

Для создания модели с клеточными делениями была использована Динамическая грамматика, реализованная в пакете Plenum для Mathematica® (Yosiphon, Mjolsness, 2007), аналогично (Mironova *et al.*, 2010). Рост и перемещение клеток вдоль оси, активный транспорт и диффузия сигнальных веществ между клетками, а также диссипация веществ описаны непрерывными функциями. Переходы между дискретными событиями описаны стохастическими правилами. На рис. 2, в представлена схема клеточного цикла, моделируемого в данной работе. Время нахождения клетки в фазе G1 и G2 – τ, определяется в соответствии с функциями распределения вероятностей прохождения контрольных точек $P_{G1/S}(\tau)$ и $P_{G2/M}(\tau)$:

$$P_{G1/S}(\tau) = \rho_{GP}(r_i) * \rho_{g1/s}(a_i) * \\ * \exp\left(-\rho_{GP}(r_i) * \rho_{g1/s}(a_i) * \tau\right) \quad (10)$$

$$P_{G2/M}(\tau) = \rho_{g2/m}(c_i) * P(x_1, x_2|x) * \\ * \exp\left(-\rho_{g2/m}(c_i) * P(x_1, x_2|x) * \tau\right) \quad (11)$$

Для индивидуальной клетки $P_{G1/S}(\tau)$ зависит от ее размера

$$\rho_{GP}(r_i) = \left(1 + \exp\left(-\frac{r_i - r_{min}}{T}\right)\right)^{-1}$$

и текущей концентрации ауксина

$$\rho_{g1/s}(a_i) = \begin{cases} 1, if\ a_i < T_a^{G1/S} \\ 0, else \end{cases}.$$

$P_{G2/M}(\tau)$, в свою очередь, зависит от концентрации цитокинина:

$$\rho_{g2/m}(c_i) = \begin{cases} 1 - c_i/T_c^{G2/M}, if\ c_i < T_c^{G2/M} \\ 0{,}0001, else \end{cases}, \quad (12)$$

где $P_{G1/S}(\tau)$, $P_{G2/M}(\tau)$ – функции распределения вероятности прохождения G1/S и G2/M; a_i, c_i – концентрации ауксина, цитокинина в клетке i; r_i – размер клетки i; $r_{min} = 1{,}5$ – минимальный размер клетки способной к делению; $T = 0{,}01$ – параметр оценки вклада размера клетки в вероятность перехода G1/S; $T_a^{G1/S}$ – пороговое значение концентрации ауксина для перехода G1/S; $T_c^{G2/M}$ – пороговое значение концентрации цитокинина для перехода G2/M; $P(x_1, x_2|x)$ – функция вероятности положения центров дочерних клеток (x_1 и x_2) после деления клетки с координатой центра x.

Скорости роста клеток в фазах G1 и G2 различаются – $r_1(t)$ и $r_2(t)$:

$$r_1(t) = \begin{cases} r_0 + K_{growth} * t, if\ r_1(t) < r_{max} \\ 0, else \end{cases} \quad (13)$$

$$r_2(t) = \begin{cases} r_0 + 3K_{growth} * t, if\ r_2(t) < r_{max} \\ 0, else \end{cases}, \quad (14)$$

где $K_{growth} = 1/10^4$ – константа роста; $r_{max} = 5$ – максимальный радиус, которого может достичь клетка.

3. Численный расчет модели

В работе проведены численный расчет и анализ двух типов решений модели: статического и динамического. При решении дифференциальных уравнений модели использован численный метод интегрирования, встроенный в пакет Mathematica©, NDSolve с предусмотренными по умолчанию параметрами. При реализации статического решения модели клетки ($N = 20$) не росли и не делились, интенсивности потоков ауксина и цитокинина оставались постоянными (значения параметров модели k и $l = 0$).

В качестве начальных данных использованы равномерные распределения концентраций в клетках: $a_i = 1{,}0$, $c_i = 0{,}001$ и $PIN1_i = 0$, $i = 1, \ldots, N$. Размер клеток в начальный этап времени вычислен как два радиуса клетки (r_i). Радиус выбран случайным образом из интервала от 0,3 до 0,65 условных единиц размера. Исходя

из радиуса клетки рассчитано положение ее центра (x_i) на координатной оси x. Численный расчет велся до t = 30 000.

При реализации динамического решения в модель включались правила динамики клеточного цикла (10)–(14), а также были использованы ненулевые коэффициенты k и l для описания возрастания потоков ауксина и цитокинина во времени. Расчет для динамического решения модели начинали с 15 клеток с распределениями концентраций как для стационарного решения модели: a_i = 1,0, c_i = 0,001 и $PIN1_i$ = 0. В этот момент все клетки находились в фазе G1 клеточного цикла.

РЕЗУЛЬТАТЫ

В данной работе проведено дальнейшее развитие модели Мироновой и соавт. (Mironova *et al.*, 2010) для описания механизма формирования структуры меристемы корня (колумеллы, пролиферационного и переходного доменов) с учетом гормональной регуляции клеточного цикла. Нами был предложен минимальный регуляторный контур (рис. 2, *в*), включающий в себя регуляцию контрольных точек G1/S и G2/M ауксином и цитокинином. Выбранный контур не противоречит экспериментальным данным и, как показано ниже методами математического моделирования, объясняет механизм самоорганизации МЗ корня в соответствии с градиентами концентраций морфогенов.

Решение модели для постоянного числа клеток *N*

В системе Mathematica© были описаны процессы изменения концентрации ауксина (4)–(6), белка PIN1 (1)–(3) и цитокинина (7)–(9). Для анализа паттернов распределений морфогенов использовано стационарное решение модели с N = 20, которое реализовано при начальном равномерном распределении концентраций ауксина, цитокинина и PIN1 белка.

Нами были подобраны значения параметров (см. таблицу) так, чтобы в процессе расчета модели происходила самоорганизация максимума концентрации ауксина в 4-й клетке от кончика корня и двух максимумов концентрации цитокинина в основании и кончике корня (рис. 3, *г*). Эти максимумы сохраняли свое положение сколь угодно долго в течение расчета, что позволяет нам говорить о формировании стационарного решения в соответствии с экспериментальными данными (рис. 1, *в*, *г* и 3, *г*).

Решение модели для динамического числа клеток *N*

На основе предложенных правил регуляции клеточного цикла (10)–(14) получено решение модели с динамическим числом клеток. Расчет велся начиная с N = 15 клеток при растущих во времени потоках ауксина и цитокинина. Расчеты проведены со значениями параметров,

Значение некоторых параметров модели, влияющих на распределение концентраций морфогенов, белка PIN1 и регуляцию клеточного цикла

Параметры для описания динамики веществ в модели								
ауксин			цитокинин			PIN1		
α_0	0,06	*cu*/*tu*	β_0	0,08	*cu*/*tu*	S_p	0,001	1/*tu*
k	0,0001	*cu*/*tu*^{2}	l	0,01	*cu*/*tu*^{2}	D_p	0,001	1/*tu*
σ_0	0,002	*cu*	$K_{s,c}$	0,01	1/*tu*	T_a^{sp}	0,1	*cu*
$K_{d,a}$	0,0055	1/*tu*	$K_{d,c}$	0,008	1/*tu*	T_a^{dp}	0,9	*cu*
D_a	0,06	1/*tu*	D_c	0,08	1/*tu*			
$K_{at,a}$	0,26	1/*tu*	T_a^{sc}	4	tu			
Параметры для описания прохождения контрольных точек								
$T_a^{G1/S}$	1,2	*cu*	$T_c^{G2/M}$	0,5	*cu*	r_{min}	1,5	*dl*

Примечание. Единицы: *cu* – концентрации, *tu* – времени, *dl* – длины.

указанными в таблице. Кроме того, были подобраны значения параметров *k* и *l*, отвечающих за увеличение потоков ауксина и цитокинина во времени соответственно. Подбор осуществлялся так, что распределение ауксина с дистальным максимумом концентрации и распределение цитокинина с дистальным и проксимальным максимумами не менялись качественно при росте клеточного ансамбля.

В процессе расчета максимум концентрации ауксина достаточно быстро формировался в 4-й клетке и сохранялся на протяжении численного эксперимента, что полностью соответствует работе (Mironova *et al.*, 2010) и экспериментальным данным (рис. 1, *б*). Распределение концентрации цитокинина, а именно наличие двух максимумов концентрации, качественно соответствует стационарному решению модели (рис. 3, *г*, *д*) и экспериментальными данными (рис. 1, *в*, *г*, 3, *б*). Самоорганизация и поддержание распределения цитокинина в корне во времени с учетом клеточной динамики показаны впервые.

В решении модели мы наблюдали формирование трех областей, различающихся по статусу клетки в клеточном цикле. На рис. 3, *е* видно, что клетки колумеллы находятся в фазе G1; клетки переходного домена расположены в фазе G2; клетки пролиферационного домена присутствуют как в фазе G1, так и в фазе G2.

Такое распределение клеток по принципу фазовой принадлежности хорошо соотносится с экспериментальными данными (рис. 3, *б*). Эти три подзоны кончика корня сохранялись при расчете сколь угодно продолжительное время, что свидетельствует о непротиворечивости предложенного нами механизма гормональной регуляции клеточного цикла (см. рис. 2, *в*) и его роли в формировании устойчивых в развитии подразделений кончика корня.

ОБСУЖДЕНИЕ

В данной работе предложен минимальный механизм регуляции фитогормонами клеточного цикла в меристематической зоне корня растений, согласно которому ауксин регулирует прохождение контольной точки G1/S, а цитокинин – контрольной точки G2/M. Биологическую основу предложенного механизма составляют следующие данные:

1. У *Arabidopsis* рецептор ауксина ABP1 необходим для регуляции G1/S перехода через CYCD/RBR путь (Tromas *et al.*, 2009). Инактивация

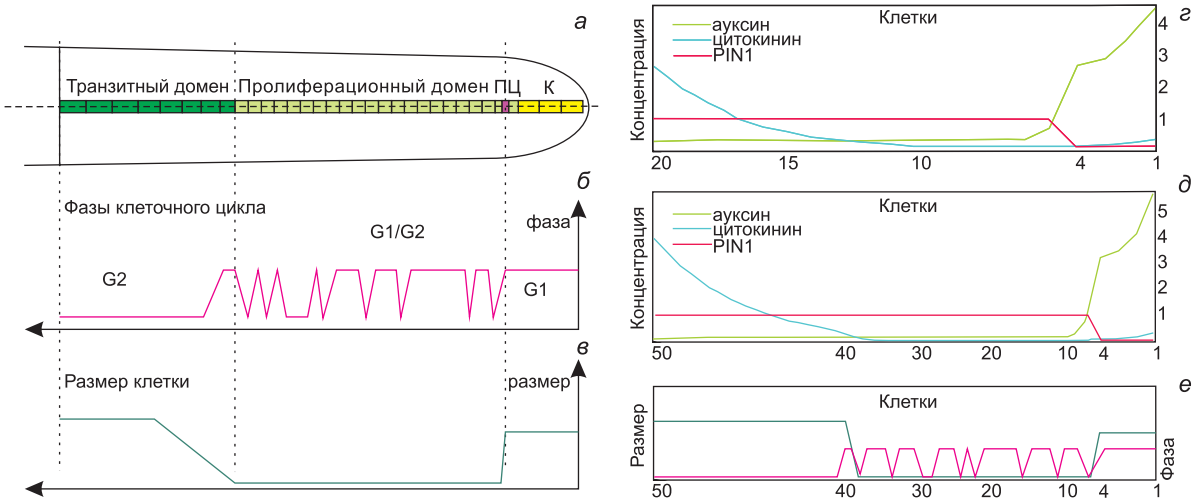


Рис. 3. *In vivo* и *in silico* распределение клеток по фазам клеточного цикла и размерам вдоль центральной оси кончика корня: *а* – расположение подзон кончика корня вдоль центральной оси корня; *б* – распределение клеток по фазам клеточного цикла вдоль центральной оси корня; *в* – распределение клеток по скоростям роста; *г* – стационарное решение модели для $N = 20$; *д* – динамическое решение модели в момент времени $t = 27\,916$; *е* – результат динамического распределения клеток по фазам и скоростям роста.

ABP1 ведет к прекращению делений в меристеме корня, и эти деления не восстанавливаются даже при больших дозах ауксина (Tromas *et al.*, 2009). Через рецепторы TIR1 и SKP2A ауксин вызывает протеолизную деградацию белков E2FC/DPB и Aux/IAA соответственно (Jurado *et al.*, 2010; Del Pozo, Manzano, 2014). E2FC/DPB являются репрессорами клеточного цикла, а Aux/IAA подавляет клеточный ответ на ауксин. Таким образом, ауксин, связываясь с тремя типами своих рецепторов (ABP1, TIR1 и SKP2A), приводит к снятию блокировки с перехода G1/S.

2. В корне *Arabidopsis* цитокинин контролирует переход G2/M, индуцируя экспрессию гена *CDC2* (Hemerly *et al.*, 1993). У *ahk2*, *ahk3*, *ahk4* – мутантов по генам рецепторов цитокинина – укорочен корень в результате уменьшения клеточных делений в МЗ (Higuchi *et al.*, 2004). В клетках кончика корня этих мутантов значимо уменьшено число диплоидных клеток и увеличено число тетраплоидных, что также указывает на то, что цитокинин регулирует переход G2/M.

С помощью математического моделирования мы протестировали, является ли этот механизм достаточным для объяснения формирования структуры меристемы корня.

На первом этапе мы провели моделирование самоорганизации распределений ауксина и цитокинина вдоль нерастущего корня, имеющего постоянное число клеток, и подобрали значения параметров, при которых наблюдается распределение ауксина в корне (рис. 3, *г*), аналогичное описанному ранее (Grieneisen *et al.*, 2007; Mironova *et al.*, 2010), а также экспериментально наблюдаемое распределение цитокинина (Zürcher *et al.,* 2013). На втором этапе мы исследовали достаточность механизма гормональной регуляции клеточного цикла для зонирования меристематической зоны в соответствии с экспериментальными данными. В численном расчете модели растущего корня мы наблюдали формирование в градиентах концентрации ауксина и цитокинина трех доменов клеток с принципиально различающейся клеточной динамикой (рис. 3, *б*, *е*).

Из расчетов модели следует, что пролиферационный домен формируется и сохраняется на всем протяжении моделирования между максимумом концентрации ауксина, с одной стороны, и максимумом концентрации цитокинина, с другой. Такой механизм формирования пролиферационного домена во времени показан нами впервые.

Однако стоит отметить, что в настоящей модели нам не удалось получить пролиферационного домена фиксированного размера – он рос с ростом «корня», а значит, для понимания механизмов организации ниши стволовых клеток в кончике корня необходимо дальнейшее исследование дополнительных факторов регуляции клеточной динамики.

БЛАГОДАРНОСТИ

Работа поддержана грантом фонда «Династия» для молодых биологов и Российским научным фондом 14-14-00734.

ЛИТЕРАТУРА

Вольтерра В. Математическая теория борьбы за существование. М.: Наука, 1976. 286 с.

Barrio R.A., Romero-Arias J.R., Noguez M.A. *et al.* Cell Patterns Emerge from Coupled Chemical and Physical Fields with Cell Proliferation Dynamics: The *Arabidopsis thaliana* Root as a Study System // PLoS Comput. Biol. 2013. V. 9 (5). P. e1003026.

Bhalerao R.P., Eklöf J., Ljung K. *et al.* Shoot-derived auxin is essential for early lateral root emergence in Arabidopsis seedlings // Plant J. 2002. V. 29 (3). P. 325–332.

Bishopp A., Lehesranta S., Vaten A. *et al.* Phloem-transported cytokinin regulates polar auxin transport and maintains vascular pattern in the root meristem // Curr. Biol. 2011. V. 21. P. 927–932.

Breuer C., Braidwood L., Sugimoto K. Endocycling in the path of plant development // Current Opinion Plant Biology. 2014. V. 17. P. 78–85.

Del Pozo J.C., Manzano C. Auxin and the ubiquitin pathway. Two players-one target: the cell cycle in action // J. Exp. Bot. 2014. V. 65 (10). P. 2617–2632.

Dello Ioio R., Nakamura K., Moubayidin L. *et al.* A Genetic Framework for the Control of Cell Division and differentiation in the Root Meristem // Science. 2008. V. 322. P. 1380–1384.

Dolan L., Janmaat K., Willemsen V. *et al.* Cellular organisation of the *Arabidopsis thaliana* root // Development. 1993. V. 119. P. 71–84.

Grieneisen V.A., Xu J., Marée A.F. *et al.* Auxin transport sufficient for maximum and gradient guiding root growth // Nature. 2007. V. 449 (7165). P. 1008–1013.

Hemerly A.S., Ferreira P., De Almeida E.J. *et al.* Cdc2a expression in Arabidopsis is linked with competence for cell division // Plant Cell. 1993. V. 5. P. 1711–1723.

Higuchi M., Pischke M.S., Mähönen A.P. *et al.* In planta functions of the Arabidopsis cytokinin receptor family // Proc. Natl. Acad. Sci. 2004. V. 101. P. 8821–8826.

Ivanov V.B., Dubrovsky J.G. Longitudinal zonation pattern in plant roots: conflicts and solutions // Trends Plant Sci. 2013. V. 18 (5). P. 237–243.

Jurado S., Abraham Z., Manzano C. *et al.* The Arabidopsis cell cycle F-box protein SKP2A binds to auxin // Plant Cell. 2010. V. 22. P. 3891–3904.

Kuderova A., Urbankova I., Valkova M. *et al.* Effects of conditional IPT-dependent cytokinin overproduction on root architecture of Arabidopsis seedlings // Plant Cell. Physiol. 2008. V. 49. P. 570–582.

Lotka A.J., Dublin L.I. On the true rate of natural increase as exemplified by the population of the United States // J. American statistical association. 1925. V. 20 (150).

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 Botany. 2012. V. 110 (2). P. 349–360.

Mironova V.V., Omelyanchuk N.A., Yosiphon G. *et al.* A plausible mechanism for auxin patterning along the developing root // BMC Systems Biology. 2010. V. 4. (98).

Sabatini S., Beis D., Wolkenfelt H. *et al*. An Auxin-Dependent Distal Organizer of Pattern and Polarity in the *Arabidopsis* Root // Cell. 1999. V. 99 (5). P. 463–472.

Sablowski R., Dornelas M. Interplay between cell growth and cell cycle in plants // J. Exp. Bot. 2014. V. 65 (10). P. 2703–2714.

Tromas A., Braun N., Muller P. *et al.* The auxin binding protei 1 is required for differential auxin respon-ses mediating root growth // PLoS One. 2009. V. 4 (9). P. e6648.

Yosiphon G., Mjolsness E. Plenum. 2007. http://computableplant.ics.uci.edu/theses/guy/downloads/papers/thesis.

Zürcher E., Tavor-Deslex D., Lituiev D. *et al.* A robust and sensitive synthetic sensor to monitor the transcriptional output of the cytokinin signaling network in planta // Plant Physiol. 2013. V. 161 (3). P. 1066–1075.

MATHEMATICAL MODEL OF PHYTOHORMONE REGULATION OF ROOT MERISTEMATIC ZONE FORMATION

V.V. Lavrekha[1], N.A. Omelyanchuk[1], V.V. Mironova[1, 2]

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

Summary

The apical meristem located at the root tip of a plant is one of the most convenient objects to study the organization of the stem cell niche. In the root apical meristem, mitotically inactive cells of the quiescent center coexist with intensely dividing cells, which lose this ability at a certain distance from the quiescent center. It is known that plant hormones auxin and cytokinin play an important role in the regulation of this structure formation, but the mechanisms maintaining the dynamics of this structure remain unknown. We propose a mathematical model that summarizes experimental data on the distribution of auxin and cytokinin along the root longitudinal axis and their role in cell cycle regulation.

Key words: *Arabidopsis thaliana*, mathematical modelling, auxin, cytokinin, cell cycle.

УДК 577.112:004.021

ВОССТАНОВЛЕНИЕ АМИНОКИСЛОТНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ ЦИКЛИЧЕСКИХ ПЕПТИДОВ ИЗ МАСС-СПЕКТРОВ

© 2014 г. **Э.С. Фомин**

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

Поступила в редакцию 4 сентября 2014 г. Принята к публикации 6 октября 2014 г.

Метод масс-спектрометрии − один из физических методов исследования протеомов различных организмов, позволяющий решать как задачи идентификации биологических макромолекул, так и секвенирования пептидных цепочек в случаях, когда нет информации о геномах либо эта информация крайне ограничена. В настоящее время существует множество компьютерных программ для поддержки исследований в этой области. Тем не менее, несмотря на высокую активность, имеется только незначительный прогресс в создании программ, позволяющих решать задачи *de novo* секвенирования для циклических пептидов, к которым относятся наиболее эффективные антибиотики, противоопухолевые агенты, иммунодепрессанты, токсины и множество пептидов с неизвестными функциями, синтезируемые в клетке по нерибосомальному пути. Предложен эффективный алгоритм для решения задачи секвенирования циклических пептидов, который позволяет восстанавливать последовательности большой (до 160 аминокислотных остатков) длины.

Ключевые слова: масс-спектрометрия, секвенирование циклических пептидов, проблема beltway.

ВВЕДЕНИЕ

Протеомика – наука о протеомах – долгое время развивалась благодаря методу электрофореза, который позволяет разделить макромолекулы, различающиеся по молекулярной массе, пространственной конфигурации и электрическому заряду за счет разной скорости их диффундирования в буферном растворе под действием электрического тока (Остерман, 1981). Этот метод используют почти в каждой биохимической лаборатории, но для больших протеомных проектов в настоящее время определяющую роль играют другие, более мощные физические методы исследования, такие как ядерный магнитный резонанс (ЯМР) и метод масс-спектрометрии (МС).

Метод ЯМР основан на поглощении электромагнитной энергии ЯМР чувствительными ядрами (такими как, например, ^{1}H или ^{13}C) в сильном магнитном поле. Поскольку разные атомы в ближайшем окружении любого ядра по-разному экранируют внешнее магнитное поле, то положение резонансных линий одних и тех же ядер в зависимости от локального окружения различается, и по величине сдвига резонансных линий можно судить о том, какие атомы и на каком расстоянии от ЯМР-чувствительных ядер они находятся. Метод позволяет решать не только задачи идентификации, как метод электрофореза, но и задачи восстановления первичной нуклеотидной и аминокислотных последовательностей, определения пространственной структуры белков (Wuthrich, 1986), и, более того, он позволяет исследовать динамические свойства молекул − константы скорости химических реакций и величины энергетических барьеров внутримолекулярного вращения (Lambert *et al.*, 2000). Тем не менее, несмотря на широкие возможности, метод ЯМР еще не стал рабочим инструментом для каждой лаборатории из-за высокой цены спектрометров и ряда присущих ему недостатков, которые включают

требование обогащения образцов ЯМР-чувствительными радиоактивными ядрами ^{13}C, ^{15}N и ^{17}O, необходимость выполнения экспериментов в растворителе, где нет водородов (D_2H, CCl_4 и др.), и низкую чувствительность. Низкая чувствительность, как следствие, приводит к необходимости иметь большой объем очищенного исследуемого вещества (не менее миллиграмм) и к длительному времени проведения эксперимента для накопления статистики.

Метод МС в отличие от ЯМР для исследования протеомов различных организмов оптимален с точки зрения соотношения «затраты/результат». Он позволяет выполнить идентификацию белков и получить количественные соотношения между различными белками (Aebersold, Mann, 2003), выявить аминокислотный состав и их последовательность в пептидных цепочках (Hubbard, Jones, 2010). Белки идентифицируются обычно в одностадийном МС-эксперименте, когда массы белков, выявленные из масс-спектра («отпечатки пальцев»), сравнивают с полученными выборкой из компьютерных баз данными последовательностей (Pappin *et al.*, 1993). Для задач секвенирования используют техники тандемной МС, в которых индивидуальные пептиды, отселектированные и накопленные после первого этапа МС, на втором этапе подвергают дальнейшей фрагментации и анализируют полученные фрагменты. Количество этапов селектирования и фрагментации обозначают надстрочным индексом *n*, например, для тандемной масс-спектрометрии принято обозначение $МС^2$ или МС/МС, для спектрометрии с большим числом этапов − обозначение $МС^n$. Одним из существенных преимуществ МС/МС над методом ЯМР являются более низкие требования к количеству исследуемого вещества (достаточно пикограмм), что весьма важно для биологических экспериментов, где стоимость получения образцов высока.

Анализ большого количества масс-спектрометрических данных, получаемых в области протеомики, – узкое место многих проектов. Современные установки могут генерировать до десятков тысяч ионных фрагментов в час. Совокупность вычислительных задач, связанных с сопоставлением этих фрагментов с пептидными последовательностями, удалением шума, идентификацией пептидных цепочек и восстановлением последовательности пептидов, представляет собой серьезный вызов для биоинформатики. В силу статистической природы экспериментальных спектров, наличия в них пропусков или ложных выбросов решение вышеупомянутых задач не является строго однозначным и может приводить к ошибочным результатам, накапливающимся в литературе и базах данных, запуская процесс их деградации и затрудняя дальнейший анализ новых данных. По этой причине актуальность разработки современных подходов, валидации баз данных, развития существующих программ и увеличение их эффективности с течением времени только увеличиваются.

Биоинформационные подходы для анализа масс-спектров

В настоящее время существующие биоинформационные подходы для анализа масс-спектрометрических данных разбиты на две категории:

- идентификация макромолекул с использованием баз данных (поиск сходства между спектром ионного фрагмента и теоретическими и/или экспериментальными спектрами пептидных цепочек, сохраненными в библиотеках);
- *de novo* секвенирование (восстановление пептидных цепочек прямым образом из МС/МС спектров).

Для больших протеомных проектов использование баз данных является основным способом идентификации образцов, другие же стратегии дают привлекательную альтернативу в особых ситуациях, например, когда исследуемый образец ранее не был зафиксирован в базах данных. Множество различных программ, поддерживающих стратегию идентификации пептидных цепочек через использование баз данных, разработано к настоящему времени (Eng *et al.*, 1994; Clauser *et al.*, 1999; Perkins *et al.*, 1999; Zhang *et al.*, 2002; Colinge *et al.*, 2003; Craig, Beavis, 2004).

Программы загружают спектр фрагмента пептида и ранжируют его относительно теоретических спектров, конструируемых для пептидов из баз данных. Количество возможных

решений ограничено согласно критериям, задаваемым пользователем, таким как точность совпадения масс фрагментов, типы разрешенных посттрансляционных модификаций и прочее. Лучшие найденные решения подвергают дальнейшему контролю методами статистического анализа (Benjamini, Hochberg, 1995; Keller *et al.*, 2002; Storey, Tibshirani, 2003; Elias, Gygi, 2007). Ряд схем ранжирования решений, описанных в литературе, включает использование спектральных корреляционных функций (Eng *et al.*, 1994), количество сходных фрагментов (Perkins *et al.*, 1999; Craig, Beavis, 2004) и статистически вычисленные частоты их встречаемости (Colinge *et al.*, 2003) или использует эмпирически подобранные правила (Dančík *et al.*, 1999), полученные с помощью технологий машинного обучения.

К настоящему времени из-за большого объема систематических исследований протеомов огромного числа организмов и большого сходства между протеомами различных организмов велика вероятность того, что исследуемый образец либо его родственные формы уже когда-либо экспериментально были изучены и занесены в ту или иную базу данных. Это позволило использовать стратегии поиска, основанные на сравнении с ранее сделанными экспериментальными данными, и разработать соответствующие программы (Craig *et al.*, 2005, 2006; Frewen *et al.*, 2006). Описанные подходы существенно более быстры, чем генерация для задач сравнения образцов теоретических спектров, и могут стать первым эффективным фильтром в задачах идентификации.

Стратегии *de novo* секвенирования (прямое восстановление первичной последовательности пептидных цепочек из масс-спектров) используют с начала 2000-х гг. (Johnson, Taylor, 2002; Ma *et al.*, 2003; Frank, Pevzner, 2005). Их главное преимущество состоит в том, что они не заменимы в тех случаях, когда либо нет информации о геномах, либо эта информация существенно ограничена, либо если подходы, основанные на поиске аналогов в базах данных, в чем-то не сработали.

Таким образом, *de novo* секвенирование может применяться к белкам, которые имеют полиморфизмы, либо к искусственно модифицированным пептидным цепочкам. В отличие от задач идентификации с поиском по базам данных, стратегии *de novo* секвенирования пептидов чрезвычайно затратны в вычислительном плане и требуют масс-спектры высочайшего качества с минимальным количеством шума.

De novo секвенирование циклических цепочек

В настоящее время наблюдается большой прогресс в создании компьютерных программ для анализа и валидации спектров МС/МС (Nesvizhskii *et al.*, 2007; Allmer, 2011). Следует обратить внимание на то, что большинство программ для решения задачи *de novo* секвенирования пептидных последовательностей работает с линейными незамкнутыми цепочками. Это связано с преобладанием подобного рода биологических макромолекул в природе. Тем не менее замкнутые в кольцо пептиды также существуют. Например, такие пептиды представляют антибиотики, синтезируемые по нерибосомальному пути, − ванкомицин, даптомицил, тироцидин и прочие. Эти антибиотики необычны тем, что их синтез в почвенных микроорганизмах не следует основному постулату молекулярной биологии «от ДНК к матричной РНК и далее к пептидной макромолекуле», они вообще не закодированы в ДНК. Вместо этого в ДНК закодированы некоторые белки (синтетазы), которые и собирают эти антибиотики (Marahiel *et al.*, 1993; Sieber, Marahiel, 2005). Подобным способом закодирована широкая область циклических пептидов, которая включает не только антибиотики, но и противоопухолевые агенты, иммунодепрессанты, токсины и множество пептидов с неизвестными функциями. Интересно также и то, что большинство пептидов, синтезируемых по нерибосомальному пути, включает нестандартные остатки, например, тироцидин включает орницин (Orn), ванкомицин включает гликозилированные остатки, причем общее число таких строительных блоков потенциально может доходить до нескольких сотен.

De novo секвенирование циклических пептидов сталкивается с новой, неожиданной на первый взгляд, дополнительной сложностью по сравнению с *de novo* секвенированием линейных пептидов − вероятность успешного восстановления последовательности компью-

терными подходами с увеличением длины последовательности становится экспоненциально малой (Jaganathan, Hassibi, 2013). Сущность этой неожиданности разъясняется только с использованием математических подходов.

В математической постановке задача секвенирования последовательности сводится к давно известной среди математиков задаче восстановления множества целых чисел из мультимножества парных расстояний между ними, причем в этой задаче различают два четко выделенных случая:

- все точки исходного множества расположены на отрезке ограниченной длины (в англоязычной литературе на этот случай ссылаются как на проблему turnpike);
- все точки исходного множества расположены на кольце ограниченной длины (проблема beltway).

Следует заметить, что к данной математической постановке сводится множество задач из кристаллографии (Millane, 1990), астрономии (Dainty, Fienup, 1987), оптики (Walther, 1963; Jaganathan *et al.*, 2013), обработки сигналов (Rabiner, Juang, 1993) и даже теории музыки (Rahn, 1994). В биоинформатике задачи, сводимые к проблемам turnpike и beltway, обнаружены в задачах картирования сайтов рестрикции ДНК (Stefik, 1978; Allison, Yee, 1988; Pandurangan, Ramesh, 2002) и *de novo* секвенирования пептидов (Chen *et al.*, 2000; Mohimani *et al.*, 2011). Ввиду высокой практической значимости в 1977 г. эти проблемы были перечислены в списке основных проблем вычислительной геометрии (Shamos, 1977). Следует отметить, что до сих пор, несмотря на более полувековой интерес математиков к решению этих проблем, нет доказательства, принадлежат ли эти задачи к классу nondeterministic polinomial (NP), т. е. к классу задач, для которых не существует решения с полиномиальной вычислительной сложностью и построение эффективного алгоритма для произвольных данных невозможно, либо все же эти задачи не относятся к классу NP и построение эффективного алгоритма возможно.

Исчерпывающее решение обозначенных проблем дано в работах Skiena с соавт. (1990) и Lemke с соавт. (2003). Было показано, что число уникальных решений для одномерной проблемы turnpike находится в диапазоне

$$1/2n^{0,8107144} \le H_1(n) \le 1/2n^{1,2324827}$$

и для проблемы beltway в диапазоне

$$\exp\left(2^{\frac{ln\, n}{ln\, ln\, n} + o(1)}\right) \le S_1(n) \le 1/2n^{n-2} \quad , \qquad (1)$$

где n – число элементов в исходной последовательности точек. Полученные формулы могут быть интерпретированы следующим образом:

- проблема turnpike скорее всего не принадлежит к строгому NP-классу; вероятность получения решения в полиномиальное время в произвольном случае высока; число случаев, для которых решение не может быть получено в полиномиальное время, экспоненциально мало;
- проблема beltway скорее всего принадлежит к строгому NP-классу; вероятность получения решения в полиномиальное время в произвольном случае экспоненциально мала; число случаев, для которых решение не может быть получено в полиномиальное время, велико.

По сути, работа Lemke с соавт. (2003) обосновала известные из практики факты: для проблемы turnpike возможно построение алгоритма с вычислительной сложностью $O(N^2)$ (Dakic, 2000) для большинства данных, для проблемы beltway такого алгоритма построить не удалось и вычислительная сложность предложенных в литературе алгоритмов для большинства данных равна $O(N^N \log N)$ (Lemke a *et al.*, 2003). Именно по этой причине для линейных пептидов (задача *de novo* секвенирования сводится к проблеме turnpike) сделан большой прогресс в области разработки программ, а для циклических пептидов (сводится к проблеме beltway) – такие программы имеются в единичных экземплярах, и их результаты не всегда приводят к убедительным выводам. Например, неудачей завершилась попытка восстановления циклических пептидов микроорганизма *Oscillatoria* sp. из МС-спектров (Ng *et al.*, 2009).

В данной работе рассмотрено узкое, но наименее разработанное в литературе подмножество задач, связанных с восстановлением циклических пептидных цепочек из масс-спектров (задачи, сводимые к проблеме beltway). Вычислительные ресурсы для решения подобных задач с ростом длины последовательности

возрастают экспоненциально. Например, если допустить, что последовательность длиной равной одному элементу восстанавливается за один машинный такт (10^{-9} с), то последовательность длиной в 10 элементов будет восстановлена за 10^{10} тактов, что равно 10 с; последовательность длиной в 15 элементов будет восстановлена за 15^{15} тактов, или за 13,8 лет; последовательность длиной в 20 элементов будет восстановлена за 20^{20} тактов, или за 3,3 млрд лет. Поскольку математически доказано, что для проблемы beltway избежать экспоненциального роста вычислительных затрат невозможно, то целью исследований в этой области биоинформатики может быть только разработка подходов, которые ограничивают скорость этого роста. Как будет показано далее, возможно существенно ограничить скорость роста и получить эффективный алгоритм, который позволяет при доступных вычислительных ресурсах восстанавливать циклические последовательности длиной до 160 элементов.

Предлагаемый путь решения проблемы beltway

Получение эффективного алгоритма для решения проблемы может быть основано на ряде ограничений:

- Ограничимся восстановлением последовательностей, состоящих из 18 заранее известных элементов из множества $\Omega = \{57^{Gly}, 71^{Ala}, 87^{Ser}, 97^{Pro}, 99^{Val}, 101^{Thr}, 103^{Cys}, 113^{Ile, Leu}, 114^{Asn}, 115^{Asp}, 128^{Gln, Lys}, 129^{Glu}, 131^{Met}, 137^{His}, 147^{Phe}, 156^{Arg}, 163^{Tyr}, 186^{Trp}\}$, которые представляют собой веса стандартных аминокислотных остатков (без учета H_2O) (Lide, 1991). Заметим, что веса аминокислотных остатков {Ile, Leu} и {Gln, Lys} совпадают, и по этой причине число элементов во множестве Ω меньше на 2, чем общее число стандартных аминокислотных остатков.
- Предположим, что используемый для восстановления последовательности масс-спектр идеален, т. е. не содержит пропусков, дубликатов и лишних элементов, характерных для реального экспериментального спектра.

Идеализация задачи является основным приближением. Реальные спектры могут содержать пропуски (недостаток чувствительности аппаратуры), лишние элементы (загрязнения образца) и дубликаты (различная вероятность разрыва пептида в том или ином месте). Все эти эффекты в настоящей работе не учтены. Основная цель работы состоит в обеспечении ограничения экспоненциального роста вычислительных затрат и достижении максимальной длины последовательности, которая может быть восстановлена при современных вычислительных ресурсах.

Для решения задачи использован следующий подход. Назовем идеальным масс-спектром S полное множество масс всех подпоследовательностей различной длины от 1 до N, образованных одно- и двукратными разрывами некоторой циклической последовательности $\{m_1, m_2, .., m_N\}$, состоящей из N элементов $m_i \in \Omega$. К примеру, такой масс-спектр включает массы подпоследовательностей длины 3: $\{m_2, m_3, m_4\}$ (образована двумя разрывами исходной циклической цепочки в позициях $m_1 \downarrow m_2$ и $m_4 \downarrow m_5$) и $\{m_N, m_1, m_2\}$ (образована разрывами в позициях $m_{N-1} \downarrow m_N$ и $m_2 \downarrow m_3$). Также спектр Ω включает N одинаковых масс $M = \Sigma_i m_i$ для подпоследовательностей $\{m_1, m_2, .., m_N\}$, $\{m_2, m_3, .., m_N, m_1\}$, .., $\{m_N, m_1, m_2, .., m_{N-1}\}$, образованных однократными разрывами циклической цепочки в позициях $m_N \downarrow m_1$, $m_1 \downarrow m_2$, .., $m_{N-1} \downarrow m_N$ соответственно.

Назовем любую n-параметрическую циклическую последовательность $\{m_1, m_2, .., m_n, M - \Sigma_i^n m_i\}$, где $n < N$ и $m_i \in \Omega$, частичным решением, если ее масс-спектр S_n является подмножеством полного масс-спектра S, $S_n \in S$. Совокупность всех частичных решений длиной $n = N$ очевидно образует полное решение исходной задачи. Следует заметить, что, в силу требования цикличности, одно и то же частичное решение любой длины $2 < n \leq N$ может быть записано в $2n$ вариантах, различающихся циклическим сдвигом элементов и/или их инверсией. Например, последовательности: $\{m_1, m_2, m_3\}$, $\{m_2, m_3, m_1\}$, $\{m_3, m_1, m_2\}$, $\{m_3, m_2, m_1\}$, $\{m_2, m_1, m_3\}$ и $\{m_1, m_3, m_2\}$ представляют собой одну и ту же циклическую последовательность.

Единственное частичное решение с нулевым числом параметров (ранга 0) тривиально. Им является последовательность {M}, которая имеет один элемент с массой, равной массе всей искомой последовательности. Частичные решения

ранга 1 строятся на основе частичного решения ранга 0 делением элемента M на две части в виде $\{m_1, M - m_1\}$, где $m_1 \in \Omega$, и сохранением в полученном множестве тех последовательностей, чей спектр S_1 является подмножеством полного спектра задачи, $S_1 \in S$. Полное число N_1 частичных решений, получаемых таким способом, ограничено числом элементов $|\Omega|$ во множестве Ω, $N_1 \le |\Omega|$. Процесс дробления последнего элемента последовательности продолжается далее, образуя частичные решения ранга 2: $\{m_1, m_2, M - m_1 - m_2\}$, ранга 3: $\{m_1, m_2, m_3, M - m_1 - m_2 - m_3\}$ и т. д. Так продолжается до тех пор, пока не будут построены частичные решения ранга N, полное множество которых и является решением задачи. Следует заметить, что полное число возможных частичных решений N_n ранга n с увеличением значения n растет экспоненциально, $N_n \le |\Omega|^n$. Некоторые существенные детали алгоритма, позволяющие снизить скорость экспоненциального роста, приведены в приложении.

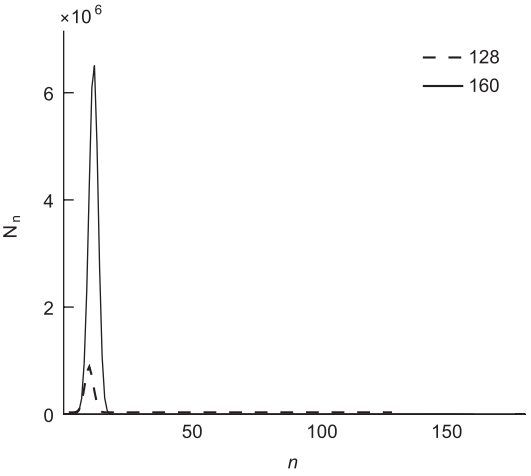


Рис. 1. Зависимость числа частичных решений разного ранга n для случайных последовательностей длиной в 128 (штриховая линия) и 160 (сплошная линия) элементов.

РЕЗУЛЬТАТЫ

На рис. 1 показано, как велико и как сильно меняется число частичных решений разного ранга на примере восстановления двух случайных последовательностей в 128 и 160 элементов. Как можно видеть, поначалу число частичных решений при увеличении длины n резко возрастает, а затем так же резко падает, образуя резкий пик при малых значениях $n \sim 10$. Величина этого пика составляет $\sim 0{,}9 \times 10^6$ для последовательности в 128 элементов и $\sim 6{,}6 \times 10^6$ для последовательности в 160 элементов. Подобный пик вполне ожидаем, так как он образован двумя противодействующими факторами: (1) экспоненциальным ростом числа решений $N_n \le |\Omega|^n$ и (2) резким падением вероятности того, что произвольная сгенерированная последовательность длины n имеет спектр, который является подмножеством заданного спектра.

На рис. 2 показано, как зависит время получения решения от числа элементов в искомой последовательности. Для получения этих данных было сгенерировано более 300 случайных последовательностей с элементами $m_i \in \Omega$ в диапазоне $n \in [30, 160]$. Ось абсцисс дана в логарифмическом масштабе. Как видно из

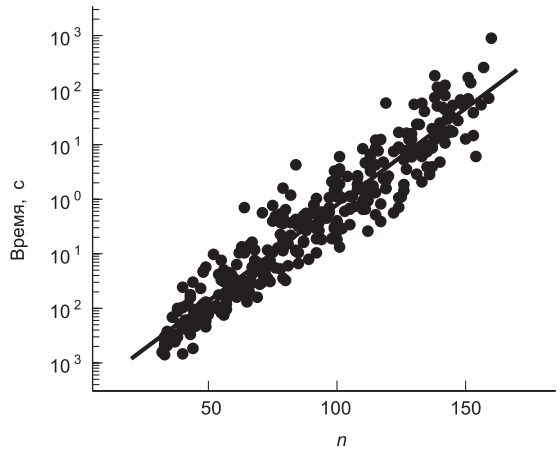


Рис. 2. Время восстановления циклической последовательности в зависимости от ее длины n.

графика, точки хорошо ложатся на прямую линию. Регрессионная прямая, дающая лучшее согласие с данными, может быть записана в виде уравнения, комбинирующего параметры $\log_{10} T$ и n: $\log_{10} T = 0{,}353n - 3{,}63$. Уравнение позволяет сделать оценку времени восстановления последовательности в зависимости от ее длины. Например, оценка для последовательности длиной в 250 элементов дает величину порядка 3 дней, а для последовательностей в 500 элементов − порядка 9 млн лет.

ВЫВОДЫ

Результаты демонстрируют, что разработанный нами алгоритм, детали которого описаны в приложении, достаточно эффективен. Он позволяет решить проблему beltway для последовательностей длиной до 160 элементов в пределах нескольких минут на персональном компьютере, что является очень хорошим результатом по ограничению экспоненциального роста для подобного рода задач. Для сравнения можно упомянуть работы (Ng *et al.*, 2009; Mohimani *et al.*, 2011), в которых решались задачи секвенирования циклических последовательностей существенно меньшей длины (~10), хотя и для реальных масс-спектров. Таким образом, полученные в данной работе результаты с существенным запасом превышают требования, возникающие в сегодняшних задачах *de novo* секвенирования. Это, в свою очередь, позволяет двигаться в направлении решения задач, интересных практически, т. е. к задачам восстановления последовательностей из реальных масс-спектров, содержащих пропуски, дубликаты и лишние данные.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке междисциплинарными интеграционными проектами СО РАН № 130, 39, 47, а также проектом фундаментальных исследований СО РАН VI.61.1.2.

ЛИТЕРАТУРА

Остерман Л.А. Методы исследования белков и нуклеиновых кислот: электрофорез и ультрацентрифугирование. М.: Наука, 1981. 286 с.

Aebersold R., Mann M. Mass spectrometry-based proteomics // Nature. 2003. V. 422. P. 198−207.

Allison L., Yee C.N. Restriction site mapping is in separation theory // Comput. Appl. Biol. Sci. 1988. V. 4. P. 97–101.

Allmer J. Algorithms for the de novo sequencing of peptides from tandem mass spectra // Expert Review of Proteomics. 2011. V. 8. No. 5. P. 645–657.

Benjamini Y., Hochberg Y. Controlling the false discovery rate – a practical and powerful approach to multiple testing // J. R. Stat. Soc. Ser. B. Methodol. 1995. V. 57. P. 289–300.

Chen T., Kao M., Tepel M., Rush J., Church G.M. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry // Proc. of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). San Francisco. CA. 2000. P. 389–398.

Clauser K.R., Baker P., Burlingame A.L. Role of accurate mass measurement (+/-10 ppm) in protein identification strategies employing MS or MS/MS and database searching // Anal. Chem. 1999. V. 71. P. 2871–2882.

Colinge J., Masselot A., Giron M. *et al.* OLAV: Towards high-throughput tandem mass spectrometry data identification // Proteomics. 2003. V. 3. P. 1454–1463.

Craig R., Beavis R.C. TANDEM: matching proteins with tandem mass spectra // Bioinformatics. 2004. V. 20. P. 1466–1467.

Craig R., Cortens J.P., Beavis R.C. The use of proteotypic peptide libraries for protein identification // Rapid Commun. Mass Spectrom. 2005. V. 19. P. 1844–1850.

Craig R., Cortens J.C., Fenyo D., Beavis R.C. Using annotated peptide mass spectrum libraries for protein identification // J. Proteome Res. 2006. V. 5. P. 1843–1849.

Dainty J.C., Fienup J.R. Phase Retrieval and Image Reconstruction for Astronomy. Image Recovery: Theory and Application, 1987. P. 231–275.

Dakic T. On the Turnpike Problem. PhD Thesis. Simon Fraser University, 2000.

Dančík V., Addona T.A., Clauser K.R., Vath J.E., Pevzner P.A. De Novo Peptide Sequencing via Tandem Mass Spectrometry // J. Computational Biology. 1999. V. 6. No. 3-4. P. 327–342.

Elias J.E., Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry // Nat. Methods. 2007. V. 4. P. 207–214.

Eng J.K., McCormack A.L., Yates J.R. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database // J. Am. Soc. Mass Spectrom. 1994. V. 5. P. 976–989.

Frank A., Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling // Anal. Chem. 2005. V. 77. P. 964–973.

Frewen B.E., Merrihew G.E., Wu C. *et al.* Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries // Anal. Chem. 2006. V. 78. P. 5678–5684.

Hubbard S.J., Jones A.R. Proteome Bioinformatics. Humana Press, 2010.

Jaganathan K., Hassibi B. Reconstruction of Integers from Pairwise Distances // Acoustics, Speech and Signal Processing (ICASSP). IEEE International Conference. 2013. P. 5974–5978.

Jaganathan K., Oymak S., Hassibi B. Sparse phase retrieval: Uniqueness guarantees and recovery algorithms. arXiv:1311.2745 [cs, math], Nov. 2013. [Online]. Available: http://arxiv.org/abs/1311.2745.

Johnson R.S., Taylor J.A. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry // Mol. Biotechnol. 2002. V. 22. P. 301–315.

Keller A., Nesvizhskii A.I., Kolker E., Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search // Anal. Chem. 2002. V. 74. P. 5383–5392.

Lambert B., Jacques V., Shivanyuk A. *et al.* Calix[4]arenes as selective extracting agents. An NMR dynamic and conformational investigation of the lanthanide (III) and thorium (IV) complexes // Inorg. Chem. 2000. V. 39. No. 10. P. 2033−2041.

Lemke P., Skiena S.S., Smith W.D. Reconstructing Sets From Interpoint Distances // Discrete Computational Geometry Algorithms Combinatorics. 2003. V. 25. P. 597–631.
Lide D.R. Handbook of Chemistry and Physics. 72nd Ed. CRC Press. Boca Raton. FL., 1991.
Ma B., Zhang K., Hendrie C. *et al.* PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry // Rapid Commun. Mass Spectrom. 2003. V. 17. P. 2337–2342.
Marahiel M.A., Nakano M.M., Zuber P. Regulation of peptide antibiotic production in Bacillus // Mol Microbiol. 1993. V. 7. No. 5. P. 631–636.
Millane R.P. Phase retrieval in crystallography and optics // J. Opt. Soc. Am. A. 1990. V. 7. No. 3. P. 394–411.
Mohimani H., Liu W.T., Yang Y.L. *et al.* Multiplex De Novo Sequen-cing of Peptide Antibiotics // J. Comp. Biol. 2011. V. 18. No. 11. P. 1371–1381.
Nesvizhskii A.I., Vitek O., Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry // Nature methods. 2007. V. 4. No. 10. P. 787–797.
Ng J., Bandeira N., Liu W. *et al.* Dereplication and de novo sequencing of nonribosomal peptides // Nat. Methods. 2009. V. 6. P. 596–599.
Pandurangan G., Ramesh H. The restriction mapping problem revisited // J. Computer System Sciences. 2002. V. 65. P. 526–544.
Pappin D.J.C., Hojrup P., Bleasby A.J. Rapid identification of proteins by peptide-mass fingerprinting Transportable // Current Biology. 1993. V. 3. P. 327−332.
Perkins D.N., Pappin D.J.C., Creasy D.M., Cottrell J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data // Electrophoresis. 1999. V. 20. P. 3551–3567.
Rabiner L., Juang B.H. Fundamentals of Speech Recognition. Signal Processing Series. Prentice Hall. 1993.
Rahn J. Possible and impossible melodies: Some formal aspects of contour // Journal Music Theory. 1994. V. 36. No. 2. P. 259–279.
Shamos M.I. Problems in computational geometry. CMU. Pittsburgh. PA, 1977.
Sieber S., Marahiel M. Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics // Chem. Rev. 2005. V. 105. P. 715–738.
Skiena S.S., Smith W.D., Lemke P. Reconstructing sets from interpoint distances // Proc. Sixth ACM Symposium Computational Geometry. Berkeley. CA, 1990. P. 332–339.
Stefik M. Inferring DNA structures from segmentation data // Artif. Intell. 1978. V. 11. P. 85–114.
Storey J.D., Tibshirani R. Statistical significance for genomewide studies // Proc. Natl. Acad. Sci. USA. 2003. V. 100. P. 9440–9445.
Walther A. The question of phase retrieval in optics // Opt. Acta. 1963. V. 10. P. 41–49.
Wuthrich K. NMR of Proteins and Nucleic Acids. John Wiley and Sons. N. Y., 1986.
Zhang N., Aebersold R., Schwilkowski B. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data // Proteomics. 2002. V. 2. P. 1406–1412.

ПРИЛОЖЕНИЕ

Детали алгоритмов

Графы частичных решений ранга n и сеть частичных решений

Построим ненаправленный граф всех частичных решений ранга 1 G_1 следующим образом. Вершинами графа являются все уникальные частичные решения $\{w_1, w_2, .., w_n\}$ ранга 1, где n – полное число решений ранга 1. Пометим вершины кортежами с тремя элементами вида $\langle w_i, \varnothing, \varnothing\rangle$, где w_i – вес соответствующей вершины. Две произвольные вершины графа $\langle w_i, \varnothing, \varnothing\rangle$ и $\langle w_j, \varnothing, \varnothing\rangle$ соединим ребром, если последовательность $\{w_i, w_j\}$ является частичным решением ранга 2. Пометим такое ребро кортежом с тремя элементами $\langle w_i + w_j, i, j\rangle$, где $w_i + w_j$ – вес данного ребра. Построенный таким образом граф может включать петли, например ребро, соединяющее вершину w_k с собой и соответствующее частичному решению $\{w_k, w_k\}$ ранга 2, но не может включать кратные ребра.

На следующем шаге мы строим ненаправленный граф G_2 всех частичных решений ранга 2. Вершинами данного графа объявляем все ребра графа G_1. Соединяем ребром любые две вершины графа G_2, если выполняются следующие условия: (1) соответствующие ребра графа G_1 являются соседями, т. е. они имеют общую вершину, (2) объединение двух частичных решений ранга 2, связанных с данными вершинами, представляет собой решение ранга 3. Любое новое ребро графа G_2 помечается кортежом с тремя элементами $\langle w_i + w_j - w_{i\cap j}, i, j\rangle$, где w_i и w_j – веса соответствующих ребер графа G_1, $w_{i\cap j}$ – вес их общей вершины $i \cap j$. Граф G_2 также может включать петли, но не может включать кратные ребра. Более того, петли в графе G_2 могут быть кратными. В самом деле, ребро $\langle w, k, m\rangle \in G_1$ может быть связано с самим собой через вершину $k \in G_1$, порождая ребро $\langle w + w - w_k, k, k\rangle$ в G_2, и через вершину $m \in G_1$, порождая ребро $\langle w + w - w_m, m, m\rangle$.

Подобным образом мы определим ненаправленный граф G_n всех частичных решений ранга n. Все вершины графа G_n образуются из ребер графа G_{n-1}. Любые две вершины $i, j \in G_n$ связываются ребром, если выполняются условия: (1) соответствующие ребра графа G_{n-1} являются

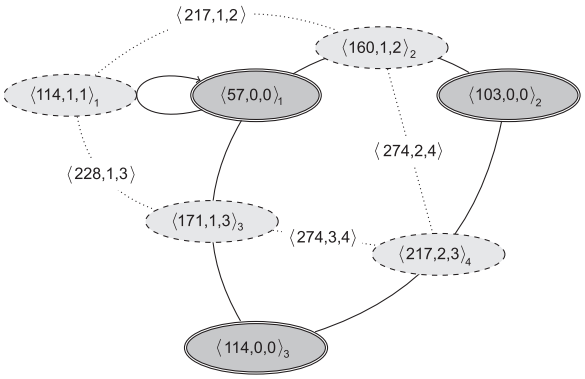


Рис. 3. Пример двух ненаправленных графов G_1 (сплошная и штриховая линии) и G_2 (штриховая и пунктирная линии) для последовательности {57, 114, 103, 57}.

соседями, т. е. имеют общую вершину, (2) объединение двух частичных решений ранга $n - 1$, связанных с данными вершинами, представляет собой решение ранга n. Вес такого ребра рассчитан по формуле $w_{i \cup j} = w_i + w_j - w_{i \cap j}$.

Заметим, что поскольку граф G_n любого ранга n строится на графе G_{n-1} предыдущего ранга $n - 1$, то совокупность всех графов образует сеть графов. На рис. 3 представлен пример объединения в сеть двух графов G_1 и G_2 для последовательности {57, 114, 103, 57}. Вершины графа G_1 выделены сплошными линиями, ребра, которые одновременно являются вершинами графа G_2, – штриховыми линиями, ребра графа G_2 – пунктирными линиями.

Оценка полного числа операций при поиске решения

Алгоритм нахождения решений основывается на последовательном построении сети из графов G_1, G_2, ... G_N. Совокупность вершин графа G_N дает полное решение поставленной задачи. Сделаем оценку полного числа операций при получении полного решения. Данное число равно $N(G_1 \cup G_2 \cup .. \cup G_N) = \Sigma N(G_k)$, где $N(G_k)$ – число операций для построения графа G_k.

Для естественного алгоритма, в котором ребра графа строятся между всеми парами вершин с дальнейшей их проверкой на то, являются ли они частичными решениями или нет, число операций $N(G_k)$ для построения ребер графа вычисляется как $N(G_k) = N_k^2\, k^2$, где N_k – число вершин в графе G_k. В этой формуле первый сомножитель N_k^2 обусловлен необходимостью проверки всех пар вершин, а второй сомножитель отражает число операций на построение спектра потенциального частичного решения длины k (аналогичное построение всех циклических пар расстояний между элементами последовательности длиной k). Общее число полученных ребер графа G_k будет равно $N(E) \le N_k^2$. Эти полученные ребра являются вершинами следующего графа G_{k+1}, что позволяет построить рекуррентную зависимость для вычисления полного числа операций построения сети. Общее число операций для «естественного» алгоритма равно $N(G_1 \cup G_2 \cup .. \cup G_N) \le \le \Sigma\, N_1^{2k}\, k^2$, где N_1 – число вершин в начальном графе G_1. Значение N_1 известно, $N_1 \le 18$, где 18 есть число элементов в последовательности весов стандартных аминокислотных остатков.

В алгоритме, который представлен в этой работе, число шагов существенно меньше, чем в «естественном» алгоритме. Это обусловлено тем, что ребра графа G_k строятся только между теми вершинами, которые являются соседями в графе меньшего ранга G_{k-1}, т. е. в графе G_{k-1} они связаны общей вершиной. Вычислительные эксперименты на большом числе случайных последовательностей показывают, что среднее число подобных соседей $<m>_k$ у любой вершины графа ранга k много меньше, чем число вершин, т. е. $<m>_k << N_k$ и обычно находится в диапазоне от 3 до 10. Это приводит к тому, что рост числа вершин при переходе от графа ранга k к графу ранга $k + 1$ перестает быть квадратичным. Другим способом ограничения числа операций является то, что спектр любого частного решения в алгоритме строится на базе уже рассчитанного спектра соседней вершины, т. е. требует не k^2 операций, а всего лишь k операций. Построив аналогичную рекуррентную зависимость для расчета полного числа операций, получим, что алгоритм требует $N(G_1 \cup G_2 \cup .. \cup G_N) \le N_1^{\,2}\, \Sigma <m>_k^{k-1}$.

Таким образом, построение сети графов частичных решений, когда каждый элемент сети строится на базе уже построенных соседних элементов, позволяет существенно снизить вычислительную сложность алгоритма от $\Sigma\, N_1^{2k}\, k^2$ до $N_1^{\,2}\, \Sigma <m>_k^{k-1}$.

RECONSTRUCTION OF AMINO ACID SEQUENCES OF CYCLIC PEPTIDES FROM THEIR MASS SPECTRA

E.S. Fomin

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

Summary

Mass spectrometry is a physical method, which can be applied to the investigation of proteomes of different organisms. It allows us both to solve the problem of identification of biological macromolecules and to sequence peptide chains in cases where information on the genomes is scarce or absent. Currently, there are many software programs to support research in this area. Nevertheless, in spite of all efforts, there is little progress in the development of programs able to solve the problem for de novo sequencing of cyclic peptides, which are most effective antibiotics, antitumor agents, immunosuppressants, toxins, and a vast number of nonribosomal peptides with unknown functions. In this paper, an effective algorithm for solving the problem of de novo sequencing cyclic peptides is proposed. The algorithm allows us to reconstruct sequences of lengths up to 160 amino acid residues.

Key words: mass spectrometry, sequencing of cyclic peptides, beltway problem.

УДК 663.15

ТЕХНОЛОГИЯ ОСАХАРИВАНИЯ БИОМАССЫ МИСКАНТУСА ПРИ ПОМОЩИ КОММЕРЧЕСКИХ ФЕРМЕНТНЫХ ПРЕПАРАТОВ

© 2014 г. **Т.Н. Горячковская[1, 2], К.Г. Старостин[1, 2], И.А. Мещерякова[1], Н.М. Слынько[1, 2], С.Е. Пельтек[1, 2]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: peltek@bionet.nsc.ru;
[2] ООО Линия солнца, Новосибирск, Россия

Поступила в редакцию 1 октября 2014 г. Принята к публикации 21 октября 2014 г.

Одним из ключевых путей снижения себестоимости биотехнологического производства является разработка для культивирования микроорганизмов дешевых субстратов, не конкурирующих с продуктами питания. В статье проанализированы возможности использования коммерчески доступных препаратов с целлюлозолитической активностью для осахаривания биомассы мискантуса сорта Сорановский – новой технической культуры, внесенной в реестр сельскохозяйственных культур РФ в 2013 г., в сравнении с осахариванием биомассы других травянистых растений – канареечника тростниковидного, кендыря ланцетовидного и сиды гермафродитной. Для ферментативного гидролиза были использованы коммерчески доступные препараты целлюлаз грибного происхождения: ксиланаза из *Thermomyces lanuginosus*, целлюлаза из *Aspergillus niger*, целлобиаза и целлюлаза из *Pen. verruculosum*. Ферментативному гидролизу предшествовала предобработка щелочной перекисью. Самой легко гидролизуемой из исследованных нами оказалась биомасса канареечника. Различными комбинациями ферментов удалось добиться 100-процентной конверсии в пересчете на массу гидролизуемых компонентов, что соответствует 70 % конверсии в пересчете на биомассу для всех образцов биомассы.

Ключевые слова: мискантус, гидролиз растительной биомассы, «зеленая химия», гликозид гидролаза, целлюлаза, ксиланаза.

ВВЕДЕНИЕ

Современный уровень науки в области молекулярной биологии обеспечивает широкомасштабное внедрение в промышленность биотехнологий. Имеющееся на планете видовое разнообразие растений представляет неисчерпаемые возобновляемые ресурсы для таких отраслей промышленности, как производство биотоплива, целлюлозы, исходных компонентов для крупнотоннажной химии. Развитие технологий «зеленой химии» приведет к существенному снижению негативного антропогенного воздействия на окружающую среду, появлению экологически-дружественных технологических процессов, рациональному природопользованию. Уже сегодня в ряде стран успешно функционирует биотехнологическое производство биоэтанола, полимолочной кислоты и 1,3-пропандиола (Erickson *et al.*, 2012). Одним из ключевых путей снижения себестоимости биотехнологического производства является разработка для культивирования микроорганизмов дешевых субстратов, не конкурирующих с продуктами питания. Интенсивное внедрение биотехнологий в промышленность в первую очередь ограничено тем, что значительные успехи в селекции микроорганизмов и ферментативных технологиях разрабатываются, как правило, без учета себестоимости получения растительной биомассы и процессов очистки целевого продукта. В понятие «биомасса» включают самые разнообразные растительные источники и даже органические отходы. Следует отметить, что при разработке технологии эффективного осахаривания растительной

биомассы необходимо ориентироваться на конкретный вид растений, тогда возможно получить значительно более стабильный и экономически эффективный технологический процесс.

В этой статье проанализированы возможности использования коммерчески доступных препаратов с целлюлозолитической активностью для осахаривания биомассы мискантуса сорта Сорановский – новой технической культуры, внесенной в реестр сельскохозяйственных культур РФ в 2013 г., с целью получения дешевых субстратов для культивирования микроорганизмов. Мискантус представляет собой быстрорастущий злак, неприхотливый к условиям выращивания. Это многолетнее растение дает стабильные урожаи биомассы 10–15 т/га. Для сравнения были взяты образцы других высокоурожайных по биомассе растений.

Гидролиз растительной биомассы можно осуществить методами химии с использованием сильных кислот и щелочей, физики – измельчение, воздействие давлением и высокой температурой, биотехнологии (ферментативный гидролиз) и микробиологии. Краеугольным камнем в вопросе деполимеризации полисахаридов клеточной стенки растений является себестоимость получения сахаросодержащего субстрата. Ферментативные реакции энергетически менее затратны и экологически безопасны, однако их эффективность в значительной степени определяется доступностью субстрата (волокна целлюлозы) и его структурой. Поэтому целью настоящей работы была разработка комбинированного процесса, включающего как механохимические предобработки, так и ферментативную стадию осахаривания предобработанной биомассы, для того чтобы обеспечить максимальную степень конверсии биомассы в сахара.

МАТЕРИАЛЫ И МЕТОДЫ

Для исследований использованы образцы биомассы мискантуса сорт Сорановский, канареечника тростниковидного, кендыря ланцетовидного и сиды гермафродитной, выращенных на экспериментальных полях ИЦиГ СО РАН, урожая 2013 г.

Мискантус сорт Сорановский – многолетнее травянистое растение, размножающееся вегетативным способом, через корневища, с прямостоячими, облиственными стеблями до 300 см высотой (урожайность зеленой массы – 75–80 т/га). Биомасса содержит 44 % целлюлозы, 23 % лигнина и 26 % гемицеллюлозы (Слынько и др., 2013).

Канареечник тростниковидный – многолетнее злаковое кормовое растение до 2 м высоты, имеет стелющиеся корни, линейные листья шириной до 2 см (содержание целлюлозы – 44,2 %, урожайность зеленой массы – 30–35 т/га).

Сида гермафродитная – растение из семейства мальвовых, рыхлокорневищное, стебли достигают высоты 300–350 см (содержание целлюлозы 40 %, урожайность зеленой массы 39–45 т/га).

Кендырь ланцетовидный – растение из семейства кутровых. Стебель высотой 80–120 см, в верхней части ветвистый (содержание целлюлозы 70 %, урожайность зеленой массы – нет данных).

Помол проводили измельчителем МАН-30 (производства ЗАО МВМ, РФ). Порошки смешивали с водой в соотношении жидкая фаза к твердой – ЖТ, мл/г, равном 10. Для ферментативного гидролиза использованы коммерчески доступные препараты: ЦеллолюксА и Целлолюкс F (НПО «Сиббиофарм»), ксиланаза из *Thermomyces lanuginosus*, целлюлаза из *Aspergillus niger* (Sigma), а также любезно предоставленные А.П. Синицыным Целлобиаза F10 и Целлюлаза B1 из *Pen. verruculosum.*

Рассеивание на фракции проводили на ротапе (шейкере-рассеивателе фракций) через сита 300 меш (с диаметром отверстий ~50 мкм), 200 меш (~71 мкм) и 150 меш (~100 мкм) со скоростью вращения 100 мин$^{-1}$ при одновременном встряхивании с частотой 180 мин$^{-1}$ в течение 20 мин. Общее количество восстанавливающих сахаров определяли колориметрическим методом с использованием 3,5-динитросалициловой кислоты (ДНСК-реагент). Долю конверсии биомассы в сахара определяли в пересчете на холоцеллюлозу из расчета содержания холоцеллюлозы в растительной биомассе 70 %.

РЕЗУЛЬТАТЫ

Порошки биомассы получали путем измельчения соломы. Следует отметить, что использование твердых добавок в процессе измельчения

приводит к уменьшению тонины помола, однако вносит балластные вещества в реакционную смесь. На рис. 1 приведен фракционный состав помолотой с различными добавками биомассы мискантуса. Добавки брали в соотношении 10 % по массе. Ранее нами было показано, что после помола с речным песком и поташем процесс ферментативного гидролиза биомассы мискантуса происходит эффективнее в 2,0 и 2,3 раза, соответственно (Слынько и др., 2013). Для ферментирования были использованы образцы биомассы мискантуса, канареечника и кендыря, помолотые без добавок. К концу вегетации стебель сиды грубеет и деревенеет, поэтому для увеличения эффективности гидролиза образец биомассы сиды был измельчен с добавкой речного песка (10 %).

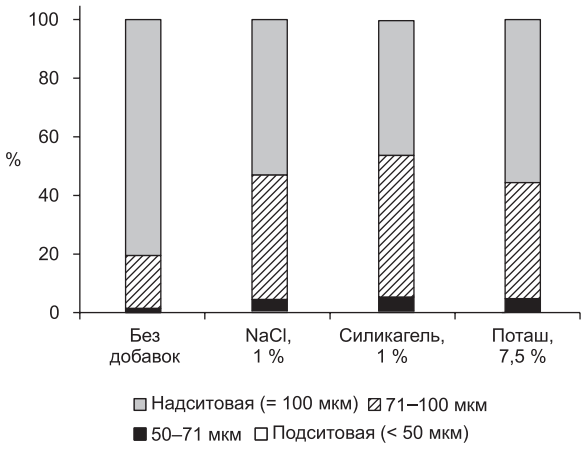


Рис. 1. Фракционный состав помолотой с различными добавками биомассы мискантуса.

После помола были последовательно проведены щелочная обработка 1 % $Ca(OH)_2$ при 100 °С, обработка щелочной перекисью, ферментативный гидролиз целлюлазами ЦеллолюксА и ЦеллолюксF. На рис. 2 приведены результаты анализа гидролизатов биомассы мискантуса на содержание восстанавливающих сахаров после обработки перекисью водорода в различных концентрациях. Исходя из приведенных на рис. 2 данных далее для предобработки биомассы использовали концентрацию перекиси водорода 4 %. Аналогичную предобработку провели для всех образцов биомассы. Ферментативный гидролиз проводили 72 ч для всех образцов. Отношение массовых долей

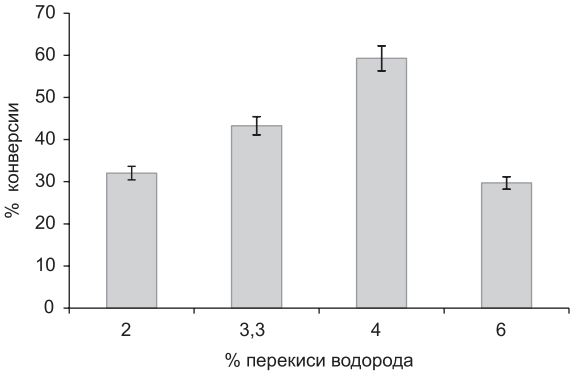


Рис. 2. Подбор концентрации перекиси водорода для обработки биомассы перед ферментативным гидролизом. Долю конверсии (в %) определяли по соотношению взятой исходно биомассы к концентрации общих восстанавливающих сахаров в гидролизате.

ферментативной смеси к биомассе и состав ферментативных смесей указаны в подписях к рисункам. Гидролиз целлюлазой Целлолюкс А (ЦелА) в комбинации с целлобиазой F10 (F10) и ксиланазой из *T. lanuginosus* (Ху) проводили при 55 °С, гидролиз целлюлазой из *A. niger* (ЦелAn) в комбинации с ксиланазой из *T. lanuginosus* (Ху) при 37 °С (рис. 3 и 4).

ОБСУЖДЕНИЕ

Гликозид гидролазы – большой класс ферментов, осуществляющих широкий спектр реакций, включая расщепление целлюлозы и гемицеллюлозы до моносахаридов (Bhalla *et al.*, 2013). Эффективный гидролиз целлюлоз требует совместного действия эндо- и экзоглюканаз, взаимодействующих с нерастворимым субстратом, и β-глюкозидаз, расщепляющих олигосахара. Эндоглюканазы случайным образом разрушают внутренние гликозидные связи, тем самым быстро увеличивая количество восстанавливающих концов цепей полисахаридов. Экзоглюканазы отщепляют олигосахара (главным образом, целлобиозу) с восстанавливающего или невосстанавливающего концов, что приводит к быстрому высвобождению олигосахаров, но медленному уменьшению длины полимера (Zhang *et al.*, 2006). Для эффективного гидролиза растительной биомассы необходим гидролиз целлобиозы, так как целлобиоза ингибирует эндо- и экзоглюканазы (Shen *et al.*,

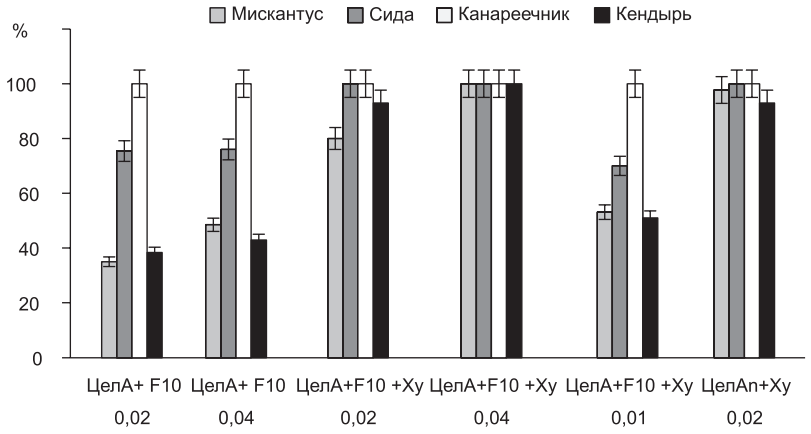


Рис. 3. Ферментативный гидролиз биомассы целлюлазой Целлолюкс А (ЦелА) в комбинации с целлобиазой F10 (F10) и ксиланазой из *T. lanuginosus* (Ху); и целлюлазой из *A. niger* (ЦелАn) в комбинации с ксиланазой из *T. lanuginosus* (Ху). По горизонтали указано массовое соотношение ферментативного комплекса и биомассы в реакционной смеси (г/г). По вертикали указаны проценты гидролизованной холоцеллюлозы.

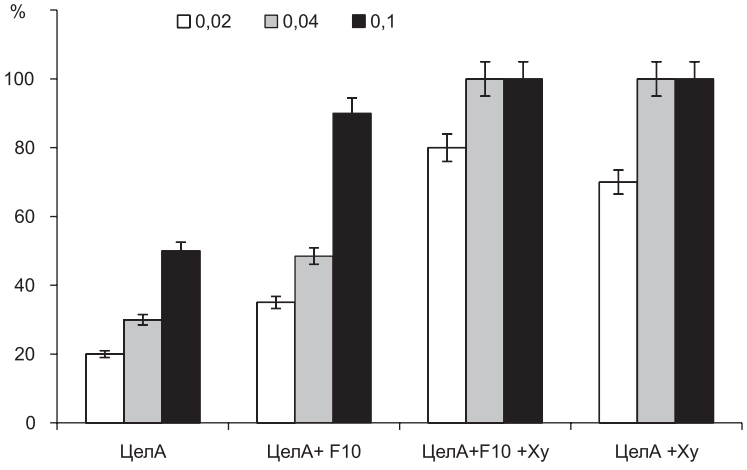


Рис. 4. Ферментативный гидролиз биомассы мискантуса целлюлазой Целлолюкс А (ЦелА) в комбинации с целлобиазой F10 (F10) и ксиланазой *T. lanuginosus* (Ху). По горизонтали указано массовое соотношение ферментативного комплекса и биомассы в реакционной смеси (г/г). По вертикали указаны проценты гидролизованной холоцеллюлозы.

2008). Расщепление целлобиозы до глюкозы осуществляется β-глюкозидазами. Именно этой активностью обладает целлобиаза F10, и ее добавление в состав реакционной смеси повышает эффективность гидролиза более чем в 1,5 раза (рис. 4). Первичная клеточная стенка растений состоит из целлюлозных фибрилл, погруженных в матрикс, в состав которого входят другие полисахариды. Лигноцеллюлозная биомасса содержит примерно 70 % полисахаридов, состоящих из остатков гексозы (целлюлоза) и пентозы (гемицеллюлоза) (Aristidou, Penttila, 2000). При полном гидролизе этих полисахаридов образуется смесь гексоз (глюкоза, галактоза, манноза) и пентоз (арабиноза, ксилоза) (Kumar *et al.*, 2008; Schädel *et al.*, 2010). Суммарное содержание целлюлозы и гемицеллюлозы (70 %) характеризует предельно возможную долю конверсии биомассы в сахара. Основной компонент гемицеллюлозы является разветвленным полимером, основа которого состоит из остатков D-ксилопираноз, соединенных β-1,4-связью. Ферментативный гидролиз гемицеллюлозы требует большого количества

активностей, в первую очередь эндо-β-1,4-ксиланазной. Для этой цели мы использовали ксиланазу из *T. lanuginosus* (Xy).

На рис. 3 видно, что обогащение реакционной смеси ксиланазой обеспечивает полный гидролиз холоцеллюлозы всех образцов биомассы. Причем для гидролиза биомассы сиды и канареечника можно исключить целлобиазу из реакционной смеси, содержащей ксиланазу. Небольшой целлобиазной активности, которой обладает один из ферментов реакционной смеси, оказалось достаточно. Для полного гидролиза биомассы мискантуса и кендыря добавка к смеси целлобиазы обязательна.

ЦеллолюксА, согласно описанию производителя, представляет собой комплексный ферментный препарат грибного происхождения, содержит в своем составе комплекс ферментов целлюлазно-глюканазно-ксиланазного действия. Однако этого комплекса недостаточно для полного гидролиза биомассы, даже при использовании соотношения фермента к биомассе 0,1 (рис. 4). Самой легко гидролизуемой из исследованных нами оказалась биомасса канареечника. Для полного гидролиза достаточно ферментативной смеси как на основе целлюлазы Целлолюкс А, так и на основе целлюлазы из *A. niger* с добавкой либо целлобиазы, либо ксиланазы, причем потребовалась минимальная из исследованных доза (см. рис. 3).

Использованные нами измельчение на мельнице и предобработка щелочной перекисью обеспечили оптимальные условия для последующего ферментативного гидролиза. Только для биомассы сиды механическая предобработка была усилена добавлением абразивного агента (песка). Основная цель предобработки – растворить гемицеллюлозу и сделать целлюлозу более доступной для ферментов. Ферментативные препараты без дополнительных предобработок не обеспечивают полный гидролиз биомассы. Так, например, В.В. Будаевой с соавт. (2013) ферментативный гидролиз пеллет из рапсовой соломы без химической предобработки позволил получить лишь 31 % конверсии биомассы. Для предобработки могут быть использованы как кислота (разбавленная или концентрированная), так и щелочь, но применение концентрированной кислоты менее привлекательно для производства по экологическим соображениям (Wyman, 1996). В зависимости от температуры процесса, в реакционной системе при кислотной предобработке могут быть обнаружены такие продукты деградации углеводных полимеров и лигнина, как фурфурол, HMF и фенольные соединения, которые ингибируют стадии ферментации (Saha *et al.*, 2005; Beg *et al.*, 2001).

Оптимальные условия для удаления лигнина из состава биомассы создает обработка щелочной перекисью, поэтому в условия щелочной предобработки ($NaOH$ / $Ca(OH)_2$) добавляется окисляющий агент – кислород или H_2O_2 (Saha, Cotta, 2006). Тайские авторы провели сравнительный анализ гидролиза биомассы 18 различных травянистых растений, произрастающих в Таиланде. Доля конверсии биомассы в сахара составила для различных трав 50–62 % в пересчете на биомассу, что соответствовало 70–80 % в пересчете на массу гидролизуемых компонентов (холоцеллюлозы). В наших экспериментах удалось добиться 100-процентной конверсии в пересчете на массу гидролизуемых компонентов, что соответствует 70 % конверсии в пересчете на биомассу.

Различают три направления развития производства биомассы: увеличение общего количества биомассы, произведенной на гектар в год, поддержание устойчивой продуктивности при минимизация затрат и увеличение количества конечных продуктов, которое может быть произведено из единицы биомассы. В качестве потенциальных энергетических растений исследуют водоросли и высшие растения. Введение в агрокультуру новых видов растений, дающих большие урожаи биомассы с высоким содержанием целлюлозы и низким содержанием лигнина, выращиваемых традиционными методами сельского хозяйства, может оказаться перспективным направлением развития агропромышленного комплекса.

БЛАГОДАРНОСТИ

Работа поддержана бюджетным проектом VI.58.1.3 и грантом фонда Сколково № МГ 4/14.

ЛИТЕРАТУРА

Будаева В.В., Макарова Е.И., Скиба Е.А., Сакович Г.В., Симирский В.В., Лисовский Д.Л., Ивашкевич О.А. Исследование кислотного и ферментативного гидролиза пеллет из рапсовой соломы // Ползуновский вестник. 2013. № 3. С. 173–179.

Слынько Н.М., Горячковская Т.Н., Шеховцов С.В., Банникова С.В., Бурмакина Н.В., Старостин К.В., Розанов А.С., Нечипоренко Н.Н., Вепрев С.Г., Шумный В.К., Колчанов Н.А., Пельтек С.Е. Биотехнологический потенциал новой технической культуры – мискантус сорт Сорановский // Вавиловский журнал генетики и селекции. 2013. Т. 17. № 4/1. С. 765–771.

Aristidou A., Penttila M. Metabolic engineering applications to renewable resource utilization // Current Opinion Biotechnology. 2000. V. 11 (2). P. 187–198.

Beg Q.K., Kapoor M., Mahajan L., Hoondal G.S. Microbial xylanases and their industrial applications: a review // Appl. Microbiol. Biotechnol. 2001. V. 56. P. 326–338.

Bhalla A., Bansal N., Kumar S., Bischoff K.M., Sani R.K. Improved lignocellulose conversion to biofuels with thermophilic bacteria and thermostable enzymes // Bioresour Technol. 2013. V. 128. P. 751–759.

Erickson B., Nelson, J.E., Winters P. Perspective on opportunities in industrial biotechnology in renewable chemicals // Biotechnol. J. 2012. V. 7. P. 176–185.

Kumar R., Singh S., Singh O.V. Bioconversion of lignocellulosic biomass: Biochemical and molecular perspectives // J. Ind. Microbiol. Biotechnol. 2008. V. 35. P. 377–391.

Saha B.C., Cotta M.A. Ethanol production from alkaline peroxide pretreated enzymatically saccharified wheat straw // Biotechnol. Prog. 2006. V. 22. P. 449–453.

Saha B.C., Iten L.B., Cotta M.A., Wu Y.V. Dilute acid pretreatment, enzymatic saccharification and fermentation of wheat straw to ethanol // Proc. Biochem. 2005. V. 40. P. 3693–3700.

Schädel C., Blöchl A., Richter A., Hoch G. Quantification and monosaccharide composition of hemicelluloses from different plant functional types // Plant Physiology Biochemistry. 2010. V. 48 (1). P. 1–8.

Shen Y., Zhang Y., Ma T., Bao X., Du F., Zhuang G., Qu Y. Simultaneous saccharification and fermentation of acid-pretreated corncobs with a recombinant Saccharomyces cerevisiae expressing b-glucosidase // Biores. Technol. 2008. V. 99. P. 5099–5103.

Wyman C.E. Handbook on bioethanol: production and utilization. Taylor Francis. Washington, 1996. P. 417.

Zhang P.Y., Himmel M.E., Mielenz J.R. Outlook for cellulase improvement, screening and selection strategies // Biotechnol. Adv. 2006. V. 24. P. 452–481.

TECHNOLOGY OF MISCANTHUS BIOMASS SACCHARIFICATION WITH COMMERCIALLY AVAILABLE ENZYMES

T.N. Goryachkovskaya[1, 2], K.V. Starostin[1, 2], I.A. Meshcheryakova[1], N.M. Slynko[1, 2], S.E. Peltek[1, 2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: peltek@bionet.nsc.ru;
[2] Sunline LLC, Novosibirsk, Russia

Summary

We analyzed the possibility of using commercially available enzymes with cellulosolytic activity for saccharification of miscanthus biomass, Soranovsky variety, a new crop registered in Russia in 2013, in comparison to the saccharification of biomasses of *Phalaris arundinacea, Thrachomitum lancifolium*, and *Sida hermaphrodita*. For enzymatic hydrolysis, we used commercially available fungal cellulases: *Thermomyces lanuginosus* xylanase, *Aspergillus niger* cellulase, and *Pen. verruculosum* cellobiase and cellulase. A biomass was ground and incubated in alkaline peroxide. The highest rate of hydrolysis was observed with the *Phalaris arundinacea* biomass. We tested various combinations of enzymes and achieved 100 % conversion for all samples relative to the weight of hydrolyzable components, which corresponds to 70 % conversion of biomass.

Keywords: Miscanthus, hydrolysis of plant biomass, “green chemistry”, glycoside, hydrolase, cellulase, xylanase.

УДК 579.66

РЕКОМБИНАНТНЫЕ ШТАММЫ *SACCHAROMYCES CEREVISIAE* ДЛЯ ПОЛУЧЕНИЯ ЭТАНОЛА ИЗ РАСТИТЕЛЬНОЙ БИОМАССЫ

© 2014 г. **А.С. Розанов, А.В. Котенко, И.Р. Акбердин, С.Е. Пельтек**

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

Поступила в редакцию 16 октября 2014 г. Принята к публикации 23 октября 2014 г.

Saccharomyces cerevisiae является наиболее подходящим и используемым организмом для промышленного получения биоэтанола из сахаров, так как дрожжи имеют высокие темпы роста, ферментации и наработки этанола в анаэробных условиях, а также они устойчивы к высоким концентрациям этанола и низким значениям pH. Наиболее перспективным источником сахаров считается лигноцеллюлозная биомасса. Сахара, полученные из лигноцеллюлозной биомассы, являются смесью гексоз и пентоз. Однако используемые штаммы *S. cerevisiae* слабо приспособлены к сбраживанию пентасахаридов, в связи с чем необходима оптимизация метаболизма существующих в настоящее время продуцентов биоэтанола, направленная на использование пентасахаров. В работе представлен обзор существующих в мире подходов, разработанных для решения этой задачи с помощью рекомбинантных штаммов *S. cerevisiae*.

Ключевые слова: *Saccharomyces cerevisiae,* лигноцеллюлозная биомасса, утилизация ксилозы, биоэтанол, штаммы-продуценты, генетическая модификация.

ВВЕДЕНИЕ

Использование дрожжей вида *Saccharomyces cerevisiae* считается крайне перспективным подходом для преобразования растительной биомассы в жидкое топливо, в основном биоэтанол. Кроме того, в настоящее время в мире активно разрабатываются модифицированные дрожжи вида *S. cerevisiae*, способные производить и другие продукты, кроме биоэтанола, такие как бутанол, молочная кислота и сукцинат (Steen *et al*., 2008; Jayaram *et al.*, 2014; Mimitsuka *et al*., 2014). В плане получения топливного этанола из растительной биомассы основные усилия исследовательских групп направлены на создание процесса получения биоэтанола низкой стоимости из лигноцеллюлозной биомассы. Согласно расчетам, таким характеристикам могут удовлетворять технологии, в которых часть этапов, например, осахаривание и ферментация, выполняются одновременно (Ojeda *et al*., 2011). Несмотря на развитие знаний и технологий, направленных на разработку продуцентов микробиального происхождения, *S. cerevisiae* остаются наиболее перспективным и востребованным в этом направлении видом, в связи с чем актуален вопрос их модификации для формирования на основе рекомбинантных штаммов интегрированных биотехнологических процессов. Существующие в настоящее время штаммы нуждаются в доработке следующих свойств (Geddes *et al*., 2011): использование пентасахаров, оптимизация биосинтеза этанола, наработка белков для их использования в прямой конверсии лигноцеллюлозы в этанол или для предварительного гидролиза.

Введение и оптимизация пути утилизации пентасахаров в дрожжах вида *S. cerevisiae*

Биоэтанол получают из сахаро- и крахмалосодержащего сырья. В результате сельскохозяйственной деятельности кроме сахаров,

получаемых в виде крахмала и дисахаридов, остаются отходы в виде лигноцеллюлозной биомассы, которую необходимо утилизировать. В связи с этим лигноцеллюлоза является перспективным источником сахаров с низкой стоимостью. Лигноцеллюлоза состоит из трех основных компонентов: целлюлозы, гемицеллюлозы и лигнина, из них только целлюлоза и гемицеллюлоза могут быть использованы в качестве сырого материала для получения этанола. В результате гидролиза получают смесь сахаров, основными компонентами которой являются глюкоза и ксилоза. Содержание ксилозы довольно высоко в травах и древесине, поэтому для налаживания экономически выгодного процесса переработки лигноцеллюлозной биомассы в этанол требуется эффективный продуцент этанола, способный утилизировать помимо глюкозы еще и ксилозу.

В силу специфики используемого до настоящего времени субстрата – крахмал- и сахорозосодержащих растений, были отобраны микроорганизмы, эффективно сбраживающие гексозы. Наиболее технологически эффективными продуцентами оказались *S. cerevisiae* (Lin, Tanaka, 2006), но дикие штаммы *S. cerevisiae* не способны использовать ксилозу в качестве источника углерода. На схеме изображен метаболизм ксилозы, осуществляемый грибами и бактериями. Процесс превращения ксилозы в

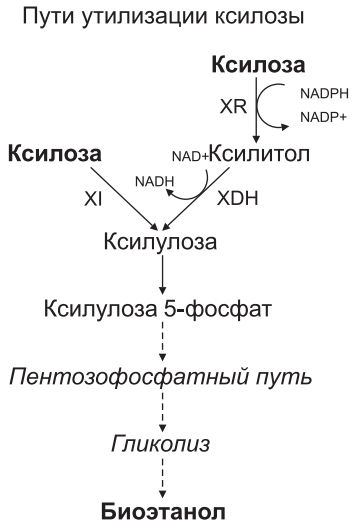


Схема включения ксилозы в метаболизм дрожжей.

ксилулозу большинства грибов и ксилозоутилизирующих дрожжей (*Pichia stipitis*, *Pachysolen tannophilus* и *Candida shehatae*) проходит в два этапа: на первом этапе работает фермент NADPH-зависимая ксилозоредуктаза (XR, EC 1.1.1.307), осуществляющий превращение ксилозы в ксилитол, на втором этапе с помощью фермента NAD^+-зависимая ксилитолдегидрогеназа (XDH, EC 1.1.1.B19) ксилитол переходит в ксилулозу. Далее фермент ксилулокиназа (XK, EC 2.7.1.17) проводит фосфорилирование ксилулозы с образованием ксилулозы-5-фосфат, дальнейший метаболизм проходит через пентозофосфатный шунт. С другой стороны, существует еще один путь утилизации ксилозы, представленный в бактериях. При этом ксилоза напрямую изомеризуется в ксилулозу с помощью фермента ксилозоизомераза (XI, EC 5.3.1.5). Далее, так же как и у грибов, ксилулоза фосфорилируется ксилулокиназой (XK) в ксилулозо-5-фосфат и переходит в пентозофосфатный путь.

Экспрессия экзогенных ферментов ксилозоредуктаза и ксилитолдегидрогеназа

В 1990 г. была впервые получена линия дрожжей *S. cerevisiae*, способная расти на среде с ксилозой в качестве единственного источника углерода (Kötter *et al*., 1990). Тогда это было достигнуто за счет генетической модификации: экспрессии ферментов XR и XDH, принадлежащих другому виду дрожжей *P. stipitis*. Таким образом, была получена линия дрожжей *S. cerevisiae*, способная перерабатывать и глюкозу, и ксилозу, но, к сожалению, концентрация целевого продукта – этанола – была низкой. Позже это было объяснено тем, что ферменты, участвующие в превращении ксилозы в ксилулозу, имеют разную коферментную специфичность. XR в качестве кофермента может использовать NADH и NADPH, а фермент XDH – только NAD^+. Поэтому возникает избыток кофермента $NADP^+$ и недостаток NAD^+. В результате, в анаэробных условиях возникает несбалансированность коферментов, приводящая к преимущественному образованию ксилитола, а не биоэтанола (Kötter, Ciriacy, 1993). К настоящему времени было предпринято множество попыток оптимизации биосинтеза этанола, основанного

на гетерологичной экспрессии ферментов XR и XDH. В одной из работ 2012 г. был получен мутантный ген, кодирующий фермент ксилозоредуктаза (XR^{MUT}) с измененной кофакторной специфичностью – вместо NADH и NADPH полученный фермент мог использовать только NADH (Lee *et al.*, 2012). Результатом исследования было создание линии дрожжей *S. cerevisiae* с повышенным уровнем экспрессии XR^{MUT}, что позволило уменьшить накопление ксилитола.

Также было показано, при делеции гена *PHO13*, кодирующего фермент, проводящий дефосфорилирование ксилулозы-5-фосфата (X5P), или мутации, приводящей к потере активности этого фермента, происходит улучшение фенотипических свойств продуцента (Kim *et al.*, 2013a).

Были описаны и другие подходы к преодолению дисбаланса коферментов. Используя инженерию белков, была получена линия дрожжей *S. cerevisiae*, содержащая мутантный фермент ксилитол дегидрогеназа (XDH) с измененным предпочтением кофермента – $NADP^+$ вместо NAD^+, что позволило уменьшить накопление ксилитола и увеличить выход этанола (Khattab *et al.*, 2013).

В рамках другого исследования с помощью методов направленного мутагенеза был получен набор рекомбинантных линий *S. cerevisiae*, содержащих новый набор мутантных генов: строго NADPH-зависимых XR и $NADP^+$-зависимых XDH, а также с увеличенным уровнем экспрессии эндогенной XK.

Еще один подход, с помощью которого удалось увеличить выход этанола на 60% – экспрессия глюкосомальной NADH-зависимой фумарат редуктазы (FRD, EC 1.3.1.6) из *Trypanosoma brucei* (Salusjärvi *et al.*, 2013). Экспрессия FRD позволяет понизить уровень накопления ксилитола, так как в результате работы этого фермента образуется NAD^+, что приводит к установлению соответствующего окислительно-восстановительного баланса.

В более ранней работе была описана линия дрожжей *S. cerevisiae*, экспрессирующая ген фермента GAPDH (EC 1.2.1.12) из *Kluyveromyces lactis*, что позволило добиться увеличения пула NADPH (Bera *et al.*, 2011). Экспрессия этого гена в клетке *S. cerevisiae* позволила уменьшить накопление ксилитола на 40 %. В настоящее время существуют подходы, основанные на определении типа скрещивания штаммов *S. cerevisiae* (Kim *et al.*, 2013b). С применением этого метода в результате скрещивания двух гаплоидных линий дрожжей была получена диплоидная гетерозиготная линия, объединяющая молекулярно-генетические характеристики двух гаплоидных линий для увеличения биосинтеза этанола.

Экспрессия экзогенного фермента ксилозоизомераза

Еще один из способов активации метаболизма ксилозы в дрожжах *S. cerevisiae* – гетерологичная экспрессия гена, кодирующего фермент ксилозоизомераза (XI), но этот путь является довольно сложным для реализации. При первых попытках активации метаболизма этим путем не удавалось добиться экспрессии функционально-активного белка, что, скорее всего, было связано с неправильным упаковкой белка, а также посттрансляционными модификациями (Matsushika *et al.*, 2009).

В 1996 г. впервые была получена линия *S. cerevisiae*, экспрессирующая функциональный белок XI бактериального происхождения *Thermus thermophiles* (Walfridsson *et al.*, 1996). Несмотря на то что экспериментально была показана наработка функционального белка, не удалось добиться его высокой активности, поэтому уровень потребления ксилозы был довольно низок.

Позже в геном *S. cerevisiae* были встроены гены фермента XI грибного происхождения: из грибов рода *Piromyces* (Kuyper *et al.*, 2003) и позже *Orpinomyces* (Madhavan *et al.*, 2009). Экспериментально была показана наработка функционального белка XI на довольно высоком уровне, но скорость роста дрожжей на ксилозе оставалась очень низка.

Также было проведено встраивание оптимизированного гена *XI* бактериального происхождения из *Burkholderia cenocepacia* (De Figueiredo *et al.*, 2013). В результате удалось добиться пятикратного увеличения уровня потребления ксилозы и приблизительно 1,5-кратного увеличения уровня наработки этанола. Стоит отметить, что не было также замечено накопления ксилитола в клетке. Полученные

данные свидетельствуют о том, что экспрессия кодон-оптимизированной ксилозоизомеразы позволяет преодолеть окислительно-восстановительный дисбаланс и перерабатывать ксилозу до ксилулозы. Выход этанола может быть ограничен следующими факторами: одиночная копия встроенного гена *xylA,* а также низкая активность нативного гена ксилулокиназы (XK) (Yu *et al*., 2011).

В 2013 г. была получена линия дрожжей *S. cerevisiae* с интегрированным геном *XI Clostridium cellulovorans*, кодируемый фермент которого был представлен на внешней поверхности клеточной стенки (Ota *et al*., 2013). Полученная линия дрожжей хорошо росла на среде, содержащей ксилозу в качестве единственного источника углерода, и напрямую продуцировала этанол из ксилозы в анаэробных условиях.

Подходы к улучшению ферментации

Помимо включения пути превращения ксилозы в ксилулозу для получения эффективного продуцента биоэтанола необходимо также провести ряд генетических модификаций центрального метаболизма дрожжей *S. cerevisiae*. Основные из них: увеличение уровня экспрессии эндогенной ксилулокиназы, встраивание транспортеров ксилозы и модификация пентозофосфатного пути.

Увеличение уровня экспрессии ксилулокиназы

Дикие штаммы дрожжей *S. cerevisiae* способны утилизировать ксилулозу – кетоизомер ксилозы, но с довольно низкой скоростью (Wang, Schneider, 1980). На первом этапе происходит фосфорилирование ксилулозы ферментом ксилулокиназа с образованием ксилулозо-5-фосфата, далее образовавшееся соединение поступает в пентозофосфатный путь. Невысокая скорость потребления ксилулозы может быть объяснена низким уровнем активности нативной ксилулокиназы, что также может оказывать негативное влияние на ферментацию ксилозы рекомбинантными штаммами *S. cerevisiae* (Deng, Ho*,* 1990).

Встраивание белков-транспортеров ксилозы

У дрожжей вида *S. cerevisiae* нет специфичных транспортеров ксилозы, и транспорт ксилозы в клетку, в основном, происходит путем диффузии через неспецифичные транспортеры гексоз, кодируемые семейством генов *HXT* (Kruckeberg, 2006). Эти транспортеры обладают значительно меньшей аффинностью к ксилозе по сравнению с глюкозой, поэтому ее потребление начинается лишь после истощения запасов глюкозы (Kötter, Ciriacy*,* 1993). Поэтому встраивание в геном дрожжей специфичных транспортеров ксилозы может оказать положительное влияние на ферментацию ксилозы.

Эксперименты по экспрессии гетерологичных транспортеров ксилозы были проведены на основе выделения следующих генов транспортеров *Gxf1*, *Sut1* и *At5g59250* из *Candida intermedia*, *Pichia stipitis* и *Arabidopsis thaliana* соответственно (Hector *et al.*, 2008; Katahira *et al.*, 2008; Runquist *et al.*, 2009). Позже был проведен их сравнительный анализ, и было показано, что линия с транспортером Gxf1 обладает наибольшей скоростью потребления ксилозы, а также и наибольшей скоростью роста (Runquist *et al.*, 2010).

В качестве альтернативного пути были предприняты попытки изменения специфичности транспортера с гексоз на ксилозу путем изменения мотива, обнаруженного в первом трансмембранном домене (Young *et al.*, 2014). Однако все же транспорт ксилозы посредством полученных мутантных белков ингибируется глюкозой.

Модификация пентозофосфатного пути

Единственный путь включения ксилулозы в гликолиз идет через пентозофосфатный путь (ПФП). Однако по сравнению с другими видами дрожжей интенсивность ПФП у *S. cerevisiae* является низкой (Fiaux J. *et al*., 2003). Поэтому для получения эффективного продуцента биоэтанола необходимо увеличить активность следующих ферментов неокислительного пентозофосфатного пути: трансальдолаза (EC 2.2.1.2), транскетолаза (EC 2.2.1.1), рибулозо-5-фосфат 3-эпимераза (EC 5.1.3.1) и рибозо-5-

фосфат кето-изомераза (EC 5.3.1.6) (Kuyper *et al*., 2005a). Для получения линии, эффективно ферментирующей ксилозу до этанола, и ее применения в промышленном производстве необходимо оптимизировать уровень экспрессии ферментов всего ферментативного пути (Lu, Jeffries, 2007).

Другие подходы по модификации дрожжей, утилизирующих ксилозу

В дополнение к методам направленной метаболической инженерии для получения линии дрожжей, утилизирующих ксилозу, также применяются методы естественной селекции и спонтанного мутагенеза (Sauer, 2001). Настоящий подход был применен и к линиям, экспрессирующим ксилозоредуктазу и ксилитолдегидрогеназу (Sonderegger, Sauer, 2003), и ксилозоизомеразу (Kuyper *et al.*, 2004).

Применение данного подхода позволяет улучшить свойства ксилозоутилизирующих линий: например, увеличить скорость потребления ксилозы (Liu, Hu, 2010); ускорить рост в анаэробных условиях (Zhou *et al.*, 2012); вывести линии, устойчивые к различным ингибирующим факторам (Çakar *et al.*, 2005; Almeida *et al.*, 2007); улучшить ферментацию смеси глюкозы и ксилозы (Kuyper *et al.*, 2005b).

Экспрессия белков, участвующих в деградации лигноцеллюлозной биомассы в дрожжах

Преобразование лигноцеллюлозного материала в биоэтанол и другие продукты с применением биотехнологических методов требует ферментативной конверсии биополимеров до моносахаров, которые могут быть ассимилированы микроорганизмами. При этом значительное увеличение стоимости конечного продукта происходит из-за стоимости используемых ферментативных препаратов. Снижения затрат на эту статью расходов можно добиться несколькими путями: повышением эффективности используемых ферментативных комплексов, наработкой ферментативных комплексов в процессе ферментации и совместным проведением ферментативного гидролиза и ферментации.

Дрожжи известны как хорошие продуценты ферментов и используются для получения рекомбинантных белков медицинского и промышленного назначения. Основной причиной популярности использования дрожжей являются легкость культивирования и достаточно активный синтез белка (Çelik, Çalik, 2012). Наиболее перспективными для получения при спиртовом брожении являются белки, используемые в ферментативном гидролизе растительной биомассы.

В этом направлении ведется значительное количество работ: были получены продуценты отдельных белков, участвующих в разрушении лигноцеллюлозной биомассы. Для изучения возможности наработки белков, необходимых для разрушения лигноцеллюлозной биомассы в дрожжах, были созданы продуценты основных ферментов, участвующих в разрушении лигноцеллюлозной биомассы: эндоглюканаза (EC 3.2.1.4) (Chen *et al.*, 2012; Mormeneo *et al.*, 2012; Wilde *et al.*, 2012), экзоглюканаза (EC 3.2.1.91) (Cho *et al.*, 1999; Ilmén *et al.*, 2011), β-глюкозидаза (EC 3.2.1.21) (Wilde *et al.*, 2012; Gurgu *et al.*, 2011), ксиланазы (EC 3.2.1.8) (Fujii *et al.*, 2011; Karaoglan *et al.*, 2014; Kirikyali, Connerton, 2014). В связи с перспективой использования дрожжей для ферментации пентасахаров в их геном были клонированы ферменты, обладающие гликолитическими активностями, направленными на разрушение гемицеллюлозы (Ahmed *et al.*, 2009).

В ходе использования дрожжей в качестве продуцентов белков, обладающих ферментативной активностью, выяснилось, что из-за гипергликозилирования, присущего многим видам дрожжей, снижается выход активного рекомбинантного белка. Этот факт послужил импульсом для проведения серии работ, посвященных исследованию мутантных штаммов с нокаутами части генов ферментов, участвующих в процессах гликозилирования в клетках *S. cerevisiae* (Kitagawa *et al.*, 2011; Wang *et al.*, 2013; Xu *et al.*, 2014). В этих работах удалось повысить наработку целевых белков за счет нокаута генов ферментов, участвующих в гликозилировании белка в процессе белкового синтеза *S. cerevisiae*. Введение в геном ферментов, обладающих функциональными активностями для разрушения растительной биомассы, имеет

основной целью не только получение штамма-продуцента, который мог бы ферментировать не отдельные сахара, а предобработанную в разной степени лигноцеллюлозу, но и также продукцию ферментативных комплексов в ходе сбраживания сахаров.

Помимо создания штаммов-продуцентов отдельных ферментов были получены продуценты нескольких ферментов целлюлозолитического комплекса. В большинстве случаев были проведены исследования гидролитических возможностей этих штаммов на различных компонентах растительной биомассы. Вработе Ван Вука с соавт. (Van Wyk *et al.*, 2010) процессивная эндоглюконаза Cel9A из *Thermobifida fusca* была экспрессирована в клетках *S. cerevisiae* совместно с белками из *Trichoderma reesei*: двумя эндоглюконазами cel5A (egII), cel7B (egI) и двумя экзоглюконазами cel6A (cbhII), cel7A (cbhI). Во всех случаях наблюдалось увеличение активности ферментов целлюлозолитического комплекса. В работе Ямада с соавт. (Yamada *et al.*, 2011) был создан штамм, эффективно экспрессирующий три основных фермента, участвующих в разложении целлюлозы. С использованием этого штамма на среде, содержащей предобработанную фосфорной кислотой целлюлозу, удалось получить 7,6 г/л биоэтанола за 72 ч ферментации.

Для изучения процесса получения спирта из микрокристаллической целлюлозы были выделены несколько штаммов, экспрессирующих: *T. aurantiacus* EGI (эндоглюконаза), *T. reesei* CBHII (экзоглюконаза) и *Aspergillus aculeatus* BGLI (β-глюкозидаза). Далее проводились исследования по их совместному культивированию. Было показано, что смесь этих линий в соотношении 6:2:1 дает в 1,3 раза больше спирта по сравнению со смесью в равных пропорциях. Данная система была неустойчива в промышленных масштабах при длительном культивировании, но очень удобна в плане изучения оптимизации пропорций ферментов (Baek *et al.*, 2012).

Эффективность работы целлюлозолитических комплексов может зависеть не только от активности отдельных субъединиц, но и от эффективности их синергии. Для проверки предположения о том, что эффективность работы ферментативного комплекса может быть повышена в результате пространственного сближения активных центров, были проведены исследования, посвященные изучению эффекта иммобилизации белковых молекул на поверхности дрожжевой клетки.

В геном *S. cerevisiae* была клонирована группа генов *A. aculeatus* β-глюкозидазы (BGL1) и *T. reesei* эндоглюканазы II (EGII) с якорным доменом GPI. В результате нарабатываемые ферменты были иммобилизованы на поверхности клетки. Для повышения эффективности работы комплекса была увеличена активность β-глюкозидазы, что привело к увеличеню целлюлозолитической активности белкового комплекса в 106 раз по сравнению с коктейлем из свободных ферментов (Matano *et al.*, 2012; Inokuma *et al.*, 2014). В работе Катахира с соавт. был разработан штамм, продуцирующий ксиланазу II (XYNII) из *T. reesei* QM9414 и β-ксилозидазу (XylA) из *Aspergillus oryzae* NiaD300, которые иммобилизовались на поверхности клетки. Основным продуктом при разрушении ксилана была ксилоза, в то время как ди- и трисахариды содержались в очень низкой концентрации (Katahira *et al.*, 2004).

Кроме введения доменов, обладающих ферментативной активностью, была изучена эффективность включения в геном *S. cerevisiae* генов вспомогательных ферментов, участвующих в процессе разложения лигноцеллюлозы. Так в работе Накатини с соавт. в дрожжах были совместно экспрессированы поверхностно закрепленные целлюлазы и “expansin-like proteins”. По сравнению с исходным вариантом, в котором экспрессировались только поверхностно закрепленные целлюлазы, получено увеличение их целлюлазной активности в 2,9 раза и увеличение выхода биоэтанола в 1,4 раза (Nakatani *et al.*, 2013).

Наиболее интересной системой для связывания доменов, участвующих в разрушении лигноцеллюлозной биомассы, в настоящее время является целлюлосома. В работе Гойяла с соавт. (Goyal *et al.*, 2011) разработан консорциум штаммов, продуцирующих ферменты: эндоглюконазу, экзоглюконазу, β-глюкозидазу и скаффолдинг протеин. Полученные ферменты самоорганизовывались в комплекс – миницеллюлосому и закреплялись на клетках дрожжей. Исследования полученного ферментативного

комплекса показали, что организация миницеллюлосомы в три раза увеличивает гидролиз целлюлозы и увеличивает выход спирта. Были проведены исследования по получению гемицеллюлозо-разрушающих миницеллюлосом на поверхности дрожжевых клеток. Для этого разработаны штаммы *S. cerevisiae*, в геном которых был включен ген скаффолдинг протеина, содержащие от одного до трех типов кохезинов, а также были клонированы химерные белки, содержащие соответствующие С-концевые докерины (Sun *et al.*, 2012).

Как для продукции активных ферментов, так и для получения функциональных докеринов необходимо снижение гликозилазной активности синтетического аппарата дрожжей. Был проведен скрининг штаммов, дефицитных по отдельным генам, обеспечивающих гликозилирование в клетках *S. cerevisiae.* Показано, что для некоторых мутантов характерно повышение количества сформированных целлюлосом (Suzuki *et al.*, 2012). Использование дрожжей как потенциальных продуцентов сопряжено с рядом трудностей, но, несмотря на это, исследователи разрабатывают все более эффективные продуценты белковых комплексов, направленных на деградацию растительной биомассы с использованием *S. cerevisiae.*

ЗАКЛЮЧЕНИЕ

S. cerevisiae является одним из наиболее изученных, с точки зрения молекулярно-генетических механизмов регуляции метаболизма, модельных объектов современной биологии. Дрожжи обладают явным преимуществом по сравнению с другими модельными объектами – наличием функциональных свойств, позволяющих их легко культивировать в условиях суспензионной культуры, несмотря на то что они являются эукариотическими организмами (Geddes *et al*., 2011). При этом *S. cerevisiae* зарекомендовали себя как устойчивый микроорганизм, способный существовать в жестких условиях технологических процессов (Geddes *et al*., 2011). Как продуцент этанола *S. cerevisiae* остаются абсолютным лидером по эффективности наработки целевого продукта, что и послужило причиной многочисленных попыток создания на их основе суперпродуцента, способного к использованию растительной биомассы или сахаров из нее для биосинтеза этанола (Geddes *et al*., 2011).

В литературе приведены многочисленные примеры экспериментальных подходов, направленных на изменение отдельных свойств дрожжей при помощи модификации их генома, и результатов их применения, примеры этих работ представлены выше. Достаточно продвинулся как методический, так и практический уровень исследований по изменению субстратной специфичности штаммов; получены новые варианты штаммов *S. cerevisiae,* требующие, конечно, дальнейшего развития и оптимизации, но тем не менее способные полноценно ферментировать ксилозу (Ota *et al*., 2013; Kim *et al*., 2013b). Несмотря на достигнутые результаты в этой области, вопрос о способности модифицированных вариантов штаммов дрожжей к эффективному гидролизу растительной биомассы остается открытым. В ряде работ показано, что принципиально дрожжи могут быть генетически модифицированы для того, чтобы получить способность к ферментации растительной биомассы (Baek *et al.*, 2012). Однако в настоящее время не найдены оптимальные пути, которые позволили бы создать соответствующий штамм для промышленного производства биоэтанола, в том числе. Безусловно, можно утверждать, что это направление будет активно развиваться на основе увеличения масштаба вносимых в геном дрожжей изменений и развития соответствующей теоретической базы для моделирования метаболизма *S. cerevisiae* с учетом вносимых в геном регуляторных модификаций.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке бюджетного проекта VI.61.1.2.

ЛИТЕРАТУРА

Ahmed S., Riaz S., Jamil A. Molecular cloning of fungal xylanases: an overview // Applied Microbiology Biotechnology. 2009. V. 84. No. 1. P. 19–35.

Almeida J.R., Modig T., Petersson A. *et al*. Increased tolerance and conversion of inhibitors in lignocellulosic hydrolysates by *Saccharomyces cerevisiae* // J. Chemical Technology biotechnology. 2007. V. 82. No. 4. P. 340–349.

Baek S.H., Kim S., Lee K. *et al*. Cellulosic ethanol production by combination of cellulase-displaying yeast cells // Enzyme Microbial Technology. 2012. V. 51. No. 6. P. 366–372.

Bera A., Ho N., Khan A., Sedlak M. A genetic overhaul of *Saccharomyces cerevisiae* 424A (LNH-ST) to improve xylose fermentation // J. industrial microbiology biotechnology. 2011. V. 38. No. 5. P. 617–626.

Çakar Z., Seker U., Tamerler C. *et al*. Evolutionary engineering of multiple-stress resistant *Saccharomyces cerevisiae* // FEMS yeast research. 2005. V. 5. No. 6-7. P. 569–578.

Çakar Z., Turanlı Y., Alkım C., Yılmaz Ü. Evolutionary engineering of *Saccharomyces cerevisiae* for improved industrially important properties // FEMS yeast research. 2012. V. 12. No. 2. P. 171–182.

Çelik E., Çalık P. Production of recombinant proteins by yeast cells // Biotechnology advances. 2012. V. 30. No. 5. P. 1108–1118.

Chen X., Meng K., Shi P. *et al*. High-level expression of a novel Penicillium endo-1, 3 (4)-β-d-glucanase with high specific activity in *Pichia pastoris* // J. industrial microbiology biotechnology. 2012. V. 39. No. 6. P. 869–876.

Cho K.M., Yoo Y.J., Kang H.S. δ-Integration of endo/exo-glucanase and β-glucosidase genes into the yeast chromosomes for direct conversion of cellulose to ethanol // Enzyme Microbial Technology. 1999. V. 25. No. 1. P. 23–30.

De Figueiredo V., de Mello V., Reis V. *et al*. Functional expression of *Burkholderia cenocepacia* xylose isomerase in yeast increases ethanol production from a glucose–xylose blend // Bioresource Technology. 2013. V. 128. P. 792–796.

Deng X., Ho N. Xylulokinase activity in various yeasts including *Saccharomyces cerevisiae* containing the cloned xylulokinase gene // Applied Biochemistry Biotechnology. 1990. V. 24. No. 1. P. 193–199.

Fiaux J., ъakar Z.P ., Sonderegger M. *et al*. Metabolic-flux profiling of the yeasts *Saccharomyces cerevisiae* and *Pichia stipitis* // Eukaryotic cell. 2003. V. 2. No. 1. P. 170–180.

Fujii T., Yu G., Matsushika A. *et al*. Ethanol production from xylo-oligosaccharides by xylose-fermenting *Saccharomyces cerevisiae* expressing β-xylosidase // Bioscience, biotechnology, biochemistry. 2011. V. 75. No. 6. P. 1140–1146.

Geddes C.C., Nieves I.U., Ingram L.O. Advances in ethanol production // Current opinion biotechnology. 2011. V. 22. No. 3. P. 312–319.

Goyal G., Tsai S.L., Madan B. *et al*. Simultaneous cell growth and ethanol production from cellulose by an engineered yeast consortium displaying a functional mini-cellulosome // Microb. Cell Fact. 2011. V. 10. P. 89.

Gurgu L., Lafraya A., Polaina, J., Marín-Navarro J. Fermentation of cellobiose to ethanol by industrial Saccharomyces strains carrying the β-glucosidase gene (BGL 1) from *Saccharomycopsis fibuligera* // Bioresource technology. 2011. V. 102. No. 8. P. 5229–5236.

Hector R.E., Qureshi N., Hughes S. *et al*. Expression of a heterologous xylose transporter in a *Saccharomyces cerevisiae* strain engineered to utilize xylose improves aerobic xylose consumption // Applied microbiology biotechnology. 2008. V. 80. No. 4. P. 675–684.

Ilmén M., Den Haan R., Brevnova E. *et al*. High level secretion of cellobiohydrolases by *Saccharomyces cerevisiae* // Biotechnol Biofuels. 2011. V. 4. P. 30.

Inokuma K., Hasunuma T., Kondo A. Efficient yeast cell-surface display of exo-and endo-cellulase using the SED1 anchoring region and its original promoter // Biotechnology biofuels. 2014. V. 7. No. 1. P. 8.

Jayaram V., Cuyvers S., Verstrepen K. *et al*. Succinic acid in levels produced by yeast (*Saccharomyces cerevisiae*) during fermentation strongly impacts wheat bread dough properties // Food chemistry. 2014. V. 151. P. 421–428.

Karaoglan M., Yildiz H., Inan M. Screening of signal sequences for extracellular production of *Aspergillus niger* xylanase in *Pichia pastoris* // Biochemical Engineering J. 2014.

Katahira S., Fujita Y., Mizuike A. *et al*. Construction of a xylan-fermenting yeast strain through codisplay of xylanolytic enzymes on the surface of xylose-utilizing *Saccharomyces cerevisiae* cells // Applied Environmental Microbiology. 2004. V. 70. No. 9. P. 5407–5414.

Katahira S., Ito M., Takema H. *et al*. Improvement of ethanol productivity during xylose and glucose co-fermentation by xylose-assimilating *S. cerevisiae* via expression of glucose transporter Sut1 // Enzyme Microbial Technology. 2008. V. 43. No. 2. P. 115–119.

Khattab S., Saimura M., Kodaki T. Boost in bioethanol production using recombinant *Saccharomyces cerevisiae* with mutated strictly NADPH-dependent xylose reductase and NADP-dependent xylitol dehydrogenase // J. biotechnology. 2013. V. 165. No. 3. P. 153–156.

Kim S., Skerker J.M., Kang W. *et al*. Rational and evolutionary engineering approaches uncover a small set of genetic changes efficient for rapid xylose fermentation in *Saccharomyces cerevisiae* // PloS one. 2013a. V. 8. No. 2. P. e57048.

Kim S., Lee K., Kong I. *et al*. Construction of an efficient xylose-fermenting diploid *Saccharomyces cerevisiae* strain through mating of two engineered haploid strains capable of xylose assimilation // J. Biotechnology. 2013b. V. 164. No. 1. P. 105–111.

Kirikyali N., Connerton I.F. Heterologous expression and kinetic characterisation of *Neurospora crassa* β-xylosidase in *Pichia pastoris* // Enzyme microbial technology. 2014. V. 57. P. 63–68.

Kitagawa T., Kohda K., Tokuhiro K. *et al*. Identification of genes that enhance cellulase protein production in yeast // J. biotechnology. 2011. V. 151. No. 2. P. 194–203.

Kötter P., Ciriacy M. Xylose fermentation by *Saccharomyces cerevisiae* // Applied microbiology and biotechnology. 1993. V. 38. No. 6. P. 776–783.

Kötter P., Amore R., Hollenberg C.P., Ciriacy M. Isolation and characterization of the *Pichia stipitis* xylitol dehydrogenase gene, XYL2, and construction of a xylose-utilizing *Saccharomyces cerevisiae* transformant // Current genetics. 1990. V. 18. No. 6. P. 493–500.

Kruckeberg A.L. The hexose transporter family of *Saccharomyces cerevisiae* // Archives microbiology. 1996. V. 166. No. 5. P. 283–292.

Kuyper M., Harhangi, H.R., Stave A. *et al*. High level functional expression of a fungal xylose isomerase: the key to efficient ethanolic fermentation of xylose by *Saccharomyces cerevisiae*? // FEMS Yeast Research. 2003. V. 4. No. 1. P. 69–78.

Kuyper M., Hartog M., Toirkens M. *et al*. Metabolic engineering of a xylose isomerase expressing *Saccharomyces cerevisiae* strain for rapid anaerobic xylose fermentation // FEMS Yeast Research. 2005a. V. 5. No. 4-5. P. 399–409.

Kuyper M., Toirkens M., Diderich J. *et al*. Evolutionary engineering of mixed-sugar utilization by a xylose-fermenting *Saccharomyces cerevisiae* strain // FEMS Yeast Research. 2005b. V. 5. No. 10. P. 925–934.

Kuyper M., Winkler A., Dijken J., Pronk J. Minimal metabolic engineering of *Saccharomyces cerevisiae* for efficient anaerobic xylose fermentation: a proof of principle // FEMS yeast research. 2004. V. 4. No. 6. P. 655–664.

Lee S., Kodaki T., Park Y. *et al*. Effects of NADH-preferring xylose reductase expression on ethanol production from xylose in xylose-metabolizing recombinant *Saccharomyces cerevisiae* // J. Biotechnology. 2012. V. 158.

Lin Y., Tanaka S. Ethanol fermentation from biomass resources: current state and prospects // Applied microbiology biotechnology. 2006. V. 69. No. 6. P. 627–642.

Liu E., Hu Y. Construction of a xylose-fermenting *Saccharomyces cerevisiae* strain by combined approaches of genetic engineering, chemical mutagenesis and evolutionary adaptation // Biochemical Engineering J. 2010. V. 48. No. 2. P. 204–210.

Lu C., Jeffries T. Shuffling of promoters for multiple genes to optimize xylose fermentation in an engineered *Saccharomyces cerevisiae* strain // Applied environmental microbiology. 2007. V. 73. No. 19. P. 6072–6077.

Madhavan A., Tamalampudi S., Ushida K. *et al*. Xylose isomerase from polycentric fungus Orpinomyces: gene sequencing, cloning, and expression in *Saccharomyces cerevisiae* for bioconversion of xylose to ethanol // Applied microbiology biotechnology. 2009. V. 82. No. 6. P. 1067–1078.

Matano Y., Hasunuma T., Kondo A. Display of cellulases on the cell surface of *Saccharomyces cerevisiae* for high yield ethanol production from high-solid lignocellulosic biomass // Bioresource technology. 2012. V. 108. P. 128–133.

Matsushika A., Inoue H., Kodaki T., Sawayama S. Ethanol production from xylose in engineered *Saccharomyces cerevisiae* strains: current state and perspectives // Applied Microbiology Biotechnology. 2009. V. 84. No. 1. P. 37–53.

Mimitsuka T., Sawai K., Kobayashi K. *et al*. Production of d-lactic acid in a continuous membrane integrated fermentation reactor by genetically modified *Saccharomyces cerevisiae*: Enhancement in d-lactic acid carbon yield // J. bioscience bioengineering. 2014.

Mormeneo M., Pastor F., Zueco J. Efficient expression of a *Paenibacillus barcinonensis* endoglucanase in *Saccharomyces cerevisiae* // J. industrial microbiology biotechnology. 2012. V. 39. No. 1. P. 115–123.

Nakatani Y., Yamada R., Ogino C., Kondo A. Synergetic effect of yeast cell-surface expression of cellulase and expansin-like protein on direct ethanol production from cellulose // Microb. Cell Fact. 2013. V. 12. P. 66.

Ojeda K., Sánchez E., El-Halwagi M., Kafarov V. Exergy analysis and process integration of bioethanol production from acid pre-treated biomass: comparison of SHF, SSF and SSCF pathways // Chemical Engineering J. 2011. V. 176. P. 195–201.

Ota M., Sakuragi H., Morisaka H. *et al*. Display of *Clostridium cellulovorans* xylose isomerase on the cell surface of *Saccharomyces cerevisiae* and its direct application to xylose fermentation // Biotechnology Progress. 2013. V. 29. No. 2. P. 346–351.

Runquist D., Hahn-Hagerdal B., Radstrom P. Comparison of heterologous xylose transporters in recombinant *Saccharomyces cerevisiae* // Biotechnol Biofuels. 2010. V. 3. No. 5.

Runquist D., Fonseca C., Radstrцm P. *et al*. Expression of the Gxf1 transporter from *Candida intermedia* improves fermentation performance in recombinant xylose-utilizing *Saccharomyces cerevisiae* // Applied Microbiology Biotechnology. 2009. V. 82. No. 1. P. 123–130.

Salusjärvi L., Kaunisto S., Holmström S. *et al*. Overexpression of NADH-dependent fumarate reductase improves D-xylose fermentation in recombinant *Saccharomyces cerevisiae* // J. industrial microbiology biotechnology. 2013. V. 40. No. 12. P. 1383–1392.

Sauer U. Evolutionary engineering of industrially important microbial phenotypes // Metabolic Engineering. Springer Berlin Heidelberg, 2001. P. 129–169.

Sonderegger M., Sauer U. Evolutionary engineering of *Saccharomyces cerevisiae* for anaerobic growth on xylose // Applied and environmental microbiology. 2003. V. 69. No. 4. P. 1990–1998.

Steen E.J., Chan R., Prasad N. *et al*. Metabolic engineering of *Saccharomyces cerevisiae* for the production of n-butanol // Microb. Cell Fact. 2008. V. 7. No. 1. P. 36.

Sun J., Wen F., Si T. *et al*. Direct conversion of xylan to ethanol by recombinant *Saccharomyces cerevisiae* strains displaying an engineered minihemicellulosome // Applied environmental microbiology. 2012. V. 78. No. 11. P. 3837–3845.

Suzuki H., Imaeda T., Kitagawa T., Kohda K. Deglycosylation of cellulosomal enzyme enhances cellulosome assembly in *Saccharomyces cerevisiae* // J. Biotechnology. 2012. V. 157. No. 1. P. 64–70.

Van Wyk N., Den Haan R., Van Zyl W.H. Heterologous co-production of *Thermobifida fusca* Cel9A with other cellulases in *Saccharomyces cerevisiae* // Applied microbiology biotechnology. 2010. V. 87. No. 5. P. 1813–1820.

Walfridsson M., Bao X., Anderlund M. *et al*. Ethanolic fermentation of xylose with *Saccharomyces cerevisiae* harboring the *Thermus thermophilus* xylA gene, which expresses an active xylose (glucose) isomerase // Applied environmental microbiology. 1996. V. 62. No. 12. P. 4648–4654.

Wang P., Schneider H. Growth of yeasts on D-xylulose // Canadian J. microbiology. 1980. V. 26. No. 9. P. 1165–1168.

Wang T.Y., Huang, C.J., Chen H.L. *et al*. Systematic screening of glycosylation-and trafficking-associated gene knockouts in *Saccharomyces cerevisiae* identifies mutants with improved heterologous exocellulase activity and host secretion // BMC biotechnology. 2013. V. 13. No. 1. P. 71.

Wilde C., Gold N.D., Bawa N. *et al*. Expression of a library of fungal β-glucosidases in *Saccharomyces cerevisiae* for the development of a biomass fermenting strain // Applied microbiology biotechnology. 2012. V. 95. No. 3. P. 647–659.

Xu L., Shen Y., Hou J. *et al*. Promotion of Extracellular Activity of Cellobiohydrolase I from *Trichoderma reesei* by Protein Glycosylation Engineering in *Saccharomyces cerevisiae* // Curr Synthetic Sys Biol. 2014. V. 2. No. 111. P. 2332–0737.1000111.

Yamada R., Taniguchi N., Tanaka T. *et al*. Direct ethanol production from cellulosic materials using a diploid strain of *Saccharomyces cerevisiae* with optimized cellulase expression // Biotechnol Biofuels. 2011. V. 4. No. 8.

Young E.M., Tong A., Bui H. *et al*. Rewiring yeast sugar transporter preference through modifying a conserved protein motif // Proc. Natl Academy Sciences. 2014. V. 111. No. 1. P. 131–136.

Yu J., Singh D., Liu N. *et al*. Construction of a Glucose and Xylose Co-Fermenting Industrial *Saccharomyces cerevisiae* by Expression of Codon-Optimized Fungal Xylose Isomerase // J. Biobased Materials Bioenergy. 2011. V. 5. No. 3. P. 357–364.

Zhou H., Cheng J.S., Wang B.L. *et al*. Xylose isomerase overexpression along with engineering of the pentose phosphate pathway and evolutionary engineering enable rapid xylose utilization and ethanol production by *Saccharomyces cerevisiae* // Metabolic engineering. 2012. V. 14. No. 6. P. 611–622.

RECOMBINANT STRAINS OF *SACCHAROMYCES CEREVISIAE* FOR ETHANOL PRODUCTION FROM PLANT BIOMASS

A.S. Rozanov, A.V. Kotenko, I.R. Akberdin, S.E. Peltek

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

Symmary

Saccharomyces cerevisiae is the most appropriate and the most widely used model organism for industrial production of ethanol from sugars, because yeasts (1) have high rates of growth, fermentation and biosynthesis of ethanol under anaerobic conditions and (2) are tolerant of high concentrations of ethanol and low pH values. Currently, the most promising source of sugar is lignocellulosic biomass. Sugars derived from it are a mixture of hexoses and pentoses. However, *S. cerevisiae* strains in current use are poorly adapted to pentasaccharide fermentation. Therefore, it is necessary to optimize the metabolism of currently available bioethanol producers for pentasaccharide consumption. The article presents an overview of existing approaches designed to solve this problem by using recombinant *S. cerevisiae* strains.

Key words: *Saccharomyces cerevisiae,* lignocellulosic biomass, xylose utilization, bioethanol, producer strain, genetic modification.

УДК 579:66:553.988:579.222

ТЕОРЕТИЧЕСКИЕ И ПРАКТИЧЕСКИЕ АСПЕКТЫ ПРОБЛЕМЫ БИОЛОГИЧЕСКОГО ОКИСЛЕНИЯ УГЛЕВОДОРОДОВ МИКРООРГАНИЗМАМИ

© 2014 г. **А.В. Брянская, Ю.Е. Уварова, Н.М. Слынько, Е.А. Демидов, А.С. Розанов, С.Е. Пельтек**

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

Поступила в редакцию 30 сентября 2014 г. Принята к публикации 23 октября 2014 г.

В статье рассмотрены теоретические вопросы биологического окисления углеводородов нефти от алканов до полициклических ароматических углеводородов. Показаны механизмы биохимических процессов разложения компонентов нефти и сделан обзор данных, представленных в популярных базах данных. Подробно описаны результаты исследований микробных сообществ естественных нефтепроявлений кальдеры Узон. Впервые изучены экофизиологические характеристики микроорганизмов нефтедеструкторов, выделенных из термальных источников нефтепроявлений кальдеры Узон.

Ключевые слова: биологическое окисление, нефть, углеводороды, микроорганизмы, кальдера Узон.

ВВЕДЕНИЕ

Изучение особенностей биологического окисления углеводородов нефти микробными сообществами необходимо как для решения фундаментальных задач микробиологии, биохимии, экологии, так и для практического применения в области биотехнологии. Биотехнологический подход к переработке нефти позволяет устранять результаты загрязнений нефтепродуктами почвы и воды, облегчать процессы добычи и переработки нефти и получать нефтепродукты, легко утилизируемые микроорганизмами (Нуртдинова, 2005).

Местообитаниями естественных комплексов нефтеокисляющих микроорганизмов служат экстремальные экосистемы, такие как: естественные выходы и нефтепроявления на поверхности почвы, в водоемах, нефтеносные пласты почв, антропогенно-загрязненные почвы и воды. Рост микроорганизмов на нефти как единственном источнике углеводородов предполагает наличие у них соответствующих ферментных систем для деградации углеводородов и механизмов подавления токсического действия нефти. Изучение свойств микроорганизмов, утилизирующих углеводороды нефти, позволяет расширять знания о биохимии, экологии и физиологии микроорганизмов; находить новые метаболические пути деградации трудноутилизируемых субстратов; составлять карты метаболических превращений компонентов нефти; выделять и описывать свойства ферментов, разрушающих углеводороды; использовать исследуемые микроорганизмы для создания эффективных биотехнологических и биоремедиационных процессов.

Разнообразие микроорганизмов, способных к утилизации нефти, обусловлено высокой конкуренцией и большим количеством путей деградации различных фракций нефти (Тимергазина, Переходова, 2012). Микроорганизмы обладают свойством избирательного отношения к различным углеводородам; эта способность определяется различием в структуре углеводородов, а также количеством

углеродных атомов, входящих в эту структуру. В природных условиях микроорганизмы образуют консорциумы, составляя единую цепь окисления углеводородов нефти. Каждый из микроорганизмов консорциума, обладая специфичными ферментными системами, направленными на использование определенного субстрата (как самих углеводородов, так и их производных) использует данный субстрат в своем метаболизме. Поэтому при совместном воздействии микроорганизмов консорциума происходит извлечение как большего количества, так и более широкого спектра нефтяных углеводородов. В работах, посвященных процессам биологического окисления нефти и нефтепродуктов, рассмотрены преимущественно микроорганизмы, принадлежащие к родам: *Rhodococcus* (Чугунов и др., 2000; Margesin *et al.*, 2003), *Pseudomonas* (Baryshnikova *et al.*, 2001; Hamme, Ward, 2001), *Azotobacter* (Градова и др., 2003), *Bacillus* (Стабникова и др., 1995; Rahman *et al.*, 2002), *Arthrobacter* (Логинов и др., 2004), *Acinetobacter* (Hanson *et al.*, 1997), *Mycobacterium, Actinomyces, Nocardia* и др. (Андреева и др., 2006). Также в ряде работ встречается описание штаммов дрожжей, утилизирующих нефть (Суржко и др., 1995; Андреева и др., 2006). Целью данной работы было изучение ряда теоретических и практических аспектов биологического окисления углеводородов микроорганизмами на примере углеводородокисляющих организмов нефтяных полей полуострова Камчатка.

ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ ПРОБЛЕМЫ БИОЛОГИЧЕСКОГО ОКИСЛЕНИЯ УГЛЕВОДОРОДОВ МИКРООРГАНИЗМАМИ

Механизмы деградации углеводородов нефти микроорганизмами

В углеводородной части нефти выделяют основные группы: метановые (алканы, циклоалканы), нафтеновые и ароматические. К более тяжелым фракциям нефти относятся асфальтосмолистая (асфальтены, смолы) и золистая (Сафиева, 1998). В зависимости от месторождения нефть имеет различный количественный состав данных химических групп углеводородов. Так, например, бакинская нефть богата циклопарафинами и сравнительно бедна предельными углеводородами. Значительно больше предельных углеводородов в грозненской и ферганской нефти. Пермская нефть содержит большое количество ароматических углеводородов (Большаков, Бейко, 1988). При изучении процессов биологического окисления нефти в качестве субстратов выбирают вещества, преобладающие в составе нефтей различных месторождений; отдельные нефтяные фракции и вещества, доступные для деградации широким спектром групп микроорганизмов, с целью изучения различных метаболических путей.

Наиболее изучены пути деградации микроорганизмами алканов, так как это одни из самых доступных для деградации соединений, которые могут служить единственным источником углерода и энергии для сапрофитных микобактерий и родственных им организмов, для ряда видов псевдомонад, нескольких видов дрожжей и некоторых грибов (Павликова, 2004). Микробная деградация алканов возможна благодаря наличию в клетке структур, обеспечивающих поглощение гидрофобного и не растворимого в воде субстрата. Ферменты микроорганизмов, осуществляющие деградацию алканов, относятся к классу оксидоредуктаз смешанных функций (оксигеназ) и связаны с мембранными структурами клеток. Оксигеназы катализируют включение одного атома кислорода из его молекулярной формы в концевую метильную группу углеводорода (Ветрова, 2010). Углеводороды, имеющие в составе молекулы разветвленную цепь атомов, практически недоступны биохимическому окислению, так как взаимодействие «субстрат – фермент» затруднено из-за конформации молекул субстрата (Dutta, Harayama, 2001). На сегодня в базе данных KEGG представлено описание ферментных систем, использующих в качестве субстрата неразветвленные алканы, для следующих родов микроорганизмов: *Enterobacter, Serratia, Pseudomonas, Psychrobacter, Burkholderia* и др.

Циклоалканы поддаются биологическому разложению труднее алканов, что связано с наличием цикла, который окисляется сложнее, чем молекулы с линейной структурой (Суржко, 1999). Штаммы, способные деградировать циклоалканы, имеют специфические фер-

ментные системы, окисляющие циклогексан до циклогексанола, а его – до адипиновой, валериановой, муравьиной кислот. Белок, катализирующий первую реакцию, гомологичен бутанмонооксигеназе (Ветрова и др., 2013). На рис. 1 приведен пример реакции превращения циклогексана в циклогексанол под действием бутанмонооксигеназы, представленной в базе данных KEGG.

В работе Тарановой и Ждановой (1996) показано, что олефины легко окисляются микроорганизмами. Ферментные системы микроорганизмов при окислении олефинов образуют насыщенные кислородом продукты, в которых двойная связь оказывается неразрушенной. Поскольку продуктом распада является тетрадеценовая кислота, считают, что имеет место прямая атака на метильную группу олефина. Кроме этого описан еще один путь окисления олефинов – эпоксидирование двойной связи (Cooper, Goldenberg, 1987). Ароматические углеводороды наиболее токсичны для живых организмов (Емельянова, 20 09). Микроорганизмы гидроксилируют ароматические углеводороды с последующим раскрытием бензольного кольца. Субстраты, полученные в ходе подобных реакций, легко утилизируются до продуктов цикла Кребса через о- и м-расщепление (Connors, Barnsley, 1982).

Алкилированные бензолы окисляются значительно интенсивнее, чем сам бензол. При этом соединение окисляется в основном за счет боковых алкильных цепей. Наиболее изучены пути деградации толуола (Dockyu et al., 2002). На сегодня известны пять путей деградации толуола, которые начинаются либо с атаки монооксигеназы или диоксигеназы на ароматическое кольцо, либо с окисления метильной группы с последующей атакой ароматического кольца. Формирующиеся гидроксилированные ароматические углеводороды подвергаются расщеплению с разрывом ароматического кольца, образуя карбоксилированные соединения, которые при наличии в штамме ферментов дальнейшего окисления могут утилизироваться до продуктов цикла Кребса (Marchai *et al.*, 2003).

Существование большого количества путей деградации моноароматических углеводородов связано с конкурентными взаимоотношениями микроорганизмов и их адаптацией к изменяющимся условиям окружающей среды (Киреева и др., 2002). При культивировании микроорганизмов в анаэробных условиях наиболее конкурентоспособным является путь с участием толуол-2-монооксигеназы. При культивировании микроорганизмов на субстрате с малым количеством толуола наименее конкурентоспособным является путь с ферментом толуол-4-монооксигеназой. В анаэробных условиях при культивировании микроорганизмов на субстрате с малым количеством толуола наименее конкурентоспособен путь с ферментом толуол монооксигеназой (Балашова и др., 1997).

В различных базах данных (KEGG, NCBI, GenNet и др.) описаны генные сети метаболизма большого числа ароматических углеводородов. Так, в базе данных KEGG представлены генные сети и метаболиты для многих реакций деградации моноароматических углеводородов нефти микроорганизмами. В качестве примера

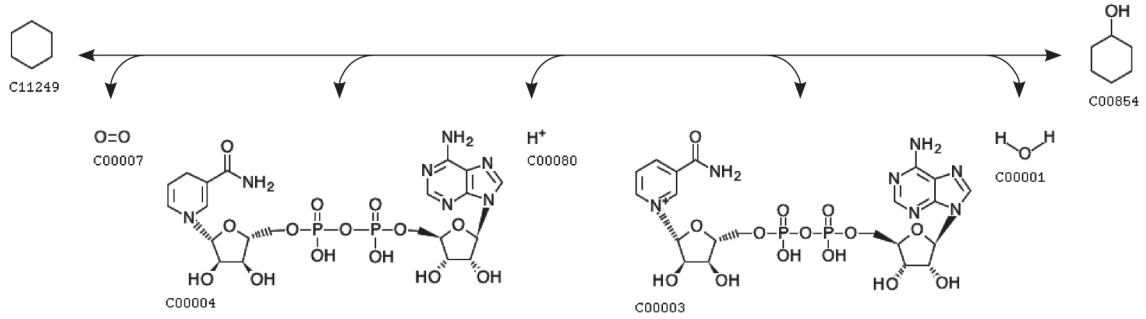


Рис. 1. Схема превращения циклогексана (C11249) в циклогексанол (C00854) под действием бутанмонооксигеназы (фермент на рисунке не указан). Реакция является NADH-зависимой. NAD^+, NADH (C00004, C00003), H^+ (C00080), H_2O (C00001), O_2 (C00007).

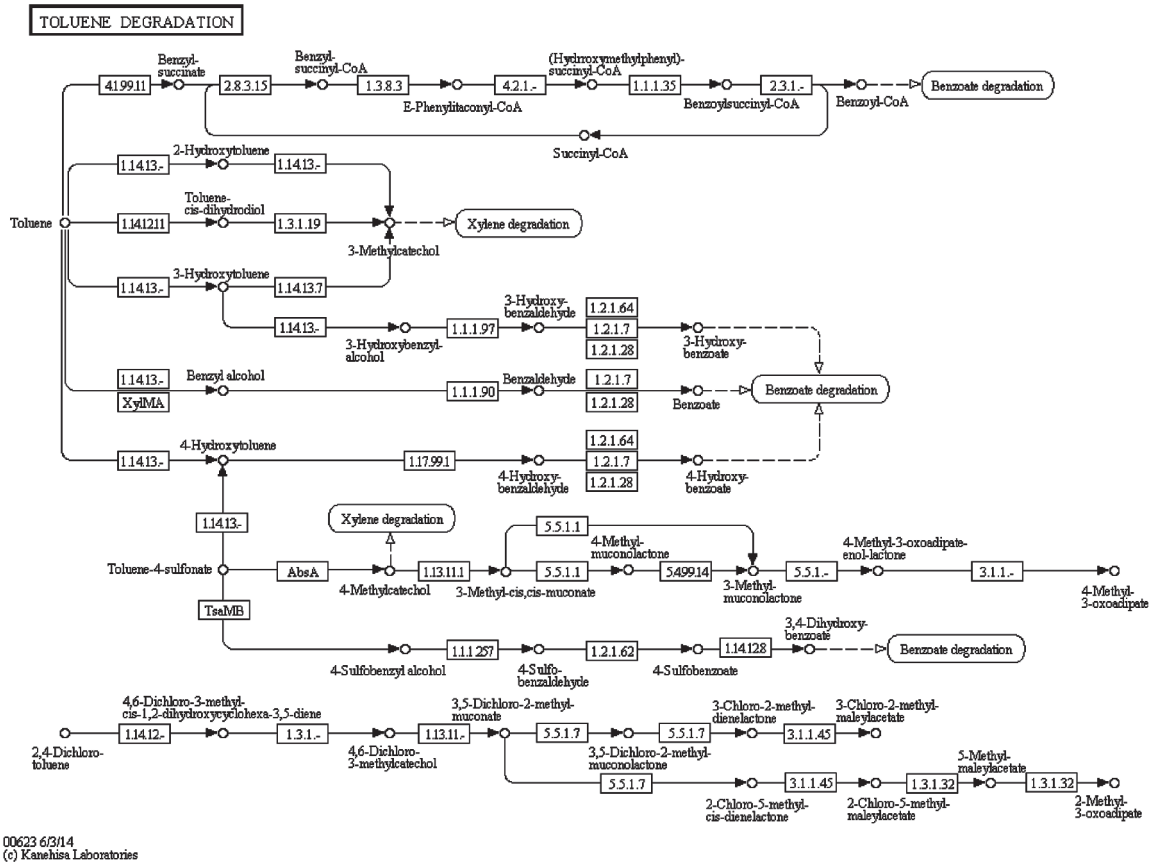


Рис. 2. Метаболическая карта преобразования толуола углеводородокисляющими организмами.

представлена карта метаболизма толуола, указаны ферменты, участвующие на каждом этапе реакции и связи с другими цепями деградации, например, бензоата (рис. 2).

Фермент бензоат/толуат диоксигеназа расщепляет циклические углеводороды, такие как камфора, толуол, салицилат, нафталин, деградирует толуол до 3-метилкатехола, который в дальнейшем поступает в сеть метаболизма ксилола.

Существуют фундаментальные различия в механизмах расщепления полициклических ароматических молекул, осуществляемых различными классами микроорганизмов (Нечаева, 2009). Бактерии и некоторые зеленые водоросли окисляют полициклические ароматические углеводороды (ПАУ), используя оба атома молекулярного кислорода (реакция катализируется диоксигеназой), при этом получается *цис*-гидродиол, который затем подвергается гидрогенизации, образуя катехол. Некоторые грибы способны окислять ПАУ с помощью цитохрома Р-450 монооксигеназ посредством включения одного из атомов молекулы кислорода в ПАУ. Скорость деградации ПАУ обратно пропорциональна числу колец в молекуле. Это связано с низкой водной растворимостью, которая снижается с увеличением числа ароматических колец. Ферментативная атака колец ПАУ происходит только в аэробных условиях (Александров, 2010).

Но некоторые ферментативные системы, такие как метан монооксидазы и лигнин пероксидазы, участвуют в анаэробном разложении ПАУ. Штаммы *Pseudomonas* и *Flavobacterium* способны окислять антрацен и фенантрен, образуя в качестве промежуточных продуктов салициловую кислоту и пирокатехин (Борзенков и др., 2006; Коршунова, Егорова, 2010). В качестве примера приведена схема деградации диоксинов углеводородокисляющими микроорганизмами, опубликованная в базе данных KEGG (рис. 3). Таким образом, в литературе имеются данные

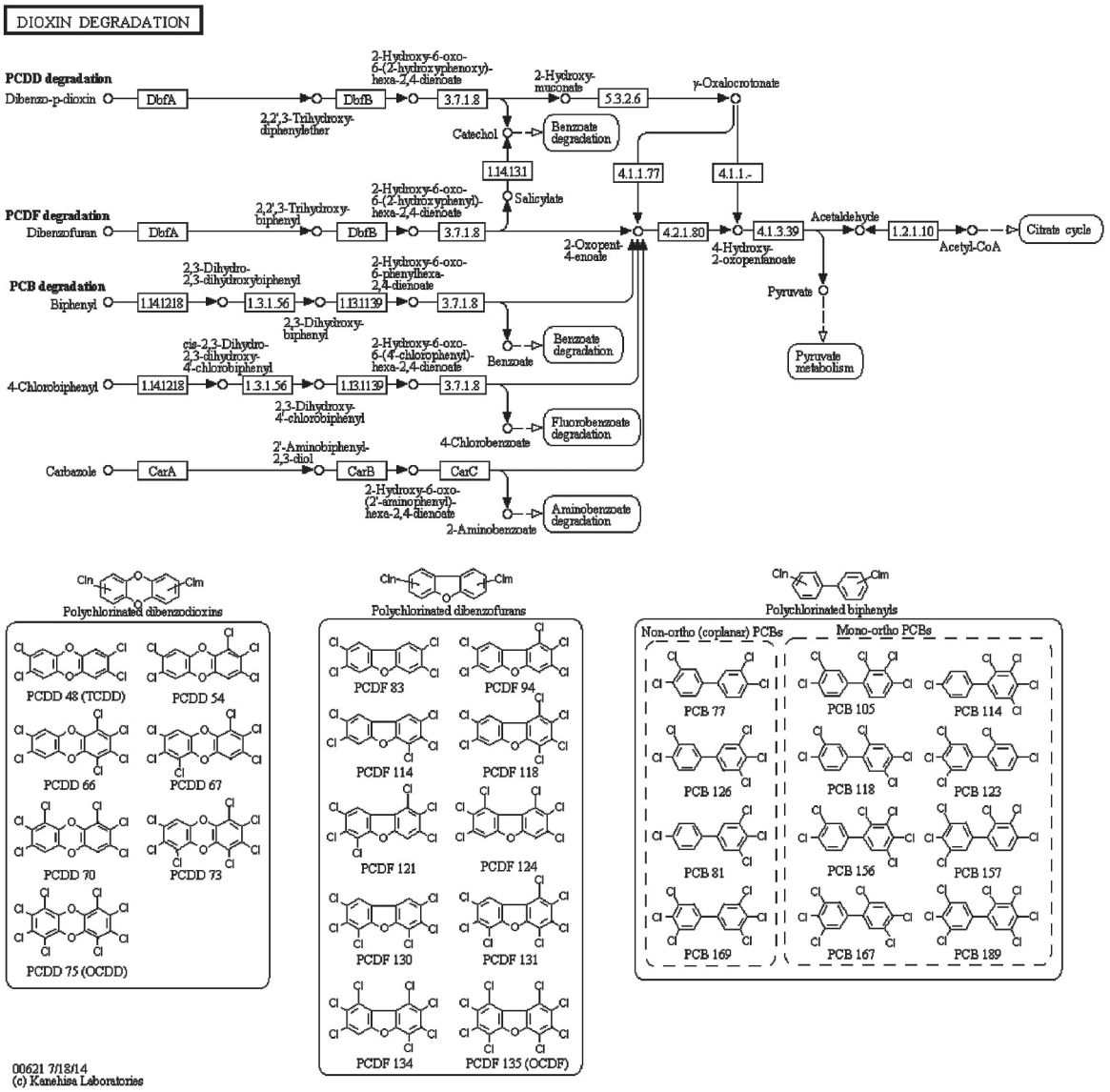


Рис. 3. Схема деградации диоксинов микроорганизмами.

о биодеградации многих компонентов нефти, таких как алканы, циклоалканы, моноароматические и полиароматические соединения. Установлены метаболические пути использования компонентов нефти такими родами микроорганизмов, как *Nocardia*, *Pseudomonas*, *Mycobacterium*, *Micrococcus, Enterobacter, Serratia, Psychrobacter, Burkholderia;* выявлены ферменты и генные сети, участвующие в метаболизме компонентов нефти, построены метаболические карты, доступные в популярных базах данных. Показано, что различные штаммы микроорганизмов имеют избирательную способность к окислению углеводородов, что объясняется большим количеством путей деградации различных фракций нефти. Так, например, выделенные и описанные J. Tausz и M. Peter *Bacterium aliphaticum* и *Bacterium aliphaticum liquefaciens* окисляли н-гексан, н-октан, декан, гексан, триконтан и тетра триоктан, а выделенная ими же *Bacterium paraffinicum* окисляла только высшие гомологи этого ряда, начиная с гексадекана (Feist, Hegeman, 1969; Кодина, 1988; Cerniglia, 1992; Кошелева и др., 2000).

ПРАКТИЧЕСКИЕ АСПЕКТЫ ПРОБЛЕМЫ БИОЛОГИЧЕСКОГО ОКИСЛЕНИЯ УГЛЕВОДОРОДОВ МИКРООРГАНИЗМАМИ

Кальдера вулкана Узон – район естественных нефтепроявлений

Уникальной экосистемой, в которой были обнаружены источники естественного выхода нефти на поверхность, является кальдера вулкана Узон (Курило-Камчатский вулканический пояс). В кальдере Узон сосредоточены многие уникальные проявления, связанные с деятельностью неглубоко залегающих магматических очагов, таких как: самостоятельные выходы нефти в источниках с сульфатно-хлоридно-кальциевым составом, скопления рудных минералов – сульфидов мышьяка, сурьмы, железа, ртути, образующихся в настоящее время (Карпов и др., 2013), высокотемпературные хлоридно-натриевые термальные источники с высоким содержанием рудных элементов – As, Sb, Hg, Au, Ag, сероводородсодержащие и углекислые источники. В приповерхностной зоне термальных полей здесь формируется оруденение As–Sb–Hg-состава, а на глубине первых сотен метров предполагается золотосеребряное оруденение (Карпов и др., 2008). Нефть кальдеры Узон имеет уникальный вещественный состав, по исследованиям НИИ вулканологии и сейсмологии РАН совместно со Швейцарским федеральным техническим университетом Цюриха, возраст узонской нефти составляет около 1 000 лет (Varfolomeev *et al.*, 2011).

В групповом составе нефтепроявлений из кальдеры вулкана Узон доминируют углеводороды (УВ) (90–93 %). Среди них по массе насыщенных УВ в 2 раза больше, чем ароматических (Конторович и др., 2011). Концентрация гетероциклических соединений составляет 7–10 %. Асфальтенов в изученных образцах очень мало (не более 0,3 %). Во фракции насыщенных углеводородов идентифицированы н-алканы C_{10}–C_{37}, алифатические изопренаны – C_{13}–C_{25}, стераны (C_{21}–C_{22} и C_{27}–C_{30}) и терпаны (C_{19}–C_{35}). Соотношение концентраций н-алканов С27 и С17 в нефти < 0,2. В составе изоалканов идентифицированы монометилалканы и изопренаны. Среди алифатических изопренанов преобладают фитан и пристан (до 53 % от суммы изопренанов). Отношение концентраций пристана к фитану (Pr/Ph) в нефтепроявлениях < 0,5. Концентрация нормальных алканов превышает изопренаны в три раза. Высокомолекулярные циклоалканы из кальдеры вулкана Узон представлены стеранами, терпанами и углеводородами гомологического ряда алкилциклогексанов. В ароматической фракции нефтепроявлений присутствуют фенантрены, метилфенантрены, моно- и триароматические стероиды, а также дибензотиофены (Бескровный и др., 1970). Среди этих соединений по концентрации преобладают триароматические стероиды (51,96–80,40 %). Концентрация фенантренов и дибензотиофенов не превышает 2,69 и 0,91 % от суммы полициклических ароматических соединений. Концентрации алкилбензолов и метилбензолов сопоставимы с концентрациями алкилциклогексанов (Лукин, Пиковский, 2004).

Состояние изученности природных комплексов нефтеокисляющих микроорганизмов Камчатки

Различные коллективы исследователей занимались изучением микрофлоры термальных полей кальдеры Узон, Долины гейзеров и других выходов термальных вод на поверхность в Курило-Камчатском вулканическом поясе (Kublanov *et al.*, 2009; Заварзин, 2010; Марданов, Равин, 2012; Бонч-Осмоловская, 2013, и др.). Была частично изучена и микрофлора районов нефтепроявлений кальдеры Узон (Mardanov *et al.*, 2009; Гумеров, 2011). Установлено огромное разнообразие микроорганизмов, населяющих эти экосистемы. В зависимости от периода изучения и методической базы данные микробные сообщества изучались классическими микробиологическими методами, молекулярно-биологическими, методами геномики, протеомики и биоинформатики (Лобкова, Лобков, 2003). В результате этих работ получены массивы данных о микроорганизмах Камчатки в целом и районов нефтепроявлений в частности (Марданов и др., 2008).

Так, например, коллективом исследователей под руководством Е.А. Бонч-Осмоловской были исследованы эколого-функциональные свойства микроорганизмов-термофилов из термальных источников Долины гейзеров и кальдеры

Узон (Perevalova *et al.*, 2005; Slepova *et al.*, 2006; Kublanov *et al.*, 2009; Bonch-Osmolovskaya *et al.*, 2011, и др.). В этих работах показано, что богатое разнообразие термопроявлений в кальдере Узон, отличающееся широким диапазоном температуры, pH, окислительно-восстановительного потенциала, солевого и микроэлементного состава воды и т. д., определяет высокое разнообразие термофильных прокариот, обитающих в этих источниках. Установлено, что в исследованных экосистемах широко распространены такие группы микроорганизмов как *Actinobacteria, Bacteroidetes, Aquificales, Deinococcus-Thermus, Thermodesulfobacteria, Verrucomicrobia, Firmicutes* и др.

Современными методами метагеномики хорошо изучен ряд микробных сообществ термальных источников кальдеры Узон, различающихся температурой и значениями pH среды. Так, в работе В.М. Гумерова (2011) проведен метагеномный анализ сообществ микроорганизмов следующих источников: Заварзин, 1884, 1810, 1805, 1807, Бурлящий (Гумеров, 2011). Более подробно остановимся на описании микробных сообществ источников 1884 и Бурлящего, которые расположены вблизи нефтяного поля кальдеры Узон.

Микробное сообщество источника 1884, представляющего собой искусственно вырытую, заполненную грунтовой водой яму, в месте выноса на поверхность углеводородов термальными водами, имело необычный состав. В нем доминировали не бактерии, а археи, составлявшие более 70 % всех микроорганизмов. Почти 90 % обнаруженных архей относились к различным линиям, не имеющим культивируемых представителей. Источник 1884 характеризовался высоким содержанием архей порядка *Fervidicoccales*. Около трети (30 %) обнаруженных последовательностей принадлежали бактериям. Доминирующими были типы *Proteobacteria (*род *Acidithiobacillus*) и *Verrucomicrobia.* Преобладание данных групп микроорганизмов обеспечивает первичную продукцию органических веществ в отсутствие фотосинтеза за счет использования субстратов вулканического происхождения, к которым относятся метан, водород и восстановленные соединения серы. Представители рода *Acidithiobacillus* окисляют неорганические соединения серы и/или металлы. *Verrucomicrobia* используют в качестве источника углерода метан. Другие группы микроорганизмов, обнаруженные в сообществе, являются органотрофами (*Fervidicoccales, Geobacillus, Actinobacteria*), или же их функциональная роль не может быть предсказана, исходя из таксономической принадлежности (некультивируемые линии бактерий и архей). Поэтому можно предположить, что в сообществе присутствуют неизвестные группы термофильных литоавтотрофов либо это сообщество зависит от притока органических веществ извне, с дождевыми водами, поступающими из окружающих более холодных районов и/или от углеводородов из глубинных слоев с геотермальным потоком.

Источник Бурлящий, расположенный рядом с основным нефтяным полем кальдеры Узон, также был подробно изучен. В высокотемпературном Бурлящем доминируют всего две группы хемолитоавтотрофных микроорганизмов: *Aquificales* среди бактерий (69 %) и *Thermoproteales* среди архей (91 %), причем последняя группа представлена почти исключительно родом *Pyrobaculum*, который отсутствовал в менее горячей точке 1884. Различия в значениях температуры и pH источников обусловливают разнообразие процессов первичной продукции и деструкции органических веществ.

Так как 1884 и Бурлящий относятся к высокотемпературным источникам и располагаются в местах выноса углеводородов на поверхность почвы, можно предположить, что микроорганизмы, составляющие данные сообщества, имеют ферментные системы, позволяющие утилизировать компоненты нефти, такие как высокомолекулярные н-алканы, циклоалканы, полициклические ароматические углеводороды, терпены, пристан, фитан и другие трудноутилизируемые соединения, и/или выживать в местообитаниях с их высокой концентрацией. Подобные микроорганизмы имеют способность осуществлять ряд биохимических превращений углеводородов нефти, который не был известен ранее. В работе Слуцкой с соавт. (2012) показано, что метагеномные исследования Камчатки позволяют находить новые группы не культивировавшихся ранее микроорганизмов с уникальными биохимическими свойствами. Из проб, взятых в источниках с очень низкими

значениями pH и высокими температурами, выделяются перспективные для биотехнологии штаммы, утилизирующие компоненты нефти при экстремальных условиях культивирования. Знания о естественных сообществах микроорганизмов, полученных в результате метагеномного анализа, позволяют более эффективно составлять консорциумы микроорганизмов для препаратов биоремедиации и биотехнологий нефтепереработки (Слуцкая и др., 2012).

ЭКОЛОГО-ФУНКЦИОНАЛЬНЫЕ АСПЕКТЫ БИОЛОГИЧЕСКОГО ОКИСЛЕНИЯ УГЛЕВОДОРОДОВ МИКРООРГАНИЗМАМИ

Выделение и характеристика нефтедеструкторов Камчатки

В 2010–2012 гг. в ИЦиГ СО РАН проводились экспедиционные работы в кальдере вулкана Узон с целью поиска и изучения природных микробных сообществ экстремальных экосистем. В период исследований было изучено более 100 различных биотопов, в том числе из района нефтепроявлений (рис. 4).

В результате исследований были отобраны образцы воды, почвы и микробных сообществ, из которых выделены в коллекцию микроорганизмов ИЦиГ СО РАН более 300 чистых и накопительных культур экстремофилов, из них – более 30 штаммов-нефтедеструкторов. Штаммы микроорганизмов-нефтедеструкторов изначально выделялись на средах с оригинальной камчатской нефтью, затем, ввиду недостаточного количества данного субстрата, были переведены на другие источники углеводородов.

В качестве минеральной среды использовали среду Ворошиловой – Дианової, следующего состава г/л: NH_4NO_3 – 1,0, K_2HPO_4 – 1,0, KH_2PO_4,– 1,0, $MgSO_4$ – 0,2, $CaCl_2$, – 0,02, $FeCl_3$ – две капли концентрированного раствора, вода дистиллированная – 1000 мл, агар 15 г. Культивирование проводили в условиях термостата при 37 °С в течение 1–21 суток. В качестве источников углеводородов использовали сырую нефть Западно-Сибирского месторождения, дизельное топливо, вазелиновое масло, скипидар. В качестве полноценной питательной среды для поддержания коллекции микроорганизмов использовали мясопептонный агар и мясопептонный бульон.

Из более чем 30 штаммов нефтедеструкторов было выделено 16 штаммов, активно растущих на сырой нефти, и 23 штамма, более активно растущих на дизельном топливе. Фенотипическую характеристику штаммов проводили на основании результатов биохимических тестов и микроскопирования. Размеры клеток и их подвижность были определены при помощи микроскопов фирмы Carl Zeiss. Филогенетическую характеристику проводили на основании анализа нуклеотидных последовательностей генов 16S рРНК, результаты секвенирования сравнивали

Рис. 4. Район нефтепроявлений (место отбора проб), кальдера вулкана Узон: слева нефтяное поле, справа источник Ящерица.

с последовательностями базы данных BLAST и на основании этого строили филогенетические деревья методом минимальной эволюции.

Для изучения способности штаммов к использованию углеводородов мясопептонный агар или среду Ворошиловой – Дианової разливали в чашки Петри. После застывания среды на поверхность наносили 50 мкл нефти либо другого источника углеводов (дизельное топливо и др.), затем в центр чашки высевали культуру. Штаммы культивировали в течение 24 ч при температуре 37 °С. После окончания культивирования измеряли диаметр зоны просветления. Диаметр зоны просветления свидетельствует о количестве углеводородов, использованных исследуемым штаммом в процессе метаболизма.

Для определения способности к росту при различных значениях pH готовили среду Ворошиловой – Дианової. После застывания среды на поверхность наносили 50 мкл нефти, в центр чашки высевали культуру. Культивировали в течение 48 ч при температуре 37 °С. Отсутствие/наличие роста оценивали визуально. Для определения способности к росту микроорганизмов на субстрате с низкими значениями pH (2 и 4) в среду добавляли H_2SO_4 до необходимого значения pH (2 или 4), добавляли 3 % нефти и культивировали 48 ч при 37 °С.

Определение способности роста на различных субстратах проводили на агаризованной среде Ворошиловой – Дианової. После застывания среды на ее поверхность наносили 50 мкл источника углеводорода (нефть, вазелиновое масло, дизельное топливо, керосин), культуру высевали штрихом. Штаммы культивировали от 7 до 21 суток при температуре 37 °С. После культивирования способность к росту на различных субстратах оценивали визуально.

Изучение ростовых характеристик штаммов на различных субстратах проводили на планшетном спектрофотометре ×Mark BioRad по оптической плотности, с интервалом в 1 ч в течение 16 ч, при длине волны 590 нм. Штаммы культивировали при температуре 37 °С на среде Ворошиловой – Дианової с добавлением дизельного топлива или вазелинового масла.

Изучение свойств естественного комплекса нефтеокисляющих микроорганизмов кальдеры Узон

Для 11 штаммов была установлена таксономическая принадлежность (рис. 5). Выявлено, что большинство штаммов, деградирующих нефть, принадлежит к роду *Bacillus*, семейству Bacillaceae, классу Bacilli, типу Firmicutes.

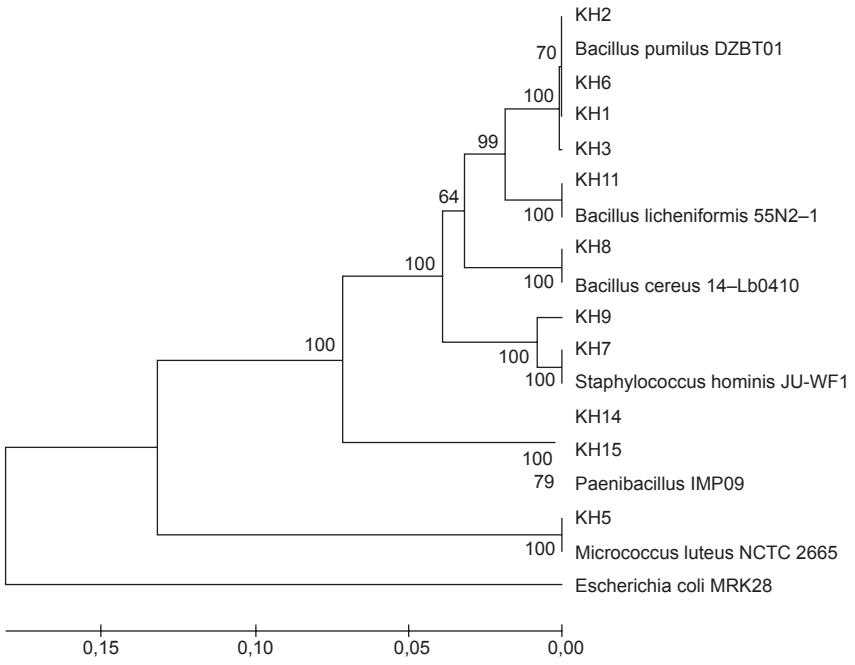


Рис. 5. Филогенетическое дерево, построенное на основании последовательностей гена 16S рРНК.

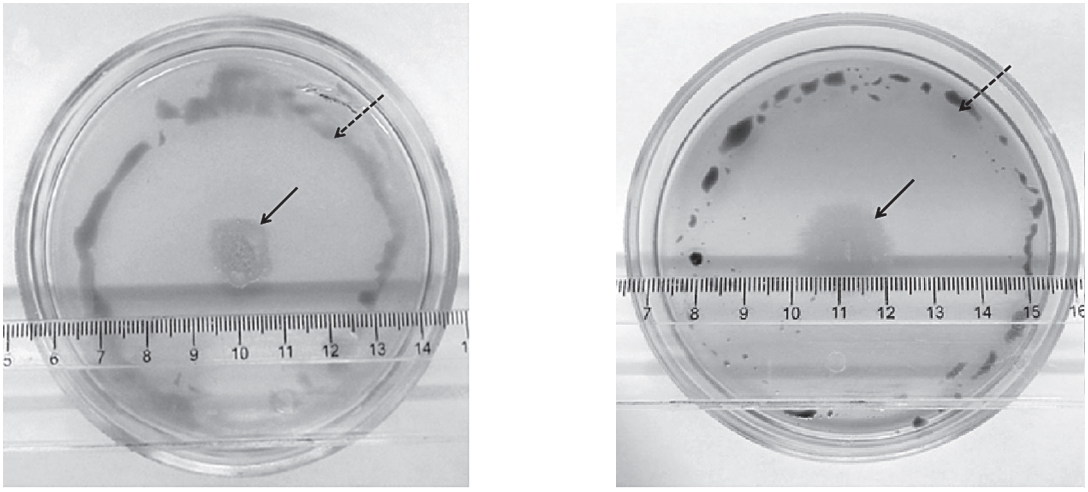


Рис. 6. Рост различных штаммов микроорганизмов (слева – КН2, справа – КН3) на сырой нефти. Сплошными стрелками показаны колонии. Штриховыми стрелками обозначена граница зоны просветления.

Была изучена способность микроорганизмов к окислению углеводородов при различных значениях температуры и pH среды. Также были проведены эксперименты по использованию микроорганизмами таких источников углерода, как нефть, дизельное топливо, вазелиновое масло, скипидар, глюкоза (рис. 6).

Наибольшие диаметры зоны просветления при различных pH наблюдали у штаммов КН2, КН3, КН6, КН10. У штаммов КН1, КН5, КН9, КН11, КН12 в данном эксперименте наблюдали меньшие диаметры зоны просветления. Штаммы КН2, КН3, КН9, КН11, КН13, КН14 имели наибольшие диаметры зоны просветления в эксперименте при культивировании микроорганизмов при различных значениях температур. В результате проведенных экспериментов установлены штаммы, обладающие способностью к интенсивной деградации нефти при высоких и низких значениях температуры культивирования и при высоких и низких значениях pH.

В ряде экспериментов была определена способность исследуемых штаммов к использованию различных источников углерода. Установлено, что пять штаммов при температуре 37 °C росли на минеральной среде Ворошиловой – Дианової с вазелиновым маслом, дизельным топливом, нефтью и глюкозой. Штаммы КН1-КН7, КН9, КН11, КН15, КН16 обладали способностью к использованию скипидара в качестве единственного источника углерода. При определении скорости роста штаммов на минеральной среде с добавлением вазелинового масла было определено, что наибольшей скоростью роста обладали штаммы КН1, КН2, КН3, КН5, КН6, КН9, КН12, КН13. Наибольшей скоростью роста на минеральной среде с добавлением дизельного топлива обладали штаммы КН1, КН5, КН7, КН9. На среде с дизельным топливом рост штаммов был слабее, чем на среде с вазелиновым маслом, потому что вазелиновое масло состоит из легкоутилизируемых парафинов, а дизельное топливо относится к более тяжелым фракциям. На графиках приведены кривые роста для штаммов КН1 и КН9, выращенных на двух различных источниках углерода (рис. 7).

ЗАКЛЮЧЕНИЕ

В настоящее время наиболее изучены процессы биологического окисления алканов и ароматических соединений, так как эти вещества могут использоваться микроорганизмами в качестве единственного источника углерода и энергии и имеют структуру, доступную для широкого спектра ферментов микроорганизмов. Наименее изучены пути деградации полициклических ароматических углеводородов, разветвленных алканов и других веществ, имеющих более сложную химическую структуру. Установлено, что в процессах деградации нефти и ее соединений участвуют различные группы микроорганизмов. В настоящее время в научной

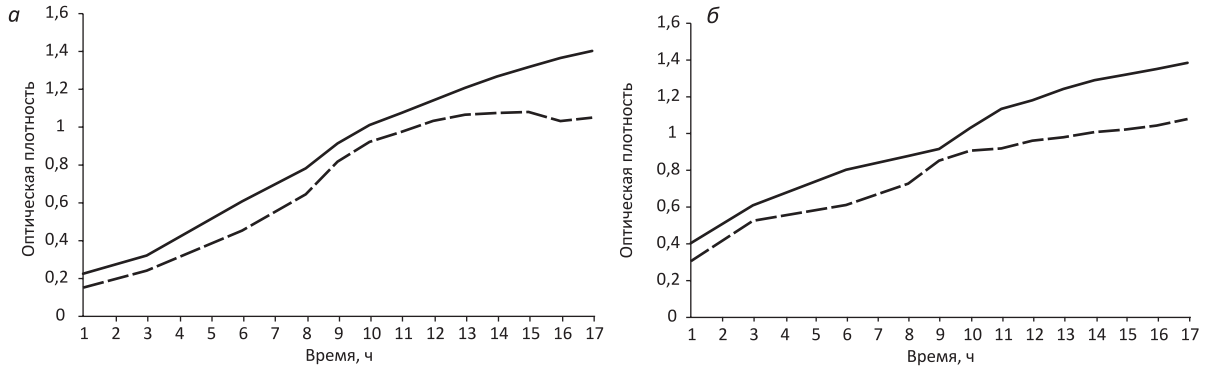


Рис. 7. Кривые роста микроорганизмов на среде Ворошиловой – Дианової с добавлением вазелинового масла (сплошная линия) и дизельного топлива (шриховая): *а* – штамм КН1, *б* – КН9.

литературе имеются значительные массивы данных о микроорганизмах Камчатки в целом и районов нефтепроявлений в частности. Исследованы эколого-функциональные свойства микроорганизмов-термофилов из термальных источников Долины гейзеров и кальдеры Узон. Показано, что богатое разнообразие термопроявлений в кальдере Узон определяет высокое разнообразие прокариот, обитающих в этих источниках. В процессе изучения естественного комплекса нефтеокисляющих микроорганизмов районов нефтепроявлений установлено, что ряд микроорганизмов обладают способностью к синтезу органического вещества за счет использования различных субстратов, например соединений серы и металлов, метана. Большинство же микроорганизмов данных систем являются органотрофами, использующими в своем метаболизме широкий спектр соединений.

Исследования микроорганизмов, выделенных из районов нефтепроявлений Камчатки, расширяют знания о процессах окисления углеводородов, позволяют проводить поиск новых путей окисления, а также возможности деградации трудноутилизируемых компонентов нефти, увеличивают объем информации о нефтедеструкторах в целом и нефтедеструкторах Камчатки в частности. В настоящей работе охарактеризованы штаммы, выделенные из районов нефтепроявлений кальдеры Узон и эффективно разлагающие нефтепродукты. Выявлены некоторые эколого-функциональные аспекты биологического окисления углеводородов исследуемыми микроорганизмами. Выделенные штаммы обладают способностью к деградации углеводородов при различных условиях культивирования, способностью к росту на различных углеводородсодержащих субстратах, способностью к росту в широком диапазоне температуры и pH, что позволяет предположить наличие у исследуемых микроорганизмов специфических ферментных систем, не изученных ранее. Исследование путей метаболизма этих штаммов является важной теоретической и прикладной задачей.

Комплексный подход к исследованиям углеводородокисляющих микробных сообществ, включающий в себя генетическую и метаболическую составляющие, важен для биотехнологии, поскольку дает более полную картину знаний об исследуемых экологических системах и процессах, а также позволяет находить наиболее эффективные пути утилизации нефтепродуктов из различных нефтяных месторождений. Проведенные исследования по изучению теоретических и практических аспектов проблемы биологического окисления углеводородов микроорганизмами-нефтедеструкторами позволят перейти к решению важных биотехнологических и биоремедиационных задач, таких как производство промышленно-важных веществ для добычи и переработки нефти, очистка нефтезагрязненных почв и вод и др.

БЛАГОДАРНОСТИ

Авторы благодарят доктора геолого-минералогических наук Г.А. Карпова и сотрудников Кроноцкого государственного природного биосферного заповедника за содействие в ор-

ганизации работ в кальдере Узон.

Работа выполнена при финансовой поддержке интеграционных проектов СО РАН № 92, 93, 94; бюджетного проекта VI.58.1.3.

ЛИТЕРАТУРА

Александров А.Ю. Влияние состава среды и условий культивирования на рост углеводородокисляющих микроорганизмов: дис. ... канд. биол. наук. Волгоград, 2010.

Андреева И.С., Емельянова Е. К., Загребельный С. Н. и др. Психротолерантные штаммы-нефтедеструкторы для биоремедиации почв и водной среды // Биотехнология. 2006. № 1. С. 43.

Балашова Н.В., Кошелева И.А., Филонов А.Е. и др. Штамм *Pseudomonas putida* BS3701 – деструктор фенантрена и нафталина // Микробиология. 1997. Т. 66. С. 488–493.

Бескровный Н.С., Лебедев Б.А., Главатских С.Ф. Металлы и нефть в гидротермальных растворах кальдеры Узон // Современные минералообразующие растворы. Петропавловск-Камчатский, 1970. С. 21–22.

Большаков Г.Ф., Бейко О.А. Химический состав нефтей Западной Сибири. Новосибирск: Наука, 1988. 285 с.

Бонч-Осмоловская Е.А. Микробные сообщества глубинных подземных местообитаний Южной Африки и Западной Сибири: биологическое разнообразие и биотехнологический потенциал // Научно-исследовательский отчет по государственному контракту № 11.519.11.2029. М., 2013. 92 с.

Борзенков И.А., Милехина Е.И., Готоева М.Т. и др. Свой-ства углеводородокисляющих бактерий, изолированных из нефтяных месторождений Татарстана, Западной Сибири, Вьетнама // Микробиология. 2006. Т. 75. № 1. С. 82–89.

Ветрова А.А. Биодеградация углеводородов нефти плазмидосодержащими микроорганизмами-деструкторами: дис. ... канд. биол. наук. М., 2010.

Ветрова А.А., Иванова А.А. Филонова А.Е. и др. Биодеструкция нефти отдельными штаммами и прин-ципы составления микробных консорциумов для очистки окружающей среды от углеводородов нефти // Известия ТулГУ. Естеств. науки. 2013. № 2-1. С. 241–257.

Градова Н.Б., Горнова И.Б., Эддауди Р., Салина Р.Н. Использование бактерий рода *Azotobacter* при биоремедиации нефтезагрязненных почв // Прикл. биохим. микробиол. 2003. Т. 39. № 3. С. 318–321.

Гумеров В.М. Молекулярный анализ биоразнообразия микроорганизмов термальных источников Камчатки: дис…. канд. биол. наук. М., 2011.

Емельянова Е.К. Микроорганизмы природных биоценозов для биоремедиации почв и водных объектов Сибири, загрязненных нефтепродуктами: дис. ... канд. биол. наук, Кольцово, 2009.

Заварзин Г.А. Начальные этапы эволюции биосферы // Вестник Российской академии наук. 2010. Т. 80. № 12. С. 1085–1098.

Карпов Г.А., Бонч-Осмоловская Е.А., Заварзин Г.А., Лупикина Е.Г. К характеристике термофильных микроорганизмов кальдеры Узон (Восточная Камчатка) // Сохранение биоразнообразия Камчатки и прилегающих морей. Петропавловск-Камчатский: Камчатпресс, 2008. № 280. С. 109–112.

Карпов Г.А., Мороз Ю.Ф., Николаева А.Г. Геохимия гидротерм и глубинное строение кальдеры Узон // Труды Кроноцкого государственного природного биосферного заповедника. Воронеж, 2013. С. 163.

Киреева Н.А., Новоселова Е.И., Онегова Т.С. Активность каталазы и дегидрогеназы в почвах, загрязненных нефтью и нефтепродуктами // Агрохимия. 2002. № 8. С. 64–72.

Кодина Л.А. Геохимическая диагностика нефтянного загрязнения почвы // Восстановление нефтеза-грязненных почвенных экосистем. М.: Наука, 1988. С. 112–122.

Конторович А.Э., Бортникова С.Б., Карпов Г.А. и др. Кальдера вулкана Узон (Камчатка) – уникальная природная лаборатория современного нафтидогенеза // Геология и геофизика. 2011. Т. 52. № 8. С. 986–990.

Коршунова И.О., Егорова Д.О. bph-Гены галотолерантных бактерий рода *Rhodococcus*, контролирующие первый этап окисления бифенила // Биология будущего: традиции и инновации. Екатеринбург, 2010. С. 98.

Кошелева И.А., Балашова Н.В., Измалкова Т.Ю. и др. Деградация фенантрена мутантными штаммами – деструкция нафталина // Микробиология. 2000. № 6. С. 783–789.

Лобкова Л.Е., Лобков Е.Г. Роль биологических компонентов в экосистемах термальных полей Узона и Долины гейзеров и некоторые аспекты охраны термальных биогеоценозов. // Сохранение биоразнообразия Камчатки и прилегающих морей: Петропавловск-Камчатский: КамчатНИРО, 2003. С. 258–262.

Логинов О.Н., Нуртдинова Л.А., Бойко Т.Ф. и др. Оценка эффективности нового биопрепарата «Ленойл» для биоремедиации нефтезагрязненных почв // Биотехнология. 2004. № 1. С. 77–82.

Лукин А.Е., Пиковский Ю.И. Новые данные об изотопном составе гидротермальной нефти (кальдера Узон на Камчатке) // Докл. АН. 2004. Т. 398. № 1. С. 90–93.

Марданов А.В., Равин Н.В., Бонч-Осмоловская Е.А., Скрябин К.Г. Определение и анализ новых геномов термофильных архей // Генетика микроорганизмов и биотехнология. 2008. Т. 20. С. 62.

Марданов А.В., Равин Н.В. Роль геномики в исследовании разнообразия и эволюции архей // Биохимия. 2012. Т. 77. № 8. С. 965–980.

Нечаева И.А. Биодеградация углеводородов нефти психротрофными микроорганизмами-деструкторами: дис. ... канд. биол. наук, Пущино, 2009.

Нуртдинова Л.А. Исследование процессов ремедиации нефтезагрязненных природных объектов с использованием биопрепарата «Ленойл»: дис. ... канд. биол. наук, Уфа, 2005.

Павликова Т.А. Деградация нефти ассоциацией аэробных углеводородокисляющих микроорганизмов в различных типах почв: дис. ... канд. биол. наук. М., 2004.

Сафиева Р.З. Физикохимия нефти: физико-химические основы технологии переработки нефти: дис. ... д-ра тех. наук, М., 1998.

Слуцкая Э.С., Безсуднова Е.Ю., Марданов А.В. и др. Характеристика новой М42 аминопептидазы из кренархеи *Desulfurococcus kamchatkensis* // Доклады Академии наук. 2012. Т. 442. С. 551–554.

Стабникова Е.В., Селезнева М.В., Рева О.Н. и др. // Прикл. биохим. микробиол. 1995. Т. 31. № 5. С. 534 – 539.

Суржко Л.Ф. Очистка природных и сточных вод от нефтезагрязнений иммобилизованными углеводородокисляющими микроорганизмами: дис. ... канд. тех. наук. СПб., 1999.

Суржко Л.Ф., Финельштейн З.И. Баскунов Б.П. и др. Утилизация нефти в почве и воде микробными клетками // Микробиология. 1995. Т. 64. № 3. С. 393–398.

Таранова Л.В., Жданова Е.Б. Влияние бактерий и дрожжей на биохимическое окисление нефти. // Нефть и газ Западной Сибири: Тез. докл. междунар. научн.-техн. конф. Тюмень, 1996. Т. 2. С. 126.

Тимергазина И.Ф., Переходова Л.С. К проблеме биологического окисления нефти и нефтепродуктов углеводородокисляющими микроорганизмами // Нефтегазовая геология. Теория и практика. 2012. Т. 7. № 1.

Чугунов В.А., Ермоленко З.М., Жиглецова С.К. и др. Разработка и испытания биосорбента «Экосорб» на основе ассоциации нефтеокисляющих бактерий для очистки нефтезагрязненных почв // Прикладная биохимия и микробиология. 2000. Т. 36. № 6. С. 666–671.

Baryshnikova L.M., Grishchenkov V.G., Arinbasarov M.U. *et al*. Biodegradation of oil products by individual degrading strains and their associations in liquid media // Applied Biochemistry Microbiology. 2001. V. 37. No. 5. P. 463–468.

Bonch-Osmolovskaya E.A., Kochetkova T.V., Rusanov I.I. *et al.* Anaerobic transformation of carbon monoxide by microbial communities of Kamchatka hot springs // Extremophiles. 2011. V. 15. No. 3. P. 319–325.

Cerniglia C.E. Biodegradation of polycyclic aromatic hydrocarbons // Biodegradation. 1992. V. 3. P. 351–368.

Connors M.A., Barnsley E.A. Naphthalene plasmid in *Pseudomonas* // J. Bacteriol. 1982. V. 149. P. 1096.

Cooper D.G., Goldenberg B.G. Surface-active agents from. two *Bacillus* species // Appl. Environ. Microbiol. 1987. V. 53. P. 224–229.

Dockyu K., Young-Soo K., Seong-Ki K. *et al.* Monocyclic aromatic hydrocarbon degradation by *Rhodococcus* sp. strain DK1 // Applied environmental Microbiology. 2002. No. 7. P. 3270–3278.

Dutta T. K., Harayama S. Biodegradation of n-alkylcycloalkanes and n-alkylbenzenes via new pathways in *Alcanivorax* sp. strain MBIC 4326 // Appl. Envir. Microbiol. 2001. V. 67. No. 4. P. 1970–1974.

Feist C.F., Hegeman G.D. Phenol and benszoate metabolism by *Pseudomonas putida* of tangential pathways // J. Bacteriology. 1969. V. 100. P. 869–877.

Hamme J., Ward O. Physical and metabolic interactions of Pseudomonas sp. strain JA5-B45 and *Rhodococcus sp.* strain F9-D79 during growth on crude oil and effect of a chemical surfactant on them // Appl. Environ. Microbiol. 2001. V. 69. P. 4874–4879.

Hanson K.G., Nigam A., Kapadia M., Desai A.J. Bioremediation of crude oil contamination using *Acinetobacter* sp. A3 // Curr. Microbiol. 1997. V. 35. No. 3. P. 191–193.

Kublanov I.V., Perevalova A.A., Slobodkina G.B. *et al.* Biodiversity of thermophilic prokaryotes with hydrolytic activities in hot springs of Uzon caldera, Kamchatka (Russia) // App. Env. Microbiology. 2009. V. 75. No. 1. P. 286–291.

Marchai R., Penet S., Solano-Screna F., Vandecasteele J.P. Gasoline and diesel oil biodegradation // Oil Gas Science Technology. 2003. V. 58. No. 4. P. 441–448.

Mardanov A.V., Ravin N.V., Svetlitchnyi V.A. *et al.* Metabolic versatility and indigenous origin of the archaeon *Thermococcus sibiricus*, isolated from a Siberian oil reservoir, as revealed by genome analysis // Appl. Environ. Microbiol. 2009. V. 75. P. 4580–4588.

Margesin R., Labbé D., Schinner F. *et al.* Characterization of hydrocarbon-degrading microbial populations in contaminated and pristine alpine soils // Appl. Environ. Microbiol. 2003. V. 69. P. 3085–3092.

Perevalova A.A., Svetlichny V.A., Kublanov I.V. *et al.* Desulfurococcus fermentans sp. nov., a novel hyperther-mophilic archaeon from a Kamchatka hot spring, and emended description of the genus *Desulfurococcus* // International journal systematic evolutionary microbiology. 2005. V. 55. No. 3. P. 995–999.

Rahman K.S., Rahman T., Lakshmanaperumalsamy P., Banat I.M. Occurrence of crude oil degrading bacteria in gasoline and diesel station soils // J. Basic Microbiol. 2002. V. 42. No. 4. P. 284–291.

Slepova T.V., Sokolova T.G., Lysenko A.M. *et al. Carboxydocella sporoproducens* sp. nov., a novel anaerobic CO-utilizing/H2-producing thermophilic bacterium from a Kamchatka hot spring // Inter. J. Systematic Evolutionary Microbiology. 2006. V. 56. No. 4. P. 797–800.

Varfolomeev S.D., Karpov G.A., Synal H.-A. *et al.* The youngest natural oil on earth // Doklady Chemistry. MAIK Nauka/Interperiodica. 2011. V. 438. No. 1. P. 144–147.

THEORETICAL AND PRACTICAL ISSUES OF BIOLOGICAL OXIDATION OF HYDROCARBONS BY MICROORGANISMS

A.V. Bryanskaya, Yu.E. Uvarova, N.M. Slynko, E.A. Demidov, A.S. Rozanov, S.E. Peltek

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

Summary

The paper deals with the theoretical issues of biological oxidation of oil hydrocarbons from alkanes to polycyclic aromatics. We analyze the mechanisms of biochemical processes of decomposition of oil components and provide an overview of data from common databases. Studies of microbial communities of natural oil seeps in the Uzon caldera are described in detail. It is the first study of ecophysiological characteristics of oil-degrading microorganisms isolated from thermal oil seeps of the caldera.

Key words: biological oxidation, oil, hydrocarbons, microorganisms of the Uzon caldera.

УДК 577.323.7

ЗАВИСИМОСТЬ РАЗМЕРОВ ГЛОБУЛЫ ДНК В ГАЗОВОЙ ФАЗЕ ОТ ДЛИНЫ ЦЕПИ

© 2014 г. **Т.Н. Горячковская[1], А.С. Козлов[2], В.М. Попик[3], Н.А. Колчанов[1, 4], С.Е. Пельтек[1]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: peltek@bionet.nsc.ru;
[2] Федеральное государственное бюджетное учреждение науки Институт кинетики и горения Сибирского отделения Российской академии наук, Новосибирск, Россия;
[3] Федеральное государственное бюджетное учреждение науки Институт ядерной физики им. Г.И. Будкера Сибирского отделения Российской академии наук, Новосибирск, Россия;
[4] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

Поступила в редакцию 5 ноября 2014 г. Принята к публикации 10 ноября 2014 г.

Современные тенденции использования ДНК в нано- и биотехнологиях ставят задачу разработки новых методов анализа молекул ДНК на основе развивающейся приборной базы. Нами разработан метод мягкой неразрушающей абляции для перевода молекул ДНК в аэрозольную фазу при помощи терагерцевого излучения. В настоящей работе с помощью диффузионного спектрометра аэрозолей были проведены измерения размеров наночастиц ДНК в газовой фазе. Изменения, происходящие с ДНК в газовой фазе, были визуализированы при помощи атомно-силовой микроскопии (АСМ). Сопоставление измерений диффузионных размеров аэрозольных частиц плазмиды pUC18 и измерений с применением АСМ дает основания предполагать, что в газовой фазе происходит процесс конденсации молекул ДНК. Построена модель согласно закономерностям, предложенным современными представлениями о процессе конденсации ДНК и формирования глобулы. Теоретические расчеты хорошо совпали с экспериментальными результатами. Экспериментально оцененная персистентная длина ДНК в газовой фазе составила около 0,5 нм, что свидетельствует об отсутствии распределенного заряда на поверхности ДНК в газовой фазе и неионизирующем характере терагерцевого излучения. Исследование конформационных состояний ДНК в газовой фазе позволит расширить знания о закономерностях компактизации ДНК в естественных и искусственных условиях.

Ключевые слова: конденсация ДНК, персистентная длина, упаковка ДНК, десорбция ДНК, атомно-силовая микроскопия, измерение размеров аэрозольных частиц.

ВВЕДЕНИЕ

Для вторичной и третичной структуры молекул ДНК характерен высокий полиморфизм, зависящий от нуклеотидного состава, гидратации и катионного окружения (Manning, 1978; Hagerman, 1988; Neidle, 1994; Shotton *et al.*, 1997; Mazur, 2006). Ионное окружение влияет на способность ДНК накручиваться на нуклеосомы, упаковываться внутри вирусного капсида или связывать транскрипционные факторы. В ряде случаев в основе этого лежит изменение структурных и механических свойств ДНК. Число возможных конформаций макромолекулы возрастает с увеличением длины полимера, и гибкость молекулы по-разному проявляется на коротких и длинных участках макромолекулы.

В разбавленных растворах в присутствии мультивалентных катионов или нейтральных полимеров происходит искусственная конденсация ДНК (Wilson, Bloomfield, 1979; Raspaud *et al.*, 1998). Относительно ДНК термин «кон-

денсация» обычно применяют в отношении мономолекулярной компактизации как противопоставление термину «агрегация», подразумевающему образование структур, состоящих из нескольких молекул. Интерес к изучению конденсации ДНК *in vitro* в последние годы вызван бурным развитием генной терапии, биоэлектроники и поиском способов использования ДНК в нанотехнологиях. Молекулы ДНК могут формировать различные пространственные формы.

Среди таких форм наиболее распространены формы B, A и Z двуспиральной ДНК. Эти структуры различаются длиной витка спирали, количеством нуклеотидов на виток, углом наклона плоскостей азотистых оснований к оси спирали и могут быть право- и левозакрученными. В определенных условиях молекулы с относительно протяженными олигопуриновыми и олигопиримидиновыми участками могут формировать тройные спирали, стабилизированные хугстиновскими взаимодействиями.

Существуют триплексные и пентаплексные структуры, искусственно созданные с использованием изогуанин-изоцитозиновых последовательностей (Piacenza, Grimme, 2004). Способность ДНК к самосборке представляет интерес для наноконструирования. С использованием ДНК созданы наносистемы размерами порядка 100 нм (Zheng *et al.*, 2009). Эти структуры получены в водных растворах, их дальнейшее использование в нанотехнологиях предполагает поиск способов нанесения на поверхности и манипулирования вне водной среды.

Поведение ДНК в растворах достаточно хорошо изучено, построены адекватные модели. Результаты симуляции молекулярной динамики поведения двойной спирали ДНК в водяных парах при температуре 100 °C показали, что испарение ДНК даже в условиях высокой температуры не приводит к критическим изменениям структурных, энергетических и динамических свойств двойной цепи. Цепи остаются связанными, сохраняется небольшая спиральность и большая часть водородных связей ДНК – ДНК (Rueda *et al.*, 2003).

Современные тенденции использования ДНК в нано- и биотехнологиях ставят задачу разработки новых методов анализа молекул ДНК на основе развивающейся приборной базы. К таким методам, позволившим получить новые знания о молекулах ДНК, относятся в частности АСМ и манипулирование единичными молекулами ДНК с помощью лазерного пинцета (Conwell *et al.*, 2003; Lyubchenko, Shlyakhtenko, 2009). Нами был разработан метод мягкой неразрушающей абляции для перевода молекул ДНК в аэрозольную фазу при помощи терагерцевого излучения (Пельтек и др., 2009). Перевод структур ДНК в газовую фазу позволит получить информацию об их размерах и свойствах, провести осаждение на твердые поверхности. Существует несколько подходов к объяснению абляции как физического явления. Каждый из них имеет свои ограничения и не является универсальным. Абляция вещества сопровождается рядом сопутствующих эффектов (конденсация пара, диспергирование жидкой фазы и др.), которые представляют технологический интерес. Лазерная абляция сегодня активно применяется для разнообразных целей: исследовательских, производственных и медицинских. Она также используется для получения наночастиц металлов (Wang *et al.*, 1998; Либенсон и др., 2000; Senkan *et al.*, 2006; Barbara *et al.*, 2007).

В настоящей работы молекулы ДНК переведены в газовую фазу методом мягкой неразрушающей абляции под действием терагерцевого излучения, с помощью диффузионного спектрометра аэрозолей в газовой фазе измерены размеры частиц, образуемых молекулами ДНК фагов T7 и λ, плазмиды pUC18, а также фрагментами ДНК фага λ, полученными при гидролизе ферментами *Hind*III и *Bss*T1I, проведен анализ десорбированного материала с помощью атомно-силовой микроскопии.

МАТЕРИАЛЫ И МЕТОДЫ

Работа проведена на уникальной установке Сибирского центра синхротронного и терагерцевого излучения – лазере на свободных электронах (ЛСЭ), разработанной и созданной в Институте ядерной физики им. Г.И. Будкера СО РАН. Диапазон длин волн излучения лежит в пределах 120–240 мкм, что соответствует 2,5–1,25 ТГц (Gavrilov *et al*., 2007). Для измерения размеров аэрозольных наночастиц использовали разработанный в ИХКиГ СО РАН диффузионный спектрометр аэрозолей (ДСА). Диапазон

измеряемого размера частиц 0,003–0,2 мкм, максимальная измеряемая концентрация частиц 5×10^5 см$^{-3}$. В сравнении с электронным микроскопом, ДСА имеет расхождение 5 % по измерению среднего диаметра частиц и 3 % по средней ширине их распределения (Анкилов и др., 2000). Для получения изображений ДНК использовали атомно-силовой микроскоп ИНТЕГРА-ПРИМА, производства ЗАО «НТ-МДТ», принадлежащий ИЯФ СО РАН. Измерения проводили в полуконтактном режиме, использовали кантилеверы марки NSG01_DLC. Для приготовления препаратов использовали слюду марки «С0 Н» 30 Ч 40 ГОСТ 7134-82 (Балашовский слюдяной комбинат). Пластину слюды расслаивали с помощью бритвенного лезвия и использовали свежесколотую поверхность. Из газовой фазы наночастицы отбирали с помощью вакуумного пробоотборника. Наночастицы осаждали непосредственно на свежесколотую поверхность слюды, для этого пластину слюды размерами 3×3 мм размещали внутри вакуумного пробоотборника.

В работе использованы препараты ДНК плазмиды pUC18, ДНК фагов T7 (39936 п. н.) и λ (48502 п. н.), HindIII и BssT1I гидролизаты ДНК фага λ, производства ООО «СибЭнзим». Для определения размеров наночастиц, образуемых молекулами ДНК в газовой фазе, были проведены эксперименты с фрагментами ДНК фага λ. Электрофоретическим методом из коммерчески доступных препаратов ДНК фага λ, гидролизованной рестриктазами *Hind*III и *Bss*T1I, были препаративно выделены отдельные фрагменты ДНК-гидролизатов и проведена их десорбция. Для этого каждый фрагмент ДНК и ДНК фагов облучали отдельно терагерцевым излучением с длиной волны 128 мкм. В эксперименте использованы следующие фрагменты ДНК (длина ДНК, п. н. (рестриктаза)): 1489(*Bss*T1I), 2027(*Hind*III), 2322(*Hind*III), 3472(*Bss*T1I), 4361(*Hind*III), 6557(*Hind*III), 9416(*Hind*III), 23130(*Hind*III).

Для приготовления препаратов плазмиды pUC18 для АСМ использовали раствор 5 мкг/мл ДНК в воде; 3 мкл раствора наносили на свежесколотую слюду и высушивали в беспылевых условиях.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Для эксперимента использовали водные растворы ДНК плазмиды pUC18, фага T7 и фага λ, фрагменты ДНК фага λ, как описано выше. На алюминиевую подложку наносили 20 мкл водного раствора каждого образца в отдельности и облучали излучением ЛСЭ длиной волны 128 мкм до высыхания капли. Десорбцию частиц регистрировали при помощи ДСА и накапливали в буферной емкости с азотом. Отбор проб из газовой фазы для АСМ проводили дважды. Первую пробу отбирали непосредственно после начала абляции. Для этого поток частиц плазмиды в азоте направляли на пластину свежесколотой слюды.

После отбора пробы поток частиц перенаправляли в буферную емкость и через 20 мин отбирали вторую пробу. Для этого поток частиц из буферной емкости направляли на новую пластину свежесколотой слюды, которую закрепляли в вакуумном пробоотборнике.

На рис. 1, *а* приведены результаты АСМ препарата плазмиды pUC18 до абляции, видны отдельные молекулы плазмиды в виде скрученных нитей. На рисунке видно, что целые молекулы плазмиды в водном растворе имеют суперскрученную, но не конденсированную структуру, т. е. молекулы плазмиды находятся на начальных стадиях конденсации.

Измерения проводили в полуконтактном режиме. С помощью ДСА были проведены измерения размеров наночастиц, получаемых в газовой фазе при десорбции этого же препарата ДНК, размер составил 20,7 нм. По данным АСМ видно, что длина отдельных структур сильно варьирует и линейные размеры существенно превышают размер наночастиц, образуемых ДНК плазмиды pUC18 в газовой фазе. Сопоставление измерений диффузионных размеров аэрозольных частиц плазмиды pUC18 и измерений с помощью АСМ дают основания предполагать, что в аэрозольной фазе происходит процесс конденсации молекул ДНК. Из рис. 1 видно, что прошедшие через газовую фазу молекулы плазмиды приобретают более компактную структуру, богатую суперскрученными участками. Частицы ДНК, собранные на слюду сразу после абляции, имеют диаметр порядка 200 нм и высоту менее 18 нм (рис. 1, *б*), а через 20 мин

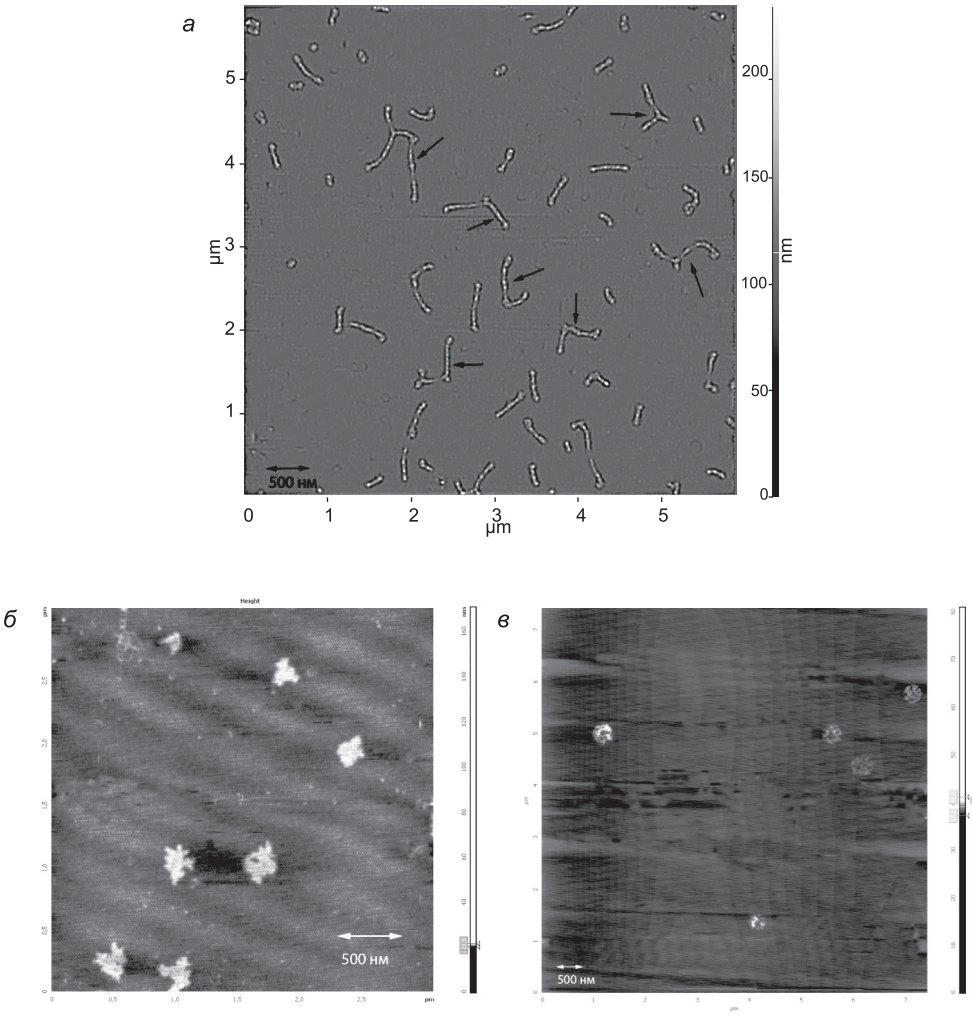


Рис. 1. Атомно-силовая микроскопия препарата плазмиды pUC18: *а* – препарат в водном растворе (5 мкг/мл) до абляции, высота структур 2,5 нм. Стрелками отмечены молекулы плазмиды, имеющие вид скрученных нитей; *б* – после начала абляции проведен отбор образца на слюду из газовой фазы; *в* – молекулы плазмиды после 20 мин пребывания в буферной емкости.

пребывания в буферной емкости они становятся несколько тоньше (менее 10 нм) и имеют более упорядоченное строение (рис. 1, *в*).

В газовой фазе наиболее вероятной формой ДНК является глобула. При осаждении наночастиц ДНК на слюде с помощью вакуумного пробоотборника глобула «распластана» по зараженной поверхности слюды, поэтому размеры структур на слюде не совпадают с размерами наночастиц ДНК в газовой фазе. Однако заметное упорядочивание структуры, наблюдаемое через 20 мин пребывания молекул ДНК плазмиды pUC18 в газовой фазе, свидетельствует в пользу того, что в газовой фазе происходит конденсация ДНК. Аэрозоль – неустойчивая система. Он подвержен постоянным изменениям. С течением времени в аэрозоле происходит укрупнение взвешенных частиц. При столкновениях между частицами под действием броуновского движения, неодинаковой скорости седиментации, под влиянием электростатических сил и гравитации происходит коагуляция аэрозоля. Отдельные частицы агрегируют, и число «свободных» частиц уменьшается, а размер частиц увеличивается (Ветошкин, Таранцева, 2004).

Свойства аэрозолей определяются природой вещества, из которого состоят частицы, природой газовой среды и общей массой частиц, содержащихся в единице объема. На рис. 2 приведена зависимость средних значений размеров

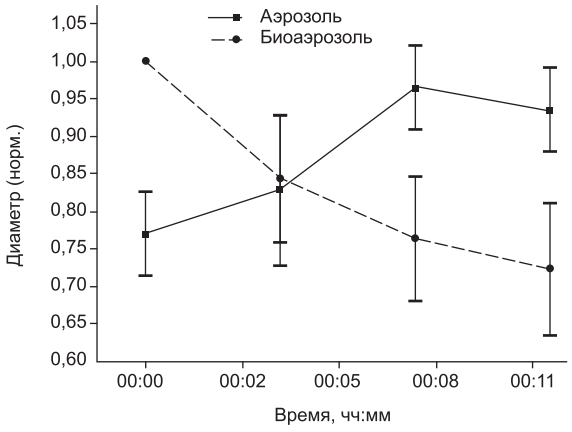


Рис. 2. Поведение «минерального» и «биологического» аэрозолей после экспозиции под излучением ЛСЭ, средние значения размеров.

аэрозольных частиц от времени пребывания в буферной емкости. Для эксперимента были взяты «минеральные» вещества (коллоидное золото, алмазная пудра, фуллереноподобные структуры молибдена и поливинилимидазола) и биополимеры (молекулы ДНК и белков). В случае «биологических» макромолекул наблюдается уменьшение размеров аэрозольных частиц. Такое аномальное поведение наночастиц биополимеров в аэрозоле свидетельствует о конформационных преобразованиях в газовой фазе. Все использованные в эксперименте биополимеры имеют линейную структуру и способны находиться в разных конформационных состояниях.

Видно, что в случае минеральных аэрозолей происходит укрупнение взвешенных частиц, в то время как биополимеры демонстрируют аномальное поведение аэрозольных частиц в зависимости от продолжительности пребывания аэрозольных частиц в буферной емкости. На основании полученных результатов можно предположить, что биополимеры продолжают сворачиваться, будучи в аэрозольной фазе.

Гибкость молекулы и ее внутренняя динамика играют важную роль в биологических функциях ДНК. Известно, что конденсация имеет значение для процессов вирусной трансфекции, регуляции активности генов, апоптоза и сохранения бактериального генома при неблагоприятных условиях внешней среды (Murphy, Zimmerman, 1995; Bloomfield, 1996; Schmutz *et al.*, 1999). В природе плотная упаковка ДНК характерна для фаговых частиц: нативный диаметр частиц фага λ составляет 50 нм, соответственно, плотность упаковки составит 0,734 п. н./нм^3. Исходя из средних линейных размеров одной пары оснований 3,34 Å и около 2 нм диаметра ДНК легко определить предел упаковки – 1,05 п. н./нм^3. В таблице приведены расчеты плотности упаковки ДНК в различных природных структурах.

Упаковка ДНК в эукариотических клетках стабилизирована гистонами и другими хромосомными белками. В нативных структурах форма и плотность упаковки ДНК зависят от первичной последовательности и окружения, состоящего из заряженных белков, полиаминов, ионов двухвалентных и щелочных металлов. Так, например, нуклеосома содержит сегмент двухцепочечной ДНК около 200 п. н. и имеет диаметр около 10 нм. Плотность упаковки составляет 1,17 п. н./нм^3, что свидетельствует о возможности превышения рассчитанного геометрического предела.

Конденсация двухцепочечной ДНК, в результате которой линейные размеры молекулы уменьшаются в десятки тысяч раз, происходит

Различные виды упаковки ДНК

Организм	Длина ДНК	Диаметр упаковки, мкм	Размер, т. п. н.	Плотность упаковки, п. н./нм^3	Структура
Бактериофаг T4	54 мкм	0,1	170	0,17	Капсид
E. coli	1,4 мкм	1	5000	0,005	Бактериальная хромосома
Человек	1,3 м	10	$3,3 \times 10^6$	0,0033	Набор хромосом (диплоидное ядро)

одним из двух способов: путем сфероидальной намотки (вирусы) либо через образование сверхспиральной (суперскрученной) ДНК (вирусы, про- и эукариоты). Внутри фагового капсида параллельные витки ДНК упорядоченно уложены внутри тороидальной структуры. Расстояние между соседними двумя спиралями ДНК составляет порядка одного – двух диаметров молекулы воды. *In vitro* при таком расстоянии между молекулами при ионной силе физиологического раствора ДНК образует жидкокристаллическую фазу.

Упакованная внутри фага ДНК обладает энергией упругости. Так, для фага Ф29 внутреннее давление создает силу порядка 50 пН, чего хватает для выхода части нити в момент инфицирования (Klimenko *et al.*, 1967). Для сравнения, сила, которую надо приложить в перпендикулярном направлении для разделения двойной цепи ДНК фага λ, составляет 15 пН для пары GC и 10 пН для пары AT (Rouzina, Bloomfield, 2001). Тороидальные конденсированные структуры, образуемые ДНК в растворах, напоминают фаговую упаковку. В обоих случаях задействованы одни и те же физические взаимодействия: изгибы, уменьшение энтропии, модификация кулоновских взаимодействий (Post, Zimm, 1979). Кроме тороидальных структур конденсированная ДНК может образовывать палочки. Упаковку в такие структуры контролируют полиамины и специфические белки (Murphy, Zimmerman, 1995; Sarkar *et al.*, 2007).

Размеры частиц ДНК в значительной степени зависят от способа упаковки, который определяется ионным окружением, концентрацией и последовательностью ДНК. С увеличением концентрации катионов магния или натрия в растворе размеры тороидальной структуры возрастают.

Показано также, что ДНК с искусственным изогнутым концом за счет введения в последовательность А-трактов формирует тор меньшего диаметра, несмотря на большую длину молекулы (Conwell *et al.*, 2003). В эксперименте авторы использовали линеаризованную ДНК плазмиды размером 2961 п. н. (3kbDNA) и ее же с добавочным фрагментом 720 п. н., содержащим 60 А-трактов. Каждый А-тракт (dAAAAAA) формирует изгиб ДНК на 13°.

Согласно закономерностям, предложенным современными моделями процесса конденсации ДНК и формирования глобулы, расчет размеров глобулярной наночастицы, формируемой линейным полимером приблизительно описывается следующей функцией:

$$R = \sqrt{N\,U}\ , \qquad (1)$$

где R – диаметр глобулы, N – количество звеньев цепи, U – длина звена, на котором отсутствует гибкость цепи (сегмент Куна) (Флори, 1971; Хохлов, Кучанов, 2000; Тейф, Ландо, 2001). Модели полимерных цепей для описания систем с объемными взаимодействиями и теория перехода клубок – глобула подробно изложены в работах академика А.Р. Хохлова (Хохлов, Кучанов, 2000). В идеальной свободно-сочлененной цепи ориентация двух соседних звеньев независима.

Если же полимерная молекула обладает некоторой межзвенной жесткостью и ориентация двух соседних звеньев коррелирована, то при описании такой цепи используется понятие сегмента Куна U (или персистентной длины l), что позволяет описывать статистику конформационных состояний такой молекулы, используя описание свободно-сочлененной цепи со звеном, равным сегменту Куна.

Персистентная длина – это контурная длина между звеньями полимера, направления которых различаются на 1 рад (57°). Величина персистентной длины определяется выражением

$$l = l_0 \exp(\Delta_e/(kT)), \qquad (2)$$

где Δ_e – разница энергии между минимумами на кривой зависимости внутренней энергии от угла вращения (определяет термодинамическую гибкость макромолекулы), $\Delta_e > 0$, $l_0 \sim 10^{-10}$ м (т. е. порядка длины химической связи), k – постоянная Больцмана, T – температура. Величина сегмента Куна связана с персистентной длиной соотношением $U = 2l$.

В случае, когда рассматриваемым полимером является ДНК, минимальный размер сегмента Куна равен длине одной пары оснований, что составляет около 0,35 нм. С увеличением изгибной жесткости ДНК будет увеличиваться длина куновского сегмента (U) и уменьшаться количество звеньев цепи (N).

В случае, когда сегмент Куна содержит n нуклеотидных пар, $U = 0{,}35n$ (нм), а ко-

личество звеньев цепи уменьшится в n раз: $N = L/n$, где L – длина цепи в нуклеотидных парах. Подставив эти выражения в (1), получим следующую формулу для оценки размеров глобулы:

$$R = \sqrt{L/n}\, 0{,}35n \qquad (3)$$

Жесткость молекулы задается параметром n, с увеличением которого растут размер куновского сегмента и персистентная длина, а количество независимых звеньев уменьшается. Для различных по жесткости молекул, т. е. при разных значениях n, наклон теоретической кривой будет разным. С увеличением n наклон кривой уменьшается. На рис. 3 приведен график зависимости размеров глобулы ДНК от длины цепи.

Как видно из рисунка, наименьшие размеры имеют аэрозольные частицы, образованные фрагментами ДНК длиной до 10 000 п. н., наибольшие аэрозольные частицы образованы фрагментами ДНК длиной 23 130 п. н. и полными линейными геномами фагов T7 и λ. Четыре экспериментальные точки соответствуют n, равному 3. ДНК фага λ формирует глобулу размером меньше теоретически предсказываемого. Также в меньшую сторону отклоняются размеры наночастиц, образованных фрагментами ДНК размерами 1 500–3 500 пар оснований. Вероятными причинами этих отклонений может быть как несовершенство предложенной формулы, так и различия в нуклеотидном составе. Таким образом, экспериментально оцененная длина сегмента Куна для ДНК в газовой фазе составляет около 1 нм, что соответствует персистентной длине 0,5 нм.

По литературным данным, персистентная длина, измеренная для различных растворов, варьирует от нескольких нанометров до контурной длины ДНК (полностью жесткая «палка»). Персистентная длина значительно уменьшается с увеличением концентрации катионов, в частности, в присутствии Na^+, K^+, Mg^{2+}, Ca^{2+} и полиаминов. Так, $Co(NH_3)^{3+}$ в концентрации 2 мкМ уменьшает персистентную длину молекулы ДНК фага λ почти в четыре раза (рис. 4) (Baumann *et al.*, 1997).

Жесткость молекулы в растворе определяется поверхностным отрицательным зарядом, распределенным по цепи ДНК. Катионы экранируют отрицательные заряды, которые несет двойная спираль, изгибы ее в этих условиях требуют меньших затрат энергии, и персистентная длина начинает уменьшаться. В присутствии катионов отрицательные заряды нейтрализуются, и молекула становится более гибкой. По проведенным нами выше оценкам в газовой фазе персистентная длина ДНК составляет менее 1 нм, что свидетельствует об отсутствии распределенного заряда на поверхности ДНК в газовой фазе и неионизирующем характере терагерцевого излучения. Исследование конформационных состояний ДНК в газовой фазе позволит расширить знания о закономерностях компактизации ДНК в естественных и искусственных условиях.

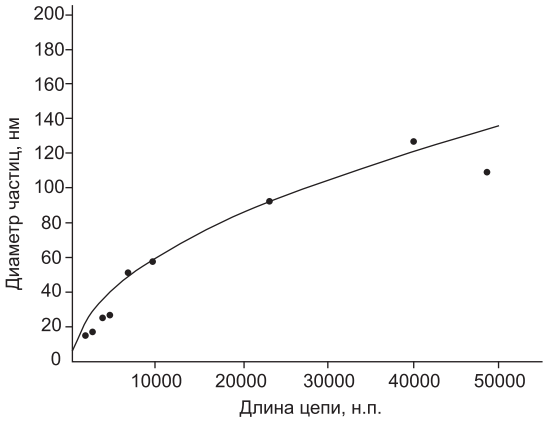


Рис. 3. Зависимость размеров глобулы ДНК от длины цепи. Линией представлены теоретические данные, построенные по формуле (2), точками – экспериментальные измерения. Теоретический расчет проведен для n, равного 3.

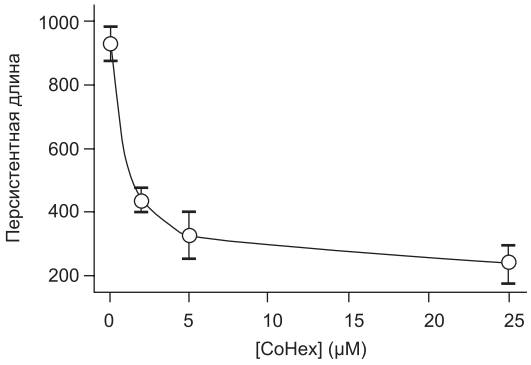


Рис. 4. Зависимость персистентной длины ДНК фага λ от концентрации ионов $Co(NH_3)^{3+}$ (CoHex) (Baumann *et al.*, 1997).

БЛАГОДАРНОСТИ

Работа поддержана проектом Президиума РАН 24.62.

ЛИТЕРАТУРА

Анкилов А.Н., Бакланов А.М., Козлов А.С., Малышкин С.Б. Определение концентрации аэрозолеобразующих веществ в атмосфере // Оптика атмосферы и океана. 2000. Т. 13. № 6-7. С. 644–647.

Ветошкин А.Г., Таранцева К.Р. Технология защиты окружающей среды (теоретические основы). Пенза, 2004. 249 с.

Либенсон М.Н., Шандыбина Г.Д., Шахмин А.Л. Химический анализ продуктов абляции наносекундного диапазона // Журнал технической физики. 2000. Т. 70. Вып. 9. С. 124–127.

Пельтек С.Е., Попик В.М., Горячковская Т.Н., Мордвинов В.А., Петров А.К. Способ абляции целевой ДНК с поверхности ДНК-биочипов. Патент РФ № 2410439. 2009.

Тейф В.Б., Ландо Д.Ю. Конденсация ДНК, вызванная адсорбцией лигандов // Молекулярная биология. 2001. Т. 35. № 1. С. 117–119.

Флори П.Д. Статистическая механика цепных молекул. М., 1971.

Хохлов А.Р., Кучанов С.И. Лекции по физической химии полимеров. М., 2000.

Barbara A., Shehadeh-Masha'our R., Garzozi H.J. Laser ablation in eyes with congenital nystagmus // J. Refract. Surg. 2007. V. 23 (6). P. 623–625.

Baumann C.G., Smith S.B., Bloomfield V.A., Bustamante C. Ionic effects on the elasticity of single DNA molecules // Proc. Natl. Acad. Sci. USA. 1997. V. 94. P. 6185– 6190.

Bloomfield V.A. DNA condensation // Curr. Opin. Struct. Biol. 1996. V. 6 (3). P. 334–341.

Conwell C.C., Vilfan I.D., Hud N.V. Controlling the size of nanoscale toroidal DNA condensates with static curvature and ionic strength. // PNAS. 2003. V. 100 (16). P. 9296–9301.

Gavrilov N.G., Knyazev B.A., Kolobanov E.I. *et al.* Status of the Novosibirsk high-power terahertz FEL // Nuclear instruments and methods in physics research. Sec. A. 2007. V. 575 (1/2). P. 54–57.

Hagerman P.J. Flexibility of DNA // Ann. Rev. Biophys.Biophys. Chem. 1988. V. 17. P. 265–286.

Klimenko S.M., Tikchonenko T.I., Andreev V.M. Packing of DNA in the head of bacteriophage T2 // J. Mol. Biol. 1967. V. 23 (3). P. 523–533.

Lyubchenko Y.L., Shlyakhtenko L.S. AFM for analysis of structure and dynamics of DNA and protein-DNA complexes // Methods. 2009. V. 47 (3). P. 206–213.

Manning G.S. The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides // Q. Rev. Biophys. 1978. V. 11. P. 179–246.

Mazur A.K. Evaluation of Elastic Properties of Atomistic DNA Models // Biophysical J. 2006. V. 91. P. 4507–4518.

Murphy L.D., Zimmerman S.B. Condensation and cohesion of lambda DNA in cell extracts and other media: implications for the structure and function of DNA in prokaryotes // Biophys. Chem. 1995. V. 57 (1). P. 71–92.

Neidle S. DNA structure and recognition. Oxford: IRL Press, 1994. 147 p.

Piacenza M., Grimme S. Systematic quantum chemical study of DNA-base tautomers // J. Comput Chem. 2004. V. 25 (1). P. 83–99.

Post C.B., Zimm B.H. Internal condensation of a single DNA molecule // Biopolymers. 1979. V. 18. P. 1487–1501.

Raspaud E., Olvera de la Cruz M., Sikorav J.-L., Livolant F. Precipitation of DNA by Polyamines: A Polyelectrolyte Behavior // Biophysical J. 1998. V. 74. P. 381 393.

Rouzina I., Bloomfield V. Force-Induced Melting of the DNA Double Helix 1. Thermodynamic Analysis // Biophysical Journal. 2001. V. 80. P. 882–893.

Rueda M., Kalko S.G., Luque F.J., Orozco M. The structure and dynamics of DNA in the gas Phase // J. Am. Chem. Soc. 2003. V. 125 (26). P. 8007–8014.

Sarkar T., Vitoc I., Mukerji I., Hud N.V. Bacterial protein HU dictates the morphology of DNA condensates produced by crowding agents and polyamines // Nucleic Acids Research. 2007. V. 35 (3). P. 951–961.

Schmutz M., Durand D., Debin A. *et al.* DNA packing in stable lipid complexes designed for gene transfer imitates DNA compaction in bacteriophage // PNAS. 1999. V. 96 (22). P. 12293–12298.

Senkan S., Kahn M., Duan S., Ly A., Leidholm C. High-throughput metal nanoparticle catalysis by pulsed laser ablation // Catalysis Today. 2006. V. 117. P. 291–296.

Shotton M.W., Pope L.H., Forsyth T. *et al.* A high-angle neutron fibre diffraction study of the hydration of deuterated A-DNA // Biophys. Chem. 1997. V. 69. P. 85–96.

Wang W., Lin J., Schwartz D. Scanning force microscopy of DNA molecules elongated by convective fluid flow in an evaporating droplet // Biophys J. 1998. V. 75 (1). P. 513–520.

Wilson R., Bloomfield V. Counterion-induced condensation of deoxyribonucleic acid. A light-scattering study // Biochemistry. 1979. V. 18. P. 2192–2196.

Zheng J., Birktoft J.J., Chen Y. *et al.* From Molecular to Macroscopic via the Rational Design of a Self-Assembled 3D DNA Crystal // Nature. 2009. V. 461. P. 74–77.

DEPENDENCE OF A GAS-PHASE DNA GLOBULE SIZE ON CHAIN LENGTH

T.N. Goryachkovskaya[1], A.S. Kozlov[2], V.M. Popik[3], N.A. Kolchanov[1, 4], S.E. Peltek[1]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: peltek@bionet.nsc.ru;
[2] Institute of Chemical Kinetics and Combustion SB RAS, Novosibirsk, Russia;
[3] Budker Institute of Nuclear Physics SB RAS, Novosibirsk, Russia;
[4] Novosibirsk National Research State University, Novosibirsk, Russia

Summary

Modern trends in using DNA in nano- and biotechnologies generated the need for new methods of analyzing DNA molecules with up-to-date equipment. We developed a method of mild nondestructive ablation with terahertz radiation for bringing DNA molecules to aerosol. DNA nanoparticles were measured in the gas phase with a diffusion aerosol spectrometer. Changes that happen to DNA in the gas phase were visualized by atomic force microscopy (AFM). Comparison of diffusion sizes of plasmid pUC18 aerosol particles with those obtained by AFM indicated that DNA molecules experienced condensation in the gas phase. We constructed a model on the base of modern concepts of DNA condensation and globule formation. The predictions matched well the experimental data. The persistence DNA length estimated in the gas phase was about 0.5 nm. This fact points to the absence of distributed charge on the DNA surface in the gas phase and the nonionizing habit of terahertz radiation. Study of DNA conformations in the gas phase will add to the understanding to DNA compactness under natural and artificial conditions.

Key words: DNA condensation, persistence length, DNA package, DNA desorption, atomic force microscopy, aerosol particle size measurement.

УДК 575.858

ЭФФЕКТИВНОСТЬ ИСПОЛЬЗОВАНИЯ ГЕНОВ *BMY2, WAXY* И ВНУТРЕННИХ ТРАНСКРИБИРУЕМЫХ СПЕЙСЕРОВ ГЕНОВ РИБОСОМНЫХ РНК В КАЧЕСТВЕ МАРКЕРОВ ДЛЯ ИЗУЧЕНИЯ ГЕНЕТИЧЕСКОГО РАЗНООБРАЗИЯ ВИДОВ РОДА *ELYMUS*

© 2014 г. **Н.А. Шмаков[1], Д.А. Афонников[1, 2], П.А. Белавин[1], А.В. Агафонов[3]**

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

Поступила в редакцию 17 октября 2014 г. Принята к публикации 28 ноября 2014 г.

Elymus L. – род семейства Poaceae, включает исключительно полиплоидные виды. Виды рода распространены на всех континентах, не менее половины встречаются в Евразии, которая считается местом происхождения рода. Тем не менее видовое разнообразие, генетические особенности отдельных видов и их эволюционные взаимосвязи во многих частях Евразии, в частности на Дальнем Востоке Российской Федерации, до сих пор не исследованы. В связи с этим представляется перспективным изучение эволюционных взаимоотношений видов, произрастающих в данном регионе. В ходе работы проанализированы имеющиеся в базах данных последовательности двух ядерных генов и внутренних транскрибируемых спейсеров генов рибосомных РНК некоторых видов *Elymus*, встречающихся на Дальнем Востоке. Выявлено, что ядерные гены более пригодны для установления филогении на межвидовом уровне. В работе также показано, что последовательности гена *waxy*, принадлежащие различным гапломам, демонстрируют заметные различия и в силу этого могут быть использованы в качестве маркера для установления геномной конституции видов *Elymus*. Наконец, систематическое положение *E. Kamczadalorum* как отдельного вида было подтверждено.

Ключевые слова: *Elymus*, филогения, микроэволюция, генетические маркеры.

ВВЕДЕНИЕ

Elymus L. (пырейник) – крупнейший род семейства Poaceae, трибы Triticeae, включает 150–200 таксонов видового ранга. В пределах России, по последним данным, насчитывается 53 вида, большинство из которых распространены в Сибири и на Дальнем Востоке (Цвелев, Пробатова, 2010). Все виды *Elymus* – аллополиплоиды. Также они отличаются большой частотой спонтанной межвидовой и межродовой гибридизации с образованием жизнеспособных и часто фертильных форм. Многие виды под влиянием условий среды проявляют большую морфологическую пластичность, нередко она затрагивает диагностические признаки. Все это очень затрудняет установление систематического и эволюционного положения видов (Агафонов, 2004).

Поскольку виды *Elymus* напрямую не используют в сельском хозяйстве, этот род не так хорошо изучен. Однако возможности по его применению в различных сферах сельского хозяйства достаточно велики. Многие виды *Elymus* являются перспективными кормовыми интродуцентами. Кроме того, *Elymus* – полностью аллополиплоидный род, и его изучение может пролить свет на эволюционное значение полиплоидизации. Полиплоидия характерна для 50–70 %, а, возможно, и более, цветковых

растений (Soltis *et al.*, 2009), в том числе культурных, и более глубокое понимание этого явления может помочь в получении новых сельскохозяйственных сортов.

Гаплом – один моноплоидный набор хромосом (Löve, 1982). В различных геномных конституциях *Elymus* представлены пять гапломов, из которых три встречаются у видов, произрастающих на территории Дальнего Востока Российской Федерации. Гаплом St отмечен у всех видов *Elymus*. Его донором является род *Pseudoroegneria* Nevski. Происхождение гаплома Y на данный момент не установлено: выдвигаются предположения, что его донором мог быть род *Peridictyon* Seberg, Fred & Baden (Fan *et al.*, 2013) либо что он, как и гаплом St, может происходить от рода *Pseudoroegneria* (Okito, 2008). Широко представлен гаплом H, его донором является род *Hordeum* L. (Dewey, 1971). Ранее существовала гипотеза о независимом происхождении гаплома H у евроазиатских и североамериканских видов, однако она не получила экспериментального подтверждения (Mason-Gamer *et al.*, 2010).

Изучение видов *Elymus*, произрастающих в Сибири и на Дальнем Востоке, на Камчатке и на Сахалине, весьма перспективно и актуально по многим аспектам. В частности, большое практическое значение имеет изучение филогенетических отношений между этими видами. Перспективным подходом для решения этой задачи является построение филогении с использованием в качестве маркеров низкокопийных ядерных генов (Mason-Gamer *et al.*, 1998; Mason-Gamer, 2013) и внутренних транскрибируемых спейсеров (ITS) генов рибосомных РНК – ITS1-5,8 S-ITS2 (Mort *et al.*, 2007). Низкокопийные ядерные гены наследуются по обеим родительским линиям, они мало подвержены конвергентной эволюции. Наличие экзон-интронной структуры предоставляет участки, в которых нуклеотидные замены накапливаются с разными скоростями, благодаря чему можно устанавливать филогению на разных уровнях, основываясь на структурах интронов, межродовом и выше – по структуре экзонов. ITS – широко используемые филогенетические маркеры, отличаются простотой использования в молекулярно-генетических исследованиях за счет очень большой копийности и наличия универсальных праймеров для амплификации. В базах данных аннотировано большое количество последовательностей ITS для множества видов, что позволяет сравнивать новые результаты с уже имеющимися данными. В ходе работы были проанализированы последовательности двух ядерных генов и внутренних транскрибируемых спейсеров генов рибосомных РНК некоторых видов *Elymus*, встречающихся в Сибири и на Дальнем Востоке с целью установления пригодности их использования в качестве филогенетических маркеров. Цель исследования – установление пригодности таких генетических маркеров, как низкокопийные ядерные гены и спейсеры ITS, для изучения филогенетических отношений внутри рода *Elymus*.

МАТЕРИАЛЫ И МЕТОДЫ

Для проверки эффективности использования низкокопийных ядерных генов в качестве маркеров нами были выбраны: (1) участок со второго по пятый экзон гена *bmy2*, кодирующего β-амилазу; (2) участок с девятого по четырнадцатый экзон гена *waxy*, кодирующего гранул-связанную синтазу крахмала GBSSI. Дополнительно использовали последовательности внутренних транскрибируемых спейсеров в генах рибосомных РНК – ITS1 и ITS2. Для каждого маркера с использованием базы данных NCBI Nucleotide была сформирована выборка, содержащая последовательности данных маркеров из геномов нескольких видов *Elymus*, встречающихся на территории Дальнего Востока и Камчатки. Для каждой из отобранных последовательностей при помощи сервиса BLAST nucleotide (Altschul *et al.*, 1990) мы оценили сходство с последовательностями соответствующих маркеров геномных доноров – родов *Pseudoroegneria* и *Hordeum*. На основании максимальной гомологии определяли принадлежность последовательности к одному из двух гапломов – St или H. Кроме того, последовательности, имеющие сходство более чем 99 % с какой-либо другой последовательностью, уже включенной в выборку, исключали из анализа. Также в выборки включили последовательности соответствующих генов *Hordeum jubatum* и *Pseudoroegneria spicata*. В качестве аутгрупп выбрали последовательности указанных генов *Secale cereale*.

Выявленные последовательности выборок были выравнены программой ClustalW (Larkin *et al.*, 2007). Слишком короткие последовательности (менее 70 % от средней длины последовательностей соответствующего маркера) не рассматривали. Оставшиеся последовательности были использованы для анализа. Детальная информация о выбранных для анализа последовательностях и видах, к которым они принадлежат, отражена в табл. 1–3. Кроме того, в

Таблица 1

Выборка последовательностей маркера ITS1-ITS2

Вид	Распространение	Клон	Длина, пн	Обозначение
E. canadensis	Сев. Америка	1	602	canadensis 1
E. caninus	Европа, Центральная Азия, Южная Сибирь	clone 1	698	caninus PI564910 1
		clone 2	702	caninus PI564910 2
		E1-1	605	caninus E1-1
		E1-2	605	caninus E1-2
		E1-3	605	caninus E1-3
		E1-4	605	caninus E1-4
		E1-5	605	caninus E1-5
		E1-6	605	caninus E1-6
E. ciliaris	Восточная Азия, Дальний Восток	clone 1	696	ciliaris H7000 1
		clone 5	699	ciliaris H7000 5
E. dahuricus	Южная Сибирь, Центральная и Восточная Азия	XM-14	703	dahuricus XM-14
		GS-21	703	dahuricus GS-21
		XJ-22	703	dahuricus XJ-22
		NM-3	701	dahuricuscilindricus NM-3
		HB-8	702	dahuricuscilindricus HB-8
		WZ-11	703	dahuricuscilindricus WZ-11
		XM-18	703	dahuricuscilindricus XM-18
E. gmelini	Восточная Азия, Дальний Восток	clone 1	698	gmelinii H1033 1
		clone 4	702	gmelinii H1033 4
E. hystrix	Северная Америка	1	603	hystrix 1
		2	603	hystrix 2
		3	603	hystrix 3
E. mutabilis	Сибирь	wx17	600	mutabilis wx17
E. sibiricus	Сибирь, Дальний Восток, Казахстан, Китай	clone 1	607	sibiricus 1
		QH-24	703	sibiricus SD-12
		SC-27	703	sibiricus ZH-24
		SD-12	703	sibiricus WZ-26
		WZ-26	703	sibiricus SC-27
E. trachycaulus	Камчатка, Сев. Америка	wx19	601	trachucaulus wx19
E. virginicus	Северная Америка	wx111	598	virginicus wx111
		wx112	602	virginicus wx112
H. jubatum		A	598	H. jubatum BCC2055(H2324) clone a
		C	598	H. jubatum BCC2055(H2324) clone c
		E	598	H. jubatum BCC2055(H2324) e
P. spicata		–	703	P. spicata
		PI232124	601	P. spicata PI 232124
		3	702	P. spicata PI232124 3
		4	702	P. spicata PI232124 4
S. cereale		pAHScc9	713	S. cerealecereale pAHScc9
		–	601	S. cereale

Таблица 2

Выборка последовательностей маркера *bmy2*

Вид	Распространение	Клон	Длина, пн	Гаплом	Обозначение
E. canadensis	Сев. Америка	4d	1418	H	canad_4d_H
		4a	1428	St	canad_4a_S
E. caninus	Европа, Центральная Азия, Южная Сибирь	1a	1218	St	canin_1a_S
		5d	1430	St	canin_5d_S
		1d	1421	H	canin_1d_H
		4b	1414	H	canin_4b_H
E. ciliaris	Восточная Азия, Дальний Восток	1g	1430	St	cilia_1g_S
		5e	1433	St	cilia_5e_S
		2f	1433	St	cilia_2f_S
		5a	1434	St	cilia_5a_S
E. gmelini	Центральная Азия, Дальний Восток	1e	1432	St	gmeli_1e_S
		1f	1363	St	gmeli_1f_S
E. mutabilis	Сибирь, Сев. Казахстан, Сев. Китай, Монголия	1a	1366	St	mutab_1a_S
		2h	1430	St	mutab_2a_S
		1c	1418	H	mutab_1c_H
		2c	1422	H	mutab_2c_H
E. sibiricus	Сибирь, Дальний Восток	3g	1401	St	sibir_3g_S
		1b	1423	H	sibir_1b_H
E. trachycaulus	Евразия, Сев. Америка	1a	1429	St	trach_1a_S
H. jubatum		1a	1434		H. jubatum-
		2c	1412		H. jubatum-
P. spicata		1a	1322		P. spicata-
		6b	1367		P. spicata-
S. cereale		1a	1445		S. cereale-
		1b	1426		S. cereale-

анализ был взят эндемичный камчатский вид *E. kamczadalorum*. При помощи адаптированного метода подготовки амплифицированных фрагментов ДНК путем клонирования в Т-вектор по методу Сэнгера проведено секвенирование последовательностей гена гранул-связанной синтазы крахмала I waxy. Для секвенирования использовали следующую пару праймеров: прямой – GGCACCGGGAAGAAGAAGTT; обратный – GGCGAGCGGCGCGATCCCTCGCC.

Применяли следующий профиль ПЦР: плавление 95 °С, 2 мин; 40 циклов – 30 с 95 °С, 30 с 60 °С, 60 с 72 °С; финальная элонгация – 72 °С, 10 мин. Всего выделили 10 копий гена *waxy*. Восемь последовательностей были отфильтрованы по причине фрагментарности, две оставшиеся поступили в дальнейший анализ.

С целью выявления наиболее достоверных моделей нуклеотидных замен выравнивания обработали в программе jModelTest 0.1.1 (Posada, 2008). Установлено, что для маркеров *bmy2* и *waxy* оптимальна модель эволюции Кимуры, в то время как для маркера ITS оптимальной моделью оказалась модель Джукса – Кантора. Далее в программе PhyML 3.0 (Guindon *et al.*, 2010) построили филогенетические деревья для каждого выравнивания. Филогенетические деревья были визуализированы при помощи программы Archaeopteryx (Han, Zmasek, 2009).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Интерпретация филогенетического дерева, построенного на основании маркера ITS (рис. 1), представляет некоторые сложности. Так, многие последовательности группируются согласно видовой принадлежности, в том числе и *H. jubatum*, один из гипотетических доноров гаплома H. Более того, кластер *H. jubatum* имеет бутстреп-поддержку 100 % и отделен от большинства последовательностей ветвями очень большой длины, что указывает на боль-

Таблица 3

Выборка последовательностей маркера *waxy*

Вид	Распространение	Клон	Длина, пн	Гаплом	Обозначение
E. canadensis	Сев. Америка	4a	1209	H	canad-4a-H
		1 (4)	1251	St	canad-14
E. caninus	Европа, Центральная Азия, Южная Сибирь	6	1241	H	canin-6-H
		1a	1190	H	canin-1a-H
		1	1251	St	canin-1-St
		1n	1235	St	canin-1n-S
E. ciliaris	Восточная Азия, Дальний Восток	1b	1231	St	cilia-1b-S
		1g	1226	St	cilia-1g-S
E. gmelinii	Центральная и Восточная Азия, Дальний Восток	1a	1213	St	gmeli-1a-S
		1b	1231	St	gmeli-1b-S
E. mutabilis	Сибирь, Сев. Казахстан, Сев. Китай, Монголия	1a	1205	H	mutab-1a-H
		1c	1225	St	mutab-1c-S
E. sibiricus	Сибирь, Дальний Восток	1b	1228	St	sibir-1b-S
		3a	1228	St	sibir-3a-S
E. trachycaulus	Северная Америка, Евразия	1b1	1240	St	trach-1b1-
		3a	1216	H	trach-3a-H
		1a1	1218	H	trach-1a1-H
		3b	1218	H	trach-3b-H
		3d	1240	St	trach_3d-S
H. jubatum		1a	1221		H. jubatum-
P. spicata		6a	1251		P. spicata-
		3a	1255		P. spicata-
		4a	1251		P. spicata-
S. cereale		1a	1240		S. cereale-

шое количество замен между ними. Однако полученная модель филогении говорит, что последовательности *H. jubatum* обособились в последнюю очередь, что противоречит данным других исследований.

Кроме того, большинство узлов высокого порядка имеет бутстреп-поддержку меньше 30 %. Последовательности многих видов разнесены по далеко отстоящим кластерам. Так, из четырех последовательностей *P. spicata* три сгруппированы в одном кластере, однако четвертая попадает в другой кластер. Такие же разделения наблюдаются для других видов *Elymus* (*E. dahuricus*, *E. sibiricus*, *E. caninus*). Фактически, филогенетическое дерево не дает четкого разделения по гаплотипам. Для дерева также характерна малая длина ветвей, разделяющих разные виды *Elymus*. Это свидетельствует о практически полной непригодности данного маркера в филогенетических исследованиях рода *Elymus*. В публикации Alvarez и Wendel (2003) обсуждены затруднения, с которыми можно столкнуться, выстраивая филогению по ITS. Одно из них – конвергентная эволюция, которая в случае ITS может удалить из генома последовательности, доставшиеся с одним из предковых геномов. Учитывая также, что ITS подвержены гомоплазии, возможность полной потери одной из предковых последовательностей ITS и рДНК становится весьма вероятной. В литературе нет упоминаний о широком применении ITS для исследования рода *Elymus*, за исключением публикации Liu с соавт. (2006). Однако это исследование в основном посвящено гаплому St, и делать выводы о влиянии конвергентной эволюции на утрату разнообразия между ITS, унаследованных от разных геномных доноров, сложно.

Филогенетическое дерево последовательностей маркера *bmy2* (рис. 2) дает достоверное (бутстреп-поддержка составляет 100 %) разделение последовательностей *Elymus* вместе с последовательностями доноров соответствующих геномов по гаплотипам. Последователь-

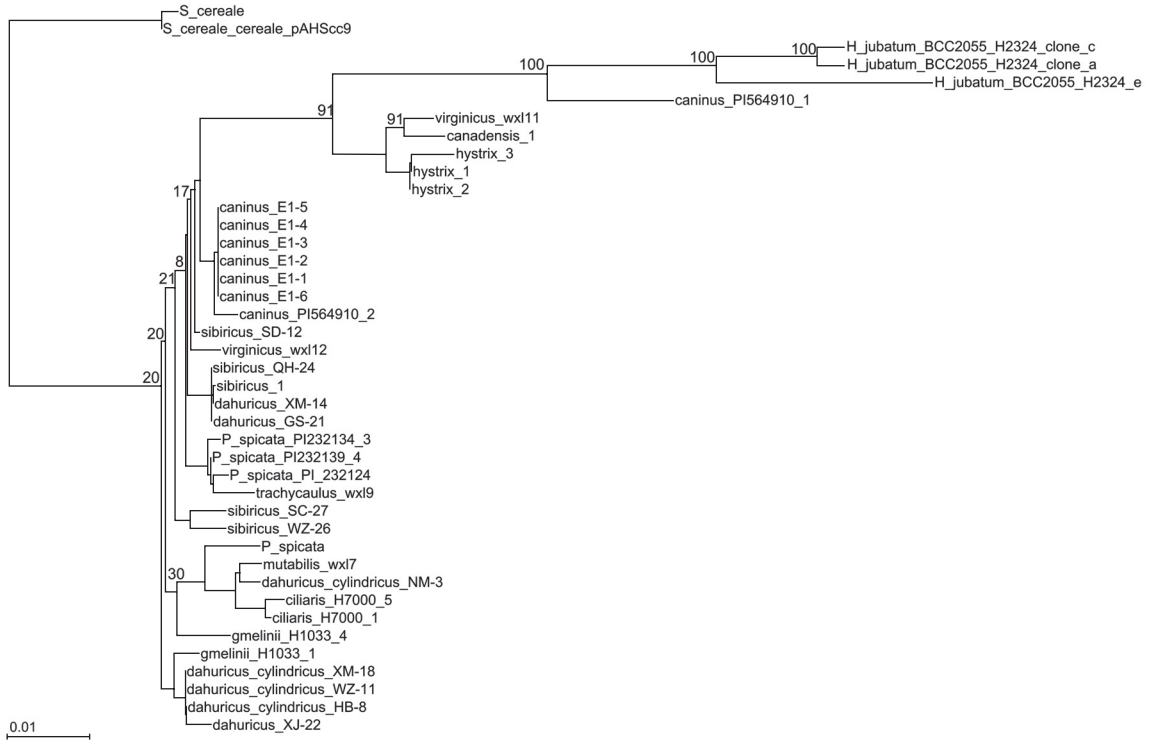


Рис. 1. Филогенетическое дерево, реконструированное на основе последовательностей маркера ITS. Использована программа PhyML 3.0.

ности каждого геномного донора отделяются от последовательностей *Elymus* в соответствующих кластерах с бутстреп-поддержкой 100 и 92–99 %. На основании этого можно сделать вывод, что последовательности маркера *bmy2*, принадлежащие гапломам H и St, при филогенетическом анализе могут с высокой степенью достоверности быть отнесены к кластерам, представляющим соответствующие гапломы. Для обоих гапломов последовательности группируются по видам, в частности, для *E. caninus*, *E. ciliaris* в гапломе St. Следует отметить, что последовательности генетически близких видов *E. caninus* и *E. mutabilis* формируют хорошо выделяющиеся кластеры для гапломов как H, так и St.

На филогенетическом дереве, построенном по маркеру *waxy* (рис. 3), можно наблюдать разделение последовательностей по гапломам St и H с бутстреп-поддержкой 100 %. Последовательности геномных доноров группируются вместе с последовательностями *Elymus* в соответствующие клады. Однако на этом дереве последовательности обоих геномных доноров «смешиваются» с последовательностями *Elymus* соответствующих гапломов. Три из четырех последовательностей *P. spicata* группируются в один кластер, в который вместе с ними попадает последовательность *E. canadensis* (canad-1-4). Четвертая последовательность *P. spicata* не включена в этот кластер. Из последовательностей гаплома H наиболее выраженный кластер формируют последовательности *E. caninus,* за исключением последовательности canin-6-H. Среди последовательностей гаплома St три представителя *E. caninus* группируются в отдельный кластер с последовательностями *E. sibiricus* и *E. mutabilis*, причем эта кластеризация имеет бутстреп-поддержку 96 %.

Как в группе гаплома H, так и в группе St выделяются в отдельный кластер последовательности североамериканских видов *E. trachycaulus* и *E. canadensis*, с бутстреп-поддержкой 99 и 84 % соответственно. Однако в группе гаплома St к ним присоединяется с бутстреп-поддержкой 99 % последовательность gmeli-1a-S. Дополнительно мы провели анализ выравнивания с целью выявления характерных

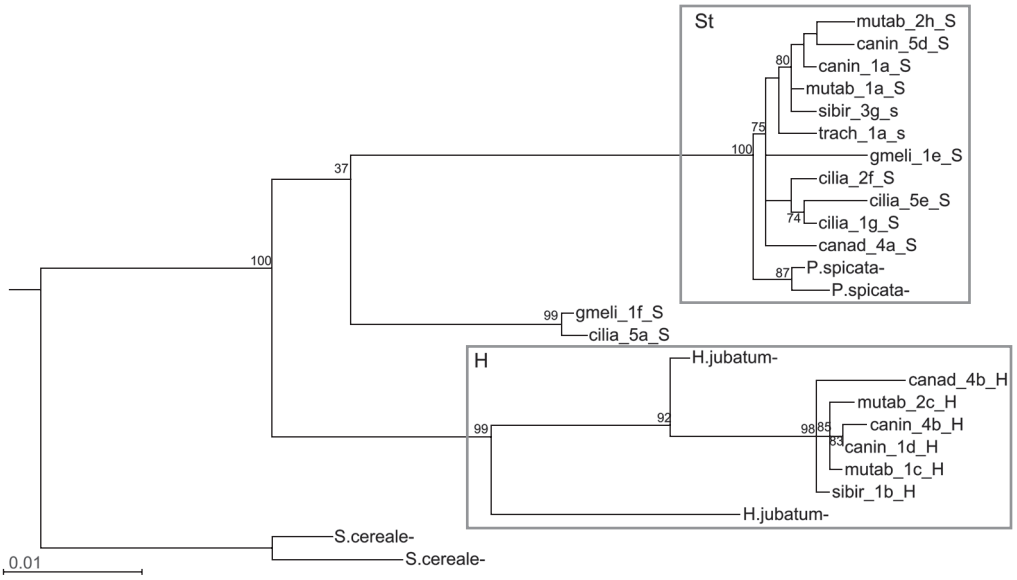


Рис. 2. Филогенетическое дерево, реконструированное по последовательностям маркера *bmy2*. Выделены последовательности, относящиеся к гапломам St и H. Использована программа PhyML 3.0.

участков, различающихся между гапломами и консервативных внутри одного гаплома (рис. 4). Для маркера *waxy* такая последовательность была обнаружена. Это консервативный мотив длиной в 39 нуклеотидов, 3′-CATAATTWTTTT GGGTTTAAATGGTGGTTTGCACAACAAT-5′, в позициях с 1073 по 1111 последовательности ‘clone 6’ *E. caninus*, соответствующей гаплому H (обозначение canin-6-H, см. рис. 3). В том же положении выравнивания у последовательностей гаплома St наблюдается фрагмент 3′-GTCGTCTCTGGTTYAGGATACAYTTCCC AGAACAACGAAGA-5′.

Дополнительный анализ Blast Nucleotide NCBI показал, что эти последовательности не имеют гомологии с последовательностями каких-либо других генов *Elymus* из числа опубликованных в базе данных NCBI Nucleotide. Отметим, что эти последовательности расположены в 11-м интроне гена GBSSI, поэтому их вариабельность не влияет напрямую на структуру белка.

В качестве модельного объекта нами выбран эндемик п-ва Камчатка *E. kamczadalorum* (Nevski) Tzvelev, который, по данным ресурса The Plant List (http://www.theplantlist.org/), до настоящего времени считается синонимом североамериканского вида *E. trachycaulus* (Link) Gouldex Shinners. Ранее была показана специфичность белковых профилей *E. kamczadalorum* по сравнению с заносными образцами *E. trachycaulus* с территории Евразии (Агафонов, Баум, 2000). Тем не менее, учитывая ряд морфологических и биохимических отличий и географическую изолированность, было необходимо подтвердить StH-геномную конституцию *E. kamczadalorum*, поскольку на п-ве Камчатка широко распространен StY-геномный вид *E. gmelinii*, который гипотетически мог участвовать в становлении микроэволюционной обособленности *E. kamczadalorum*.

Среди копий гена *waxy* этого вида, поддавшихся адекватному выравниванию и уложившихся в дендрограмму, маркеры гаплома St не обнаружены. Вместе с тем выявлены два маркера гаплома H. Отсюда можно сделать вывод, что эндемичный вид *E. kamczadalorum* обладает геномной конституцией StStHH. Попадание образцов *E. kamczadalorum* в кластер с североамериканскими видами *E. trachycaulus* и *E. canadensis* можно расценивать как неслучайное, поскольку флора Камчатки филогенетически связана с континентальной североамериканской. Таким образом, полученные результаты демонстрируют преимущество ядерных генов *bmy2* и *waxy* как маркеров для изучения генетического разнообразия видов рода *Elymus* по сравнению с ITS. Эти маркеры позволяют более

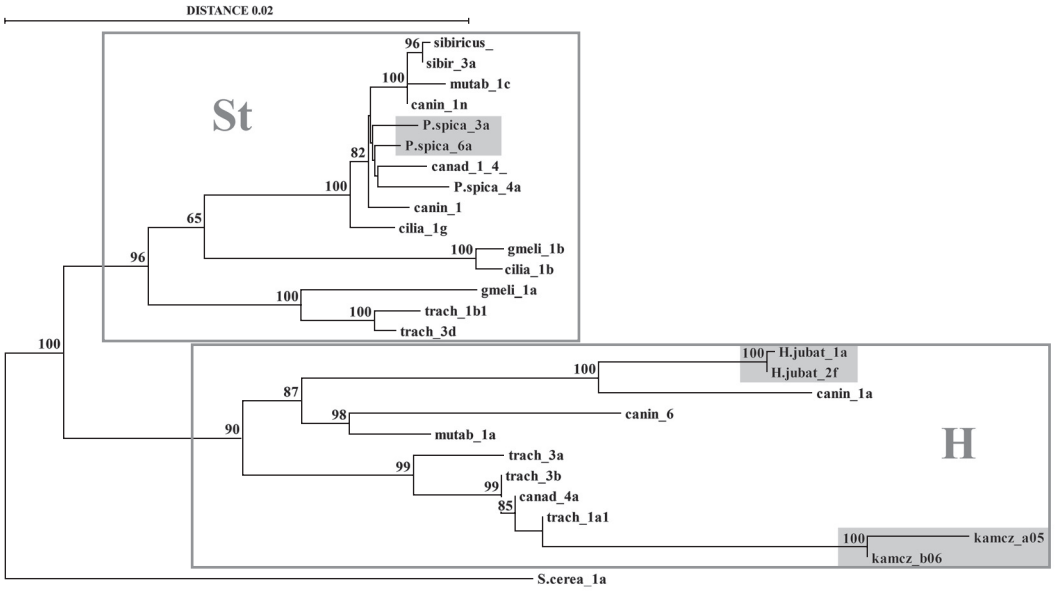


Рис. 3. Филогенетическое дерево, реконструированное на основе последовательностей маркера *waxy*. Выделены последовательности гапломов St, H, Y. Внутри гаплома H отмечены последовательности *E. kamczadalorum.*

```
canad-4a-H      AAGGTACATAATTTTTTTGGGTTTAAATGGTGGTTTGCACAACAAT---------TTAAGAC-TACA---------- TGGCTCAATGGCGGTTT
canin-5b-H      AAGGTACATAATTTTTTTGGGTTTAAATGGTGGTTTGCACAACAAT---------TTAAGAC-TACA---------- TGGCTCAATGGCGGTTT
canin-4b-H      AAGGTACATAATTTTTTTGGGTTTAAATGGTGGTTTGCACAACAAT---------TTAAGAC-TACA---------- TGGATCAATGGCGGTTT
canin-2a-H      AAGGTACATAATTTTTTTGGGTTTAAATGGTGGTTTGCACAACAAT---------TTAAGAC-TACA---------- TGGCTCAATGGCGGTTT
canin-6-H       AAGGTACATAATTATTTTGGGTTTAAATGGTGGTTTGCACAACAAT---------TTAAGAC-TACAC--GGCTAGTCGTGTTCGATACATGGCTCAATGGCGGTTT
canin-1a-H      AAGGTACATAATTTTTTTGGGTTTAAATGGTGGTTTGCACAACAAT---------TTAAGAC-TACA---------- TGGCTCAATGGCGGTTT
trach-3a-H      AAGGTACATAATTTTTTTGGGTTTAAATGGTGGTTTGCACAACAAT---------TTAAGAC-TACA---------- TGGCTCAATGGCGGTTT
mutab-1a-H      AAGGTACATAATTATTTTGGGTTTAAATGGTGGTTTGCACAACAAT---------TTAAGAC-TACA---------- TGGCTCAATGGCGGTTT
H.jubatum-      AAGGTACATAATTATTTTGGGTTTAAATGGTGGTTTGCACAAATAT---------TTAAGAC-TACA---------- TGGCTCAATGGCGGTTT
canin-1-St      AAGGTAAGTCGTCTCT---GGTTTAGGATACATTTCCCAGAACAACGAAGA ---GTTAAGAC-TACA--ATGGTGCTTGTGTTCGATGCA TCCATTAATGGTGGTTT
canin-5a-S      AAGGTAAGTCGTCTCT---GGTTTAGGATACATTTCCCAGAACAACTAAGA ---GTTAAGAC-TACA--ATGGTGCTTGTGTTCGATGCA TCCATTAATGGTGGTTT
canin-2b-S      AAGGTAAGTCGTCTCT---GGTTTAGGATACATTTCCCAGAACAACGAAGA ---GTTAAGAC-TACA--ATGGTGCTTGTGT TCGATGCA TCCATTAATGGTGGTTT
canin-1n-S      AAGGTAAGTCGTCTCT---GGTTTAGGATACATTTCCCAGAACAACGAAGA ---GTTAAGAC-TACA--ATGGTGCTTGTGTTCGATGCA TCCATTAATGGTGGTTT
canad-1(4)      AAGGTAAGTCGTCTCT---GGTTTAGGATACATTTCCCAGAACAACGAAGA ---GTTAAGAC-TACA--ATGGTGCTTGTGTTCGATGCA TCCATTAATGGTGGTTT
canad-4c-S      AAGGTAAGTCGTCTCT---GGTTCAGTATACACTTCCCGGAACAACTAGTAAGAG TTAAGAC-TACA--CCGCTGCTCGTGTTCGATGTA TCCATTAATGGTGGCTT
sibir-3a-S      AAGGTAAGTCGTCTCT---GGTTTAGGATACATTTCCCAGAACAACGAAGA ---GTTAAGAC-TACA--ATGGTGCTTGTGTTCGATGCA TCCATTAATGGTGGTTT
sibir-1b-S      AAGGTAAGTCGTCTCT---GGTTTAGGATACATTTCCCAGAACAACGAAGA ---GTTAAGAC-TACA--ATGGTGCTTGTGTTCGATGCA TCCATTAATGGTGGTTT
gmeli-1a-S      AAGGTACATAATTTCT---GGTTCAGGATACACTTCCCAGAACAATTAACA ---GTTAAGAC-TACA--TAGCTGCTCGAGTTTGAGACA TCCATTAATGGTGGCTT
gmeli-1b-S      AAGGTAAGTCGTCTCT---GGTTTAGGATGCATTTCCCACAACAACTAAGA ---GTTAAGAC-TACACAATGGTGCTCGTGTTCGATGCA TCCATTAATGGTGGCTT
cilia-1b-S      AAGGTAAGTCGTCTCT---GGTTTAGGATGCATTTCCCACAACAACTAAGA ---GTTAAGAC-TACACAATGGTGCTCGTGTTCGATGCA TCCATTAATGGTGGCTT
cilia-1g-S      AAGGTAAGTCGTCTCT---GGTTTAGGATACATTTCCCAGAACAACGAAGA ---GTTAAGAC-TACA--ATGGTGCTTGTGTTCGATGCA TCCATTAATGGTGGTTT
mutab-1c-S      AAGGTAAGTCGTCTCT---GGTTTAGGATACATTTCCCAGAACAACGAAGA ---GTTAAGAC-TACA--ATGGTGCTTGTGTTCGATGCA TCCATTAATGGTGTTTT
trach-1b1       AAGGTAAGTCGTCTCT---GGTTCAGTATACACTTCCCGGAACAACTAGTAAGAG TTAAGAC-TACA--CCGCTGCTCGTGTTCGATGTA TCCATTAATGGTGGCTT
P.spicata-      AAGGTAAGTCGTCTCT---GGTTTAGGATACATTTCCCAGAACAACGAAGA ---GTTAAGAC-TACA--ATGGTGCTTGTGTTCGATGCA TCCATTAATGGTGGTTT
P.spicata-      AAGGTAAGTCGTCTCT---GGTTTAGGATACATTTCCCAGAA CAACGAAGA ---GTTAAGAC-TACA--ATGGTGCTTGTGTTCGATGCA TCCATTAATGGTGGTTT
P.spicata-      AAGGTAAGTCGTCTCT---GGTTCAGTATACACTTCCCGGAACAACTAGTAAGAG TTAAGAC-TACA--CCGCTGCTCGTGTTCGATGTA TCCATTAATGGTGGCTT
P.spicata-      AAGGTAAGTCGTCTCT---GGTTTAGGATACATTTCCCAAAACAACGAAGA ---GTTAAGAC-TACA--ATGGTGCTTGTGTTCGATGCA TCCATTAATGGTGGTTT
P.spicata-      AAGGTAAGTCGTCTCT---GGTTTAGGATACATTTCCCAAAACAACGAAGA ---GTTAAGAC-TACA--ATGGTGCTTGTGTTCGATGCA TCCATTAATGGTGGTTT
```

Рис. 4. Участок гена *waxy*, содержащий гаплом-специфичные последовательности. Светло-серым выделена последовательность, характерная для гаплома H, темно-серым – для гаплома St.

надежно идентифицировать организмы по их видовой принадлежности и отнести последовательность к тому или иному гаплому.

В данном контексте существует отчетливая проблема в отношении многих видов рода *Elymus*, описанных с территории России. Состав гапломов, или геномная конституция (ГК), установлен только для тех видов, ареал которых выходит за пределы территории РФ и которые были включены в интенсивные цитогенетические исследования конца XX в. (Dewey, 1984; Wang *et al.*, 1994).

У большинства видов геномная конституция остается неизвестной. На наш взгляд, первоначальный акцент должен быть сделан на определении геномной конституции методами молекулярного маркирования, поскольку при этом подразумевается поиск и таксономическая идентификация живого материала неизученных видов. Но следует осознавать, что среди существующих данных о молекулярно-генетических характеристиках известных видов присутствует высокая доля материала с ошибочным определением видовой принадлежности. Еще больший риск включить в исследования ошибочно идентифицированный материал существует при работе с редкими или сомнительными видами из маргинальных местообитаний.

ЗАКЛЮЧЕНИЕ

В ходе работы выявлено, что ядерные гены более пригодны для установления филогении на межвидовом уровне. Показано, что последовательности гена *waxy*, принадлежащие различным гапломам, демонстрируют заметные различия и в силу этого могут быть использованы в качестве маркера для установления геномной конституции видов *Elymus*.

БЛАГОДАРНОСТИ

Работа выполнена при частичной поддержке бюджетного проекта VI.61.1.2 (Н.А. Шмаков, Д.А. Афонников).

ЛИТЕРАТУРА

Агафонов А.В. Система рекомбинационных и интрогрессивных генпулов StH-геномных видов рода *Elymus* L. Северной Евразии: дис. д-ра биол. наук. Центральный Сибирский ботанический сад, Новосибирск, 2004.

Агафонов А.В., Баум Б.Р. Индивидуальная изменчивость и репродуктивные свойства половых гибридов внутри комплекса *Elymus trachycaulus* (Poaceae: Triticeae) и близких таксонов. 1. Полиморфизм запасных белков эндосперма у биотипов Северной Америки и Евразии // Turczaninowia. 2000. Т. 3. Вып. 1. С. 63–75.

Цвелев Н.Н., Пробатова Н.С. Роды *Elymus* L., *Elytrigia* Desv., *Agropyron* Gaertn., *Psathyrostachys* Nevski и *Leymus* Hochst. (Poaceae: Triticeae) во флоре России // Комаровские чтения. Владивосток: Дальнаука, 2010. Вып. 57. С. 5–102.

Altschul S.F., Gish W., Miller W. *et al.* Basic local alignment search tool // J. Mol. Biol. 1990. V. 215. P. 403–410.

Alvarez I.A., Wendel J.F. Ribosomal ITS sequences and plant phylogenetic inference // Molecular Phylogenetics Evolution. 2003. V. 29. P. 417–434.

Dewey D.R. Synthetic hybrids of Hordeum bogdanii with Elymuscanadensis and Sitanionhystrix // American Journal Botany. 1971. V. 58. P. 902–908.

Dewey D.R. The genomic system of classification as a guide to intergeneric hybridization with the perennial Triticeae. Gene manipulation in plant improvement. N. Y.: Plenum Publ. Corp., 1984. P. 209–279.

Fan X., Sha L., Dong Z. *et al.* Phylogenetic relationships and Y genome origin in *Elymus* L. sensu lato (Triticeae; Poaceae) based on single-copy nuclear *Acc1* and *Pgk1* gene sequences // Molecular Phylogenetics Evolution. 2013. V. 69. Issue 3. P. 919–928.

Guindon S., Dufayard J., Lefort V. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0 // Systematic Biology. 2010. V. 59 (3). P. 307–321.

Han M.V., Zmasek C.M. phyloXML: XML for evolutionary biology and comparative genomics // BMC Bioinformatics. 2009. V. 10. P. 356.

Larkin M.A., Blackshields G., Brown N.P. *et al.* ClustalW and ClustalX version 2 // Bioinformatics. 2007. V. 23 (21).

Liu Q., Ge S., Tang H. *et al.* Phylogenetic relationships in Elymus (Poaceae: Triticeae) based on the nuclear ribosomal internal transcribed spacer and chloroplast trnL-F sequences // New Phytologist. 2006. V. 170. P. 411–420.

Löve A. Genetic evolution of the wheatgrasses // New Zealand J. Bot. 1982. V. 20. P. 169–186.

Mason-Gamer R. Phylogeny of a genomically diverse group of *Elymus* (Poaceae) allopolyploids reveals multiple levels of reticulation. Plos ONE, 2013.

Mason-Gamer R., Burns M., Naum M. Reticulate evolutionary history of a complex group of grasses: phylogeny of *Elymus* StStHH allotetraploids based on three nuclear genes. Plos ONE, 2010.

Mason-Gamer R., Weil C.F., Kellog E.A. Granule-bound starch synthase: structure, function and phylogenetic utility // Mol. Biol. Evol. 1998. V. 15 (12). P. 1658–1673.

Mort M., Archibald J., Randle C. *et al.* Inferring phylogeny at low taxonomic levels: utility of rapidly evolving cpDNA and nuclear ITS loci // American Journal Botany. 2007. V. 94 (2). P. 173–183.

Okito P. Origin of the Y genome in Elymus. All Graduate Theses and Dissertation. Paper 95, 2008.

Posada D. jModelTest: phylogenetic model averaging // Mol. Biol. Evol. 2008. V. 25 (7). P. 1253–1256. doi: 10.1093/molbev/msn083.

Soltis E.D., Albert V.A., Leebens-Mack J., Bell C.D. Polyploidy and angiosperm diversification // American Journal Botany. 2009. V. 96 (1). P. 336–348.

The Plant List (2013). Version 1.1. Published on the Internet; http://www.theplantlist.org/ (accessed 1st January).

Wang R., von Bothmer R., Dvorak J. *et al.* Genome symbols in the Triticeae (Poaceae) // Proc. 2nd Int. Triticeae Symp. Logan, Utah, USA, 1994. P. 29–34.

THE SUITABILITY OF THE *BMY2* AND *WAXY* GENES AND INTERNAL TRANSCRIBED SPACERS OF rRNA AS MARKERS FOR STUDYING GENETIC VARIABILITY IN *ELYMUS* SPECIES

N.A. Shmakov[1], D.A. Afonnikov[1, 2], P.A. Belavin[1], A.V. Agafonov[3]

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

Summary

Elymus L. is a genus of the Poaceae family, which includes only polyploid species. It is widespread over all continents, with at least half of the species occurring in Asia, and this continent is considered to be its motherland. However, the diversity, genetic characteristics, and evolutionary interactions among *Elymus* species of some regions of Asia are still vague, and the Far East of Russia is one of such territories. Thus, investigation of evolutionary relations among species of Far East and Kamchatka is promising. In this work, several sequences of two nuclear genes and rDNA Internal Transcribed Spacers annotated in databases are analyzed. Nuclear genes sequences are shown to be more useful in building phylogeny at the interspecies level. Also, a region of the nuclear gene *waxy* is shown to vary among different haplomes. This variation makes it useful in investigating the genome constitutions of novel *Elymus* species. Finally, systematical status of E. kamczadalorum as a species was proven valid.

Key words: *Elymus*, phylogeny, microevolution, genetic markers.

УДК 575.852.112

ЧИСЛО ГОМОЛОГОВ НЕКОТОРЫХ ФЕРМЕНТОВ БИОСИНТЕЗА ТРИПТОФАНА У РАСТЕНИЙ КОРРЕЛИРУЕТ С ДОЛЕЙ БЕЛКОВ, АССОЦИИРОВАННЫХ С ТРАНСКРИПЦИЕЙ

© 2014 г. **И.И. Турнаев[1], И.Р. Акбердин[1], В.В. Суслов[1], Д.А. Афонников[1, 2]**

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

Поступила в редакцию 17 октября 2014 г. Принята к публикации 1 декабря 2014 г.

Путь биосинтеза триптофана (ПБТ) универсален у большинства известных организмов, хотя и отсутствует у животных и некоторых эубактерий. У растений этот путь консервативен, но для разных видов наблюдается различное количество паралогов ферментов, участвующих в этом пути. В настоящей работе исследована возможная роль изменения числа паралогов ПБТ в процессе эволюции. Для этого проведена идентификация паралогов ферментов этого пути в известных полногеномных последовательностях и оценена статистическая связь между числом паралогов ПБТ и сложностью организмов. Показано, что сложность организмов достоверно коррелирует с числом гомологов ферментов синтеза триптофана у растений как для всех гомологов ферментов этого пути суммарно, так и для гомологов трех из шести ферментов этого пути ASA/ASB, PAI и IGPS. Выявленные зависимости могут быть обусловлены тем, что рост сложности организации растений и увеличение числа гомологов ферментов синтеза триптофана являются механизмами эволюционной адаптации к изменчивым условиям наземной среды обитания.

Ключевые слова: путь биосинтеза триптофана, филогенетические сети, морфологическая сложность организмов, адаптация, изменчивость внешних условий.

ВВЕДЕНИЕ

Триптофан синтезируется бактериями, грибами и растениями и необходим для биосинтеза белков. Кроме того, у растений триптофан служит предшественником таких веществ, как фитоалексины, глюкозинолаты, ряд алкалоидов, участвующих в процессах защиты от патогенов и вредителей, а также ауксина, ключевого гормона морфогенеза растений (Radwanski, Last, 1995). Кроме того, триптофан, как аминокислота, является субстратом для синтеза белков. Следует отметить, что реакция синтеза белков двадцатисубстратная, по числу аминокислот. При этом ситуация, когда аминокислоты за счет их низких концентраций могли бы быть регуляторами синтеза белка, является аминокислотным голоданием, в ответ на который у эукариот клетка снижает скорости синтеза рРНК и тРНК примерно в 10–15 раз, тем самым искусственно создавая избыток аминокислот для реакции белкового синтеза (Картель и др., 2011).

В то же время в одной или двух субстратных реакциях субстраты часто становятся регуляторами собственных реакций. Соответственно триптофан, будучи в пути биосинтеза ауксина у растений субстратом односубстратной реакции, играет роль ее регулятора. Таким образом, триптофан участвует в специфической регуляции экспрессии генов, так как ауксин – регулятор транскрипции ряда генов (Zhao, 2012).

У растений ПБТ из хоризмата включает шесть последовательных реакций, контролируемых ферментами ASA/ASB (антранилат синтаза α/антранилат синтаза β), TRP (PAT, фосфорибозил антранилат-трансфераза), PAI

(фосфорибозил-антранилат-изомераза), IGPS (индол-3-глицерол фосфат синтаза), TSA (триптофан синтеза α), TSB (триптофан синтеза β) (Radwanski, Last, 1995). Гены, кодирующие ферменты ПБТ, несмотря на участие в таком консервативном процессе, как биосинтез белка, у ряда таксонов претерпели дупликации, роль которых до конца не выяснена.

Число генов-компонентов генных сетей в ходе эволюции увеличивается за счет дупликаций генов, с их последующей дивергенцией (Teichmann, Babu, 2004). В ряде случаев за счет дупликации генов образуются мультигенные семейства, включающие гены, кодирующие в одном организме белки с перекрывающимися функциями. К таким семействам относятся гемоглобины, иммуноглобины, антигены гистосовместимости, актины, тубулины, кератины, коллагены, белки теплового шока, клейкие белки слюны, белки хориона, белки кутикулы, желточные белки, фазеолины (запасной гликопротеин семян фасоли), белки YUCCA растений, также как гены гистонов, рибосомных и транспортных РНК (Feliner, Rossello, 2012). Таким образом, ситуация множественности генов, кодирующих белки с перекрывающимися функциями, широко распространена, и паралоги ферментов путей биосинтеза, подобных ПБТ, относятся к таким белкам.

В работе Vogel и Chortia (2006) на 36 видах, представляющих разные таксоны эукариот, было показано, что расширение белковых семейств (оценивалось по изменению общего числа белковых доменов в семействе) коррелирует с ростом числа клеточных типов. Так, положительная корреляция была показана для 194 из 1 219 исследованных семейств белков. Известно, что в процессе эволюции усложнение организма связано с увеличением сложности регуляторной компоненты его генома (Колчанов и др., 2004). Поэтому, если функции триптофана как субстрата для синтеза регуляторных низкомолекулярных соединений существенны для растений, можно предположить, что увеличение числа гомологов ферментов ПБТ, т. е. сложности этого пути, будет коррелировать с увеличением сложности растений в процессе их эволюции. В настоящей работе проведен анализ зависимости между сложностью организации растений и числом паралогов генов ПБТ в геномах 24 видов растений, принадлежащих к разным таксонам. Показано, что между числом паралогов ферментов ПБТ и сложностью организации растений существует значимая положительная корреляция, что свидетельствует о важной регуляторной функции триптофана у растений. Выявленная зависимость может быть связана с необходимостью реализации дифференциальной экспрессии генов-паралогов ПБТ для динамического изменения уровня триптофана в процессе ответа растений на изменяющиеся условия внешней среды.

МАТЕРИАЛЫ И МЕТОДЫ

Формирование выборок гомологов ферментов пути биосинтеза триптофана у растений

Для анализа были взяты последовательности белков ферментов ПБТ у *Arabidopsis thaliana*: ASA (идентификаторы TAIR (Lamesch *et al.*, 2012) AT5G05730, AT2G29690); ASB (AT1G25220, AT5G57890); PAT(TRP) (AT5G17990); PAI (AT1G07780, AT5G05590, AT1G29410); IGPS (AT2G04400, AT5G48220); TSA (AT4G02610, AT3G54640); TSB (AT5G54810, AT4G27070). С помощью программы BLASTP 2.2.29+, (e- value $<10^{-40}$) проведен поиск гомологов этих ферментов в 24 полностью секвенированных геномах растений из базы данных (БД) PLAZA (Van Bel *et al.*, 2012): пяти зеленых водорослей, одного мха, одного плауна, четырех однодольных растений и 13 двудольных. Для каждого фермента проводили реципрокный поиск, при котором объединялись результаты поиска по всем его паралогам.

Идентификация паралогов ферментов пути биосинтеза триптофана у растений

Множественное выравнивание последовательностей выявленных гомологов проводили с помощью программы Mafft 7.110 (Katoh, Toh, 2008). Проверяли наличие и целостность в последовательностях ключевых консервативных доменов, информация о которых была взята из БД CDD версии 3.10 (Marchler-Bauer *et al.*, 2013). Последовательности со значительными нарушениями доменов (крупные вставки или

делеции) удаляли из выборки. Для более точной идентификации ортологов и паралогов в последовательностях белковых семейств мы реконструировали филогенетические сети методом ProteinMLdist/NeighborNet из пакета SplitsTree4 (v. 4.12.8) (Huson, Bryant, 2006), поскольку известно, что филогенетические сети более адекватно отражают филогенетические отношения при наличии в эволюции горизонтального переноса генов, что можно ожидать при анализе эволюции белковых семейств у растений (Richardson, Palmer, 2007).

Для распознавания гомологов ферментов ПБТ нами намеренно был выбран порог распознавания BLASTP 2.2.29+, (e-value < 10^{-40}), позволяющий идентифицировать как белки ферментов ПБТ, так и родственные им белки, выполняющие функции, не связанные с синтезом триптофана. Чтобы в построенных филогенетических сетях определить границу, отделяющую белки ПБТ от родственных им белков с другими функциями, мы для каждого из отдельных кластеров филогенетической сети определяли белки с известными функциями на основе имеющихся литературных данных и информации из БД CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). Последовательности белков, принадлежащие к кластерам, в которых (1) оказались белки с функциями, не относящимися к синтезу триптофана, и (2) не было белков ПБТ, удаляли из выборки, после чего по новым выборкам заново строили филогенетические сети.

Затем из выборок удаляли последовательности, положение которых на графе филогенетической сети нарушало общепринятую топологию филогенетического дерева растений. Проводили несколько итераций этой процедуры, до тех пор пока нарушения не устранялись. Отметим, что среди удаленных таким образом последовательностей многие содержали делеции и вставки среднего размера в консервативных участках белков. Описанный отбор последовательностей позволил выявить среди гомологов ферментов устойчивые группы белков, с высокой вероятностью выполняющие сходную функцию. Если в полногеномных данных гомологи какого-либо фермента не обнаруживались, для определения их возможного наличия в геноме мы проводили дополнительный поиск в последовательностях EST организма, представленных в БД GenBank и Ensemble (EST последовательности в большинстве представляют фрагменты мРНК, кодирующих белки, что делает неэффективным их использование для точного определения количества гомологов и реконструкции филогении).

Анализ корреляций между сложностью организмов и количеством паралогов ферментов пути биосинтеза триптофана

Для оценки сложности организмов мы использовали параметр $F_{БАТ}$ – отношение количества белков, ассоциированных с транскрипцией (БАТ), к общему числу белков организма (Lang *et al.*, 2010). К числу БАТ относятся транскрипционные факторы и другие регуляторы транскрипции. $F_{БАТ}$ может быть точно оценен на основе полногеномных данных и хорошо коррелирует с такой широко известной характеристикой сложности организмов, как число клеточных типов (Там же). Данные по количеству белков БАТ растений вышеуказанных таксонов (зеленые водоросли, мхи, плауны, одно- и двудольные) взяты из статьи Lang с соавт. (2010). Для оценки значимости взаимосвязи между числом паралогов и количеством белков БАТ мы использовали коэффициент корреляции Пирсона.

РЕЗУЛЬТАТЫ

Количество выявленных гомологов ферментов ПБТ в геномах растений представлено в табл. 1. Оказалось, что водоросли и плауны имеют по одному паралогу каждого фермента, мхи и сосудистые растения – от 1 до 9 паралогов. Девять паралогов ASA выявлено в геноме *Malus domestica*. Для ферментов ПБТ, у которых в полногеномных данных не было найдено гомологов, был проведен анализ библиотек EST. В некоторых случаях (PAI у *Mallus domestica* и ASA у *Chlamidomonas reinhardtii*) гомологичные последовательности не обнаружены ни в полногеномных данных, ни в библиотеках EST, что, вероятно, связано с неполной представленностью генов геномов данных организмов как в полногеномных проектах, так и в библиотеках EST этих видов. Анализ взаимосвязи между

Таблица 1

Наличие гомологов ферментов синтеза триптофана (строки) в геномах растений (столбцы)

Кол-во последовательностей в семействе			**53**		**43**		**40**		**30**		**37**		**49**		**40**	
Двудольные	*Lotus japonicus*		1		1		1		2		**E-50**		1		2	
	Medicago truncatula		4		**E-91**		1		1		1		**E-111**		2	
	Glycine max		4		3		1		2		2		2		3	
	Malus domestica		9		3		5		–		3		4		5	
	Fragaria vesca		2		1		5		2		**E-59**		4		2	
	Manihot esculenta		2		2		1		1		2		2		2	
	Ricinus communis		2		1		2		1		1		1		1	
	Populus trichocarpa		2		2		2		2		1		2		2	
	Arabidopsis thaliana		3		5		1		3		2		2		2	
	Arabidopsis lyrata		3		5		5		1		3		2		2	
	Carica papaya		2		1		2		E-88		2		2		1	
	Theobroma cacao		2		1		1		1		4		1		1	
Однодольные	*Vitis vinifera*		2		1		2		1		1		1		1	
	Oryza sativa spp japonica		2		2		1		1		2		5		1	
	Brachypodium distarhion		2		1		1		1		2		5		1	
	Sorghum bicolor		1		1		1		2		3		2		1	
	Zea mays		3		5		1		2		2		6		2	
Плауны	*Selaginella moellendorffii*		1		1		1		1		**E-56**		1		**E-152**	
Мхи	*Physcomitrella patens*		2		2		1		1		1		1		4	
Зеленые водоросли	*Ostreococcus lucimarinus*		1		1		1		1		1		1		1	
	Ostreococcus tauri		1		1		1		1		1		1		1	
	Volvox carteri		1		1		1		1		1		1		1	
	Chlamidomonas reinhardtii	–	1		1		1		1		1		1		1	
Таксон	ферменты пути биосинтеза триптофана	ASA	ASB		TRP		PAI		IGPS		TSA		TSB			

Примечание. Цифры в ячейках таблицы – количество гомологов фермента, выявленное при анализе полных геномов. Знак E с цифрами – значения уровня сходства (e-value) с последовательностями ферментов *A. thaliana* для гомологов, выявленных по EST. Прочерк обозначает, что гомологи данного фермента не обнаружены ни среди полногеномных данных, ни в библиотеках EST.

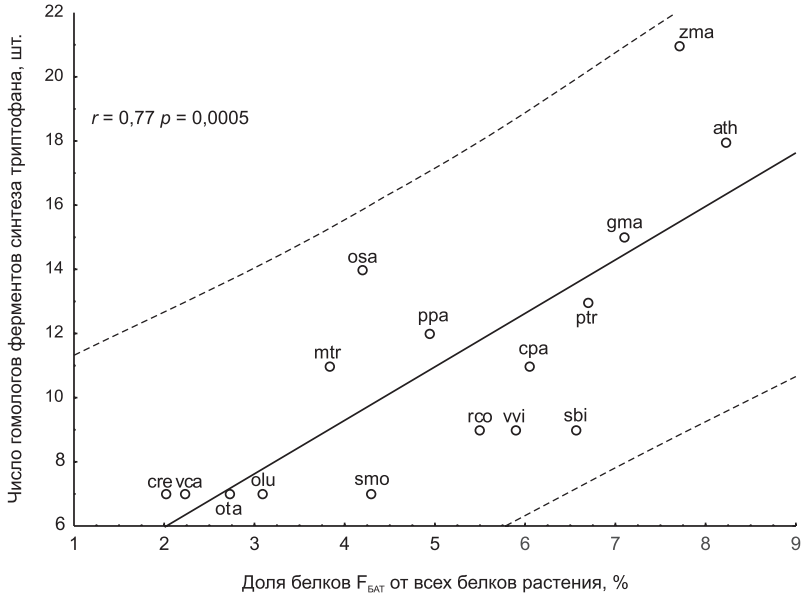
параметром FБАТ растений и числом гомологов ферментов ПБТ выявил статистически значимую (p = 0,0005) положительную корреляцию как для отдельных ферментов (ASA/ASB (обе субъединицы), PAI и IGPS; табл. 2), так и для их суммарного числа (см. рисунок).

Таблица 2

Зависимость числа гомологов пути биосинтеза триптофана от параметра $F_{БАТ}$ для каждого из ферментов этого пути

Семейство белков	*r*	*p*
ASA	0,56	0,0025*
ASB	0,69	0,031*
TRP	0,29	0,27
PAI	0,79	0,0033*
IGPS	0,59	0,015*
TSA	0,43	0,094
TSB	0,38	0,15

Примечание. Представлены коэффициенты корреляции (*r*) и уровни статистической значимости (*p*) для этих двух параметров. * достоверные значения ($p < 0,05$).



Зависимость числа гомологов ферментов пути биосинтеза триптофана от доли белков БАТ среди всех белков растения, $F_{БАТ}$. По оси *X* отложено значение параметра $F_{БАТ}$. По оси *Y* – суммарное число гомологов по всем ферментам пути биосинтеза триптофана. Коэффициент корреляции и уровень его значимости приведены на графике. Штриховыми линиями обозначены границы 95 % доверительного интервала.

cre – *Chlamidomonas reinhardtii*, ota – *Ostreococcus tauri*, olu – *Ostreococcus lucimarinus*, vca – *Volvox carteri*, ppa – *Physcomitrella patens*, smo – *Selaginella moellendorffii*, zma – *Zea mays*, sbi – *Sorghum bicolor*, osa – *Oryza sativa spp japonica*, vvi – *Vitis vinifera*, ath – *Arabidopsis thaliana*, ptr – *Populus trichocarpa*, cpa – *Carica papaya*, rco – *Ricinus communis*, gma – *Glycine max*, mtr – *Medicago truncatula*.

ОБСУЖДЕНИЕ

Сравнительный анализ функций белков в парах паралогов ферментов ПБТ, выявленных с помощью нокаутного анализа их генов у некоторых видов семенных растений, свидетельствует о специфичной экспрессии разных гомологов одного фермента синтеза триптофана в зависимости от условий среды. Например, у *A. thaliana* ген *ASA2* экспрессируется на конститутивном базальном уровне, а уровень экспрессии его паралога *ASA1* в десять раз выше и может дополнительно повышаться в ответ на ранение или инфицирование бактериальными патогенами (Niyogi, Fink, 1992). Такие различия характерны и для паралогов *ASA1*, *ASA2* у *Ruta graveolens* (Bohlmann *et al.*, 1995), а также *OASA1*, *OASA2* у однодольного растения *Oryza sativa japonica* (Tozawa *et al.*, 2001). Ген, кодирующий TSB2 у *A. thaliana*, продуцирует только 10 % мРНК триптофан синтазы β в тканях листа, экспрессируется конститутивно на базальном уровне и необходим для роста растения при недостатке освещения. Его паралог, ген, кодирующий TSB1, экспрессирует 90 % мРНК триптофан синтазы β, но лишь при ярком освещении (Last *et al.*, 1991). Эти данные демонстрируют важное значение дупликаций в ПБТ: один ген из пары экспрессируется на конститутивном базальном уровне и требуется для синтеза белков и вторичных метаболитов триптофана на минимальном уровне, необходимом, например, для биосинтеза белков, а второй ген пары экспрессируется индуцибельно на высоком уровне в ответ на внешние условия, благоприятные для быстрого роста и развития растения, или на стрессовые факторы (бактериальную инфекцию, ранения, недостаток освещения и др.), повышая концентрацию триптофана, часть молекул которого может служить субстратом для синтеза ауксина, участвуя в регуляторных процессах.

Растения в ходе эволюции были вынуждены приспосабливаться к увеличению изменчивости условий внешней среды (резкое увеличение пессимальности условий и амплитуды их изменений при выходе растений из воды на сушу) за счет усложнения их морфологии: увеличение количества тканей (появление корней, листьев, развитого стебля, тканей, позволивших освоить сушу папоротникам, хвощам, плаунам; переход от спор к семенам – у семенных папоротников, позволивший растениям оторваться от экосистем с высокой влажностью; появление тканей цветов и плодов у покрытосеменных растений) и усложнения молекулярно-генетических систем растений. Это усложнение морфологии позволило растениям занимать в процессе эволюции все большее число экологических ниш, а также освоить ниши со значительными колебаниями условий (жара – холод, дожди – засуха и т. д.). Это обстоятельство может объяснить выявленные нами положительные корреляции между долей генов БАТ (мерой сложности организма) и числом гомологов ферментов синтеза триптофана как для всего пути биосинтеза в целом, так и для трех (ASA/ASB, PAI и IGPS) из шести его ферментов в частности.

БЛАГОДАРНОСТИ

Работа поддержана грантом РНФ 14-14-00734.

ЛИТЕРАТУРА

Картель Н.А., Макеева Е.Н., Мезенко А.М. Генетика = Genetics: энциклопедический словарь. Минск: Беларус навука, 2011. 758 с.

Колчанов Н.А., Суслов В.В., Гунбин К.В. Моделирование биологической эволюции: регуляторные генетические системы и кодирование биологической организации // Информационный вестник ВОГИС. 2004. Т. 8. № 2. С. 86–99.

Bohlmann J., DeLuca V., Eilert U., Martin W. Purification and cDNA cloning of anthranilate synthase from Ruta graveolens: modes of expression and properties of native and recombinant enzymes // Plant J. 1995. V. 7. No. 3. P. 491–501.

Feliner G.N., Rossello J.A. Concerted evolution of multigene and homoelogous recombination families. Plant genome diversity. Springer Wein Heidelberg, N. Y., Dordrecht, London, 2012. 171 p.

Huson D.H., Bryant D. Application of phylogenetic networks in evolutionary studies // Mol Biol Evol. 2006. V. 23. No. 2. P. 254–267.

Katoh K., Toh H. Recent developments in the MAFFT multiple sequence alignment program // Brief Bioinform. 2008. V. 9. No. 4. P. 286–298.

Lamesch P., Berardini T.Z., Li D. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools // Nucleic Acids Res. 2012. V. 40. P. D1202–D1210.

Lang D., Weiche B., Timmerhaus G. *et al.* Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and

correlation with complexity // Genome Biol. Evol. 2010. V. 19. No. 2. P. 488–503.
Last R.L., Bissinger P.H., Mahoney D.J. *et al.* Tryptophan mutants in Arabidopsis: the consequences of duplicated tryptophan synthase beta genes // Plant Cell. 1991. V. 3. No. 4. P. 345–358.
Marchler-Bauer A., Zheng C., Chitsaz F. *et al.* CDD: conserved domains and protein three-dimensional structure // Nucleic Acids Research. 2013. V. 41. P. D348–D352.
Niyogi K.K., Fink G.R. Two anthranilate synthase genes in Arabidopsis: defense-related regulation of the tryptophan pathway // Plant Cell. 1992. V. 4. No. 6. P. 721–733.
Radwanski E.R., Last R.L. Tryptophan biosynthesis and metabolism: biochemical and molecular genetics // Plant Cell. 1995. V. 7. P. 921–934.
Richardson A.O., Palmer J.D. Horizontal gene transfer in plants // Journal experimental Botany. 2007. V. 58. No. 1. P. 1–9.
Teichmann S.A., Babu M.M. Gene regulatory network growth by duplication // Nat Genet. 2004. V. 36. No. 5. P. 492–496.
Tozawa Y., Hasegawa H., Terakawa T., Wakasa K. Characterization of rice anthranilate synthase alpha-subunit genes OASA1 and OASA2. Tryptophan accumulation in transgenic rice expressing a feedback-insensitive mutant of OASA1 // Plant Physiol. 2001. V. 126. No. 4. P. 1493–1506.
Van Bel M., Proost S., Wischnitzki E. *et al.* Dissecting plant genomes with the PLAZA comparative genomics platform // Plant Physiology. 2012. V. 158. No. 2. P. 590–600.
Vogel C., Chothia C. Protein family expansions and biological complexity // PLoS Comput Biol. 2006. V. 2. No. 5. P. e48.
Zhao Y. Auxin biosynthesis: a simple two-step pathway converts tryptophan to indole-3-acetic acid in plants // Mol. Plant. 2012. V. 5. No. 2 P. 334–338.

THE NUMBER OF HOMOLOGS OF SOME ENZYMES IN THE TRYPTOPHAN BIOSYNTHESIS PATHWAY CORRELATES WITH THE PROPORTION OF PROTEINS ASSOCIATED WITH TRANSCRIPTION IN PLANTS

I.I. Turnaev[1], I.R. Akberdin[1], V.V. Suslov[1], D.A. Afonnikov[1, 2]

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

Summary

The tryptophan biosynthesis pathway (TBP) is ubiquitous in most known organisms, being absent only from animals and some bacteria. It is conserved in plants, although various species differ in the number of TBP enzyme paralogs. In the current work we investigated a putative possible role of changes in the number of paralogs of TBP enzymes in the course of plant evolution. We identified TBP enzyme paralogs in plant species with fully sequenced genomes and estimated the relationship between its number and organismal complexity. It is shown that organismal complexity significantly correlates with the total number of TBP paralogs and for some enzymes specifically (ASA/ASB, PAI, and IGPS). We suggest that such a relationship arises because both organismal complexity and the increasing number of paralogs may be important for the evolutionary adaptation of land plants to variable environmental conditions.

Key words: tryptophan biosynthesis pathway, phylogenetic network, morphological complexity of organisms, adaptation, variability of environmental conditions.

УДК 573.22

HIGH-PERFORMANCE SIMULATIONS OF POPULATION-GENETIC PROCESSES IN BACTERIAL COMMUNITIES USING THE HAPLOID EVOLUTIONARY CONSTRUCTOR SOFTWARE

© 2014 г. **Z.S. Mustafin[1], Yu.G. Matushkin[1, 2], S.A. Lashin[1, 2]**

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

Received October 17, 2014. Accepted for publication November 27, 2014.

Three high-performance versions of the Haploid Evolutionary Constructor program are presented (http://evol-constructor.bionet.nsc.ru). The software was designed for simulating the functioning and evolution of microbial communities. These high-performance versions are to be run on systems with shared and distributed memory, using CPU and/or GPU. Almost linear acceleration has been achieved on clusters and multi-core CPU. On GPU systems, the simulation time was reduced to several minutes (dozens of hours on CPU).

Key words: Microbial community, parallelization, simulation, evolution, optimization.

INTRODUCTION

Simulation of evolutionary processes occurring in bacterial communities is one of the vital tasks of modern systems biology. Bacteria perform a vast majority of processes in nature. Many bacterial species are utilized to meet human needs. However, some bacterial communities reach such a high genetic diversity and huge population size that they cannot be investigated under laboratory conditions. Theoretical studies including mathematical modeling and simulation may be helpful in these cases. Information technology development has led to the appearance of a variety of programs devoted to modeling and simulation of various aspects of bacterial communities' life. Evolution and functioning of such communities in certain conditions (including infeasible for laboratories) can be modeled for purposes of medicine, fundamental and applied science.

In recent years, many papers on modeling and simulation of various features of bacterial communities have been published. Some papers are focused on the biological sense of simulation results, such as interconnections between individual and populational features of communities and their members (Kutalik *et al.*, 2005) or mechanisms of biodiversity sustainability in various fitness landscapes and at various mutation rates (Beardmore *et al.*, 2011). Others studied mathematical and programming features (Ashlock, McEachern, 2011; Bihary *et al.*, 2012). Applicability and advantages/disadvantages of agent-based (DeAngelis, Mooij, 2005) approaches, or cellular automata (Esteban, Rodrнguez-Patyn, 2011) have been also analyzed and compared with classical ODE and PDE equations. In spite of the multitude of modeling methods and software packages for simulation of bacterial communities, most of them consider a system under study at only one level of biological organization. Furthermore, few of them use modern technologies for high-performance computations.

This study is dedicated to the development of high-performance methods for simulating the functioning and evolution of bacterial communities (more generally, communities of unicellular haploid microorganisms). The method has been implemented as part of the Haploid Evolutionary Constructor software package (hereafter referred to as the HEC, http://evol-constructor.bionet.nsc.ru). HEC models are multiscale, and they include submodels describing different levels of biological organization: genetic, metabolic, population, and ecological (Lashin *et al.*, 2011; Lashin, Matushkin,

2012). Such composite models consume a lot of computational resources, especially in the case of communities of extremely broad genetic diversity (about 10^8 various allelic combinations in a population considering 10–100 model genes), which results in long simulation time. The paper also presents high-performance algorithms for HEC and test results. Three high-performance implementations have been made: OpenMP (http://openmp.org), MPI (http://www.open-mpi.org), and CUDA (http://www.nvidia.com/object/cuda_home_new.html).

HEC software package

HEC uses the multiscale modeling approach (Ayton *et al.*, 2007; Martins *et al.*, 2010). Four layers of biological organization are considered: genetic, metabolic, populational, and ecological (Lashin, Matushkin, 2012). Models of every layer can be implemented with various mathematical techniques (differential equations, automata, graphs, etc.). For each layer, libraries of submodels are released (Lashin, Matushkin, 2012). Notably, models of gene networks can also be implemented as HEC plugins. Such a multilayered approach allows users to study various aspects of a bacterial community within an integral framework.

Polymorphic population is described via the "*generalized population genome*" and the genetic spectrum technique (Lashin *et al.*, 2010), which affords a valuable decrease of the computational time as compared to classical agent-based approaches. This method also ensures comparable accuracy. An organism (cell) is characterized by a set of traits, each of which determines the process of either synthesis or utilization of a particular metabolite (substrate). In HEC, the whole network of those processes is assumed to be a "gene network" of the cell. Such a "gene network" can be formally implemented by using, for example, differential equations. Parameters of such a gene network are assumed to be *genes*, whereas particular values of these genes are assumed to be *alleles*. Cells belonging to the same population may possess different allelic combinations (ACs). The total number of ACs in a population characterizes its genetic diversity. Various ACs may be differently efficient in substrate synthesis and/or utilization, which results in different fitnesses and reproduction rates of subpopulations in the entire polymorphic population. HEC allows simulation of mutations, horizontal transfer of genes and gene loss. The last two change the set of metabolic reactions and, thereby, the gene network of a cell, generating a new strain/species. This feature allows HEC to model speciation and evolution of biodiversity in the community, which can be simulated either in complete-mixed or in spatially distributed environments.

The variation in reproduction rates depending upon both genetic and environmental factors allows us to simulate a wide range of evolutionary modes including neutral evolution. Other features of HEC are the simulation of phage infections (Lashin *et al.*, 2011) and gene networks (Lashin, Matushkin, 2012). Integration of the gene network concept with the HEC opens exciting possibilities for investigation of gene network evolution at the over-genetic and over-organism levels of biological organization, such as populational or ecological.

The genetic diversity of a community impacts the computational time

The most time-consuming procedure in the HEC computational process is the simulation of the reproduction of populations. When a broadly diverse (10^6–10^8 unique ACs) community is simulated, almost all computational time is consumed by this function (Fig. 1).

Figure 2 shows an example of a generalized population genome in HEC. It is just a multidimensional distribution of allelic frequencies for all genes present in cells of this population. This

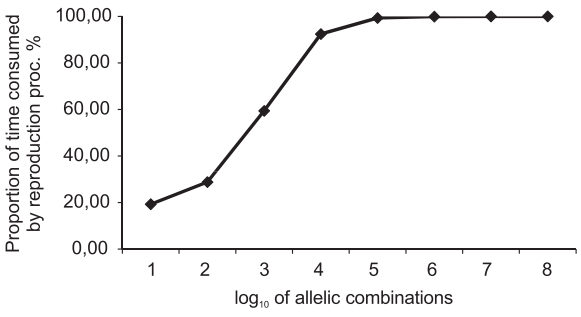


Fig. 1. Proportion of time consumed by the reproduction procedure in relation to the overall computational time.

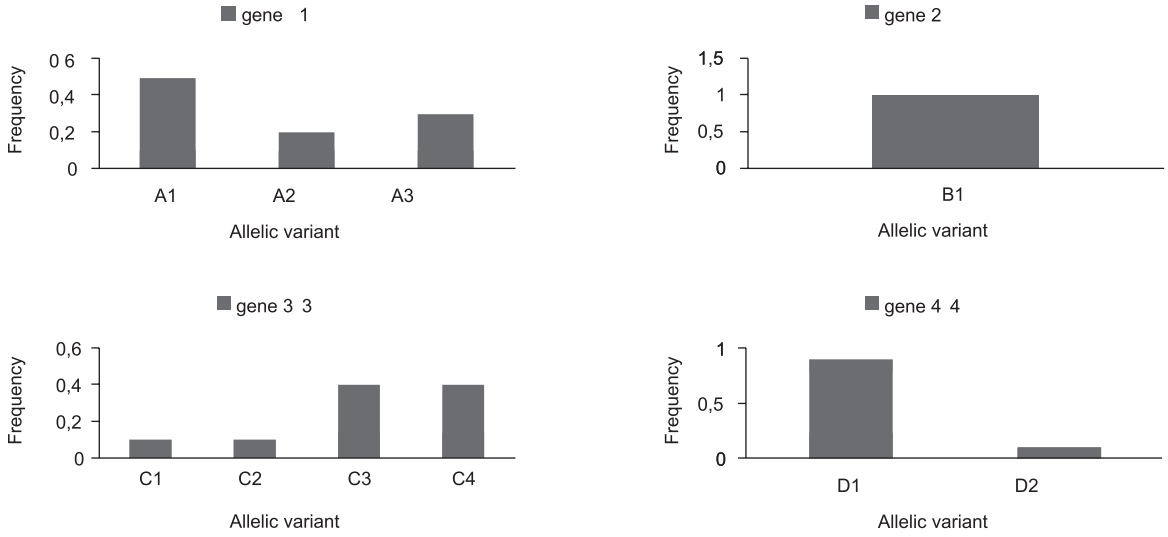


Fig. 2. Example of a generalized population genome in the HEC:

A1(0.5) A2(0.2) A3(0.3)	distribution of alleles for gene 1
B1(1)	distribution of alleles for gene 2
C1(0.1) C2(0.1) C3(0.4) C4(0.4)	distribution of alleles for gene 3
D1(0.9) D2(0.1)	distribution of alleles for gene 4

generalized genome contains four genes, which have three, one, four, and two possible allelic variants present in the population, respectively. By multiplying the numbers of allelic variants, we obtain the total AC number (for example, $3 \times 1 \times \times 4 \times 2 = 24$, as shown below).

In order to calculate the total population size change, it is necessary to calculate it for each AC subpopulation (cells with identical genome), and then rearrange new allelic frequencies in the population (Lashin *et al.*, 2010). This routine requires cycles and cycles of the same function calls where only the AC changes. Thus, it can be and should be parallelized.

High-performance versions of the HEC

The initial reproduction procedure used a recursive algorithm for iteration over allelic combinations, which was unsuitable for parallelization. The iteration algorithm and internal data representation were modified (Mustafin *et al.*, 2012), which resulted in an elegant parallelization scheme (fig. 3) upon which high-performance implementations could be made. Several high-performance versions of the reproduction procedure were developed and tested: OpenMP, MPI, and CUDA ones.

The OpenMP version has been developed for the desktop version of HEC primarily along with a graphic user interface. It is effective in modeling mid- and high-diverse communities (> 100 AC). Computations for models of low-diverse communities (< 100) themselves take less time than data exchange between processes. It is ineffective to parallelize such models. The optimal number of parallel threads for this version should be divisible by the number of populations. It is also desirable that simulated populations should have roughly equal levels of genetic diversity in order to obtain the optimal thread load. The OpenMP version is suited for high frequency/few-core processors (In contrast with MPI, OpenMP gives an acceleration even when the number of threads exceeds the number of processor cores.)

An MPI version to be used with console versions of HEC is being developed. It is effective when it comes to models of genetic diversity more than 100 ACs. Communities of any number of populations with any genetic diversity can be simulated with minimal efficiency loss (as the genetic diversity increases, the data exchange time defies evaluation). Models with high genetic diversity (> 10^5 ACs) show linear efficiency growth. The tendency breaks only when the software is run with

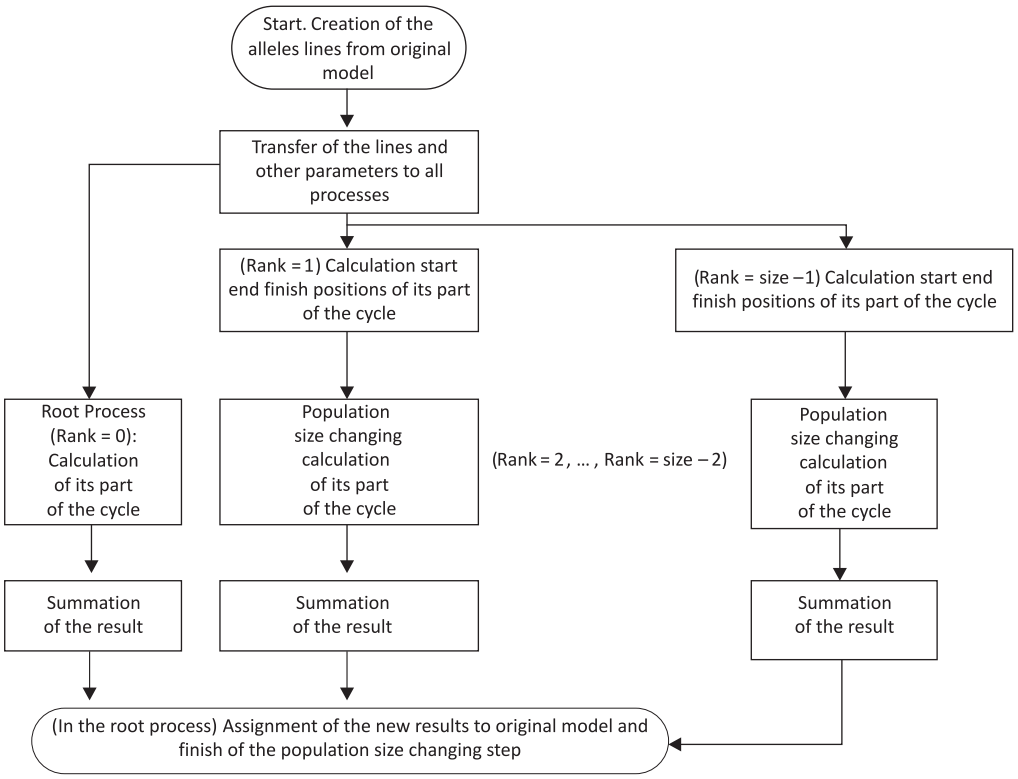


Fig. 3. Parallelization scheme for the reproduction procedure.

The MPI version implies that each thread receives its portion of allelic combinations, performs calculations, and returns data to the root node (MPI_BROADCAST and MPI_REDUCE are used). The CUDA version saves data to the memory of the video card, and then GPU performs calculations and data return to the main process.

very high numbers of parallel threads (typically for supercomputer clusters). The MPI version is suited for supercomputer clusters and personal computers with several core processors. It can also replace the OpenMP version, while in most cases it shows just as good results as OpenMP.

The CUDA version is developed to run on computers with NVidia CUDA graphic accelerators. It is effective only on models of high and extremely high genetic diversity. If AC is less than 10^5, the version is awfully ineffective due to memory allocation (on a video accelerator) and copying data takes more time than computation procedures. An obvious advantage of the CUDA version is the possibility to use extremely high numbers of threads (more than 1000-fold compared to OpenMP or MPI). With this version, models of genetic diversity exceeding 10^6 ACs can be calculated much faster than with any other. Simulation results and tests are presented in the following sections (see also tables 1–4).

Test calculations

Test simulations were performed on the AMD Phenom II ×6 1055T (6-core) processor and the NVIDIA GTX 570 video accelerator. The MPI version was also tested on the NKS-30T supercomputer cluster (http://bioinformatics.bionet.nsc.ru/). The tests were performed on the set of test models published previously (Lashin *et al.*, 2010; Lashin, Matushkin, 2012). Furthermore, special load tests were used.

Table 1 shows the test results of the MPI version. Simulation time depended on several parameters, the main of which were AC number (as the measure of community genetic diversity) and the number of iterations (i.e. generations) per simulation run. Parallelization efficiency also depends on these parameters. In low genetic diversity (1–100 ACs) models, parallelization is ineffective, as the execution time of the reproduction procedure is low with respect to the overall execution time. In the range from 100 to 1000 ACs, the paralleliza-

Table 1

Test results for the MPI version on AMD Phenom II ×6 1055T (2.8 GHz)

Number of allelic combinations	Iterations (generations)	Average simulation time, s; parallelization efficiency, % Number of parallel threads					
		1	2	3	4	5	6
10^3	25 000	53 s	45 s – 58 %	44 s – 40 %	44 s – 30 %	45 s – 23 %	52 s – 17 %
5×10^3	10 000	73 s	46 s – 79 %	38 s – 64 %	34 s – 54 %	33 s – 44 %	37 s – 33 %
10^4	5 000	68 s	39 s – 87 %	31 s – 73 %	27 s – 63 %	25 s – 54 %	27 s – 42 %
10^5	500	85 s	48 s – 89 %	35 s – 81 %	28 s – 76 %	24 s – 71 %	23 s – 62 %
10^6	50	162 s	88 s – 92 %	61 s – 89 %	48 s – 84 %	40 s – 81 %	37 s – 73 %
10^7	4	199 s	101 s – 99 %	69 s – 96 %	55 s – 90 %	46 s – 87 %	44 s – 75 %

The number of generations per simulation decreases with the growth of genetic diversity. It was made in order to make the test last for several minutes. The values are rounded up or down to the nearest integer.

Table 2

Test results for the OpenMP version on AMD Phenom II ×6 1055T (2.8 GHz)

Number of allelic combinations	Iterations (generations)	Average simulation time, s; parallelization efficiency, % Number of parallel threads			
		1	2	4	8
10^3	25 000	49 s	39 s – 63 %	37 s – 33 %	35 s – 18 %
5×10^3	10 000	65 s	41 s – 80 %	29 s – 56 %	28 s – 29 %
10^4	5 000	59 s	35 s – 84 %	23 s – 64 %	21 s – 35 %
10^5	500	75 s	49 s – 77 %	23 s – 82 %	21 s – 45 %
10^6	50	147 s	81 s – 91 %	45 s – 82 %	34 s – 54 %
10^7	4	186 s	99 s – 94 %	73 s – 64 %	48 s – 48 %

Table 3

Comparison of parallelization efficiency for the OpenMP and MPI versions

Number of allelic combinations	Iterations (generations)	Parallelization efficiency, %			
		2 threads		4 threads	
		MPI	OpenMP	MPI	OpenMP
10^3	25 000	58	63	30	33
5×10^3	10 000	79	80	54	56
10^4	5 000	87	84	63	64
10^5	500	89	77	76	82
10^6	50	92	91	84	82
10^7	4	99	94	90	64

Table 4

Comparison of the best computational times obtained using various high-performance versions of HEC

Number of allelic combinations	Iterations (generations)	Minimal computational time, s (optimal number of parallel threads)		
		MPI	OpenMP	CUDA
10^2	25,000	23 (1)	22 (1)	318 (1000)
10^3	25,000	43 (3)	35 (8)	335 (500)
5×10^3	10,000	33 (5)	28 (8)	128 (1000)
10^4	5000	25 (5)	21 (8)	65 (500)
10^5	500	23 (6)	21 (8)	12 (1000)
10^6	50	37 (6)	34 (8)	3 (10,000)
10^7	4	43 (6)	47 (8)	3 (5000, 10,000, 50,000)
10^8	1	×	×	5 (50,000, 100,000)

× Failure with AMD Phenom II ×6 1055T

tion effect becomes apparent and depends on the number of iterations per simulation. In models of high-average genetic diversity (1000–10,000 ACs), parallelization is more effective, and the iteration number exerts next to no effect on efficiency. In models with AC > 10,000, efficiency does not depend on the number of iterations. With AC increase, the total efficiency approaches unity, but the more parallel threads run, the less efficiency is kept, although the computation time decreases continuously.

Test results for the OpenMP version are presented in Table 2. The main difference from MPI is that OpenMP loses more efficiency when the number of threads grows. Comparison of these two versions is shown in Table 3.

Finally, the test results for the CUDA version were compared with OpenMP and MPI. We compared minimal obtained times for each version (Table 4). In the table, the optimal number of threads for each table cell is shown in parentheses. Thus, the CUDA version can significantly speed up simulations in computationally costly (AC>10^5) tasks.

Choosing the optimal parallel version

We classified parallel versions described above according to the most suitable simulation tasks for each model. They are listed below:

(1) Models of communities with genetic diversity less than 100 ACs are better simulated with the use of non–parallel HEC because the reproduction procedure in such a simple case takes only a small amount of the total time. Sometimes parallel versions even increase the simulation time. Only the OpenMP version works with the same efficiency (no data exchange), while the MPI version takes 15–20 % more time. We strongly recommend not using the CUDA version in simple tasks, as it may increase the running time.

(2) Models of communities with genetic diversity of 100–10,000 ACs are better simulated with the use of either MPI or OpenMP versions. The efficiency of both models grows in proportion to the increase of AC. The efficiency of the CUDA version grows even more, but its overall computational time is longer than in the above case.

(3) Models of communities with genetic diversity of 10^5–10^6 AC are suitable for all three versions.

(4) Finally, communities of extremely high genetic diversity (> 10^6 AC) are better simulated with the CUDA version. It takes much less time for simulation than with MPI or OpenMP.

CONCLUSION

In this paper, we present several high-performance versions of the HEC software packages, available at our web site. Under appropriate conditions (models of "optimal" genetic diversity), they provide nearly linear acceleration. Parallel versions are classified according to the most suitable situations for their usage. The non-parallel version is the best for simple models of low genetic diversity. All the three parallel versions are appropriate for models of intermediate genetic diversity. Finally,

the CUDA version is the best for extremely diverse communities. We think that the last case is the most interesting for large-scale theoretical studies, and we hope that high-performance versions of HEC presented in this paper will allow users to investigate more complex and diverse microbial communities and will produce new interesting biological results.

ACKNOWLEDGMENTS

The study was supported by the following grants: RFBR 12-07-00671-a, project VI.61.1.2, and interdisciplinary SB RAS project 47.

LITERATURE

Ashlock D., McEachern A. A simulation of bacterial communities // IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE. 2011. P. 1–8.

Ayton G.S., Noid W.G., Voth G.A. Multiscale modeling of biomolecular systems: in serial and in parallel // Curr. Opin. Struct. Biol. 2007. V. 17. No. 2. P. 192–198.

Beardmore R.E., Gudelj I., Lipson D.A., Hurst L.D. Metabolic trade–offs and the maintenance of the fittest and the flattest // Nature. 2011. V. 472. No. 7343. P. 342–346.

Bihary D., Kerenyi A., Gelencser Z. *et al*. Simulation of communication and cooperation in multispecies bacterial communities with an agent based model // Scalable Comput. Pract. Exp. 2012. V. 13. No. 1. P. 21–28.

DeAngelis D.L., Mooij W.M. Individual-based modeling of ecological and evolutionary processes // Annu. Rev. Ecol. Evol. Syst. 2005. V. 36. No. 1. P. 147–168.

Esteban P.G., Rodríguez-Patón A. Simulating a Rock – Scissors – Paper Bacterial Game with a Discrete Cellular Automaton // New Challenges on Bioinspired Applications, Lecture Notes in Computer Science / Eds. Ferrández J.M., Álvarez Sánchez J.R., De la Paz F., Toledo F.J. Berlin Heidelberg: Springer, 2011. P. 363–370.

Kutalik Z., Razaz M., Baranyi J. Connection between stochastic and deterministic modelling of microbial growth // J. Theor. Biol. 2005. V. 232. No. 2. P. 285–299.

Lashin S.A., Matushkin Y.G. Haploid evolutionary constructor: new features and further challenges // In Silico Biol. 2012. V. 11. No. 3–4. P. 125–135.

Lashin S.A., Matushkin Y.G., Suslov V. V., Kolchanov N.A. Evolutionary trends in the prokaryotic community and prokaryotic community-phage systems // Russ. J. Genet. 2011. V. 47. No. 12. P. 1487–1495.

Lashin S.A., Suslov V.V., Matushkin Yu.G. Comparative modeling of coevolution in communities of unicellular organisms: adaptability and biodiversity // J. Bioinform. Comput. Biol. 2010. V. 8. No. 3. P. 627–643.

Martins M.L., Ferreira S.C., Vilela M.J. Multiscale models for biological systems // Curr. Opin. Colloid Interface Sci. 2010. V. 15. No. 1–2. P. 18–23.

Mustafin Z.S., Matushkin Y.G., Lashin S.A. Haploid Evolutionary Constructor: parallelization and high performance simulations of evolution of prokaryotic communities // Russ. J. Genet. Appl. Res. 2012. V. 16. No. 4/1. P. 825–829.

Федеральное государственное унитарное предприятие «ПОЧТА РОССИИ»

Бланк заказа периодических изданий Ф. СП-1

АБОНЕМЕНТ на газету / **журнал** **4 2 1 5 3**

(индекс издания)

Вавиловский журнал

(наименование издания)

генетики и селекции

Количество Комплектов

на 20__год по месяцам

1	2	3	4	5	6	7	8	9	10	11	12

Куда

(почтовый индекс) (адрес)

Кому

(фамилия, инициалы)

---------- Линия отреза ----------

ПВ | Место | Литер

ДОСТАВОЧНАЯ КАРТОЧКА **4 2 1 5 3**

(индекс издания)

На Газету / Журнал **Вавиловский журнал генетики и селекции**

(наименование издания)

Стоимость

подписки	руб.	Кол-во комп-лектов	
переадресовки	руб.		

на 20__год по месяцам

1	2	3	4	5	6	7	8	9	10	11	12

	город
	Село
почтовый индекс	область
	район
код улицы	улица
дом / корпус	квартира

Фамилия И.О.

Отредактировано и подготовлено к печати
в редакционно-издательском отделе ИЦиГ СО РАН

Ответственный секретарь выпуска: Е.А. Ананько
Редактор: Т.Ф. Чалкова
Корректор: Н.А. Митарновская
Дизайн: А.В. Харкевич
Компьютерная графика и верстка: А.В. Харкевич

Подписано в печать 15.12.2014 г.
Формат бумаги 60×84 1/8. Усл.-печ. л. 23,95. Уч.-изд. л. 22,47
Тираж 250 экз. Заказ 299

Отпечатано в типографии ФГУП «Издательство СО РАН»
630090, г. Новосибирск, Морской проспект, 2