Научный рецензируемый журнал

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

Основан в 1997 г.
Периодичность 6 выпусков в год

Учредители

Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук»

Межрегиональная общественная организация Вавиловское общество генетиков и селекционеров

Сибирское отделение Российской академии наук

December 2015 • 19 • 6
Scientific Peer Reviewed Journal

VAVILOV JOURNAL OF GENETICS AND BREEDING

Vavilovskii Zhurnal Genetiki i Selektsii

Founded in 1997
Published 6 times annually

Founders
Federal State Budget Scientific institution "The Federal Research Center Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences"
The Vavilov Society of Geneticists and Breeders
Siberian Branch of the Russian Academy of Sciences

Vavilov Journal of Genetics and Breeding

Content • 2015 • 19 • 6

Ontologies

Genomics and Polymorphism analysis

Plant Bioinformatics

Computer Simulation

Вавиловский журнал генетики и селекции

содержание • 2015 • 19 • 6

Онтологии

Геномика и анализ полиморфизмов

Биоинформатика растений

Компьютерное моделирование

Николай Александрович Колчанов
академик, директор ИЦиГ СО РАН

Николай Леонтьевич Подколодный
заведующий Центром коллективного пользования «Биоинформатика»

Развитие новых высокопроизводительных экспериментальных технологий в молекулярной биологии и генетике привело к (1) накоплению больших объемов молекулярно-биологических данных; (2) необходимости массового анализа этих данных методами биоинформатики и системной биологии; (3) интеграции гигантских объемов гетерогенных экспериментальных биологических данных (Big data) для получения комплексного представления о структурно-функциональных особенностях различных иерархических уровней организации биологических систем (молекулярно-генетический, клеточный, организменный, экосистемный); (4) широкому применению методов математического моделирования для изучения механизмов генетического контроля биологических систем и процессов; (5) необходимости высокопроизводительных (суперкомпьютерных) вычислений при решении перечисленных выше задач.

Настоящий выпуск журнала содержит статьи по ряду актуальных направлений биоинформатики и системной компьютерной биологии.

Раздел «Онтологии» представляет обзор биологических онтологий и их применение для решения задач биоинформатики и системной биологии; описаны задачи, при решении которых использование онтологий дает ощутимый эффект (семантическая интеграция и анализ гетерогенных экспериментальных данных, построение математических моделей молекулярно-генетических систем и процессов, компьютерная интерпретация молекулярно-генетических знаний и др.).

В разделе «Геномика и анализ полиморфизмов» представлены статьи, в которых описаны: применение высокопроизводительных вычислений для выявления функциональных сигналов в регуляторных районах генов прокариот; результаты анализа влияния генетической изменчивости промоторов генов человека на связывание белка TBP; закономерности распределения контекстной сложности геномных районов, содержащих однонуклеотидные полиморфизмы человека, а также роль однонуклеотидных полиморфизмов в генах при доместикации свиньи.

Раздел «Биоинформатика растений» включает статью по идентификации микросателлитных локусов по данным секвенирования BAC-клонов и их картированию на геном мягкой пшеницы, что актуально для выявления полиморфных маркеров для участков хромосом, определяющих хозяйственно ценные признаки, а также обзор механизмов устойчивости растений к патогенам, связанных с особенностями строения их клеточной стенки.

Раздел «Компьютерное моделирование» содержит статьи по применению компьютерного моделирования в таких областях биоинформатики и системной биологии, как дизайн малых химических соединений – ингибиторов белков; моделирование биологически активных пептидов; решение обратных задач математической биологии в применении к биомедицине, а также компьютерное агентное моделирование популяций микробных сообществ.

Вавиловский журнал генетики и селекции. 2015;19(6):652-660
Doi 10.18699/VJ15.090

Онтологии в биоинформатике и системной биологии

Н.Л. Подколодный[1, 2, 3], О.А. Подколодная[1]

1 Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия 2 Федеральное государственное бюджетное учреждение науки Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия
3 Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный

Компьютерное моделирование в настоящее время становится центральной научной парадигмой системной биологии и основным инструментом для теоретического исследования и понимания механизмов функционирования сложных живых систем. Увеличение количества и сложности этих моделей приводит к необходимости их коллективной разработки, повторного использования, верификации, описания вычислительного эксперимента и его результатов. При разработке форматов представления знаний для математического моделирования биологических систем активно применяют онтологическое моделирование предметной области. В этом смысле онтологию, связанную со всей совокупностью форматов, обеспечивающих поддержку исследований в системной биологии, в частности компьютерное моделирование биологических систем и процес-сов, можно считать первым приближением к онтологии системной биологии. В обзоре кратко представлены особенности предметной области (биоинформатика, системная биология, биомедицина), основные мотивации в развитии онтологий и наиболее важные примеры онтологического моделирования и семантического анализа на разных уровнях иерархии знаний: молекулярно-генетическом, клеточном, тканевом, органов и организма. Биоинформатика и системная биология являются прекрасным полигоном для отработки технологий и эффективного использования онтологического моделирования. Создание нескольких десятков базовых ссылочных онтологий и их верификация позволяют использовать эти онтологии в качестве источников знаний для интеграции и построения более сложных моделей предметной области, ориентированных на решение конкретных задач биомедицины и биотехнологии. Дальнейшая формализация и накопление онтологических знаний, а также использование формальных методов их анализа могут поднять весь цикл научных исследований в области системной биологии на новый технологический уровень.

Ключевые слова: онтологическое моделирование; биоинформатика; системная биология.

HOW TO CITE THIS ARTICLE?
Podkolodnyy n.l ., Podkolodnaya o .A. o ntologies in bioinformatics and systems biology. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2015;19(6):652-660. Doi 10.18699/V J15.090

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?
Подколодный Н.л., Подколодная О.А. Онтологии в биоинформатике и системной биологии. Вавиловский журнал генетики и селекции. 2015;19(6):652-660. Doi 10.18699/VJ15.090

Received 28.09.2015
Accepted for publication 29.10.2015

e-mail: pnl@bionet.nsc.ru

Ontologies in bioinformatics and systems biology

N.L. Podkolodnyy[1, 2, 3], O.A. Podkolodnaya[1]

1 Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia 2 Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia
3 Novosibirsk State University, Novosibirsk, Russia

c omputer simulation is now becoming a central scientific pa adigm of systems biology and the basic tool for the theoretical study and understanding of the complex mechanisms of living systems. The increase in the number and complexity of these models leads to the need for their collaborative development, reuse of models, and their verific tion, and the description of the computational experiment and its results. on tological modeling is used to develop formats for knowledge-oriented mathematical modeling of biological systems. in this sense, ontology associated with the entire set of formats, supporting research in systems biology, in particular, computer modeling of biological systems and processes can be regarded as a first app oximation to the ontology of systems biology. This review summarizes the features of the subject area (bioinformatics, systems biology, and biomedicine), the main motivation for the development of ontologies and the most important examples of ontological modeling and semantic analysis at different levels of the hierarchy of knowledge: the molecular genetic level, cellular level, tissue levels of organs and the body. Bioinformatics and systems biology is an excellent ground for testing technologies and effici t use of ontological modeling. Several dozens of verified basic eference ontologies now represent a source of knowledge for the integration and development of more complex domain models aimed at addressing specific issues in biomedicine and biotechnology. Further formalization and ontological accumulation of knowledge and the use of formal methods of analysis can take the entire cycle of research in systems biology to a new technological level.

Key words: ontological modeling; bioinformatics; systems biology.

Появление качественно новых возможностей для проведения исследований, основанных на использовании высокопроизводительных экспериментальных технологий массового параллельного секвенирования ДНК, многолокусного генотипирования, многопараметрического профилирования экспрессии генов с использованием ДНК-чипов, ChiP-on-chip технологий, протеомных и метаболомных технологий и др., привело к накоплению беспрецедентно больших массивов экспериментальных данных и знаний.

Огромный объем молекулярно-биологической информации, ее сложность и наличие большого числа барьеров: технологических, информационных, ресурсных и т. д. затрудняют ее анализ, систематизацию и применение для решения конкретных задач биоинформатики, биотехнологии, фармакологии, персонализированной медицины и др. Чтобы освоить, систематизировать и эффективно использовать такого рода информацию, необходимы новые подходы к обработке больших данных (BIG DATA), в частности автоматические методы семантической интеграции гетерогенных данных, одним из основных этапов которой является согласование понятий предметной области, способов их описания и использования (сопоставление, обработка данных и т. д.). Такое согласованное описание конкретной предметной области называют онтологией.

Разработка онтологии является сложным и затратным процессом. Первый этап этого процесса – онтологический анализ и моделирование предметной области, включая создание словаря терминов, точных их определений и взаимосвязей между ними, правил и ограничений, согласно которым на базе введенной терминологии формируются достоверные утверждения, описывающие состояние изучаемого объекта.

Для чего же нужны онтологии? Онтологии позволяют представить понятия в таком виде, что они становятся пригодными для машинной обработки и вследствие этого используются в качестве посредника между пользователем и информационной системой или между членами научного сообщества при обмене данными. Молекулярному биологу важно иметь возможность описывать молекулярные события, взаимодействующие компоненты, роли, которые играют эти компоненты в молекулярных событиях и процессах, оценивать гипотезы. Биоинформатик заинтересован в интеграции данных, компьютерной аннотации, моделировании процессов и систем. Общезначимыми потребностями является использование онтологий в образовании.

Системная биология как научная дисциплина как раз и возникла с появлением возможностей построения портретных моделей биологических систем и процессов на основе интеграции и совместного компьютерного анализа большого объема такого рода принципиально новых экспериментальных данных, описывающих поведение молекулярно-генетических систем в целом. Предметом исследования в системной биологии являются биологические объекты и формируемые ими сложные, иерархически организованные сети взаимодействий, контролируемые информацией, закодированной в геномах (Kitano, 2002).

В связи с этим в настоящее время онтологический анализ становится одним из основных инструментов биоинформатики и системной биологии, используемым для семантической интеграции экспериментальных данных и знаний с целью построения «единой картины мира» (Подколодный, 2011).

Формальное представление онтологий

В информатике термин «онтология» означает концептуальную модель представления объектов, свойств объектов и отношений между ними (Chandrasekaran et al., 1999). Онтология включает набор понятий (терминов) предметной области, их определений и атрибутов, а также связанное с ними множество аксиом и правил вывода (Gruber, 1995).

Таким образом, формальная модель онтологии – это упорядоченная тройка конечных множеств $O = <T, R, F>$, где T – конечное и непустое множество классов и концептов (понятий, терминов) предметной области как части реального мира, рассматриваемой в пределах заданного контекста (в нашем случае – биоинформатика и системная биология), которую описывает онтология O; R – конечное множество отношений между концептами заданной предметной области; F – конечное множество функций интерпретации, заданных на понятиях и/или отношениях онтологии O или аксиом, используемых для моделирования утверждений, которые всегда являются истинными, что ограничивает интерпретацию и обеспечивает корректное использование понятий.

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

Базы знаний предметной области с позиции дескриптивной логики подразделяются на общие знания о множестве классов понятий предметной области, свойствах и отношениях между ними (terminological knowledge, или T-Box) и знания об индивидуальных объектах (экземплярах класса), их свойствах и связях с другими объектами (assertional knowledge, или A-Box), т. е. описывают предметную область на уровне конкретных данных (база данных). В базе знаний обе компоненты взаимосвязаны.

В общем случае создание прикладных онтологий, ориентированных на конкретную предметную область, может быть существенно ускорено, если использовать ранее разработанные канонические (ссылочные) онтологии для построения онтологических классов и отношений между ними.

В частности, в качестве такого рода ссылочной онтологии может быть онтология верхнего уровня или онтология базовых знаний, в которой описываются наиболее общие концепты (пространство, время, материя, объект, система, состояние, поведение, событие, процесс, действие, структура, функция и т. п.) и отношения («часть – целое», «общее – частное», «является подклассом», «оказывает воздействие», «является причиной», «приводит к», «регулирует», «связан с», «похоже на», а также простран-

Types of *part_of* relations proposed by (winston et al, 1987) and their specifying properties

Types of *part_of* relations	Properties		
	«Functional»	«Homeomerous»	«Separable»
component/integral-object	+	–	+
Member/collection	–	–	+
Portion/Mass	–	+	+
Stuff object	–	–	–
Feature/Activity	+	–	–
Place/Area	–	+	–

ственные и темпоральные отношения и т. п.). Эти концепты не зависят от конкретной проблемы или области, поэтому представляется разумным унифицировать их для больших сообществ пользователей.

Проект open Biological ontologies

Целью проекта Open Biological Ontologies (OBO) является разработка унифицированных подходов для создания онтологий, методов их интеграции, а также инструментальных средств работы с ними (Bada, Hunter, 2007; Smith et al., 2007). В OBO содержится информация об онтологиях и проектах, которые выполняются в области биологии (http://obofoundry.github.io/).

В настоящее время в OBO описано более 70 онтологий по различным направлениям, включая анатомию, биохимию, биологические процессы, функции и последовательности, заболевания, окружающую среду, экспериментальные доказательства, фенотип, белки, таксономии и др. (Schober et al., 2009).

Для обеспечения совместимости разрабатываемых в биомедицине онтологий в рамках проекта OBO предложены рекомендации по стандартизации, используемым онтологическим отношениям. Задаются формальные свойства отношений, которые можно использовать при логическом выводе новых утверждений. В частности, принимается, что отношения *part_of* и *is_a* транзитивны, рефлексивны, антисимметричны.

Однако на самом деле в зависимости от дальнейшего уточнения семантики отношений и специфики их применения, свойства этих отношений могут не выполняться. Даже такие распространенные отношения, как *part_of* и *is_a*, имеют на практике различные интерпретации. В этом случае свойства транзитивности могут нарушаться. С этим столкнулись разработчики проекта Gene Ontology (GO), когда приступили к формализованной проверке ее онтологии (Smith et al., 2003).

Ниже представлены некоторые проблемы, которые возникли в GO при интерпретации отношения *part_of*:

P1. *A part_of B* означает: *A* иногда является частью *B*, т. е. для каждого *A* в некоторое время t *A* является частью *B*.

Пример: «replication fork» *part_of* «nucleoplasm» («вилка репликации» наблюдается в определенной фазе клеточного цикла).

P2. *A part_of B* означает: *A* может быть частью *B*. Класс *A* является частью класса *B* тогда и только тогда, если существует подкласс $C \subset B$, в котором все экземпляры *A* включены как часть в экземпляры *C* и все экземпляры класса *C* имеют как часть экземпляры класса *A*.

Пример: «flagellum» *part_of* «cell» (некоторые типы клеток включают как часть жгутики).

P3. *A part_of B* означает: *A* всегда является частью *B*.

Пример: «membrane» *part_of* «cell» (мембрана является частью любой клетки).

Для разрешения этих проблем в настоящее время в GO вводятся более специфичные отношения типа: *is_localized_in* или *is_involved_in*.

Аналогично в GO возникли проблемы интерпретации отношения *is_a*. Хорошо структурированная классификация может быть получена путем замены отношения *is_a* на специальные типы, например: *has_role*, *is_dependent_on*, *is_involved_in*, *contributes_to*, *is_located_in*, а также добавления различных категории сущностей: *sites*, *constituents*, *roles*, *functions*, *qualities*.

В общем случае для разрешения этих проблем требуется уточнение семантики этих отношений, используемой при разработке конкретной онтологии.

Существуют различные попытки уточнения семантики отношений *part_of* и *is_a* и их классификации для разрешения типовых конфликтов и нарушений свойств этих отношений. В статье (Winston, et al., 1987) введено 6 различных типов отношения *part_of* (таблица) на основе следующих свойств или критериев:

- «Functional» выполняется, когда части в специфической пространственной или временной локализации выполняют такую же функциональную роль, как и целое.
- «Homeomerous» выполняется, когда части подобны каждая друг другу и целому, которому принадлежит.
- «Separable» выполняется, когда части могут быть физически не связаны и хотя бы в принципе отделены от целого, которое они составляют.

Проблемы с транзитивностью отношения *part_of* возникают, когда комбинируются различные типы отношений *part_of*. В общем случае транзитивность должна предполагаться, по крайней мере, тогда, когда используются отношения *part_of* того же типа, т. е. имеют те же свойства.

В зависимости от особенностей предметной области можно использовать другие наборы свойств, определяющих классы отношения *part_of*, например: configurational, encapsulated, exchangeable, functional, homeomerous, homogeneous, mandatory, canonically necessary, removable, segmental, separable, shareable.

В биоинформатике и системной биологии традиционно широко используются представления онтологий на языке OBO. В последнее время многие онтологии транслируются в представление на языке OWL (Ontology Web Language) (Stevens et al., 2007). Главной проблемой при таких преобразованиях является наличие ошибок, противоречий и нарушений интерпретаций отношений. Использование формальных методов поиска противоречий, неполноты позволяет существенно улучшить качество описаний.

развитие онтологий в биоинформатике

В настоящее время в области биологии разработано несколько сотен онтологий, которые можно использовать для описания и интеграции знаний, а также вывода новых знаний.

В частности, разработаны и активно используются биоинформационные ресурсы и онтологии, позволяющие описывать молекулярные структуры, функции, процессы и генные сети (GO).

Онтологии MIAPE (Minimum Information About a Proteomics Experiment) (Taylor et al., 2007) и MIMIx (Minimum Information required for reporting a Molecular Interaction eXperiment) (Orchard et al., 2007) предложены рабочей группой Human Proteome Organization (HUPO) для описания протеомных исследований и экспериментов по молекулярным взаимодействиям соответственно.

Небольшая онтология верхнего уровня BFO (Basic Formal Ontology) предназначена для разработки онтологий, ориентированных на поиск и интеграцию научных данных. BFO уже использовалась для разработки более 130 онтологий в разных предметных областях (http://ifomis.uni-saarland.de/bfo/).

База знаний ChEBI (Chemical Entities of Biological Interest) включает онтологию молекулярных объектов – природных соединений или синтетических продуктов, воздействующих на процессы в живых организмах, включая любые уникальные по структуре или изотопному составу атомы, молекулы, ионы, ионные пары, радикалы, ион-радикалы, комплексы, конформеры и т. п. (https://www.ebi.ac.uk/chebi/). ChEBI в настоящее время (release 131) включает 46477 полностью аннотированных молекулярных объектов.

Онтология клеточных типов CL (Cell Type Ontology), по сути, является структурированным контролируемым словарем, включающим описание клеточных типов различных видов организмов – от прокариот до млекопитающих (http://www.obofoundry.org/cgi-bin/detail.cgi?id=cell).

Онтологии, разработанные в рамках KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa et al., 2004), ориентированы на широкую область молекулярной биологии – от генов и белков до метаболических и генных сетей. База знаний TAMBIS (Transparent Access to Multiple Bioinformatics Information Source, http://www.cs.man.ac.uk/~stevensr/tambis/) предоставляет пользователям-биологам единую точку доступа к мировым источникам биологической информации, которая интегрируется в рамках этой системы с помощью онтологического описания (Stevens et al., 2000). EcoCyc – база научных данных, в которой накапливается информация, полученная в результате аннотирования научных публикаций по геному *E. coli*, регуляции транскрипции ее генов, транспортным и метаболическим путям (Karp et al., 2014).

Sequence Ontology (SO) включает множество понятий и контролируемых словарей, используемых для описания свойств и первичной аннотации нуклеотидных или белковых последовательностей, структурного представления этих аннотаций в геномных базах данных, мутаций в обоих типах последовательностей и на более высоком уровне (Cunningham et al, 2015).

В рамках Mouse Atlas Project развиваются цифровой атлас и база данных об экспрессии генов мыши и клеточных линиях с описанием анатомической локализации клеток (Graham et al., 2015).

Онтология MGED (Microarray and Gene Expression Data) используется для описания экспериментов и данных по экспрессии генов (Whetzel et al., 2006).

MIAME (Minimum Information About a Micro-array Experiment) используется для описания экспрессионных данных (Brazma et al., 2001).

В базе знаний PharmGKB (https://www.pharmgkb.org/) представлена информация по фармакогенетике (Klein et al., 2001).

Целью проекта Cell Cycle Ontology (Antezana et al., 2009) является расширение существующих онтологий, связанных с клеточным циклом, для интеграции и управления знаниями о его компонентах и регуляторных аспектах. В качестве источников этих знаний используются уже существующие ресурсы (GO, UniProt, IntAct, BIND, NCBI taxonomy и др.). Интеграция и комбинация этих знаний позволяют представлять наиболее полную картину процессов деления клетки.

Анатомические и морфологические онтологии для модельных организмов являются важнейшими примерами онтологического моделирования, в котором используется большое разнообразие пространственно-временных отношений.

В частности, основная модель анатомии FMA (Foundational Model of Anatomy ontology) является ссылочной онтологией, включающей понятия и отношения, которые используются для описания структурной организации тела человека на различных уровнях – от макромолекул, клеток, ткани, органа до организма с учетом онтогенеза, и предназначенной для символического компьютерного моделирования анатомических структур (Rosse et al., 2003; Rosse et al., 2007).

Онтологические знания в FMA представлены в виде фреймов и хранятся в реляционной базе данных. FMA (http://sig.biostr.washington.edu/projects/fm/) включает около 75 тыс. анатомических классов, более 130 тыс. уникальных терминов, более 205 тыс. фреймов и 174 уникальных слота, которые используются для представления различных типов отношений, атрибутов и атрибутных отношений. Сеть отношений FMA содержит более 2,5 млн экземпляров отношений, свыше 1 млн экземпляров классов, около 450 тыс. связей между классами.

Gene ontology

В качестве примера одного из самых успешных проектов создания онтологии можно привести GO (http://www.geneontology.org/).

В состав GO входят 3 раздела:

- Молекулярная функция (Molecular function) – элементарная активность/задача или роль, которую может выполнять ген, продукт гена в каких-либо биологических процессах, например, «catalytic activity» или «Toll receptor binding».
- Биологические процессы (Biological process) описывают серию событий, реализующих одну или более организованных ансамблей молекулярных функций. В отличие от функции процесс должен иметь несколько различающихся этапов. Например: «pyrimidine metabolic process».
- Клеточные компоненты (Cellular component) как часть анатомической структуры, в которой описывается локализация гена или его продукта в организме на уровнях клеточных структур и макромолекулярных комплексов (например, «nucleus», «membrane») или групп продуктов генов (например, «ribosome», «proteasome» или «protein dimer»).

По сути, GO позволяет описывать знания о том, какую функцию выполняет ген или его продукт (РНК, белок) в том или ином биологическом процессе и той или иной клеточной структуре.

GO содержит более 40 тыс. понятий (The Gene Ontology Consortium, 2015), включая:

- Biological process – около 30 тыс.;
- Molecular function – более 10 тыс.;
- Cellular component – 3 758.

На основе GO разработан ресурс GOA (Gene Ontology Annotation, http://www.ebi.ac.uk/GOA), который используется для аннотации белков из UniProtKB (UniProt Knowledgebase). В настоящее время GOA включает 368 млн аннотаций GO для почти 54 млн белков из 480 тыс. таксономических групп (Huntley et al., 2015).

Основные отношения между понятиями, которые используются в GO, это *is_a, part_of* и *regulates*.

- *is_a* – простое отношение «класс – подкласс», где A *is_a* B означает, что A является подклассом B.
- *part_of* – отношение «часть – целое». Выражение A *part_of* B означает, что если A существует, то A всегда является частью B.
- *Regulates*, *positively_regulates* и *negatively_regulates* описывают отношения между биологическими процессами, молекулярными функциями или биологическими свойствами.

В GO описываются иерархические отношения, однако граф отношений не является деревом. Одно понятие может иметь несколько предков. Свойство транзитивности отношений, используемых в GO, позволяет строить решетку отношений между понятиями и выполнять логический вывод о свойствах понятий и их отношениях (Srinivas, 2009).

Онтология системной биологии

Компьютерные модели становятся центральной научной парадигмой системной биологии и основным инструментом для теоретического исследования и понимания механизмов функционирования сложных живых систем. С увеличением числа и размеров этих моделей возрастает необходимость коллективной разработки, повторного использования моделей, их верификации, описания вычислительного эксперимента и его результатов.

При разработке форматов представления знаний для математического моделирования биологических систем активно применяют онтологическое моделирование предметной области. В этом смысле онтологию, связанную со всей совокупностью форматов, обеспечивающих поддержку исследований в системной биологии, в частности компьютерное моделирование биологических систем и процессов, можно считать первым приближением к онтологии системной биологии.

В настоящем разделе мы рассмотрим онтологии системной биологии, которые ориентированы на описание генетических систем и их моделей.

Многие онтологии не разрабатываются с нуля, а интегрируют (или компилируют) уже разработанные ранее частные онтологии, которые используются в качестве базовых источников знаний. К ним относятся формализация структуры математических моделей (SBML), стандартизация описания компонент модели с кинетической и биологической точки зрения (SBO, GO, UniProt). Однако формализация только структуры компьютерной модели недостаточна для поддержки всей технологической цепочки вычислительного эксперимента и компьютерного моделирования в системной биологии. Необходимо формализовать прагматические и динамические аспекты процесса моделирования.

Примером специализированного проекта по созданию онтологии для системной биологии является Systems Biology Ontology (SBO) (http://www.ebi.ac.uk/sbo/). Цель проекта – разработка контролируемых словарей и онтологий, ориентированных на решение задач системной биологии, особенно в контексте компьютерного моделирования.

SBO состоит из 6 ортогональных контролируемых словарей, включающих описание ролей участников реакций (например, «reactant», «product», «modifier»), значения количественных параметров моделей реакций (например, «kinetic constant»), точную классификацию математических выражений, которые описывают систему (например, «mass action rate law»), тип используемой среды моделирования (например, «logical framework», «discrete framework»), типы сущностей системы (например, «macromolecule», «enzyme», «ligand») и взаимодействий в ней (например, «process», «biochemical reaction», «genetic interaction», «relationship»). Количественные параметры и математические выражения описываются на языке MathML 3.0 (http://www.w3.org/TR/MathML3/).

Математическая модель может быть проаннотирована в SBO на любой стадии жизненного цикла (от времени создания до расширения и модификации модели) путем последовательного расширения ее семантики.

Можно привести примеры других форматов представления знаний и связанных с ними онтологий системной биологии:

- Systems Biology Markup language (SBML, http://sbml.org) – формат представления структуры биологических моделей (Hucka et al., 2003).
- Biological Pathway Exchange Language (BioPAX, http://www.biopax.org) – формат для описания и интеграции

информации о молекулярных взаимодействиях и биологических процессах (Demir et al., 2010). Этот подход используется для представления знаний в существующих базах данных (BioCyc, BIND, WIT, aMAZE, KEGG, Reactome и т. д.), важных для описания механизмов экспрессии генов.

- Minimal Information Requested In the Annotation of biochemical Models (MIRIAM) (Le Novère et al., 2005) – формат для стандартизации минимального множества информации, необходимой для аннотации модели и обеспечения возможности коллективной аннотации, курирования и развития, а также повторного использования моделей.
- Simulation Experiment Description Markup Language (SED-ML, http://sed-ml.org/) – формат для описания экспериментов по моделированию и обмену результатами моделирования независимо от использованного языка спецификации модели и среды моделирования (Waltemath et al., 2011).
- TErminology for the Description of DYnamics (TEDDY, http://www.ebi.ac.uk/compneursrv/teddy) (Chelliah, Endler, 2009) – онтология для описания динамического поведения биологической системы или динамического явления, управления элементами биологической модели и системы в системной и синтетической биологии. В частности, используя TEDDY, можно количественно описывать характеристики биологической системы или модели: тип осцилляции (хаотическая, периодическая, квазипериодическая и т. д.), области осцилляции, период, точки покоя или нестабильности, бифуркации, зависящие от параметров, функциональные мотивы (например, отрицательные обратные связи) и т. д.
- Kinetic Simulation Algorithm Ontology (KiSAO, http://biomodels.net/kisao/) – онтология для описания алгоритмов моделирования кинетических процессов (Chelliah, Endler, 2009). Классификация алгоритмов моделирования биологических моделей в KiSAO построена с использованием различных категорий и с учетом версий алгоритмов. Например, детерминированные или стохастические правила, пространственные или непространственные подходы, дискретные или непрерывные переменные, фиксированные или адаптивные временные шаги и т. д.
- Systems Biology Graphical Notation (SBGN, http://www.sbgn.org) – формат для графического представления биологических систем и процессов (Dada et al., 2010).
- SBRML (www.comp-sys-bio.org/tiki-index.php?page=SBRML) – формат для сохранения результатов моделирования (Orchard et al., 2007).
- CellML (www.cellml.org) – формат для описания математических моделей биологических систем и процессов (Lloyd et al., 2004). Для представления математических выражений в CellML используется язык MathML.

Ссылочные онтологии, GO, SO, Chemical Entities of Biological Interest (ChEBI), FMA, FMP, CPRO, PaTO, Pro, RnaO, CARO, описывающие биологические системы на различных уровнях, могут быть интегрированы в единую онтологию для описания объекта исследования системной биологии (организм, орган, ткани, клетки).

Применение онтологий в системной биологии и биоинформатике

В целом в биоинформатике и системной биологии можно выделить следующие задачи, в решении которых применение онтологий дает ощутимый эффект (Bodenreider, Stevens, 2006; Bodenreider, 2008; Beck et al., 2009; Noy et al., 2009; Подколодный, 2011).

1. Интерпретация молекулярно-генетических знаний, семантическая интерпретация методов анализа данных и моделей в системной биологии. В частности, анализ обогащения генов терминами из GO (GO Enrichment Analysis) используется для интерпретации данных (например, функциональное описание множества генов), контроля качества, систематизации и отбора данных.
2. Приоритизация генов, белков, биомаркеров и т. д.
3. Анализ сходства и кластеризация объектов. В качестве примера можно привести анализ уровня экспрессии десятков тысяч генов в различных клеточных ситуациях, при разных состояниях и на различных этапах развития клетки, ткани, органа или организма. После выделения группы генов со схожими паттернами экспрессии (коэкспрессирующиеся гены) возникает задача описания этих групп. Использование GO позволяет описать, в реализации каких функций участвуют гены, входящие в кластер (Khatri, Draghici, 2005). По сути, используя онтологии, можно количественно оценивать семантическое сходство объектов предметной области.
4. Поддержка интероперабельности и обмена знаниями:

унифицированный доступ к множествам гетерогенных источников данных;

поиск релевантной информации в документах. Онтология в этом случае задает структуру для аннотации содержания документа с семантической информацией, а также обеспечивает индексирование и связывание фактов, описанных в базах данных (Shah et al., 2009);

интеграция информации из различных источников и создание больших баз знаний;

комбинирование экспериментальных данных и знаний из онтологии для формирования баз знаний;

интероперабельность, поддержка коммуникации (между людьми и организациями) и обмена знаниями (между людьми и/или системами);

анализ текстов и семантический анализ (Chapman, Cohen, 2009);

приобретение знаний, извлечение знаний, неявных и явных отношений между сущностями в аннотированных источниках, аналитика.

5. Создание новых онтологий на основе повторного использования базовых канонических онтологий и различного типа операции с ними, включая сопоставление (ontology matching), слияние (ontology merging), отображение (ontology mapping), выравнивание (ontology alignment) и т. д.
6. Обеспечение непротиворечивости и корректности представления знаний. Поддержка процесса построения онтологий, включая любые типы автоматического вывода для поиска ошибок и выявления новых отношений. Количество понятий и отношений в современных онтологиях исчисляется сотнями тысяч, поэтому ручная проверка невозможна. Эксперт в этом случае

проверяет противоречия и результаты, полученные путем формального вывода на онтологиях (Livingston et al., 2015).

7. Поддержка индуктивного вывода для извлечения дополнительных знаний из множества фактов и тестирование гипотез. Например, в работе Подколодного с коллегами (2012) представлены подходы к онтологическому моделированию механизмов регуляции транскрипции генов и показаны примеры реконструкции гипотетических механизмов регуляции транскрипции с учетом информации о строении регуляторных районов генов и функциях регуляторных белков, присутствующих в заданных клетках или тканях на определенной стадии развития.
8. Повышение аргументации методов биоинформатики, включая точное описание биомедицинских экспериментальных протоколов, методов анализа данных и моделирования биологических процессов и систем и т. д. (Chen et al., 2007).

Можно привести примеры успешного применения дескриптивной логики и логического вывода для формальных онтологий в области биомедицины. В обзоре Кит с коллегами (Keet et al., 2007) приводится список типовых сценариев, полученных путем анализа и обобщения примеров использования средств логического вывода. При этом отмечено, что стандартные средства Racer, Pellet, FaCT++ не позволяют выполнять многие сценарии на реальных биомедицинских онтологиях из-за их большого объема и сложности. Поэтому весьма актуальным остается развитие новых эффективных программных средств логического вывода и подходов к формализации биологических знаний.

Программные средства для Go

В настоящее время среди всех онтологий в области биоинформатики и системной биологии наиболее широко используется GO. Для этого разработано большое число программных средств:

AMIGO (http://amigo.geneontology.org/amigo), QuickGO (http://www.ebi.ac.uk/QuickGO/), Protein2GO и Ontology Lookup Service (http://www.ebi.ac.uk/ontology-lookup/) – поиск и просмотр GO и аннотаций данных (Carbon et al., 2009).

OBO-Edit – просмотр и редактирование онтологических описаний (Day-Richter et al., 2007).

Blast2GO (https://www.blast2go.com/) – функциональная аннотация неизвестных последовательностей путем поиска гомологий с помощью BLAST и анализа GO аннотаций полученных результатов (Conesa, Gotz, 2008).

GoPubMed – поиск биологических текстов на основе GO и Medical Subject Headings (MeSH, http://www.nlm.nih.gov/mesh/introduction.html). GoPubMed связывает GO с базой рефератов PubMed и отвечает на вопросы: Что? Кто? Где? Когда? (http://www.gopubmed.org/web/gopubmed/).

Onto-Tools (http://vortex.cs.wayne.edu/) – набор сервисов, включающий: Onto-Express, Onto-Compare, Onto-Design, Onto-Translate, Onto-Miner, Pathway-Express, Promoter-Express, nsSNPCounter, TAQ и OE2GO, – профилирование множества генов, сравнение экспрессионных данных, компьютерное сопровождение процесса конструирования ДНК-чипов, включая выбор множества генов на основе их функции, процессов, в которых эти гены участвуют, или клеточных компонент, где эти гены экспрессируются, и т. д.

GOToolBox – анализ результатов ДНК-чиповых экспериментов (http://genome.crg.es/GOToolBox/) (Martin et al., 2004).

Наиболее часто GO используется для анализа обогащения описания генов терминами генной онтологии. Для решения этой задачи разработаны несколько десятков программных систем, среди которых можно выделить наиболее популярные (ftp://ftp.geneontology.org/pub/go/www/GO.tools.microarray.shtml):

- Gorilla (Gene Ontology enRIchment anaLysis and visualizAtion tool) (http://cbl-gorilla.cs.technion.ac.il/) – Web-приложение для идентификации обогащенных терминов GO в ранжированных списках генов.
- PANTHER (http://pantherdb.org/) – классификации белков (и их генов) в соответствии с информацией из GOA, эволюционными взаимосвязями, взаимодействиями, участием в метаболических и генных сетях и т. д. Как часть проекта GO эта система регулярно обновляется и использует наиболее актуальные данные (Thomas et al., 2003; Huaiyu et al., 2003; Mi et al., 2005).
- DAVID (http://david.abcc.ncifcrf.gov/) – Web-сервис для аннотации и анализа экспрессионных и протеомных данных, полученных с помощью высокопроизводительных экспериментальных технологий, а также обогащения описания генов. Основной недостаток DAVID – это использование устаревших данных GO (отставание – 3–4 года).
- Веб-сервис WebGestalt (http://bioinfo.vanderbilt.edu/webgestalt/) используется в функциональной геномике, протеомике и широкомасштабных генетических исследованиях для функциональной аннотации больших групп генов, например, групп дифференциально экспрессирующихся и коэкспрессирующихся генов и т. д.

Бурное развитие экспериментальных технологий в области молекулярной биологии привело к тому, что онтологическое моделирование становится базовым методом в биоинформатике и системной биологии для интеграции и анализа гетерогенных экспериментальных данных и использования их для построения математических моделей молекулярно-генетических систем и процессов.

Биоинформатика и системная биология являются прекрасным полигоном для отработки технологий и эффективного использования онтологического моделирования. Создание нескольких десятков базовых ссылочных онтологий и их верификация позволяют использовать эти онтологии в качестве источников знаний для интеграции и построения более сложных моделей предметной области, ориентированных на решение конкретных задач биомедицины. В качестве примера можно привести базу знаний KaBOB (Livingston et al., 2015), в которой интегрированы знания из 14 гетерогенных источников – баз данных по генам и их гомологам, белкам, лекарствам, генетическим ассоциациям, регуляторным последовательностям, белок-белковым взаимодействиям и др. Интеграция данных и поиск информации в этой базе знаний основаны на онтологии, объединяющей:

1. Basic Formal Ontology (BFO) (http://purl.obolibrary.org/obo/bfo.owl);
2. BRENDA Tissue / Enzyme Source (BTO) (http://purl.obolibrary.org/obo/bto.owl);
3. ChEBI (http://purl.obolibrary.org/obo/chebi.owl);
4. Cell Type Ontology (CL) (http://purl.obolibrary.org/obo/cl.owl);
5. Gene Ontology (GO) (http://purl.obolibrary.org/obo/go.owl);
6. Information Artifact Ontology (IAO) (http://purl.obolibrary.org/obo/iao.owl);
7. Protein-Protein Interaction Ontology (MI) (http://purl.obolibrary.org/obo/mi.owl) ;
8. Mammalian Phenotype Ontology (MP) (http://purl.obolibrary.org/obo/mp.owl);
9. NCBI Taxonomy (http://purl.obolibrary.org/obo/ncbitaxon.owl);
10. Ontology for Biomedical Investigation (OBI) (http://purl.obolibrary.org/obo/ obi.owl);
11. Protein Modification (MOD) (http://purl.obolibrary.org/obo/mod.owl);
12. Protein Ontology (PR) (http://purl.obolibrary.org/obo/pr.owl);
13. Relation Ontology (RO) (http://purl.obolibrary.org/obo/ro.owl);
14. Sequence Ontology (SO) (http://purl.obolibrary.org/obo/so.owl).

Для представления знаний использовался язык OWL. Описание онтологии KaBOB составляет более 13 млн RDF-триплетов. Для поиска используется язык SPARQL 1.1. Реализованы две версии базы знаний KaBOB только для человека и для человека и ряда модельных организмов. В итоге система позволяет выполнять запросы на множестве гетерогенных данных, сопоставляя знания на различных уровнях описания биологических систем. В качестве примера можно привести следующий запрос: «Какие гены или продукты генов у человека, локализованные в митохондриях, вовлечены в процесс окислительного фосфорилирования и являются мишенями лекарств? Что это за лекарства?». Архитектура системы KaBOB и применяемые информационные технологии позволяют расширять используемые онтологии и базы данных и в перспективе решать сложные биоинформационные задачи.

Таким образом, дальнейшие формализация и накопление онтологических знаний, а также применение формальных методов их анализа могут поднять весь цикл научных исследований в области системной биологии на новый технологический уровень.

Acknowledgments

This work was supported by the Russian Science Foundation, project 14-24-00123.

Conflict of interest

The authors declare no conflict of interest.

References

Antezana E., Egaña M., Blondé W., Illarramendi A., Bilbao I., De Baets B., Stevens R., Mironov V., Kuiper M. The Cell Cycle Ontology: an application ontology for the representation and integrated analysis of the cell cycle process. Genome Biol. 2009;10(5):R58. DOI 10.1186/gb-2009-10-5-r58

Bada M., Hunter L. Enrichment of OBO ontologies. J. Biomed. Inform. 2007;40:300-315.

Beck T., Morgan H., Blake A., Wells S., Hancock J.M., Mallon A.-M. Practical application of ontologies to annotate and analyze large-scale raw mouse phenotype data. BMC Bioinformatics. 2009;10(Suppl5): S2. DOI 10.1186/1471-2105-10-S5-S2

Bodenreider O. Biomedical ontologies in action: Role in knowledge management, data integration and decision support. IMIA Yearbook Medical Informatics, 2008.

Bodenreider O., Stevens R. Bio-ontologies: Current trends and future directions. Brief Bioinform. 2006;7(3):256-274.

Brazma A., Hingamp P., Quackenbush J., Sherlock G., Spellman P., Stoeckert C., Aach J., Ansorge W., Ball C.A., Causton H.C., Gaasterland T., Glenisson P., Holstege F.C., Kim I.F., Markowitz V., Matese J.C., Parkinson H., Robinson A., Sarkans U., Schulze-Kremer S., Stewart J., Taylor R., Vilo J., Vingron M. Minimum information about a microarray experiment (MIAME)–toward standards for microarray data. Nat. Genet. 2001;29:365-371.

Carbon S., Ireland A., Mungall C.J., Shu S., Marshall B., Lewis S. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009;25(2):288-289. DOI 10.1093/bioinformatics/btn615

Chandrasekaran B., Josephson J.R., Benjamins V.R. What are ontologies and why do we need them? IEEE Intelligent Systems. 1999;14(1): 20-26. DOI 10.1109/5254.747902

Chapman W.W., Cohen K.B. Current issues in biomedical text mining and natural language processing. J. Biomed. Inform. 2009;42:757-759. DOI 10.1016/j.jbi.2009.09.001

Chelliah V., Endler L., Juty N., Laibe C., Li C., Rodriguez N., Le Novère N. Data integration and semantic enrichment of systems biology models and simulations. Data integration in the life sciences. Lecture Notes in Computer Science. 2009;5647:5-15.

Chen Q., Chen Y.-P.P., Zhang C. Detecting inconsistency in biological molecular databases using ontologies. Data Min. Knowl. Disc. 2007;15:275-296. DOI 10.1007/s10618-007-0071-0

Conesa A., Gotz S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. Int. J. Plant Genomics. 2008;2008:619832. DOI 10.1155/2008/619832

Cunningham F., Moore B., Ruiz-Schultz N., Ritchie G.R., Eilbeck K. Improving the Sequence Ontology terminology for genomic variant annotation. J. Biomed Semantics. 2015; 6:32. DOI 10.1186/s13326-015-0030-4

Dada J.O., Spasić I., Paton N.W., Mendes P. SBRML: a markup language for associating systems biology data with models. Bioinformatics. 2010;26(7):932-938. DOI 10.1093/bioinformatics/btq069

Day-Richter J., Harris M.A., Haendel M. Gene ontology OBO-edit working group, Lewis S. OBO-Edit – an ontology editor for biologists. Bioinformatics. 2007;23(16):2198-2200.

Demir E., Cary M.P., Paley S. et al. The BioPAX community standard for pathway data sharing. Nat. Biotechnol. 2010;28:935-942. DOI 10.1038/nbt.1666

Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucl. Acids Res. 2015;43(Database issue):D1049-D1056. DOI 10.1093/nar/gku1179

Graham E., Moss J., Burton N., Armit C., Richardson L., Baldock R.D. The atlas of mouse development eHistology resource. Development. 2015;142:1909-1911. DOI 10.1242/dev.124917

Gruber T.R. Toward principles for the design of ontologies used for knowledge sharing. Int. J. Human-Computer Studies.1995;43(5/6): 907-928.

Huaiyu M.I., Lazareva-Ulitsky B., Loo R., Kejariwal A., Vandergriff J., Rabkin S., Guo N., Muruganujan A., Doremieux O., Campbell M.J., Kitano H., Thomas P.D. The PANTHER database of protein families, subfamilies, functions and pathways. Nucl. Acids Res. 2005; 33(suppl.1):D284-D288. DOI 10.1093/nar/gki078

Hucka M., Finney A., Sauro H.M., Bolouri H., Doyle J.C., Kitano H., Arkin A.P., Bornstein B.J., Bray D., Cornish-Bowden A., Cuellar A.A.,

Dronov S., Gilles E.D., Ginkel M., Gor V., Goryanin I.I., Hedley W.J., Hodgman T.C., Hofmeyr J.H., Hunter P.J., Juty N.S., Kasberger J.L., Kremling A., Kummer U., Le Novère N., Loew L.M., Lucio D., Mendes P., Minch E., Mjolsness E.D., Nakayama Y., Nelson M.R., Nielsen P.F., Sakurada T., Schaff J.C., Shapiro B.E., Shimizu T.S., Spence H.D., Stelling J., Takahashi K., Tomita M., Wagner J., Wang J. SBML Forum. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics. 2003;19(4):524-31.

Huntley R.P., Sawford T., Mutowo-Meullenet P., Shypitsyna A., Bonilla C., Martin M.J., O'Donovan C. The GOA database: Gene Ontology annotation updates for 2015. Nucl. Acids Res. 2015;43(Database issue):D1057-D1063. DOI 10.1093/nar/gku1113

Kanehisa M., Goto S., Kawashima S., Okuno Y., Hattori M. The KEGG resource for deciphering the genome. Nucl. Acids Res. 2004;32 (Database issue):D277-80.

Karp P.D., Weaver D., Paley S., Fulcher C., Kubo A., Kothari A., Krummenacker M., Subhraveti P., Weerasinghe D., Gama-Castro S., Huerta A.M., Muñiz-Rascado L., Bonavides-Martinez C., Weiss V., Peralta-Gil M., Santos-Zavaleta A., Schröder I., Mackie A., Gunsalus R., Collado-Vides J., Keseler I.M., Paulsen I. The EcoCyc Database. Ecosal Plus. 2014;2014. DOI 10.1128/ecosalplus

Keet C.M., Roos M., Marshall M.S. A survey of requirements for automated reasoning services for bio-ontologies in owl. Workshop on OWL: Experiences and Directions. Innsbruck, Austria, 2007.

Khatri P., Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics. 2005; 21(18):3587-3595.

Kitano H. Systems biology: a brief overview. Science. 2002;295:1662-1664.

Klein T.E., Chang J.T., Cho M.K., Easton K.L., Fergerson R., Hewett M., Lin Z., Liu Y., Liu S., Oliver D.E., Rubin D.L., Shafa F., Stuart J.M., Altman R.B. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenomics J. 2001;1: 167-170.

Le Novère N., Finney A., Hucka M., Bhalla U.S., Campagne F., Collado-Vides J., Crampin E.J., Halstead M., Klipp E., Mendes P., Nielsen P., Sauro H., Shapiro B., Snoep J.L., Spence H.D., Wanner B.L. Minimum information requested in the annotation of biochemical models (MIRIAM). Nat. Biotechnol. 2005;23(12):1509-1515.

Livingston K.M., Bada M., Baumgartner W.A. Jr, Hunter L.E. KaBOB: ontology-based semantic integration of biomedical databases. BMC Bioinformatics. 2015;16:126. DOI 10.1186/s12859-015-0559-3

Lloyd C.M., Halstead M.D.B., Nielsen P.F. CellML: its future, present and past. Progr. Biophys. Mol. Biol. 2004;85(2/3):433-450.

Martin D., Brun C., Remy E., Mouren P., Thieffry D., Jacq B. GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome Biol. 2004;5(12):R101.

Mi H., Lazareva-Ulitsky B., Loo R., Kejariwal A., Vandergriff J., Rabkin S., Guo N., Muruganujan A., Doremieux O., Campbell M.J., Kitano H., Thomas P.D. The PANTHER database of protein families, subfamilies, functions and pathways. Nucl. Acids Res. 2005;1:33 (Database issue):D284-8.

Noy N.F., Shah N.H., Whetzel P.L., Dai B., Dorf M., Griffith N., Jonquet C., Rubin D.L., Storey M.A., Chute C.G., Musen M.A. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucl. Acids Res. 2009;37:W170-W173. DOI 10.1093/nar/gkp440

Orchard S., Salwinski L., Kerrien S., Montecchi-Palazzi L., Oesterheld M., Stümpflen V., Ceol A., Chatr-aryamontri A., Armstrong J., Woollard P., Salama J.J., Moore S., Wojcik J., Bader G.D., Vidal M., Cusick M.E., Gerstein M., Gavin A.C., Superti-Furga G., Greenblatt J., Bader J., Uetz P., Tyers M., Legrain P., Fields S., Mulder N., Gilson M., Niepmann M., Burgoon L., De Las Rivas J., Prieto C., Perreau V.M., Hogue C., Mewes H.W., Apweiler R., Xenarios I., Eisenberg D., Cesareni G., Hermjakob H. The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat. Biotechnol. 2007;25(8):894-898.

Podkolodnyy N.L. Ontological modeling in bioinformatics and systems biology. Trudy Vtorogo simpoziuma "Ontologicheskoe modelirovanie" [Proceedings of the 2nd Symposium "Ontological modeling"]. Moscow, IPI RAN, 2011:233-269.

Podkolodnyy N.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. Information suppotr of research on transcriptional regulatory mechanisms: an ontological approach. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2012; 16(4/1):742-755.

Rosse C., Mejino J.L.Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J. Biomed. Inform. 2003;36(6): 478-500.

Rosse C., Mejino J.L.V. The Foundational Model of Anatomy Ontology. Anatomy Ontologies for Bioinformatics: Principles and Practice. Eds A. Burger, D. Davidson, R. Baldock. N.Y.: Springer, 2007.

Schober D., Smith B., Lewis S.E., Kusnierczyk W., Lomax J., Mungall C., Taylor C.F., Rocca-Serra P., Sansone S.A. Survey-based naming conventions for use in OBO foundry ontology development. BMC Bioinformatics. 2009;27;10:125. DOI 10.1186/1471-2105-10-125

Shah N.H., Jonquet C., Chiang A.P., Butte A.J., Chen R., Musen M.A. Ontology-driven indexing of public datasets for translational bioinformatics. BMC Bioinformatics. 2009;10(Suppl. 2):S1. DOI 10.1186/1471-2105-10-S2-S1

Smith B., Williams J., Schulze-Kremer S. The Ontology of the Gene Ontology. AMIA Annual Symp. Proceedings, 2003.

Smith B., Ashburner M., Rosse C., Bard J., Bug W., Ceusters W., Goldberg L.J., Eilbeck K., Ireland A., Mungall C.J., OBI Consortium, Leontis N., Rocca-Serra P., Ruttenberg A., Sansone S.A., Scheuermann R.H., Shah N., Whetzel P.L., Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotech. 2007;25(11):1251-1255.

Srinivas K. OWL Reasoning in the Real World: Searching for Godot. Proc. of the 22nd Intern. Workshop on Description Logics (DL 2009), Oxford, UK, July 27–30, 2009.

Stevens R., Aranguren M.E., Wolstencroft K., Sattler U., Drummond N., Horridge M., Rector A. Using OWL to model biological knowledge. Int. J. Human-Computer Studies http://portal.acm.org/citation.cfm?id=1247774. 2007;65(7):583-594.

Stevens R., Baker P., Bechhofer S., Ng G., Jacoby A., Paton N.W., Goble C.A., Brass A. TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics. 2000;16(2):184-185.

Taylor C.F., Paton N.W., Lilley K.S., Binz P.A., Julian R.K.Jr., Jones A.R., Zhu W., Apweiler R., Aebersold R., Deutsch E.W., Dunn M.J., Heck A.J., Leitner A., Macht M., Mann M., Martens L., Neubert T.A., Patterson S.D., Ping P., Seymour S.L., Souda P., Tsugita A., Vandekerckhove J., Vondriska T.M., Whitelegge J.P., Wilkins M.R., Xenarios I., Yates J.R. 3rd, Hermjakob H. The minimum information about a proteomics experiment (MIAPE). Nat. Biotechnol. 2007;25(8):887-893.

Thomas P.D., Campbell M.J., Kejariwal A., Mi H., Karlak B., Daverman R., Diemer K., Muruganujan A., Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13(9):2129-2141. PubMed PMID:12952881; PubMed Central PMCID: PMC403709.

Waltemath D., Adams R., Bergmann F.T., Hucka M., Kolpakov F., Miller A.K., Moraru I.I., Nickerson D., Snoep J.L., Le Novère, N. Reproducible computational biology experiments with SED-ML – the simulation experiment description markup language. BMC Systems Biol. 2011;5:198.

Whetzel P.L., Parkinson H., Causton H.C., Fan L., Fostel J., Fragoso G., Game L., Heiskanen M., Morrison N., Rocca-Serra P., Sansone S.A., Taylor C., White J., Stoeckert C.J.Jr. The MGED Ontology: a resource for semantics-based description of microarray experiments. Bioinformatics. 2006;22(7):866-73.

Winston M.E., Chaffin R., Herrman D. A taxonomy of part-whole relations. Cognitive Sci. 1987;11:417-444.

Вавиловский журнал генетики и селекции. 2015;19(6):661-667
Doi 10.18699/VJ15.087

Использование графических ускорителей для выявления функциональных сигналов в регуляторных районах генов прокариот

О.В. Вишневский[1, 2], А.В. Бочарников[2], А.А. Романенко[2]

1 Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия
2 Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия

Различные методы выявления значимых контекстных сигналов широко используются для поиска сайтов связывания транскрипционных факторов и выявления структурно-функ-циональной организации регуляторных районов генов. Такие методы не требуют ни предварительного выравнивания выборки анализируемых последовательностей, ни экспери-ментальной информации о точном расположении сайтов связывания транскрипционных факторов. Широкое распро-странение получили методы поиска контекстных сигналов, основанные на выявлении вырожденных олигонуклеотидных мотивов, записанных в 15-буквенном коде номенклатуры IUPAC (International Union of Pure and Applied Chemistry). Существенной сложностью использования вырожденных мотивов является их огромное разнообразие, что заставляет исследователей применять различные эвристические подходы, не гарантирующие нахождение наиболее значимого сигнала. Появление высокопроизводительных вычислитель-ных систем, основанных на использовании графических ускорителей, сделало возможным применение точных полнопереборных методов для выявления значимых мотивов. Нами разработана новая система выявления значимых вырожденных олигонуклеотидных мотивов заданной длины в регуляторных районах генов, основанная на использовании широко распространенных графических ускорителей и обеспечивающая поиск сигнала с наибольшей значимостью. Показана высокая эффективность использования графических ускорителей (GPU) в сравнении с расчетами на центральном процессоре (CPU). С использованием предложенного подхода проанализированы регуляторные районы генов B. subtilis, E. coli, H. pylori, M. gallisepticum, M. genitalium и M. pneumoniae. Для каждого вида прокариот были выявлены наборы вырожденных мотивов и проведена их классификация на основе сходства с сайтами связывания транскрипционных факторов E. coli.

Ключевые слова: вырожденный олигонуклеотидный мотив; регуляция транскрипции; регуляция трансляции; CUDA; графические ускорители.

HOW TO CITE THIS ARTICLE?
Vishnevsky o.V., Bocharnikov A.V., r omanenko A.A. The use of graphics accelerators to detect functional signals in the regulatory regions of prokaryotic genes. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2015;19(6):661-667. Doi 10.18699/V J15.087

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?
Вишневский О.В., Бочарников А.В., Романенко А.А. Использование графических ускорителей для выявления функциональных сигналов в регуляторных районах генов прокариот. Вавиловский журнал генетики и селекции. 2015;19(6):661-667. Doi 10.18699/VJ15.087

The use of graphics accelerators to detect functional signals in the regulatory regions of prokaryotic genes

O.V. Vishnevsky[1, 2], A.V. Bocharnikov[2], A.A. Romanenko[2]

1 Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
2 Novosibirsk State University, Novosibirsk, Russia

Various methods for identification of significant contextual signals are widely used to search for transcription factor binding sites and to identify the structural and functional organization of regulatory regions. These methods do not require any pre-alignment of the sample sequences analyzed or experimental information about the exact location of transcription factor binding sites. Methods of searching for contextual signals, based on the identi-fication of degenerate oligonucleotide motives recorded in the 15-letter IUPAC code have become widespread. An essential problem with degenerate motifs is their great diversity, which makes the researchers apply heuristics which do not guarantee that the most significant signal will be found. The development of high-performance computing systems based on the use of graphics cards has made it possible to use the exact exhaustive methods to identify significant motifs. We have developed a new system for identifying significant degenerate oligonucleo- tide motifs of a given length in the regulatory regions based on the use of widespread graphics cards that provides a search for the signal with the greatest signifi-cance. High efficiency of the GPU compared with CPU was demonstrated. Using the proposed approach, we ana-lyzed the regulatory regions of B. subtilis, E. coli, H. pylori, M. gallisepticum, M. genitalium and M. pneumoniae genes. Sets of degenerate motifs have been identified for each species of prokaryotes. They were classified on the basis of similarity with the transcription factor binding sites of E. coli.

Key words: degenerated oligonucleotide motif; transcription regulation; translation regulation; CUDA; GPU.

Received 22.09.2015
Accepted for publication 16.10.2015

e-mail: oleg@bionet.nsc.ru

Выявление функциональных сигналов в регуляторных районах генов является важной задачей современной биоинформатики и необходимо как для понимания базовых механизмов регуляции транскрипции и трансляции, так и для определения специфических особенностей функционирования регуляторных районов.

Существующие методы, как правило, основываются на использовании ранее полученной экспериментальной информации о локализации сайтов связывания транскрипционных факторов (ССТФ), собранной в специализированных базах данных (Kolchanov et al., 2002; Matys et al., 2006; Portales-Casamar et al., 2010), или на сравнении анализируемых последовательностей и выявлении в них относительно схожих участков (Vishnevsky, Kolchanov, 2005).

Гигантский рост баз данных в связи с появлением методов высокопроизводительного секвенирования ДНК (Elnitski et al., 2003) и огромное разнообразие регуляторных сигналов требуют разработки новых высокопроизводительных компьютерных методов для их выявления и анализа.

Наибольшее распространение получили методы выявления значимых контекстных сигналов, основанные на анализе частот *l*-плетов (*l*-letter substrings) (Pesole et al., 2000); деревьев суффиксов (suffix trees) (Marsan et al., 2000); поиске максимальной клики в графе, построенном на основе дистанции редактирования (edit distance) *l*-плетов (Pevzner, Sze, 2000); локальном множественном выравнивании, основанном на «жадном» (greedy) алгоритме (Hertz, Stormo, 1999); методе EM (Expectation–Maximization) (Grundy et al., 1996) и стохастическом отборе (stochastic sampling) (Lawrence et al., 1993). Результатом работы таких методов, как правило, являются позиционные весовые матрицы или олигонуклеотидные мотивы, записанные в 4- (A, T, G, C) или 15-буквенном коде IUPAC (A, T, G, C, R = G/A, Y = T/C, M = A/C, K = G/T, W = A/T, S = G/C, B = T/G/C, V = A/G/C, H = A/T/C, D = A/T/G, N = A/T/G/C).

Применение олигонуклеотидных мотивов является одним из наиболее ранних и широко распространенных подходов, но их использование затруднено гигантским разнообразием возможных вариантов. Так, мотив длиной 8, записанный в 15-буквенном коде IUPAC, имеет $15^8 \sim 2{,}5 \times 10^9$ различных вариантов записи в 4-буквенном коде. Это приводит к необходимости использования различных эвристических подходов (Mrázek et al., 2002; Vishnevsky, Kolchanov, 2005). Однако такие подходы не гарантируют достижения глобального минимума и нахождения наиболее представленного и значимого мотива. Для применения точного полнопереборного метода необходимо использование высокопроизводительных вычислительных систем, основанных на массивно-параллельных алгоритмах.

На данный момент наиболее популярными системами для параллельных вычислений в биоинформатике являются:

- FPGA (Field-Programmable Gate Array) – программируемая логическая интегральная схема (ПЛИС) (Baker, Prasanna, 2006; Yooseph et al., 2007).
- Cell/BE (Cell Broadband Engine) – процессор, используемый в Sony PlayStation 3 (Fomin, Alemasov, 2009).
- GPU (Graphics Processing Unit) – графический ускоритель (Manavski, Valle, 2008; Sukhwani, Herbordt, 2009).
- CPU (Central Processing Unit) – универсальный центральный процессор и компьютерные кластеры на его основе (Vishnevsky, Kolchanov, 2005).

Хотя FPGA и кластеры CPU обладают очень высокими вычислительными мощностями, они до сих пор являются очень дорогим и не для всех доступным решением. Cell/BE более доступны, но не обладают необходимой вычислительной мощью. По соотношению цена/качество наиболее приемлемым и распространенным решением являются вычислительные платформы, основанные на графических ускорителях. Поэтому для разработки метода выявления вырожденных олигонуклеотидных мотивов мы использовали графические ускорители и технологию CUDA.

Графические ускорители стали одной из наиболее распространенных компонент современных компьютеров. Изначально разработанные для обработки компьютерной графики, они стали одним из наиболее мощных инструментов компьютерного анализа, благодаря своей относительно небольшой стоимости и высокой производительности. Это было достигнуто за счет фундаментальных изменений в архитектуре чипа. Она была оптимизирована для одновременного параллельного расчета огромного числа относительно простых операций. За счет уменьшения размера кэш-памяти на чипе, упрощения арифметико-логических устройств (Arithmetically-Logic Unit, ALU), относительного уменьшения количества ALU, работающих с двойной точностью, было значительно увеличено общее количество ALU на чипе. В результате чип GPU может содержать сотни вычислительных ядер, одновременно обрабатывающих поток данных.

Одной из проблем, возникающих при работе с большим количеством вычислительных потоков, является синхронизация. Поэтому все ALU на GPU, именуемые скалярными процессорами (Scalar Processor, SP), собраны в группы, называемые потоковыми мультипроцессорами (Streaming Multiprocessor, SM). Все программные нити внутри одного SM имеют доступ к общей разделяемой (shared) памяти. Доступ к ней осуществляется в сотни раз быстрее, чем к глобальной памяти GPU.

Фактически графические ускорители являются процессорами, построенными на основе технологии SIMD (Single Instruction Multiple Data). Это означает, к сожалению, что не все алгоритмы могут быть эффективно реализованы с использованием GPU.

Несмотря на высокую производительность графических ускорителей, долгое время использование их для решения задач обработки неграфических данных было затруднено. Применение платформ OpenGL и DirectX API требовало переформулирования математических алгоритмов в терминах обработки графики. Появление технологии CUDA (Compute Unified Device Architecture), предложенной NVIDIA, позволило существенно упростить разработку программ для графических ускорителей. CUDA является расширением языка C, позволяющим создавать многопотоковые приложения для устройств, поддерживающих эту архитектуру (NVIDIA CUDA programming guide 3.2. http://developer.download.nvidia.com/).

Table 1. Binary representation of the 15-letter IUPAc code for letters in a motif)

l etters of the IUPAc code		A	T	G	c	r	Y	M	K	w	S	B	H	V	D	n
Transcription of letters in the IUPAc code		**a**	**t**	**G**	**C**	**G/a**	**t /C**	**a /C**	**G/t**	**a /t**	**C/G**	**!a**	**!G**	**!t**	**!C**	**n**
n ucleotide	**a**	1	0	0	0	1	0	1	0	1	0	0	1	1	1	1
	t	0	1	0	0	0	1	0	1	1	0	1	1	0	1	1
	G	0	0	1	0	1	0	0	1	0	1	1	0	1	1	1
	C	0	0	0	1	0	1	1	0	0	1	1	1	1	0	1
c ode		1	2	4	8	5	10	9	6	3	12	14	11	13	7	15

Поскольку программа на CUDA запускается на GPU, процессоре с SIMD архитектурой, она должна содержать ядро, состоящее из вычислительных операций, одновременно рассчитываемых на GPU в виде множественных нитей (threads). Каждая нить имеет уникальный идентификатор, который может быть использован для получения соответствующих данных из памяти для обработки. Нити объединены в блоки (blocks), в свою очередь объединенные в сеть (grid). Это сделано для облегчения программной реорганизации нитей на структуру обрабатываемых данных. Например, нити могут быть организованы в виде как одномерной, так и двух- и трехмерной решетки. Нити, объединенные в блок, могут взаимодействовать друг с другом, синхронизироваться и использовать одну разделяемую память. Нити в блоке выполняются на мультипроцессоре SM одновременно группами, называемыми *warp*.

На основе технологии CUDA нами разработан новый метод выявления значимых вырожденных олигонуклеотидных мотивов, записанных в 15-буквенном коде IUPAC, позволяющий использовать в расчетах широко распространенные графические ускорители. С использованием предложенного подхода были проанализированы регуляторные районы генов *B. subtilis*, *E. coli*, *H. pylori*, *M. gallisepticum*, *M. genitalium* и *M. pneumoniae*. Для каждого вида прокариот были выявлены наборы вырожденных мотивов и проведена их классификация на основе сходства с сайтами связывания транскрипционных факторов *E. coli*. Мотивы, полученные таким образом, могут являться целью дальнейшего экспериментального анализа.

Материалы и методы

Нами предложен алгоритм расчета представленности мотива M длины l в выборке D, состоящей из N_{seq} последовательностей длины L_{seq}, основанный на оценке соответствия мотива M каждой из $N_{seq} \cdot (L_{seq} - l + 1)$ позиций выборки D. Для этого каждый символ мотива в 15-буквенном коде IUPAC записывается в виде целого числа от 1 до 15 (табл. 1), а каждый нуклеотид выборки анализируемых последовательностей D записывается в виде целого числа от 0 до 3 (табл. 2).

В этом случае соответствие между мотивом M длины l и районом $[i; i+l]$ анализируемой последовательности, записанной 4-буквенном коде, может быть оценено с помощью операции побитового сдвига вправо. При этом если буквы в позиции мотива M и анализируемой нуклеотидной последовательности соответствуют друг другу, то побитовый сдвиг вправо бинарного представления символа мотива M (табл. 1) на число, соответствующее бинарному представлению нуклеотида (табл. 2), выдаст 1, в противном случае – 0. Таким образом, если все символы мотива и сравниваемого участка последовательности соответствуют друг другу, произведение результатов побитового сдвига для всех позиций будет равным 1. Подобный подход позволяет существенно ускорить оценку соответствия мотива и нуклеотидной последовательности.

Table 2. Binary representation of 15-letter IUPAc code for letters in a nucleotide sequence

n ucleotide	**a**	**t**	**G**	**C**
c ode	0	1	2	3

Для оценки представленности всех 15^l возможных мотивов все рассматриваемые мотивы разбиваются на группы, равные количеству потоков в потоковом блоке, а каждый потоковый блок обрабатывает свою нуклеотидную последовательность. При этом все нуклеотидные последовательности анализируемой выборки D размещались в текстурной памяти, что позволило существенно ускорить доступ к этим последовательностям. Последовательность, с которой работает блок, копировалась в разделяемую память, поскольку доступ к ней существенно быстрее, чем к глобальной памяти.

Загрузка последовательностей из текстурной памяти производится всеми потоками блока. Размер разделяемой памяти на мультипроцессоре ограничивает длину последовательностей в ~14 тыс. нуклеотидов, что достаточно для решения большинства задач по анализу регуляторных районов генов. Сократить количество итераций обращения к текстурной памяти можно за счет использования упакованных типов данных. В нашем случае вместо char (один 8-битный символ) использовался uchar4 (четыре 8-битных символа), то есть, например, для загрузки одной последовательности длины $L = 2\,000$ нуклеотидов 512 потоками нам потребуется четыре итерации обращений к текстурной памяти для char и только одна – для uchar4.

Затем каждый поток в блоке проверяет встречаемость одного мотива в одной последовательности нуклеотидов и запоминает результат в глобальной памяти. В случае использования запакованных типов данных (uchar4) каждый поток может обрабатывать одновременно четыре последо-

Table 3. Properties of samples from regulatory regions of prokaryotic genes

Species	N_{seq}	N_{mot}	N_{TFBS}
B. subtilis	4 109	388	27
E. coli	4 173	334	42
H. pylori	1 565	454	35
M. gallisepticum	725	486	24
M. genitalium	212	452	28
M. pneumoniae	687	423	35

вательности. После этого запускается другое ядро на GPU, которое вычисляет встречаемость обработанной порции мотивов во всех последовательностях нуклеотидов. Пока на GPU идет обработка мотивов, на CPU готовится следующая их порция, и процесс повторяется.

После того как процесс расчета представленности для всего множества мотивов проведен, производятся оценка значимости полученных мотивов согласно биномиальному критерию (Vishnevsky, Kolchanov, 2005) и расчет их представленности в выборке случайных последовательностей. Случайная выборка генерировалась с частотами нуклеотидов, соответствующими частотам нуклеотидов в анализируемой выборке. Мотивы, не удовлетворяющие граничным критериям, удалялись из рассмотрения, а среди оставшихся мотивов выбирался наиболее значимый. Позиции этого мотива маскировались в выборке анализируемых последовательностей, и процесс оценки значимости оставшихся мотивов производился заново. Затем среди найденных мотивов выявлялся следующий по значимости мотив, производилась маскировка позиций его расположения в выборке последовательностей, и цикл поиска значимых мотивов повторялся до тех пор, пока в анализе оставались мотивы, удовлетворяющие граничным критериям.

Предложенный алгоритм был реализован в виде компьютерной программы на языке CUDA. Программа может работать в операционных системах Windows и Linux и позволяет оценивать представленность в заданной выборке нуклеотидных последовательностей всех вырожденных олигонуклеотидных мотивов длиной 8, записанных в 15-буквенном коде IUPAC. Программа обладает интерфейсом, в котором пользователь может задать границы окна в анализируемой выборке, граничный уровень значимости и представленности в выборке случайных последовательностей. Можно указать такие параметры случайной выборки, как количество последовательностей в ней и необходимость использования частот нуклеотидов, характерных для анализируемой выборки последовательностей. Поиск может проводиться как в прямой, так и комплементарной цепях ДНК. На вход программы подается выборка нуклеотидных последовательностей, записанных в формате FASTA. На выходе – набор полученных вырожденных олигонуклеотидных мотивов, удовлетворяющих заданным критериям.

В качестве примера использования предложенного метода нами проведен поиск вырожденных олигонуклеотидных мотивов длиной 8 в регуляторных районах генов шести видов прокариот. Для этого из базы данных GenBank (Benson et al., 2013) были экстрагированы выборки [–100; +25] районов относительно старта трансляции для *B. subtilis*, *E. coli*, *H. pylori*, *M. gallisepticum*, *M. genitalium* и *M. pneumoniae*. Поиск проводился в трех окнах шириной 50 нуклеотидов: [–100; –50], [–75; –25] и [–25; +25] относительно старта трансляции. Хорошо видно (табл. 3), что построенные выборки значительно различались по количеству содержащихся в них последовательностей N_{seq} – от 212 для *M. genitalium* до 4 173 для *E. coli*.

В каждой из построенных выборок проводился поиск значимых вырожденных олигонуклеотидных мотивов. В качестве достоверных рассматривались мотивы, чья представленность в выборке регуляторных районов превышала 10 % (для отбора относительно слабо вырожденных мотивов), а вероятность наблюдения по случайным причинам не превышала 10^{-8}. Затем была проведена классификация выявленных мотивов с использованием базы данных сайтов связывания транскрипционных факторов *E. coli* (Osada et al., 2004).

результаты и обсуждение

Оценка производительности работы программы на разных вычислительных устройствах

Для того чтобы получить оценки производительности работы программы, не зависящие ни от длины и количества анализируемых последовательностей, ни от длины мотивов, будем измерять производительность в количестве сравнений позиций мотивов с позициями выборки последовательностей за единицу времени. Для набора мотивов M и выборки последовательностей D производительность G вычисляется следующим образом:

$$G = \frac{|M| \cdot |D|}{t \cdot 10^9} = \frac{l \cdot N_{mot} \cdot N_{seq} \cdot (L_{seq} - l + 1)}{t \cdot 10^9},$$

где $|M|$ – суммарное количество всех букв в наборе вырожденных олигонуклеотидных мотивов, $|D|$ – суммарное количество всех позиций (в выборке последовательностей), в которых возможно сравнение с мотивами, t – время работы в секундах, N_{seq} – количество последовательностей в выборке, L_{seq} – длина последовательностей, l – длина мотивов.

С помощью предложенной меры мы провели оценку эффективности использования разработанного метода на различных графических ускорителях и CPU. В качестве CPU использовался четырехъядерный процессор i7-950

с частотой 3 ГГц. Распараллеливание расчетов между различными ядрами CPU проводилось с использованием библиотеки POSIX Threads. На рис. 1 приведены сравнительные оценки производительности *G* на различных вычислительных платформах на выборке из 10 последовательностей в зависимости от их длины.

Из рис. 1 хорошо видно, что наиболее производительным решением является сдвоенная Tesla C1060, которая суммарно обладает 480 процессорами. Отметим, что все графические ускорители имеют худшую производительность при малой длине последовательности. Это можно объяснить тем, что внутренний цикл становится достаточно коротким и запуск вычислительного ядра происходит чаще. Tesla C2050 обладает лучшей производительностью при длине последовательности 32 символа. При длине последовательностей от 50 нуклеотидов производительность всех GPU стабилизируется. Показано, что метод хорошо шкалируется между различными вычислительными устройствами и его производительность растет с ростом выборки. В среднем отставание CPU i7-950 от сдвоенной GPU C1060 составляет 18 раз. Кроме того, при наличии 4 ядер и частоты 3ГГц i7-950 в 1,5 раза быстрее, чем бюджетная видеокарта geForce210 с 16 процессорами и частотой 1,4 ГГц (данные не приведены). Производительность одного графического ускорителя Tesla C2050 примерно в 70 раз выше, чем производительность одного ядра CPU i7-950.

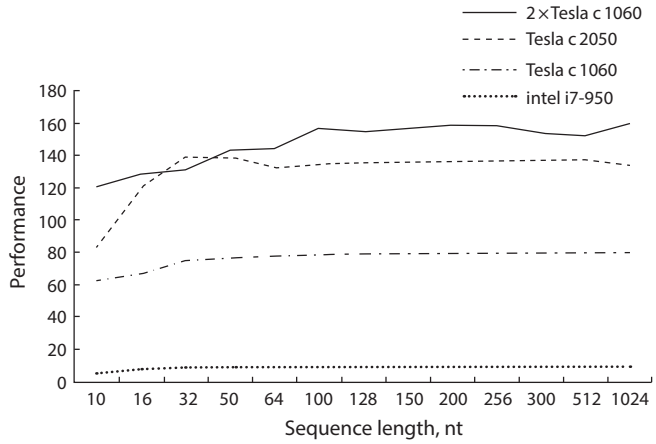


Fig. 1. Performance evaluation *G* of the program identifying degenerate oligonucleotide motifs vs. sequence length. The performance was evaluated on various devices.

Выявление вырожденных олигонуклеотидных мотивов в регуляторных районах генов прокариот

Для каждой из построенных выборок регуляторных районов были получены сотни вырожденных мотивов N_{mot} (табл. 3). Их количество варьировало от 334 для *E. coli* до 486 для *M. gallisepticum*. Отметим, что количество выявляемых мотивов N_{mot} и размер анализируемой выборки N_{seq} не демонстрируют явной корреляции между собой. В табл. 4 приведен пример мотивов, найденных в регуляторных районах генов *M. genitalium*. Например, мотив **AAAYWAAA** представлен в 11 % ($F = 0{,}11$) анализируемых последовательностей, в 3 % ($Q = 0{,}03$) случайных последовательностей, а вероятность его наблюдения по случайным причинам в анализируемой выборке – 10^{-989} ($-P = 984$). Для оценки величины Q использовались выборки, состоящие из 700 случайных последовательностей, сгенерированных с частотами нуклеотидов, характерными для частот нуклеотидов выборки регуляторных районов генов соответствующего вида прокариот. Отметим, что следующий за ним по значимости мотив **AAAAYWAR** является практически полным подобием **AAAYWAAA** со сдвигом и добавлением к нему нуклеотида **A** на 3′-конце, т. е. функционально значимым для регуляторных районов *M. genitalium*, видимо, является контекстный сигнал длины большей, чем восемь нуклеотидов. Такие сигналы можно выявлять в наборе полученных мотивов с использованием различных методов оценки подобия и рассматривать отдельно.

Затем мы провели классификацию полученных мотивов с использованием базы данных сайтов связывания транскрипционных факторов *E. coli* (Osada et al., 2004) с уровнем значимости $p < 10^{-4}$. Как и ожидалось, наибольшее количество ССТФ, демонстрирующих значимое сходство с выявленными мотивами, наблюдалось для *E. coli* ($N_{TFBS} = 42$). Можно предположить, что часть ССТФ-специфичных мотивов, полученных для других видов прокариот, соответствуют транскрипционным факторам – ортологам транскрипционных факторов, найденных в *E. coli* (табл. 5). Оставшаяся часть выявленных мотивов может соответствовать как видоспецифичным ССТФ, отсутствующим в базе данных ССТФ *E. coli*, так и некоторым структурным физико-химическим особенностям регуляторных районов генов прокариот, таким, например, как короткие поли-A / поли-T тракты, приводящие к формированию участков повышенной «плавкости» (easily melting sites) и специфическому изгибу ДНК.

Нам показалось интересным оценить относительное сходство и взаиморасположение вырожденных мотивов в регуляторных районах генов эволюционно близких и эволюционно удаленных видов прокариот. Ранее (Vishnevsky et al., 2011) нами был предложен метод усредненной оценки межвидового олигонуклеотидного сходства регуляторных районов генов H_{Oli}. Этот метод учитывает как степень вырожденности мотивов, так и характер их расположения в регуляторных районах. На рис. 2 показаны оценки такого сходства H_{Oli}, полученные для разных видов, рассчитанные на основе вырожденных мотивов, полученных для *M. genitalium*. Для нижней оценки величины олигонуклеотидного сходства использовалась выборка, состоящая из 700 случайных последовательностей, сгенерированных с частотами нуклеотидов, характерными для частот нуклеотидов выборки регуляторных районов генов *M. genitalium*. Хорошо видно, что максимальные значения олигонуклеотидного сходства с регуляторными районами генов *M. genitalium* наблюдаются для наиболее близких к нему видов, таких как *M. gallisepticum* ($H_{Oli} = 0{,}53$) и *M. pneumoniae* ($H_{Oli} = 0{,}5$), в то время как эволюционно удаленные от него виды, такие как *B. subtilis* и *E. coli*, имеют существенно меньшие значения H_{Oli} – 0,22 и 0,08 соответственно. Кроме того, с использованием метода парного выравнивания мы оценили усредненную меж-

Table 4. An example of motifs found in [–100; –50] regions of *M. genitalium* with reference to the translation start

Motif	*F*	*Q*	*–P*
AAAYWAAA	0.11	0.03	989
AAAAYWAR	0.14	0.05	804
AAAMAASM	0.12	0.02	603
SAAAAAMW	0.12	0.04	558
AWYTNWTT	0.21	0.10	392
WAAYTRWT	0.14	0.07	374
NWAGMAAA	0.14	0.07	348
TRCWANWK	0.12	0.08	34
RKMMTTWK	0.12	0.08	34
YAASYTWN	0.12	0.08	34
MWYTWTKS	0.12	0.05	34
WKWWGSWK	0.18	0.08	30
SMASMANT	0.11	0.05	30
AASNSYTW	0.11	0.05	26

Designations: *F*, motif occurrences in the sample analyzed; *Q*, motif occurrences in the sample of random sequences generated with the nucleotide frequencies specific o the analyzed sample; *P*, decimal logarithm of the probability of observing the motif by chance.

Table 5. An example of motifs found in the [–100; –50] regions of *H. pylori* relative to the translation start that demonstrates a significa t similarity to *E.coli* TFBSs

Motif	TFBS	*–P*
SNSAWRAA	metr	6
KRRGAWWR	lrp	5
NWMTTWAA	nagc	4
WWKSSMTT	metJ	4
AWAAYNSM	argr	4

P, decimal logarithm of observation probability of the motif in E. coli TFBSs by chance.

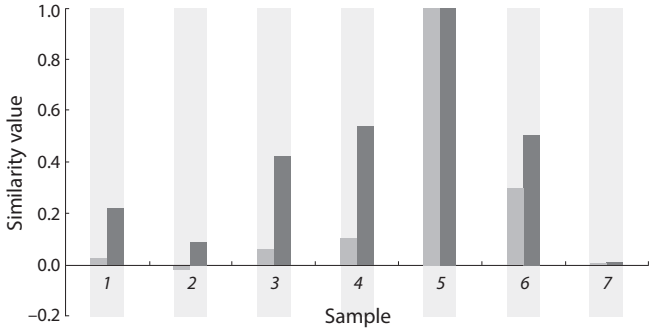


Fig. 2. Values of the average oligonucleotide similarity H_{Oli} (dark box) and the averaged interspecies homology H_{align} (light box) in the regulatory regions of *M. genitalium* genes with the regulatory regions of genes of other prokaryotes.

1, *B. subtilis*; *2*, *E. coli*; *3*, *H. pylori*; *4*, *M. gallisepticum*; *5*, *M. genitalium*; *6*, *M. pneumoniae*; *7*, random sample.

видовую гомологию H_{align} регуляторных районов генов различных видов прокариот согласно Vishnevsky et al., (2011). На рис. 2 показаны оценки усредненной гомологии регуляторных районов генов 5 видов прокариот с регуляторными районами генов *M. genitalium*. Видно, что усредненная гомология с *M. genitalium* резко снижена даже для эволюционно близких к нему видов, а для эволюционно далеких видов она находится на случайном уровне.

Таким образом, на основе анализа величин H_{Oli} и H_{align} можно сделать вывод, что общий контекст регуляторных районов генов может практически полностью изменяться в ходе эволюции и разделения видов, в то время как регуляторный код, основанный на присутствии в этих районах специфических контекстных сигналов, остается в значительной степени консервативным. Можно предположить, что это обусловлено как относительной консервативностью сайтов связывания транскрипционных факторов, так и сходством физико-химических структурных особенностей регуляторных районов генов.

На основе технологии CUDA нами разработан программный комплекс для итерационного выявления в регуляторных районах генов наборов значимых вырожденных

олигонуклеотидных мотивов фиксированной длины, записанных в 15-буквенном коде IUPAC. Предложенный метод может использовать для вычислений широко распространенные бюджетные графические ускорители. Было показано, что применение GPU позволяет в десятки раз увеличить производительность системы в сравнении с распространенными CPU процессорами. На базе предложенного подхода с использованием языков Perl, C++ и CUDA нами разрабатывается Интернет-доступный сайт, который даст исследователям новый инструмент для изучения регуляторных районов генов.

Проведенный анализ регуляторных районов генов *B. subtilis*, *E. coli*, *H. pylori*, *M. gallisepticum*, *M. genitalium* и *M. pneumoniae* выявил сотни значимых олигонуклеотидных мотивов. Оценка сходства полученных мотивов с сайтами связывания транскрипционных факторов *E. coli* показала, что только часть мотивов демонстрируют достоверное сходство с этими ССТФ. Мы полагаем, что оставшиеся неклассифицированные мотивы могут соответствовать сайтам связывания видоспецифичных транскрипционных факторов или разнообразным структурным особенностям регуляторных районов, необходимым для нормального протекания процессов транскрипции или трансляции. Полученные мотивы могут являться мишенями для дальнейшего экспериментального анализа.

Выявленные нами мотивы были использованы для оценки усредненного межвидового сходства регуляторных районов генов прокариот различной эволюционной удаленности друг от друга. Оказалось, что, хотя регуляторные районы генов разных видов прокариот в ходе эволюционного расхождения и накопления мутаций различаются крайне сильно, они продолжают сохранять в относительно консервативном виде специфические контекстные сигналы, обеспечивающие регуляцию базовых молекулярно-генетических процессов.

Acknowledgments

Computations were done at the Bioinformatics Shared Access Center of the Institute of the Cytology and Genetics; Siberian Supercomputer Center of SB RAS; and the Supercomputer Center of the Novosibirsk State University. This work was supported by the Institute of Cytology and Genetics, Budgeted Project VI.61.1.2.

Conflict of interest

The authors declare no conflict of interest.

References

Baker Z.K., Prasanna V.K. An architecture for efficient hardware data mining using reconfigurable computing systems. 14th Annual IEEE Symp. on Field-Programmable Custom Computing Machines, 2006.

Benson D.A., Cavanaugh M., Clark K., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W. GenBank. Nucl. Acids Res. 2013;41(Database issue):D36-42.

Elnitski L., Hardison R.C., Yang S., Kolbe D., Eswara P., O'Connor M.J., Schwartz S., Miller W. Chiaromonte F. Distinguishing regulatory DNA from neutral sites. Genome Res. 2003;13(1):64-72.

Fomin E.S., Alemasov N.A. Implementation of a non-bonded interaction calculation algorithm for the cell architecture. Lect. Notes Comput. Sci. 2009;5698:399-405.

Grundy W.N., Bailey T.L., Elkan C.P. ParaMEME: a parallel implementation and a web interface for a DNA and protein motif discovery tool. CABIOS. 1996;12:303-310.

Hertz G.Z, Stormo G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics. 1999;15:563-577.

Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucl. Acids Res. 2002;30:312-317.

Lawrence C.E., Altschul S.F., Boguski M.S., Liu J.S., Neuwald A.F., Wootton J.C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 1993;262:208-214.

Manavski S.A., Valle G. CUDA compatible GPU cards as efficient hardware accelerators for Smith–Waterman sequence alignment. BMC Bioinformatics. 2008;26;9 Suppl 2:S10.

Marsan L., Sagot M.F. Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. J. Comput. Biol. 2000;7:345-362.

Matys V., Kel-Margoulis O.V., Fricke E., Liebich I., Land S., Barre-Dirrie A., Reuter I., Chekmenev D., Krull M., Hornischer K., Voss N., Stegmaier P., Lewicki-Potapov B., Saxel H., Kel A.E., Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucl. Acids Res. 2006;34:D108-10.

Mrázek J., Gaynon L.H., Karlin S. Frequent oligonucleotide motifs in genomes of three streptococci. Nucl. Acids Res. 2002;19:4216-4221.

NVIDIA CUDA programming guide 3.2. [http://developer.download.nvidia.com/compute/cuda/3_2/toolkit/docs/CUDA_C_Programming_Guide.pdf]

Osada R., Zaslavsky E., Singh. M. Comparative analysis of methods for representing and searching for transcription factor binding sites. Bioinformatics 2004;20(18):3516-3525.

Pesole G., Liuni S., Dsouza M. PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. Bioinformatics. 2000;16:439-450.

Pevzner P.A., Sze S.H. Combinatorial approaches to finding subtle signals in DNA sequences. Proc. of the 8th Int. Conf. on Intelligent Systems for Molecular Biology (ISMB). 2000.

Portales-Casamar E., Thongjuea S., Kwon A.T., Arenillas D., Zhao X., Valen E., Yusuf D., Lenhard B., Wasserman W.W., Sandelin A. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucl. Acids Res. 2010;38:D105-10.

Sukhwani B., Herbordt M.C. GPU acceleration of a production molecular docking code. Proc. of 2nd Workshop on General Purpose Processing on Graphics Processing Units. 2009.

Vishnevsky O.V., Gunbin K.V., Bocharnikov A.V., Berezikov E.V. Analysis of the conservative motifs in promoters of miRNA genes, expressed in different tissues of mammalians. Evolutionary Biology Concepts, Molecular and Morphological Evolution. 2011.

Vishnevsky O.V., Kolchanov N.A. ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters. Nucl. Acids Res. 2005;33(Web Server issue):417-22.

Yooseph S., Sutton G., Rusch D.B., Halpern A.L., Williamson S.J., Remington K., Eisen J.A., Heidelberg K.B., Manning G., Li W., Jaroszewski L., Cieplak P., Miller C.S., Li H., Mashiyama S.T., Joachimiak M.P., van Belle C., Chandonia J.M., Soergel D.A., Zhai Y., Natarajan K., Lee S., Raphael B.J., Bafna V., Friedman R., Brenner S.E., Godzik A., Eisenberg D., Dixon J.E., Taylor S.S., Strausberg R.L., Frazier M., Venter J.C. The sorcerer II global ocean sampling expedition: expanding the universe of protein families. PLoS Biol. 2007:5(3):e16.

Genomics and Polymorphism analysis
original article

Вавиловский журнал генетики и селекции. 2015;19(6):668-674
Doi 10.18699/VJ15.092

Фланкирующие повторы мономеров определяют пониженную контекстную сложность сайтов однонуклеотидных полиморфизмов в геноме человека

Н.С. Сафронова[1, 2], М.П. Пономаренко[1, 2], И.И. Абнизова[3], Г.В. Орлова[1], И.В. Чадаева[1], Ю.Л. Орлов[1, 2]

1 Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия 2 Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия 3 Центр Сенгера, Кембридж, Великобритания

Исследование зависимости частоты возникновения мутаций в геноме человека выполнено на примере набора документированных однонуклеотидных полиморфизмов (ОНП) из проекта «1 000 геномов». Рассмотрены задачи разработки новых компьютерных методов статистического анализа генетических текстов на основе оценок сложности последовательности символов. Показано применение профилей сложности в скользящем окне к анализу сайтов, содержащих однонуклеотидные полиморфизмы в геноме человека. Установлено локальное понижение сложности текста в районе ОНП. На основе анализа профилей сложности в участках, содержащих ОНП, показано, что фланкирующие повторы мономеров определяют пониженную контекстную сложность сайтов однонуклеотидных полиморфизмов в геноме человека. Эффект локального понижения уровня сложности текста последовательностей фланкирующих сайты ОНП подтвержден для данных о полиморфизмах в геномах крысы и мыши. Определены различия в контекстной организации для кодирующих и регуляторных последовательностей, которые отражаются в сложности текста нуклеотидных последовательностей, содержащих ОНП. Изменения в частоте точковых мутаций были ранее показаны для последовательностей, содержащих микросателлиты. С использованием более общего математического аппарата и более полных данных в работе показана насыщенность политрактами и простыми повторяющимися последовательностями локального геномного окружения участков, содержащих ОНП. Определены олигонуклеотиды с повышенной частотой встречаемости в геномном окружении ОНП у человека, показана их связь с политрактами. Присутствие политрактов может свидетельствовать о большей вероятности разрыва двойной цепи ДНК в этой точке, приводящей к повышению частоты замен нуклеотидов. Полученные оценки были определены при помощи разработанного ранее комплекса компьютерных программ, который кроме оценки сложности фазированных выборок позволяет эффективно определять частотный спектр олигонуклеотидов фиксированной длины, производить сравнение частот олигонуклеотидов в выборках большого объема.

Ключевые слова: ОНП; геном; нуклеотидные последовательности; повторы; энтропия; мутации.

Received 17.09.2015
Accepted for publication 27.10.2015

e-mail: taschasafronova@bionet.nsc.ru; orlov@bionet.nsc.ru

Flanking monomer repeats define lower context complexity of sites containing single nucleotide polymorphisms in the human genome

N.S. Safronova[1, 2], M.P. Ponomarenko[1, 2], I.I. Abnizova[3], G.V. Orlova[1], I.V. Chadaeva[1], Y.L. Orlov[1, 2] . Абнизова[3], Г.В. Орлова[1], И.В. Чадаева[1], Ю.Л. Орлов[1, 2]

1 Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
2 Novosibirsk State University, Novosibirsk, Russia
3 Sanger Center, Cambridge, UK

we have investigated a mutation frequency within the human genome for the set of known single nucleotide polymorphisms (SnPs) from the "1000 genomes" project. we have developed and applied novel statistical computational methods to analyze genetic text based on its complexity. A complexity profiling in a sliding window is applied to the sites containing single nucleotide polymorphisms within the human genome. A local decrease in text complexity level in SnP-containing sites has been shown. Analysis of the complexity profiles for SnP-containing sites shows that flan ing monomer repeats define a l wer context complexity of sites containing SnPs within the human genome. An effect of local decrease in text complexity in SnP-containing sites is confi med by analysis of polymorphisms in the rat and mouse genomes. we have found context differences between coding and regulatory sequences. These differences refle t a complexity of SnP-containing loci. The changes in point mutation frequency were shown previously for microsatellite containing sequences. Using enhanced mathematical tools and larger data sets this work shows enrichment of polytracks and simple sequence repeats in local genome surroundings of SnP containing sites. we have found high-frequency oligonucleotides within genomic regions containing SnPs. Such oligonucleotides are related to nucleotide polytracks. The presence of poly-A tracks might be associated with an increased probability of double helix DnA breaks around mutable loci and following fix tion of nucleotide changes. The complexity estimates were computed using a previously developed program tool. This tool allows for both (i) complexity estimation of phased samples, and (ii) rapid and effective identific tion of the frequency spectrum of oligonucleotides with fi ed lengths, and a comparison of oligonucleotide frequencies in different samples.

Key words: SnP; genome; nucleotide sequences; repeats; entropy; mutations.

HOW TO CITE THIS ARTICLE?
Safronova n .S., Ponomarenko M.P., Abnizova i.i., o rlova G.V., c hadaeva i.V., o rlov Y.I . Flanking monomer repeats define lower context complexity of sites containing single nucleotide polymorphisms in the human genome. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2015;19(6):668-674. Doi 10.18699/VJ15.092

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?
Сафронова Н.С., Пономаренко МП, Абнизова И.И., Орлова Г.В., Чадаева И.В., Орлов Ю.л. Фланкирующие повторы мономеров определяют пониженную контекстную сложность сайтов однонуклеотидных полиморфизмов в геноме человека. Вавиловский журнал генетики и селекции. 2015;19(6):668-674. Doi 10.18699/VJ15.092

Исследование нуклеотидных полиморфизмов в геноме имеет большое значение для фундаментальной и прикладной медицинской генетики. Изучение нуклеотидных замен изменений, а также предпосылок их возникновения по нуклеотидным последовательностям может позволить ответить на ряд важных молекулярно-биологических вопросов о ходе мутационного процесса, достигнуть успехов в предсказании и лечении генетических заболеваний, связанных с естественной изменчивостью генома. Развиваются международные проекты HapMap (International HapMap 3 Consortium, 2010), «1 000 Genomes» (http://www.1000genomes.org/), идут национальные и региональные исследования генетической изменчивости (Sidore et al., 2015; UK10K Consortium, 2015).

Исследование зависимости частоты возникновения мутаций в геноме человека было выполнено нами на примере набора документированных однонуклеотидных полиморфизмов из проекта «1 000 геномов». Однонуклеотидные полиморфизмы (ОНП, или SNP – Single Nucleotide Polymorphism) – однонуклеотидные различия последовательностей ДНК в геноме или участке генома. Такие полиморфизмы имеют большое значение при изучении различных заболеваний, что требует развития биоинформационных ресурсов анализа (International HapMap 3 Consortium, 2010).

За последние два десятилетия создан большой набор программных продуктов, направленных на изучение свойств и структуры последовательностей ДНК и белков (Babenko et al., 1999; Орлов, 2012; Игнатьева и др., 2015). Одной из важных проблем исследования геномной ДНК является анализ сложности генетических текстов с помощью математических оценок, учитывающих эволюционные ограничения на изменение последовательности (Орлов, 2004; Orlov, Potapov, 2004; Orlov et al., 2006). Исследование сложности текста нуклеотидных и аминокислотных последовательностей как независимой универсальной характеристики имеет свой широкий круг применений: от анализа строения регуляторных районов генов до анализа расположения повторов в полных геномах (Chuzhanova et al., 2002; Orlov et al., 2006; Trifonov et al., 2012).

Связь так называемых «горячих точек» мутаций в геномах с окружающим нуклеотидным контекстом была показана для разных организмов (Rogozin et al., 1991, 2001; Rogozin, Kolchanov, 1992). Изменения в частоте точковых мутаций были ранее показаны для последовательностей, содержащих микросателлиты (Siddle et al., 2011). Ряд исследований, рассматривающих ди- и тринуклеотидные повторы в геноме, показал увеличение скорости мутаций в районе повтора (Vowles, Amos, 2004; Siddle et al., 2011). Опираясь на более общий математический аппарат и более полные данные (Safronova et al., 2015), мы показываем насыщенность политрактами и простыми повторяющимися последовательностями локального геномного окружения участков, содержащих ОНП.

Материалы и методы

В работе использовали базу данных dbSNP и базу документированных однонуклеотидных полиморфизмов из проекта «1 000 геномов». Нуклеотидные последовательности загружали из ресурса UCSC Genome Browser (genome.ucsc.edu). Для численной оценки сложности применяли алгоритмы разработанной ранее компьютерной программы (Orlov, Potapov, 2004): расчет числа операций, необходимых для сжатия текста алгоритмом Лемпеля и Зива (Gusev et al., 1999), расчет комбинаторной (лингвистической) сложности текста (Troyanskaya et al., 2002). Эти алгоритмы были дополнены оценками: 1) энтропии символов (Wootton, Federhen, 1996); 2) присутствия политрактов; 3) частоты чередования символов в последовательности.

Каждое определение сложности текста учитывает определенные аспекты его структурной организации. Энтропия представляет неравномерность олигонуклеотидного состава. Алгоритмы оценки сложности используют представление последовательности символов в виде *l*-граммного дерева (Orlov et al., 2002), которое позволяет рассчитывать число всех возможных 4^l слов длины l в последовательности, определять позиции этих слов и с минимальными затратами компьютерного времени выполнять операции по расчету сложности текста и поиску гомологии.

Операционная сложность, или сложность по методу Лемпеля и Зива – это число операций копирования (дупликаций коротких последовательностей), необходимых для порождения текста из некоторого базового текста

(самой последовательности) в так называемом сложностном разложении (Gusev et al., 1999):

$$C_{LZ}(S) = S[1; i_1]\,S[i_1+1; i_2]\,\ldots\,S[i_{m-1}+1; L], \quad (1)$$

где $S[i_{k-1}+1; i_k]$ – фрагмент исходной строки, генерируемый на k-м шаге из уже сгенерированной последовательности, слева направо, L – длина текста (в нуклеотидах), m – минимальное число фрагментов (число операций копирования), определяющее сложность текста.

Анализ такой сложности учитывает все типы повторов, включая инверсии и симметричные повторы, и дает полную картину всех, а не только локальных, взаимозависимостей в тексте (Orlov, Potapov, 2004; Orlov et al., 2006). Чем короче повторы, тем больше операций копирования и больше сложность текста. Операционная сложность хорошо подходит для поиска повторов в протяженных последовательностях. Она наиболее близка к общему определению сложности, по Колмогорову, как длины кратчайшей программы, необходимой для генерации символьной последовательности.

Комбинаторная, или лингвистическая, сложность определяется как отношение числа встретившихся слов к числу всех возможных слов в последовательности заданной длины (Troyanskaya et al., 2002):

$$C_{ling} = \frac{\sum_1^L N_i}{\sum_1^L N_{max_i}}, \quad (2)$$

где N_i – число встретившихся слов длины i, N_{max_i} – максимальное число возможных слов длины i в последовательности длины L, которое равно минимуму из двух чисел: числа всех возможных слов длины i для используемого алфавита, либо числа слов длины i, которые можно разместить в последовательности длины L.

Чем меньше повторов всех длин, чем больше разнообразие найденных слов, тем больше сложность последовательности. Комбинаторная сложность учитывает существенные особенности текста, позволяет быстро находить простые участки (например короткие повторы) в последовательности.

Сложность частот чередования символов оценивает относительное число смен одного символа на другой в линейном расположении. Чем больше это число, тем более перемешаны символы и тем меньше политрактов – слов, состоящих из одного символа. Таким образом, эта мера позволяет эффективно находить локальную насыщенность политрактов. В то же время не учитывается разнообразие словаря возможных слов. Эта сложность рассчитывается по формуле:

$$C_{var} = Ch/(L-1), \quad (3)$$

где Ch – количество смен текущего символа, L – количество символов в строке (длина последовательности).

Рассмотрим величину C_{var} на примере коротких олигонуклеотидов длины 6: олигонуклеотид AAAAAA в интуитивном понимании проще, чем AAATAA, который в свою очередь проще ATGCGA; таким образом, в рамках введенной меры AAAAAA имеет сложность 0, AAATAA – 0,4 (2/5), ATGCGA – 1 (5/5).

Стандартный метод – оценка энтропии символов (энтропии Шеннона) в тех же обозначениях:

$$CE = -\sum_{i=1}^{K}(n_i/L)\log_K(n_i/L), \quad (4)$$

здесь L – размер последовательности, n_i – число встретившихся символов (нуклеотидов) i-го типа, i = 1, 2, ..., K, K – размер алфавита (K = 4 для ДНК, K = 20 для аминокислот). Поскольку логарифм берется по основанию K, все значения сложности содержатся в интервале (0; 1).

Для анализа коротких повторов в тексте (от 3 до 10 нт) рассмотрены меры сложности олигонуклеотидов. Таким образом, рассматривается сложность элементов (слов, олигонуклеотидов), которые могут быть выделены в исследуемых участках генома. Например, для олигонуклеотида AAATAAA, состоящего из повтора, за исключением одного нуклеотида, T, энтропия составляет 0,296, для ATATATA – 0,493, для TTATGCGC, содержащего уже все четыре нуклеотида, энтропия равна 0,979. Более подробно см. в Дополнительных материалах[1].

Ранее была рассмотрена связь различных мер сложности на геномных последовательностях, было отмечено, что все меры коррелируют между собой, при этом мера Лемпеля–Зива наиболее эффективно выделяет длинные повторы, а мера чередования символов более чувствительна к наличию политрактов (Орлов, 2004; Orlov et al., 2006).

результаты

Нуклеотидные последовательности из генома человека, содержащие аннотированные ОНП, были загружены из UCSC Table Browser (http // genome.ucsc.edu.). Мы использовали последовательности размером 101 нт (1 позиция НП и +/–50 нт окружение). Был выполнен расчет профилей сложности в скользящем окне размеров 10, 7, 5 нт. Рассматривалось отдельно четыре класса ОНП – замена A на другую букву (нуклеотид), C – на другой нуклеотид и т. д. Построенные профили сложности были усреднены.

Замечено, что в районе мутации в скользящем окне профиля происходит резкое изменение сложности (Safronova et al., 2015) (рис. 1).

Как показал расчет стандартного отклонения значений сложности, при большем числе проанализированных последовательностей такое изменение профиля статистически значимо ($p < 0{,}01$). Дальнейший анализ показал, что сложность последовательности в скользящем окне возрастает к точке полиморфизма, а после нее симметрично убывает, в районе ОНП происходит скачок сложности. Было замечено, что у всех наблюдаемых организмов в районе мутации профиль сложности имеет очень схожее поведение: сначала идет убывание, затем следует скачок и снова убывание (яма–пик–яма).

Отмечено локальное симметричное понижение сложности в районе вокруг точки полиморфизма. Дальнейший частотный анализ показал высокую насыщенность области около полиморфизма (+/–10 нт) поли-A и поли-T трактами. Анализ дисперсии значений сложности в точке ОНП и на флангах (5 нт) показал статистическую значимость различия средних значений (рис. 1).

Для модельных организмов – мыши и крысы – были получены схожие результаты, подтвержден эффект, обнаруженный на более полных данных по человеку (рис. 2).

[1] Дополнительные материалы см. в Приложении 1 по адресу: http://www.bionet.nsc.ru/vogis/download/pict-2015-12/appx1.pdf

На рис. 2 показано симметричное изменение профиля сложности с пиком на точке полиморфизма для участков, содержащих ОНП в геноме крысы.

Мера вариабельности C_{var} (мера чередования символов) в скользящем окне также для участков ОНП в геноме человека представлена на рис. 3.

Профиль меры чередования символов имеет тот же вид, что и сложность по Лемпелю – Зиву (число повторяющихся фрагментов), представленная на рис. 1, с локальными минимумами на флангах точки ОНП до 10 нт.

Локальное понижение сложности текста отмечено и с помощью других мер – лингвистической сложности, энтропии и меры чередования нуклеотидов. Минимальные значения сложности соответствуют последовательностям, состоящим почти целиком из одной повторяющейся единицы (тандемного повтора), самое минимальное значение дает политракт (однобуквенный повтор), например $(A)_n$. На рис. 4 приведены профили встречаемости наиболее представленных олигонуклеотидов, в данном случае политрактов, в последовательностях, фазированных относительно точки ОНП.

Присутствие сигнала TAA в горячих точках мутаций было показано еще в ранней работе И.Б. Рогозина и Н.А. Колчанова на ограниченной выборке данных (Rogozin, Kolchanov, 1992). В нашей работе представлен расширенный набор олигонуклеотидов, содержащих короткие повторы нуклеотидов A и T, непосредственно фланкирующих точки ОНП, которые также связаны с повышенным уровнем мутаций.

С использованием набора методов оценок сложности текста были рассчитаны значения сложности для нуклеотидных последовательностей различных типов – белок-кодирующих (экзоны), некодирующих (интроны) и регуляторных (промоторы и энхансеры). Ранее было показано различие мер сложности для кодирующих и некодирующих последовательностей (Orlov et al., 2006). Белок-кодирующие последовательности несут большую нагруженность сигналами различных типов (триплетный код аминокислот, информация о вторичной и пространственной структуре

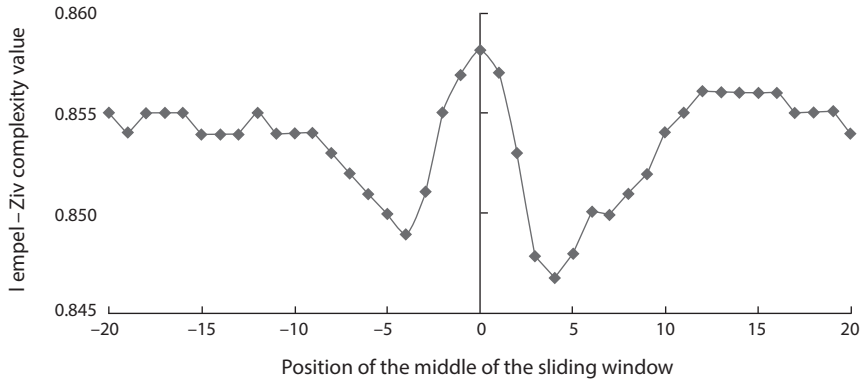


Fig. 1. Profile of l empel – Ziv complexity values in the sliding window for Sn P sites in the human genome.

w indow size 7 nt. r eplacements of nucleotide A in Sn P for three other nucleotides.

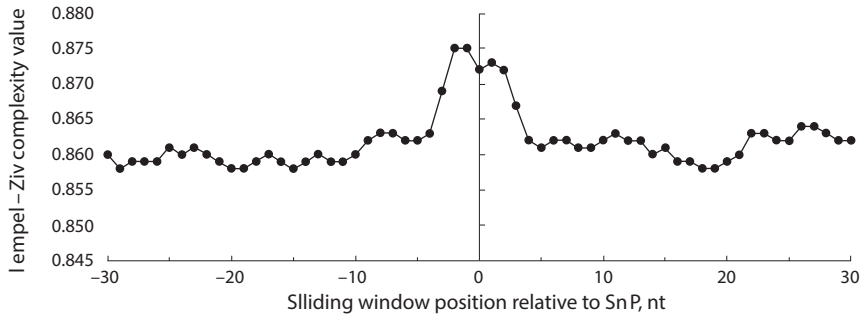


Fig. 2. Profile of l empel – Ziv complexity value changes in the 7-nt sliding window for Sn P sites in the rat genome.

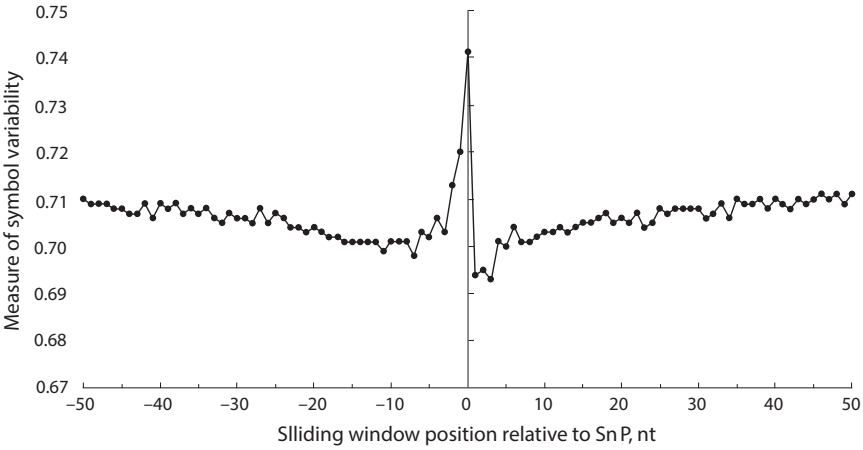


Fig. 3. Profile of the f equency of symbol changes in sequence (monomer variability) for Sn P sites in the human genome.

белка), и, следовательно, большие сложность текста и энтропию символов (Trifonov et al., 2012). Некодирующие последовательности (интроны), напротив, свободны от сигналов структуры белка, соответственно, имеют меньшую сложность. В то же время регуляторные последовательности содержат ряд размытых контекстных сигналов – о сайтах связывания белков, сайте посадки транскрипционного комплекса, но не о кодировании белка. Регуляторные последовательности занимают промежуточное положение между интронами

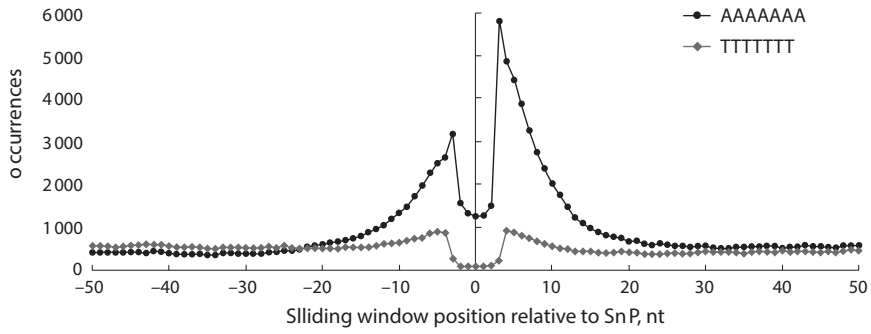


Fig. 4. Profile of 7-bp long pol -A tracts occurrence in Sn P-containing sites in the human genome.

и экзонами. Для более коротких последовательностей результаты различий уже не видны. Поэтому мы использовали построение профилей в скользящем окне для позиций однонуклеотидных полиморфизмов.

Для сайтов, содержащих ОНП, минимальные значения сложности соответствуют последовательностям, состоящим почти целиком из одной повторяющейся единицы (тандемного повтора), самое минимальное значение дает политракт. Таким образом, для участков низкой сложности, фланкирующих ОНП, характерно наличие простых участков (содержащих короткие тандемные повторы, политракты), что подтверждает ранние результаты (Rogozin et al., 1991; Ponomarenko et al., 2002; Medvedeva et al., 2013).

Обсуждение

В связи с возможными молекулярными механизмами патогенеза ОНП в кластерах повторяющейся ДНК представляется интересным, прежде всего, что в обзоре Поляновского с коллегами (2012) к числу самых важных причин устойчивости к моноклональным антителам-ингибиторам онкогенов-мишеней были отнесены наследственные ОНП и соматические мутации в кодирующих районах как этих онкогенов, так и генов для их рецепторов и эффекторов сигнальных путей. Поскольку ранее была открыта статистически значимая корреляция между эффективностью соматического мутагенеза в его горячих точках с количеством содержащих их несовершенных повторов (Rogozin et al., 1991), то степень риска канцерогенеза может быть оценена по числу повторов в анцестральном аллеле гена, в границах которых локализованы ассоциированные с раком ОНП.

Кроме того, большинство инвертированных повторов и комплементарных палиндромов локализованы в белок-кодирующих районах генов, что, как было показано, связано с неравномерностью использования кодонов и оптимизацией вторичной структуры кодируемых ими мРНК по ее устойчивости к ее деградации (Karlin et al., 1989). Это означает, что в числе молекулярных механизмов патогенеза ОНП в границах инвертированных повторов и комплементарных палиндромов в белок-кодирующих районах генов может быть рассогласование координированно экспрессирующихся генов в генной сети. В свою очередь, конвергентное возникновение прямых повторов в белок-кодирующих районах генов связано с кодированием элементов вторичной структуры (α-спиралей и β-нитей) белковых глобул. Это означает, что к числу молекулярных механизмов патогенеза, вызванного ОНП в границах прямых повторов в белок-кодирующих районах генов, могут быть отнесены нарушения пространственных структур белковых глобул.

Наконец, в работе Babenko с коллегами (1999) был обнаружен специфический повторяющийся контекст локального окружения сайта в кор-промоторах генов человека для ТАТА-связывающего белка (TBP), образующего анкерный комплекс для РНК полимеразы II (Ponomarenko et al., 2013a, b). Поэтому в числе молекулярных механизмов патогенеза ОНП в повторяющейся ДНК промоторов генов человека могут быть изменения сродства TBP к этим промоторам и в итоге – изменение экспрессии этих генов (Пономаренко и др., 2008; Савинкова и др., 2009). Таким образом, контекстные оценки сложности текста для геномных участков могут использоваться как характеристики для оценки возникновения полиморфизмов и возможных нарушений регуляции экспрессии генов.

На языке C++ был разработан комплекс программ, который позволяет эффективно определять частотный спектр олигонуклеотидов заданной фиксированной длины, выполнять сравнение частот олигонуклеотидов в различных выборках. С помощью этого программного обеспечения рассчитана контекстная сложность геномных районов, содержащих ОНП человека по базе данных dbSNP. Установленная насыщенность политрактами области точки полиморфизма свидетельствует о более высокой вероятности разрыва двойной спирали ДНК в этой точке, что подтверждает ранее показанные результаты (Siddle et al., 2011; Lenz et al., 2014). Фланги прямых мононуклеотидных повторов подвержены мутациям. Изменения идут в сторону увеличения энтропии, т. е. при возникновении полиморфизмов природа избегает длинных повторов мононуклеотидов и сложность геномного текста в таком участке вновь возрастает. Таким образом, с информационной точки зрения, происходит ослабление контекстных мононуклеотидных сигналов в геноме (Safronova et al., 2015). Пониженная сложность фланкирующих ОНП точки последовательностей наблюдалась также по доступным данным у мыши и крысы, и, следовательно, может иметь общий характер для геномов эукариот (Medvedeva et al., 2013; Lenz et al., 2014).

Интересно отметить, что повышенная частота полиморфизмов в последовательностях, фланкируемых мономерами, может быть связана с открытостью нуклеосомной упаковки в таких участках. Действительно, политракты статистически чаще встречаются в линкерных участках между соседними нуклеосомами, чем в последовательностях ДНК внутри нуклеосомной упаковки (Орлов и

др., 2006; Goh et al., 2010; Trifonov et al., 2012). Анализ контекстной сложности и насыщенности повторами в геномной ДНК содержащих ОНП участков с разной функциональной нагрузкой (регуляторные последовательности, белок-кодирующие последовательности) требует более детального исследования. Данный анализ может быть расширен на протяженные геномные последовательности (Babenko et al., 2015). В дальнейшем планируется интегрировать разработанную программу в комплекс для расчета сложности текста в виде дополнительного программного модуля, провести работу по улучшению и оптимизации пользовательского интерфейса программы и встроить ее в модули анализа геномных последовательностей, заданных генов и участков хромосомных контактов (Орлов и др., 2012; Кулакова и др., 2015; Спицина и др., 2015).

Acknowledgments

Computations were done at the Bioinformatics Shared Access Center of the Institute of the Cytology and Genetics; Siberian Supercomputer Center of SB RAS. This work was supported by the Russian Foundation for Basic Research, project 14-04-01906 (Development of programs for genomic analysis); joint project 15-54-53091 of the Russian Foundation for Basic Research and the National Natural Science Foundation of China (Analysis of polymorphisms in humans); and the Institute of Cytology and Genetics, Budgeted Project VI.61.1.2.

Conflict of interest

The authors declare no conflict of interest.

References

Babenko V.N., Kosarev P.S., Vishnevsky O.V., Levitsky V.G., Basin V.V., Frolov A.S. Investigating extended regulatory regions of genomic DNA sequences. Bioinformatics. 1999;15(7/8):644-653. DOI 10.1093/bioinformatics/15.7.644

Babenko V.N., Matvienko V.F., Safronova N.S. Implication of transposons distribution on chromatin state and genome architecture in human. J. Biomol. Struct. Dyn. 2015;33(1):10-11. DOI 10.1080/07391102.2015.1032559

Chuzhanova N.A., Krawczak M., Thomas N., Nemytikova L.A., Gusev V.D., Cooper D.N. The evolution of the vertebrate beta-globin gene promoter. Evolution. 2002;56(2):224-232.

Goh W.S., Orlov Y., Li J., Clarke N.D. Blurring of high-resolution data shows that the effect of intrinsic nucleosome occupancy on transcription factor binding is mostly regional, not local. PLoS Comput. Biol. 2010;6(1):e1000649. DOI 10.1371/journal.pcbi.1000649

Gusev V.D., Nemytikova L.A., Chuzhanova N.A. On the complexity measures of genetic sequences. Bioinformatics. 1999;15(12):994-999. DOI 10.1093/bioinformatics/15.12.994

Ignatieva E.V., Podkolodnaya O.A., Orlov Y.L., Vasiliev G.V., Kolchanov N.A. Regulatory genomics: Combined experimental and computational approaches. Genetika = Genetics. 2015;51(4):409-429.

International HapMap 3 Consortium, Altshuler D.M., Gibbs R.A., Peltonen L., Dermitzakis E., Schaffner S.F., Yu F., Peltonen L., Dermitzakis E., Bonnen P.E., Altshuler D.M., Gibbs R.A., de Bakker P.I., Deloukas P., Gabriel S.B., Gwilliam R., Hunt S., Inouye M., Jia X., Palotie A., Parkin M., Whittaker P., Yu F., Chang K., Hawes A., Lewis L.R., Ren Y., Wheeler D., Gibbs R.A., Muzny D.M., Barnes C., Darvishi K., Hurles M., Korn J.M., Kristiansson K., Lee C., McCarrol S.A., Nemesh J., Dermitzakis E., Keinan A., Montgomery S.B., Pollack S., Price A.L., Soranzo N., Bonnen P.E., Gibbs R.A., Gonzaga-Jauregui C., Keinan A., Price A.L., Yu F., Anttila V., Brodeur W., Daly M.J., Leslie S., McVean G., Moutsianas L., Nguyen H., Schaffner S.F., Zhang Q., Ghori M.J., McGinnis R., McLaren W., Pollack S., Price A.L., Schaffner S.F., Takeuchi F., Grossman S.R., Shlyakhter I., Hostetter E.B., Sabeti P.C., Adebamowo C.A., Foster M.W., Gordon D.R., Licinio J., Manca M.C., Marshall P.A., Matsuda I., Ngare D., Wang V.O., Reddy D., Rotimi C.N., Royal C.D., Sharp R.R., Zeng C., Brooks L.D., McEwen J.E. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-58. DOI 10.1038/nature09298

Karlin S., Ost F., Blaisdell B.T. Patterns in DNA and amino-acid sequences and their statistical significance. Mathematical methods for DNA sequences. Ed. M.S. Waterman. Boca Raton: CRC Press, 1989.

Kulakova E.V., Spitsina A.M., Orlova N.G., Dergilev A.I., Svichkarev A.V., Safronova N.S., Chernykh I.G., Orlov Y.L. Program analysis of genomic sequence data, obtained through technologies ChIP-seq, ChIA-PET and Hi-C. Programmnye sistemy: teoriya i prilozheniya = Program Systems: Theory and Applications. 2015;6(2):129-148.

Lenz C., Haerty W., Golding G.B. Increased substitution rates surrounding low-complexity regions within primate proteins. Genome Biol. Evol. 2014;6(3):655-665. DOI 10.1093/gbe/evu042

Medvedeva S.A., Panchin A.Y., Alexeevski A.V., Spirin S.A., Panchin Y.V. Comparative Analysis of Context-Dependent Mutagenesis Using Human and Mouse Models. BioMed Res. Intern. 2013;2013. Article ID 989410

Orlov Y.L. Analiz regulyatornykh genomnykh posledovatelnostey s pomoshchyu kompyuternykh metodov otsenok slozhnosti geneticheskikh tekstov. Diss. kand. biol. nauk. [Analysis of regulatory genome sequences using computer methods of genetic text complexity. Cand. biol. sci. diss.]. Novosibirsk, 2004.

Orlov Y.L., Bragin A.O., Medvedeva I.V., Podkolodnaia O.A., Khlebodarova T.M., Kolchanov N.A. ICGenomics: Software for analysis of symbol genomics sequences. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2012;16(4/1):732-741.

Orlov Y.L., Filippov V.P., Potapov V.N., Kolchanov N.A. Construction of stochastic context trees for genetic texts. In Silico Biology. 2002;2(3):257-262.

Orlov Y.L., Levitskii V.G., Smirnova O.G., Gunbin K.V., Demenkov P.S., Vishnevsky O.V., Levitsky V.G., Oshchepkov D.Y., Podkolodnyi N.L., Afonnikov D.A., Grosse I., Kolchanov N.A. Statistical analysis of nucleosome formation sites. Biofizika = Biophisics (Moscow). 2006;51(4):608-614.

Orlov Y.L., Potapov V.N. Complexity: an internet resource for analysis of DNA sequence complexity. Nucl. Acids. Res. 2004;32(Web Server issue):W628-633. DOI 10.1093/nar/gkh466

Orlov Y.L., Te Boekhorst R., Abnizova I.I. Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. J. Bioinform. Comput. Biol. 2006;4:523-536. DOI 10.1142/S0219720006001801

Polanovski O.L., Lebedenko E.N., Deyev S.M. ERBB oncogenes as targets for monoclonal antibodies. Biokhimia = Biochemistry (Moscow). 2012;77(3):289-311.

Ponomarenko J.V., Orlova G.V., Merkulova T.I., Gorshkova E.V., Fokin O.N., Vasiliev G.V., Frolov A.S., Ponomarenko M.P. rSNP_Guide: an integrated database-tools system for studying SNPs and site-directed mutations in transcription factor binding sites. Hum. Mutat. 2002;20(4):239-248. DOI 10.1002/humu.10116

Ponomarenko M., Mironova V., Gunbin K., Savinkova L. Hogness Box. Brenner's Encyclopedia of Genetics. 2nd edn. Eds S. Maloy, K. Hughe. San Diego: Acad. Press, Elsevier Inc. 2013a;3:491-494. DOI 10.1016/B978-0-12-374984-0.00720-8

Ponomarenko M., Savinkova L., Kolchanov N. Initiation Factors. Brenner's Encyclopedia of Genetics, 2nd ed. Eds S. Maloy, K. Hughes. San Diego: Acad. Press, Elsevier Inc. 2013b;4:83-85. DOI 10.1016/B978-0-12-374984-0.00798-1

Ponomarenko P.M., Savinkova L.K., Drachkova I.A., Lysova M.V., Arshinova T.V., Ponomarenko M.P., Kolchanov N.A. A step-by-step model of TBP/TATA box binding allows predicting human hereditary diseases by single nucleotide polymorphism. Doklady RAN = Proceedings of the Russian Academy of Sciences. 2008;419(6):828-832.

Putta P., Orlov Y.L., Podkolodnyy N.L., Mitra C.K. Relatively conserved common short sequences in transcription factor binding sites and miRNA. Vavilov Journal of Genetics and Breeding. 2011;15(4): 750-756.

Rogozin I.B., Kolchanov N.A. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. Biochim. Biophys. Acta. 1992;1171(1):11-18. DOI 10.1016/0167-4781(92)90134-L

Rogozin I.B., Pavlov Y.I., Bebenek K., Matsuda T., Kunkel T.A. Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. Nat. Immunol. 2001;2(6):530-536. DOI 10.1038/88732

Rogozin I.B., Solovyov V.V., Kolchanov N.A. Somatic hypermutagenesis in immunoglobulin genes. I. Correlation between somatic mutations and repeats. Somatic mutation properties and clonal selection. Biochim. Biophys. Acta. 1991;1089(2):175-182. DOI10.1016/0167-4781(91)90005-7

Safronova N.S., Babenko V.N., Orlov Y.L. 117 Analysis of SNP containing sites in human genome using text complexity estimates. J. Biomol. Struct. Dyn. 2015;33(1):73-74. DOI 10.1080/07391102. 2015.1032750

Savinkova L.K., Ponomarenko M.P., Ponomarenko P.M., Drachkova I.A., Lysova M.V., Arshinova T.V., Kolchanov N.A. TATA box polymorphisms in human gene promoters and associated hereditary pathologies. Biokhimiya=Biochemistry (Moscow). 2009;74(2): 149-163.

Siddle K.J., Goodship J.A., Keavney B., Santibanez-Koref M.F. Bases adjacent to mononucleotide repeats show an increased single nucleotide polymorphism frequency in the human genome. Bioinformatics. 2011;27(7):895-898. DOI 10.1093/bioinformatics/btr067

Sidore C., Busonero F., Maschio A., Porcu E., Naitza S., Zoledziewska M., Mulas A., Pistis G., Steri M., Danjou F., Kwong A., Ortega Del Vecchyo V.D., Chiang C.W., Bragg-Gresham J., Pitzalis M., Nagaraja R., Tarrier B., Brennan C., Uzzau S., Fuchsberger C., Atzeni R., Reinier F., Berutti R., Huang J., Timpson N.J., Toniolo D., Gasparini P., Malerba G., Dedoussis G., Zeggini E., Soranzo N., Jones C., Lyons R., Angius A., Kang H.M., Novembre J., Sanna S., Schlessinger D., Cucca F., Abecasis G.R. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat. Genet. 2015; 47(11):1272-1281. DOI 10.1038/ng.3368

Spitsina A.M., Orlov Y.L., Podkolodnaya N.N., Svichkarev A.V., Dergilev A.I., Chen M., Kuchin N.V., Chernykh I.G., Glinskij B.M. Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing. Programmnye sistemy: teoriya i prilozheniya=Program Systems: Theory and Applications. 2015;6:1(23):157-174.

Trifonov E.N., Volkovich Z., Frenkel Z.M. Multiple levels of meaning in DNA sequences, and one more. Ann. N.Y. Acad Sci. 2012;1267: 35-38. DOI 10.1111/j.1749-6632.2012.06589.x

Troyanskaya O.G., Arbell O., Koren Y., Landau G.M., Bolshoy A. Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. Bioinformatics. 2002;18(5):679-688. DOI 10.1093/bioinformatics/18.5.679

UK10K Consortium; Walter K., Min J.L., Huang J., Crooks L., Memari Y., McCarthy S., Perry J.R., Xu C., Futema M., Lawson D., Iotchkova V., Schiffels S., Hendricks A.E., Danecek P., Li R., Floyd J., Wain L.V., Barroso I., Humphries S.E., Hurles M.E., Zeggini E., Barrett J.C., Plagnol V., Richards J.B., Greenwood C.M., Timpson N.J., Durbin R., Soranzo N. The UK10K project identifies rare variants in health and disease. Nature. 2015;526:82-90. DOI 10.1038/nature14962

Vowles E.J., Amos W. Evidence for widespread convergent evolution around human microsatellites. PLoS Biol. 2004;2:E199. DOI 10.1371/journal.pbio.0020199

Wootton J.C., Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol. 1996;266:554-571. DOI 10.1016/S0076-6879(96)66035-2

Вавиловский журнал генетики и селекции. 2015;19(6):675-681.
Doi 10.18699/VJ15.085

Прогноз и верификация влияния SNP rs367781716 на взаимодействие TATA-связывающего белка с промотором гена *ABCA9* человека

О.В. Аркова[1], И.А. Драчкова[1], Т.В. Аршинова[1], Д.А. Рассказов[1], В.В. Суслов[1], П.М. Пономаренко[2], М.П. Пономаренко[1, 3], Н.А. Колчанов[1, 3], Л.К. Савинкова[1]

1 Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия
2 Детский госпиталь Лос-Анджелеса, Университет Южной Калифорнии, США
3 Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия

Высокопроизводительное секвенирование ДНК, в том числе в ходе проекта «1 000 геномов», открыло возможность для учета локусов и SNPs (Single Nucleotide Polymorphism – SNP)
в медицине. Это позволяет врачам улучшить лечение. Однако десяткам миллионов неаннотированных SNPs соответствует гигантское число ложноположительных (ложноотрицательных) кандидатных SNP-маркеров, отбираемых компьютерными методами для сравнения их частот у пациентов с нормой.
Это способствует недооценке значимых для медицины SNPs и затратам на проверку нейтральных SNPs. Предшествующие медицинским испытаниям опыты по проверке потенциально кандидатных SNP-маркеров могут исключить нейтральные SNPs. С помощью ранее созданного Web-сервиса SNP_ TATA_ Comparator был найден неаннотированный SNP rs367781716 – замена референсного T (норма) на минорный C в позиции –37 перед стартом транскрипта с гена ABCA9, достоверно снижающий сродство его промотора к TATA-связывающему белку (TBP). Это соответствует недостаточности продукта гена ABCA9, транспортера ATP-связывающей кассеты A9. Для экспериментальной проверки этого rs367781716 мы измерили методом гель-ретардации скорости образования (ka) и рас-пада (kd) комплексов TBP с олигонуклеотидами, идентичными аллелям «– 37C» и «–37T» гена ABCA9. Установлено, что скорость образования комплексов TBP/TATA, ka , для минорного аллеля в 2,4 раза ниже, чем для референсного. Экспериментальное значение изменения равновесной константы диссоциации (KD = kd /ka), характеризующей аффинность TBP к промотору, содержащему TATA-бокс, и его прогноз с использованием Web-сервиса SNP_ TATA_Comparator совпали с точностью до погрешности измерений и расчетов. Измерены время полураспада и свободная энергия Гиббса комплекса TBP с промотором ABCA9. Обсуждаются возможные фенотипические проявления потенциально кандидатного SNP-маркера rs367781716.

Ключевые слова: ABCA9; промотор; TATA-связывающий белок; сродство; SNP.

Received 05.08.2015
Accepted for publication 12.10.2015

e-mail: savinkl@mail.ru

Prediction and verification of the influence of the rs367781716 SNP on the interaction of TATA-binding protein with the promoter of the human *ABCA9* gene

O.V. Arkova[1], I.A. Drachkova[1], T.V. Arshinova[1], D.A. Rasskazov[1], V.V. Suslov[1], P.M. Ponomarenko[2], M.P. Ponomarenko[1, 3], N.A. Kolchanov[1, 3], L.K. Savinkova[1]

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

[2] Children's Hospital Los Angeles, University of Southern California, CA 90027, USA

[3] Novosibirsk State University, Novosibirsk, Russia

The high-throughput sequencing project "1 000 Genomes" made it possible to catalog and utilize genetic loci and single nucleotide polymorphisms (SNPs) in medicine. Analysis of SNP markers (significantly frequent differences of individual genomes of patients from the reference human genome) allows physicians to optimize treatment. On the other hand, tens of millions of unannotated SNPs correspond to a gigantic number of false positive (false negative) candidate SNP markers that are selected by computer methods for comparison of their frequency in patients with that in healthy people. This approach contributes to undervaluation of clinically relevant SNPs and to unnecessary computational expenses (on verifica-tion of neutral SNPs). Preclinical empirical verification of possible candidate SNP markers may eliminate neutral SNPs from the dataset. In the present study, we found, using the SNP_TATA_Comparator web service, the unannotated SNP rs367781716: the substitution of ancestral T (health) with minor C at position –37 before the transcription initiation site of the ABCA9 gene. This SNP significantly reduces affinity of TATA-binding protein (TBP) for this gene's promoter and corresponds to a defi-ciency (low protein level) of the ABCA9 gene product (the transporter ATP-binding cassette A9) in patients with the –37C allele. For preclinical empirical verification of rs367781716, we used an electrophoretic mobility shift assay (EMSA) to measure the rates of formation (ka) and decay (kd) of the complexes of TBP with an oligonucleotide matching either allele –37C or –37T of the ABCA9 gene. We found that the rate of formation (ka) of the TBP/TATA complex for the minor allele is 2.4-fold lower than that for the ancestral allele. We calculated the empirical value of the change in the equilibrium constant of dissociation (KD = kd /ka), which characterizes binding affinity of TBP for a promoter containing the TATA box. This empirical value matched the value predicted by SNP_TATA_Comparator within the margin of error of the measurements and calculations. We also determined the half-life and Gibbs free energy of the complex of TBP with the ABCA9 promo-ter. Possible phenotypic manifestations of the candidate SNP marker rs367781716 are discussed.

Key words: ABCA9; promoter; TATA-binding protein; affinity; SNP.

HOW TO CITE THIS ARTICLE?

Arkova o .V., Drachkova i.A., Arshinova T.V., r asskazov D.A., Suslov V.V., Ponomarenko P.M., Ponomarenko M.P., Kolchanov n .A., Savinkova l .K. Prediction and verification of the influence of the rs367781716 Sn P on the interaction of тАтА-binding protein with the promoter of the human *ABCA9* gene. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2015;19(6):675-681. Doi 10.18699/VJ15.085

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?

Аркова О.В., Драчкова И.А., Аршинова т.В., Рассказов Д.А., Суслов В.В., Пономаренко П.М, Пономаренко М.П., Колчанов Н.А., Савинкова л.К. Прогноз и верификация влияния SnP rs367781716 на взаимодействие тАтА-связывающего белка с промотором гена *ABCA9* человека. Вавиловский журнал генетики и селекции. 2015;19(6):675-681. Doi 10.18699/VJ15.085

Транспорт различных молекул через липидные мембраны является важной функцией всех живых организмов. К семейству транспортных белков АТФ-связывающей кассеты (ABC – ATP Binding Cassette) относятся белки, многие из которых переносят самые разные соединения: пептиды, холестерины и стерины, желчные кислоты, ретиноиды, ионы и сахара и т. д. (Dean, Allikmets, 2001). Мутации в генах *ABC* вызывают различные заболевания, в том числе муковисцидоз, Штаргардта дегенерацию желтого пятна, и нарушения метаболизма липопротеинов и липидов (Oram, Vaughan, 2006). Кроме того, продукты многих генов этого семейства делают вклад в генетические и онкологические заболевания человека, участвуют в формировании лекарственной устойчивости (Dean, Allikmets, 2001). Гены, кодирующие ABC-белки, рассредоточены по всему геному и обладают высокой аминокислотной гомологией среди всех эукариот. В 2002 г. было открыто подсемейство белков-транспортеров ABCA9 макрофагов человека (Piehler et al., 2002). Анализ структуры генома показал, что ген *ABCA9* состоит из 39 экзонов, занимающих область примерно в 85 тыс. п. о. на хромосоме 17q24.2 (Piehler et al., 2002). Предполагается, что наряду с другими ABC-транспортерами, ABCA9 участвует в гомеостазе липидов (Ye et al., 2008; Oram, Vaughan, 2006).

В работе большой группы авторов (Hedditch et al., 2014) получены интересные и противоречивые результаты о том, что высокие уровни экспрессии *ABCA1*, *ABCA6*, *ABCA8* и *ABCA9* в первичных опухолях статистически значимо ассоциированы со снижением выживаемости больных серозным раком яичников. При исследовании гена *ABCA10*, высоко гомологичного генам *ABCA9* и *ABCA6*, показано, что его экспрессия повышена в опухолях нескольких типов, в том числе в фолликулярной лимфоме (Baecklund et al., 2014), и влияет на исход заболеваний. Hendig с коллегами (2010) показали экспрессию 37 генов транспортеров ABC в кожных фибробластах при эластической псевдосаркоме. Семь генов – *ABCA6*, *ABCA9*, *ABCA10*, *ABCB5*, *ABCC2*, *ABCC9* и *ABCD2* – были индуцированы, в то время как экспрессия одного гена, *ABCA3*, была снижена, по сравнению с контрольной группой, по крайней мере в 2 раза.

Высокопроизводительное секвенирование геномной ДНК, в том числе в ходе проекта «1 000 геномов» (Colonna et al., 2014), развитие подхода GWAS (Genome-Wide Association Study) для локусов (Welter et al., 2014) и построение полногеномных карт для гаплотипов (International HapMap 3 Consortium et al., 2010), ассоциированных с заболеваниями, открыли возможность учета геномов пациентов в медицинской практике, так называемой постгеномной предиктивно превентивной персонализированной медицине, PPPM (Mallal et al., 2002; Trovato, 2014). Но десяткам миллионов неаннотированных SNPs соответствует гигантское число ложноположительных (ложноотрицательных) кандидатных SNP-маркеров, отбираемых компьютерными методами для очень затратных биомедицинских испытаний путем сравнения их частот на различных пациентах с нормой. Это способствует недооценке значимых для медицины SNPs, но имеющих статистически слабый сигнал (Liu, Xuan, 2015), например, из-за их крайне редкой встречаемости и/или их патогенной манифестации лишь в гаплотипах при некоторых сочетаниях с другими SNPs (Kaniwa et al., 2005) и бесполезных затрат на испытания нейтральных «молчащих SNPs» (Yoo et al., 2015). Поэтому предшествующие медицинским испытаниям опыты по проверке отобранных *in silico* кандидатных SNP-маркеров могли бы исключить из их числа по крайней мере нейтральные «молчащие» SNPs.

Ранее на примерах более 70 известных SNP-маркеров заболеваний человека (Пономаренко и др., 2009), селекци-

онно-ценных признаков животных и растений (Суслов и др., 2010), а также 146 SNPs единственного функционального ТАТА бокса в геноме ВИЧ-1 (Suslov et al., 2010) мы разработали компьютерный метод оценки достоверности аллельных различий экспрессии гена (Пономаренко и др., 2008) и подтвердили его прогнозы в наших оригинальных экспериментах с использованием гель-ретардации (Electrophoretic Mobility Shift Assay, EMSA) при равновесных (Savinkova et al., 2013) и неравновесных (Drachkova et al., 2014) условиях *in vitro*, а также *in vitro* в режиме «реального времени» с использованием (Драчкова и др., 2012) поверхностного плазмонного резонанса (Surface Plasmon Resonance, SPR) на биосенсоре ProteOn™ XPR36 (Bio-Rad Lab., Inc., США) и переноса флюоресцентной резонансной энергии (Fluorescence Resonance Energy Transfer, FRET) (Аркова и др., 2014) и метода остановленной струи (Stopped-flow) на спектрометре SX20 (Applied Photophysics, Великобритания). Дополнительно, мы подтвердили прогнозы нашего метода (Пономаренко и др., 2008) на данных независимых экспериментов на генах-паралогах семейства ARF (Auxin Response Factors) арабидопсиса и риса (Миронова и др., 2010), транскриптома арабидопсиса через один час после обработки растения 1 μM ауксином (Ponomarenko, Ponomarenko, 2015), транскриптома мозга человека (Пономаренко и др., 2014), а также 68 других опытов разных авторов (Пономаренко и др., 2010). Наконец, в завершение исчерпывающих экспериментальных проверок метода (Пономаренко и др., 2008) мы создали общедоступный Web-сервис SNP TATA Comparator (Рассказов и др., 2013), http://beehive.bionet.nsc.ru/cgi-bin/mgs/tatascan/start.pl, для самостоятельного применения пользователями сети Интернет.

В этой работе мы экспериментально охарактеризовали неаннотированный (не ассоциированный с каким-либо нарушением здоровья человека) SNP rs367781716, –37T→C, промотора к альтернативному старту транскрипции в позиции chr17:66985252 (минус-цепь) гена *ABCA9*, который был детектирован экспериментально (Piehler et al., 2002), но не был до сих пор документирован в базе данных GENECODE v.19 (Harrow et al., 2012). С помощью Web-сервиса (Рассказов и др., 2013) мы предсказали для этого SNP rs367781716 достоверное снижение сродства TBP к промотору. Данные независимых опытов *in vitro* и *in vivo* (Pugh, 2000; Stewart, Stargell, 2001; Mogno et al., 2010) указывают на то, что такому снижению сродства TBP/промотор соответствует недостаточность у пациента-носителя аллеля –37C продукта ABCA9, транспортера ATP-связывающей кассеты A9. В свою очередь, этот *in silico* прогноз риска недостаточности экспрессии транспортера, предположительно участвующего в липидном гомеостазе, в случае аллеля –37C позволяет нам предположить rs367781716 в качестве кандидатного SNP-маркера для нарушений липидного обмена. В качестве экспериментальной проверки этого потенциально кандидатного SNP-маркера мы измерили методом гель-ретардации (EMSA) скорости образования (k_a) и распада (k_d) комплексов TBP с олигонуклеотидами, идентичными аллелям –37T (норма) и –37C (недостаток синтеза белка гена *ABCA9*). Установлено, что константа скорости образования комплексов, k_a, в 2,4 раза ниже нормы. Прогноз с помощью Web-сервиса SNP_TATA_Comparator (Рассказов и др., 2013) относительного изменения кажущейся константы диссоциации ($K_D = k_d/k_a$) и экспериментальное значение этой величины совпали в пределах точности эксперимента и расчетов. В работе также измерены время полураспада и свободная энергия Гиббса комплекса TBP с промотором *ABCA9*, в контексте которых обсуждаются вероятностные фенотипические проявления SNP-маркера rs367781716.

Материалы и методы

Получение рекомбинантного TBP

Рекомбинантный TBP человека (hTBP) экспрессировали в клетках *Escherichia coli* BL21(DE3) с плазмиды pAR3038-hTBP, любезно предоставленной профессором B. Pugh (Center for Gene Regulation, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, США), выделяли и очищали его, как описано статье Драчковой с коллегами (2010).

Последовательности ДНК промотора гена *ABCA9* человека

Координаты обоих промоторов гена *ABCA9* в референсном геноме человека (hg19) были взяты из Eukaryotic Promoter Database, EPD (Dreos et al., 2015). Между позициями –70 и –20 этих промоторов расположен район доказанных сайтов TBP-связывания (Ponomarenko et al., 2013a, b), с помощью Web-сервиса «UCSC Genome Browser» (Dreszer et al., 2012) были взяты все 9 неаннотированных SNPs из базы данных dbSNP v. 142 (NCBI Resource Coordinators), как это показано на рис. 1, *а*. С помощью Web-сервиса (Рассказов и др., 2013) было предсказано значимое ($\alpha < 10^{-7}$) изменение сродства TBP к промоторам гена *ABCA9* для SNP rs367781716 (рис. 1, *б*), для всех 8 остальных SNPs – незначимое, вследствие чего мы исключили их из дальнейшего анализа (данные не показаны). Последовательности ДНК для аллелей, –37T (норма) и –37C для SNP rs367781716 промотора к альтернативному старту транскрипции chr17:66985252 (минус-цепь) с гена *ABCA9* (Piehler et al., 2002), взяты из базы данных Ensembl, как это показано на рис. 1, *б*.

Получение меченых ^{32}P олигодезоксирибонуклеотидов

В работе использовали олигодезоксирибонуклеотиды (ОДН) длиной 26 п. о., идентичные аллелям –37T (норма) и –37C SNP rs367781716, которые были синтезированы и дополнительно очищены электрофорезом в ПААГ (BIOSYN, Новосибирск). Получение меченых и немеченых двуцепочечных ОДН подробно описано в статье Драчковой с коллегами (2010).

Определение константы скорости образования и диссоциации комплексов TBP/ДНК

Для комплексов рекомбинантного TBP человека и ОДН измеряли константу скорости образования (k_a) и распада (k_d) с помощью четырех концентраций ОДН при 25 °C в буфере, содержащем 20 mM HEPES-KOH (pH 7,6), 5 mM $MgCl_2$, 70 mM KCl, 1 mM DTT, 100 μg/mL BSA, 0,01 %

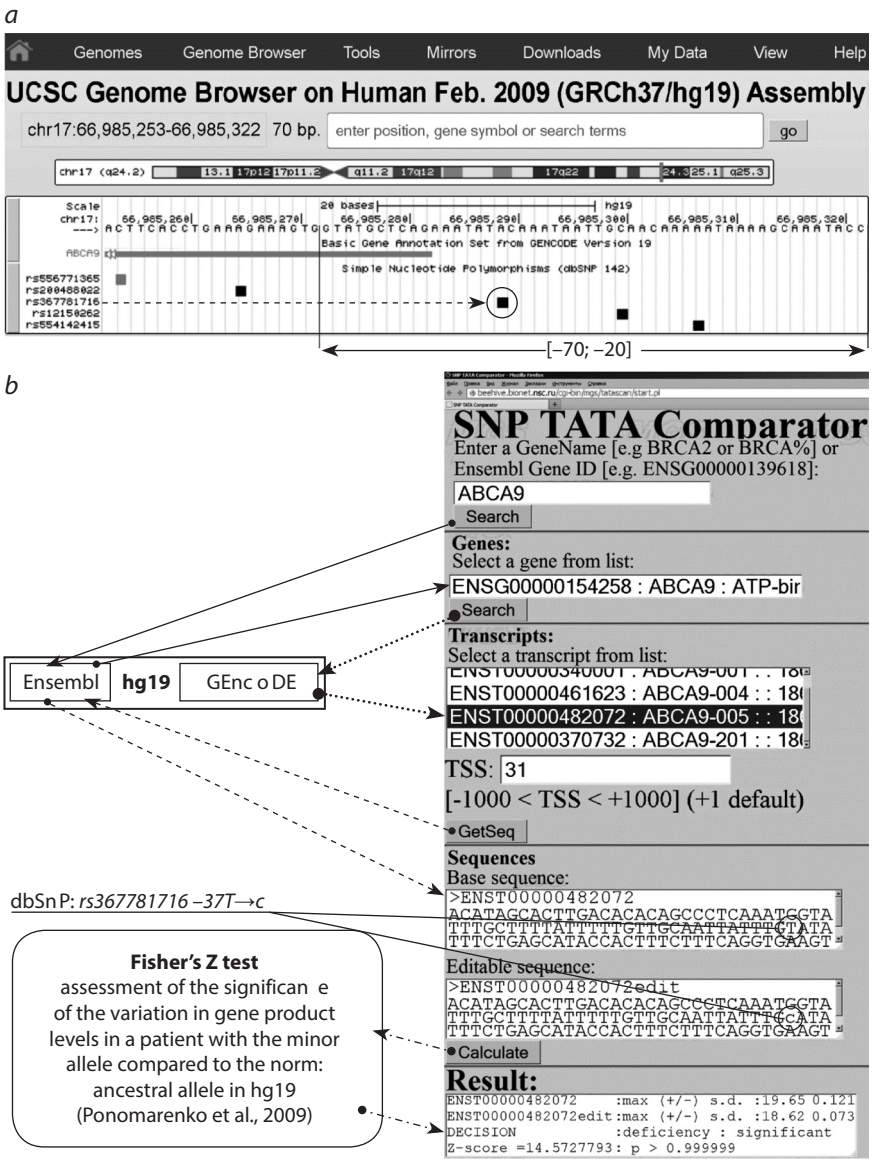


Fig. 1. (*a*) The location of the unannotated SnP rs367781716 (substitution –37T→c) in the promoter of the human *ABCA9* gene and (*b*) its computational analysis by means of the web service SnP_TATA_comparator (Rasskazov et al., 2013).

NP-40, и 5 % глицерин, TBP (обычно 0,3 nM). Реакционную смесь с TBP и ОДН хранили во льду. Каждый эксперимент по определению константы ассоциации состоял из 32 реакций связывания (8 временных точек × 4 концентрации ОДН). Все 4 реакции запускали одновременно добавлением TBP и помещали в термостат при 25 °C. Все реакционные смеси одновременно наносили на гель.

Комплексы TBP/ОДН отделяли от свободных ОДН методом EMSA в 5 % ПААГ на трис-глициновом буфере (pH 8,3) при температуре 10 °C и напряженности поля 25 В/см в течение 40 мин. Гели высушивали и экспонировали с экраном Imaging Screen-K (Kodak) для фосфоимиджера Molecular Imager PharosFX Plus (Bio-Rad). Затем экран сканировали на фосфоимиджере и с помощью программы Quantity One – 4.5.0 (Bio-Rad) количественно анализировали радиоавтографы (рис. 2, *а*).

Константы скоростей, k_a и k_d, определяли с помощью программы GraphPad Prism 5 software (Equation: Association kinetics (two ligand concentrations) на основе зависимости изотерм связывания TBP/ОДН от концентрации ОДН при 0,3 nM TBP (рис. 2, *б*). Из полученных экспериментальных величин k_a и k_d оценили величину кажущейся равновесной константы диссоциации ($K_D = k_d/k_a$), характеризующую аффинность TBP/ОДН, а также время полураспада ($t_{1/2} = \ln 2/k_d$) и изменение свободной энергии Гиббса ($\Delta G^0 = -RT \ln K_A$, где R – универсальная газовая константа, T – абсолютная температура, $K_A = k_a/k_d$).

результаты и обсуждение

Полученный с помощью Web-сервиса (Рассказов и др., 2013) прогноз относительного изменения в 2,4 раза кажущейся равновесной константы диссоциации K_D комплекса TBP с минорным аллелем –37C промотора гена *ABCA9* приведен в таблице.

Как можно видеть, экспериментальная оценка этой величины совпала с прогнозом *in silico* в пределах точности используемых расчетов и измерений: Δ, характеризующая изменение значения $-\ln K_D$ комплексов TBP с референсным и минорным аллелями, равна 0,88 по прогнозу и в результате экспериментальной верификации. Это означает, что с использованием двух тест-систем, *in silico* и *in vitro*, мы установили, что неаннотированный ранее SNP rs367781716 достоверно нарушает связывание TBP с промотором гена *ABCA9* человека на одном из первых этапов инициации транскрипции этого гена (Ponomarenko et al., 2013a, b).

Согласно результатам независимых опытов *in vitro* и *in vivo* (Pugh, 2000; Stewart, Stargell, 2001; Mogno et al., 2010), такому снижению сродства TBP к промотору гена *ABCA9* у пациента, «маркированного» аллелем –37C (SNP rs367781716), соответствует недостаточность кодируемого этим геном белка-транспортера ATP-связывающей кассеты A9, участвующего в гомеостазе липидов. Поэтому мы предлагаем SNP rs367781716 в качестве потенциального кандидатного SNP-маркера для нарушений липидного обмена. Это является экспериментально-компьютерным прогнозом данной работы, который может быть верифицирован по биомедицинским стандартам и протоколам.

Экспериментальные оценки изменения кинетических характеристик связывания TBP с промотором рефе-

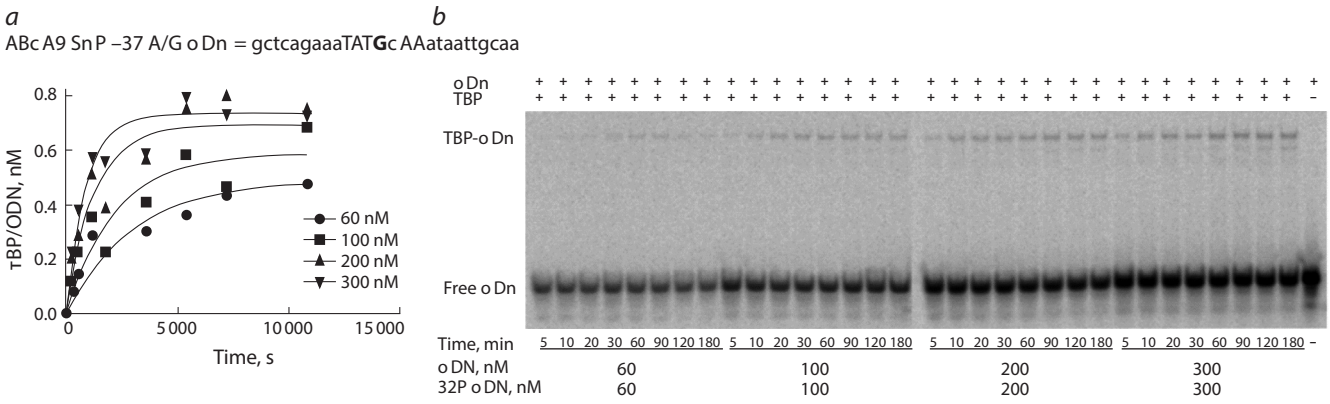


Fig. 2. Kinetic analysis of the interaction of TATA-binding protein (TBP) with a TATA-containing oligonucleotide that matches the minor allele c in the TATA box of the *ABCA9* gene.

(a) Dependence of isotherms of the association reaction on the concentration of the oligonucleotide. isotherms of the association of TBP with the TATA box, as well as k_a and k_d, were calculated from the electrophoretic image with GraphPad Prism 5 (http//graphpad-prism.software.informer.com/5.01). (*б*) Electrophoretic images from which the association isotherms were calculated. c oncentration of TBP was 0.3 nM in all experiments; concentrations of the oligonucleotide are indicated in the figu e.

Kinetic parameters of the interaction of TATA-binding protein with an oligonucleotide matching either the reference allele or a minor allele of the TATA box in the promoter of the human *ABCA9* gene

Allele of *ABCA9*	Prediction		Experimental data						
	$-\ln K_D$, ln-unit	Δ, ln-unit	K_D, nM	$-\ln K_D$ ln-unit	Δ, ln-unit	k_a, $\times 10^3 M^{-1}c^{-1}$	k_d, $\times 10^{-4}c^{-1}$	$t_{1/2}$, min	–ΔG, kcal/mol
–37T (wT, hg19)	19.66	–0.88	70 ± 20	16.47	–0.88	7 ± 1	4.6 ± 0.8	25.1	9.8
T–37C	18.78		170 ± 30	15.59		2.9 ± 0.4	4.9 ± 0.7	23.6	9.2

hg19 is the reference human genome. The TATA-containing sequence of the promoter of *ABCA9* (26 bp) that we use here is 5′-ttgcaattatTTG**T**ATAtttctgagc-3′. it corresponds to the current version of the reference human genome. T-37c is a substitution in the TATA box (5′-ttgcaattatTTG**C**ATAtttctgagc-3′). The kinetic parameters shown in the table were calculated from the electrophoretic images and curves shown in Fig. 2.

ренсного и минорного аллелей гена *ABCA9* представлены в таблице, в которой показано, что скорость образования, k_a, комплекса TBP с минорным вариантом TATA-бокса снижается в 2,4 раза, тогда как скорость его распада, k_d, определяющая время полужизни ($t_{1/2}$) этого комплекса, изменяется незначительно: с 25 мин в норме (референсный аллель) до 23,6 мин для минорного аллеля. В свою очередь, свободная энергия Гиббса (ΔG), которая характеризует спонтанное протекание реакции взаимодействия TBP с промотором гена *ABCA9* человека, уменьшается для минорного аллеля и свидетельствует о дестабилизации комплекса TBP с этим промотором вследствие ухудшения последовательности TATA-бокса и снижения сродства между ними.

Независимые опыты многих авторов (Пономаренко и др., 2010) показали, что, несмотря на множество регуляторных событий *in vivo* на разных этапах экспрессии генов (транскрипция, процессинг, трансляция, тканеспецифическая регуляция и др.), оценки сродства, скоростей образования и распада комплексов TBP с промоторами достоверно коррелируют с тяжестью моногенных (например, β-талассемия) и сложных комплексных (например, инфаркт миокарда, рак легкого, состояние иммунодефицита) заболеваний, для которых известны SNP-маркеры доказанных сайтов TBP-связывания в геноме человека. Исходя из этого, можно предположить, что уменьшение сродства и скорости образования комплексов TBP с минорным вариантом –37C промотора гена *ABCA9* с большой степенью вероятности может иметь фенотипическое проявление. Известно, что ген *ABCA9* экспрессируется во многих тканях и органах (Piehler et al., 2002), где его белковые продукты выполняют характерную для класса генов *ABCA* роль белков транспортеров, регулируемых холестерином. Снижение экспрессии этого гена вследствие уменьшения сродства TBP/TATA может привести к повышению риска нарушений липидного гомеостаза, затрагивающего многие органы и ткани в организме человека. Поскольку максимальная экспрессия гена *ABCA9* у человека установлена в сердце, головном мозге и эмбриональных тканях (Piehler et al., 2002), то предлагаемый кандидатный SNP-маркер rs367781716 может быть наиболее вероятно связан с их патологиями. Вызванное SNP –37T→C уменьшение сродства TBP к промотору гена *ABCA9* и соответствующая недостаточность по кодируемому этим геном белку-транспортеру ATP-связывающей кассеты A9 могут снизить отток холестерина через мембраны макрофагов этих органов и его перенос на APOA1 (Wang et al., 2007). Это, в свою очередь, может привести к накапливанию липопротеинов низкой плотности в стенке сосудов, которое сопровождается отложением холестерина и липопротеинов в интиме сосудов, и их закупорке. Эти процессы могут повысить риск возникновения

атеросклероза, болезни Альцгеймера и широкого спектра сердечно-сосудистых заболеваний. Снижение экспрессии и уровня белка ABCA9 может также повысить вероятность развития общих воспалительных и метаболических расстройств. Как отмечалось выше, ранее обнаружена прямая зависимость между уровнем экспрессии *ABCA9* (наряду с такими генами, как *ABCA6*, *ABCA10*, *ABCB5*, *ABCC2*, *ABCC9* и *ABCD2*) и развитием эластической псевдосаркомы (Hendig et al., 2010) и фолликулярной лимфомы (Baecklund et al., 2014). Суперэкспрессия *ABCA9* (а также *ABCA1*, *ABCA6* и *ABCA8*) значимо ассоциирована со снижением выживаемости больных серозным раком яичников (Hedditch et al., 2014). Многие авторы отмечают противоречивость полученных результатов сложившемуся представлению о роли *ABCA9* и гомологичных ему белков-транспортеров ABCA6, ABCA8, ABCA10 и ABCA1 и других, объясняемую индивидуальной вариабельностью их экспрессии, хотя это похоже на закономерность, которой необходимо найти объяснение.

Таким образом, разработанный Web-сервис SNP_TATA_Comparator позволяет выявлять SNPs с потенциалом функциональной значимости, прогнозировать их влияние на аффинность взаимодействия TBP с TATA-элементами кор-промоторов генов. Экспериментальная проверка показывает хорошую корреляцию компьютерного прогноза с результатами *in vitro*, что дает основания для проведения экспериментов по изучению влияния исследуемых SNPs на экспрессию репортерного гена и уточнению их возможного влияния на здоровье человека.

Acknowledgments

AT acknowledges the support of data collection and computerized analysis by basic project VI.58.1.2. AO, DI, SL, and PM acknowledge the support of the experiments by the Russian Foundation for Basic Research, project 14-04-00485. RD and KN acknowledge the support of the Web service design by the Russian Science Foundation, project 14-24-00123.

Conflict of interest

The authors declare no conflict of interest.

References

Arkova O.V., Kuznetsov N.A., Fedorova O.S., Kolchanov N.A., Savinkova L.K. Real-time interaction between TBP and the TATA box of the human triosephosphate isomerase gene promoter in the norm and pathology. Acta Naturae. 2014;6(2):36-40.

Baecklund F., Foo J.N., Bracci P., Darabi H., Karlsson R., Hjalgrim H., Rosenquist R., Adami H.-O., Glimelius B., Melbye M., Conde L., Liu J., Humphreys K., Skibola C.F., Smedby K.E. A comprehensive evaluation of the role of genetic variation in follicular lymphoma survival. BMC Med. Genet. 2014;15:113. DOI 10.1186/s12881-014-0113-6

Calpe-Berdiel L., Zhao Y., de Graauw M., Ye D., van Santbrink P.J., Mommaas A.M., Foks A., Bot M., Meurs I., Kuiper J., Mack J.T., Van Eck M., Tew K.D., van Berkel T.J. Macrophage ABCA2 deletion modulates intracellular cholesterol deposition, affects macrophage apoptosis, and decreases early atherosclerosis in LDL receptor knockout mice. Atherosclerosis. 2012;223(2):332-341. DOI 10.1016/j.atherosclerosis.2012.05.039

Colonna V., Ayub Q., Chen Y., Pagani L., Luisi P., Pybus M., Garrison E., Xue Y., Tyler-Smith C.; The 1000 Genomes Project Consortium, Abecasis G.R., Auton A., Brooks L.D., DePristo M.A., Durbin R.M., Handsaker R.E., Kang H.M., Marth G.T., McVean G.A. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome Biol. 2014;15(6):R88. DOI 10.1186/gb-2014-15-6-r88

Dean M., Allikmets R. Complete characterization of the human ABC gene family. J. Bioenerg. Biomembr. 2001;33(6):475-479. DOI 10.1023/A:1012823120935

Dean M., Hamon Y., Chimini G. The human ATP-binding cassette (ABC) transporter superfamily. J. Lipid Res. 2001;42(7):1007-1017.

Drachkova I.A., Arshinova T.V., Ponomarenko P.M., Merkulova T.I., Kolchanov N.A., Savinkova L.K. Effect of TATA box polymorphisms in the human β-globin gene promoter associated with β-thalassemia on the interaction of the TATA-binding protein. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2010;14(4):698-705.

Drachkova I., Savinkova L., Arshinova T., Ponomarenko M., Peltek S., Kolchanov N. The mechanism by which TATA-box polymorphisms associated with human hereditary diseases influence interactions with the TATA-binding protein. Hum. Mutat. 2014;35(5):601-608. DOI 10.1002/humu.22535

Drachkova I.A., Shekhovtsov S.V., Peltek S.E., Ponomarenko P.M., Arshinova T.V., Ponomarenko M.P., Merkulova T.I., Savinkova L.K., Kolchanov N.A. Surface plasmon resonance study of the interaction between the human TATA-box binding protein and the TATA element of the NOS2A gene promoter. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2012;16(2):391-396.

Dreos R., Ambrosini G., Perier R.C., Bucher P. The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. Nucl. Acids Res. 2015;43(Database issue):D92-D96. DOI 10.1093/nar/gku1111

Dreszer T.R., Karolchik D., Zweig A.S., Hinrichs A.S., Raney B.J., Kuhn R.M., Meyer L.R., Wong M., Sloan C.A., Rosenbloom K.R., Roe G., Rhead B., Pohl A., Malladi V.S., Li C.H., Learned K., Kirkup V., Hsu F., Harte R.A., Guruvadoo L., Goldman M., Giardine B.M., Fujita P.A., Diekhans M., Cline M.S., Clawson H., Barber G.P., Haussler D., James Kent W. The UCSC Genome Browser database: extensions and updates 2011. Nucl. Acids Res. 2012:40 (Database issue):D918-D923. DOI 10.1093/nar/gkr1055

Harrow J., Frankish A., Gonzalez J.M., Tapanari E., Diekhans M., Kokocinski F., Aken B.L., BarrellD., Zadissa A., Searle S., Barnes I., Bignell A., Boychenko V., Hunt T., Kay M., Mukherjee G., Rajan J., Despacio-Reyes G., Saunders G., Steward C., Harte R., Lin M., Howald C., Tanzer A., Derrien T., Chrast J., Walters N., Balasubramanian S., Pei B., Tress M., Rodriguez J.M., Ezkurdia I., van Baren J., Brent M., Haussler D., Kellis M., Valencia A., Reymond A., Gerstein M., Guigó R., Hubbard T.J. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760-1774. DOI 10.1101/gr.135350.111

Hedditch E.L., Gao B., Russell A.J., Lu Y., Emmanuel C. , Beesley J., Johnatty S.E., Chen X., Harnett P., George J.; Australian Ovarian Cancer Study Group, Williams R.T., Flemming C., Lambrechts D., Despierre E., Lambrechts S., Vergote I., Karlan B., Lester J., Orsulic S., Walsh C., Fasching P., Beckmann M.W., Ekici A.B., Hein A., Matsuo K., Hosono S., Nakanishi T., Yatabe Y., Pejovic T., Bean Y., Heitz F., Harter P., du Bois A., Schwaab I., Hogdall E., Kjaer S.K., Jensen A., Hogdall C., Lundvall L., Engelholm S.A., Brown B., Flanagan J., Metcalf M.D., Siddiqui N., Sellers T., Fridley B., Cunningham J., Schildkraut J., Iversen E., Weber R.P., Berchuck A., Goode E., Bowtell D.D., Chenevix-Trench G., deFazio A., Norris M.D., MacGregor S., Haber M., Henderson M.J. ABCA transporter gene expression and poor outcome in epithelial ovarian cancer. J. Natl. Cancer Inst. 2014;106(7):dju149. DOI 10.1093/jnci/dju149

Hendig D., Langmann T., Kocken S., Zarbock R., Szliska C., Schmitz G., Kleesiek K., Götting C. Gene expression profiling of ABC transporters in dermal fibroblasts of pseudoxanthoma elasticum patients identifies new candidates involved in PXE pathogenesis. Lab. Invest. 2008;88(12):1303-1315. DOI 10.1038/labinvest.2008.96

International HapMap 3 Consortium, Altshuler D.M., Gibbs R.A., Peltonen L., Dermitzakis E., Schaffner S.F, Yu F., Peltonen L., Dermitzakis E., Bonnen P.E., Altshuler D.M., Gibbs R.A., de Bakker P.I., Deloukas P., Gabriel S.B., Gwilliam R., Hunt S., Inouye M., Jia X., Palotie A., Parkin M., Whittaker P., Yu F., Chang K., Hawes A., Lewis L.R., Ren Y., Wheeler D., Gibbs R.A., Muzny D.M., Barnes C., Darvishi K., Hurles M., Korn J.M., Kristiansson K., Lee C., McCarrol S.A., Nemesh J., Dermitzakis E., Keinan A., Montgomery S.B., Pollack S., Price A.L., Soranzo N., Bonnen P.E., Gibbs R.A., Gonzaga-Jauregui C., Keinan A., Price A.L., Yu F., Anttila V., Brodeur W., Daly M.J., Leslie S., McVean G., Moutsianas L., Nguyen H., Schaffner S.F., Zhang Q., Ghori M.J., McGinnis R., McLaren W., Pollack S., Price A.L., Schaffner S.F., Takeuchi F., Grossman S.R., Shlyakhter I., Hostetter E.B., Sabeti P.C., Adebamowo C.A., Foster M.W., Gordon D.R., Licinio J., Manca M.C., Marshall P.A., Matsuda I., Ngare D., Wang V.O., Reddy D., Rotimi C.N., Royal C.D., Sharp R.R., Zeng C., Brooks L.D., McEwen J.E. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-58. DOI 10.1038/nature09298

Kaniwa N., Kurose K., Jinno H., Tanaka-Kagawa T., Saito Y., Saeki M., Sawada J., Tohkin M., Hasegawa R. Racial variability in haplotype frequencies of UGT1A1 and glucuronidation activity of a novel single nucleotide polymorphism 686C> T (P229L) found in an African-American. Drug Metab. Dispos. 2005;33(3):458-465. DOI 10.1124/dmd.104.001800

Liu C., Xuan Z. Prioritization of cancer-related genomic variants by SNP association network. Cancer Inform. 2015;14(Suppl. 2):57-70. DOI 10.4137/CIN.S17288

Mallal S., Nolan D., Witt C., Masel G., Martin A.M., Moore C., Sayer D., Castley A., Mamotte C., Maxwell D., James I., Christiansen F.T. Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. Lancet. 2002;359(9308):727-732. DOI 10.1016/S0140-6736(02)07873-X

Mironova V.V., Omelyanchuk N.A., Ponomarenko P.M., Ponomarenko M.P., Kolchanov N.A. Specific/nonspecific binding of TBP to promoter DNA of the auxin response factor genes in plants correlated with ARFs function on gene transcription (activator/repressor). Doklady RAN=Proceedings of the Russian Academy of Sciences. 2010;433(4):549-554.

Mogno I., Vallania F., Mitra R.D., Cohen B. TATA is a modular component of synthetic promoters. Genome Res. 2010;20(10):1391-1397. DOI 10.1101/gr.106732.110

Oram J., Vaughan A. ATP-Binding cassette cholesterol transporters and cardiovascular disease. Circ. Res. 2006;99(10):1031-1043. DOI 0.1161/01.res.0000250171.54048.5c

Piehler A., Kaminski W.E., Wenzel J., Langmann T., Schmitz G. Molecular structure of a novel cholesterol-responsive A subclass ABC transporter, ABCA9. Biochem. Biophys. Res. Commun. 2002;295(2):408-416. DOI 10.1016/S0006-291X(02)00659-9

Ponomarenko M., Mironova V., Gunbin K., Savinkova L. Hogness Box. Brenner's Encyclopedia of Genetics. Eds S. Maloy, K. Hughes. San Diego: Acad. Press, Elsevier Inc, 2013a;3:491-494. DOI 10.1016/B978-0-12-374984-0.00720-8

Ponomarenko M., Savinkova L., Kolchanov N. Initiation Factors. Brenner's Encyclopedia of Genetics. Eds S. Maloy, K. Hughes. San Diego: Acad. Press, Elsevier Inc. 2013b;4:83-85. DOI 10.1016/B978-0-12-374984-0.00798-1

Ponomarenko M.P., Suslov V.V., Gunbin K.V., Ponomarenko M.P., Vishnevsky O.V., Kolchanov N.A., Identification of the relationship between variability of expression of signaling pathway genes in the human brain and affinity of TATA-binding protein to their promoters. Vavilovskii Zhurnal Genetiki i Selektsii=Vavilov Journal of Genetics and Breeding. 2014;18(4/3):1219-1230.

Ponomarenko P.M., Ponomarenko M.P. Sequence-based prediction of transcription upregulation by auxin in plants. J. Bioinform. Comput. Biol. 2015;13(1). Art.1540009. DOI 10.1142/S0219720015400090

Ponomarenko P.M., Ponomarenko M.P., Drachkova I.A., Lysova M.V., Arshinova T.V., Savinkova L.K., Kolchanov N.A. Prediction of the affinity of the TATA-binding protein to TATA boxes with single nucleotide polymorphisms. Molekulyarnaya biologiya=Molecular Biology (Moscow). 2009;43(3):512-520.

Ponomarenko P.M., Savinkova L.K., Drachkova I.A., Lysova M.V., Arshinova T.V., Ponomarenko M.P., Kolchanov N.A. A step-by-step model of TBP/TATA box binding allows predicting human hereditary diseases by single nucleotide polymorphism. Doklady RAN=Proceedings of the Russian Academy of Sciences. 2008;419(6):828-832.

Ponomarenko P.M., Suslov V.V., Savinkova L.K., Ponomarenko M.P., Kolchanov N.A. A precise equation of equilibrium of four steps of TBP binding with the TATA box for prognosis of phenotypic manifestation of mutations. Biofizika=Biophysics (Moscow). 2010;55(3);400-414.

Pugh B.F. Control of gene expression through regulation of the TATA-binding protein. Gene. 2000;255(1):1-14. DOI 10.1016/S0378-1119(00)00288-2

Rasskazov D.A., Gunbin K.V., Ponomarenko P.M., Vishnevsky O.V., Ponomarenko M.P., Afonnikov D.A. SNP_TATA_Comparator: Web-service for comparison of SNPs within gene promoters associated with human diseases using the equilibrium equation of the TBP/TATA complex. Vavilovskii Zhurnal Genetiki i Selektsii=Vavilov Journal of Genetics and Breeding. 2013;17(4/1):599-606.

Savinkova L.K., Drachkova I.A., Arshinova T.V., Ponomarenko P.M., Ponomarenko M.P., Kolchanov N.A. An experimental verification of the predicted effects of promoter TATA-box polymorphisms associated with human diseases on interactions between the TATA boxes and TATA-binding protein. PLoS One. 2013;8(2). Art.e54626. DOI 10.1371/journal.pone.0054626

Stewart J.J., Stargell L.A. The stability of the TFIIA-TBP-DNA complex is dependent on the sequence of the TATAAA element. J. Biol. Chem. 2001;276(32):30078-30084. DOI 10.1074/jbc.M105276200

Suslov V.V., Ponomarenko P.M., Efimov V.M., Savinkova L.K., Ponomarenko M.P., Kolchanov N.A. SNPs in the HIV-1 TATA box and the AIDS pandemic. J. Bioinform. Comput. Biol. 2010;8(3):607-625. DOI 10.1142/S0219720010004677

Suslov V.V., Ponomarenko P.M., Ponomarenko M.P., Drachkova I.A., Arshinova T.V., Savinkova L.K., Kolchanov N.A. TATA box polymorphisms in genes of commercial and laboratory animals and plants associated with selectively valuable traits. Genetika=Genetics (Moskow). 2010;46(4):448-457.

Trovato G.M. Sustainable medical research by effective and comprehensive medical skills: overcoming the frontiers by predictive, preventive and personalized medicine. EPMA J. 2014;5(1):14. DOI 10.1186/1878-5085-5-14

Wang X., Collins H.L., Ranalletta M., Fuki I.V., Billheimer J.T., Rothblat G.H., Tall A.R., Rader D.J. Macrophage ABCA1 and ABCG1, but not SR-BI, promote macrophage reverse cholesterol transport *in vivo*. J. Clin. Invest. 2007;117(8):2216-2224.

Welter D., MacArthur J., Morales J., Burdett T., Hall P., Junkins H., Klemm A., Flicek P., Manolio T., Hindorff L., Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucl. Acids Res. 2014;42(Database issue):D1001-D1006. DOI 10.1093/nar/gkt1229

Ye D., Hoekstra M., Out R., Meurs I., Kruijt J.K., Hildebrand R.B., Van Berkel T.J.C., Van Eck M. Hepatic cell-specific ATP-binding cassette (ABC) transporter profiling identifies putative novel candidates for lipid homeostasis in mice. Atherosclerosis. 2008;196(2):650-658.

Yoo S.S., Jin C., Jung D., Choi Y., Choi J.E., Lee W.K., Lee S.Y., Lee J., Cha S.I., Kim C.H., Seok Y., Lee E., Park J.Y. Putative functional variants of XRCC1 identified by RegulomeDB were not associated with lung cancer risk in a Korean population. Cancer Genet. 2015;208(1/2):19-24. DOI 10.1016/j.cancergen.2014.11.004

Вавиловский журнал генетики и селекции. 2015;19(6):682-690
Doi 10.18699/VJ15.089

Влияние однонуклеотидных полиморфных замен в районах позиционирования РНК-полимеразы II на сродство к ним TBP в генах циркадных часов человека

О.А. Подколодная[1], Д.А. Рассказов[1], Н.Л. Подколодный[1, 2, 3], Н.Н. Подколодная[1, 3,]
В.В. Суслов[1], Л.К. Савинкова[1], П.М. Пономаренко[4], М.П. Пономаренко[1, 3]

1 Федеральное государственное бюджетное научное учреждение
«Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия
2 Федеральное государственное бюджетное учреждение науки Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия
3 Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия 4 Детский госпиталь Лос-Анджелеса, Университет Южной Калифорнии, США

Генетическая вариабельность в системе циркадных часов проявляется в фенотипической изменчивости физиологических функций и поведения, а также в нарушениях функционирования не только самих часов, но и других систем, приводящих к развитию серьезных патологических состояний. В данной работе был проведен анализ влияния однонуклеотидных полиморфных замен (ОНП), локализованных в области [–70, –20] от старта транскрипции, на сродство TATA-связывающего белка (TATA-binding protein, TBP) к промотору в двух группах генов, являющихся компонентами системы циркадных часов человека. Первую группу составляют гены ядра циркадного осциллятора (11 генов), вторую – гены ближайшего регуляторного окружения циркадного осциллятора (21 ген), для сравнения взята группа функционально отличающихся генов (31 ген). Для оценки *in silico* изменения константы диссоциации и, следовательно, сродства TBP/промотор при мутациях был использован web-сервис SnP_TATA_ comparator. В результате показано, что в первой группе генов количество ОНП-маркеров снижения сродства TBP/промотор значимо ниже количества ОНП-маркеров увеличения сродства ($\alpha < 10^{-3}$), в то время как в группе сравнения наблюдается противоположная картина: ОНП-маркеров уменьшения сродства TBP/промотор значимо больше, чем ОНП-маркеров увеличения сродства ($\alpha < 10^{-6}$). Наблюдаемая особенность может быть специфической характеристикой генов циркадного осциллятора, влияющей на его устойчивость при генетической вариабельности анализируемой области промоторов. Полученные предсказания могут играть важную роль для отбора кандидатных ОНП- маркеров различных патологий, связанных с нарушением системы циркадных часов, для дальнейшей проверки их в экспериментальных исследованиях, а также при верификации математических моделей циркадного осциллятора.

Ключевые слова: циркадный ритм; промотор; ОНП; TATA-связывающий белок (TBP); сродство TBP/промотор.

Received 29.09.2015
Accepted for publication 06.11.2015

The effects of SNPs in the regions of positioning RNA polymerase II on the TBP/promoter affinity in the genes of human circadian clock

O.A. Podkolodnaya[1], D.A. Rasskazov[1], N.L. Podkolodnyy[1, 2, 3],
N.N. Podkolodnaya[1, 3], V.V. Suslov1, L.K. Savinkova[1],
P.M. Ponomarenko[4], M.P. Ponomarenko[1, 3]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
[2] Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
[4] Children's Hospital Los Angeles, University of Southern California, USA

Genetic variability in the genes of circadian clock is manifested as the phenotypic variability of physiological functions and behavior as well as disorders of the func-tion of not only the clock but also other systems, leading to the development of a pathologies. we ana-lyzed the influence of SNPs localized in the [–70, –20] region from the transcription start site of the gene on TBP / promoter affinity in two groups of genes that are components of the system of human circadian clock. The first group comprises the genes of the circadian oscillator core (11 genes); the second, the genes of the nearest regulatory environment of the circadian oscillator (21 genes). A group for comparison included genes with another function (31 genes). The SnP_TATA_ comparator web service was used for prediction of the effect of SnPs in the regions of positioning of rnA polymerase ii on the dissociation constant for TBP / promoter. it was shown that the number of SnP markers reducing the TBP / promoter affinity in the first group of genes significantly lower than the number of SnP markers increasing affinity ($\alpha < 10^{-3}$). The reverse was true of the comparison group: SnP markers reduced TBP / promoter affinity to a significantly greater extent than the SnP marker increased affinity ($\alpha < 10^{-6}$). This property may be a characteristic feature of genes of the circadian oscillator. These predictions are important for identification of candidate SnP markers of various pathologies associated with the dysfunction of circadian clock genes for further testing them in experimental and clinical studies, as well as for verification of mathematical models of the circadian oscillator.

Key words: circadian rhythm; the promoter; SnP; TATA-binding protein (TBP); the affinity of TBP / promoter; gene expression.

e-mail: opodkol@bionet.nsc.ru

HOW TO CITE THIS ARTICLE?

Podkolodnaya o .A., r asskazov D.A., Podkolodnyy n .I ., Podkolodnaya n .n ., Suslov V.V., Savinkova l .K., Ponomarenko P.M., Ponoma-renko M.P. The effects of Sn Ps in the regions of positioning r n A polymerase ii on the TBP/promoter affinity in the genes of hu-man circadian clock. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2015;19(6):682-690. Do i 10.18699/VJ15.089

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?

Подколодная О.А., Рассказов Д.А., Подколодный Н.л., Подколодная Н.Н., Суслов В.В., Савинкова л.К., Пономаренко П.М, Пономаренко МП. Влияние однонуклеотидных полиморфных замен в районах позиционирования РНК-полимеразы ii на сродство к ним TBP в генах циркадных часов человека. Вавиловский журнал генетики и селекции. 2015;19(6):682-690. Doi 10.18699/VJ15.089

Циркадные часы (ЦЧ) млекопитающих представляют собой систему самоподдерживающихся осцилляторов, функционирующих под управлением центрального циркадного пейсмейкера, локализованного в супрахиазматических ядрах гипоталамуса. Они синхронизируют все процессы в живых организмах – от транскрипции генов до поведения, обеспечивая их временную адаптацию к 24-часовым земным суткам. Минимальный набор из семи групп генов: Clock (ген *CLOCK*), Bmal (ген *BMAL1*, *BMAL2*), Per (гены *PER1*, *PER2*), Cry (гены *CRY1*, *CRY2*), CKI (гены *CKIε*, *CKIδ*), Rev-erb (ген *REV-ERBA*) и Ror (гены *RORA* и *RORC*) формирует ядро молекулярно-генетического механизма циркадных часов, функционирование которого обеспечивается обратными связями между его компонентами (Reppert, Weaver, 2001; Brown et al., 2012; Kim, Forger, 2012). Кроме того, компоненты ядра циркадного осциллятора (ЦО) связаны обратными связями с рядом генов, не входящих в его ядро. Наличие таких связей дополнительно способствует устойчивости функционирования часов. Кроме того, эти гены могут служить точками входа внешних сигналов, модулирующих параметры циркадных часов в ответ на внешние сигналы, такие как свет, пища и др. (Reppert, Weaver, 2001; Brown et al., 2012; Chen, Yang, 2014).

Система циркадных часов связана с широким спектром физиологических систем организма. Поэтому можно ожидать, что генетическая вариабельность в системе циркадных часов может проявляться в фенотипической изменчивости физиологических функций и поведения, а также в нарушениях функционирования не только самих часов, но и других систем, приводящих к развитию патологических состояний. Это подтверждено, прежде всего, экспериментами на генетических моделях животных, которые продемонстрировали, что помимо изменения параметров ЦЧ (амплитуды, фазового ответа на внешние сигналы, периода свободнотекущего ритма) у мутантных животных проявляются такие расстройства, как метаболический синдром, нарушение в системе глюконеогенеза и липогенеза, нарушение функции почек и термогенеза, развитие опухолей и др. (см. обзоры Ko, Takahashi, 2006; Sahar, Sassone-Corsi, 2012). Кроме того, исследования в области генетической эпидемиологии выявили ассоциации однонуклеотидных полиморфных замен (ОНП) генов циркадных часов с широким спектром патологических состояний (Подколодная, 2014; Kettner et al., 2014). Значительная часть таких ОНП локализована в некодирующих областях генов, обеспечивающих регуляцию экспрессии. Функциональная аннотация регуляторных ОНП, а также анализ их проявления на уровне экспрессии генов представляются важной задачей, поскольку многие из таких ОНП могут быть маркерами патологических состояний. В данной работе мы сосредоточили внимание на области проксимального промотора [–70, –20 от старта транскрипции], соответствующей месту связывания TBP (TATA – Binding Protein). Используя созданный нами ранее Web-сервис SNP_TATA_Comparator (Рассказов и др., 2013), который предоставляет возможность в автоматическом режиме предсказывать изменения величин константы диссоциации TBP/промотор, мы проанализировали влияние ОНП, локализованных в этой области, что позволило оценить их потенциальное влияние на экспрессию соответствующих генов. Такой подход существенно облегчает отбор кандидатных ОНП-маркеров ряда патологий для дальнейшей проверки их в экспериментальных исследованиях.

Материалы и методы

Источником данных о генах, участвующих в регуляции циркадного ритма, послужила генная сеть «Circadian Rhythm» (Подколодная и др., 2014). На ее основе было составлено два списка генов, один из которых содержал гены ядра циркадного осциллятора (I группа), второй – гены его ближайшего окружения (II группа). Список генов ядра циркадного осциллятора совпадает с описанным в литературе (Дополнительные материалы 1[1]). Список генов ближайшего окружения (Доп. материалы 2) включал гены, являющиеся дополнительными регуляторами осциллятора. Некоторые из них связаны с генами ядра отрицательными обратными связями, другие участвуют

[1] Дополнительные материалы 1–3 см. в Приложении 2 по адресу: http://www.bionet.nsc.ru/vogis/download/pict-2015-12/appx2.pdf

в регуляции на различных уровнях экспрессии генов ядра или деградации их продуктов, являются точками входа регуляторных сигналов, а также выполняют другие регуляторные функции. Для прогноза влияния ОНП в области проксимальных промоторов [–70, –20 от старта транскрипции] на сродство к ним ТВР в этих двух группах генов человека мы использовали Web-сервис SNP_TATA_Comparator (Рассказов и др., 2013). При работе Web-сервиса основные варианты последовательностей извлекаются из базы данных Ensembl (Zerbino et al., 2015) на основе использования границ транскриптов, взятых из базы данных GENCODE (Frankish et al., 2015) референсного генома человека hg19. Минорные варианты создаются автоматически путем внесения в основные варианты последовательностей соответствующих замен, делеций и/или вставок нуклеотидов из базы данных dbSNP (вып. 142). SNP_TATA_Comparator позволяет оценивать сродство ТВР/промотор, выраженное как $-\ln(K_D) \pm \delta$ в логарифмических единицах и в nM (здесь $\ln(K_D)$ – логарифм константы диссоциации комплекса ТВР/промотор и δ – стандартное отклонение $\ln(K_D)$), а также проводить сравнение значений, полученных для анцестральных и минорных вариантов последовательности (Рассказов и др., 2013).

Для оценки значимости повышения или понижения количества ОНП использовался метод сравнения наблюдаемого количества с ожидаемым по случайным причинам количеством, соответствующим 5 %-му уровню, на основе биномиального распределения. Значимость различия количества ОНП, увеличивающих и уменьшающих сродство, рассчитывалась на основе оценки вероятности реализации по случайным причинам наблюдаемого количества ОНП при условии выполнения нулевой гипотезы об их равенстве в генеральной совокупности. Эти расчеты также выполнялись с использованием биномиального распределения. Для оценки значимости взаимосвязи между двумя переменными в таблице сопряженности размером 2×2 использовался точный критерий Фишера.

В соответствии с экспериментальными данными значимый рост сродства ТВР/промотор можно рассматривать как предсказание суперэкспрессии гена, снижение – как признак дефицитной экспрессии (Mogno et al., 2010).

результаты и обсуждение

Основные результаты, полученные нами, представлены в табл. 1. Первая часть ее содержит информацию о генах группы ядра циркадного осциллятора человека, вторая – о генах его ближайшего регуляторного окружения. Из I группы пять генов кодируют транскрипционные факторы, два – ферменты и четыре – регуляторные белки. Отметим, что в таблице представлены только те ОНП, которые значимо изменяют константу диссоциации ТВР/промотор. В графе K_D представлены полученные нами оценки константы диссоциации комплекса ТВР для основного и минорного типов последовательности, а в графах Z и α –статистические оценки достоверности изменения константы диссоциации ТВР/промотор для каждого ОНП. В графе Δ этой таблицы отмечено предполагаемое влияние данного ОНП на уровень экспрессии гена. Кроме того, в таблице содержатся данные из научных публикаций, свидетельствующие о состояниях организма, при которых выявлены такие изменения экспрессии соответствующих генов. Рассмотрим более подробно данные, представленные в табл. 1, на примере гена *PER1,* кодирующего белок PERIOD 1 (PER1), субъединицу гетеродимерного комплекса PER/CRY, который является основным негативным компонентом циркадного осциллятора, подавляющим активность транскрипционного фактора CLOCK/BMAL1 за счет белок-белковых взаимодействий. Из 54 ОНП, выявленных в исследуемом районе различных транскриптов этого гена, только два (rs2518024 и rs3027175), согласно прогнозу SNP_TATA_Comparator, могут оказывать влияние на сродство ТВР/промотор. Эти ОНП локализованы в области [–70; –20] относительно старта транскрипции мРНК № 5 (референсный геном человека hg19). При этом первый ОНП увеличивает сродство ТВР/промотор, а второй – уменьшает его. Как отмечалось выше, увеличение сродства ТВР/промотор можно рассматривать как предсказание увеличения экспрессии гена, а снижение его – как предсказание уменьшения экспрессии. Как известно, высокий уровень экспрессии гена *PER1* ингибирует рост раковых клеток (Gery et al., 2006; Cao et al., 2009; Kettner et al., 2014). Сниженный уровень экспрессии *PER1* выявлен в тканях раковых опухолей желудка и простаты человека (Cao et al., 2009; Zhao et al., 2014). В то же время у пациентов с высокой экспрессией этого гена, страдающих раком желудка, отмечается более длительное время выживания (Zhao et al., 2014). Этот ген рассматривают как супрессор опухолей, один из механизмов действия которого определяется влиянием на чувствительность клеток к индуцированному разрушением ДНК апоптозу (Gery et al., 2006; Kettner et al., 2014). Отметим также, что в исследованиях на генетических моделях мыши ($Per^{-/-}$) наблюдается снижение 3D обучаемости и увеличение проявлений гепатотоксичности этанола (Jilg et al., 2010; Wang et al., 2013). Следовательно, можно предположить, что данные ОНП гена *PER1* могут рассматриваться как кандидатные ОНП-маркеры ряда патологий.

Аналогичная информация представлена в табл. 1 для остальных генов ядра циркадного осциллятора, в анализируемой области промоторов которых выявлены ОНП, влияющие на сродство ТВР/промотор. В целом в данной группе генов выявлено 282 ОНП, из них только для 14 предсказаны статистически значимые изменения сродства ТВР/промотор (два ОНП значимо снижали этот показатель, что в согласии с нашими предположениями может уменьшать экспрессию соответствующих генов, а 12 – увеличивали).

Данные, представленные в табл. 1 в графе «Патология», свидетельствуют о том, что дисрегуляция экспрессии генов ядра циркадного осциллятора может служить маркером широкого круга патологических состояний, таких как различные формы рака, заболевания легких, сердечно-сосудистой системы, нейродегенративные заболевания и др. Такое разнообразие может быть следствием того, что циркадные часы осуществляют регуляцию большого количества процессов, протекающих в организме, и, кроме того, являются интеграторами различных сигналов, получаемых ими за счет обратных связей как локально, на клеточном уровне, так и на уровне организма, благодаря их иерархической системе организации.

Table 1. Candidate SnPs-markers of pathologies that are important for the affinity of TBP to the promoters of genes regulating the circadian rhythm

Gene (n_{SnPs})	rnA (oth.)	dbSnP rel.142	5′-flan	**hg19** / *min*	flan -3′	K_D, nM **hg19** / *min*	Δ	Z	α	condition	reference(s)
Group i: Genes of the circadian oscillator core											
CLOCK (38)	no 1	rs**192518038**	aggacctaag	**g** / *t*	ctagcgctct	**63** / *29*	↑	14	10^{-7}	Elevated risk of heart attacks in diabetes (*CLOCK*-mutant mice)	oishi et al., 2005
		rs537333415	gcctccagga	**c** / *t*	ctaaggctag	**63** / *45*	↑	7	10^{-7}		
CRY2 (13)	no 1	rs575588903	ctaagggtgg	**a** / *g*	gttgcggcgt	**27** / *25*	↑	2	0.05	chemoresistance and poor patient survival in colorectal cancer; diffuse subcortical glioma	luo et al., 2012; Fang et al., 2015
	no 2	rs529410313	agctgtcagt	**c** / *a*	ttgcaagtca	**22** / *18*	↑	3	10^{-2}		
PER1 (54)	no 5	rs2518024	gtgctctgga	**g** / *a*	ttaaaccagc	**17** / *8*	↑	12	10^{-7}	longer survival of patients with gastric cancer	Zhao et al., 2014
	no 5	rs3027175	ccagcaggtg	**c** / *t*	tctggagtta	**17** / *19*	↓	2	0.05	Prostate cancer, hippocampal defect, impaired 3D learning, and poor tolerance of ethanol in the liver (*PER1*$^{-/-}$ mice)	cao et al., 2009; Jilg et al., 2010; wang et al., 2013
PER2 (11)	no 5	rs181985043	gcagctccac	**c** / *a*	ctagtgacag	**29** / *12*	↑	16	10^{-7}	risk of Q fever in men; growth suppression of tumor cells (cell line S-180)	Miyazaki et al., 2010; Mehraj et al., 2012
RORA (21)	no 2	rs**374778785**	attatccccc	**a** / *t*	tactcctccc	**34** / *28*	↑	4	10^{-3}	risk of emphysema and its development to lung cancer in smokers (mice, smoking machine TE-10)	Shi et al., 2012
RORC (**20**)	no 1	rs568650510	actccttttc	**c** / *t*	ctgcctgctg	**55** / *25*	↑	14	10^{-7}	risk of asthma (30 pediatric patients) and neurological manifestations of Behçet's disease (18 patients)	Hamzaoui et al., 2011a, b
CSNK1E (36)	no 1	rs369188273	ccctcccctc	**c** / *t*	gcgcccgctc	**288** / *195*	↑	7	10^{-7}	risk of MYc-dependent carcinogenesis, ovarian cancer, elevated Abeta peptide production in the brain	Flajolet et al., 2007; rodriguez et al., 2012; Toyoshima et al., 2012
	no 6	rs558609213	tcttttcttg	**c** / *t*	atccctgcag	**30** / *9*	↑	21	10^{-7}		
	no 1	rs2899302	cgagaaaact	**g** / *c*	cgcgaggcct	**288** / *335*	↓	3	10^{-2}	Susceptibility to opioids (*CSNK1E*$^{-/-}$ mice)	Bryant et al., 2012
CSNK1D (51)	no 1	rs540139460	gcagggtcgg	**g** / *a*	aggaggcctg	**253** / *145*	↑	10	10^{-7}	risk of breast cancer (27 samples from the bank of snap-frozen surgically removed tumors)	Abba **et al.,** 2007
	no 2	rs4313857	gccccgccgg	**g** / *a*	ttgctagggg	**57** / *32*	↑	8	10^{-7}		
Group ii: Genes closest to the circadian oscillator											
BHLHE40 (8)	no 1	rs527901110	gcgccccccа	**c** / *g*	ccaactgggc	**79** / *94*	↓	3	10^{-3}	oesophageal squamous cell carcinoma	wong **et al.,** 2011
BHLHE41 (**8**)	no 1	rs532670734	ctgccgttcg	**c** / *t*	ctgccgttcg	**86** / *55*	↑	9	10^{-7}	risk of pancreatic cancer (human pancreatic cancer cells BxPc-3)	Sato **et al., 2012**
NGFR (10)	no 2	rs532346435	tgctgactaa	**c** / *t*	gccgctggtt	**24** / *13*	↑	6	10^{-7}	risk of breast cancer (human breast cancer cell line McF-7)	wilmet et al., 2011
ID2 (17)	no 201	rs377457836	gcccgctcgt	**c** / *t*	ttgatagacg	**7** / *6*	↑	4	10^{-3}	More intense tumor cell migration, metastasis (cell line U87MG)	coma et al., 2010
FBXL3 (19)	no 1	rs558539957	ggcgccattt	**t** / *c*	agactcgctc	**22** / *51*	↓	13	10^{-7}	reduced amplitude and longer circadian rhythm period (*FBXL3*$^{-/-}$ mouse, *FBXL3*$^{-/-}$ cells with the use of human criSPr/cas9)	Hirano et al., 2013; Korge et al., 2015
		rs368477256	cgagccctgg	**g** / *c*	cgccatttta	**22** / *63*	↓	12	10^{-7}		

End of Table 1

Gene (n_{SnPs})	rn A (oth.)	dbSnP rel.142	5′-flan	**hg19** / *min*	flan -3′	K_D, nM **hg19** / *min*	Δ	Z	α	condition	reference(s)
SIRT1 (13)	no. 1	rs561608453	ggggtttaaa	**t** / *a*	ctcccgcagc	**7** / *5*	↑	3	10^{-2}	reduced risk of amyotrophic lateral sclerosis (mice overexpressing ***SIRT1***)	watanabe et al., 2014
		rs200894138	tggggtttaa	**a** / *c*	tctcccgcag	**7** / *10*	↓	7	10^{-7}	risk of severe lung damage during inflammation (*SIRT1*$^{-/-}$ mice)	Gao et al., 2015
HLF (29)	no. 1	rs568118114	tcactcttgt	**c** / *g*	agggccgcgg	**34** / *29*	↑	3	10^{-3}	Acute lymphoblastic leukemia (transgenic mice with *HLF* overexpression)	Yamasaki et al., 2010
	no. 5	rs546257513	tccaaagtaa	**a** / *g*	taggaagatg	**6** / *9*	↓	5	10^{-7}	cardiac hypertrophy and left ventricular dysfunction associated with low blood pressure (*DBP*$^{-/-}$*TEF*$^{-/-}$*HLF*$^{-/-}$ mice)	wang et al., 2010
		rs571526277	caaagtaaat	**a** / *t*	ggaagatgtt	**6** / *10*	↓	10	10^{-7}		
PPARA (20)	no. 1	rs140255482	agagaacaac	**c** / *t*	gtaatcactt	**36** / *23*	↑	8	10^{-7}	risk of cardiovascular diseases in mice receiving low-fat diet (model of spontaneous hypertension)	Burgueno et al., 2013
	no. 8	rs79623011	ctgggtgatt	**t** / *a*	ataaacaaca	**3** / *4*	↓	6	10^{-7}	Decreased longevity and elevated risk of age-dependent disorders (*PPARA*$^{-/-}$ mice)	Howroyd et al., 2004
PPARG (22)	no. 1	rs569533856	ccgggggggct	**g** / *a*	gaaaggcgaa	**79** / *15*	↑	11	10^{-7}	risk of adipogenic steatosis in the mouse liver (transgenic mice overexpressing *PPARG*)	Yu et al., 2003
FBXW7 (37)	no. 2	rs374622342	ccgcttcttc	**c** / *t*	tcagtaccgc	**26** / *16*	↑	8	10^{-7}	Better response to chemotherapy in colorectal cancer cells (patients and *FBXW7*$^{-/-}$ HcT116 cell line)	Fang et al., 2015
		rs371186818	ttagcgacac	**g** / *a*	agcaccgctt	**26** / *23*	↑	2	0.05		
		rs570804767	tttagcgaca	**c** / *a*	gagcaccgct	**26** / *18*	↑	6	10^{-7}		
	no. 13	rs527901110	atttacattt	**[12 n.o.]** / *–*	atttttatct	**3** / *2*	↑	6	10^{-7}		
	no. 3	rs549308580	ttcagtagtt	**t** / *g*	ataattcctg	**2** / *8*	↓	18	10^{-7}	Poor survival in breast cancer (406 patients, Japan)	Yumimoto et al., 2015
	no. 13	rs35003923	agattccttt	**a** / *g*	tattgaacat	**3** / *5*	↓	8	10^{-7}		
NAMPT (19)	no. 1	rs564443257	gcgtgcatca	**c** / *a*	gtcctcctcc	**53** / *36*	↑	12	10^{-7}	risk of thyroid cancer (mouse cell lines)	Sawicka-Gutaj et al., 2015
	no. 2	rs556790067	gcagtgactt	**a** / *c*	agcaacggag	**15** / *24*	↓	3	10^{-3}	reduced risk of inflammatory vessel damage in myocardial atherosclerosis (*APOE*$^{-/-}$ atherosclerosis model (mouse), FK-866 inhibitor of *NAMPT*)	nencioni et al., 2014
	no. 3	rs114947395	caaaaatata	**t** / *c*	actgacttca	**2** / *4*	↓	11	10^{-7}		
RBM4 (25)	no. 2	rs552014762	tcacctcccc	**c** / *t*	ttctactcag	**10** / *9*	↑	2	0,05	reduction of apoptosis in breast cancer cells (**MCF-7**)	lin et al., 2014b
	no. 3	rs368067876	ctcggcattg	**c** / *a*	gcggtaagcc	**63** / *33*	↑	8	10^{-7}		
	no. 2	rs11551411	ggaacgccgg	**c** / *t*	cttttacttt	**10** / *13*	↓	5	10^{-3}	Hyperlipidemia accompanied with reduced mass of interscapular brown adipose tissue (*RBM4*$^{-/-}$ mice)	lin et al., 2014a
PER3 (27)	no. 1	rs532828969	ctgtctgttc	**c** / *t*	atttgtccct	**19** / *11*	↑	10	10^{-7}	risk of colorectal cancer and metastasis in liver cancer (202 patients)	oshima et al., 2011
	no. 3	rs373227456	gctgctgacc	**g** / *a*	gcacgcggcg	**166** / *131*	↑	4	10^{-3}		
	no. 1	rs172933	gtctgttcca	**t** / *c*	ttgtcccttg	**19** / *24*	↓	4	10^{-3}	risk of recurrent Er -positive breast cancer; hyperactivity when exposed to chronic light regimens (*PER3*$^{-/-}$ mice)	climent et al., 2010; Hasan et al., 2011; Pereira et al., 2014

n_{SnPs}, the number of analyzed SnPs; rnA, protein-encoding transcripts with sequenced 5′-ends and TSI indices below 3, enumerated according to ENSEMBL; **hg19**, major allele; *min*, minor allele; K_D, the estimate (rasskazov et al., 2013) of the dissociation constant K_D кof the TBP-promoter complex corresponding to in vitro conditions (Savinkova et al., 2013). Δ, change: (↑), excess; (↓), deficiency; Z, Z-score; α, significance.

Table 2. Candidate SnPs markers affecting the TBP/promoter affinity in genes of the circadian oscillator, its regulatory environment, and the reference group

Group of genes	number of genes	number of SnPs	SnPs that increase affinity		SnPs that decrease affinity	
			number of SnPs $n(i, 1)$	The significance of differences expected by chance with 5% probability	number of SnPs $n(i, 2)$	The significance of differences expected by chance with 5% probability
i. Genes of the circadian oscillator core	11	282	$n(1, 1) = 12$	>0.3	$n(1, 2) = 2$	$<10^{-4}$ (–)
The significance of differences in the number of SnPs increasing and decreasing the affinity				$<10^{-3}$ (*)		
ii. Genes of the regulatory environment of the circadian oscillator core	21	341	$n(2, 1) = 16$	>0.4	$n(2, 2) = 13$	$>0,1$
The significance of differences in the number of SnPs increasing and decreasing the affinity				$>0,2$		
iii. reference group	31	202	$n(3, 1) = 13$	>0.2	$n(3, 2) = 37$	$<10^{-12}$ (+)
The significance of differences in the number of SnPs increasing and decreasing the affinity				$<10^{-6}$ (*)		

Designations: SnP, single nucleotide polymorphism; $n(i, 1)$ and $(i, 2)$, numbers of SnP markers increasing and decreasing the affinity to the i^{th} group of genes, respectively; +, significant increase of the SnP number compared to the number expected by chance; –, significant reduction of the SnP number compared to the number expected by chance; *, significant difference between the numbers of SnP markers that increase and decrease the affinity to a particular group of genes.

По такой же схеме был проведен анализ генов II группы, которая была охарактеризована нами как ближайшее регуляторное окружение ядра циркадного осциллятора. Группа включает 21 ген, среди которых гены, кодирующие транскрипционные факторы (9), регуляторные белки (3), ферменты (3), компоненты убиквитинлигазного комплекса (4), РНК-связывающий белок (1) и рецептор ростового фактора (1). Из 341 проанализированных ОНП 29 потенциально способны оказывать влияние на исследуемый показатель (табл. 1). Оказалось, что в этой группе количество ОНП, предположительно увеличивающих экспрессию гена, несколько превышает количество понижающих ее (16 против 13). Патологии, при которых наблюдается изменение экспрессии анализируемых генов, так же как и в предыдущей группе, весьма разнообразны. Помимо состояний, связанных с нарушениями циркадного ритма, это различные формы рака, нейродегенеративные, сердечно-сосудистые, обменные и другие заболевания. Аналогично предыдущей группе генов патологические последствия может иметь как увеличение, так и уменьшение экспрессии генов этой группы (табл. 1).

Выявленные в этом исследовании ОНП промоторов потенциально способны влиять на сродство ТВР/промотор в генах обеих групп и оказывать негативное воздействие на их регуляцию и функцию, что может приводить к формированию патологических состояний. Можно рекомендовать эти ОНП для дальнейших исследований в качестве кандидатных ОНП-маркеров.

Результаты сравнения достоверности влияния ОНП на сродство ТВР/промотор в исследованных группах представлены в табл. 2. Как можно видеть, в группе генов ядра циркадного осциллятора доля ОНП-маркеров дефицитной экспрессии гена (снижения сродства ТВР/промотор) значимо ниже 5 %-го порога ожидаемой по случайной причине ($\alpha<10^{-4}$), тогда как доля ОНП-маркеров суперэкспрессии (увеличения сродства ТВР/промотор) тех же генов не отличается от ожидаемого по случайным причинам значения ($\alpha>0,3$). При этом доли ОНП-маркеров дефицитной экспрессии и суперэкпрессии в группе генов ядра циркадного осциллятора достоверно различаются ($\alpha<10^{-3}$). Одной из причин, объясняющих низкую частоту ОНП-маркеров дефицитной экспрессии в группе генов ядра циркадного осциллятора, может быть предположение о наличии отрицательного эволюционного отбора по данному показателю в связи с меньшей приспособленностью организмов с ОНП, уменьшающими уровень экспрессии генов циркадного осциллятора.

В группе генов регуляторного окружения ядра циркадного осциллятора доли ОНП-маркеров обоих типов изменения сродства ТВР/промотор не отличаются от ожидаемого по случайной причине значения, а также не различаются между собой.

Уровень значимости различий количеств кандидатных ОНП-маркеров генов ядра циркадного осциллятора и генов его ближайшего окружения по точному критерию Фишера $\alpha(1,2)<0,086$, что может рассматриваться как тенденция: количество ОНП-маркеров снижения сродства ТВР/промотор генов I группы ниже количества ОНП-маркеров снижения сродства ТВР/промотор II группы.

Необходимо отметить, что мы выбрали для сравнения кандидатные ОНП-маркеры, которые изменяют константу K_D с уровнем значимости $<0,05$ (см. табл. 1). Выбор уровня значимости достаточно условный. Если принять в рассмотрение только ОНП-маркеры, которые меняют K_D с уровнем значимости $<0,01$, то для этих более достоверных ОНП-маркеров получится измененная 2×2 таблица сопряженности со значениями: $n(1,1) = 11$, $n(1,2) = 1$; $n(2,1) = 14$; $n(2,2) = 13$ и уровнем значимости различий между I и II группами по точному критерию Фишера $\alpha(1,2)<0,0028$. Это подкрепляет указанную тенденцию.

Для того чтобы проверить, является ли наблюдаемое в исследованной от начала транскрипта области [–70, –20]

снижение доли ОНП-маркеров, потенциально снижающих экспрессию, характеристической особенностью промоторов генов циркадного осциллятора, мы провели сравнение полученных результатов с результатами аналогичного исследования, проведенного нами ранее в группе генов, отличных по функциональной принадлежности от исследуемой группы. Эта группа из 31 гена была представлена в обзоре Ponomarenko с коллегами (2015) и сформирована на основе поиска в научных публикациях ОНП-маркеров в области, соответствующей сайту связывания ТВР в промоторах генов человека, ассоциированных с различными патологиями. Полный список генов этой группы можно найти в Доп. материалах 3. С помощью нашего Web-сервиса в этих генах были выявлены соседние с этими маркерами кандидатные ОНП-маркеры.

Полученная суммарная выборка содержала 203 ОНП, локализованные в промоторах 31 гена человека. Из них только 50 оказались способны потенциально воздействовать на сродство ТВР/промотор соответствующих генов (13 – увеличивали, 37 – понижали) (табл. 2). Как можно видеть, в этой группе генов, в отличие от рассмотренных выше, наблюдается другая картина: доля ОНП-маркеров суперэкспрессии также недостоверно превышала 5 %-й порог *a priory* ожидаемой их доли ($\alpha > 0{,}2$), тогда как доля ОНП-маркеров дефицитной экспрессии тех же генов оказалась достоверно высокой ($\alpha < 10^{-12}$). Данный результат в определенной степени соответствует таковым, полученным при сопоставлении информации из проектов «1 000 Genomes Project» и «ENCODE», свидетельствующим, что в целом по геному количество ОНП, повреждающих сайты связывания транскрипционных факторов, значительно выше, чем улучшающих его (1 000 Genomes Project Consortium et al., 2012). Кроме того, исследования Kasowski с коллегами (2010) свидетельствуют о том, что ОНП, попадающие в сайт связывания транскрипционного фактора NF-κB или Pol II, значимо чаще способствуют снижению, чем повышению связывания их с мутированным ДНК мотивом (при сравнении с референсным геномом) (Kasowski et al., 2010). То есть, возвращаясь к нашим результатам, можно отметить, что ожидаемым было бы преобладание в исследуемой нами области промоторов доли ОНП, уменьшающих сродство ТВР/промотор, как это показано в группе сравнения, в то время как в группах генов циркадного осциллятора наблюдается обратная картина.

Количество ОНП-маркеров увеличения и уменьшения сродства в I группе генов значимо отличается от такового в контрольной группе по точному критерию Фишера ($\alpha(1,3) < 10^{-4}$). Различия II группы генов от контрольной также оказались статистически значимыми ($\alpha(2,3) < 0{,}00021$).

Таким образом, полученные закономерности (уменьшение доли ОНП-маркеров снижения сродства ТВР/промотор) могут быть специфической характеристикой генов циркадного осциллятора, влияющей на робастность циркадного осциллятора при генетической вариабельности анализируемой области промоторов.

Особенности распределения количества кандидатных ОНП-маркеров в группе ближайшего регуляторного окружения циркадного осциллятора могут быть объяснены большей функциональной гетерогенностью ее генов. В эту группу входят гены, находящиеся под непосредственным контролем ядра циркадного осциллятора, имеющие с ним обратные связи и демонстрирующие ритмический характер экспрессии мРНК (например, *PPARα/γ*, *PER3*, *ID2* и др.), и гены, продукты которых участвуют в процессах, обеспечивающих деградацию компонент осциллятора, передачу к нему различных внешних сигналов (*FBXW7*, *FBXL3*, *SIRT1* и др.). По этому признаку данная группа занимает промежуточное положение между группами ядра циркадного осциллятора и сравнения.

Полученные предсказания могут играть важную роль для отбора кандидатных ОНП-маркеров различных патологий, связанных с нарушением генов циркадного ритма, для дальнейшей проверки их в экспериментальных исследованиях, а также при верификации математических моделей циркадного осциллятора.

Выявленные в данной работе особенности генов ЦО требуют дальнейшего анализа. В частности, полезным может оказаться проведение аналогичных исследований в специально сформированных группах генов, схожих и контрастных по функциональным характеристикам с генами циркадных часов. Кроме того, изучение структурных характеристик промоторов, таких как их нуклеотидный контекст, в частности GC-состав, наличие множественных стартов транскрипции и т. д., возможно, прольет свет на природу наблюдаемых особенностей.

Acknowledgments

DAR, NLP, OAP, and NNP acknowledge the support of the Web service design, data collection and preprocessing, and result interpretation by the Russian Science Foundation, project 14-24-00123. MPP acknowledges the support of data analysis by the Russian Foundation for Basic Research, project 14-04-00485.

Conflict of interest

The authors declare no conflict of interest.

References

1 000 Genomes Project Consortium; Abecasis G.R., Auton A., Brooks L.D., DePristo M.A., Durbin R.M., Handsaker R.E., Kang H.M., Marth G.T., McVean G.A. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56-65. DOI 10.1038/nature11632

Abba M.C., Sun H., Hawkins K.A., Drake J.A., Hu Y., Nunez M.I., Gaddis S., Shi T., Horvath S., Sahin A., Aldaz C.M. Breast cancer molecular signatures as determined by SAGE: correlation with lymph node status. Mol. Cancer Res. 2007;5(9):881-890.

Brown S.A., Kowalska E., Dallmann R. (Re)inventing the circadian feedback loop. Dev. Cell. 2012;22(3):477-87. DOI 10.1016/j.devcel.2012.02.007

Bryant C.D., Parker C.C., Zhou L., Olker C., Chandrasekaran R.Y., Wager T.T., Bolivar V.J., Loudon A.S., Vitaterna M.H., Turek F.W., Palmer A.A. Csnk1e is a genetic regulator of sensitivity to psychostimulants and opioids. Neuropsychopharmacology. 2012;37(4):1026-1035. DOI 10.1038/npp.2011.287

Burgueno A.L., Gianotti T.F., Mansilla N.G., Pirola C.J., Sookoian S. Cardiovascular disease is associated with high-fat-diet-induced liver damage and up-regulation of the hepatic expression of hypoxia-inducible factor 1α in a rat model. Clin. Sci. (Lond). 2013;124(1):53-63. DOI 10.1042/CS20120151

Cao Q., Gery S., Dashti A., Yin D., Zhou Y., Gu J., Koeffler H.P. A role for the clock gene per1 in prostate cancer. Cancer Res. 2009;69(19):

7619-7625. DOI 10.1158/0008-5472.CAN-08-4199

Chen L., Yang G. PPARs integrate the mammalian clock and energy metabolism. PPAR Res. 2014;2014:653017. DOI 10.1155/2014/653017

Climent J., Perez-Losada J., Quigley D.A., Kim I.J., Delrosario R., Jen K.Y., Bosch A., Lluch A., Mao J.H., Balmain A. Deletion of the PER3 gene on chromosome 1p36 in recurrent ER-positive breast cancer. J. Clin. Oncol. 2010;28(23):3770-3778. DOI 10.1200/JCO.2009.27.0215

Coma S., Amin D.N., Shimizu A., Lasorella A., Iavarone A., Klagsbrun M. Id2 promotes tumor cell migration and invasion through transcriptional repression of semaphorin 3F. Cancer Res. 2010;70(9):3823-3832. DOI 10.1158/0008-5472.CAN-09-3048

Fang L., Yang Z., Zhou J., Tung J.Y., Hsiao C.D., Wang L., Deng Y., Wang P., Wang J., Lee M.H. Circadian clock gene CRY2 degradation is involved in chemoresistance of colorectal cancer. Mol. Cancer Ther. 2015;14(6):1476-1487. DOI 10.1158/1535-7163.MCT-15-0030

Flajolet M., He G., Heiman M., Lin A., Nairn A.C., Greengard P. Regulation of Alzheimer's disease amyloid-beta formation by casein kinase I. Proc. Natl Acad. Sci. USA. 2007;104:4159-4164.

Frankish A., Uszczynska B., Ritchie G.R., Gonzalez J.M., Pervouchine D., Petryszak R., Mudge J., Fonseca N., Brazma A., Guigo R., Harrow J. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. BMC Genomics. 2015;16(Suppl.8):S2. DOI 10.1186/1471-2164-16-S8-S2

Gao R., Ma Z., Hu Y., Chen J., Shetty S., Fu J. Sirt1 restrains lung inflammasome activation in a murine model of sepsis. Am. J. Physiol. Lung Cell Mol. Physiol. 2015;308(8):L847-L853. DOI 10.1152/ajplung.00274.2014

Gery S., Komatsu N., Baldjyan L., Yu A., Koo D., Koeffler H.P. The circadian gene per1 plays an important role in cell growth and DNA damage control in human cancer cells. Mol. Cell. 2006;22(3):375-382.

Hamzaoui A., Maalmi H., Berraies A., Abid H., Ammar J., Hamzaoui K. Transcriptional characteristics of CD4 T cells in young asthmatic children: RORC and FOXP3 axis. J. Inflamm. Res. 2011a;4:139-146.

Hamzaoui K., Borhani Haghighi A., Ghorbel I.B., Houman H. RORC and Foxp3 axis in cerebrospinal fluid of patients with neuro-Behçet's disease. J. Neuroimmunol. 2011b;233(1/2):249-253. DOI 10.1016/j.jneuroim.2011.01.012

Hasan S., van der Veen D.R., Winsky-Sommerer R., Dijk D.J., Archer S.N. Altered sleep and behavioral activity phenotypes in PER3-deficient mice. Am. J. Physiol. Regul. Integr. Comp. Physiol. 2011;301(6):R1821-R1830. DOI 10.1152/ajpregu.00260.2011

Hirano A., Yumimoto K., Tsunematsu R., Matsumoto M., Oyama M., Kozuka-Hata H., Nakagawa T., Lanjakornsiripan D., Nakayama K.I., Fukada Y. FBXL21 regulates oscillation of the circadian clock through ubiquitination and stabilization of cryptochromes. Cell. 2013;152(5):1106-1118. DOI 10.1016/j.cell.2013.01.054

Howroyd P., Swanson C., Dunn C., Cattley R.C., Corton J.C. Decreased longevity and enhancement of age-dependent lesions in mice lacking the nuclear receptor peroxisome proliferator-activated receptor alpha (PPARalpha). Toxicol. Pathol. 2004;32(5):591-599. DOI 10.1080/01926230490515283

Jilg A., Lesny S., Peruzki N., Schwegler H., Selbach O., Dehghani F., Stehle J.H. Temporal dynamics of mouse hippocampal clock gene expression support memory processing. Hippocampus. 2010;20(3):377-388. DOI 10.1002/hipo.20637

Kasowski M., Grubert F., Heffelfinger C., Hariharan M., Asabere A., Waszak S.M., Habegger L., Rozowsky J., Shi M., Urban A.E., Hong M.Y., Karczewski K.J., Huber W., Weissman S.M., Gerstein M.B., Korbel J.O., Snyder M. Variation in transcription factor binding among humans. Science. 2010;328(5975):232-235. DOI 10.1126/science.1183621

Kettner N.M., Katchy C.A., Fu L. Circadian gene variants in cancer. Ann. Med. 2014;46(4):208-220. DOI 10.3109/07853890.2014.914808

Kim J.K., Forger D.B. A mechanism for robust circadian timekeeping via stoichiometric balance. Mol. Syst. Biol. 2012;8:630. DOI 10.1038/msb.2012.62

Ko C.H., Takahashi J.S. Molecular components of the mammalian circadian clock. Hum. Mol. Genet. 2006;15(Spec No 2):R271-277.

Korge S., Grudziecki A., Kramer A. Highly efficient genome editing via CRISPR/Cas9 to create clock gene knockout cells. J. Biol. Rhythms. 2015;30(5):389-395. DOI 10.1177/0748730415597519

Lin J.C., Lin C.Y., Tarn W.Y., Li F.Y. Elevated SRPK1 lessens apoptosis in breast cancer cells through RBM4-regulated splicing events. RNA. 2014b;20(10):1621-1631. DOI 10.1261/rna.045583.114

Lin J.C., Tarn W.Y., Hsieh W.K. Emerging role for RNA binding motif protein 4 in the development of brown adipocytes. Biochim. Biophys. Acta. 2014a;1843(4):769-779. DOI 10.1016/j.bbamcr.2013.12.018

Luo Y., Wang F., Chen L.A., Chen X.W., Chen Z.J., Liu P.F., Li F.F., Li C.Y., Liang W. Deregulated expression of cry1 and cry2 in human gliomas. Asian Pac. J. Cancer Prev. 2012;13(11):5725-5728.

Mehraj V., Textoris J., Capo C., Raoult D., Leone M., Mege J.L. Overexpression of the Per2 gene in male patients with acute Q fever. J. Infect. Dis. 2012;206(11);1768-1770. DOI 10.1093/infdis/jis600

Miyazaki K., Wakabayashi M., Hara Y., Ishida N. Tumor growth suppression in vivo by overexpression of the circadian component, PER2. Genes Cells. 2010;15(4):351-358. DOI 10.1111/j.1365-2443.2010.01384.x

Mogno I., Vallania F., Mitra R., Cohen B. TATA is a modular component of synthetic promoters. Genome Res. 2010;20(10):1391-1397. DOI 10.1101/gr.106732.110

Nencioni A., da Silva R.F., Fraga-Silva R.A., Steffens S., Fabre M., Bauer I., Caffa I., Magnone M., Sociali G., Quercioli A., Pelli G., Lenglet S., Galan K., Burger F., Vazquez Calvo S., Bertolotto M., Bruzzone S., Ballestrero A., Patrone F., Dallegri F., Santos R.A., Stergiopulos N., Mach F., Vuilleumier N., Montecucco F. Nicotinamide phosphoribosyltransferase inhibition reduces intraplaque CXCL1 production and associated neutrophil infiltration in atherosclerotic mice. Thromb. Haemost. 2014;111(2):308-322. DOI 10.1160/TH13-07-0531

Oishi K., Ohkura N., Amagai N., Ishida N. Involvement of circadian clock gene Clock in diabetes-induced circadian augmentation of plasminogen activator inhibitor-1 (PAI-1) expression in the mouse heart. FEBS Lett. 2005;579(17):3555-3559.

Oshima T., Takenoshita S., Akaike M., Kunisaki C., Fujii S., Nozaki A., Numata K., Shiozawa M., Rino Y., Tanaka K., Masuda M., Imada T. Expression of circadian genes correlates with liver metastasis and outcomes in colorectal cancer. Oncol. Rep. 2011;25(5):1439-1446. DOI 10.3892/or.2011.1207

Pereira D.S., van der Veen D.R., Gonçalves B.S., Tufik S., von Schantz M., Archer S.N., Pedrazzoli M. The effect of different photoperiods in circadian rhythms of per3 knockout mice. Biomed. Res. Int. 2014;2014:170795. DOI 10.1155/2014/170795

Podkolodnaya O.A. Molecular and genetic aspects of interactions of the circadian clock and the energy producing substrate metabolism in mammals. Genetika = Genetics (Moscow). 2014;50(2):125-137.

Podkolodnaya O.A. Podkolodnaya N.N. Podkolodnyy N.L. The mammalian circadian clock: gene regulatory network and computer analysis. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2014;18(4/2):928-938.

Ponomarenko M., Rasskazov D., Arkova O., Ponomarenko P., Suslov V., Savinkova L., Kolchanov N. How to use SNP_TATA_Comparator to find a significant change in gene expression caused by the regulatory SNP of this gene's promoter via a change in affinity of the TATA-binding protein for this promoter. Biomed. Res. Int. 2015;2015:359835 (in press).

Rasskazov D.A., Gunbin K.V., Ponomarenko P.M., Vishnevsky O.V., Ponomarenko M.P., Afonnikov D.A. SNP_TATA_Comparator: Web-service for comparison of SNPs within gene promoters associated with human diseases using the equilibrium equation of the TBP/TATA complex. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2013;17(4/1):599-606.

Reppert M., Weaver D.R. Molecular analysis of mammalian circadian rhythms. Annu. Rev. Physiol. 2001;63:647-676.

Rodriguez N., Yang J., Hasselblatt K., Liu S., Zhou Y., Rauh-Hain J.A., Ng S.K., Choi P.W., Fong W.P., Agar N.Y., Welch W.R., Berkowitz R.S., Ng S.W. Casein kinase I epsilon interacts with mitochondrial proteins for the growth and survival of human ovarian cancer cells. EMBO Mol. Med. 2012;4:952-963. DOI 10.1002/emmm.201101094

Sahar S., Sassone-Corsi P. Regulation of metabolism: the circadian clock dictates the time. Trends Endocrinol. Metab. 2012;23(1):1-8. DOI 10.1016/j.tem.2011.10.005

Sato F., Kawamura H., Wu Y., Sato H., Jin D., Bhawal U.K., Kawamoto T., Fujimoto K., Noshiro M., Seino H., Morohashi S., Kato Y., Kijima H. The basic helix-loop-helix transcription factor DEC2 inhibits TGF-β-induced tumor progression in human pancreatic cancer BxPC-3 cells. Int. J. Mol. Med. 2012;30(3):495-501. DOI 10.3892/ijmm.2012.1037

Savinkova L.K., Drachkova I.A., Arshinova T.V., Ponomarenko P.M., Ponomarenko M.P., Kolchanov N.A. An experimental verification of the predicted effects of promoter TATA-box polymorphisms associated with human diseases on interactions between the TATA boxes and TATA-binding protein. PLoS One. 2013; 8(2):e54626. DOI 10.1371/journal.pone.0054626

Sawicka-Gutaj N., Waligorska-Stachura J., Andrusiewicz M., Biczysko M., Sowinski J., Skrobisz J., Ruchala M. Nicotinamide phosphorybosiltransferase overexpression in thyroid malignancies and its correlation with tumor stage and with survivin/survivin DEx3 expression. Tumour Biol. 2015;36(10):7859-7863. DOI 10.1007/s13277-015-3506-z

Shi Y., Cao J., Gao J., Zheng L., Goodwin A., An C.H., Patel A., Lee J.S., Duncan S.R., Kaminski N., Pandit K.V., Rosas I.O., Choi A.M., Morse D. Retinoic acid-related orphan receptor-α is induced in the setting of DNA damage and promotes pulmonary emphysema. Am. J. Respir. Crit. Care Med. 2012;186(5):412-419. DOI 10.1164/rccm.201111-2023OC

Toyoshima M., Howie H.L., Imakura M., Walsh R.M., Annis J.E., Chang A.N., Frazier J., Chau B.N., Loboda A., Linsley P.S., Cleary M.A., Park J.R., Grandori C. Functional genomics identifies therapeutic targets for MYC-driven cancer. Proc. Natl Acad. Sci. USA. 2012;109(24):9545-9550. DOI 10.1073/pnas.1121119109

Wang Q., Maillard M., Schibler U., Burnier M., Gachon F. Cardiac hypertrophy, low blood pressure, and low aldosterone levels in mice devoid of the three circadian PAR bZip transcription factors DBP, HLF, and TEF. Am. J. Physiol. Regul. Integr. Comp. Physiol. 2010; 299(4):R1013-R1019. DOI 10.1152/ajpregu.00241.2010

Wang T., Yang P., Zhan Y., Xia L., Hua Z., Zhang J. Deletion of circadian gene Per1 alleviates acute ethanol-induced hepatotoxicity in mice. Toxicology. 2013;314(2/3):193-201. DOI 10.1016/j.tox.2013.09.009

Watanabe S., Ageta-Ishihara N., Nagatsu S., Takao K., Komine O., Endo F., Miyakawa T., Misawa H., Takahashi R., Kinoshita M., Yamanaka K. SIRT1 overexpression ameliorates a mouse model of SOD1-linked amyotrophic lateral sclerosis via HSF1/HSP70i chaperone system. Mol. Brain. 2014;7:62. DOI 10.1186/s13041-014-0062-1

Wilmet J.P., Tastet C., Desruelles E., Ziental-Gelus N., Blanckaert V., Hondermarck H., Le Bourhis X. Proteome changes induced by overexpression of the p75 neurotrophin receptor (p75NTR) in breast cancer cells. Int. J. Dev. Biol. 2011;55(7-9):801-809. DOI 10.1387/ijdb.113345jw

Wong V.C., Ko J.M., Qi R.Z., Li P.J., Wang L.D., Li J.L., Chan Y.P., Chan K.W., Stanbridge E.J., Lung M.L. Abrogated expression of DEC1 during oesophageal squamous cell carcinoma progression is age- and family history-related and significantly associated with lymph node metastasis. Br. J. Cancer. 2011;104(5):841-849. DOI 10.1038/bjc.2011.25

Yamasaki N., Miyazaki K., Nagamachi A., Koller R., Oda H., Miyazaki M., Sasaki T., Honda Z.I., Wolff L., Inaba T., Honda H. Identification of Zfp521/ZNF521 as a cooperative gene for E2A-HLF to develop acute B-lineage leukemia. Oncogene. 2010;29(13):1963-1975. DOI 10.1038/onc.2009.475

Yu S., Matsusue K., Kashireddy P., Cao W.Q., Yeldandi V., Yeldandi A.V., Rao M.S., Gonzalez F.J., Reddy J.K. Adipocyte-specific gene expression and adipogenic steatosis in the mouse liver due to peroxisome proliferator-activated receptor gamma1 (PPARgamma1) overexpression. J. Biol. Chem. 2003;278(1):498-505. DOI 10.1074/jbc.M210062200

Yumimoto K., Akiyoshi S., Ueo H., Sagara Y., Onoyama I., Ueo H., Ohno S., Mori M., Mimori K., Nakayama K.I. F-box protein FBXW7 inhibits cancer metastasis in a non-cell-autonomous manner. J. Clin. Invest. 2015;125(2):621-635. DOI 10.1172/JCI78782

Zerbino D.R., Wilder S.P., Johnson N., Juettemann T., Flicek P.R. The Ensembl regulatory build. Genome Biol. 2015;16:56. DOI 10.1186/s13059-015-0621-5

Zhao H., Zeng Z.L., Yang J., Jin Y., Qiu M.Z., Hu X.Y., Han J., Liu K.Y., Liao J.W., Xu R.H., Zou Q.F. Prognostic relevance of Period1 (Per1) and Period2 (Per2) expression in human gastric cancer. Int. J. Clin. Exp. Pathol. 2014;7(2):619-630.

Вавиловский журнал генетики и селекции. 2015;19(6):691-698
Doi 10.18699/VJ15.083

Биомедицинские и кандидатные SNP-маркеры для хронопатологий могут достоверно изменять сродство TATA-связывающего белка к промоторам генов человека

Д.А. Рассказов[1], Н.Л. Подколодный[1, 2, 3], О.А. Подколодная[1], Н.Н. Подколодная[1, 3], В.В. Суслов[1], Л.К. Савинкова[1], П.М. Пономаренко[4], М.П. Пономаренко[1, 3]

1 Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия 2 Федеральное государственное бюджетное учреждение науки Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия
3 Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия 4 Детский госпиталь Лос-Анджелеса, Университет Южной Калифорнии, США

Компьютерный анализ миллионов неаннотированных SnPs (Single nucleotide Polymorphisms) из проекта «1 000 геномов» может ускорить поиск биомедицинских SnP-маркеров. Анализ при помощи web-сервиса SnP_TATA_comparator SnPs сайтов связывания TAtA-связывающего белка (тBP) сочетали с поиском хронопатологий по ключевым словам так, чтобы биохимические маркеры хронопатологий соответствовали изменениям экспрессии генов, содержащих эти SnPs. Для промоторов 14 генов человека в районе [–70; –20] (район доказанных сайтов связывания TBP) были найдены биомедицинские и кандидатные SnP-маркеры нарушений циркадного ритма, которые могут достоверно (Z-тест) изменять сродство тBP к этим промоторам. В их числе: rs17231520, rs569033466 (хронопатологии печени); rs35036378 (хронопатология поведенческой активности); rs549858786 (хронопатология экспрессии *IL1B* при ревматоидном артрите); rs563207167, rs11557611, rs5505 (хронопатологии баланса «опухоль – хозяин», кровяного давления и репродуктивной системы); rs1143627 (циркадность диагностики и терапии биполярного расстройства); rs16887226, rs544850971 (неустойчивость к эндотоксинам из-за дисбаланса циркадной и иммунной систем); rs367732974, rs549591993 (циркадность обострений сердечной недостаточности); rs563763767 (циркадность случаев инфаркта миокарда); rs2276109, rs572527200 (циркадность приступов астмы); rs34223104, rs563558831 и rs10168 (циркадные оптимумы терапии метотрексатом и циклофосфамидом); rs397509430, rs33980857, rs34598529, rs33931746, rs33981098, rs34500389, rs63750953, rs281864525, rs35518301, rs34166473 (циркадность синдрома беспокойных ног и нейросенсорной тугоухости). Проверка этих 32 SNP-маркеров по медицинским стандартам может способствовать предиктивно-превентивной персонифицированной медицине.

Ключевые слова: TATA-связывающий белок (TBP); сайт связывания TBP; SNP; промотор; сродство TBP к промотору; значимость; патология; циркадный ритм; SNP-маркер; предиктивно-превентивная персонифицированная медицина.

Received 10.09.2015
Accepted for publication 12.10.2015

e-mail: pon@bionet.nsc.ru

Biomedical and candidate SNP markers of chronopathologies can significantly change affinity of TATA-binding protein for human gene promoters

D.A. Rasskazov[1], N.L. Podkolodnyy[1, 2, 3], O.A. Podkolodnaya[1], N.N. Podkolodnaya[1, 3], V.V. Suslov[1], L.K. Savinkova[1], P.M. Ponomarenko[4], M.P. Ponomarenko[1, 3]

1 Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
2 Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia
3 Novosibirsk State University, Novosibirsk, Russia
4 Children's Hospital Los Angeles, University of Southern California, USA

computational analysis of millions of unannotated SnPs from the 1000 Genomes Project may speed up the search for biomedical SnP markers. we combined the analysis of SnPs in the binding sites of тATA- binding protein (тBP) using a previously described web service (http://beehive.bionet.nsc.ru/cgi-bin/mgs/tatascan/start.pl) with a keyword search for biochemical markers of chronopathologies, which correspond to clinical manifestations of these SnPs. in the [–70; –20] region of promoters of 14 human genes (location of proven binding sites of TBP), we found 32 known and candidate SNP markers of circadian-rhythm disturbances, including rs17231520 and rs569033466 (both: risk of chronopathologies in liver); rs35036378 (behavioral chronoaberrations); rs549858786 (rheuma-toid arthritis with a chronoaberration of *IL1B* expression); rs563207167, rs11557611, and rs5505 (all three: chronopa-thologies of the tumor – host balance, blood pressure, and the reproductive system); rs1143627 (bipolar disorder with circadian dependence of diagnosis and treatment); rs16887226 and rs544850971 (both: lowered resistance to endotoxins because of the imbalance between the circa-dian and immune systems); rs367732974 and rs549591993 (both: circadian dependence of heart attacks); rs563763767 (circadian dependence of myocardial infarction); rs2276109 and rs572527200 (both: circadian dependence of asthma attacks); rs34223104, rs563558831, and rs10168 (circadian optima of treatment with methotrexate and cyclophosphamide); and rs397509430, rs33980857, rs34598529, rs33931746, rs33981098, rs34500389, rs63750953, rs281864525, rs35518301, and rs34166473 (all: neurosensory hearing loss and restless legs syndrome). For these SNPs, we evaluated α (significance) of changes in the affinity of TBP for promoters, where increased affinity corresponds to overexpression of the genes, and decreased affinity to deficient expression (Z-test). Verification of these 32 SnP markers according to clinical standards and proto-cols may advance the field of predictive preventive perso-nalized medicine.

Key words: тATA-binding protein (TBP); TBP-binding site; SnP; promoter; тBP-promoter affinity; statistical significance; pathology; circadian rhythm; SnP marker; predictive preventive personalized medicine.

HOW TO CITE THIS ARTICLE?

r asskazov D.A., Podkolodnyy n.l ., Podkolodnaya o .A., Podkolodnaya n.n., Suslov V.V., Savinkova l .K., Ponomarenko P.M., Ponoma-renko M.P. Biomedical and candidate SnP markers of chronopathologies can significantly change affinity of тАтА-binding protein for human gene promoters. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2015;19(6):691-698. Doi 10.18699/VJ15.083

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?

Рассказов Д.А., Подколодный Н.л., Подколодная О.А., Подколодная Н.Н., Суслов В.В., Савинкова л.К., Пономаренко П.М, Пономаренко МП. Биомедицинские и кандидатные SnP-маркеры для хронопатологий могут достоверно изменять сродство тАтА-связывающего белка к промоторам генов человека. Вавиловский журнал генетики и селекции. 2015;19(6):691-698. Doi 10.18699/VJ15.083

Суточная (циркадная) ритмика экспрессии выявлена у примерно 10^4 генов плацентарных (Zhang et al., 2014). Ритм около 24 ч задает центральный эндогенный водитель ритма: молекулярно-генетические осцилляторы нейронов супрахиазматических ядер гипоталамуса синхронизуются сутками через ретиногипоталамический путь и накладывают эту ритмику на осцилляторы периферийные – молекулярно-генетически идентичные, но работающие с собственной ритмикой в клетках органов, тканей и систем тканей. Далее осцилляторы ритмизуют множество генов через экспрессию тканеспецифичных транскрипционных факторов (краткодействующая регуляция) или ремоделинг хроматина (долгодействующая регуляция) (Padmanabhan et al., 2012; Eckel-Mahan et al., 2013). Подстройка периферийной осцилляции под общий циркадный ритм синхронизует различные системы организма. Поэтому десинхрония может отягчить/вызвать патологии непосредственно не взаимодействующих систем: аутоиммунные патологии могут быть следствием десинхронии иммунной защиты организма от экзотоксинов и систем вывода/метаболизма эндотоксинов-аналогов (Wang et al., 2015), хронофармакология находит циркадные оптимумы диагностики (Marckmann et al., 1993) и терапии (Ohdo et al., 1997; Gorbacheva et al., 2005).

В «догеномную эру» связывание SNPs с заболеваниями – редкая удача, в настоящее время оно является одной из целей проекта «1 000 геномов» (Delaneau et al., 2014), результаты которого dbSNP собирает и ранжирует SNPs по встречаемости (Sherry et al., 2001). Самые частые вносятся в референсный геном человека (hg19) как анцестральные, отражая в Ensembl (Zerbino et al., 2015) и GENCODE (Frankish et al., 2015). Минорные SNPs в генах, вовлеченных в данный патологический процесс, можно найти с помощью Web-сервиса «UCSC Genome Browser» (Haeussler et al., 2015), визуализирующего полногеномную карту. Последующее рутинное генотипирование их у множества пациентов выявляет среди минорных SNP биомедицинские маркеры, достоверно часто ассоциированные с данной патологией (Abbas et al., 2006). Общепринятый подход биоинформатиков – оценка сходства проекций миллионов неаннотированных SNPs «1 000 геномов» и тысяч биомедицинских SNP-маркеров на карты геномов и метаболомов, генов и их продуктов в норме, патологии или при терапии (Chen et al., 2014). Это ускоряет поиск кандидатных SNP-маркеров лишь при множестве уже найденных биомедицинских SNP-маркеров или широкой инвариантности дефектов (например, любое повреждение белка). Оба условия взаимосвязаны и выполняются для SNPs кодирующей части генов. Напротив, SNP-маркеров регуляторных районов генов описано мало из-за специфичности их патогенных проявлений (Zerbino et al., 2015). При этом примерно 10 % регуляторных SNP-маркеров (Савинкова и др., 2009) найдены в районе [–70; –20] п. о. от старта транскрипции, т. е. в районе сайтов связывания TATA-связывающего белка (TBP). Это отвечает экспериментально установленным фактам: 1) TBP необходим для инициации транскрипции (Martianov et al., 2002); 2) сродство TBP к промотору гена позитивно коррелирует с уровнем экспрессии этого гена (Mogno et al., 2010). Опыты по иммунопреципитации хроматина (ChIP) подтвердили, что у мыши большинство промоторов несут сайты связывания TBP (Choukrallah et al., 2012), что совпало с *in silico* прогнозом и выборочными *in vivo* тестами для человека (Yang et al., 2011). Поэтому для поиска кандидатных SNPs в промоторах мы предлагаем дополнить общепринятую оценку сходства положения SNP-маркеров и неаннотированных SNPs *in silico* прогнозом сродства «TBP/промотор» для этих SNPs (см. рисунок). Web-сервис SNP_TATA_Comparator, представленный нами ранее (Рассказов и др., 2013), подвел итог циклу работ по поиску данных о SNPs сайтов связывания TBP в промоторах человека (Савинкова и др., 2009), прогнозу связи «SNP-патология» (Пономаренко и др., 2009) и проверке таких прогнозов *in vitro*: 1) в «реальном вре-

мени» (Аркова и др., 2014); 2) в равновесных (Savinkova et al., 2013) и 3) в неравновесных (Drachkova et al., 2014) условиях.

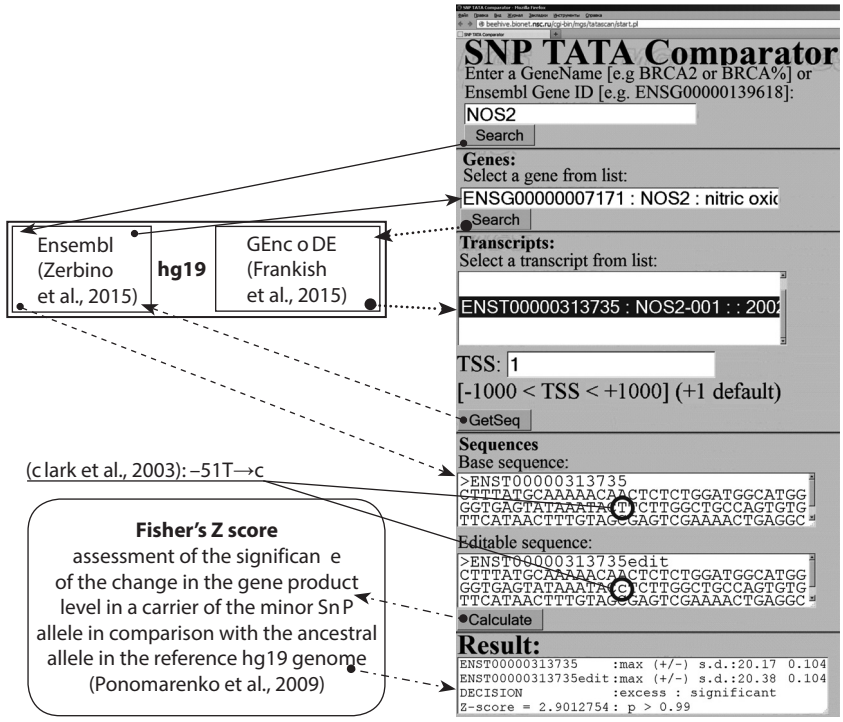
В данной работе был применен этот Web-сервис к неаннотированным SNPs сайтов TBP-связывания, сходным с биомедицинскими SNP-маркерами по способности менять сродство TBP к тем же промоторам. Проведен поиск статей о биохимических маркерах хронопатологий, идентичных по изменениям экспрессии генов с этими SNPs, и обнаружено 32 кандидатных SNP-маркера, способных влиять на хронопатогенез путем изменения сродства TBP к промоторам 14 генов человека. Их проверка по медицинским стандартам может быть полезной для предиктивно-превентивной персонифицированной медицины.

Материалы и методы

Анализировали фрагменты ДНК проксимальных промоторов длиной 90 п.о. для анцестральных (hg19) и 89 минорных (min) вариантов 14 генов человека из обзора Савинковой с коллегами (2009). Анцестральные варианты брали из Ensembl (Zerbino et al., 2015) и GENECODE (Frankish et al., 2015), минорные делали из анцестральных, как показано на рисунке, в окнах «Base sequence» и «Editable sequence». По каждой ДНК мы оценили сродство TBP к промотору, $-\ln(K_D) \pm \delta$, в ln-единицах (окно «Result») и в nM (табл.) (здесь: K_D и δ – константа диссоциации комплекса TBP/промотор и ее стандартное отклонение). Сравнили сродство TBP к аллелям hg19 и min промотора в рамках Z-теста в составе стандартного статистического пакета R (Waardenberg et al., 2015), $Z = abs[\ln(K_D^{(min)}/K_D^{(hg19)}]/[\delta^2_{(min)} + \delta^2_{(hg19)}]^{1/2}$. Была предсказана (строка «Decision») при значимом росте сродства TBP к промотору суперэкспрессия гена, при спаде – дефицитная экспрессия (Mogno et al., 2010).

результаты

Результаты для всех 32 найденных – описанных в литературе и кандидатных – биомедицинских SNP-маркеров хронопатологий даны на рисунке и в таблице. Рассмотрим их детально на примере гена индуцибельной NO-синтазы *NOS2*. В его промоторе



In silico search for a candidate SnP marker for chronopathologies with the web-service SnP_TATA_c omparator (r asskazov et al., 2013) by the example of the biomedical SnP –51T→c in the promoter of the *NOS2* gene, which marks chronopathologies observed in epilepsy (Hofstra, de weerd, 2009).

замена g**tataaata**c$(T \rightarrow C)_{-51}$cttggctgc (далее: –51T→C, выделено жирным – TATA-бокс, канонический сайт TBP-связывания) описана Clark с коллегами (2003) как SNP-маркер устойчивости к малярии и риска эпилепсии, при которой наблюдают нарушения циркадных ритмов сердечно-сосудистой системы, сна – бодрствования, температуры тела и уровня мелатонина, кортизола и гормона роста (Hofstra, de Weerd, 2009). Сродство TBP к минорному аллелю –51T было 20,17±0,10 ln ед. (в табл. – 2 nM) значимо ($Z = 2{,}9$, $\alpha < 0{,}01$) выше такового для анцестрального аллеля –51C: 19,40±0,10 ln ед. (1 nM). Согласно опыту (Mogno et al., 2010), росту сродства TBP к промотору соответствуют суперэкспрессия *NOS2* и избыток NO у пациента с –51T (табл.: «↑»).

Это согласуется с клиническими данными роста уровня NO как компонента врожденного иммунитета при устойчивости к малярии (Clark et al., 2003) и как нейромедиатора при эпилепсии (Gonzalez-Martinez et al., 2009) и хронопатологиях (Hofstra, de Weerd, 2009). Мы нашли клинические наблюдения (Kaya et al., 2004) роста уровня NO как биохимического маркера ремиссии панического расстройства с циркадностью симптомов и поэтому прогнозируем SNP *NOS2*:–51T→C как кандидатный SNP-маркер этой хронопатологии.

Для остальных 13 генов человека мы получили аналогичные результаты (таблица), описание которых можно найти в Дополнительных материалах[1].

Обсуждение

Связывание TBP обязательно перед стартом транскрипции любой мРНК (Martianov et al., 2002). Было показано, что среди 68 биомедицинских и кандидатных SNP-маркеров, нарушающих TBP-связывание (Ponomarenko et al., 2015) есть 32 SNP-маркера для хронопатологий. На основе биоинформатического анализа TBP-связывания были подтверждены ранее описанные SNP-маркеры хронопатологий и болезней с циркадным оптимумом терапии и предложены кандидатные SNP-маркеры и гипотезы их фенотипического

[1] Дополнительные материалы см. в Приложении 3 по адресу: http://www.bionet.nsc.ru/vogis/download/pict-2015-12/appx3.pdf

In silico prediction of changes in the TBP/promoter affin y by biomedical and candidate SnP markers of human chronopathologies

Gene	dbSnP or reference	5′-flan	**hg19** / *min*	3′-flan	$K_{D'}$, nM **hg19** / *min*	Δ	Z	α	Known (literature) and hypothetical (this work) disorders associated with the circadian rhythm	r eference(s)
NOS2	c lark et al., 2003	gtataaatac	**t** / *c*	tcttggctgc	**2** / *1*	↑	3	10^{-2}	The risk of epilepsy in which the whole circadian rhythm can be completely changed; resistance to malaria; and, also *hypothetically*, the phase of remission of panic disorder whose circadian dependence can be reduced	Gonzalez-Martinez et al., 2009; Hofstra, de weerd, 2009; c lark et al., 2003; Эта работа; Kaya et al., 2003
APOA1	Matsunaga et al., 1999	tgcagacata	**a** / *c*	ataggccctg	**3** / *4*	↓	5	10^{-7}	Hypoalphalipoproteinemia breaks the expression of circadian oscillator genes in the liver; hematuria, fatty liver; obesity. *Hypothetically*: higher risk of postprandial atherosclerosis development in diabetes	Matsunaga et al., 1999; Gabas-r ivera et al., 2013; This work; o ka et al., 2007; Hirayama et al., 2010
CETP	Plengpanich et al., 2011	cgtgggggct	**18 п.о.** / –	gggctccagg	**4** / *7*	↓	7	10^{-7}	Hyperalphalipoproteinemia reduces the risk of postprandial athero-sclerosis development in diabetes	Hirayama et al., 2010; Plengpanich et al., 2011
	rs17231520	ggggctgggc	**g** / *a*	gacatacata	**4** / *2*	↑	10	10^{-7}	Additionally, *hypothetically*, hypoalphalipoproteinemia breaks the expression of circadian oscillator genes in the liver, which increases the risk of atherosclerosis in addition to the higher risk of postprandial atherosclerosis development in diabetes	This work
	rs569033466	atacatatac	**g** / *a*	ggctccaggc	**4** / *3*	↑	4	10^{-3}		
ESR2	rs35036378	cctctcggtc	**t** / *g*	ttaaaaggaa	**6** / *8*	↓	5	10^{-3}	A female mice showed disrupted circadian rhythms of daily behavioral activity, as well as a pronounced circadian optimum of treatment with tamoxifen against ESr 2-deficie t tumor pT1	Philips et al., 2012; Sieuwerts et al., 2010; r oyston et al., 2014; Binkhorst et al., 2015
DHFR	rs10168	ctgcacaaat	**g** / *a*	gggacgaggg	**15** / *9*	↑	9	10^{-7}	A pronounced circadian optimum of treatment with tamoxifen against leukemia in children with resistance to the drug	o hdo et al., 1997; Al-Shakfa et al., 2009
CYP2B6	rs34223104	gatgaaattt	**t** / *c*	ataacagggt	**4** / *10*	↓	15	10^{-7}	Better response to cyclophosphamide when administered in a circadian mode	Gorbacheva et al., 2005; Zukunft et al., 2005
	rs563558831	tgaaatttta	**t** / *c*	aacagggtgc	**4** / *10*	↓	13	10^{-2}	*Hypothetically:* better response to cyclophosphamide when administered in a circadian mode	This work
IL1B	rs1143627	ttttgaaagc	**c** / *t*	ataaaaacag	**5** / *2*	↑	15	10^{-7}	c ircadian optimum for diagnosis and treatment of the major depressive disorder, which may be shifted in cases of fat and carbohydrate-rich diets; gastric ulcer; chronic gastritis; non-small cell lung cancer; gastric cancer; hepatocellular carcinoma; and Graves' disease. *Hypothetically*, a diet-dependent circadian optimum for diagnosis and treatment of bipolar disorder.	El-o mar et al., 2000; w ang et al., 2003; Hayashi et al., 2009; Martinez-c arrillo et al., 2010; w u et al., 2010; Borkowska et al., 2011; Ávila Moraes et al., 2013; Pivovarova et al., 2015 This work; c arter, 2007
	rs549858786	tgaaagccat	**a** / *t*	aaaacagcga	**7** / *5*	↓	8	10^{-7}	*Hypothetically*, rheumatoid arthritis, which can, in turn, break the circadian rhythm of the expression of the *IL1B* gene	This work; c hikanza et al., 1992; Yamazaki et al., 2012
MMP12	rs2276109	gatatcaact	**a** / *g*	tgagtcactc	**11** / *14*	↓	3	10^{-2}	l ower risk of systemic sclerosis and psoriasis, as well as asthma, whose symptoms worsen in a circadian way at night and in early morning	Hunninghake et al., 2009; Manetti et al., 2010; **Starodubtseva et al., 2011;** Durrington et al., 2014
	rs572527200	gatgatatca	**a** / *g*	ctatgagtca	**11** / *14*	↓	3	10^{-2}	*Hypothetically*, lower risk of systemic sclerosis and psoriasis as well as asthma, whose symptoms worsen in a circadian way at night	This work
INS	rs5505	agatcactgt	**c** / *t*	cttctgccat	**53** / *44*	↑	4	10^{-3}	r isk of neonatal diabetes, as well as risk of hyperinsulinemia, which disturbs the circadian rhythms of the reproductive system, blood pressure and the "tumor – host" balance	Bianchi et al., 1997; Sherry et al., 2001; Blask et al., 2014; Mereness et al., 2015

End of table

Gene	dbSnP or reference	5′-flan	**hg19** / *min*	3′-flan	K_D, nM **hg19** / *min*	Δ	Z	α	Known (literature) and hypothetical (this work) disorders associated with the circadian rhythm	r eference(s)
INS	rs563207167	tcagccctgc	**c** / *t*	tgtctcccag	**53** / *44*	↑	4	10^{-3}	*Hypothetically*, risk of neonatal diabetes, as well as risk of hyperinsulinemia, which disturbs the circadian rhythms of the reproductive system, blood pressure and the "tumor – host" balance	This work
	rs11557611	gatcactgtc	**c** / *t*	ttctgccatg	**53** / *60*	↓	2	0,05	*Hypothetically*: risks of hypothalamic amenorrhea	This work ; l aughlin et al., 1998
F7	Kavlie et al., 2003	ccttggaggc	**a** / *c*	gagaactttg	**53** / *62*	↓	3	10^{-2}	l iability to moderate bleeding, whose symptoms worsen in a circadian way in winter and with frequent jet lags	Kavlie et al., 2003; c olognesi et al., 2007
	rs367732974	aactttgccc	**g** / *a*	tcagtcccat	**53** / *47*	↑	2	0,05	*Hypothetically*, risks of (i) heart failure events with circadian preference, (ii) postprandial thrombogenesis, and (iii) transformation of primary colorectal cancer into metastatic	This work ; **c arvalho de Sousa** et al., 1989; Marckmann et al., 1993; Tang et al., 2009
	rs549591993	gcccgtcagt	**c** / *a*	ccatggggaa	**53** / *25*	↑	13	10^{-7}		
F3	rs563763767	ccctttatag	**c** / *t*	gcgcggggca	**3** / *2*	↑	6	10^{-7}	r isks of venous thromboembolism and myocardial infarction with circadian preference in the early morning in the elderly	Arnaud et al., 2000; Haus, 2007; o ishi et al., 2011
HBB	rs397509430	gggctgggca	**t** / –	atacaacagt	**5** / *29*	↓	34	10^{-7}	Elevated resistance to malaria, c ooley's anemia (β-thalassemia), which increases the risk of restless leg syndrome, whose symptoms worsen in a circadian way. r isk of sensorineural hearing loss as a complication of treatment of thalassemia with deferoxamine	Bannerman et al., 1986; Sun et al., 1991; Sherry et al., 2001; Thio et al., 2008; Unger et al., 2009
	rs33980857	gggctgggca	**t** / *a, g, c*	atacaacagt	**5** / *21*	↓	27	10^{-7}		
	rs34598529	ggctgggcat	**a** / *g*	aaagtcaggg	**5** / *18*	↓	24	10^{-7}		
	rs33931746	gctgggcata	**a** / *g, c*	aagtcagggc	**5** / *11*	↓	14	10^{-7}		
	rs33981098	agggctgggc	**a** / *g, c*	taaaagtcag	**5** / *9*	↓	10	10^{-7}		
	rs34500389	cagggctggg	**c** / *a, t, g*	ataaaagtca	**5** / *6*	↓	3	10^{-2}		
	rs63750953	ctgggcataa	**aa** / –	gtcagggcag	**5** / *8*	↓	9	10^{-7}	*Hypothetically*, elevated resistance to malaria, c ooley's anemia (β-thalassemia), which increases the risk of restless legs syndrome, whose symptoms worsen circadian at night. r isk of sensorineural hearing loss as a complication of deferoxamine use	This work
	rs281864525	tgggcataaa	**a** / *c*	gtcagggcag	**5** / *7*	↓	*7*	10^{-7}		
HBD	rs35518301	caggaccagc	**a** / *g*	taaaaggcag	**4** / *8*	↓	11	10^{-7}	Elevated resistance to malaria, c ooley's anemia (β-thalassemia) with circadian restless leg syndrome and sensorineural hearing loss as a complication of deferoxamine use	Bannerman et al., 1986; Sun et al., 1991; Thio et al., 2008; Unger et al., 2009
	rs34166473	aggaccagca	**t** / *c*	aaaaggcagg	**4** / *8*	↓	18	10^{-7}	*Hypothetically*, elevated resistance to malaria, c ooley's anemia (β-thalassemia), with circadian restless leg syndrome. r isk of sensorineural hearing loss as a complication of deferoxamine use	This work
StAR	rs16887226	cagccttcag	**c** / *t*	gggggacatt	**10** / *10*	=	0	–	Hypertension in diabetes (StAr deficien y, EMSA: distorted binding of an unknown transcription factor but not of TBP). *Hypothetically*, lower resistance to endotoxins due to a lack of a mediator (StAr) between the circadian and immune systems	c asal et al., 2006 This work ; w ang et al., 2015
	rs544850971	tcagcggggg	**a** / *g*	catttaagac	**10** / *12*	↓	5	10^{-2}	*Hypothetically*, lower resistance to exotoxins and hypertension in diabetes	This work

Designations: **hg19**, ancestral allele; *min*, minor allele; K_D, an estimate (r asskazov et al., 2013) of the dissociation constant K_D of the TBP-promoter complex corresponding to *in vitro* conditions (Savinkova et al., 2013); Δ, change: (↑), excess; (↓), deficien y; Z, Z-score; $\alpha \equiv 1 - p$, significan e (where p is the probability rate shown in the r esult window of the figu e); EMSA, electrophoretic mobility shift assay.

эффекта (появление хронопатологии, выраженность ответа на терапию при циркадной ритмике, см. таблицу). Следовательно, анализ связи «TBP–промотор» может дать врачам меру обоснованности поиска SNP-маркеров среди SNPs в области проксимальных промоторов. Доказать же значимость найденных SNP-маркеров может лишь достоверно высокая их частота у пациентов с учетом возрастного, гендерного и этнического состава их когорт, климата, экологии, условий и стиля их жизни (Yoo et al., 2014). Спектр связей SNPs сайтов связывания TBP с хронопатологиями включает как нарушения циркадности (e. g., rs35036378) и риск заболеваний (e. g., rs549858786), так и внутрисуточные (e. g., rs397509430), сезонные (e. g., *F7*: –33A→C), гендерные (e. g., rs11557611) и возрастные (e. g., rs563763767) особенности патологий, циркадные оптимумы диагностики (e. g., *NOS2*: –51T→C) и терапии (e. g., rs34223104), их сдвиг диетой пациента (e. g., rs1143627), дисбаланс систем в организме (e. g., rs16887226), включая «опухоль–хозяин» (e. g., rs5505), риск ряда осложнений патологий (e. g., rs17231520). Это соответствует нашему каталогу (Подколодная, Подколодный, 2013) биомедицинских SNP-маркеров в генах циркадного ритма, установленных чаще в регуляторных районах и в связи с циркадными особенностями неоплазм (e. g., нарушен суточный ритм при раке молочной железы (Cash et al., 2015)), психических (e. g., синдром Смит–Магениса (Mullegama et al., 2015)), нейродегенеративных (e. g., болезнь Альцгеймера (Wu et al., 2007)), аутоиммунных (e. g., ревматоидный артрит (Kouri et al., 2013)), воспалительных, метаболических заболеваний и старения.

Наконец, статистическая значимость кандидатных SNP-маркеров варьировала от высокой (e. g., $\alpha < 10^{-7}$ для rs10168) до пороговой (e. g., $\alpha < 0{,}05$ для rs549591993). Поэтому необходима их проверка по медицинским стандартам, для которой биоинформатический анализ может дать пределы точности. В таблице даны оценки K_D (в nM) для комплексов TBP/ДНК в условиях *in vitro* (Savinkova et al., 2013). Они были от 1 до 60 nM при их разнице для вариантов SNP менее 2 % этого диапазона, это вне рамок точности их измерений, ±10 %, без их оценок *a priori*. Поэтому оценки K_D в таблице – это необходимая для проверки по медицинским стандартам часть кандидатного SNP-маркера хронопатологий.

Верификация предсказанных кандидатных SNP-маркеров сайтов связывания TBP в промоторах генов человека в рамках медицинских стандартов и протоколов будет способствовать предиктивно-превентивной персонализированной медицине.

Acknowledgments

RD, PNL, PO, and PNN acknowledge the support of Web service design by the Russian Science Foundation, project 14-24-00123. PM acknowledges the support of data collection by the Russian Foundation for Basic Research, project 14-04-00485. SL acknowledges the support of data analysis by project VI.58.1.2.

Conflict of interest

The authors declare no conflict of interest.

References

Abbas A., Lechevrel M., Sichel F. Identification of new single nucleotid polymorphisms (SNP) in alcohol dehydrogenase class IV ADH7 gene within a French population. Arch. Toxicol. 2006;80(4):201-205. DOI 10.1007/s00204-005-0031-7

Al-Shakfa F., Dulucq S., Brukner I., Milacic I., Ansari M., Beaulieu P., Moghrabi A., Laverdiere C., Sallan S., Silverman L.B., Neuberg D., Kutok J.L., Sinnett D., Krajinovic M. DNA variants in region for noncoding interfering transcript of dihydrofolate reductase gene and outcome in childhood acute lymphoblastic leukemia. Clin. Cancer Res. 2009;15(22):6931-6938. DOI 10.1158/1078-0432.CCR-09-0641

Arkova O.V., Kuznetsov N.A., Fedorova O.S., Kolchanov N.A., Savinkova L.K. Real-time interaction between TBP and the TATA box of the human triosephosphate isomerase gene promoter in the norm and pathology. Acta Naturae. 2014;6(2):36-40.

Arnaud E., Barbalat V., Nicaud V., Cambien F., Evans A., Morrison C., Arveiler D., Luc G., Ruidavets J.B., Emmerich J., Fiessinger J.N., Aiach M. Polymorphisms in the 5' regulatory region of the tissue factor gene and the risk of myocardial infarction and venous thromboembolism: the ECTIM and PATHROS studies. Etude Cas-Temoins de l'Infarctus du Myocarde. Paris Thrombosis case-control Study. Arterioscler. Thromb. Vasc. Biol. 2000;20(3):892-898.

Ávila Moraes C., Cambras T., Diez-Noguera A., Schimitt R., Dantas G., Levandovski R., Hidalgo M.P. A new chronobiological approach to discriminate between acute and chronic depression using peripheral temperature, rest-activity, and light exposure parameters. BMC Psychiatry. 2013;13:77. DOI 10.1186/1471-244X-13-77

Bannerman R.M., Garrick L.M., Rusnak-Smalley P., Hoke J.E., Edwards J.A. Hemoglobin deficit: an inherited hypochromic anemia in the mouse. Proc. Soc. Exp. Biol. Med. 1986;182(1):52-57.

Bianchi S., Bigazzi R., Nenci R., Campese V. Hyperinsulinemia, circadian variation of blood pressure and end-organ damage in hypertension. J. Nephrol. 1997;10(6):325-333.

Binkhorst L., Kloth J.S., de Wit A.S., de Bruijn P., Lam M.H., Chaves I., Burger H., van Alphen R.J., Hamberg P., van Schaik R.H., Jager A., Koch B.C., Wiemer E.A., van Gelder T., van der Horst G.T., Mathijssen R.H. Circadian variation in tamoxifen pharmacokinetics in mice and breast cancer patients. Breast Cancer Res. Treat. 2015;152(1):119-128. DOI 10.1007/s10549-015-3452-x

Blask D., Dauchy R., Dauchy E., Mao L., Hill S.M., Greene M.W., Belancio V.P., Sauer L.A., Davidson L. Light exposure at night disrupts host/cancer circadian regulatory dynamics: impact on the Warburg effect, lipid signaling and tumor growth prevention. PLoS One. 2014;9(8):e102776. DOI 10.1371/journal.pone.0102776

Borkowska P., Kucia K., Rzezniczek S., Paul-Samojedny M., Kowalczyk M., Owczarek A., Suchanek R., Medrala T., Kowalski J. Interleukin-1beta promoter (-31T/C and –511C/T) polymorphisms in major recurrent depression. J. Mol. Neurosci. 2011;44(1);12-16. DOI 10.1007/s12031-011-9507-5

Carter C.J. Multiple genes and factors associated with bipolar disorder converge on growth factor and stress activated kinase pathways controlling translation initiation: implications for oligodendrocyte viability. Neurochem. Int. 2007;50(3):461-490. DOI 10.1016/j.neuint.2006.11.009

Carvalho de Sousa J., Bruckert E., Giral P., Soria C., Chapman J., Truffert J., Dairou F., De Gennes J.L., Caen J.P. Coagulation factor VII and plasma triglycerides. Decreased catabolism as a possible mechanism of factor VII hyperactivity. Haemostasis. 1989;19(3):125-130.

Casal A., Sinclair V., Capponi A.M., Nicod J., Huynh-Do U., Ferrari P. A novel mutation in the steroidogenic acute regulatory protein gene promoter leading to reduced promoter activity. J. Mol. Endocrinol. 2006;37(1):71-80. DOI 10.1677/jme.1.02082

Cash E., Sephton S.E., Chagpar A.B., Spiegel D., Rebholz W.N., Zimmaro L.A., Tillie J.M., Dhabhar F.S. Circadian disruption and biomarkers of tumor progression in breast cancer patients awaiting surgery. Brain Behav. Immun. 2015;48:102-114. DOI 10.1016/j.bbi.2015.02.017

Chen C.Y., Chang I.S., Hsiung C.A., Wasserman W.W. On the identification of potential regulatory variants within genome wide asso-

ciation candidate SNP sets. BMC Med. Genomics. 2014;11(7):34. DOI 10.1186/1755-8794-7-34

Chikanza I.C., Petrou P., Kingsley G., Chrousos G., Panayi G.S. Defective hypothalamic response to immune and inflammatory stimuli in patients with rheumatoid arthritis. Arthritis Rheum. 1992;35(11): 1281-1288.

Choukrallah M.A., Kobi D., Martianov I., Pijnappel W.W., Mischerikow N., Ye T., Heck A.J., Timmers H.T., Davidson I. Interconversion between active and inactive TATA-binding protein transcription complexes in the mouse genome. Nucl. Acids Res. 2012;40(4): 1446-1459. DOI 10.1093/nar/gkr802

Clark I., Rockett K.A., Burgner D. Genes, nitric oxide and malaria in African children. Trends Parasitol. 2003;19(8):335-337. DOI 10.1016/S1471-4922(03)00147-8

Colognesi I., Pasquali V., Foa A., Renzi P., Bernardi F., Bertolucci C., Pinotti M. Temporal variations of coagulation factor VII activity in mice are influenced by lighting regime. Chronobiol. Int. 2007;24: 305-313. DOI 10.1080/07420520701282307

Delaneau O., Marchini J.; 1000 Genomes Project Consortium. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat. Commun. 2014;5:3934. DOI 10.1038/ncomms4934

Drachkova I., Savinkova L., Arshinova T., Ponomarenko M., Peltek S., Kolchanov N. The mechanism by which TATA-box polymorphisms associated with human hereditary diseases influence interactions with the TATA-binding protein. Hum. Mutat. 2014;35(5):601-608. DOI 10.1002/humu.22535.

Durrington H.J., Farrow S.N., Loudon A.S., Ray D.W. The circadian clock and asthma. Thorax. 2014;69(1):90-92. DOI 10.1136/thoraxjnl-2013-203482

Eckel-Mahan K., Sassone-Corsi P. Epigenetic regulation of the molecular clockwork. Prog. Mol. Biol. Transl. Sci. 2013;119:29-50. DOI 10.1016/B978-0-12-396971-2.00002-6

El-Omar E.M., Carrington M., Chow W.H., McColl K.E., Bream J.H., Young H.A., Herrera J., Lissowska J., Yuan C.C., Rothman N., Lanyon G., Martin M., Fraumeni J.F. Jr, Rabkin C.S. Interleukin-1 polymorphisms associated with increased risk of gastric cancer. Nature. 2000;404(6776):398-402. DOI 10.1038/35006081

Frankish A., Uszczynska B., Ritchie G.R., Gonzalez J.M., Pervouchine D., Petryszak R., Mudge J., Fonseca N., Brazma A., Guigo R., Harrow J. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. BMC Genomics. 2015;16(Suppl. 8):S2. DOI 10.1186/1471-2164-16-S8-S2

Gabas-Rivera C., Martinez-Beamonte R., Rios J.L., Navarro M.A., Surra J.C., Arnal C., Rodriguez-Yoldi M., Osada J. Dietary oleanolic acid mediates circadian clock gene expression in liver independently of diet and animal model but requires apolipoprotein A1. J. Nutr. Biochem. 2013;24:2100-2109. DOI 10.1016/j.jnutbio.2013.07.010

Gonzalez-Martinez J.A., Moddel G., Ying Z., Prayson R.A., Bingaman W.E., Najm I.M. Neuronal nitric oxide synthase expression in resected epileptic dysplastic neocortex. J. Neurosurg. 2009;110(2): 343-349. DOI 10.3171/2008.6.17608

Gorbacheva V., Kondratov R.V., Zhang R., Cherukuri S., Gudkov A.V., Takahashi J.S., Antoch M.P. Circadian sensitivity to the chemotherapeutic agent cyclophosphamide depends on the functional status of the CLOCK/BMAL1 transactivation complex. Proc. Natl Acad. Sci. USA. 2005;102(10):3407-3412. DOI 10.1073/pnas.0409897102

Haeussler M., Raney B.J., Hinrichs A.S., Clawson H., Zweig A.S., Karolchik D., Casper J., Speir M.L., Haussler D., Kent W.J. Navigating protected genomics data with UCSC Genome Browser in a box. Bioinformatics. 2015;31(5):764-766. DOI 10.1093/bioinformatics/btu712

Haus E. Chronobiology of hemostasis and inferences for the chronotherapy of coagulation disorders and thrombosis prevention. Adv. Drug Deliv. Rev. 2007;59(9/10):966-984. DOI 10.1016/j.addr.2006.11.002

Hayashi F., Watanabe M., Nanba T., Inoue N., Akamizu T., Iwatani Y. Association of the -31C/T functional polymorphism in the interleukin-1beta gene with the intractability of Graves' disease and the proportion of T helper type 17 cells. Clin. Exp. Immunol. 2009;158(3): 281-286. DOI 10.1111/j.1365-2249.2009.04034.x

Hirayama S., Soda S., Ito Y., Matsui H., Ueno T., Fukushima Y., Ohmura H., Hanyu O., Aizawa Y., Miida T. Circadian change of serum concentration of small dense LDL-cholesterol in type 2 diabetic patients. Clin. Chim. Acta. 2010;411(3/4):253-257. DOI 10.1016/j.cca. 2009.11.017

Hofstra W.A., de Weerd A.W. The circadian rhythm and its interaction with human epilepsy: a review of literature. Sleep Med. Rev. 2009;13(6)413-420. DOI 10.1016/j.smrv.2009.01.002

Hunninghake G.M., Cho M.H., Tesfaigzi Y., Soto-Quiros M.E., Avila L., Lasky-Su J., Stidley C., Melen E., Soderhall C., Hallberg J., Kull I., Kere J., Svartengren M., Pershagen G., Wickman M., Lange C., Demeo D.L., Hersh C.P., Klanderman B.J., Raby B. A., Sparrow D., Shapiro S.D., Silverman E.K., Litonjua A.A., Weiss S. T., Celedon J. C. MMP12, lung function, and COPD in high-risk populations. N. Engl. J. Med. 2009;361(27):2599-2608. DOI 10.1056/ NEJMoa0904006

Kavlie A., Hiltunen L., Rasi V., Prydz H. Two novel mutations in the human coagulation factor VII promoter. Thromb. Haemost. 2003; 90(2):194-205. DOI 10.1160/TH02-09-0050

Kaya B., Unal S., Karabulut A.B., Turkoz Y. Altered diurnal variation of nitric oxide production in patients with panic disorder. Tohoku J. Exp. Med. 2004;204(2):147-154. DOI 10.1620/tjem.204.147

Kouri V.P., Olkkonen J., Kaivosoja E., Ainola M., Juhila J., Hovatta I., Konttinen Y.T., Mandelin J. Circadian timekeeping is disturbed in rheumatoid arthritis at molecular level. PLoS One. 2013;8(1): e54049. DOI 10.1371/journal.pone.0054049

Laughlin G.A., Dominguez C.E., Yen S.S. Nutritional and endocrine-metabolic aberrations in women with functional hypothalamic amenorrhea. J. Clin. Endocrinol. Metab. 1998;83(1):25-32. DOI http://dx.doi.org/10.1210/jcem.83.1.4502

Manetti M., Ibba-Manneschi L., Fatini C., Guiducci S., Cuomo G., Bonino C., Bazzichi L., Liakouli V., Giacomelli R., Abbate R., Bombardieri S., Montecucco C., Valentini G., Matucci-Cerinic M. Association of a functional polymorphism in the matrix metalloproteinase-12 promoter region with systemic sclerosis in an Italian population. J. Rheumatol. 2010;37(9):1852-1857. DOI 10.3899/jrheum. 100237

Marckmann P., Sandstrom B., Jespersen J. Dietary effects on circadian fluctuation in human blood coagulation factor VII and fibrinolysis. Atherosclerosis. 1993;101(2):225-234. DOI 10.1016/0021-9150(93)90119-F

Martianov I., Viville S., Davidson I. RNA polymerase II transcription in murine cells lacking the TATA binding protein. Science. 2002;298(5595):1036-1039. DOI 10.1126/science.1076327

Martinez-Carrillo D.N., Garza-Gonzalez E., Betancourt-Linares R., Monico-Manzano T., Antunez-Rivera C., Roman-Roman A., Flores-Alfaro E., Illades-Aguiar B., Fernandez-Tilapa G. Association of IL1B -511C/-31T haplotype and Helicobacter pylori vacA genotypes with gastric ulcer and chronic gastritis. BMC Gastroenterol. 2010;10:126. DOI 10.1186/1471-230X-10-126

Matsunaga A., Sasaki J., Han H., Huang W., Kugi M., Koga T., Ichiki S., Shinkawa T., Arakawa K. Compound heterozygosity for an apolipoprotein A1 gene promoter mutation and a structural nonsense mutation with apolipoprotein A1 deficiency. Arterioscler. Thromb. Vasc. Biol. 1999;19(2):348-355. DOI 10.1161/01.ATV.19.2.348

Mereness A.L., Murphy Z.C., Sellix M.T. Developmental programming by androgen affects the circadian timing system in female mice. Biol. Reprod. 2015;92(4):88. DOI 10.1095/biolreprod.114.126409

Mogno I., Vallania F., Mitra R., Cohen B. TATA is a modular component of synthetic promoters. Genome Res. 2010;20(10):1391-1397. DOI 10.1101/gr.106732.110

Mullegama S.V., Pugliesi L., Burns B., Shah Z., Tahir R., Gu Y., Nelson D.L., Elsea S.H. MBD5 haploinsufficiency is associated with sleep disturbance and disrupts circadian pathways common to Smith-Magenis and fragile X syndromes. Eur. J. Hum. Genet. 2015;23(6): 781-789. DOI 10.1038/ejhg.2014.200

Ohdo S., Inoue K., Yukawa E., Higuchi S., Nakano S., Ogawa N. Chronotoxicity of methotrexate in mice and its relation to circadian

rhythm of DNA synthesis and pharmacokinetics. Jpn. J. Pharmacol. 1997;75(3):283-290. DOI 10.1254/jjp.75.283

Oishi K., Koyanagi S., Ohkura N. Circadian mRNA expression of coagulation and fibrinolytic factors is organ-dependently disrupted in aged mice. Exp. Gerontol. 2011;46(12):994-999. DOI 10.1016/j.exger.2011.09.003

Oka K., Belalcazar L.M., Dieker C., Nour E.A., Nuno-Gonzalez P., Paul A., Cormier S., Shin J.K., Finegold M., Chan L. Sustained phenotypic correction in a mouse model of hypoalphalipoproteinemia with a helper-dependent adenovirus vector. Gene Ther. 2007;14(3):191-202. DOI 10.1038/sj.gt.3302819

Padmanabhan K., Robles M.S., Westerling T., Weitz C.J. Feedback regulation of transcriptional termination by the mammalian circadian clock PERIOD complex. Science. 2012;337(6094)599-602. DOI 10.1126/science.1221592

Philips S., Richter A., Oesterreich S., Rae J.M., Flockhart D.A., Perumal N.B., Skaar T.C. Functional characterization of a genetic polymorphism in the promoter of the ESR2 gene. Horm. Cancer. 2012;3(1/2):37-43. DOI 10.1007/s12672-011-0086-2

Pivovarova O., Jurchott K., Rudovich N., Hornemann S., Ye L., Mockel S., Murahovschi V., Kessler K., Seltmann A.C., Maser-Gluth C., Mazuch J., Kruse M., Busjahn A., Kramer A., Pfeiffer A.F. Changes of dietary fat and carbohydrate content alter central and peripheral clock in humans. J. Clin. Endocrinol. Metab. 2015;100(6):2291-2302. DOI 10.1210/jc.2014-3868

Plengpanich W., Le Goff W., Poolsuk S., Julia Z., Guerin M., Khovidhunkit W. CETP deficiency due to a novel mutation in the CETP gene promoter and its effect on cholesterol efflux and selective uptake into hepatocytes. Atherosclerosis. 2011;216(2):370-373. DOI 10.1016/j.atherosclerosis.2011.01.051

Podkolodnaya O.A., Podkolodny N.L. Polimorfizmy genov tsirkadnykh chasov (PM GTsCh) [Polymorphisms of circadian clock genes (PM CCG)]. Patent RF, no. 20133621533, 2013.

Ponomarenko P.M., Ponomarenko M.P., Drachkova I.A., Lysova M.V., Arshinova T.V., Savinkova L.K., Kolchanov N.A. Prediction of the affinity of the TATA-binding protein to TATA boxes with single nucleotide polymorphisms. Molekulyarnaya biologiya = Molecular Biology (Moscow). 2009;43(3):512-520.

Ponomarenko M., Rasskazov D., Arkova O., Ponomarenko P., Suslov V., Savinkova L., Kolchanov N.A. How to use SNP_TATA_Comparator to find a significant change in gene expression caused by the regulatory SNP of this gene's promoter via a change in affinity of the TATA-binding protein for this promoter. Biomed. Res. Int. 2015;359835. DOI 10.1155/2015/359835

Rasskazov D.A., Gunbin K.V., Ponomarenko P.M., Vishnevsky O.V., Ponomarenko M.P., Afonnikov D.A. SNP_TATA_Comparator: Web-service for comparison of SNPs within gene promoters associated with human diseases using the equilibrium equation of the TBP/TATA complex. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2013;17(4/1):599-606.

Royston S., Yasui N., Kondilis A.G., Lord S.V., Katzenellenbogen J. A., Mahoney M.M. ESR1 and ESR2 differentially regulate daily and circadian activity rhythms in female mice. Endocrinology. 2014;155(7):2613-2623. DOI 10.1210/en.2014-1101

Savinkova L.K., Drachkova I.A., Arshinova T.V., Ponomarenko P.M., Ponomarenko M.P., Kolchanov N.A. An experimental verification of the predicted effects of promoter TATA-box polymorphisms associated with human diseases on interactions between the TATA boxes and TATA-binding protein. PLoS One. 2013;8(2):e54626. DOI 10.1371/journal.pone.0054626

Savinkova L.K., Ponomarenko M.P., Ponomarenko P.M., Drachkova I.A., Lysova M.V., Arshinova T.V., Kolchanov N.A. TATA box polymorphisms in human gene promoters and associated hereditary pathologies. Biokhimiya = Biochemistry (Moscow). 2009;74(2):149-163.

Sherry S.T., Ward M.H., Kholodov M., Baker J., Phan L., Smigielski E.M., Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucl. Acids Res. 2001;29(1):308-311. DOI 10.1093/nar/29.1.308

Sieuwerts A.M., Ansems M., Look M.P., Span P.N., de Weerd V., van Galen A., Foekens J.A., Adema G.J., Martens J.W. Clinical significance of the nuclear receptor co-regulator DC-SCRIPT in breast cancer: an independent retrospective validation study. Breast Cancer Res. 2010;12(6):R103. DOI 10.1186/bcr2786

Starodubtseva N.L., Sobolev V.V., Soboleva A.G., Nikolaev A.A., Bruskin S.A. Genes expression of metalloproteinases (MMP-1, MMP-2, MMP-9, and MMP-12) associated with psoriasis. Genetika = Genetics (Moscow). 2011;47(9):1254-1261.

Sun A.H., Wang Z.M., Xiao S.Z., Li Z.J., Li J.Y., Kong L.S. Red cell basic ferritin concentration in sensorineural hearing loss. ORL J. Otorhinolaryngol. Relat. Spec. 1991;53(5):270-272.

Tang J.Q., Fan Q., Wan Y.L., Liu Y.C., Wang X., Wu T., Pan Y.S., Wu W. H., Zhu J. Ectopic expression and clinical significance of tissue factor/coagulation factor VII complex in colorectal cancer. Beijing Da Xue Xue Bao. 2009;41(5):531-536. DOI 10.3969/j.issn.1671-167x.2009.05.005

Thio D., Prasad V., Anslow P., Lennox P. Marrow proliferation as a cause of hearing loss in beta-thalassaemia major. J. Laryngol. Otol. 2008;122(11):1253-1256. DOI 10.1017/S0022215107000874

Unger E.L., Earley C.J., Beard J.L. Diurnal cycle influences peripheral and brain iron levels in mice. J. Appl. Physiol. (1985). 2009;106(1):187-193. DOI 10.1152/japplphysiol.91076.2008

Waardenberg A.J., Basset S.D., Bouveret R., Harvey R.P. CompGO: an R package for comparing and visualizing Gene Ontology enrichment differences between DNA binding experiments. BMC Bioinformatics. 2015;16(1):275. DOI 10.1186/s12859-015-0701-2

Wang Y., Kato N., Hoshida Y., Yoshida H., Taniguchi H., Goto T., Moriyama M., Otsuka M., Shiina S., Shiratori Y., Ito Y., Omata M. Interleukin-1beta gene polymorphisms associated with hepatocellular carcinoma in hepatitis C virus infection. Hepatology. 2003;37(1):65-71. DOI 10.1053/jhep.2003.50017

Wang J., Luo Y., Wang K., Wang Y., Zhang X., Teng H., Sun Z. Clock-controlled StAR's expression and corticosterone production contribute to the endotoxemia immune response. Chronobiol. Int. 2015;32(3):358-367. DOI 10.3109/07420528.2014.982284

Wu Y., Zhou J., Van Heerikhuize J., Jockers R., Swaab D. Decreased MT1 melatonin receptor expression in the suprachiasmatic nucleus in aging and Alzheimer's disease. Neurobiol. Aging. 2007;28(8):1239-1247. DOI 10.1016/j.neurobiolaging.2006.06.002

Wu K., Zhou X., Zheng F., Xu X., Lin Y., Yang J. Influence of interleukin-1 beta genetic polymorphism, smoking and alcohol drinking on the risk of non-small cell lung cancer. Clin. Chim. Acta. 2010;411(19/20):1441-1446. DOI 10.1016/j.cca.2010.05.035

Yamazaki H., Takeoka M., Kitazawa M., Ehara T., Itano N., Kato H., Taniguchi S. ASC plays a role in the priming phase of the immune response to type II collagen in collagen-induced arthritis. Rheumatol. Int. 2012;32(6):1625-1632. DOI 10.1007/s00296-011-1825-y

Yang M., Laflamme K., Gotea V., Joiner C.H., Seidel N.E., Wong C., Petrykowska H.M., Lichtenberg J., Lee S., Welch L., Gallagher P.G., Bodine D.M., Elnitski L. Genome-wide detection of a TFIID localization element from an initial human disease mutation. Nucl. Acids Res. 2011;39(6):2175-2187. DOI 10.1093/nar/gkq1035

Yoo S.S., Jin C., Jung D.K., Choi Y.Y., Choi J.E., Lee W.K., Lee S.Y., Lee J., Cha S.I., Kim C.H., Seok Y., Lee E., Park J.Y. Putative functional variants of XRCC1 identified by RegulomeDB were not associated with lung cancer risk in a Korean population. Cancer Genet. 2015;208(1/2):19-24. DOI 10.1016/j.cancergen.2014.11.004

Zerbino D.R., Wilder S.P., Johnson N., Juettemann T., Flicek P.R. The Ensembl regulatory build. Genome Biol. 2015;16:56. DOI 10.1186/s13059-015-0621-5

Zhang R., Lahens N.F., Ballance H.I., Hughes M.E., Hogenesch J.B. A circadian gene expression atlas in mammals: implications for biology and medicine. Proc. Natl Acad. Sci. USA. 2014;111(45):16219-16224. DOI 10.1073/pnas.1408886111

Zukunft J., Lang T., Richter T., Hirsch-Ernst K.I., Nussler A.K., Klein K., Schwab M., Eichelbaum M., Zanger U.M. A natural CYP2B6 TATA box polymorphism (–82T→C) leading to enhanced transcription and relocation of the transcriptional start site. Mol. Pharmacol. 2005;67(5):1772-1782. DOI 10.1124/mol.104.008086

Вавиловский журнал генетики и селекции. 2015;19(6):699-706
Doi 10.18699/VJ15.088

Оценка роли однонуклеотидного полиморфизма в гене лимфотоксина бета при доместикации свиньи на основе биоинформационного и экспериментального подходов

Р.Б. Айтназаров[1, 3], Е.В. Игнатьева[1, 3], Н.Э. Базарова[4], В.Г. Левицкий[1, 3], С.П. Князев[4], Я. Гон[3, 5], Н.С. Юдин[1, 2, 3]

1 Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия
2 Федеральное государственное бюджетное научное учреждение «Научно-исследовательский институт терапии и профилактической медицины», Новосибирск, Россия
3 Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия
4 Федеральное государственное бюджетное образовательное учреждение высшего образования «Новосибирский государственный аграрный университет», Новосибирск, Россия
5 Хэйлунцзянский университет, биологический факультет, Харбин, Хэйлунцзянская провинция, Китай

В работах, выполненных на диких и лабораторных животных, показано существование компромисса между репродуктивным успехом и иммунитетом. Поэтому в процессе доместикации могли отбираться особи с повышенными репродуктивными способностями, но со сниженным иммунитетом. Пониженная реактивность иммунной системы могла в дальнейшем стать наследуемой путем фиксации в популяции генов с «неблагоприятными» мутациями. Цель исследования – изучить: 1) частоты генотипов и аллелей однонуклеотидного полиморфизма (SnP – Single nucleotide P olymorphism) SnP rs340283541 в гене цитокина лимфотоксина бета (*LTB*) у домашних свиней и диких кабанов; 2) экспрессию мРНК этого гена у миниатюрных свиней с разными генотипами; 3) провести биоинформатический анализ потенциальной функциональной роли этого SnP. Частота генотипа GG в выборке кабанов была достоверно ниже частоты данного генотипа в объединенной выборке из разных пород и популяций домашних свиней. уровень экспрессии мРНК гена *LTB* в лимфатическом узле у миниатюрных свиней с генотипом GG имел тенденцию к повышению ($p < 0,06$), по сравнению с носителями аллеля A. SNP rs340283541 входит в состав мотива ДНК с высокой степенью консервативности у 12 видов млекопитающих, что косвенно свидетельствует о его важной функциональной роли. С помощью контекстного анализа выявлено, что аллель A содержит потенциальные сайты связывания транскрипционных факторов BRN-2 и AP-1, а аллель G – факторов RFX1, ISGF3 (сайт ISRE) и USF, которые экспрессируются в клетках иммунной системы. Таким образом, в процессе доместикации свиней произошло повышение частоты генотипа GG SNP rs340283541 в 3′- области гена LTB. Генотип GG, вероятно, ассоцииро-ван с повышенным уровнем экспрессии мРНК гена LTB в ткани лимфатического узла. Повышение уровня экспрессии у свиней с генотипом GG может быть связано с образованием сайтов связывания транскрипционных факторов RFX1, ISRE, USF и/или разрушением сайтов связывания BRN-2 и AP-1. Также не исключено, что полиморфизм rs340283541 находится в неравновесии по сцеплению с другой функционально значимой мутацией.

Ключевые слова: доместикация; свинья; дикий кабан; ген; лимфотоксин бета; однонуклеотидный полиморфизм; SNP; транскрипция; сайт связывания транскрипционного фактора.

Rceived 21.09.2015
Accepted for publication 15.10.2015

e-mail: yudin@bionet.nsc.ru

Dissecting the role of single nucleotide polymorphism of lymphotoxin beta gene during pig domestication using bioinformatic and experimental approaches

R.B. Aitnazarov[1, 3], E.V. Ignatieva[1, 3], N.E. Bazarova[4], V.G. Levitsky[1, 3], S.P. Knyazev[4], Y. Gon[3, 5], N.S. Yudin[1, 2, 3]

1 Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
2 Institute of Internal and Preventive Medicine, Siberian Branch of Russian Academy of Medical Sciences, Novosibirsk, Russia
3 Novosibirsk State University, Novosibirsk, Russia
4 Novosibirsk State Agrarian University, Novosibirsk, Russia
5 Heilongjiang University College of Life Sciences, Harbin, China

Studies of wild and laboratory animals have revealed a trade-off between reproductive success and immunity. Therefore, it is likely that domestication favored selection of individuals with high reproductive performance but low immunity. The low responsiveness of the immune system could become hereditary through fixation of genes with “unfavorable” mutations in populations. The objectives of this work are: 1) determination of frequencies of geno-types and alleles of the rs340283541 SnP in the gene for the lymphotoxin beta (*LTB*) cytokine in pigs of domestic breeds and wild boars; 2) investigation of the expression of *LTB* mrnA in minipigs with different genotypes, and 3) bioinformational analysis of the putative functional role of the SnP. The frequency of the GG genotype in the wild boar sample was significantly lower than in the pooled sample of domestic pigs. The *LTB* mrnA expression rate in the lymph node of minipigs with genotype GG tended to increase ($p < 0.06$) in comparison with carriers of allele A. The rs340283541 SnP occurs in a DnA motif highly conservative among 11 mammalian species; thus, it may be of functional significance. context analysis shows that allele A has putative binding sites for transcription factors BRN-2 and AP-1, whereas allele G has binding sites for transcription factors RFX1, ISGF3 (site ISRE), and USF expressed in cells of the immune system. Thus, pig domestication was accompanied by an increase in the frequency of the GG genotype for the rs340283541 SnP, occurring in the 3′ region of the *LTB* gene. it is likely that the GG genotype is associated with elevated *LTB* mrnA expression in the lymph node tissue. This increase may be related to the formation of binding sites for RFX1, ISRE, and USF and/or disruption of binding sites for BRN-2 and AP-1. A linkage disequilibrium between rs340283541 and another functionally significant mutation in l TB is also conceivable.

Key words: domestication; pig; wild boar; gene; lymphotoxin beta; single nucleotide polymorphism; SnP; transcription; transcription factor binding site.

HOW TO CITE THIS ARTICLE?

Aitnazarov r .B., ignatieva E.V., Bazarova n .E., l evitsky V.G., Knyazev S.P., Gon Y., Yudin n .S. Dissecting the role of single nucleotide polymorphism of lymphotoxin beta gene during pig domestication using bioinformatic and experimental approaches. Vavilovskii Zhurnal Genetiki i Selektsii = **Vavilov Journal of Genetics and Breeding.** 2015;19(6):699-706. Doi 10.18699/VJ15.088

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?

Айтназаров Р.Б., Игнатьева е.В., Базарова Н.Э., левицкий В.Г., Князев С.П., Гон Я., Юдин Н.С. Оценка роли однонуклеотидного полиморфизма в гене лимфотоксина бета при доместикации свиньи на основе биоинформационного и экспериментального подходов. Вавиловский журнал генетики и селекции. 2015;19(6):699-706. Doi 10.18699/VJ15.088

В работах, выполненных на диких и лабораторных животных, показано существование компромисса между репродуктивным успехом и иммунитетом (Sheldon, Verhulst, 1996; Ardia et al., 2011; van der Most et al., 2011; Balenger, Zuk, 2014). Цитокинам принадлежит центральная роль в регуляции иммунного ответа, а также его интеграции с физиологическими функциями других систем организма – эндокринной и гематопоэтической. Одним из наиболее значительных семейств цитокинов является семейство белков фактора некроза опухолей, которое включает около 48 белков (Kim et al., 2005). Типичным представителем этого семейства является лимфотоксин бета (LTB). Он синтезируется активированными Т- и В-лимфоцитами, естественными киллерами и образует гетеротример с лимфотоксином-альфа LTA/LTB/LTB (реже LTA/LTA/LTB) и, таким образом, «заякоривает» лимфотоксин-альфа на клеточной мембране лимфоцита (Nakamura et al., 1995). Такой гетеротример функционирует как лиганд для рецептора TNFRSF3/LTBR и участвует в развитии иммунного ответа, обеспечивая межклеточную коммуникацию (Crowe et al., 1994). Считается, что основная функция лимфотоксина бета заключается в стимулировании развития лимфоидной ткани, в первую очередь лимфатических узлов (Onder et al., 2013).

Доместикация (одомашнивание) животных представляет собой гигантский биологический эксперимент, главный результат которого состоит в огромном повышении темпа и размаха изменчивости организмов (Беляев, 1981). Известно, что, хотя виды одомашненных животных относятся к далеко отстоящим систематическим группам, их изменчивость по многим признакам носит характер гомологической изменчивости (Трапезов, 2009). Одним из основных результатов доместикации является сильное повышение воспроизводительных способностей животных (Беляев, 1979). Поэтому доместицированных животных и их диких предков можно рассматривать в качестве модели для изучения механизмов компромисса между репродуктивными параметрами и иммунитетом. Можно ожидать, что в процессе доместикации и селекции на повышение хозяйственно важных признаков, преимущества при отборе имели более плодовитые особи, а стало быть, особи со сниженным иммунитетом. Пониженная реактивность иммунной системы могла в дальнейшем стать наследуемой путем фиксации в популяции генов с «неблагоприятными» для иммунореактивности мутациями.

Хорошей моделью для изучения мутаций в генах иммунной системы при доместикации являются домашняя свинья и ее дикий предок – кабан. В результате секвенирования полного генома свиньи в 3′-области гена *LTB* на расстоянии 283 п. н. от последнего экзона был обнаружен однонуклеотидный полиморфизм (SNP) – замена A на G в позиции 27547441 7-й хромосомы (rs340283541).

Цель исследования – оценить распространение SNP rs340283541 гена *LTB* в популяциях домашних свиней и диких кабанов, сравнить относительную экспрессию мРНК этого гена у свиней с разными генотипами и провести анализ потенциальной функциональной роли этого SNP методами биоинформатики.

Материалы и методы

Образцы крови и шкур диких кабанов, представляющих разные подвиды *Sus scrofa* L., получены из России (Воронежский биосферный заповедник, Воронежская область) и Украины (Николаевская и Закарпатская области). Образцы крови свиней пород ландрас, скороспелой мясной СМ-1 и сибирских миниатюрных свиней получены из хозяйств Новосибирской области. ДНК выделяли методом протеолитической обработки с последующей экстракцией фенолом. Фрагмент гена *LTB* амплифицировали с использованием праймеров 5′-TCCCCTCAGACTCAACACTGCACAC-3′ и 5′-TTCAGGCAGCTGGCAGGGAGAA-3′. Ампликон обрабатывали рестриктазой HpySE526 I (Сибэнзим, Россия). Генотип SNP rs340283541 определяли путем электрофореза продуктов рестрикции в 4 %-м полиакриламидном геле

(генотип AA – 172 п. н., генотип AG – 172, 148 и 24 п. н., генотип GG – 148 и 24 п. н.).

Для изучения экспрессии мРНК гена *LTB* в паховом лимфоузле миниатюрных свиней с различными генотипами по этому гену использовали животных в возрасте 1 мес., массой 8–12 кг. В эксперимент брали по четыре животных каждого генотипа. Амплификацию фрагмента гена *LTB* длиной 139 п. н. проводили с праймерами LTB/F 5′-AACTGGTAACAGGGACCGCT-3′ и LTB/R 5′-ATCCAAGCGCCAATGAGGT-3′. В качестве гена сравнения использовали ген *GADPH*. С помощью праймеров GAPDH/F 5′-CGTCAAGCTCATTTCCTGGTACG-3′ и GAPDH/R 5′-GGGGTCTGGGATGGAAACTGGAAG-3′ амплифицировали фрагмент размером 223 п. н.

Суммарную РНК выделяли с помощью реактива TRIzol (Invitrogen, США) согласно рекомендациям производителя. Реакцию обратной транскрипции проводили с использованием олиго-(dT)-праймера и обратной транскриптазы M-MuLV (Сибэнзим, Россия). Реакцию ПЦР в реальном времени (ПЦР-РВ) проводили с использованием набора реагентов для проведения ПЦР-РВ в присутствии красителя SYBR Green I (Синтол, Росия) по стандартной схеме. Полученные данные обрабатывали методом относительного количественного анализа Δ-ΔCt с помощью программы Rotor-Gene 6000 Series Software.

Частоты аллелей и генотипов SNP rs340283541 сравнивали с применением критерия χ^2 с поправкой Йетса. Относительный уровень экспрессии мРНК гена *LTB* у сибирских миниатюрных свиней с различным генотипом по SNP rs340283541 сравнивали с помощью однофакторного дисперсионного анализа и критерия Стьюдента.

Построение выравнивания между участком 7-й хромосомы свиньи, включающим SNP, и геномными последовательностями других млекопитающих осуществляли с использованием опций геномного браузера UCSC (https://genome.ucsc.edu/). Отображение уровня консервативности нуклеотидов в выравнивании выполняли с помощью программы WebLogo (http://weblogo.berkeley.edu/logo.cgi).

Потенциальные сайты связывания транскрипционных факторов в окрестностях SNP rs340283541 выявляли с помощью интернет-доступной программы Match-1.0 Public (http://www.gene-regulation.com/cgi-bin/pub/programs/match/bin/match.cgi), осуществляющей поиск потенциальных сайтов в нуклеотидных последовательностях на основе весовых матриц (ВМ) методом PWM. Оценку количества ложноположительных сайтов, выявляемых методом PWM, производили на геномной последовательности 1-й хромосомы свиньи. Данные о QTLs экстрагировали их базы Pig Quantitative Trait Locus Database (Pig QTLdb) (http://www.animalgenome.org/cgi-bin/QTLdb/SS/index).

Подробное описание всех процедур см. в Дополнительных материалах 1[1].

результаты

Было проведено генотипирование замены A на G в 3′-фланкирующем районе гена *LTB* свиньи (rs340283541). Частота редкого аллеля A у диких кабанов достоверно не отличалась от его частоты как у отдельных пород свиней, так и у домашних свиней в целом (Доп. материалы 2). Однако частота генотипа GG в выборке кабанов (20,0 %) была достоверно ниже частоты этого же генотипа в объединенной выборке домашних свиней (44,5 %) ($p \leq 0{,}005$).

Однофакторный дисперсионный анализ не выявил достоверного влияния генотипа SNP rs340283541 на уровень относительной экспрессии мРНК гена *LTB* в ткани пахового лимфатического узла у миниатюрных свиней (Доп. материалы 3). При этом уровень экспрессии мРНК гена *LTB* у мини-свиней с генотипом GG имел тенденцию к повышению по сравнению с носителями аллеля A (объединенная группа GA+AA) ($p < 0{,}07$).

С использованием данных, экстрагированных из геномного браузера UCSC, произведена оценка уровня консервативности участка, включающего исследуемый SNP. У свиньи SNP расположен в 3′фланкирующем районе гена *LTB* на расстоянии 284 н. п. от точки терминации транскрипции. У 11 видов млекопитающих в этом районе ДНК были выявлены участки гомологии протяженностью от 38 до 74 нуклеотидов (рисунок, *а*). Наиболее протяженный участок гомологии (74 нуклеотида) выявлен у коровы. Нуклеотид, соответствующей полиморфной позиции у свиньи, располагался у этих видов организмов на расстоянии от 280 (у овцы) до 445 (у гиббона и шимпанзе) нуклеотидов ниже 3′-конца гена *LTB*. Наиболее часто (в 10 случаях из 12) в данной позиции располагается тимин (рисунок, *а*). Лишь у двух видов в этой позиции обнаружены другие нуклеотиды: полиморфная позиция аденин/гуанин (свинья) и цитозин (мышь). Десять нуклеотидов, непосредственно прилежащих к SNP на 3′-фланге, являются высоко консервативными. На графике WebLogo данный консервативный район отображен в виде мотива TTTCTCTTGG (рисунок, *б*).

С помощью программы распознавания сайтов связывания транскрипционных факторов Match-1.0 Public были предсказаны потенциальные сайты в двух нуклеотидных последовательностях участка 3′-фланкирующей области гена *LTB* в окрестностях SNP rs340283541, соответствующих аллельным вариантам A и G. Было выявлено, что максимальные значения ВМ имели потенциальные сайты связывания транскрипционных факторов 5 типов (табл. 1). У этих потенциальных сайтов были выявлены: 1) либо значения параметра matrix similarity (сходство последовательности с матрицей) > 0,85; 2) либо значения параметра core similarity (сходство последовательности с матрицей по 5 наиболее консервативным позициям матрицы) > 0,95; 3) либо значения обоих параметров превышали вышеуказанные пороговые. Два сайта (Brn-2 и AP-1) были обнаружены в последовательности аллеля A. Три сайта (RFX1, USF, ISRE) выявлены в последовательности аллеля G. У двух из пяти сайтов связывания (Brn-2 и RFX1) оба значения ВМ превышали вышеуказанные пороговые. Сравнение участков ДНК, соответствующих предсказанным сайтам, с консенсусными последовательностями показало:

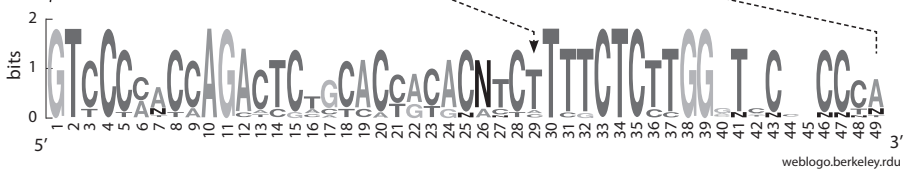
1) в четырех случаях (Brn-2, AP-1, RFX1, ISRE) из пяти полиморфный нуклеотид (A либо G) расположен в коровой части потенциального сайта, соответствующей пяти наиболее консервативным нуклеотидам матрицы;

[1] Дополнительные материалы 1–3 см. в Приложении 4 по адресу: http://www.bionet.nsc.ru/vogis/download/pict-2015-12/appx4.pdf

a

Species	The nucleotide sequence at the 3′-flank of the pig *LTB*	Genomic positions	Distance between SnP (rs340283541) and the 3′-end of the *LTB* transcript
Pig	gtcccc-tcagactcaacactgcac-tcAtttctcttggatcccacccagcagtgatagggtcattacagggg-cag	chr7:27,547,415-27,547,485	284
Cow	gtcccc-ccagactcagcactgcac-tcTtctctcttgggtccatcccagtagtggtggggtcattgcagggg-aag	chr23:27,532,139-27,532,212	306
Sheep	gtcccc-ccagactcagcactgcac-tcTtttctcttgggtccatcccagtagtggtggggtcattgca--------	chr20:26,852,623-26,852,689	280
Mouse	gtccccacaagacccccctcacatacaccCttgctcctggcttctgcct----------ctgtgaggag	chr17:35,196,687-35,196,744	411
Mouse lemur	gtccccaccagcctgtgcactgcac-tcTtttctcttgggtccttcccaac---ggaggggccattacagggg-aag	scaffold_8597:54,872-54,801	281
Rat	gttccctcaagaacccctcccatacactTtttctcctggcttctgcctt----------gctgtga-gaa	chr20:6,941,356-6,941,298	426
Gorilla	gtcccaaccagactctgcaccacac-tcTtttctcttgggtacccccca----------acagtga-gag	chr6:32,473,642-32,473,585	294
Human	gtcccaaccagactctgcaccatac-tcTtttctcttgggtacccccca----------acagtga-gag	chr6:31,548,069-31,548,012	295
Orangutan	gtcccaaccagactctgcaccacac-tcTtttctctcgggtcccccca----------gcagtga-gag	chr6:32,078,216-32,078,159	352
Gibbon	gtcctaaccagactctgcaccacgc-tcTtttctcttgggtcccccca----------acagtga-gag	chr1a:76,380,055-76,379,998	445
Tarsier	gttcccaccagattctgcaccacg---cTtttctcttggatccctccca----------cagtga-gag	KE944812v1:95,150-95,096	289
Chimpanzee	gtcccaaccagactctgcaccacac-tcTtttctcttgg----------	chr6:31,859,616-31,859,579	445

b



Conservative nucleotides in the 3′-flan ing region of the *LTB* gene in twelve mammalian species.

(*a*) Alignment of regions homologous to the region of pig chromosome 7 containing the rs340283541 SnP. The SnP position in the pig sequence and the corresponding nucleotides in sequences of other species are shown in bold capitals. (*b*) Presentation of conservative positions with webl ogo (crooks et al., 2004). The overall height of a stack indicates the sequence conservation at the corresponding position (bits on the Y axis), while the height of symbols within the stack refle ts their relative frequencies at that position. numerals on the X axis indicate numbers of positions in the alignment. The rs340283541 SnP (dashed arrow) occurs at position 29 of the pig sequence.

Table 1. Putative transcription factor binding sites predicted in the DnA region comprising SnP rs340283541

Transcription factor binding site	wM scores		Strand	nucleotide sequence of binding site* / consensus sequence**	wM scores for the DnA region with nucleotide substitution	
	Matrix similarity	core similarity			Matrix similarity	core similarity
Allele A						
Brn-2[!]	**0.857**	**1.000**	–	5′-actcATTTCtcttgga-3′ ||||||||| |||| 5′-nyknATTWYsnatgnn-3′	0.713	0.759
AP-1	0.683	**0.989**	+	5′-tgcACTCAt-3′ | ||||| 5′-ntgASTCAg-3′	< 0.550	< 0.600
Allele G						
r FX1	**0.888**	**0.982**	–	5′-tcGTTTCtcttggatccc-3′ ||||| || ||| | ||| 5′-nnGTTRCyatngynacnn-3′	0.693	0.720
iSr E	0.786	1.000	+	5′-tcGTTTCtcttggat-3′ ||||||||| 5′-caGTTTCwctttycc-3′	0.672	0.800
USF	**0.846**	0.905	+	5′- tGCACTcgtt-3′ ||| | | 5′-nnRYCACGtgrynn -3′	0.732	0.905

The capital letters denote nucleotides that correspond to fi e most conservative positions in the binding sites predicted by the Match 1.0 tool. The two wM scores that exceed predefined th esholds (matrix similarity>0.85 and core similarity>0.95) are given in bold type and underlined. (!) The false-positive rate in TFBSs prediction using PwM is less than 3E-4 at the recognition threshold corresponding to the score of the putative binding site comprising the SnP (rs340283541). * For the binding sites found in the minus (reverse) strand, we present the DnA sequence from the plus (forward) strand. in such cases, we present the inverted versions of the consensus sequences. ** Positions in the consensus sequences are indicated in the 15-letter IUPAc nucleotide code, where K = G or T; r = A or G; S = c or G; Y = c or T; and w = A or T.

Table 2. Quantitative trait loci (QTl s) co-localized with Sn P rs340283541 (position in the porcine genome chr7:27547441) that affect the immune or reproduction traits in pigs.

QTl	QTl iD in Pig QTl db	QTl Span (position on chromosome 7)
Traits related to the immune capacity		
l ymphocyte number	5469	11625414–36993248
Hemolytic complement activity (alternative pathway)	9596	11625414–38992356
Prr SV susceptibility	31803	27019705–27999311
r eproduction-associated traits		
Age at puberty	588	16365408–42509154
c orpus luteum number	24284	8014191–92220281
	31864	11136187–116028974
Teat number	5257	10763543–117929721
	584	18392887–31164815

2) во всех случаях полиморфный нуклеотид (A либо G) во фрагменте регуляторной последовательности гена *LTB* и соответствующий ему на основании выравнивания нуклеотид консенсусной последовательности совпадают;
3) замена нуклеотида в полиморфной позиции приводит к существенному снижению обоих значений BM у четырех (Brn-2, AP-1, RFX1, ISRE) из пяти рассматриваемых сайтов (табл. 1, последние две колонки). У сайта связывания USF, выявленного в аллельном варианте G, замена гуанина на аденин приводит к снижению значения matrix similarity и не влияет на значение core similarity.

Для того чтобы выявить сайт, распознанный с наименьшей ошибкой перепредсказания (False Positive Rate), для всех пяти выявленных типов сайтов в качестве оценки было рассчитано количество сайтов данного типа в последовательности хромосомы 1 свиньи. В каждом случае пороговое значение BM, использованное функцией распознавания, равнялось значению BM соответствующего потенциального сайта, выявленного в окрестностях SNP. Было обнаружено, что минимальная плотность сайтов в геномной последовательности ДНК выявляется при распознавании сайта BRN2 (в табл. 1 помечено знаком «!»).

Для того чтобы охарактеризовать потенциальную роль участка хромосомы 7 свиньи, включающего полиморфизм rs340283541, в формировании фенотипических признаков, мы проанализировали данные из базы Pig QTLdb. Было выявлено, что полиморфизм rs340283541 содержится в 161 локусе количественных признаков (QTLs). Из них три локуса влияют на три различных показателя иммунной системы (табл. 2), а другие пять QTLs ассоциированы с тремя репродуктивными параметрами животных.

Обсуждение

В результате проведенного исследования нами обнаружено, что генотип GG SNP rs340283541 у диких кабанов встречается гораздо реже, чем у домашних свиней (Доп. материалы 2). Это позволяет выдвинуть предположение, что замена A на G в этой позиции может нарушать иммунный ответ, что, согласно гипотезе компромисса, должно приводить к улучшению репродуктивных показателей у животных, гомозиготных по этой мутации. Однако против такого предположения говорит повышенный уровень экспрессии мРНК гена *LTB* у животных с генотипом GG, по сравнению с носителями аллеля A (Доп. материалы 3). Альтернативное объяснение этих фактов заключается в вероятном существовании положительной корреляции между иммунным ответом и репродуктивными признаками при доместикации свиней. Поскольку процесс доместикации, несомненно, сопровождался усилением давления со стороны паразитов, селекция на улучшение репродуктивных показателей могла быть эффективной только у генетически устойчивых животных с сильным иммунитетом.

В литературе имеется ряд работ, косвенно свидетельствующих о вероятном отборе «патологических» аллелей в генах иммунной системы при селекции сельскохозяйственных животных на повышение хозяйственно важных признаков. Например, селекция коров красной норвежской породы на увеличение удоев молока сопровождалась повышением частоты мастита, кератоза и задержания последа, что свидетельствует о снижении у них функционального состояния иммунной системы (Heringstad et al., 2007). При селекции на пониженную частоту мастита у коров той же породы происходило снижение содержания белка в молоке (Heringstad et al., 2005). Интересно, что мутация в гене *LTB* у мышей приводит к нарушению формирования волосяных фолликулов (Cui et al., 2006). Поэтому будет интересно сравнить окраску волосяного покрова у миниатюрных свиней с различными генотипами по гену *LTB*.

Консервативность района Дн К в окрестностях Sn P rs340283541

Анализ нуклеотидных последовательностей 3′-районов генов *LTB* млекопитающих показал, что SNP rs340283541 располагается на консервативном участке (рисунок). Это означает, что данный участок с высокой долей вероятности несет определенную функциональную нагрузку.

Известно, что 3′-районы генов могут участвовать в регуляции транскрипции генов. Так же, как и промоторы генов и регуляторные районы, расположенные в 5′-районах, и интронах, 3′-области генов могут включать функционально-активные сайты связывания транскрипционных факторов, оказывающих как активирующие, так и подавляющие эффекты на транскрипцию (Kolchanov et al., 2002; Меркулова и др., 2013; Игнатьева и др., 2015). Основываясь на данном представлении, мы предположили, что замена нуклеотида в полиморфной позиции SNP rs340283541 может нарушать сайт связывания транскрипционного фактора, участвующего в регуляции активности гена *LTB*. Для того чтобы проверить данное предположение, нами был произведен поиск потенциальных сайтов связывания транскрипционных факторов в последовательности ДНК, включающей SNP rs340283541. Было выявлено, что аллельный вариант A содержит потенциальные сайты связывания Brn-2 и AP-1, которые нарушаются при замене аденина на гуанин в полиморфной позиции. С другой стороны, аллельный вариант G 3′-регуляторной последовательности гена *LTB* содержит потенциальные сайты RFX1, ISRE, USF, которые не удается выявить в случае, если в полиморфной позиции присутствует аденин (табл. 1).

Характеристика предсказанных сайтов

Brn-2. Фактор Brn-2 (Brain-2) кодируется геном *POU3F2* (POU class 3 homeobox 2), который, согласно данным базы EntrezGene, в геноме свиньи еще имеет статус «MODEL» (предсказан на основе электронной аннотации и слабо изучен). Относительно белка Brn-2/Oct7 у человека и у мыши в научной литературе сложилось представление как о транскрипционном факторе, участвующем в дифференцировке нейрональных клеток (Zhao et al., 2013). Гиперпродукция фактора Brn-2 ассоциирована с усиленной пролиферацией клеток меланомы (Goodall et al., 2004). Согласно данным базы BioGPS, у человека ген *POU3F2*, кодирующий фактор Brn-2, экспрессируется не только в нейральных клетках, но и в других тканях и органах, включая лимфатические узлы. Известно, что Brn-2 может выступать в качестве репрессора транскрипционной активности гена *CDH13* (T-cadherin) в клетках меланомы (Ellmann et al., 2012). Это позволяет предположить, что в случае, если данный сайт является функциональным у животных с аллелем A, Brn-2 также может выполнять роль репрессора и замена аденина на гуанин (аллель G) может приводить к снятию репрессирующего эффекта фактора Brn-2 и активации транскрипции гена *LTB*.

AP-1. Фактор AP-1 представляет собой гетеродимер, в состав которого могут входить белки таких семейств, как c-Fos, c-Jun, ATF и JDP. Фактор AP-1 участвует в регуляции генов, экспрессирующихся в широком круге тканей, в ответ на множество факторов, включая цитокины, факторы роста, стрессовые сигналы, бактериальные и вирусные инфекции (Hess et al., 2004). У потенциального сайта связывания фактора AP-1 наблюдается наиболее существенное (по сравнению с потенциальными сайтами связывания других факторов) падение значения core similarity при замене аденина на гуанин в полиморфной позиции. Как правило, фактор AP-1 является активатором транскрипции, однако имеются свидетельства того, что AP-1 может функционировать в роли репрессора (Mittelstadt et al., 2012). Таким образом, присутствие сайта связывания AP-1 в последовательности аллеля A может быть причиной более низкого уровня транскрипции гена у животных с генотипами AA и AG, по сравнению с животными с генотипом GG.

RFX1. Ген фактора RFX1 (Regulatory factor X1) имеет статус «MODEL» в геноме свиньи согласно базе данных EntrezGene. У человека RFX1 регулирует активность генов, экспрессирующихся в различных тканях (включая лимфоидные), среди которых идентифицированы гены 2-го класса главного комплекса гистосовместимости (Fontes et al., 1997), а также гены вируса гепатита С и вируса Эпштейна–Бара (Tierney et al., 2000). Согласно данным базы GeneCards, у человека ген *RFX1* экспрессируется в лимфатических узлах, моноцитах, Т- и В-лимфоцитах.

ISRE. Согласно базе TRANSFAC MATRIX (Release 7), с сайтом ISRE (interferon-stimulated response element) (идентификатор матрицы V$ISRE_01) связывается фактор (мультимерный комплекс) ISGF-3 (Interferon-Stimulated Gene Factor-3), включающий три субъединицы из семейства STAT и одну субъединицу из семейства IRF (Kessler et al., 1990). Комплекс ISGF-3 опосредует активацию генов в ответ на интерферон-α и -β, и сайт ISRE присутствует в регуляторных районах целого ряда генов, активно экспрессирующихся в клетках лимфоидного ряда (Kolchanov et al., 2002; Ananko et al., 2007). Таким образом, появление потенциального сайта ISRE в 3′-регуляторной последовательности гена *LTB* вследствие замены аденина на гуанин (аллель G) с высокой вероятностью может повлечь за собой активацию экспрессии гена *LTB* в лимфатических узлах у животных с генотипом GG. Потенциальный сайт ISRE, выявленный в последовательности ДНК аллельного варианта G, является перспективным кандидатом для дальнейшего экспериментального исследования в связи с тем, что, как показано нами ранее, частота носителей полноразмерных геномов ретровирусов PERV типов A, B и C у сибирских мини-свиней очень велика (Айтназаров и др., 2014).

USF. Факторы семейства USF (upstream transcription factor) еще слабо изучены у свиньи. В базе EntrezGene статусы для генов *USF1* и *USF2* обозначены как «PROVISIONAL» и «MODEL». У человека и мыши белки USF1 и USF2 известны как повсеместно экспрессирующиеся факторы. USF1 и USF2 участвуют в регуляции генов, функционирующих в регуляторных сетях стрессового и иммунного ответа, клеточного цикла и пролиферации (Corre, Galibert, 2006).

Предположение о том, что выявленные нами потенциальные сайты в районе, включающем SNP rs340283541, могут быть функциональными и их повреждение в результате замены нуклеотида может влиять на экспрессию гена, должно рассматриваться с определенной долей осторожности. Это связано с тем, что компьютерные методы распознавания сайтов (включая метод PWM, реализованный ресурсом Match-1.0 Public) характеризуются существенными погрешностями распознавания (ошибками недо- и перепредсказания) (Levitsky et al., 2007). По нашим оценкам, выполненным на геномной

последовательности хромосомы 1 свиньи, с наименьшим уровнем перепредсказания (т. е. с наибольшей точностью) в нашем исследовании был выявлен сайт связывания Brn2 (табл. 1).

С другой стороны, известно, что сайты, входящие в состав композиционных элементов (пар близкорасположенных сайтов посадки транскрипционных факторов), могут обладать невысоким сродством к соответствующим факторам (распознаваться методом PWM только при низких значениях BM) и, тем не менее, быть функциональными (Kel-Margoulis et al., 2000). В этом случае стабилизация взаимодействия участка ДНК с соответствующим транскрипционным фактором может осуществляться за счет дополнительных белок-белковых контактов с другим фактором, также контактирующим с соседним участком ДНК. Учитывая это обстоятельство, а также биологический контекст, в котором может функционировать сайт связывания, мы предполагаем, что, наряду с сайтом Brn-2, для дальнейшего исследования могут быть также интересны сайты связывания факторов, специфичных для клеток иммунной системы: RFX1 и ISGF-3 (сайт ISRE).

Биоинформатический анализ локусов количественных признаков

В ряде исследований показано, что многие однонуклеотидные полиморфизмы, ассоциированные с фенотипическими признаками, не являются причинными (causative), т. е. не оказывают эффекта ни на функцию белка, ни на интенсивность экспресссии гена. Однако выясняется, что рассматриваемые полиморфные аллели являются маркерами определенных гаплотипов, в составе которых имеются причинные SNP (Seddon et al., 2010; Loos et al., 2014). Учитывая такую возможность, мы проанализировали базу данных Pig QTLdb и выявили, что участок 7-й хромосомы, на котором расположен SNP rs340283541, входит в состав трех QTLs, имеющих отношение к иммунному ответу, и пяти QTLs, имеющих отношение к репродуктивным признакам (табл. 2). Эти данные позволяют рассматривать SNP rs340283541 как маркер определенного гаплотипа, подвергающегося селекции в ходе доместикации. Однако отбору могла подвергаться мутация в другом гене, поскольку в непосредственной близости к гену *LTB* находятся гены фактора некроза опухоли альфа (*TNFa*) и лимфотоксина альфа (*LTA*), которые также кодируют белки иммунной системы. О возможном вкладе SNP rs340283541 в формирование компромисса между репродуктивным успехом и иммунитетом при доместикации говорит и то, что он также входит в состав QTL для возраста полового созревания, числа сосков и числа желтых тел в яичниках, т. е. основных признаков плодовитости у свиней.

Локус генов *LTB-TNFa-LTA* человека насыщен регуляторными элементами, участвующими в транскрипционном и пост-транскрипционном контроле (Недоспасов, Купраш, 2008). У свиньи этот локус характеризуется необычно высокой плотностью генов. Так, промежуток между генами LTB и TNF, в котором находится SNP rs340283541, составляет всего 2 461 п. н., в то время как в геноме человека (имеющего такую же длину, что и геном свиньи), межгенное расстояние в 50 % случаев превышает величину 3 949 п. н. (Djebali et al., 2012), а среднее межгенное расстояние в геномах млекопитающих составляет 91 тыс. п. н. (Zhang et al., 2014). Не исключено, что участок между генами *LTB* и *TNF* содержит регуляторные элементы и аллельные варианты полиморфизма SNP rs340283541 маркируют варианты протяженного гаплотипа, определяющие активность одного из регуляторных участков.

Таким образом, обнаружено, что при доместикации свиньи произошло изменение частоты генотипов SNP в 3′-области гена *LTB*. По-видимому, данный SNP может быть ассоциирован с различным уровнем экспрессии мРНК гена *LTB*. Биоинформатический анализ показал, что участок 3′-района гена *LTB*, окружающий SNP, консервативен и замена нуклеотида в полиморфной позиции может влиять на потенциальные сайты связывания транскрипционных факторов. Выявлено 5 типов сайтов связывания факторов (Brn-2, AP-1, RFX1, ISRE, USF). Согласно нашим оценкам, сайт фактора Brn-2 распознан с наименьшей ошибкой перепредсказания, а с биологической точки зрения наиболее интересными в плане дальнейшего исследования являются сайты связывания RFX1 и ISRE. Также не исключено, что полиморфизм rs340283541 находится в неравновесии по сцеплению с другой, пока не известной, функционально значимой мутацией в гене *LTB*.

Acknowledgments

The experiments were financially supported by the Russian Foundation for Basic Research, project 13-04-00968a. The bioinformatical analysis was supported by Budgeted Project VI.61.1.2.

The authors are grateful to A.S. Klimov (Voronezh State University) for assistance in the collection of boar fur samples.

Conflict of interest

The authors declare no conflict of interest.

References

Aitnazarov R.B., Yudin N.S., Nikitin S.V., Ermolayev V.I., Voevoda M.I. Identification of whole genomes of endogenous retroviruses in Siberian miniature pigs. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2014;18(2): 294-297.

Ananko E.A., Kondrakhin Y.V., Merkulova T.I., Kolchanov N.A. Recognition of interferon-inducible sites, promoters, and enhancers. BMC Bioinformatics. 2007;8:56.

Ardia D.R., Parmentier H.K., Vogel L.A. The role of constraints and limitation in driving individual variation in immune response. Functional Ecology. 2011;25(1):61-73. DOI 10.1111/j.1365-2435.2010.01759.x

Belyaev D.K. Destabilizing selection as a factor of variability in domestication. Priroda = Nature (Moscow). 1979;2:36-45.

Belyaev D.K. Destabiliziruyushchiy otbor kak faktor domestikatsii [Destabilizing selection as a factor in domestication]. Genetika i blagosostoyanie chelovechestva [Genetics and the wellbeing of mankind]. Moscow, 1981:53-66.

Balenger S.L., Zuk M. Testing the Hamilton-Zuk hypothesis: past, present, and future. Integr Comp. Biol. 2014;54(4):601-613. DOI 10.1093/icb/icu059

Corre S., Galibert M.D. USF as a key regulatory element of gene expression. Med. Sci. (Paris). 2006;22(1):62-67.

Crooks G.E., Hon G., Chandonia J.M., Brenner S.E. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188-1190.

Crowe P.D., VanArsdale T.L., Walter B.N., Ware C.F., Hession C., Ehrenfels B., Browning J.L., Din W.S., Goodwin R.G., Smith C.A. A lymphotoxin-beta-specific receptor. Science. 1994;264(5159):707-710.

Cui C.Y., Hashimoto T., Grivennikov S.I., Piao Y., Nedospasov S.A., Schlessinger D. Ectodysplasin regulates the lymphotoxin-beta pathway for hair differentiation. Proc. Natl Acad. Sci. USA. 2006; 103(24):9142-9147.

Djebali S., Davis C.A., Merkel A., Dobin A., Lassmann T., Mortazavi A., Tanzer A., Lagarde J., Lin W., Schlesinger F., Xue C., Marinov G.K., Khatun J., Williams B.A., Zaleski C., Rozowsky J., Röder M., Kokocinski F., Abdelhamid R.F., Alioto T., Antoshechkin I., Baer M.T., Bar N.S., Batut P., Bell K., Bell I., Chakrabortty S., Chen X., Chrast J., Curado J., Derrien T., Drenkow J., Dumais E., Dumais J., Duttagupta R., Falconnet E., Fastuca M., Fejes-Toth K., Ferreira P., Foissac S., Fullwood M.J., Gao H., Gonzalez D., Gordon A., Gunawardena H., Howald C., Jha S., Johnson R., Kapranov P., King B., Kingswood C., Luo O.J., Park E., Persaud K., Preall J.B., Ribeca P., Risk B., Robyr D., Sammeth M., Schaffer L., See L.H., Shahab A., Skancke J., Suzuki A.M., Takahashi H., Tilgner H., Trout D., Walters N., Wang H., Wrobel J., Yu Y., Ruan X., Hayashizaki Y., Harrow J., Gerstein M., Hubbard T., Reymond A., Antonarakis S.E., Hannon G., Giddings M.C., Ruan Y., Wold B., Carninci P., Guigó R., Gingeras T.R. Landscape of transcription in human cells. Nature. 2012;489(7414):101-108. DOI 10.1038/nature11233

Ellmann L., Joshi M.B., Resink T.J., Bosserhoff A.K., Kuphal S. BRN2 is a transcriptional repressor of CDH13 (T-cadherin) in melanoma cells. Lab Invest. 2012;92(12):1788-1800. DOI 10.1038/labinvest.2012.140

Fontes J.D., Jabrane-Ferrat N., Peterlin B.M. Assembly of functional regulatory complexes on MHC class II promoters *in vivo*. J. Mol. Biol. 1997;270(3):336-345.

Goodall J., Martinozzi S., Dexter T.J., Champeval D., Carreira S., Larue L., Goding C.R. Brn-2 expression controls melanoma proliferation and is directly regulated by beta-catenin. Mol. Cell Biol. 2004;24(7):2915-2922.

Heringstad B., Chang Y.M., Gianola D., Klemetsdal G. Genetic association between susceptibility to clinical mastitis and protein yield in norwegian dairy cattle. J. Dairy Sci. 2005;88(4):1509-1514.

Heringstad B., Klemetsdal G., Steine T. Selection responses for disease resistance in two selection experiments with Norwegian red cows. J. Dairy Sci. 2007;90(5):2419-2426.

Hess J., Angel P., Schorpp-Kistner M. AP-1 subunits: quarrel and harmony among siblings. J. Cell Sci. 2004;117(25):5965-5973.

Ignatieva E.V., Podkolodnaya O.A., Orlov Y.L., Vasiliev G.V., Kolchanov N.A. Regulatory genomics: integrated experimental and computer approaches. Genetika = Genetics (Moscow). 2015;51(4): 409-429.

Kel-Margoulis O.V., Romashchenko A.G., Kolchanov N.A., Wingender E., Kel A.E. COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. Nucl. Acids Res. 2000;28(1):311-315.

Kessler D.S., Veals S.A., Fu X.Y., Levy D.E. Interferon-alpha regulates nuclear translocation and DNA-binding affinity of ISGF3, a multimeric transcriptional activator. Genes Dev. 1990;4(10):1753-1765.

Kim J.Y., Moon S.M., Ryu H.J., Kim J.J., Kim H.T., Park C., Kim K., Oh B., Lee J.K. Identification of regulatory polymorphisms in the TNF-TNF receptor superfamily. Immunogenetics. 2005;57(5): 297-303.

Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. Transcription regulatory regions database (TRRD): its status in 2002. Nucl. Acids Res. 2002;30(1):312-317.

Levitsky V.G., Ignatieva E.V., Ananko E.A., Turnaev I.I., Merkulova T.I., Kolchanov N.A., Hodgman T.C. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. BMC Bioinformatics. 2007;8:481.

Loos R.J., Yeo G.S. The bigger picture of FTO: the first GWAS-identified obesity gene. Nat. Rev. Endocrinol. 2014;10(1):51-61. DOI 10.1038/nrendo.2013.227

Mittelstadt M.L., Patel R.C. AP-1 mediated transcriptional repression of matrix metalloproteinase-9 by recruitment of histone deacetylase 1 in response to interferon β. PLoS One. 2012;7(8):e42152. DOI 10.1371/journal.pone.0042152

Merkulova T.I., Ananko E.A., Ignat'eva E.V., Kolchanov N.A. Regulatory transcription codes in eukaryotic genomes. Genetika = Genetics (Moscow). 2013;49(1):37-54.

Nakamura T., Tashiro K., Nazarea M., Nakano T., Sasayama S., Honjo T. The murine lymphotoxin-beta receptor cDNA: isolation by the signal sequence trap and chromosomal mapping. Genomics. 1995; 30(2):312-319.

Nedospasov S.A., Kuprash D.V. Tumor necrosis factor and lymphotoxin: physiological function and role in cytokine and anti-cytokine therapy. Russkiy zhurnal "SPID, rak i obshchestvennoe zdorove" = Russian Journal of AIDS, Cancer, and Public Health. 2008; 12(1):69-76.

Onder L., Danuser R., Scandella E., Firner S., Chai Q., Hehlgans T., Stein J.V., Ludewig B. Endothelial cell-specific lymphotoxin-β receptor signaling is critical for lymph node and high endothelial venule formation. J. Exp. Med. 2013;210(3):465-473. DOI 10.1084/jem.20121462

Seddon J.M., Berggren K.T., Fleeman L.M. Evolutionary history of DLA class II haplotypes in canine diabetes mellitus through single nucleotide polymorphism genotyping. Tissue Antigens. 2010;75(3):218-226. DOI 10.1111/j.1399-0039.2009.01426.x

Sheldon B.C., Verhulst S. Ecological immunology: costly parasite defences and trade-offs in evolutionary ecology. Trends Ecol. Evol. 1996;11(8):317-321.

Tierney R., Kirby H., Nagra J., Rickinson A., Bell A. The Epstein-Barr virus promoter initiating B-cell transformation is activated by RFX proteins and the B-cell-specific activator protein BSAP/Pax5. J. Virol. 2000;74(22):10458-10467.

Trapezov O.V. Darwinism and the lessons of practical selection in Russia. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2009;13(2):249-297.

van der Most P.J., de Jong B., Parmentier H.K., Verhulst S. Trade-off between growth and immune function: a meta-analysis of selection experiments. Funct. Ecol. 2011;25(1):74-80. DOI 10.1111/j.1365-2435.2010.01800.x

Zhao F.Q. Octamer-binding transcription factors: genomics and functions. Front Biosci. 2013;18:1051-1071.

Zhang G., Li C., Li Q., Li B., Larkin D.M., Lee C., Storz J.F., Antunes A., Greenwold M.J., Meredith R.W., Ödeen A., Cui J., Zhou Q., Xu L., Pan H., Wang Z., Jin L., Zhang P., Hu H., Yang W., Hu J., Xiao J., Yang Z., Liu Y., Xie Q., Yu H., Lian J., Wen P., Zhang F., Li H., Zeng Y., Xiong Z., Liu S., Zhou L., Huang Z., An N., Wang J., Zheng Q., Xiong Y., Wang G., Wang B., Wang J., Fan Y., da Fonseca R.R., Alfaro-Núñez A., Schubert M., Orlando L., Mourier T., Howard J.T., Ganapathy G., Pfenning A., Whitney O., Rivas M.V., Hara E., Smith J., Farré M., Narayan J., Slavov G., Romanov M.N., Borges R., Machado J.P., Khan I., Springer M.S., Gatesy J., Hoffmann F.G., Opazo J.C., Håstad O., Sawyer R.H., Kim H., Kim K.W., Kim H.J., Cho S., Li N., Huang Y., Bruford M.W., Zhan X., Dixon A., Bertelsen M.F., Derryberry E., Warren W., Wilson R.K., Li S., Ray D.A., Green R.E., O'Brien S.J., Griffin D., Johnson W.E., Haussler D., Ryder O.A., Willerslev E., Graves G.R., Alström P., Fjeldså J., Mindell D.P., Edwards S.V., Braun E.L., Rahbek C., Burt D.W., Houde P., Zhang Y., Yang H., Wang J., Avian Genome Consortium; Jarvis E.D., Gilbert M.T., Wang J. Comparative genomics reveals insights into avian genome evolution and adaptation. Science. 2014;346(6215):1311-1320. DOI 10.1126/science.1251385

Вавиловский журнал генетики и селекции. 2015;19(6):707-714
Doi 10.18699/VJ15.086

Идентификация микросателлитных локусов по данным секвенирования BAC-клонов и их физическое картирование на хромосому 5B мягкой пшеницы

М.А. Нестеров[1], Д.А. Афонников[1], Е.М. Сергеева[1], Л.А. Мирошниченко[2], М.К. Брагина[1], А.О. Брагин[1], Г.В. Васильев[1], Е.А. Салина[1]

1 Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия 2 Федеральное государственное бюджетное учреждение науки «Институт математики им. С.Л. Соболева Сибирского отделения Российской академии наук», Новосибирск, Россия

Необходимость изучения микросателлитных локусов пшеницы, в первую очередь, обусловлена актуальностью работ по выявлению полиморфных маркеров для участков хромосом, определяющих хозяйственно ценные признаки. В настоящей работе проведено насыщение отдельных районов короткого плеча хромосомы 5B (5BS) мягкой пшеницы SSr-маркерами, разработанными по данным секвенирования BAc-к лонов. 130 клонов, отобранных случайным образом из BAC-библиотеки 5BS, были секвенированы на платформе ionTorrent и собраны в контиги с использованием программы Mir A. Характеристики сборки (N50 = 4 136 п. н.) сравнимы с таковыми для сборок генома пшеницы и родственных видов, полученными в последнее время, и приемлемы для решения задачи идентификации микросателлитных локусов. Для выявления последовательностей ДНК с повторяющейся единицей 2–4 п. н. использовался алгоритм, основанный на свойствах сложностных разложений, формирующихся в режиме скользящего окна. По данным анализа 17 770 контигов общей протяженностью 25 879 921 п. н., разработано 113, 79 и 67 маркеров микросателлитных (SSr) лок усов с повторяющейся единицей 2, 3 и 4 п. н. соответственно. SSr-марк еры с мотивом 3 п. н. были проверены на нулли-тетрасомных линиях пятой гомеологичной группы хромосом сорта пшеницы Чайниз Спринг (cS). Вы явлен 21 маркер, специфичный для хромосомы 5B. Были локализованы 8 маркеров в дистальном районе хромосомы (бин 5BS6) с использованием серии делеционных линий cS по 5BS. Д ля 8 и 4 маркеров определена локализация в интерстициальном районе в бинах 5BS5 и 5BS4 соответственно, один маркер был локализован в прицентромерном бине. Сравнительный анализ распределения тринуклеотидных микросателлитов по хромосоме 5B пшеницы и у различных видов злаков указывает на пролиферацию и поддержание количественного содержания повтора (AAG)n в процессе эволюции злаков.

Ключевые слова: мягкая пшеница; Triticum aestivum; BAC-клоны; IonTorrent; MIRA; микросателлиты; SSR-маркеры; хромосома 5B; (AAG)n.

Received 21.09.2015
Accepted for publication 20.10.2015

e-mail: salina@bionet.nsc.ru

Identification of microsatellite loci according to BAC sequencing data and their physical mapping to the bread wheat 5B chromosome

M.A. Nesterov[1], D.A. Afonnikov[1], E.M. Sergeeva[1], L.A. Miroshnichenko[2], M.K. Bragina[1], A.O. Bragin[1], G.V. Vasiliev1, E.A. Salina[1]

1 Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
2 Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia

The shortage of polymorphic markers for the regions of wheat chromosomes that encode commercially valuable traits determined the need for studying wheat microsatellite loci. in this work, SSr markers for individual regions in the short arm of bread wheat chromosome 5B (5BS) were designed based on sequencing data for BAc clones, and the regions of the corresponding chromosome were saturated with these markers. Totally, 130 randomly selected BAc clones from the 5BS library were sequenced on the ion Torrent platform and assembled in contigs using Mir A software. The assembly characteristics (n 50 = 4 136 bp) are comparable to the recently obtained data for wheat and relative species and acceptable for identific tion of microsatellite loci. An algorithm utilizing the properties of complexity decompositions in the sliding-window mode was used to detect Dn A sequences with a repeat unit of 2–4 bp. Analysis of 17 770 contigs with the total length of 25 879 921 bp allowed for designing 113, 79, and 67 microsatellite (SSr) loci with a repeat unit of 2, 3, and 4 bp, respectively. The SSr markers with a motif of 3 bp were tested using nullitetrasomic lines of c hinese Spring wheat homoeologous group 5. Thus, 21 markers specific or chromosome 5B were detected. Eight of these markers were mapped to the distal region of this chromosome (bin 5BS6) using a set of c hinese Spring deletion lines for 5BS. Eight and four markers were mapped to the interstitial region (bins 5BS5 and 5BS4, respectively). o ne marker was mapped to a pericentromeric bin. A comparative analysis of the distribution of trinucleotide microsatellites over wheat chromosome 5B and in different cereal species suggests that the $(AAG)_n$ repeat has proliferated and has been maintained during the evolution of cereals.

Key words: bread wheat; *Triticum aestivum*; BAc clones; ion Torrent; Mir A; microsatellites; chromosome 5B; SSr markers; $(AAG)_n$.

HOW TO CITE THIS ARTICLE?

nesterov M.A., Afonnikov D.A., Sergeeva E.M., Miroshnichenko l .A., Bragina M.K., Bragin A.o ., Vasiliev G.V., Salina E.A. identification of microsatellite loci according to BAc sequencing data and their physical mapping to the bread wheat 5B chromosome. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2015;19(6):707-714. Doi 10.18699/VJ15.086

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?

Нестеров МА., Афонников Д.А., Сергеева е.М, Мирошниченко л.А., Брагина МК., Брагин А.О., Васильев Г.В., Салина е.А. Идентификация микросателлитных локусов по данным секвенирования BAC-клонов и их физическое картирование на хро-мосому 5В мягкой пшеницы. Вавиловский журнал генетики и селекции. 2015;19(6):707-714. Doi 10.18699/VJ15.086

Физическое картирование хромосом, а именно прямая локализация последовательностей ДНК на хромосомы, впервые получило свое развитие с появлением методов гибридизации *in situ*. Физическое картирование можно успешно проводить с использованием различных генетических ресурсов растений, включающих анеуплоидные, замещенные, интрогрессивные и делеционные линии, в связи с чем его иногда называют цитогенетическим картированием (Sourdille et al., 2004). В век геномного секвенирования термин «физическое картирование» все чаще используется при локализации протяженных контигов ДНК или BAC-клонов в определенном участке хромосомы, при этом методы локализации могут быть различны, включая в том числе и биоинформатические подходы (Paux et al., 2008).

Мягкая пшеница *Triticum aestivum* L. ($2n = 42$) является естественным аллополиплоидом с геномной формулой BBAADD (Feldman, 2001), в образовании которого принимали участие диплоидные виды *Triticum* и *Aegilops*. Э. Сирс (Sears, 1966) показал, что хромосомы трех геномов мягкой пшеницы можно разбить на 7 гомеологических групп, внутри которых одна хромосома в экстра- (тетра-) дозе компенсирует нарушения, обусловленные отсутствием другой. Способность гомеологичных хромосом служить буфером при потере хромосом или их фрагментов привела к созданию коллекций анеуплоидных и делеционных линий, которые явились основой для нового этапа развития генетических и молекулярных работ по анализу генома пшеницы.

Первый этап таких работ был связан с использованием серии анеуплоидных линий мягкой пшеницы для изучения генетического вклада каждой хромосомы в наследование различных признаков пшеницы, локализации и распределения маркеров и генов по группам сцепления, анализа взаимодействия отдельных генов при формировании признаков (Plaschke et al., 1996).

Разработка новых коллекций, основанных на стабильности генома полиплоидных форм при потере отдельных участков хромосом, связана с созданием серии делеционных линий. Коллекции этих линий по каждому хромосомному плечу были тщательно охарактеризованы цитологическими методами (http://www.ksu.edu/wgrc/Germplasm/Deletions/). Создание перекрывающихся делеционных линий явилось необходимым этапом для анализа генома растений. В последние годы данная коллекция интенсивно используется Консорциумом по секвенированию генома пшеницы при построении физических карт индивидуальных хромосом (http://www.wheatgenome.org/Projects/IWGSC-Bread-Wheat-Projects/Physicalmapping).

Несмотря на то что число маркеров для генома пшеницы существенно увеличилось в последние пять лет, особенно с появлением первых результатов его секвенирования и SNP-чипов для анализа генома пшеницы, задача по разработке новых маркеров все еще остается актуальной (Akhunov et al., 2009; Brenchley et al., 2012). Сопоставление физических и генетических карт хромосом мягкой пшеницы указывает на неравномерность распределения молекулярных маркеров по хромосоме (Sourdille et al., 2004; Тимонова и др., 2013). Микросателлитные локусы, или тандемные повторы с повторяющей единицей от 2 до 6 п. н. (Simple Sequence Repeat, SSR) успешно используются для разработки SSR-маркеров. Преимущество этих маркеров заключается в высоком уровне полиморфизма и стабильности в получении результатов, их недостаток – более низкая частота встречаемости относительно SNP- и ISBP-маркеров (Li et al., 2002).

Короткое плечо хромосомы 5В имеет протяженность 290 млн п. н. Ряд генов, влияющих на устойчивость к грибным патогенам, скрещиваемость с рожью, мягкозерность зерна, устойчивость к гессенской мухе, локализованы на данном плече хромосомы (http://www.shigen.nig.ac.jp/wheat/komugi/genes/symbol ClassList.jsp). В настоящий момент количество маркеров на физической и генетической картах хромосомы 5BS явно недостаточно для маркирования генов и их дальнейшего использования для позиционного клонирования генов или их применения в маркер-ориентированной селекции.

Разработка маркеров на основе BAC-клонов преследует две цели, одна из которых – расширение пула маркеров, вторая – разработка маркеров непосредственно к BAC-клонам для их последующей локализации на хромосоме. Концевое секвенирование BAC-клонов активно проводилось в последние годы для решения этих задач (Paux et al., 2008). Однако при использовании данного метода эффективность разработки маркеров крайне низка из-за ограничений по протяженности и количеству секвенированных последовательностей (не более двух) для индивидуального BAC-клона.

Настоящая работа направлена на разработку новых SSR-маркеров по данным высокопроизводительного секвенирования пула BAC-клонов короткого плеча хромосомы 5В и их локализацию на физической карте хромосомы.

Материалы и методы

растительный материал

В работе использовались сорт мягкой пшеницы Чайниз Спринг; нуллитетрасомные линии N5BT5A (хромосома 5B замещена на 5A) и N5BT5D (хромосома 5B замещена на 5D); дителосомная линия Dt5BL (отсутствует короткое плечо 5BS) и набор из 10 делеционных линий мягкой пшеницы сорта Чайниз Спринг по хромосоме 5B, полученных и описанных ранее (Endo, Gill, 1996). Делеционные линии определяют 7 делеционных участков (бинов) 5BS: C-5BS3-0.41; 5BS3-0.41-0.42; 5BS2-0.42-0.43; 5BS4-0.43-0.56; 5BS8-0.56-0.71; 5BS5-0.71-0.81; 5BS6-0.81-1.00. Номенклатура делеционных линий и бинов подробно описана Л. Ки с коллегами (Qi et al., 2003). Делеционные линии были любезно предоставлены центром генетических ресурсов Канзасского университета (http://www.k-state.edu/wgrc/Germplasm/Deletions/del_index.html). Линии были размножены и проверены на наличие делеций (Тимонова и др., 2013).

Bac-клоны

BAC-клоны 5BS получены в Институте экспериментальной ботаники (Оломоук, Чехия) и любезно предоставлены профессором Я. Долежелем.

Выделение ДнК

Выделение и очистку ДНК из растений проводили согласно ранее описанной методике (Plaschke et al., 1995).

ДНК из BAC-клонов выделяли с использованием набора NucleoSpin 96 Flash (Macherey-Nagel), в основе которого лежит SDS-щелочной лизис. Методика была адаптирована под имеющуюся аппаратуру с включением дополнительной стадии очистки методом центрифугирования.

анализ пцр

Анализ проводили в реакционной смеси объемом 20 мкл, содержащей 50 нг ДНК-матрицы, 65 мМ трис-HCl (pH 8,9), 1,5 мМ $MgCl_2$, 16 мМ $(NH_4)_2SO_4$; 0,05 % Tween 20, по 0,2 мМ каждого дНТФ, 1 ед. ДНК-полимеразы Taq, 0,25 мкМ прямого и обратного специфических праймеров (Дополнительные материалы 1[1]). Реакцию проводили при денатурации – 30 с при 94 °C; отжиге – 30 с при 55 °C; полимеризации – 30 с при 72 °C; числе циклов – 35. ПЦР осуществляли в амплификаторе «BioRad T100» (США). ПЦР-фрагменты разделяли электрофорезом в 2 %-м агарозном геле или 10 % ПААГ, в зависимости от уровня полиморфизма между фрагментами.

секвенирование на платформе iontorrent

ДНК 130 индивидуально выделенных BAC-клонов была пулирована в приблизительно эквимолярных количествах. Для приготовления библиотек IonTorrent 5 мкг ДНК фрагментировано ультразвуком на приборе Covaris, фрагменты длиной 330–450 п.н. выделены на приборе LabchipXT, после репарации концов проведено лигирование адаптеров с помощью набора Ion Plus Fragment Library Kit. Библиотека мультиплицирована с 8 циклами амплификации, очищена набором AmPureXP и проверена на приборе Bioanalyzer2100. Далее проведены эмульсионная ПЦР с набором Ion PGM Template OT2 400 Kit и секвенирование на приборе IonTorrent на расчетную длину 400 п.н. с использованием чипа максимальной емкости – 318.v2.

Для сборки последовательностей ДНК была использована программа MIRA (Chevreux et al., 1999). Оценку глубины покрытия сборки генома и доли ПЦР-дупликатов проводили на основе картирования исходных секвенированных фрагментов длиной более 30 п.н. на последовательности контигов программой Bowtie2 (Langmead, Salzberg, 2012) и дальнейшей обработки выравниваний программой Samtools (Li et al., 2009).

идентификация SSr -локусов, разработка маркеров

Для выявления последовательностей с повторяющимся мотивом 2–4 п.н. использовался алгоритм, основанный на свойствах сложностных разложений, формирующихся в режиме скользящего окна (Gusev et al., 2009).

Разработка праймеров к повторяющимся мотивам 2–4 п.н. (SSR-локусам) осуществлялась с помощью программы PrimerQuest (http://eu.idtdna.com/PrimerQuest/Home/Index) при установке следующих параметров: длина ампликона (150–450 символов); длина праймера (17–25 символов); (C+G) состав (40–60 %); Tm (59–65 °C); C или G на 3′-конце праймера. Остальные параметры PrimerQuest использовались по умолчанию.

Из нескольких (до 5) пар праймеров, предложенных программой PrimerQuest, для каждого SSR-локуса выбиралась одна пара. Преимуществом пользовались праймеры с высокой сложностью: лингвистической или ДНК-ориентированной мерой LZ (Gusev et al., 1999). Исключались праймеры с неравновесным составом нуклеотидов, с наличием длинных (G+C)-, (A+G)-, (A+C)-, (A+T)-, (C+T)-, (G+T)-трактов, а также праймеры, в которых присутствовали повторяющиеся мотивы.

результаты

секвенирование Bac-клонов и сборка контигов

Количество клонов, отобранных для секвенирования, было рассчитано исходя из не менее чем стократного прочтения изучаемых последовательностей ДНК, средней длины BAC-клона около 100 т.п.н. и размера чипа IonTorrent максимальной емкости. Таким образом, расчетное число BAC-клонов составило 130.

В ходе секвенирования пула BAC-клонов были получены результаты по плотности нанесения и качеству секвенирования выше средних, наблюдаемых при использовании данного метода. Распределение длин прочитанных фрагментов (рис. 1) показало, что большинство из них имеет длину 150–400 п.н.

Общий объем полученных данных составил более 5,5 млн индивидуальных чтений со средней длиной 239, в сумме более 1,3 млрд нуклеотидов, доля последовательностей с Q > 20 составила 74,6 %. В результате было получено расчетное стократное среднее прочтение взятых в анализ BAC-клонов.

[1] Дополнительные материалы 1–3 см. в Приложении 5 по адресу: http://www.bionet.nsc.ru/vogis/download/pict-2015-12/appx5.pdf

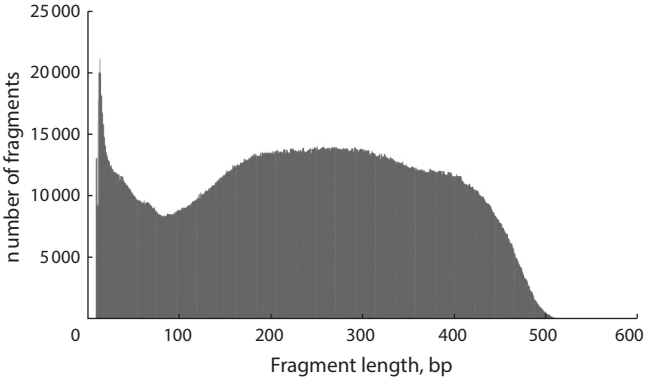


Fig. 1. l ength distribution of the sequenced fragments from the BAc library.

Одним из первых этапов при сборке последовательностей ДНК после секвенирования BAC-клонов является удаление последовательностей ДНК вектора и ДНК *E. coli*. В исходных данных было выявлено менее 2,2 % ДНК *E. coli* (*Escherichia coli* str. K12 substr. DH10B), что указывало на высокое качество выделяемой ДНК BAC-клонов.

Результаты сборки последовательностей после удаления ДНК *E. coli* с использованием программы MIRA представлены в таблице. Отметим, что характеристики для контигов, имеющих длину более 500 п. н. (~5 тыс. контигов), демонстрируют почти двукратное увеличение параметров N50 и более чем двукратное для параметров N90 и N95.

Для дополнительной оценки качества секвенирования полученных библиотек мы оценили содержание ПЦР дупликатов в них и среднюю глубину покрытия последовательности ДНК BAC-клонов прочтениями. Для этого на собранные контиги мы картировали исходные прочтения (длиной более 30 п. н.) и по данным выравнивания вычисляли указанные характеристики. Оказалось, что доля ПЦР дупликатов составила около 14 %, а средняя глубина покрытия – примерно 25 прочтений на позицию контига. Кроме того, программа MIRA выдает оценку средней глубины покрытия для контигов, имеющих длину 500 и среднее покрытие не менее 15. Значение среднего покрытия для таких качественных контигов составило ~40 прочтений на позицию. Полученные оценки позволяют заключить, что библиотеки и собранные на основе их секвенирования контиги являются пригодными для решения задачи идентификации SSR-маркеров в последовательностях BAC-клонов.

идентификация и маркирование SSr -локусов

Исходный материал (секвенированные последовательности ДНК) содержит повторяющиеся фрагменты. Некоторые последовательности повторяются целиком, другие являются фрагментами более длинных последовательностей. Как следствие, часть последовательностей ДНК, входящих в состав SSR-локусов, тоже повторяются. Для выявления повторов была разработана программа, основанная на алгоритме поиска «образцов», которая позволила удалять из анализа идентичные SSR-локусы.

На первом этапе, по данным анализа 17 770 контигов общей протяженностью 25 879 921 п. н., выявлено 253, 156 и 27 последовательностей ДНК, содержащих SSR-локусы с периодом повторности 2, 3 и 4 п. н. соответственно и длиной не менее 20 нуклеотидов. На втором этапе было проведено удаление повторяющихся мотивов, расположенных на концах контигов, повторяющихся SSR-локусов и локусов, окружение которых не позволяет разработать праймеры. В результате было отобрано 113 праймеров для SSR-локусов с повторяющимся мотивом 2 п. н., 79 – для SSR-локусов с мотивом 3 п. н. и 23 – для SSR-локусов с мотивом 4 п. н.

Для 67 пар праймеров, подобранных к SSR-локусам с повторяющимся мотивом 3 п. н., была проведена ПЦР с ДНК нулли-тетрасомных и делеционных линий и изучена возможность их использования в качестве маркеров различных участков хромосомы 5B пшеницы (Доп. материалы 1).

Оценка и физическое картирование новых SSr -маркеров

Результаты ПЦР ДНК сорта CS с праймерами, разработанными к SSR-локусам с повторяющимся мотивом 3 п. н., представлены в Доп. материалах 2. Продукты амплификации для большинства изученных маркеров были четкими и хорошо выявляемыми как в ПААГ, так и в агарозных гелях. Только четыре (Xicgc740, Xicgc1738, Xicgc1755, Xicgc1931) из 67 изученных SSR-маркеров в стандартных ПЦР условиях давали слабые фрагменты амплификации.

Физическое картирование маркеров на хромосоме 5BS проводили с использованием 7 делеционных линий (5BS6, 5BS5, 5BS1, 5BS8, 5BS4, 5BS2 и 5BS3) (рис. 2; Доп. материалы 3). В качестве контроля использовали сорт мягкой пшеницы CS и нулли-тетрасомная линия CSN5BT5A (хромосома 5B замещена на 5A).

На рис. 2 представлено расположение точек разрывов у 7 делеционных линий короткого плеча хромосомы 5B. Для примера, в делеционной линии 5BS6 отсутствует дистальный участок хромосомы до точки разрыва 5BS6 (делеционный бин 5BS6), а у делеционной линии 5BS5 делеция крупнее и распространяется до точки разрыва 5BS5 (делеционный бин 5BS5). Анализ серии делеционных линий пшеницы позволяет локализовать маркеры на определенном участке (бине) хромосомы, т. е. провести их физическое картирование. Примеры локализации маркеров в бинах 5BS6, 5BS5, 5BS4 и C-5BS3 приведены на рисунке в Доп. материалах 3.

Всего при сравнительном анализе ПЦР фрагментов CS и линии CSN5BT5A на хромосоме 5B локализован 21 SSR-маркер. Остальные 46 маркеров имели близкие по длине фрагменты амплификации между хромосомами 5-й гомеологичной группы. Анализ 21 полиморфного SSR-маркера с серией делеционных линий CS по хромосоме 5B позволил локализовать данные маркеры на физическую карту 5BS. Так, 8 маркеров (Xicgc178, Xicgc178_2, Xicgc284, Xicgc456, Xicg14c009, Xicg15c020, Xicg16c004_2, Xicg16c041) были локализованы в дистальном районе хромосомы (бин 5BS6), 8 маркеров (Xicgc122, Xicgc131, Xicgc229, Xicgc299, Xicgc307, Xicgc342,

Main parameters of the assembly of BAc library fragments

Parameter	All contigs	l ong contigs (> 500 bp)
n umber of contigs	17 770	4 780
Total consensus, bp	25 879 921	18 560 927
l ongest contig size, bp	250 809	250 809
n 50 contig size, bp	4 136	7 767
n 90 contig size, bp	504	1 431
n 95 contig size, bp	416	942

Xicgc686, Xicgc1572) – в бине 5BS5, четыре маркера (Xicgc495, Xicgc498, Xicgc69, Xicgc1699) – в бине 5BS4 и Xicgc1988 – в бине C-5BS3 (рис. 2).

Обсуждение

Подходы и результативность сборки последовательностей Дн К у злаков

Для сборки генома использовали программу MIRA (Chevreux et al., 1999). С одной стороны, выбор был обусловлен тем, что эта программа достаточно хорошо зарекомендовала себя при сборке бактериальных геномов на основе библиотек, полученных по технологии Ion Torrent (Loman et al., 2012). С другой, эта программа была ранее использована при сборке геномных фрагментов ряда растений, в частности, для сборки BAC-клонов ячменя (Pasquariello et al., 2014), подсолнечника (*Helianthus annuus* L.) (Staton et al., 2012), для сборки фрагментов генома барбадосского ореха (*Jatropha curcas*) (Sato et al., 2011), а также генома пластид гнездовки настоящей (*Neottia nidus-avis*) (Logacheva et al., 2011) (во всех указанных случаях для секвенирования применялась технология Roche 454). В обзоре Штайна и Шторнагеля (Stein, Steuernagel, 2014), посвященном работам в области секвенирования генома ячменя, эта программа характеризуется как очень удачная для анализа относительно небольших объемов данных, полученных на разных платформах (Illumina GAIIx, Hiseq, Roche 454) и содержащих большое количество повторов.

В данной работе нам удалось получить достаточно качественные результаты ассемблирования геномных фрагментов (таблица). Представляется интересным сравнить характеристики сборки с аналогичными данными по секвенированию геномных последовательностей злаковых растений, в особенности родственных пшенице. Анализ последних работ, посвященных секвенированию геномов мягкой пшеницы и родственных видов по технологии «дробовика» (shot-gun sequencing), показал, что характеристики полученной сборки BAC-клонов достаточно хорошо согласуются с результатами других групп. В частности, одной из наиболее информативных характеристик общего качества сборки является параметр N50 (значение, при котором контиги равной и большей длины составляют половину суммарной протяженности сборки). В данной работе она составила около 4 тыс. п. н., а при анализе контигов с длиной более 500 значение этого параметра стало 7 700 п. н. (таблица).

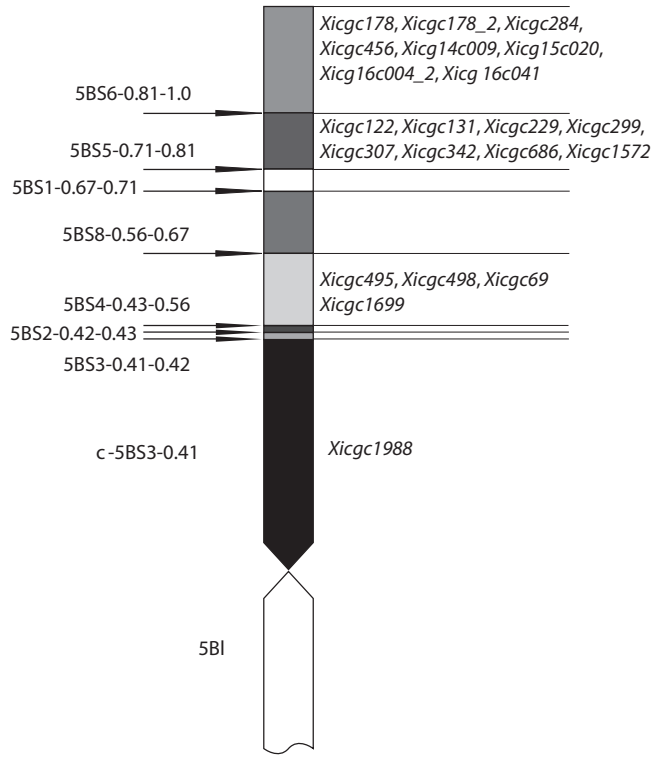


Fig. 2. l ocation of SSr markers in the deletion bins of chromosome 5BS.

Сборка генома мягкой пшеницы на основе секвенирования с низким покрытием (~5×) по технологии 454 (Brenchley et al., 2012) характеризовалась величиной N50 для контигов около 900 п. н. Для сборки генома ячменя при секвенировании методом Illumina GAIIx и 50-кратном покрытии (International Barley Genome Sequencing Consortium, 2012) значение этого параметра для контигов составило 1 425 п. н. Для сборки генома *Aegilops tauschii* при секвенировании методом Roche 454 и 90-кратном покрытии значение N50 для контигов составило ~4 500 п. н. (Jia et al., 2013). Для сборки генома *Triticum urartu* значение N50 для контигов составило 3 400 п. н. при секвенировании методом Illumina HiSequation (2000) и 91-кратном покрытии (Ling et al., 2013). В работе, посвященной секвенированию и сборке последовательностей хромосом мягкой пшеницы с использованием технологии Illumina

и с покрытием между 30× и 241×, значение параметра N50 для разных хромосом находится в пределах 0,5 ~ 4,3 kb (International Wheat Genome Sequencing Consortium, 2014). Для недавней сборки генома пшеницы, полученной Чапманом с коллегами, методом Illumina по данным нескольких линий при общем ~ 175-кратном покрытии, N50 для контигов с длиной более 1 тыс. п. н. определяется значением ~ 8 300 (Chapman et al., 2015). Таким образом, характеристики сборки, полученные в нашей работе, в целом хорошо согласуются с таковыми для недавних вариантов сборок генома (Jia et al., 2013; Ling et al., 2013; Chapman et al., 2015) и достаточны для решения поставленных в работе задач.

Следует также отметить, что практически со всеми праймерами, разработанными по результатам сборки контигов, были получены хорошие продукты амплификации как на BAC-клонах, используемых для секвенирования, так и на геномной ДНК пшеницы.

Особенности организации микросателлитных локусов хромосомы 5B

Микросателлитные последовательности ДНК встречаются с высокой частотой во всех изученных геномах эукариот (Tautz, Renz, 1984; Li et al., 2002). Средняя частота встречаемости микросателлитов в геноме эукариот оценивается как 1 на 10 000 п. н. (Brown et al., 1996). Количество копий SSR может варьировать и достигать 100 и более повторностей в одном локусе (Tautz, Renz, 1984; Li et al., 2002). Первые оценки распределения по хромосомам микросателлитных локусов были сделаны по результатам картирования генома растений и на основании прямой *in situ* гибридизации. При изучении распределения микросателлитов для ряда объектов, например, у сахарной свеклы, методом *in situ* гибридизации было показано, что они группируются в определенных районах хромосом (Schmidt, Heslop-Harrison, 1996). При картировании микросателлитов томатов показано, что они локализуются в прицентромерной области хромосом (Areshchenkova, Ganal, 1999). Однако у большинства растительных видов микросателлитные локусы имеют более равномерное распределение по геному, что способствовало их успешному использованию в качестве маркеров как для генотипирования генома растений, так и в последующем их использовании в маркер-ориентированной селекции (Mason, 2015; http://maswheat.ucdavis.edu/protocols/).

Анализ результатов секвенирования выборки BAC-клонов позволил дополнительно разработать 67 маркеров для короткого плеча хромосомы 5B, на котором расположен ряд генов, влияющих на формирование количественных признаков мягкой пшеницы и устойчивость к абиотическим и биотическим факторам внешней среды (http://www.shigen.nig.ac.jp/wheat/komugi/genes/symbolClassList.jsp). Примечательно, что 21 маркер выявил фрагменты амплификации, специфичные для короткого плеча хромосомы 5B, что позволило локализовать их на физической карте хромосомы 5B (в делеционных бинах). Наибольшее число маркеров из числа локализованных на хромосому (76 %) присутствовало в дистальном участке хромосомы (в бинах 5BS6, 5BS5) и только один маркер – в прицентромерном бине хромосомы 5BS (рис. 2; Доп. материалы 2). Аналогичные результаты по распределению маркеров были получены нами ранее при локализации SSR-маркеров из базы данных по пшенице на хромосому 5B (Тимонова и др., 2013). Однако, в отличие от ранее полученных результатов, в настоящей работе 19 % маркеров было локализовано в интерстициальном бине 5BS4. Какой-либо зависимости от структуры микросателлитного локуса и его локализации на хромосоме по данным проведенного картирования выявлено не было (Доп. материалы 2). В то же время прямая гибридизация различных групп микросателлитов на хромосомы указывает на преимущественную их локализацию в прицентромерном районе (Cuadrado et al., 2000, 2008). Обратная закономерность, выявленная нами при физическом картировании B-специфичных хромосомных локусов, указывает на то, что, по-видимому, прицентромерные районы достаточно консервативны у гомеологичных хромосом, в том числе и у 5-й группы хромосом пшеницы, а отсутствие полиморфизма не позволило локализовать большинство (70 %) разработанных SSR-маркеров на хромосоме 5BS.

Распространенность различных микросателиттов в целом по геному в пределах одного вида зависит от структуры повторяющейся единицы и от ее протяженности. Так, анализ результатов секвенирования отдельных хромосом или всего генома ряда злаковых и кукурузы (Zhang et al., 2007; Qu, Liu, 2013; Sergeeva et al., 2014) указывает на то, что наиболее часто встречаются динуклеотидные повторы, затем частота встречаемости убывает при увеличении длины повторяющегося мотива. Полученные нами результаты по анализу микросателитных локусов, по данным секвенирования выборки из 130 BAC-клонов, также указывают на более высокую распространенность динуклеотидных последовательностей (253 последовательности), по сравнению с три- (156) и тетра-нуклеотидными (27) повторяющимися последовательностями ДНК. Наблюдаемая тенденция в распространенности повторяющихся мотивов, в зависимости от длины, соответствует ожидаемому распределению мотива, обратно пропорциональному его длине. Встречаемость ди-, три- и тетра-повторов варьирует также и в зависимости от участка генома. Оценка распределения ди- и тринуклеотидных повторов в транслируемых и нетранслируемых районах генных локусов показала, что встречающиеся в кодирующих районах генов микросателлиты более чем на 87 % состоят из тринуклеотидов, причем это характерно, например, как для риса, так и для кукурузы (Zhang et al., 2007; Qu, Liu, 2013). Это связано с тем, что увеличение копийности в тринуклеотидном повторе не приводит к сдвигу рамки считывания, и в ряде случаев может поддерживаться естественным отбором. Межвидовые различия у злаков заключаются в распространенности разных тринуклеотидных повторов. Так, по данным анализа последовательностей ДНК всего генома кукурузы, выявлено, что наиболее распространенный тринуклеотидный повтор имеет структуру $(AGC)_n$, далее следуют повторы $(ACG)_n$, $(CCG)_n$, $(ATC)_n$, $(AAG)_n$ с частотой встречаемости от 2,81 до 2,48 % от общего числа ди- и тринуклеотидных повторов (Qu, Liu, 2013). У риса распределение тринуклеотидных повторов по степени встречаемости несколько иное (Zhang et al., 2007). Наиболее часто встречаются

$(CCG)_n$, затем $(AGG)_n$, $(AGC)_n$, $(ACG)_n$, $(AAG)_n$. Основные различия между видами наблюдаются по двум группам микросателлитов, а именно у кукурузы повторы $(ATC)_n$ встречаются часто, а $(AGG)_n$ – существенно реже, в то время как у риса наоборот. Общей характеристикой изученных геномов злаков (рис, кукуруза) является широкая представленность $(AGC)_n$-, $(ACG)_n$-, $(CCG)_n$-, $(AAG)_n$-повторов. Следует отметить, что в перечисленные выше группы повторов не входят тринуклеотидные мотивы, относящиеся к стоп-кодонам ядерного генома эукариот.

У пшеницы и ее предшественников широкое распространение в геноме получил $(AAG)_n$-повтор. Это подтверждается как полученными результатами (Доп. материалы 2), так и ранее проведенными исследованиями распределения тринуклеотидных микросателлитов методом прямой локализации на хромосомы олигонуклеотидов (Cuadrado et al., 2000, 2008; Adonina et al., 2015). Другой особенностью мягкой пшеницы, и преимущественно B-генома и его предшественника, является широкая представленность $(AGG)_n$-повтора (Доп. материалы 2; Cuadrado et al., 2008). $(AGG)_n$ микросателлит широко встречается и у риса (Zhang et al., 2007). Такая мозаичная пролиферация $(AGG)_n$-повтора в геномах отдельных видов злаков в процессе эволюции, возможно, происходила посредством мобильных элементов.

Таким образом, сравнительный анализ распределения тринуклеотидных микросателлитов по хромосоме 5B пшеницы и у различных видов злаков указывает на поддержание количественного содержания повтора $(AAG)_n$ в процессе эволюции злаков, в то время как содержание других тринуклеотидных повторов существенно варьирует у современных видов злаков и их предшественников.

Acknowledgments

The authors are grateful to the Centre of Plant Structural and Functional Genomics and his Head J. Dolezhel (Institute of Experimental Botany, Olomouc, Czech Republic) for providing BAC-clones.

The work was supported by the Russian Ministry of Science and Education, contract No. 14.604.21.0106 of July 7, 2014, identification number RFMEFI60414X0106.

The facilities of the Bioinformatics shared access center and the Genomics Core Facility of the SB RAS were used in this research.

Conflict of interest

The authors declare no conflict of interest.

References

Akhunov E., Nicolet C., Dvorak J. Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. Theor. Appl. Genet. 2009;119:507-517. DOI 10.1007/s00122-009-1059-5

Adonina I.G., Goncharov N.P., Badaeva E.D., Sergeeva E.M., Petrash N.V., Salina E.A. $(GAA)_n$ microsatellite as an indicator of the A genome reorganization during wheat evolution and domestication. CompCytogen. 2015;9(4):533-547. DOI 10.3897/CompCytogen.v9i4.5120

Areshchenkova T., Ganal M.W. Long tomato microsatellites are predominantly associated with centromeric regions. Genome. 1999;42:536-544.

Brenchley R., Spannagl M., Pfeifer M., Barker G.L.A., D'Amore R., Allen A.M., McKenzie N., Kramer M., Kerhornou A., Bolser D., Kay S., Waite D., Trick M., Bancroft I., Gu Y., Huo N., Luo M.-C., Sehgal S., Gill B., Kianian S., Anderson O., Kersey P., Dvorak J., McCombie W.R., Hall A., Mayer K.F.X., Edwards K.J., Bevan M.W., Hall N. Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature. 2012;491(7426):705-710. DOI 10.1038/nature11650

Brown S.M., Szewc-McFadden A.K., Kresovich S. Development and application of simple sequence repeat (SSR) loci for plant genome analysis. Methods in Genome Analysis in Plants. Boca Raton: CRC Press, 1996.

Chapman J.A., Mascher M., Buluç A., Barry K., Georganas E., Session A., Strnadova V., Jenkins J., Sehgal S., Oliker L., Schmutz J., Yelick K.A., Scholz U., Waugh R., Poland J.A., Muehlbauer G.J., Stein N., Rokhsar D.S. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. Genome Biology. 2015;16(1):26. DOI 10.1186/s13059-015-0582-8

Chevreux B., Wetter T., Suhai S. Genome sequence assembly using trace signals and additional sequence information. Computer science and biology: Proc. of the German Conference on Bioinformatics. 1999:45-56.

Cuadrado A., Schwarzacher T., Jouve N. Identification of different chromatin classes in wheat using *in situ* hybridization with simple sequence repeat oligonucleotides. Theor. Appl. Genet. 2000;101:711-717. DOI 10.1007/s001220051535

Cuadrado A., Cardoso M., Jouve N. Increasing the physical markers of wheat chromosomes using SSRs as FISH probes. Genome. 2008;51(10):809-815. DOI 10.1139/G08-065

Endo T.R., Gill B.S. The deletion stocks of common wheat. J. Hered. 1996;87(4):295-307.

Feldman M. The origin of cultivated wheat. The World Wheat Book. Paris: Lavoisier Publishing, 2001.

Gusev V.D., Miroshnichenko L.A., Chuzhanova N.A. The detection of fractal-like structures in DNA sequences. Information science and computing. Int. Book Series, No. 8: Classification, forecasting, data mining. Sofia: ITHEA, 2009.

Gusev V.D., Nemytikova L.A., Chuzhanova N.A. On the complexity measures of genetic sequences. Bioinformatics. 1999;15(12):994-999. DOI 10.1093/bioinformatics/15.12.994

International Barley Genome Sequencing Consortium. A physical, genetic and functional sequence assembly of the barley genome. Nature. 2012;491(7426):711-716. DOI 10.1038/nature11543

International Wheat Genome Sequencing Consortium. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science. 2014;345(6194):1251788. DOI 10.1126/science.1251788

Jia J., Zhao S., Kong X., Li Y., Zhao G., He W., Appels R., Pfeifer M., Tao Y., Zhang X., Jing R., Zhang C., Ma Y., Gao L., Gao C., Spannagl M., Mayer K.F.X., Li D., Pan S., Zheng F., Hu Q., Xia X., Li J., Liang Q., Chen J., Wicker T., Gou C., Kuang H., He G., Luo Y., Keller B., Xia Q., Lu P., Wang J., Zou H., Zhang R., Xu J., Gao J., Middleton C., Quan Z., Liu G., Wang J., International Wheat Genome Sequencing Consortium; Yang H., Liu X., He Z., Mao L., Wang J. *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. Nature. 2013;496(7443):91-95. DOI 10.1038/nature12028

Langmead B., Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9:357-359.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R.; 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 2009;25:2078-2079.

Li Y.-C., Korol A.B., Beiles A., Nevo E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol. Ecol. 2002;11:2453-2465. DOI 10.1046/j.1365-294X.2002.01643.x

Ling H.-Q., Zhao S., Liu D., Wang J., Sun H., Zhang C., Fan H., Li D., Dong L., Tao Y., Gao C., Wu H., Li Y., Cui Y., Guo X., Zheng S.,

Wang B., Yu K., Liang Q., Yang W., Lou X., Chen J., Feng M., Jian J., Zhang X., Luo G., Jiang Y., Liu J., Wang Z., Sha Y., Zhang B., Wu H., Tang D., Shen Q., Xue P., Zou S., Wang X., Liu X., Wang F., Yang Y., An X., Dong Z., Zhang K., Zhang X., Luo M.-C., Dvorak J., Tong Y., Wang J., Yang H., Li Z., Wang D., Zhang A., Wang J. Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature. 2013;496(7443):87-90. DOI 10.1038/nature11997

Logacheva M.D., Schelkunov M.I., Penin A.A. Sequencing and analysis of plastid genome in mycoheterotrophic orchid *Neottia nidus-avis*. Genome Biol. Evol. 2011;3:1296-1303. DOI 10.1093/gbe/evr102

Loman N.J., Misra R.V., Dallman T.J., Constantinidou C., Gharbia S.E., Wain J., Pallen M.J. Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. 2012;30(5):434-439. DOI 10.1038/nbt.2198

Mason A.S. SSR genotyping. Methods Mol. Biol. 2015;1245:77-89. DOI 10.1007/978-1-4939-1966-6_6

Pasquariello M., Barabaschi D., Himmelbach A., Steuernagel B., Ariyadasa R., Stein N., Gandolfi F., Tenedini E., Bernardis I., Tagliafico E., Pecchioni N., Francia E. The barley Frost resistance-H2 locus. Funct. Integr. Genomic. 2014;14(1):85-100. DOI 10.1007/s10142-014-0360-9

Paux E., Sourdille P., Salse J., Saintenac C., Choulet F., Leroy P., Korol A., Michalak M., Kianian S., Spielmeyer W., Lagudah E., Somers D., Kilian A., Alaux M., Vautrin S., Bergès H., Eversole K., Appels R., Safar J., Simkova H., Dolezel J., Bernard M., Feuillet C. Physical map of the 1-Gigabase bread wheat chromosome 3B. Science. 2008;322:101-104. DOI 10.1126/science.1161847

Plaschke J., Ganal M.W., Röder M.S. Detection of genetic diversity in closely related bread wheat using microsatellite markers. Theor. Appl. Genet. 1995;91:1001-1007. DOI 10.1007/BF00223912

Plaschke J., Börner A., Wendehake K., Ganal M.W., Röder M.S. The use of wheat aneuploids for the assignment of microsatellite loci. Euphytica. 1996;89:33-40. DOI 10.1007/BF00015716

Sato S., Hirakawa H., Isobe S., Fukai E., Watanabe A., Kato M., Kawashima K., Minami C., Muraki A., Nakazaki N., Takahashi C., Nakayama S., Kishida Y., Kohara M., Yamada M., Tsuruoka H., Sasamoto S., Tabata S., Aizu T., Toyoda A., Shin-i T., Minakuchi Y., Kohara Y., Fujiyama A., Tsuchimoto S., Kajiyama S., Makigano E., Ohmido N., Shibagaki N., Cartagena J.A., Wada N., Kohinata T., Atefeh A., Yuasa S., Matsunaga S., Fukui K. Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. DNA Res. 2011;18(1):65-76. DOI 10.1093/dnares/dsq030

Sears E.R. Nullisomic-tetrasomic combinations in hexaploid wheat. Chromosome manipulations and Plant Genetics. London: Oliver and Boyd, 1966. DOI 10.1007/978-1-4899-6561-5_4

Sergeeva E.M., Afonnikov D.A., Koltunova M.K., Gusev V.D., Miroshnichenko L.A., Vrána J., Kubaláková M., Poncet C., Sourdille P., Feuillet C., Doležel J., Salina E.A. Common wheat chromosome 5B composition analysis using low-coverage 454 sequencing. Plant Genome. 2014;7(2):1-16. DOI 10.3835/plantgenome2013.10.0031

Schmidt T., Heslop-Harrison J.S. The physical and genomic organization of microsatellites in sugar beet. Proc. Natl Acad. Sci. USA. 1996;93:8761-8765.

Staton S.E., Bakken B.H., Blackman B.K., Chapman M.A., Kane N.C., Tang S., Ungerer M.C., Knapp S.J., Rieseberg L.H., Burke J.M. The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements. Plant J. 2012;72(1):142-153. DOI 10.1111/j.1365-313X.2012.05072.x

Stein N., Steuernagel B. Advances in sequencing the barley genome. Genomics of plant genetic resources. Springer Netherlands, 2014. DOI 10.1007/978-94-007-7572-5_16

Sourdille P., Singh S., Cadalen T., Brown-Guedira G.L., Gay G., Qi L., Gill B.S., Dufour P., Murigneux A., Bernard M. Microsatellite-based deletion bin system for the establishment of genetic-physical map relationships in wheat (*Triticum aestivum* L.). Funct. Integr. Genomics. 2004;4:12-25. DOI 10.1007/s10142-004-0106-1

Tautz D., Renz M. Simple sequences are ubiquitious repetitive component of eukaryotic genomes. Nucl. Acid. Res. 1984;12:4127-4138. DOI: 10.1093/nar/12.10.4127

Timonova E.M., Dobrovol'skaya O.B., Sergeeva E.M., Bildanova L.L., Sourdille P., Feuillet C., Salina E.A. A comparative genetic and cytogenetic mapping of wheat chromosome 5B using introgression lines. Genetika=Genetics (Moscow). 2013;49(12):1200-1206.

Qi L., Echalier B., Friebe B., Gill B.S. Molecular characterization of a set of wheat deletion stocks for use in chromosome bin mapping of ESTs. Funct. Integr. Genomics. 2003;3:39-55. DOI 10.1007/s10142-002-0063-5

Qu J., Liu J. A genome-wide analysis of simple sequence repeats in maize and the development of polymorphism markers from next-generation sequence data. BMC Res. Notes. 2013;6:403. DOI 10.1186/1756-0500-6-403

Zhang Z., Deng Y., Tan J., Hu S., Yu J., Xue Q. A genome-wide microsatellite polymorphism database for the indica and japonica rice. DNA Res. 2007;14:37-45. DOI 10.1093/dnares/dsm005

Вавиловский журнал генетики и селекции. 2015;19(6):715-723
Doi 10.18699/VJ15.109

Клеточная стенка растений и механизмы устойчивости к патогенам

О.Г. Смирнова[1, 2], А.В. Кочетов[1, 2]

1 Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия
2 Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, России

Огромное число грибов, бактерий и вирусов потенциально способны инфицировать ткани и вызывать заболевания растений. устойчивость растений к патогенам основывается на сложной сети конститутивных и индуцированных защитных реакций, в контроле которых задействовано большое число генов. Клеточная стенка является первым препятствием, которое должны преодолеть патогенные микроорганизмы. успешная защита на уровне клеточной стенки может остановить вторжение подавляющего большинства потенциальных фитопатогенов. Разные виды растений различаются по структуре клеточной стенки. Основу клеточной стенки составляет сеть из микрофибрилл целлюлозы, пересекаемых молекулами гемицеллюлозы. В растущих частях растения эта сеть встроена в матрикс из пектиновых полисахаридов. В уже сформировавшихся тканях клеточные стенки усилены лигнином. Кроме полисахаридов, клеточная стенка содержит значительное количество белков, выполняющих структурную и ферментативную функции. Информация о многочисленных белках клеточных стенок разных видов растений представлена в базе данных wallProtDB. Каждый из компонентов клеточной стенки вносит вклад в формирование устойчивости к патогенам. В местах контакта с потенциальными патогенами происходит дополнительное укрепление клеточной стенки и накопление антимикробных вторичных метаболитов. Патогены секретируют ферменты, способные расщеплять компоненты клеточной стенки. В ответ на атаку микробов растение продуцирует ингибиторы микробных гидролитических ферментов. Растение также способно оценивать количество компонентов клеточной стенки. так, мутанты с дефицитом целлюлозы обычно имеют повышенный уровень лигнификации и усиление защитного ответа. Возникающие после действия микробных ферментов низкомолекулярные фрагменты клеточной стенки выполняют сигнальную функцию, усиливая защитную реакцию растения. таким образом, клеточная стенка является динамической структурой, способной предотвращать проникновение большинства потенциальных патогенов и запускать разные варианты иммунного ответа. Реконструкция генных сетей, контролирующих структурно-функциональную организацию клеточной стенки в процессе роста и в условиях биотических и абиотических стрессов, необходима для понимания молекулярных механизмов развития и стрессоустойчивости. В обзоре рассматриваются механизмы специфической и неспецифической устойчивости растений к патогенам различной природы, связанные с клеточной стенкой. Обсуждаются структура клеточной стенки и роль различных компонентов в детекции инвазии фитопатогенов и индукции защитных механизмов.

Ключевые слова: врожденный иммунитет; клеточная стенка; зерновые культуры; грибные патогены; листовая ржавчина, неспецифическая устойчивость.

Received 30.09.2015
Accepted for publication 21.10.2015

e-mail: planta@bionet.nsc.ru

Plant cell wall and the mechanisms of resistance to pathogens

O.G. Smirnova[1, 2], A.V. Kochetov[1, 2]

1 Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
2 Novosibirsk State University, Novosibirsk, Russia

A huge variety of phytopathogens (viruses, bacteria, fungi) are potentially able to infect plant tissues and cause diseases. n umerous plant genes control a complex network of defense mechanisms based on both constitutive and inducible processes. The cell wall is a primary barrier the pathogens have to penetrate to start the infection process. However, it is able to block invasion by most non-specific po ential pathogens. The cell wall structure may differ in various plant species. it is based on the net of cellulose microfib ils linked by hemicellulose molecules. Pectin and lignin are the other important cell wall constituents. Dozens of proteins inside the cell wall are involved in structural and metabolic processes as well as in signal transduction and regulatory circuits (more information is available in wallProtDB database). Each of these components contributes to resistance to pathogens. At the points of contact with potential pathogens cell wall structural changes and accumulation of metabolites with antimicrobial, antifungal or antiviral activities occur. Some pathogens could produce hydrolytic enzymes able to degrade cellulose and pectin to counteract these non-specific pla t resistance mechanisms. in turn, plants developed the inhibitors of pathogen-related enzymes and this "arms race" is an important part of plant evolution and host-pathogen interaction mechanisms. Plants also can evaluate the cell wall state to compensate for imbalances and deficiencie . For instance, mutants with cellulose deficien y may have a higher lignific tion rate and a stronger stress response. The cell wall is also a source of signal molecules triggering the initiation of response mechanisms. in total, the plan cell wall is a complex dynamic structure able to prevent infection by most potential (non-specific) p thogens and switch on the mechanisms of plant immune response. The reconstruction of gene networks controlling the cell wall structural and functional organization during the growth, and under normal and stressful conditions is vitally important for understanding the basic molecular mechanisms of development and stress resistance. The mechanisms of specific and non-specific pl t resistance to various phytopathogens connected to the cell wall structure are reviewed. The roles of the cell wall constituents in pathogen detection and the induction of defense mechanism are discussed

Key words: innate immunity; cell wall; crops; fungal pathogen; leaf rust; non-host resistance.

HOW TO CITE THIS ARTICLE?
Smirnova o.G., Kochetov A.V. Plant cell wall and the mechanisms of resistance to pathogens. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2015;19(6):715-723. Doi 10.18699/VJ15.109

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?
Смирнова О.Г., Кочетов АВ. Клеточная стенка растений и механизмы устойчивости к патогенам. Вавиловский журнал генетики и селекции. 2015;19(6):715-723. Doi 10.18699/VJ15.109

Растения подвергаются воздействию огромного количества грибов, микробов, вирусов, некоторые из которых способны преодолевать защитные механизмы и вызывать заболевания. Устойчивость растений к патогенам основывается на сложной сети конститутивных и индуцированных защитных барьеров, в контроле которых задействовано большое число генов. На начальных этапах заражения грибом *Zymoseptoria tritici* и до появления первых признаков заболевания септориозом наблюдается изменение экспрессии более 3 000 генов пшеницы (*Triticum aestivum*) (Rudd et al., 2015). Эти изменения затрагивают синтез защитных белков, сигнальных молекул, гормонов, фруктана, лигнина и др.

Наряду с существованием различных специализированных механизмов защиты, у всех растений существует клеточная стенка (КС) – первое препятствие, которое должны преодолеть патогены, чтобы заселить ткани растения. Многочисленные изменения могут возникать в клеточных стенках в ответ на атаку микробов (Malinovsky et al., 2014). Успешная защита на уровне КС может остановить вторжение патогенов на ранней стадии, до формирования заболевания, и может исключить необходимость в более «дорогих» защитных механизмах, таких как гибель клетки при гиперчувствительном ответе. Следовательно, изучение механизмов устойчивости, связанных с КС, и понимание того, почему эти механизмы не срабатывают при встрече с некоторыми болезнетворными микроорганизмами и вирусами, имеет фундаментальное значение.

Способ взаимодействия патогена с КС зависит от жизненного цикла патогена. Некротрофы, которые убивают клетки и питаются мертвыми тканями, обычно размягчают ткани растения с помощью гидролитических ферментов, разрушающих полимеры КС. Биотрофы и гемибиотрофы, взаимодействующие с живыми клетками растений на протяжении всего жизненного цикла или его части, обычно применяют более тонкие стратегии для взаимодействия с КС. Образующие гаустории (органы питания гриба) патогены, такие как плесневые грибы и оомицеты, могут проникать через КС, создавая питающие структуры, тесно контактирующие с нижерасположенными клетками хозяина (Szabo, Bushnell, 2001).

Гетерогенность в строении КС у разных видов растений отражается в разнообразии стратегий, которые используют патогены для ее разрушения. В частности, для этой цели служит секреция патогенами различных гидролитических ферментов. Низкомолекулярные продукты разрушения КС (DAMPs, Damage-Associated Molecular Patterns), такие как олигосахара, являются сигнальными молекулами, запускающими защитные механизмы (Boller, Felix, 2009). Таким образом, КС – динамическая структура, которая регулирует конститутивные и индуцибельные механизмы защиты, являясь источником сигнальных молекул, запускающих разные варианты иммунного ответа (Miedes et al., 2014).

Врожденный иммунитет обеспечивает устойчивость растений к большинству патогенов, в том числе и за счет распознавания характерных для патогенов сигнальных молекул PAMPs (Pathogen-Associated Molecular Patterns), таких как бактериальный флагеллин, липосахариды, бета-глюкан, хитин и гидролитические ферменты (Zipfel et al., 2014).

Сигнальные молекулы DAMPs и PAMPs, имеющие, соответственно, растительное и микробное происхождение, запускают PTI иммунитет (Pattern-Triggered Immunity), который, как правило, предотвращает колонизацию микробов. Растения распознают сигнальные молекулы DAMPs и PAMPs при помощи расположенных на поверхности клеток рецепторов PRRs (Pattern-Recognition Receptors). PRRs являются трансмембранными белками с внеклеточными доменами (Trdá et al., 2015). Один из PRR рецепторов арабидопсиса, EFR, опознает цитоплазматический белок бактерий EF-Tu по минимальному пептидному эпитопу elf18 (Furukawa et al., 2013). Трансгенная экспрессия *AtEFR* в табаке, томатах и пшенице обеспечивает узнавание ими elf18, что сопровождается индукцией генов иммунного ответа, отложением каллозы, уменьшением повреждений, вызываемых патогеном, и свидетельствует о существовании у разных видов растений высоко консервативных механизмов защитного ответа, расположенных после идентификации PAMPs (Lacombe et al., 2010; Schoonbeek et al., 2015). В сельскохозяйственной практике предварительная обработка растений изолированными защитными элиситорами может способствовать повышению устойчивости растений (Wiesel et al., 2014).

Чтобы усложнить свою идентификацию, патогенные микроорганизмы секретируют эффекторные белки. Обнаружение растением эффекторных белков приводит к индукции ETI иммунитета (Effector-Triggered Immunity). ETI часто связан с локальной программируемой гибелью клеток (гиперчувствительным ответом), что ограничивает распространение микробной атаки (Jones, Dangl, 2006).

Устойчивость к определенным расам микробов, обусловленная R (Resistance)-генами, широко используется в селекционных программах зерновых культур (Dangl et al.,

2013). Однако с появлением новых рас патогенов большая часть имеющихся R-генов не может обеспечить иммунитет, поскольку при мутации патогены могут терять эффекторы, которые опознаются продуктами этих R-генов. Это является одной из причин постоянной «гонки вооружений» – поиска новых генов устойчивости в природных популяциях близких видов и их интродукции в сорта сельскохозяйственных растений, а также вынуждает исследователей использовать биотехнологические подходы для разработки более стабильных вариантов устойчивости (Smirnova et al., 2012; Филипенко и др., 2013; Смирнова, Кочетов, 2014; Smirnova, Kochetov, 2015).

распознавание хитина

Хитин, гомополимер ацетилглюкозамина, является основным структурным компонентом клеточной стенки грибов, а также входит в состав экзоскелета насекомых, панциря ракообразных, яиц и кишечного клапана нематод (Bueter et al., 2013). Хитин относится к PAMP и является хорошей мишенью для защитной реакции растений, поскольку полимеры глюкозамина в растениях отсутствуют. Поэтому не удивительно, что эволюционно консервативная стратегия растений против грибов и насекомых базируется на секреции хитиназ, гидролитических ферментов, которые расщепляют полимеры хитина (Hadwiger, 2013). Существуют примеры эволюционной коадаптации механизмов патогенеза грибов и защитных систем растений. Биотрофный патогенный гриб *Cladosporium fulvum* нивелирует действие хитиназ благодаря секреции апопластического эффектора Avr4 – хитин-связывающего белка, который защищает целостность клеточной стенки гриба от хитиназ (van den Burg et al., 2006). Гетерологичная экспрессия Avr4 в арабидопсисе или томатах снижает распознавание хитина и тем самым повышает вирулентность некоторых патогенных грибов (van Esse et al., 2007). В свою очередь растение-хозяин синтезирует внеклеточный, закрепленный на мембране богатый лейцином белок Cf-4, который обеспечивает узнавание Avr4 и активирует гиперчувствительный ответ (Takken et al., 1999). У риса и пшеницы хитин распознается как PAMP при помощи двойной системы, состоящей из белков CERK1 (Chitin Elicitor Receptor Kinase-1) и CEBiP (Chitin Oligosaccharide Elicitor-Binding Protein) (Shimizu et al., 2010; Lee et al., 2014). У арабидопсиса только один белок, CERK1, функционирует как рецептор для распознавания хитина (Shinya et al., 2012). Биологическая активность олигомеров хитина зависит от их размеров. Гептамеры и октамеры обладают наибольшей активностью как PAMP. Октамеры хитина могут связывать две или больше молекул AtCERK1, вызывая их димеризацию, что приводит к активации рецептора (Liu et al., 2012). Чтобы предотвратить опознавание хитина, *C. fulvum* во время инфицирования также секретирует эволюционно консервативный внеклеточный белок Ecp6. Он утилизирует фрагменты хитина, высвобождаемые хитиназами растений, ограничивает связывание хитина с PRR и препятствует идентификации гриба (Bolton et al., 2008; de Jonge et al., 2010; Sanchez-Vallet et al., 2013). Пока не известно, распознается ли Ecp6 растениями. Многоступенчатые механизмы взаимоотношений между грибом и растением во время инфицирования в большинстве случаев обеспечивают устойчивость растений к заражению.

структура клеточной стенки

Большинство КС в своей основе имеют обширную несущую сеть из микрофибрилл целлюлозы, пересекаемую молекулами гемицеллюлозы (Scheller, Ulvskov, 2010). В первичных стенках растущих частей растения эта сеть встроена в матрикс из пектиновых полисахаридов. Во вторичных клеточных стенках сформировавшихся тканей пектин представлен в меньшей степени, а стенки усилены лигнином (Endler, Persson, 2011). КС разных видов растений отличаются по тонкой структуре и трехмерной архитектуре. Кроме полисахаридов, растительная КС содержит значительное количество белков, выполняющих структурную (экстенсин, гликопротеины) и ферментативную функции (Михайлова, 2007). Создана база данных WallProtDB, которая содержит информацию о 2 170 белках и ESTs, экспериментально идентифицированных в 13 видах растений в результате протеомных исследований клеточных стенок (San Clemente, Jamet, 2015).

целлюлоза

Микрофибриллы целлюлозы синтезируются большими мультимерными комплексами, состоящими из субъединиц целлюлозо-синтазы (CESAs; Kumar, Turner, 2015). Мутанты с дефицитом целлюлозы обычно имеют повышенный уровень лигнификации и защитного ответа (Cano-Delgado et al., 2003; Hamann, 2012). Так, мутант CESA3 с изменением первичной КС более устойчив к мучнистой росе (Ellis, Turner, 2001). Дефекты во вторичной КС, вызванные повреждением субъединиц CESA4, CESA7 и CESA8, также приводят к повышенной устойчивости к грибу *Plectosphaerella cucumerina* и почвенной бактерии *Ralstonia solanacearum* (Hernandez-Blanco et al., 2007). У арабидопсиса нарушение КС, вызванное ингибитором синтеза целлюлозы изоксабеном, приводит к индукции синтеза лигнина через RbohD (Respiratory Burst Oxidase Homolog D)-зависимый механизм, а тонкая настройка протекает путем негативной обратной регуляции при участии жасмоновой кислоты (Denness et al., 2011). Нарушение КС, связанное с потерей целлюлозы, включает защитные ответы и предполагает присутствие системы мониторинга целостности КС.

Гемицеллюлоза

Гемицеллюлозы – это разнообразная группа полисахаридов, состоящих из остатков пентоз и гексоз. Гемицеллюлозы укрепляют КС, взаимодействуя с целлюлозой и иногда с лигнином (Endler, Persson, 2011; Pauly et al., 2013). Ксилан является преобладающей гемицеллюлозой во вторичной КС. Некоторые фитопатогены секретируют ксиланазы, которые расщепляют содержащийся в КС ксилан до ксилоз, что нарушает и ослабляет ее (Belien et al., 2006). Гриб *Trichoderma* spp. продуцирует ксиланазу EIX (Ethylene-Inducing Xylanase), которая узнается растением как PAMP (Furman-Matarasso et al., 1999). У томатов (*Lycopersicum esculentum*) идентификация EIX осуществляется расположенными на клеточной поверхности рецепторподобными белками LeEix1 и LeEix2 (Ron,

Avni, 2004). Чтобы противостоять деградации ксилана микробными эндоксиланазами, травянистые однодольные растения продуцируют ингибиторы ксиланаз, такие как TAXI (*Triticum aestivum* Xylanase Ihibitor), XIP (Xylanase Inhibitor Protein) и TL-XI (Thaumatin-Like Xylanase Inhibitor) (Bellincampi et al., 2004; Juge, 2006). Конститутивная экспрессия TAXI-III в пшенице понижает чувствительность к *Fusarium graminearum* (Moscetti et al., 2013). Выступая в роли PAMPs, ксиланазы грибов усиливают защитный ответ (Noda et al., 2010; Sella et al., 2013).

Пектин

Пектины являются главными компонентами матрикса КС. Они представляют собой полисахариды, образованные, главным образом, остатками галактуроновой кислоты. Одними из первых ферментов, которые патогенные грибы секретируют во время инфекции, являются эндо-полигалактуроназы, которые разрушают пектин, нарушают целостность КС и обеспечивают доступ патогенов. При деградации пектина образуются фрагменты олигогалактурониды, которые в норме не присутствуют в КС и поэтому выступают в роли DAMP (Galletti et al., 2009). Сенсорами целостности пектина являются связанные с КС киназы, которые определяют присутствие олигогалактуронидов с уровнем полимеризации между 10 и 15 (Ferrari et al., 2013).

Показано, что белок RWA2 (Reduced Wall Acetylation 2) отвечает за ацетилирование пектиновых и непектиновых полимеров у арабидопсиса и нокаутные мутанты rwa2 имеют повышенную устойчивость к *Botrytis cinerea* (Manabe et al., 2011). Деацетилирование пектина и ксилоглюкана в трансгенных растениях может быть частью защитной стратегии, поскольку увеличивает доступность для ферментов деградации, продуцирующих олигосахариды, которые выступают в качестве элиситоров защитной реакции (Pogorelko et al., 2013).

Ингибиторы полигалактуроназ играют важную роль в защитном ответе, являясь модуляторами активности этих ферментов. Накопление ингибитора PGIP1 при несовместимом взаимодействии проса с *Sclerospora graminicola* может быть использовано для создания устойчивых форм проса (Prabhu et al., 2015).

Разработан метод мониторинга уровня инфекции по уровню гидролизованного пектина путем фенотипирования пектин-метилтрансфераз, пектиназ и олигогалактуронидов (Lionetti, 2015).

лигнин и фенольные компоненты

Лигнин – это ароматический полимер, который влияет на прочность и непроницаемость, располагаясь преимущественно во вторично утолщенных КС. У растений лигнин состоит преимущественно из монолигнолов: конифенилового и синапинового спиртов, дающих начало G и S единицам полимера лигнина соответственно. Реже представлен кумариловый спирт, формирующий H единицу лигнина. H единица чаще встречается у однодольных, чем у двудольных растений. У некоторых видов растений мономеры лигнина представлены в ацетилированной форме. В качестве мономеров лигнина растения также используют ряд других фенолов. Например, лигнин в соломе пшеницы имеет довольно высокий уровень флавоноида трицина (Del Río et al., 2012).

Лигнин и лигниноподобные фенольные полимеры быстро накапливаются в КС в ответ на биотические и абиотические стрессы и на нарушения ее структуры (Sattler, Funnell-Harris, 2013). Стрессы вызывают индукцию экспрессии генов фенилпропаноидного пути у различных видов растений, что приводит к лигнификации КС (Bhuiyan et al., 2007; Zhao et al., 2009). Запасание лигнина в инфицированных клетках может предотвращать распространение токсинов и ферментов патогенов в организме-хозяине и перенос воды и питательных веществ от клеток хозяина к патогену (Smith et al., 2007).

У пшениц лигнификация действует как защитный ответ при инфекции. Например, S-обогащенный лигнин накапливается во время гиперчувствительной реакции пшеницы после инфекции *Puccinia graminis* (Menden et al., 2007) и синтезируется в оболочках клеток эпидермиса пшеницы, инфицированных *Fusarium proliferatum* (Bishop et al., 2002). Напротив, не наблюдалось изменений в содержании лигнина в листьях пшеницы, инфицированных вирусом полосатой мозаики пшеницы (Kofalvi, Nassuth, 1995). У пшеницы сорта Toropi устойчивость к листовой ржавчине, вызываемой *Puccinia triticina*, формируется до образования гаусторий за счет индукции генов устойчивости, в том числе напрямую или опосредованно участвующих в лигнификации (Casassola et al., 2015). У линий тыквы (*Cucumis melo*), устойчивых к мучнистой росе, вызываемой грибом *Podosphaera fusca*, во время инфекции происходит более быстрое накопление лигнина по сравнению с чувствительными линиями. Это коррелирует с повышением уровня фермента фенилпропаноидного пути PAL (phenylalanine ammonia-lyase) (Romero et al., 2008).

Фенилпропаноидный путь, задействованный в синтезе лигнина, также участвует в синтезе многочисленных фенольных компонентов, таких как стильбены, кумарины, неолигнаны, конъюгаты фенилпропаноидов и флавоноиды. Многие из этих компонентов являются фитоалексинами – антимикробными компонентами, участвующими в защите растений (König et al., 2014).

Доказательства роли лигнина и растворимых фенолов в защите растений были получены после анализа устойчивости трансгенных растений и мутантов с измененным составом или уровнем лигнина. У хлопчатника (*Gossypium hirsutum*) обнаружена количественная связь между повышением уровня лигнина в стеблях во время инфекции грибом *Verticillium dahliae* и устойчивостью (Xu et al., 2011). Сверхэкспрессия гена хлопчатника *DIRIGENT1*, усиливающая лигнификацию, блокировала распространение *V. dahliae* (Shi et al., 2012). Трансгенные растения табака, конститутивно сверхэкспрессирующие ген *PAL*, показывали большую устойчивость к *Cercospora nicotianae* и *Phytophthora parasitica* pv. *nicotianae* (Way et al., 2002; Shadle et al., 2003). Растения табака с супрессией *PAL* имели пониженный уровень хлорогеновой кислоты и более быстрое возникновение повреждений после инфекции патогенным грибом *Cercospora nicotianae* по сравнению с растениями дикого типа. Уровень лигнина у линий с суперессией *PAL* не изучался, но можно предпо-

ложить, что повышенная чувствительность этих растений могла быть вызвана снижением количества лигнина или более тонкой КС (Maher et al., 1994).

Изучено влияние модификаций в биосинтезе лигнина на чувствительность растений к патогенам. Например, у пшениц (*Triticum monococcum*) выключение генов синтеза монолигнола (*TmPAL*, *TmCOMT*, *TmCCoAOMT* и *TmCAD*), основной структурной единицы лигнина, приводило к сверхчувствительности к грибу *Blumeria graminis* f. sp. *tritici*, вызывающему заболевание мучнистой росой (Bhuiyan et al., 2009). Повышенное накопление моно- и диферулатов в КС пшеницы и овса во время инфекции, соответственно, *Puccinia coronate* sp. *avenae* и *Agrobacterium* sp. было связано с устойчивостью к этим патогенам (Ikegawa et al., 1996; Parrott et al., 2002). У льна супрессия синтеза фермента CAD (Cinnamyl Alcohol Dehydrogenase) вызывала повышение чувствительности к сосудистому грибу *Fusarium oxysporum* (Wróbel-Kwiatkowska et al., 2007).

Однако не всегда нарушение биосинтеза лигнина приводит к снижению сопротивляемости к некоторым патогенам. Линии табака с супрессией синтеза ферментов COMT (caffeic acid O-methyltransferase) и CCoAOMT (caffeoyl-CoA O-methyltransferase) были более устойчивы к инфицированию *Agrobacterium tumefaciens* и имели меньшие размер и массу опухоли по сравнению с растениями дикого типа (Maury et al., 2010). Фенольные соединения, секретируемые такими растениями после нанесения им повреждений, не вызывали экспрессию бактериальных *Vir* генов на столь же высоком уровне, как фенольные соединения, секретируемые нормальными растениями. Другими словами, *Agrobacterium* не узнавала своего хозяина из-за различий в растворимых фенолах.

Мутанты сорго (*Sorghum bicolor* L.) bmr6 и bmr12, имеющие пониженное содержание лигнина, характеризуются нарушением первичной последовательности генов *CAD* и *COMT* и синтезом нефункциональных ферментов (Bout, Vermerris, 2003; Sattler et al., 2009). Несмотря на пониженный уровень лигнина, в зерновках этих мутантов наблюдался пониженный уровень колонизации *Fusarium* ssp. и *Alternaria alternata* (Funnell-Harris et al., 2010). Неизвестно, изменение в составе лигнина или накопление фенольных соединений вызывало повышение устойчивости у этих линий сорго.

В большинстве исследований не изучалась роль лигнина в регуляции других защитных ответов, и пока не ясно, является ли роль лигнина в регуляции специфического ответа активной или относительно пассивной.

Папиллы клеточной стенки

КС активно перестраивается и укрепляется в местах контакта с потенциальными патогенами. Активное локальное укрепление КС через формирование папилл является ранним индуцированным ответом на большое число патогенных грибов и бактерий. Папилла – это сложная структура, которая формируется между плазматической мембраной и внутренней стороной КС в месте проникновения патогена и служит физическим барьером для ограничения проникновения патогенов в протопласт. Дополнительно папиллы являются центрами накопления антимикробных вторичных метаболитов (Bednarek et al., 2009; Clay et al., 2009). Мало известно о молекулярных механизмах и клеточных процессах, участвующих в определении местоположения и сборке папилл (Underwood, 2012). Несмотря на то что у разных видов растений биохимический состав папилл может различаться, некоторые классы компонентов, такие как полимеры и белки КС, фенольные производные, активные формы кислорода и каллоза, встречаются повсеместно. Согласованность различных транспортных процессов при формировании папилл является ключевым фактором успешной защиты растений (Voigt, 2014). Быстрое формирование папилл коррелирует с повышенной устойчивостью к проникновению грибов, в то время как задержка в их формировании коррелирует с успешным проникновением грибов (Bayles et al., 1990; Collins et al., 2003).

Укрепление отдельных участков КС посредством папилл является частью иммунного ответа и, по-видимому, общим механизмом для разных видов растений (Nicaise et al., 2009).

неспецифическая устойчивость

В растениях развились сложные механизмы для защиты от неадаптированных патогенов. Неспецифическая устойчивость стабильно защищает различные виды растений от поражения подавляющим числом патогенов. Этот вид устойчивости постоянно привлекает внимание исследователей, так как обеспечивается врожденным иммунитетом растений и представляет собой наиболее надежную и долговременную форму.

Механизмы, лежащие в основе неспецифической устойчивости, остаются относительно малоисследованными по сравнению с механизмами специфической устойчивости. Процессы, участвующие в формировании неспецифической устойчивости при бактериальной инфекции, затрагивают укрепление клеточной стенки, синтез воскового налета, закрывание устьиц, синтез стерола, защитных молекул (Senthil-Kumar, Mysore, 2013). Индуцированная неспецифическая устойчивость против бактерий, грибов и оомицетов может быть разделена на два типа. При I типе не наблюдается видимых симптомов, в то время как при II типе происходит быстрый гиперчувствительный ответ с гибелью клеток. I тип более распространен, чем II тип (Mysore, Ryu, 2004; Nurnberger, Lipka, 2005).

Используя третью транспортную систему, патогенные бактерии секретируют эффекторные молекулы, под воздействием которых растительная клетка начинает производить необходимые для бактерии питательные вещества (Cunnac et al., 2009). Неспецифическая устойчивость растения может быть связана с его неспособностью изменять свой клеточный метаболизм под воздействием бактериальных эффекторов и со снижением проницаемости клеточных мембран. Нарушение синтеза стерола у растений табака и арабидопсиса приводит к повышению проницаемости мембран и выходу питательных веществ в апопласт. Повышенный уровень питательных веществ в апопласте приводил к повышенной чувствительности этих растений не только к специфическим, но и неспецифическим патогенным бактериям (Wang et al., 2012). Отсутствие необходимых для патогена питательных

веществ или доступа к питательным веществам является важной причиной неспецифической устойчивости растений (Fatima et al., 2015).

неспецифическая устойчивость к ржавчине

Устойчивость к ржавчине задействует индукцию разнообразных защитных механизмов. Хотя большинство зерновых чувствительны, по крайней мере, к одному из видов грибов, вызывающих ржавчину, рис (*Oryza sativa*) является исключением и не поражается известными видами ржавчины. После инокуляции листьев риса сорта Nipponbare грибом *P. triticina* f. sp. *tritici* (*Ptt*) только 10 % проросших спор формировали аппрессории через устьица. Через три дня вокруг аппрессорий накапливалась перекись водорода. Только 3 % аппрессорий формировали короткие гифы внутри листа, из которых только 0,2 % через 21 день после инокуляции формировали разветвленные гифы в клетках мезофилла. При этом не наблюдалось образование субстромальных везикул, материнских клеток гаусторий или гаусторий. Устойчивость риса к *Ptt* связана с изменением белкового и энергетического метаболизма, накоплением фитоалексинов, укреплением КС, ускорением репарации клетки, повышенным уровнем антиокисления и детоксификации. Более половины белков с повышенным уровнем экспрессии были связаны с работой хлоропластов и митохондрий, что предполагает важную роль этих органелл в устойчивости (Li et al., 2012).

Грибы, вызывающие ржавчину у злаков, не способны вызвать заболевание у бобовых. После инокуляции листьев бобов (*Vicia faba* L.) патогеном *Puccinia striiformis* f. sp. *tritici* (*Pst*), вызывающим желтую ржавчину у пшениц, видимых симптомов заболевания не наблюдается. Попытки инфицирования вызывали образование папилл, утолщение КС, образование активных форм кислорода, запасание каллозы и накопление фенольных соединений в КС бобов. Немногочисленные гаустории *Pst*, которые формировались в клетках бобов, были окружены активным кислородом и каллозным материалом, и такие клетки были подвергнуты гиперчувствительному ответу (Cheng et al., 2012).

Пшеница поражается несколькими видами *Puccinia*, но устойчива ко всем видам *Uromyces*. Изучена основа устойчивости пшеницы к *Uromyces fabae*, вызывающему ржавчину у бобов. Уредоспоры *U. fabae* эффективно прорастали на листьях пшеницы, но только 2 % из них формировали аппрессории через устьица. В то же время большая часть из них не могла проникнуть через клетки мезофилла пшеницы. Через четыре дня только 4 % достигших мезофилла инфекционных единиц *U. fabae* формировали гаустории. Попытки проникновения материнских клеток гаусторий вызывали утолщение КС и формирование папилл в растительных клетках, что ограничивало развитие и рост гриба. Проникшие в клетки гаустории были заключены в каллозоподобный материал и не вызывали реакции гиперчувствительности. У пшеницы наблюдалась активация нескольких генов базовой устойчивости и окислительного стресса (Zhang et al., 2011).

Данные результаты показывают многоуровневый способ защиты при неспецифической устойчивости, включая структурное и химическое укрепление КС, гиперчувствительный ответ и индукцию ряда генов. Причем, если при взаимодействии бобов и *Pst* наблюдается гиперчувствительный ответ, то при взаимодействии пшеницы и *U. fabae* гаустории были инкапсулированы и гибель клеток не наблюдалась.

Для большинства вызывающих ржавчину патогенов процесс инфицирования задерживается сразу после образования первичной материнской клетки гаусторий в тканях невосприимчивых растений (Niks, 1983; Hoogkamp et al., 1998). Исследования взаимодействий невосприимчивых растений и ржавчинных грибов, таких как арабидопсис и *Uromyces vignae*, *Puccinia triticina*, *Hemileia vastatrix* (Mellersh, Heath, 2003; Shafiei et al., 2007; Azinheira et al., 2010); пшеница и *P. hordei*, *U. fabae* (Prats et al., 2007; Zhang et al., 2011); ячмень и *P. triticina*, *P. hordei-murini*, *P. hordei-secalini*, *P. persistens* (Jafary et al., 2008); рис и *P. graminis*, *P. triticina*, *P. striiformis*, *P. hordei* и *Melampsora lini* (Ayliffe et al., 2011a, b), показали, что устойчивость к грибной ржавчине наследуется филогенетически и является активным ответом, в котором задействованы сигналы салициловой кислоты.

У мутантов арабидопсиса sid2 и NahG с пониженным уровнем салициловой кислоты наблюдалось ускоренное развитие гриба *U. vignae*, вызывающего ржавчину у вигны (Mellersh, Heath, 2003). В формировании устойчивости арабидопсиса к листовой ржавчине пшеницы, вызываемой *Ptt*, задействованы активные формы кислорода, оксид азота, салициловая кислота и фитоалексин камалексин (Shafiei et al., 2007). Для устойчивости риса, взаимодействующего с грибом стеблевой ржавчины пшеницы *P. graminis* f.sp *tritici*, характерны индукция образования перекиси водорода и отложение каллозы (Ayliffe et al., 2011b). Устойчивость, связанная с гиперчувствительным ответом в запирающих клетках устьиц арабидопсиса после проникновения через них апрессорий гриба *H. vastatrix*, вызывающего ржавчину у кофе, сопровождается накоплением фенолов, отложением каллозы и экспрессией генов устойчивости, таких как *PR1*, *PR5*, *POX* и *WRKY* (Azinheira et al., 2010).

Изучение неспецифической устойчивости способствует лучшему пониманию механизмов специфической устойчивости в связи с наличием ассоциаций между неспецифической устойчивостью растений к неадаптированным и базовой устойчивостью к адаптированным патогенам (Cheng et al., 2012).

Acknowledgments

The work was supported by the Russian Ministry of Science and Education, contract No.14.604.21.0107 of August 7, 2014, identification number RFMEFI60414X0107.

Conflict of interest

The authors declare no conflict of interest.

References

Ayliffe M., Devilla R., Mago R., White R., Talbot M., Pryor A., Leung H. Nonhost resistance of rice to rust pathogens. Mol. Plant-Microbe Interact. 2011a;24:1143-1155.

Ayliffe M., Jin Y., Kang Z.S., Persson M., Steffenson B., Wang S.P., Leung H. Determining the basis of nonhost resistance in rice to ce-

real rusts. Euphytica. 2011b;179:33-40. DOI 10.1007/s10681-010-0280-2

Azinheira H.G., Silva M.D., Talhinhas P., Medeira C., Maia I., Petitot A.S., Fernandez D. Non-host resistance responses of *Arabidopsis thaliana* to the coffee leaf rust fungus (*Hemileia vastatrix*). Botany. 2010;88:621-629.

Bayles C.J., Ghemawat M.S., Aist J.R. Inhibition by 2-deoxy-D-glucose of callose formation, papilla deposition, and resistance to powdery mildew in an *mlo* barley mutant. Physiol. Mol. Plant Pathol. 1990;36:63-72. DOI 10.1016/0885-5765(90)90092-C

Bednarek P., Piślewska-Bednarek M., Svatoš A., Schneider B., Doubský J., Mansurova M., Humphry M., Consonni C., Panstruga R., Sanchez-Vallet A., Molina A., Schulze-Lefert P. A glucosinolate metabolism pathway in living plant cells mediates broad-spectrum antifungal defense. Science. 2009;232:101-106. DOI 10.1126/science.1163732

Belien T., Van Campenhout S., Robben J., Volckaert G. Microbial endoxylanases: effective weapons to breach the plant cell-wall barrier or, rather, triggers of plant defense systems? Mol. Plant-Microbe Interact. 2006;19:1072-1081.

Bellincampi D., Camardella L., Delcour J.A., Desseaux V., D'Ovidio R., Durand A., Elliot G., Gebruers K., Giovane A., Juge N., Sørensen J.F., Svensson B., Vairo D. Potential physiological role of plant glycosidase inhibitors. Biochim. Biophys. Acta. 2004;1696(2):265-274. DOI 10.1016/j.bbapap.2003.10.011

Bhuiyan N., Liu W., Liu G., Selvaraj G., Wei Y., King J. Transcriptional regulation of genes involved in the pathways of biosynthesis and supply of methyl units in response to powdery mildew attack and abiotic stresses in wheat. Plant Mol. Biol. 2007;64:305-318.

Bhuiyan N.H., Selvaraj G., Wei Y., King J. Gene expression profiling and silencing reveal that monolignol biosynthesis plays a critical role in penetration defence in wheat against powdery mildew invasion. J. Exp. Bot. 2009;60:509-521. DOI 10.1093/jxb/ern290

Bishop D.L., Chyatterton N.J., Harrison P.A., Hatfield R.D. Changes in carbohydrate partitioning and cell wall remodeling with stress-induced pathogenesis in wheat sheaths. Physiol. Mol. Plant Pathol. 2002;61:53-63. DOI 10.1006/pmpp.2002.0416

Boller T., Felix G. A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. Annu. Rev. Plant Biol. 2009;60:379-406. DOI 10.1146/annurev.arplant.57.032905.105346

Bolton M.D., Van Esse H.P., Vossen J.H., De Jonge R., Stergiopoulos I., Stulemeijer I.J., van den Berg G.C., Borrás-Hidalgo O., Dekker H.L., de Koster C.G., de Wit P.J., Joosten M.H., Thomma B.P. The novel *Cladosporium fulvum* lysin motif effector Ecp6 is avirulence factor with orthologues in other fungal species. Mol. Microbiol. 2008; 69(1):119-136. DOI 10.1111/j.1365-2958.2008.06270.x

Bout S., Vermerris W. A candidate-gene approach to clone the sorghum *Brown midrib* gene encoding caffeic acid *O*-methyltransferase. Mol. Genet.Genomics.2003;269:205-214.DOI10.1007/s00438-003-0824-4

Bueter C.L., Specht C.A., Levitz S.M. Innate sensing of chitin and chitosan. PLoS Pathog. 2013;9:e1003080. DOI 10.1371/journal.ppat.1003080

Cano-Delgado A., Penfield S., Smith C., Catley M., Bevan M. Reduced cellulose synthesis invokes lignification and defense responses in *Arabidopsis thaliana*. Plant J. 2003;34:351-362. DOI 10.1046/j.1365-313X.2003.01729.x

Casassola A., Brammer S.P., Chaves M.S., Martinelli J.A., Stefanato F., Boyd L.A. Changes in gene expression profiles as they relate to the adult plant leaf rust resistance in the wheat cv. Toropi. Physiol Mol. Plant Pathol. 2015;89:49-54. DOI 10.1016/j.pmpp.2014.12.004

Cheng Y., Zhang H., Yao J., Wang X., Xu J., Han Q., Wei G., Huang L., Kang Z. Characterization of non-host resistance in broad bean to the wheat stripe rust pathogen. BMC Plant Biol. 2012;12:96. DOI 10.1186/1471-2229-12-96

Clay N.K., Adio A.M., Denoux C., Jander G., Ausubel F.M. Glucosinolate metabolites required for an *Arabidopsis* innate immune response. Science. 2009;323:95-100. DOI 10.1126/science.1164627

Collins N.C., Thordal-Christensen H., Lipka V., Bau S., Kombrink E., Qiu J.L., Huckelhoven R., Stein M., Freialdenhoven A., Somerville S.C., Schulze-Lefert P. SNARE-protein-mediated disease resistance at the plant cell wall. Nature. 2003;425:973-977. DOI 10.1038/nature02076

Cunnac S., Lindeberg M., Collmer A. *Pseudomonas syringae* type III secretion system effectors: repertoires in search of functions. Curr. Opin. Microbiol. 2009;12(1):53-60. DOI 10.1016/j.mib.2008.12.003

Dangl J.L., Horvath D.M., Staskawicz B.J. Pivoting the plant immune system from dissection to deployment. Science. 2013;341(6147):746-751. DOI 10.1126/science.1236011

de Jonge R., Van Esse H.P., Kombrink A., Shinya T., Desaki Y., Bours R., van der Krol S., Shibuya N., Joosten M.H., Thomma B.P. Conserved fungal LysM effector Ecp6 prevents chitin-triggered immunity in plants. Science. 2010;329:953-955. DOI 10.1126/science.1190859

Del Río J.C., Rencoret J., Prinsen P., Martínez Á.T., Ralph J., Gutiérrez A. Structural characterization of wheat straw lignin as revealed by analytical pyrolysis, 2D-NMR, and reductive cleavage methods. J. Agric. Food Chem. 2012;60(23):5922-5935. DOI 10.1021/jf301002n

Denness L., Mckenna J.F., Segonzac C., Wormit A., Madhou P., Bennett M., Mansfield J., Zipfel C., Hamann T. Cell wall damage-induced lignin biosynthesis is regulated by a reactive oxygen species- and jasmonic acid-dependent process in *Arabidopsis*. Plant Physiol. 2011;156(3):1364-1374. DOI 10.1104/pp.111.175737.

Ellis C., Turner J.G. The *Arabidopsis* mutant *cev1* has constitutively active jasmonate and ethylene signal pathways and enhanced resistance to pathogens. Plant Cell. 2001;13(5):1025-1033.

Endler A., Persson S. Cellulose synthases and synthesis in *Arabidopsis*. Mol. Plant. 2011;4(2):199-211. DOI 10.1093/mp/ssq079

Fatima U., Senthil-Kumar M. Plant and pathogen nutrient acquisition strategies. Front Plant Sci. 2015;17;6:750. DOI 10.3389/fpls.2015.00750

Ferrari S., Savatin D.V., Sicilia F., Gramegna G., Cervone F., Lorenzo G.D. Oligogalacturonides: plant damage-associated molecular patterns and regulators of growth and development. Front. Plant Sci. 2013;4:49. DOI 10.3389/fpls.2013.00049

Filipenko E.A., Kochetov A.V., Kanayama Y., Malinovsky V.I., Shumny V.K. Association between PR proteins with ribonuclease activity and plant resistance against pathogenic fungi. Vavilovskii Zhurnal Genetiki i Selektsii=Vavilov Journal of Genetics and Breeding. 2013;17(2):326-334.

Funnell-Harris D.L., Pedersen J.F., Sattler S.E. Alteration in lignin biosynthesis restricts growth of *Fusarium* spp. in brown midrib sorghum. Phytopathology. 2010;100(7):671-681. DOI 10.1094/PHYTO-100-7-0671

Furman-Matarasso N., Cohen E., Du Q., Chejanovsky N., Hanania U., Avni A. A point mutation in the ethylene-inducing xylanase elicitor inhibits the beta-1-4-endoxylanase activity but not the elicitation activity. Plant Physiol. 1999;121(2):345-351.

Furukawa T., Inagaki H., Takai R., Hirai H., Che F.S. Two distinct EF-Tu epitopes induce immune responses in rice and *Arabidopsis*. Mol. Plant Microbe Interact. 2013;27(2):113-124. DOI 10.1094/MPMI-10-13-0304-R

Galletti R., De Lorenzo G., Ferrari S. Host-derived signals activate plant innate immunity. Plant Signal. Behav. 2009;4:33-34.

Hadwiger L.A. Multiple effects of chitosan on plant systems: solid science or hype. Plant Sci. 2013;208:42-49. DOI 10.1016/j.plantsci.2013.03.007

Hamann T. Plant cell wall integrity maintenance as an essential component of biotic stress response mechanisms. Front. Plant Sci. 2012;3:77. DOI 10.3389/fpls.2012.00077

Hernandez-Blanco C., Feng D.X., Hu J., Sanchez-Vallet A., Deslandes L., Llorente F., Berrocal-Lobo M., Keller H., Barlet X., Sánchez-Rodríguez C., Anderson L.K., Somerville S., Marco Y., Molina A. Impairment of cellulose synthases required for *Arabidopsis* secondary cell wall formation enhances disease resistance. Plant Cell. 2007;19(3):890-903. DOI 10.1105/tpc.106.048058

Hoogkamp T., Chen W.Q., Niks R. Specificity of prehaustorial resistance to *Puccinia hordei* and to two inappropriate rust fungi in barley. Phytopathology. 1998;88(8):856-861. DOI 10.1094/PHYTO.1998.88.8.856

Ikegawa T., Mayama S., Nakayashiki H., Kato H. Accumulation of diferulic acid during the hypersensitive response of oat leaves to *Puccinia coronate* f. sp. *avenae* and its role in the resistance of oat tissues to cell wall degrading enzymes. Physiol. Mol. Plant Pathol. 1996;48(4):245-256. DOI 10.1006/pmpp.1996.0021

Jafary H., Albertazzi G., Marcel T.C., Niks R.E. High diversity of genes for nonhost resistance of barley to heterologous rust fungi. Genetics. 2008;178(4):2327-2339. DOI 10.1534/genetics.107.077552

Jones J.D.G., Dangl J.L. The plant immune system. Nature. 2006; 444(7117):323-329. DOI 10.1038/nature05286

Juge N. Plant protein inhibitors of cell wall degrading enzymes. Trends Plant Sci. 2006;11(7):359-367. DOI 10.1016/j.tplants.2006.05.006

Kofalvi S.A., Nassuth A. Influence of wheat streak mosaic virus infection on phenylpropanoid metabolism and the accumulation of phenolics and lignin in wheat. Physiol. Mol. Plant Pathol. 1995;47(6):365-377. DOI 10.1006/pmpp.1995.1065

König S., Feussner K., Kaever A., Landesfeind M., Thurow C., Karlovsky P., Gatz C., Polle A., Feussner I. Soluble phenylpropanoids are involved in the defense response of *Arabidopsis* against *Verticillium longisporum*. New Phytol. 2014;202(3):823-837. DOI 10.1111/nph.12709

Kumar M., Turner S. Plant cellulose synthesis: CESA proteins crossing kingdoms. Phytochemistry. 2015;112:91-99. DOI 10.1016/j.phytochem.2014.07.009

Lacombe S., Rougon-Cardoso A., Sherwood E., Peeters N., Dahlbeck D., Van Esse H.P., Smoker M., Rallapalli G., Thomma B.P., Staskawicz B., Jones J.D., Zipfel C. Interfamily transfer of a plant pattern-recognition receptor confers broad-spectrum bacterial resistance. Nat. Biotechnol. 2010;28(4):365-369. DOI 10.1038/nbt.1613

Lee W.S., Rudd J.J., Hammond-Kosack K.E., Kanyuka K. Mycosphaerella graminicola LysM effector-mediated stealth pathogenesis subverts recognition through both CERK1 and CEBiP homologues in wheat. Mol. Plant Microbe Interact. 2014;27(3):236-243. DOI 10.1094/MPMI-07-13-0201-R

Li H., Goodwin P.H., Han Q., Huang L., Kang Z. Microscopy and proteomic analysis of the non-host resistance of *Oryza sativa* to the wheat leaf rust fungus, *Puccinia triticina* f. sp. *tritici*. Plant Cell Rep. 2012;31(4):637-650. DOI 10.1007/s00299-011-1181-0

Lionetti V. PECTOPLATE: the simultaneous phenotyping of pectin methylesterases, pectinases, and oligogalacturonides in plants during biotic stresses. Front Plant Sci. 2015;6:331. DOI 10.3389/fpls.2015.00331

Liu T., Liu Z., Song C., Hu Y., Han Z., She J., Fan F., Wang J., Jin C., Chang J., Zhou J.M., Chai J. Chitin-induced dimerization activates a plant immune receptor. Science. 2012;336(6085):1160-1164. DOI 10.1126/science

Maher E.A., Bate N.J., Ni W., Elkind Y., Dixon R.A., Lamb C.J. Increased disease susceptibility of transgenic tobacco plants with suppressed levels of preformed phenylpropanoid products. Proc. Natl Acad. Sci. USA. 1994;91(16):7802-7806.

Malinovsky F.G., Fangel J.U., Willats W.G. The role of the cell wall in plant immunity. Front Plant Sci. 2014;5:178. DOI 10.3389/fpls.2014.00178

Manabe Y., Nafisi M., Verhertbruggen Y., Orfila C., Gille S., Rautengarten C., Cherk C., Marcus S.E., Somerville S., Pauly M., Knox J.P., Sakuragi Y., Scheller H.V. Loss-of-function mutation of reduced wall acetylation 2 in *Arabidopsis* leads to reduced cell wall acetylation and increased resistance to *Botrytis cinerea*. Plant Physiol. 2011;155(3):1068-1078. DOI 10.1104/pp.110.168989

Maury S., Delaunay A., Mesnard F., Cronier D., Chabbert B., Geoffroy P., Legrand M. *O*-methyltransferase(s)-suppressed plants produce lower amounts of phenolic vir inducers and are less susceptible to *Agrobacterium tumefaciens* infection. Planta. 2010;232(4):975-986. DOI 10.1007/s00425-010-1230-x

Mellersh D.G., Heath M.C. An investigation into the involvement of defense signaling pathways in components of the nonhost resistance of Arabidopsis thaliana to rust fungi also reveals a model system for studying rust fungal compatibility. Mol. Plant Microbe Interact. 2003;16(5):398-404.

Menden B., Kohlhoff M., Moerschbacher B.M. Wheat cells accumulate a syringyl-rich lignin during the hypersensitive resistance response. Phytochemistry. 2007;68(4):513-520. DOI 10.1016/j.phytochem.2006.11.011

Miedes E., Vanholme R Boerjan W Molina A. The role of the secondary cell wall in plant resistance to pathogens. Front Plant Sci. 2014;5:358. DOI 10.3389/fpls.2014.00358

Mikhaylova R.V. Matseriruyushchie fermenty mitselialnykh gribov v biotekhnologii [Macerating enzymes of mycelial fungi in biotechnology]. Minsk, Belorusskaya nauka, 2007.

Moscetti I., Tundo S., Janni M., Sella L., Gazzetti K., Tauzin A., Giardina T., Masci S., Favaron F., D'Ovidio R. Constitutive expression of the xylanase inhibitor TAXI-III delays fusarium head blight symptoms in durum wheat transgenic plants. Mol. Plant Microbe Interact. 2013;26(12):1464-1472. DOI 10.1094/MPMI-04-13-0121-R

Mysore K.S., Ryu C.M. Nonhost resistance: how much do we know? Trends Plant Sci. 2004;9(2):97-104. DOI 10.1016/j.tplants.2003.12.005

Nicaise V., Roux M., Zipfel C. Recent advances in PAMP-triggered immunity against bacteria: pattern recognition receptors watch over and raise the alarm. Plant Physiol. 2009;150(4):1638-1647. DOI 10.1104/pp.109.139709

Niks R. Comparative histology of partial resistance and the nonhost reaction to leaf rust pathogens in barley and wheat seedlings. Phytopathology. 1983;73:60-64.

Noda J., Brito N., González C. The *Botrytis cinerea* xylanase Xyn11A contributes to virulence with its necrotizing activity, not with its catalytic activity. BMC Plant Biol. 2010;10:38. DOI 10.1186/1471-2229-10-38

Nurnberger T., Lipka V. Non-host resistance in plants: new insights into an old phenomenon. Mol. Plant. Pathol. 2005;6(3):335-345. DOI 10.1111/j.1364-3703.2005.00279.x

Parrott D.L., Anderson A.J., Carman J.G. *Agrobacterium* induces plant cell death in wheat (*Triticum aestivum* L.). Physiol. Mol. Plant Pathol. 2002;60(2):59-69. DOI 10.1006/pmpp.2002.0378

Pauly M., Gille S., Liu L.F., Mansoori N., De Souza A., Schultink A., Xiong G. Hemicellulose biosynthesis. Planta. 2013;238(4):627-642. DOI 10.1007/s00425-013-1921-1

Pogorelko G., Lionetti V., Bellincampi D., Zabotina O. Cell wall integrity: targeted post-synthetic modifications to reveal its role in plant growth and defense against pathogens. Plant Signal Behav. 2013;8:e25435. DOI 10.4161/psb.25435

Prabhu S.A., Wagenknecht M., Melvin P., Gnanesh Kumar B.S., Veena M., Shailasree S., Moerschbacher B.M., Kini K.R. Immuno-affinity purification of PglPGIP1, a polygalacturonase-inhibitor protein from pearl millet: studies on its inhibition of fungal polygalacturonases and role in resistance against the downy mildew pathogen. Mol. Biol. Rep. 2015;42(6):1123-1138. DOI 10.1007/s11033-015-3850-5

Prats E., Martinez F., Rojas-Molina M., Rubiales D. Differential effects of phenylalanine ammonia lyase, cinnamyl alcohol dehydrogenase, and energetic metabolism inhibition on resistance of appropriate host and nonhost cereal-rust interactions. Phytopathology. 2007;97(12):1578-1583. DOI 10.1094/PHYTO-97-12-1578

Romero D., Rivera M.E., Cazorla F.M., Codina J.C., Fernández-Ortuño D., Torés J.A., Pérez-García A., de Vicente A. Comparative histochemical analyses of oxidative burst and cell wall reinforcement in compatible and incompatible melon-powdery mildew (*Podosphaera fusca*) interactions. J. Plant Physiol. 2008;165(18):1895-1905. DOI 10.1016/j.jplph.2008.04.020

Ron M., Avni A. The receptor for the fungal elicitor ethylene-inducing xylanase is a member of a resistance-like gene family in tomato. Plant Cell. 2004;16(6):1604-1615. DOI 10.1105/tpc.022475

Rudd J.J., Kanyuka K., Hassani-Pak K., Derbyshire M., Andongabo A., Devonshire J., Lysenko A., Saqi M., Desai N.M., Powers S.J., Hoop-

er J., Ambroso L., Bharti A., Farmer A., Hammond-Kosack K.E., Dietrich R.A., Courbot M. Transcriptome and metabolite profiling of the infection cycle of Zymoseptoria tritici on wheat reveals a biphasic interaction with plant immunity involving differential pathogen chromosomal contributions and a variation on the hemibiotrophic lifestyle definition. Plant Physiol. 2015;167(3):1158-1185. DOI 10.1104/pp.114.255927

San Clemente H., Jamet E. WallProtDB, a database resource for plant cell wall proteomics. Plant Methods. 2015;11(1):2. DOI 10.1186/s13007-015-0045-y

Sanchez-Vallet A., Saleem-Batcha R., Kombrink A., Hansen G., Valkenburg D.J., Thomma B.P., Mesters J.R. Fungal effector Ecp6 outcompetes host immune receptor for chitin binding through intrachain LysM dimerization. Elife. 2013;2:e00790. DOI 10.7554/eLife.00790

Sattler S.E., Funnell-Harris D.L. Modifying lignin to improve bioenergy feedstocks: strengthening the barrier against pathogens? Front. Plant Sci. 2013;4:70. DOI 10.3389/fpls.2013.00070

Sattler S.E., Saathoff A.J., Haas E.J., Palmer N.A., Funnell-Harris D.L., Sarath G., Pedersen J.F. A nonsense mutation in a cinnamyl alcohol dehydrogenase gene is responsible for the sorghum brown midrib 6 phenotype. Plant Physiol. 2009;150(2):584-595. DOI 10.1104/pp.109.136408

Scheller H.V., Ulvskov P. Hemicelluloses. Annu. Rev. Plant Biol. 2010;61:263-289. DOI 10.1146/annurev-arplant-042809-112315

Schoonbeek H.J., Wang H.H., Stefanato F.L., Craze M., Bowden S., Wallington E., Zipfel C., Ridout C.J. Arabidopsis EF-Tu receptor enhances bacterial disease resistance in transgenic wheat. New Phytol. 2015;206(2):606-613. DOI 10.1111/nph.13356

Sella L., Gazzetti K., Faoro F., Odorizzi S., D'Ovidio R., Schafer W., Favaron F. A *Fusarium graminearum* xylanase expressed during wheat infection is a necrotizing factor but is not essential for virulence. Plant Physiol. Biochem. 2013;64:1-10. DOI 10.1016/j.plaphy.2012.12.008

Senthil-Kumar M., Mysore K.S. Non host resistance against bacterial pathogens: retrospectives and prospects. Annu. Rev. Phytopathol. 2013;51:407-427. DOI 10.1146/annurev-phyto-082712-102319

Shadle G.L., Wesley S.W., Korth K.L., Chen F., Lamb C., Dixon R.A. Phenylpropanoid compounds and disease resistance in transgenic tobacco with altered expression of l-phenylalanine ammonia-lyase. Phytochemistry 2003;64(1):153-161. DOI 10.1016/S0031-9422(03)00151-1

Shafiei R., Hang C., Kang J.G., Loake G.J. Identification of loci controlling non-host disease resistance in Arabidopsis against the leaf rust pathogen *Puccinia triticina.* Mol. Plant Pathol. 2007;8(6):773-784. DOI 10.1111/j.1364-3703.2007.00431.x

Shi H., Liu Z., Zhu L., Zhang C., Chen Y., Zhou Y., Li F., Li X. Overexpression of cotton (*Gossypium hirsutum*) dirigent 1 gene enhances lignification that blocks the spread of *Verticillium dahlia*. Acta Biochim. Biophys. Sin. 2012;44(7):555-564. DOI 10.1093/abbs/gms035

Shimizu T., Nakano T., Takamizawa D., Desaki Y., Ishii-Minami N., Nishizawa Y., Minami E., Okada K., Yamane H., Kaku H., Shibuya N. Two LysM receptor molecules, CEBiP and OsCERK1, cooperatively regulate chitin elicitor signaling in rice. Plant J. 2010;64(2):204-214. DOI 10.1111/j.1365-313X.2010.04324.x

Shinya T., Motoyama N., Ikeda A., Wada M., Kamiya K., Hayafune M., Kaku H., Shibuya N. Functional characterization of CEBiP and CERK1 homologs in *Arabidopsis* and rice reveals the presence of different chitin receptor systems in plants. Plant Cell Physiol. 2012;53(10):1696-1706. DOI 10.1093/pcp/pcs113

Smirnova O.G., Ibragimova S.S., Kochetov A.V. Simple database to select promoters for plant transgenesis. Transgenic Res. 2012;21(2):429-437. DOI 10.1007/s11248-011-9538-2

Smirnova O.G., Kochetov A.V. Plant gene promoters responsive to pathogen invasion. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2014;18(4/1):765-775.

Smirnova O.G., Kochetov A.V. Promoters of plant genes responsive to pathogen invasion. Russ. J. Genet.: Applied Res. 2015;5(3):254-261. DOI: 10.1134/S2079059715030181

Smith A.H., Gill W.M., Pinkard E.A., Mohammed C.L. Anatomical and histochemical defence responses induced in juvenile leaves of *Eucalyptus globulus* and *Eucalyptus nitens* by *Mycosphaerella* infection. For. Pathol. 2007;37:361-373. DOI 10.1111/j.1439-0329.2007.00502.x

Szabo L.J., Bushnell W.R. Hidden robbers: the role of fungal haustoria in parasitism of plants. Proc. Natl Acad. Sci. USA. 2001;98(14):7654-7765. DOI 10.1073/pnas.151262398

Takken F.L., Thomas C.M., Joosten M.H., Golstein C., Westerink N., Hille J., Nijkamp H.J., De Wit P.J., Jones J.D. A second gene at the tomato Cf-4 locus confers resistance to *Cladosporium fulvum* through recognition of a novel avirulence determinant. Plant J. 1999;20(3):279-288. DOI 10.1046/j.1365-313X.1999.00601.x

Trdá L., Boutrot F., Claverie J., Brulé D., Dorey S., Poinssot B. Perception of pathogenic or beneficial bacteria and their evasion of host immunity: pattern recognition receptors in the frontline. Front Plant Sci. 2015;6:219. DOI 10.3389/fpls.2015.00219

Underwood W. The plant cell wall: a dynamic barrier against pathogen invasion. Front Plant Sci. 2012;3:85. DOI 10.3389/fpls.2012.00085

van den Burg H.A., Harrison S.J., Joosten M.H., Vervoort J., De Wit P.J. *Cladosporium fulvum* Avr4 protects fungal cell walls against hydrolysis by plant chitinases accumulating during infection. Mol. Plant Microbe Interact. 2006;19(12):1420-1430.

van Esse H.P., Bolton M.D., Stergiopoulos I., de Wit P.J., Thomma B.P. The chitin-binding *Cladosporium fulvum* effector protein Avr4 is a virulence factor. Mol. Plant Microbe Interact. 2007;20(8):1092-1101.

Voigt C.A. Callose-mediated resistance to pathogenic intruders in plant defense-related papillae. Front Plant Sci. 2014;5:168. DOI 10.3389/fpls.2014.00168

Wang K., Senthil-Kumar M., Ryu C.M., Kang L., Mysore K.S. Phytosterols play a key role in plant innate immunity against bacterial pathogens by regulating nutrient efflux into the apoplast. Plant Physiol. 2012;158(4):1789-1802. DOI 10.1104/pp.111.189217

Way H.M., Kazan K., Mitter N., Goulter K.C., Birch R.G., Manners J.M. Constitutive expression of a phenylalanine ammonia-lyase gene from *Stylosanthes humilis* in transgenic tobacco leads to enhanced disease resistance but impaired plant growth. Physiol. Mol. Plant Pathol. 2002;60(6):275-282. DOI 10.1006/pmpp.2002.0407

Wiesel L., Newton A.C., Elliott I., Booty D., Gilroy E.M., Birch P.R., Hein I. Molecular effects of resistance elicitors from biological origin and their potential for crop protection. Front Plant Sci. 2014;5:655. DOI 10.3389/fpls.2014.00655

Wróbel-Kwiatkowska M., Starzycki M., Zebrowski J., Oszmiañski J., Szopa J. Lignin deficiency in transgenic flax resulted in plants with improved mechanical properties. J. Biotechnol. 2007;128(4):919-934. DOI 10.1016/j.jbiotec.2006.12.030

Xu L., Zhu L., Tu L., Liu L., Yuan D., Jin L., Long L., Zhang X. Lignin metabolism has a central role in the resistance of cotton to the wilt fungus *Verticillium dahliae* as revealed by RNA-Seq-dependent transcriptional analysis and histochemistry. J. Exp. Bot. 2011;62:5607-5621.

Zipfel C. Plant pattern-recognition receptors. Trends Immunol. 2014;35(7):345-351. DOI 10.1016/j.it.2014.05.004

Zhang H., Wang C., Cheng Y., Wang X., Li F., Han Q., Xu J., Chen X., Huang L., Wei G., Kang Z. Histological and molecular studies of the non-host interaction between wheat and *Uromyces fabae*. Planta. 2011;234(5):979-991. DOI 10.1007/s00425-011-1453-5

Zhao J., Buchwaldt L., Rimmer S.R., Sharpe A., Mcgregor L., Bekkoui D., Heqedus D. Patterns of differential gene expression in *Brassica napus* cultivars infected with *Sclerotinia sclerotiorum*. Mol. Plant Pathol. 2009;10(5):635-649. DOI 10.1111/j.1364-3703.2009.00558.x

Вавиловский журнал генетики и селекции. 2015;19(6):724-730
Doi 10.18699/VJ15.084

Дизайн и проверка действия малых химических соединений, направленных на ингибирование белка FADD

Н.В. Иванисенко[1, 2], Л. Хиллерт[3], В.А. Иванисенко[1], И.Н. Лаврик[1, 3]

1 Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия
2 Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия
3 Факультет прикладных исследований воспалительных процессов, Институт экспериментальной внутренней медицины, Университет Отто фон Гюрике, Магдебург, Германия

Рецептор CD95 является одним из наиболее изученных представителей семейства рецепторов смерти. Его активация ведет к запуску апоптоза – программы программируемой клеточной гибели через образование комплекса DISC (Death-Inducing Signaling Complex – комплекс, индуцирующий смерть). Основным структурным звеном комплекса CD95 DISC является адаптерный белок FADD (Fas-Associated Death Domain – Fas-ассоциированный домен смерти), олигомеризация которого необходима для последующей активации прокаспазы-8 в рецепторном комплексе. Белок FADD характеризуется наличием домена смерти и домена DED (Death Effector Domain – эффекторный домен смерти). Домен смерти рецептора cD95 связывается с соответствующим доменом белка-адаптера FADD, а за счет связывания доменов DED происходит образование комплекса с участием прокаспазы-8, 10 и белка c-FliP. Поиск ингибиторов взаимодействия белка FADD и других ключевых компонент комплекса DISC представляет огромный интерес для исследования структурно-функциональной организации данного комплекса, молекулярных механизмов клеточной гибели и лечения нейродегенеративных заболеваний. Был осуществлен поиск малых химических соединений *in silico*, направленно взаимодействующих с доменом DED белка FADD. Для достижения данной цели были проведены молекулярное моделирование белковых комплексов и виртуальный скрининг потенциальных ингибиторов FADD, а также разработана новая методология экспериментальной проверки их биологического эффекта на клеточных линиях. Компьютерно-экспериментальный анализ позволил выявить оптимальную конформацию белка FADD для дизайна низкомолекулярных соединений, способных связываться в районе аминокислотного остатка Y25. Мы предполагаем, что дальнейшая оптимизация структур химических соединений, способных связываться с гидро-фобным карманом вблизи аминокислотного остатка Y25 FADD, позволит создать новые перспективные ингибиторы программируемой клеточной гибели.

Ключевые слова: апоптоз; CD95; FADD; молекулярное моделирование; DiSc; каспаза.

HOW TO CITE THIS ARTICLE?
ivanisenko n.V., Hillert l ., ivanisenko V.A., l avrik i.n . Design and experimental validation of the action of small molecule-based inhibitors of the FADD protein. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2015;19(6):724-730. Doi 10.18699/V J15.084

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?
Иванисенко Н.В., Хиллерт л., Иванисенко В.А., лаврик И.Н. Дизайн и проверка действия малых химических соединений, направленных на ингибирование белка FADD. Вавиловский журнал генетики и селекции. 2015;19(6):724-730. Doi 10.18699/VJ15.084

Received 14.09.2015
Accepted for publication 12.10.2015

e-mail: ivanisenko@bionet.nsc.ru

Design and experimental validation of the action of small molecule-based inhibitors of the FADD protein

N.V. Ivanisenko[1, 2], L. Hillert[3], V.A. Ivanisenko[1], I.N. Lavrik[1, 3]

C D95 is one of the best studied members of the death receptor family. Activation of cD95 leads t o the induction of the cell death programme, apoptosis, via formation of the death-inducing signaling complex (DiSc). FADD is a key adaptor protein for the formation of the cD95 DiSc and ac tivation of caspase-8 in the receptor complex. FADD comprises the death domain and the death effector domain (DED). The death domain is essential for the inter-actions of FADD with CD95, while DED is necessary for the recruitment of procaspase-8, -10 and the protein c-FliP into the DiSc. The search for the inhibitors that would block the interactions of FADD with the other core proteins of the DiSc is essen tial for the studies of the structure and function of this complex, investigation of the apoptosis mechanisms and development of new treatments for neurodegenerative diseases. in the course of this work, the screening for small inhibitors in silico that selectively interact with DED has been performed. For this purpose, the molecular modeling of the protein complexes and virtual screening of the potential inhibitors of FADD has been performed. in addition, a new technology to test the activity of these inhibitors has been developed. The computational and experimental analysis performed allowed us to characterize the optimal conformation of the FADD protein for the design of the small molecules that can bind in the region of amino acid residue Y25. we presume that further optimization of the structures of chemical compounds that can bind with the hydro-phobic pocket next to the residue Y25 of FADD will allow for the creation of the new perspective inhibitors of the programmed cell death.

Key words: apoptosis; cD95; FADD; molecular modeling; DiSc; caspase .

Апоптоз – это неотъемлемая функция всех многоклеточных организмов, которая отвечает за уничтожение всех поврежденных клеток (Krammer et al., 2007). Дефекты в системе программируемой клеточной гибели, апоптоза, были описаны для целого ряда онкологических и нейродегенеративных заболеваний. Имеются два основных пути, по которым передаются апоптотические сигналы: внутренний, или митохондриальный, и внешний, передаваемый через так называемые рецепторы смерти (DR – Death Receptor) (Lavrik, Krammer, 2012). При поступлении сигнала инициации апоптоза происходит образование ряда апоптотических сигнальных комплексов, в которых осуществляется активация основных ферментов апоптоза – каспаз, что, собственно, и является сигналом к инициации апоптоза (Lavrik et al., 2005). Поэтому изучение молекулярных механизмов апоптоза и действия каспаз, включающих в себя исследование белковых комплексов, в которых происходит их активация, является актуальнейшей задачей современных биомедицинских исследований.

В настоящее время в научной литературе представлено значительное количество работ, посвященных изучению апоптоза, индуцируемого через рецепторы смерти (Krammer et al., 2007). Рецептор CD95, который также известен как Fas или APO-1, является одним из наиболее изученных представителей данного семейства. Известно, что первый этап инициации апоптоза через рецептор CD95 заключается в образовании комплекса DISC (Death-Inducing Signaling Complex – комплекс, индуцирующий смерть) (Lavrik, Krammer, 2012). CD95 DISC – это сложный макромолекулярный комплекс, инициирующий апоптоз, который состоит из рецептора CD95, белка-адаптера FADD (Fas-associated death domain – ФАС-ассоциированный домен смерти), прокаспазы-8, 10 и белка c-FLIP (cellular FLICE (FADD-like interleukin-1 beta-converting enzyme) Ihibitory Protein – клеточный белок, ингибирующий FLICE (FADD-подобный белок-конвертирующий интерлейкин-1) (Krammer et al., 2007). Гомотипические взаимодействия играют центральную роль в формировании комплекса CD95 DISC. Домен смерти рецептора CD95 взаимодействует с соответствующим доменом белка-адаптера FADD, а за счет взаимодействия доменов DED (Death Effector Domain – эффекторный домен смерти) происходит связывание в комплекс прокаспазы-8, прокаспазы-10 и белка c-FLIP (Lavrik et al., 2005). В результате образования комплекса происходит активация каспазы-8, что и служит основным сигналом запуска апоптоза (Schleich et al., 2012).

Однако, несмотря на значительный прогресс в понимании апоптоза, по-прежнему остается невыясненным целый ряд механизмов функционирования молекулярных комплексов, индуцирующих апоптоз. В частности, это связано с тем, что структурная информация о молекулярных взаимодействиях в составе макромолекулярных белковых комплексов, инициирующих апоптоз, в большинстве случаев отсутствует. Поэтому огромное значение для исследований структуры и функции макромолекулярных комплексов, индуцирующих апоптоз, приобретают передовые технологии молекулярного моделирования, которые позволяют создавать малые молекулы, специфически ингибирующие белок-белковые взаимодействия. Данный подход позволяет осуществлять направленное воздействие на определенные белок-белковые взаимодействия в сигнальном пути апоптоза, индуцируемого через рецепторы смерти, и, таким образом, получать уникальную информацию об особенностях функционирования исследуемого сигнального пути. Широкие перспективы для изучения этих процессов предоставляют технологии виртуального молекулярного скрининга, предназначенные для идентификации малых химических соединений, эффективно ингибирующих функциональную активность белков. Компьютерное моделирование ингибиторов основных компонент сигнального пути апоптоза, индуцируемого через рецепторы смерти, и экспериментальное изучение эффекта их действия могут позволить получить новые фундаментальные знания о молекулярных механизмах биологического процесса клеточной смерти, включая знания о межмолекулярных взаимодействиях в составе апоптотических белковых платформ, и создадут основу для разработки принципиально новых лекарственных препаратов, направленных на лечение заболеваний, обусловленных дефектами в системе рецепторов смерти.

В данной работе описаны стратегии поиска малых химических соединений, направленно взаимодействующих с основным структурным белком комплекса CD95 DISC, белком FADD, и изучение с их помощью особенностей структурно-функциональной организации данного комплекса и молекулярных механизмов клеточной смерти. Для достижения данной цели были проведены молекулярное моделирование белковых комплексов и виртуальный скрининг потенциальных ингибиторов с последующей экспериментальной проверкой их биологического эффекта на клеточных линиях. В качестве основной мишени для дизайна малых химических соединений был выбран белок FADD, который является основным структурным звеном сборки комплекса CD95 DISC и, тем самым, ключевым регулятором сигнальных путей апоптоза, индуцируемых через активацию рецепторов смерти (Lavrik, Krammer, 2012). Как было отмечено выше, именно связывание белка FADD с рецептором смерти предоставляет возможность активации прокаспазы-8 в рецепторном комплексе и, соответственно, в отсутствие активного белка FADD инициация апоптоза является невозможной. Таким образом, создание специфических соединений, селективно блокирующих взаимодействие между белком FADD и другими ключевыми компонентами комплекса DISC, является приоритетной задачей современных биомедицинских исследований, решение которой позволит создать новые терапевтические подходы для лечения онкологических и нейродегенеративных заболеваний.

Материалы и методы

Процедура виртуального скрининга

Молекулярный докинг лигандов (соединений FADDin) проводился с использованием программы Glide (Schrodinger, Inc.) (Halgren et al., 2004; Friesner et al., 2004; 2006). Данная программа использует модифицированную версию функции ChemScore для оценки энергии взаимодействий «белок – лиганд». Виртуальный скри-

нинг проводился с использованием режима стандартной точности (SP), а также режима экстраточности (XP). Для проведения виртуального скрининга использовались библиотеки коммерчески доступных химических соединений, подготовленных для молекулярного докинга: ZINC NCI Diversity Set, включающая около 1,9 тыс. соединений, а также библиотека, состоящая из >3,5 млн соединений (Irwin et al., 2012).

Подготовка белка для виртуального скрининга осуществлялась с использованием модуля Protein Preparation, входящего в состав пакета Schrodinger (Sastry et al., 2013). Перед началом процедуры молекулярного докинга проводились предварительная замена Y25F, а также минимизация состояния белка с ограничением на среднеквадратичное отклонение атомов не более 0,3 Е.

Процедура проверки действия ингибиторов Fa DD

Проверка действия соединений FADDin (FADD Inhibitor – ингибитор FADD) проводилась в Т-клеточной линии Jurkat при использовании лиганда CD95 (CD95L, CD95 Ligand) в концентрации 60 нг/мл. Клеточная гибель была измерена с помощью набора CellTiter-Glo (Promega, Германия) для определения жизнеспособности клеток на основе их метаболической активности путем детекции АТФ. Клетки Jurkat (2×10^4) были обработаны соединениями FADDin в концентрациях 5 мкМ, 10 мкМ и 20 мкМ за 2 ч до добавления CD95L с последующей инкубацией с CD95L в течение 6 и 22 ч. Содержание АТФ и жизнеспособность клеток определялись согласно инструкциям производителя.

Для контроля отсутствия неспецифических эффектов на клеточную гибель растворителя к клеткам линии Jurkat добавлялся растворитель диметилсульфоксид (ДМСО/DMSO) для растворения соединений FADDin. При этом его количество соответствовало количеству DMSO при добавлении 20 мкМ соединения FADDin.

результаты и обсуждение

Поиск малых соединений, ингибирующих белок Fa DD *in silico*

Для проведения виртуального скрининга было решено использовать единственную опубликованную на сегодняшний день структуру полноразмерного белка FADD (pdb 2GF5), полученную с помощью метода ядерного магнитного резонанса (ЯМР), с использованием которой было выявлено наличие 25 различных конформаций этого белка (Carrington et al., 2006). В данной работе было показано, что аминокислотный остаток F25 играет важную роль при олигомеризации белка FADD, а также то, что аминокислотная замена F25Y заметно уменьшает эффективность такого типа олигомеризации. Более того, в работе Carrington с коллегами было сделано предположение о наличии гидрофобной полости в районе остатка F25. Поэтому в настоящей работе виртуальный скрининг был сфокусирован на поиске низкомолекулярных соединений, способных связываться с участком белка вблизи аминокислотного остатка F25. При этом была выбрана структура, содержащая аминокислотную замену F25Y (pdb 2GF5). Анализ пространственной структуры белка *in silico* показал расположение потенциальной полости для связывания низкомолекулярных соединений вблизи альфа-спиралей α1, α2 и петли, соединяющей спирали α3 и α4 (рис. 1), что и послужило предпосылкой для проведения скрининга *in silico*.

Для того чтобы выбрать конформацию белка FADD, которая наиболее эффективно позволяла бы связывать низкомолекулярные соединения с выбранным участком этого белка, был проведен виртуальный скрининг библиотеки NCI Diversity Set, состоящей из ~1,9 тыс. соединений, для всех 25 конформаций белка. Конформации № 2, 11, 13, 23 и 25 имели наибольшее среднее значение так называемой оценочной функции в режиме стандартной точности (SP Score) и были использованы для последующего виртуального скрининга большой библиотеки соединений.

В результате из коммерчески доступных соединений ZINC (>3,5 млн соединений) с использованием режима SP для каждой из указанных выше конформаций были отобраны 1 000 соединений с наилучшим значением оценочной функции SP Score. Для полученного набора соединений был проведен дополнительный молекулярный докинг с использованием режима XP. Были отобраны 100 соединений, имеющих наибольшее значение оценочной функции XP Score для всех анализируемых конформаций белка. Из полученного набора с использованием визуального анализа (Bissantz et al., 2010) были выбраны шесть соединений для последующей экспериментальной проверки *in vitro* (таблица).

Данные соединения были названы FADDin. На рис. 2 показаны конформации со связанными потенциальными низкомолекулярными ингибиторами, использованные для молекулярного докинга.

Экспериментальная проверка действия соединений

Для проверки действия соединений, полученных в результате скрининга, было решено использовать индукцию апоптоза при добавлении лиганда CD95 (CD95L) к Т-клеточной линии Jurkat. Данная форма индукции внешнего пути апоптоза широко используется в ряде исследований и сопровождается эффективным образованием комплекса CD95 DISC, активацией каспаз и индукцией апоптоза (Lavrik et al., 2012). Клеточная гибель была измерена с помощью набора CellTiter-Glo (Promega, Германия) для определения жизнеспособности клеток по их метаболической активности, основанного на детекции АТФ. При этом в случае эффективной индукции клеточной гибели происходит значительное уменьшение содержания АТФ в клетках. Соединения FADDin, согласно моделированию *in silico*, предположительно ингибируют олигомеризацию белка FADD, что должно препятствовать связыванию и последующей активации каспазы-8 в комплексе CD95 DISC и, тем самым, приводить к ингибированию апоптоза. Таким образом, в ходе экспериментальной проверки действия соединения FADDin ожидалось ингибирование клеточной гибели при индукции апоптоза через рецептор CD95 при одновременном добавлении соединений FADDin. В частности, с учетом того что ингибирование клеточной гибели не должно менять количество жизнеспособных клеток, ожидалось отсутствие снижения в них уровня содержания АТФ, поскольку для определения

жизнеспособности клеток, как упоминалось выше, использовался набор CellTiter-Glo (Promega, Германия), работа которого основана на детекции АТФ. Таким образом, индукция апоптоза измерялась по снижению содержания АТФ в клетках относительно необработанных клеток, а вывод по ингибированию апоптоза мог быть сделан только при неизменном уровне АТФ.

Были проверены несколько концентраций соединений FADDin (5, 10 и 20 μM), а также два временных интервала: 6 и 22 ч. Промежуток времени 6 ч был выбран для анализа различий на начальных этапах апоптоза, поскольку в данный временной интервал количество погибших клеток незначительно, что может позволить детекцию различий в скорости индукции апоптоза под действием соединений FADDin (рис. 3). Временной интервал 22 ч соответствует поздним этапам апоптоза, когда большинство клеток уже погибло и количество жизнеспособных клеток значительно уменьшилось (рис. 4), что дает возможность проанализировать влияние FADDin на окончательное число клеток, вошедших в апоптоз. В ходе экспериментов соединения FADDin добавляли за 2 ч до добавления CD95L, после чего в течение 6 и 22 ч проводили инкубацию, как упоминалось выше.

Добавление к клеткам почти всех соединений группы FADDin в комбинации с CD95L не вызвало ожидаемого ингибирующего эффекта на индукцию апоптоза, поскольку изменения содержания АТФ в клетках при добавлении только CD95L или в комбинации с FADDin не наблюдалось (рис. 3, 4). Более того, следует отметить, что добавление соединений FADDin без CD95L к клеткам линии Jurkat в некоторых случаях вело к снижению содержания АТФ в клетках при обработке только соединениями FADDin в концентрациях 5, 10 и 20 μM, особенно при инкубации в течение 22 ч (рис. 3, 4, столбцы 3–5). В частности, данный эффект заключается в понижении количества АТФ, что

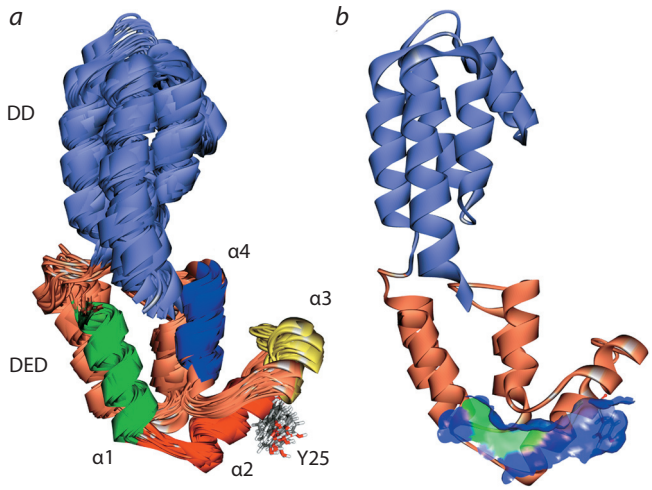


Fig. 1. The structure of the FADD protein used in this work.
(*a*) Superposition of 25 conformations of FADD obtained by n Mr (pdb 2GF5). r esidue Y25 is shown as a ball-and-stick model. (*b*) The conformation of protein #2 with the domain binding low-molecular-weight compounds. The putative binding domain is shown as a surface accessible for the solvent.

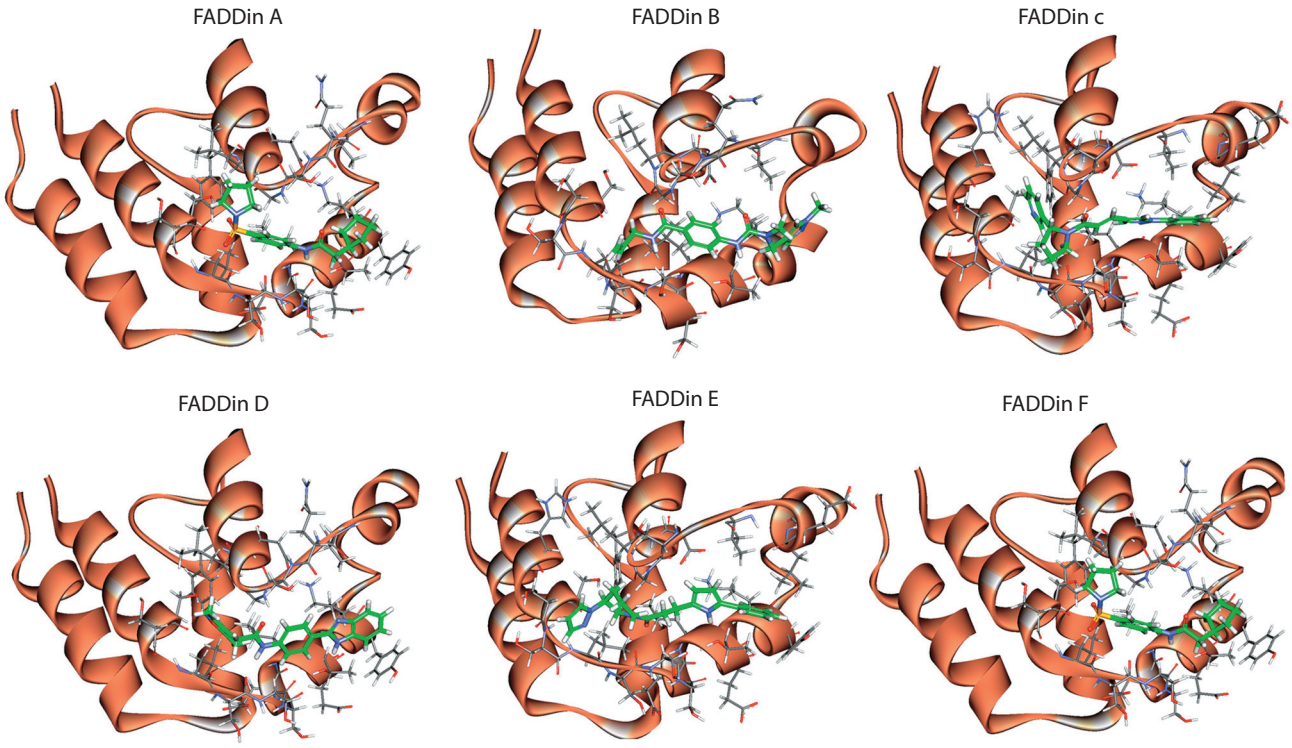
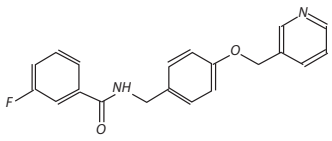
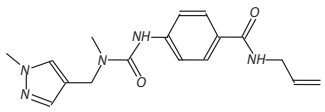
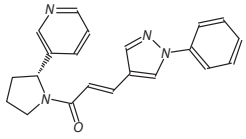
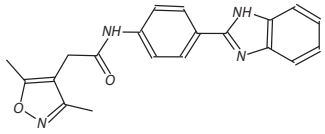
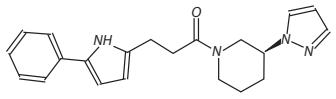
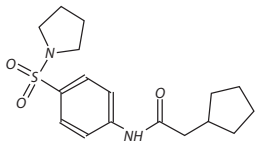


Fig. 2. The position of the binding of the best ligands selected for three n Mr conformations of the FADD protein.

c haracterization of FADDin compounds tested in this work. The ligand chemical structure, docking ranking, and the ordinal number of the FADD 2GF5 conformation used for its generation

FADDin type	Structure	identifie	SP Score	XP Score	Protein conformation no.
FADDin A		Zinc 36391506	143	27	23
FADDin B		Zinc 48159387	83	28	13
FADDin c		Zinc 12793275	575	40	23
FADDin D		Zinc 22394070	306	55	2
FADDin e		Zinc 69568542	813	13	23
FADDin F		Zinc 06271291	944	24	2

свидетельствует об индукции клеточной гибели лишь при добавлении соединений FADDin и, тем самым, о неспецифических эффектах данных соединений на индукцию клеточной гибели.

Важным исключением оказалось соединение FADDin B. Характерной чертой этого соединения было отсутствие неспецифических эффектов на индукцию клеточной гибели: изменение содержания АТФ в клетках при обработке FADDin B даже в самой высокой применяемой концентрации 20 μM в течение 6, а также 22 ч не наблюдалось (рис. 3, 4, столбец 3).

Отметим, что при сравнительном анализе неспецифических эффектов соединений FADDin на индукцию клеточной гибели наибольшую токсичность показывает FADDin A (рис. 3, 4, столбцы 3–5). При обработке этим соединением в течение 6 ч наблюдалась значительная потеря содержания АТФ в клетках, свидетельствующая о снижении их жизнеспособности, в то время как все другие соединения FADDin в течение данного интервала времени подобного действия не оказывали. Наибольшая потеря содержания АТФ наблюдалась при использовании FADDin A в концентрации 20 μM (рис. 3, столбец 3), но также следует отметить снижение содержания АТФ в клетках при использовании FADDin A в концентрациях 10 μM (рис. 3, столбец 4) и даже 5 μM (рис. 3, столбец 5). Более того, если взять за критерий отсутствие неспецифических эффектов при инкубации в течение 22 ч с концентрацией FADDin 20 μM (рис. 4, столбец 3), то по показателю клеточной гибели на первом месте будет соединение FADDin B, на втором – FADDin C, затем – FADDin E и F, за которыми последует FADDin D, и, наконец, на последнем месте, как было отмечено выше, будет находиться FADDin A. При этом только FADDin B не проявлял неспецифических эффектов соединений FADDin на индукцию клеточной гибели, в то время как все остальные соединения данной группы продемонстрировали некоторую токсичность: максимальную – в случае FADDin A и небольшую (только при инкубации в течение 22 ч в самой высокой концентрации) – в случае FADDin C.

При сравнительном анализе эффектов FADDin при комбинаторном воздействии вместе с CD95L наилучшее действие имело FADDin B. Добавление к клеткам соединения FADDin B в комбинации с CD95L вызвало небольшой ингибирующий эффект на индукцию программируемой клеточной гибели. В частности, было детектировано большее содержание АТФ в клетках во временной точке 6 ч при обработке CD95L в комбинации с соединением FADDin B в концентрациях 20 и 10 μM (рис. 3, столбцы 8, 9) по сравнению с содержанием АТФ при обработке только CD95L (рис. 3, столбцы 6, 7). При этом добавление FADDin B в концентрации 20 μM оказывало чуть больший эффект (рис. 3, столбец 8), чем в концентрации 10 μM (рис. 3, столбец 9). Как отмечалось выше, в ходе проверки действия соединений FADDin ожидалось отсутствие сни-

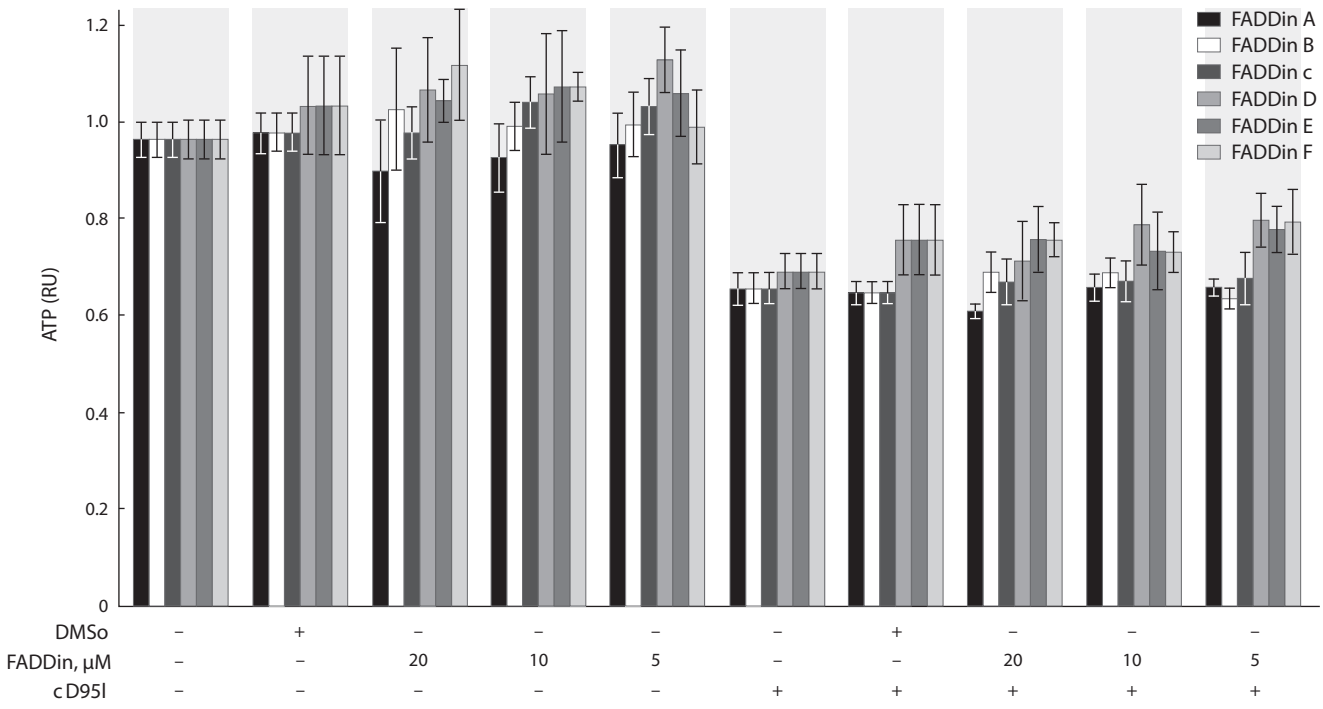


Fig. 3. Action of compounds FADDin A, B, C, D, E и F on Jurkat cells after 6-h incubation with CD95l .

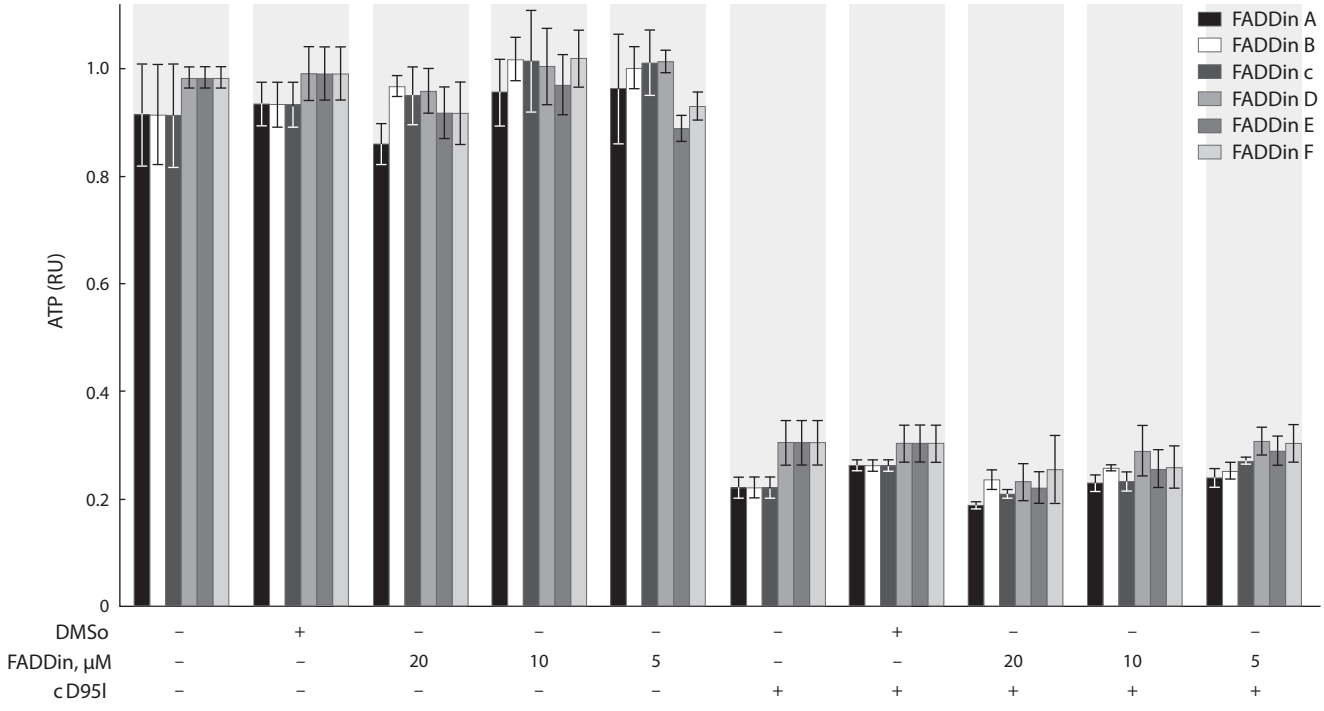


Fig. 4. Action of compounds FADDin A, B, C, D, E и F on Jurkat cells after 22-h incubation with CD95l .

жения уровня содержания АТФ в клетках при индукции апоптоза через рецептор CD95 при добавлении соединений FADDin за счет ингибирования клеточной гибели и, соответственно, повышения количества жизнеспособных клеток. Таким образом, эффекты, наблюдаемые в данном эксперименте, вполне можно отнести к ожидаемому эффекту FADDin В на ингибирование апоптоза через связывание с белком FADD.

Отметим, что ингибирующее влияние соединения FADDin В при инкубации вместе с CD95L во временной

точке 22 ч отсутствовало (рис. 4, столбцы 8, 9). По-видимому, данное явление связано с тем, что действие FADDin B в концентрациях 20 и 10 μM является недостаточно эффективным, в связи с этим его активность можно детектировать только на начальных стадиях индукции апоптоза.

Таким образом, проведенный анализ показывает, что соединение FADDin B обладает потенциалом для ингибирования апоптоза, что должно быть проверено в последующих исследованиях. Отсутствие явно выраженных неспецифических эффектов соединений FADDin на индукцию клеточной гибели позволяет сделать предположение о том, что концентрация FADDin B может быть увеличена в дальнейших экспериментах для увеличения эффективности его действия. Более того, данные экспериментального анализа позволяют предположить, что именно конформация белка FADD № 13, возможно, является активной при образовании комплекса CD95 DISC и рекрутировании прокаспазы-8 в данный комплекс, что соответствует ключевому этапу в запуске апоптоза через рецептор CD95.

Также небольшие эффекты по ингибированию апоптоза проявляли FADDin E и F при обработке вместе с CD95L в течение 6 ч в концентрации 5 μM (рис. 3, столбец 10). Также следует отметить потенциальное действие FADDin D при инкубации вместе с CD95L в течение 6 ч и концентрациях 10 и 5 μM (рис. 3, столбцы 9, 10). Если проранжировать эффект действия данных химических соединений, принимая за критерий влияние на клеточную гибель при обработке вместе с CD95L в течение 6 ч и концентрации FADDin 20 μM, то первое место, безусловно, занимает FADDin B, а последнее – FADDin A.

При этом следует отметить, что наблюдаемые эффекты по ингибированию апоптоза, индуцируемого через рецептор CD95, были не очень большими, что позволяет предложить дальнейшую химическую модификацию полученных малых химических соединений как стратегию продолжения данных исследований. Альтернативным объяснением может служить предположение, что конформации белка FADD, такие как № 2 и 23, не содержат гидрофобного кармана вблизи аминокислотного остатка Y25, благоприятного для связывания низкомолекулярных соединений. В дальнейших исследованиях предполагается проверить возможность дизайна малых химических соединений для других участков белка FADD, которые могут служить сайтами для их связывания.

Таким образом, в ходе данной работы был проведен поиск малых химических соединений *in silico*, направленно взаимодействующих с основным структурным белком комплекса CD95 DISC, FADD, и с их помощью изучены особенности структурно-функциональной организации этого комплекса и молекулярные механизмы клеточной гибели. Для достижения поставленной цели были проведены молекулярное моделирование белковых комплексов и виртуальный скрининг потенциальных ингибиторов (соединений FADDin), а также разработана новая методология экспериментальной проверки их биологического эффекта на клеточной линии Jurkat. Действие соединений FADDin на ингибирование апоптоза, индуцируемого через рецептор CD95, было не очень значительным, что позволяет предложить дальнейшую химическую модификацию соединений группы FADDin как стратегию продолжения данных исследований. Наибольшую активность продемонстрировало соединение FADDin B, что, в свою очередь, позволяет высказать предположение о том, что именно конформация белка FADD № 13, которая является мишенью для соединения FADDin B, представляет потенциал для проведения дальнейшего структурного дизайна низкомолекулярных химических ингибиторов программируемой клеточной гибели.

Acknowledgments

The work was supported by the Russian Science Foundation, project 14-44-00011 “Programmed cell death induced via death receptors: identification of molecular mechanisms initiating apoptosis by molecular simulation”.

Conflict of interest

The authors declare no conflicts of interest.

References

Bissantz C., Kuhn B., Stahl M. A medicinal chemist’s guide to molecular interactions. J. Med. Chemistry. 2010;53(14):5061-5084. DOI 10.1021/jm100950p

Carrington P.E., Sandu C., Wei Y., Hill J.M., Morisawa G., Huang T., Gavathiotis E., Wei Y., Werner M.H. The structure of FADD and its mode of interaction with procaspase-8. Mol. Cell. 2006;22(5):599-610. DOI 10.1016/j.molcel.2006.04.018

Friesner R.A., Banks J.L., Murphy R.B., Halgren T.A., Klicic J.J., Mainz D.T., Repasky M.P., Knoll E.H., Shaw D.E., Shelley M., Perry J.K., Francis P., Shenkin P.S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J. Med. Chem. 2004;47(7):1739-1749. DOI 10.1021/jm0306430

Friesner R.A., Murphy R.B., Repasky M.P., Frye L.L., Greenwood J.R., Halgren T.A., Sanschagrin P.C., Mainz D.T. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. J. Med. Chem. 2006;49(21):6177-6196. DOI 10.1021/jm051256o

Halgren T.A., Murphy R.B., Friesner R.A., Beard H.S., Frye L.L., Pollard W.T., Banks J.L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J. Med. Chem. 2004;47(7):1750-1759. DOI 10.1021/jm030644s

Irwin J.J., Sterling T., Mysinger M.M., Bolstad E.S., Coleman R.G. ZINC: a free tool to discover chemistry for biology. J. Chem. Inf. Model. 2012;52(7):1757-1768. DOI 10.1021/ci3001277

Krammer P.H., Arnold R., Lavrik I.N. Life and death in peripheral T cells. Nat. Rev. Immunol. 2007;7(7):532-542. DOI 10.1038/nri2115

Lavrik I.N., Golks A., Krammer P.H. Caspases: pharmacological manipulation of cell death. J. Clin. Invest. 2005;115(10):2665-2672. DOI 10.1172/JCI26252

Lavrik I.N., Krammer P.H. Regulation of CD95/Fas signaling at the DISC. Cell Death Differ. 2012;19(1):36-41. DOI 10.1038/cdd.2011.155

Sastry G.M., Adzhigirey M., Day T., Annabhimoju R., Sherman W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. J. Comput. Aid. Mol. Des. 2013;27(3):221-234. DOI 10.1007/s10822-013-9644-8

Schleich K., Warnken U., Fricker N., Öztürk S., Richter P., Kammerer K., Schnoelzer M., Krammer P.H., Lavrik I.N. Stoichiometry of the CD95 death-inducing signaling complex: experimental and modeling evidence for a death effector domain chain model. Mol. Cell. 2012;47(2):306-319. DOI 10.1016/j.molcel.2012.05.006

Вавиловский журнал генетики и селекции. 2015;19(6):731-737
Doi 10.18699/VJ15.101

Компьютерное моделирование пространственных структур пептидов из MUC1, способных ингибировать апоптоз

Н.В. Иванисенко[1, 2], И.Н. Лаврик[1, 3], В.А. Иванисенко[1]

1 Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия 2 Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия 3 Факультет прикладных исследований воспалительных процессов, Институт экспериментальной внутренней медицины, Университет Отто фон Гюрике, Магдебург, Германия

Поиск эффективных ингибиторов апоптоза является актуальной задачей при создании лекарственных препаратов, в том числе направленных на лечение нейродегенеративных заболеваний. Инициация апоптоза осуществляется через образование макромолекулярных комплексов, в которых происходит активация каспаз – основных ферментов, ответственных за гибель клетки. Одним из таких макромолекулярных комплексов является комплекс DiSc (D eath-inducing Signaling c omplex – комплекс, индуцирующий смерть), который играет ключевую роль при индукции так называемого внешнего пути апоптоза, в формировании которого центральное место занимает белок-адаптер FADD (Fas-Associated Death Domain– Fas-ассоциированный домен смерти). Поэтому ингибиторы белка FADD, препятствующие выполнению его функций в составе комплекса DISC, могут быть потенциальными лекарствами, подавляющими запуск апоптоза, а изучение молекулярного механизма их действия представляет высокий интерес для понимания функционирования путей передачи сигнала апоптоза. Известно, что одним из природных белков-ингибиторов FADD является протеогликан MUC1 из группы муцинов. В частности, было установлено, что два пептида из первичной структуры цитоплазматического домена MUC1 (MUC1-cD, MUC1-cytoplasmic domain) также способны ингибировать связывание каспазы-8 с FADD. Однако пространственная структура белка MUC1-cD до сих пор не расшифрована, что существенно усложняет рациональное конструирование потенциальных лекарств на основе данных пептидов. В связи с этим целью настоящей работы были компьютерное моделирование пространственных структур пептидов MUC1-cD, соответствующих фрагментам этого белка (1–20 и 46–72), а также анализ их конформационных свойств. Основное внимание в работе было уделено пептиду MUC1-CD (46–72), который способен связываться с FADD. С использованием метода молекулярной динамики в неявной воде было показано, что пептид MUC1-CD (46–72) обладает конформацией, сходной с таковой у ряда участков домена DED (Death Effector Domain – эффекторный домен смерти) белка каспазы-8. Было обнаружено как минимум 4 участка белка каспазы-8, пространственную структуру которых может принимать пептид MUC1-CD (46–72). Полученные результаты показывают, что молекулярный механизм ингибирующей активности данного пептида может заключаться в конкурентном связы-вании с FADD за счет структурного и конформационного сходства с белок-связыва-ющими участками домена DED каспазы-8.

Ключевые слова: апоптоз; программируемая клеточная гибель; FADD; MUC1; каспаза- 8; молекулярная динамика; неявная вода; обобщенная модель Борна; предсказание структуры белка.

Computer simulation of the spatial structure of MUC1 peptides capable of inhibiting apoptosis

N.V. Ivanisenko[1, 2], I.N. Lavrik[1, 3], V.A. Ivanisenko[1]

1 Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
2 Novosibirsk State University, Novosibirsk, Russia
3 Department of Translational Inflammation, Institute of Experimental Internal Medicine, Otto von Guericke University, Magdeburg, Germany

identification of new effective inhibitors of apoptosis is an important task for drug development for treatment of a number diseases including neurogenerative diseases. initiation of apoptosis occurs via the forma-tion of macromolecular protein complexes. in these complexes, activation of key enzymes in apoptosis, caspases, takes place. one of those macromolecular complexes is DiSc (death-inducing signaling complex) playing a central role in the induction of the extrinsic apoptosis pathway. The adaptor protein FADD has a major role in the formation of the DiSc. Therefore, inhibitors of FADD, preventing its function in the DISC, can act as potential drugs inhibiting apoptosis. Furthermore, the study of the mechanisms of action of these inhibitors is of great interest for understanding the mechanisms of the signal transduction pathways of apoptosis. it has been reported that a natural protein inhibitor of FADD is mucin-type 1 glycoprotein (MUC1). In particular, two fragments of the primary structure of the cytoplasmic domain of MUC1 (MUC1-cD) are capable of inhibiting the binding of caspase-8 to FADD. However, the three-dimensional structure of MUC1 has not been obtained yet. it complicates significantly the rational design of potential drugs on the basis of these peptides. In this context, the aim of the present study was *in silico* prediction of three-dimensional structures of MUC1-CD peptides corresponding to protein fragments (1-20 and 46-72), as well as analysis of their conformational properties. The main focus of the work was given to the peptide MUC1-CD (46-72), which is capable of binding to FADD. Using the methods of molecular dynamics in the implicit water it was shown that the peptide MUC1-cD (46-72) can take conformations similar to the conformations of a number of fragments of the caspase-8 DED domain. it was found that the structure of the peptide MUC1-cD (46-72) is similar to the spatial structure of at least four fragments of caspase-8. These results indicate that the molecular mechanism of the inhibitory activity of the peptide can be explained by competitive binding with FADD due to the structural and conformational similarity with the fragments of the caspase-8 DED domain.

Key words: apoptosis; programmed cell death; FADD; MUC1; caspase-8; molecular dynamics; implicit salvation; Generalized Born model; protein structure prediction.

Received 14.09.2015
Accepted for publication 21.10.2015

e-mail: ivanisenko@bionet.nsc.ru

HOW TO CITE THIS ARTICLE?

ivanisenko n.V., l avrik i.n., ivanisenko V.A. c omputer simulation of the spatial structure of MUc 1 peptides capable of inhibiting apoptosis. Vavilovskii Zhurnal Genetiki i Selektsii = **Vavilov Journal of Genetics and Breeding.** 2015;19(6):731-737. Do i 10.18699/VJ15.101

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?

Иванисенко Н.В., лаврик И.Н, Иванисенко В.А. Компьютерное моделирование пространственных структур пептидов из MUc 1, способных ингибировать апоптоз. Вавиловский журнал генетики и селекции. 2015;19(6):731-737. Doi 10.18699/VJ15.101

Апоптоз – регулируемый процесс программируемой клеточной гибели – вызывается внешними либо внутренними сигналами, которые активируют каскад цистеиновых протеаз, называемых каспазами. Пути передачи внешних сигналов апоптоза осуществляются с участием так называемых рецепторов клеточной смерти, включая рецептор фактора некроза опухоли 1-го типа (TNF-R1), а также рецепторы FAS (CD95), TRAIL-R1 и TRAIL-R2. Взаимодействие CD95 с лигандом CD95L, а также между TRAIL-R1/2 и TRAIL приводит к образованию комплекса DISC (Death-Inducing Signaling Complex – комплекс, индуцирующий смерть), в формировании которого важнейшую роль играет белок FADD (Fas-Associated Death Domain, Fas-ассоциированный домен смерти). В свою очередь, связывание каспазы-8 с FADD, которое осуществляется с помощью домена DED (Death Effector Domains – эффекторный домен смерти), сопровождается разрезанием междимерной субъединицы с образованием активной формы каспазы-8 (фрагмент p18/p10) и запуском сигнала апоптоза. Таким образом, белок FADD является ключевым звеном в пути передачи сигнала, обеспечивающим активацию каспазы-8. Ингибиторы белка FADD, препятствующие выполнению его функций в составе комплекса DISC, представляют огромный интерес для исследования молекулярных механизмов передачи сигнала апоптоза. Известно, что одним из природных белков-ингибиторов FADD является белок mucin-type гликопротеин (MUC1). Человеческий белок MUC1 экспрессируется в апикальной плазматической мембране нормальных секреторных эпителиальных клеток (Kufe et al., 1984). В работе Agata с коллегами (2008) было показано, что MUC1 сверхэкспрессируется в клетках карциномы и представлен на всей поверхности этих клеток. Незрелая форма белка MUC1 подвергается посттрансляционному автопротеолизу, в результате которого образуются две субъединицы, MUC1-N и MUC1-C (Ligtenberg et al., 1992; Levitin et al., 2005; Macao et al., 2006). MUC1-C состоит из межклеточного домена длиной 58 а. о., а также трансмембранного (28 а. о.) и цитоплазматического (MUC1-CD, 75 а. о.) доменов. Цитоплазматический домен MUC1-CD является субстратом киназ c-Src (Li et al., 2001), GSK 3b (Huang et al., 2005), Cd (Ren et al., 2002) и c-Abl (Raina et al., 2006). MUC1-CD так же напрямую взаимодействует с ключевыми для клеточной пролиферации белками, включая Wnt/β-катенин (Huang et al., 2005), p53 (Wei et al., 2005), IkB киназы (IKK-β и IKK-γ) и т. д. (Wei et al., 2005). В работе (Agata et al., 2008) было показано, что MUC1-C может напрямую связываться с каспазой-8 и доменом DED белка FADD. Эти взаимодействия MUC1-CD способны блокировать связывание FADD с каспазой-8 и тем самым препятствовать активации апоптоза. Более того, авторы определили два пептида в последовательности MUC1-CD, которые ингибировали связывание каспазы-8 с FADD. Один из этих пептидов, MUC1-CD (1–20), способен образовывать комплекс с доменом p18 каспазы-8, в то время как другой пептид, MUC1-CD (46–72), способен напрямую связываться с доменом DED белка FADD.

Последовательности этих пептидов могут быть использованы для разработки низкомолекулярных соединений ингибиторов FADD, обладающих антиапоптотическими свойствами. Однако на сегодняшний день пространственная структура MUC1-CD остается неизвестной, что существенно усложняет рациональный дизайн препаратов на основе данных пептидов. В связи с этим целью настоящей работы были компьютерное моделирование пространственных структур пептидов MUC1-CD (1–20 и 46–72) и анализ их конформационных свойств. Основное внимание было направлено на исследование конформационных свойств пептида MUC1-CD (46–72), который способен связываться с белком FADD *in vitro*.

Согласно полученным результатам большая часть последовательности пептидов принимала неупорядоченную третичную структуру, однако N-конец пептида MUC1-CD (1–20) имел тенденцию образовывать альфа-спирали. На основе анализа рассчитанных траекторий молекулярной динамики (МД) было показано, что пептид MUC1-CD (46–72) обладает конформацией, сходной с таковой ряда участков домена DED2 белка каспазы-8, имеющих сходство аминокислотных остатков с рассматриваемым пептидом. Всего было найдено не менее четырех участков белка каспазы 8, пространственную структуру которых может принимать пептид MUC1-CD (46–72). Полученные результаты подтверждают, что молекулярный механизм ингибирующей активности данного пептида состоит во взаимодействии с белком FADD в тех участках, которые связываются с доменом DED2 каспазы-8 и, таким обра-

зом, пептид может конкурировать с каспазой-8 за связывание с белком FADD.

Материалы и методы

Молекулярная динамика

Моделирование молекулярной динамики (МД) с неявным представлением воды проводили с помощью модуля pmemd.cuda пакета программ Amber 14 (Case et al., 2015) на графических картах NVIDIA Tesla M 2090 в комбинации с моделью неявной воды GB-Neck2 (Nguyen et al., 2013) с использованием атомных радиусов mbondi3 и силового поля ff14SBonlysc. Начальные структуры пептидов генерировали с использованием модуля cpptraj (AmberTools 14), затем структуры минимизировали и уравновешивали в три шага: 1 000 циклов минимизации, нагрев от 0 до 300 K в течение первых 100 пс, эквилибрация в течение первых 10 нс. Ковалентные связи, включающие атомы водорода, ограничивали с использованием алгоритма SHAKE с точностью 0,00001. Температуру контролировали с использованием термостата Лангевина с частотой столкновений $\gamma = 1{,}0$ пс$^{-1}$. Конечное моделирование структур проводили при температуре 300 K в течение 1 мкс с шагом по времени 2 фс.

а нализ траектории молекулярной динамики

Кластеризацию конформаций пептида проводили с использованием алгоритма ближайшего соседа, реализованного в инструменте MaxCluster (Herbert, Sternberg, 2014). Всего кластеризацию проводили по 3 тыс. конформаций пептида с использованием координат Cα атомов пептидов.

Структурное выравнивание пептидов проводили с помощью инструмента MultiProt (Shatsky et al., 2004), качество выравнивания характеризовалось по значениям среднеквадратичного отклонения RMSD (Root Mean Square Deviation, среднеквадратичное отклонение) и оценочной функции, учитывающей качество выравнивания первичной структуры. Расчет RMSD проводили с использованием модуля cpptraj (AmberTools 14), расчет вторичной структуры – методом DSSP (Defined Secondary Structure of Proteins) (Kabsch et al., 1983), реализованным в модуле cpptraj (AmberTools 14).

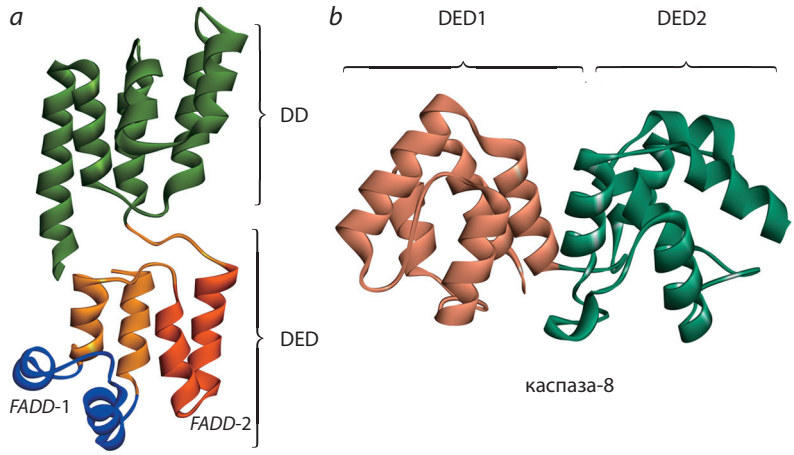


Fig. 1. Structures of FADD (pdb: 2GF5) and caspase 8 (pdb: 4ZBw) proteins.
(*a*) FADD consists of two domains: Death Domain (DD) and Death Effector Domain (DED), (*b*) caspase 8 domains: DED1 and DED2.

результаты и обсуждение

Валидация метода на примере пептидов, полученных из Fa DD

В последнее время для *in silico* предсказания третичной структуры белков и пептидов широкое распространение получили методы, использующие длинные траектории МД в неявной воде (в приближении обобщенной модели Борна). Однако, несмотря на то что метод моделирования фолдинга белков с использованием МД в неявной воде показал хорошие результаты для целого ряда белков (Nguyen et al., 2013), возможность этого подхода предсказывать конформации коротких пептидов остается не до конца изученной. Для того чтобы проверить применимость данного метода к семейству белков, содержащих домен DED, мы использовали этот подход для двух пептидов, входящих в состав домена DED белка FADD (Carrington et al., 2006). Моделирование проводилось на длительном интервале времени, равном 1 мкс. Принимая во внимание низкую вязкость растворителя при использовании моделей неявной воды (Zagrovic et al., 2003), мы предположили, что для идентификации наиболее стабильных конформаций пептида такой длины траектории будет достаточно. Идентификация всего конформационного ансамбля пептидов не являлась целью текущего исследования.

Последовательности этих пептидов, получивших название FADD-1 и FADD-2, соответствовали участкам 1–25 и 60–85 белка FADD (рис. 1). Данные участки FADD принимают вторичную структуру альфа-спираль – петля – альфа-спираль. В качестве начальной структуры этих пептидов бралась развернутая цепь. Наблюдаемые в ходе МД вторичные структуры этих пептидов после проведенного уравновешивания (100 нс) показаны на рис. 2. В ходе моделирования оба пептида принимали конформации с вторичной структурой альфа-спираль – петля – альфа-спираль, аналогичной той, которая наблюдается в составе белка. При этом для пептида FADD-2 данная вторичная структура являлась основной в траектории молекулярной динамики. С-конец пептида FADD-1 в ходе моделирования оказался наименее упорядоченным. Одной из причин может являться то, что конформация этого пептида в составе белка в значительной степени стабилизирована за счет взаимодействий между пептидом и остальной частью белка, а взаимодействия между аминокислотными остатками внутри пептида вносят меньший вклад в его стабилизацию. Стоит отметить, что в силу возможных недостатков моделирования молекулярной

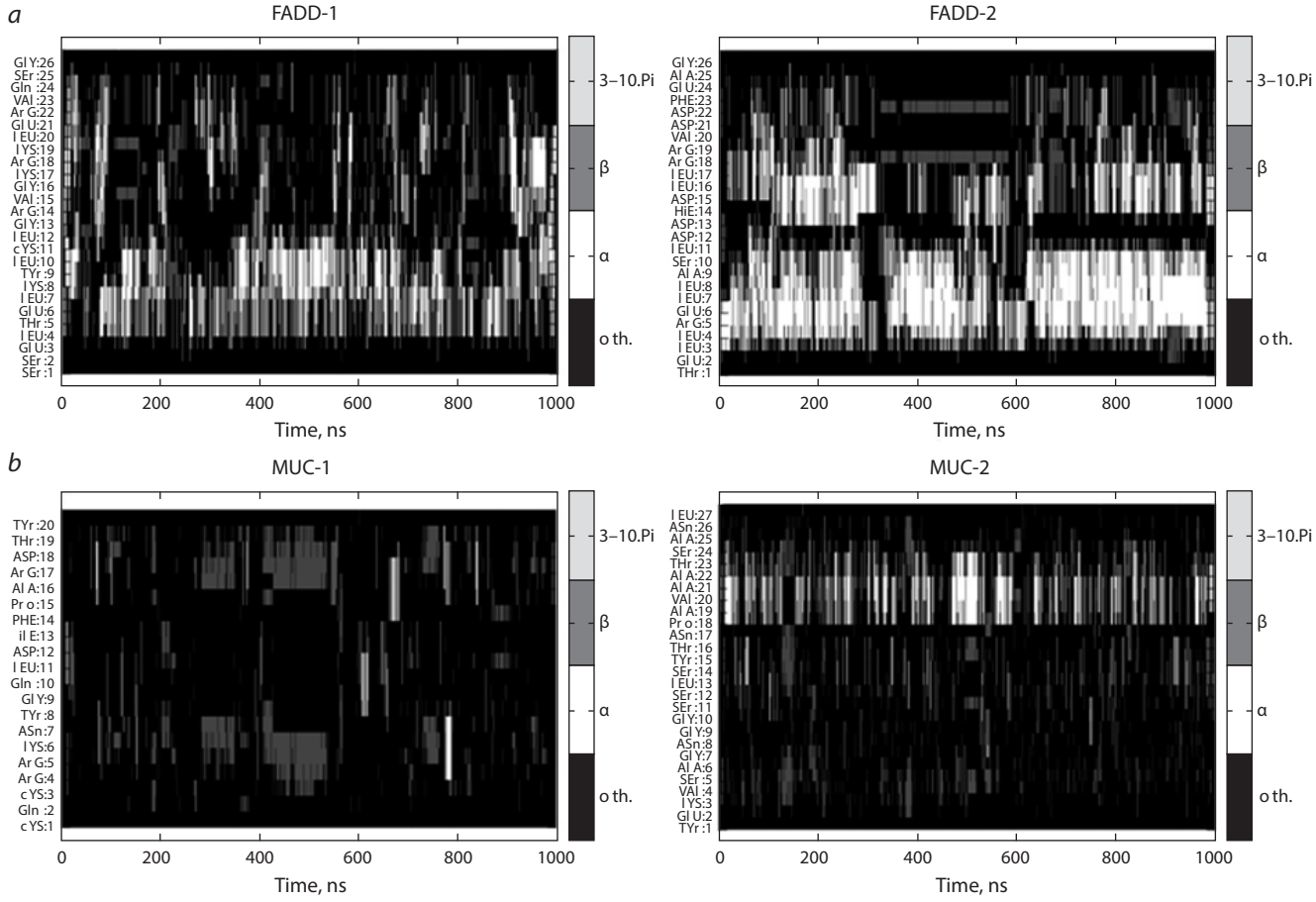


Fig. 2. Time variation of the secondary structures of peptides during molecular dynamics simulation.

(*a*) changes in the secondary structures of the FADD-1 and FADD-2 peptides. (*b*) changes in the secondary structures of the MUC-1 and MUC1-2. white bars: alpha helix; light-gray: 3–10, Pi; dark-gray: beta sheet; black: disorder or turn.

динамики в неявной воде данное утверждение требует дополнительной валидации, например, с использованием метода ЯМР (ядерный магнитный резонанс).

Для сравнения пространственных структур изолированных пептидов и соответствующих им участков полноразмерного белка был проведен расчет RMSD отклонения этих структур по координатам Cα атомов (рис. 3). При расчете RMSD не учитывали три начальных аминокислотных остатка на N- и C-концах пептидов, предполагая, что они являются неупорядоченными в растворе. Было обнаружено, что пептид FADD-1 находился в заданной конформации с RMSD < 2 Å примерно в 0,1 % точек траектории МД. При этом примерно для 2 % точек структура этого изолированного пептида имела отличие от его структуры в составе белка с RMSD < 3 Å. Для пептида FADD-2 наблюдались конформации, при которых RMSD < 2 Å в более чем 5 % точек МД и в более чем 30 % точек RMSD оставались меньше 3 Å. Таким образом, из проведенного анализа следует, что наиболее часто наблюдаемая в траектории МД третичная структура пептида FADD-2 имеет конформацию (RMSD < 3 Å), схожую с таковой данного пептида в составе полноразмерного белка. Это позволяет предположить, что пептид FADD-2 может являться структурным миметиком данного участка белка FADD. Заметим, что, хотя конформация пептида FADD-1 в составе белка не являлась доминирующей в процессе МД, метод молекулярной динамики в неявной воде позволил ее идентифицировать на относительно протяженных районах траектории МД. Принимая во внимание тот факт, что моделирование структуры изолированных пептидов проводилось *de novo* из развернутой цепи, можно заключить, что предлагаемый подход может быть эффективно использован для решения задач предсказания конформаций пептидов. В частности, мы предполагаем, что он может быть применен для анализа структур пептидов, полученных из последовательности MUC1-CD.

Анализ конформаций пептидов MUC1-1 и MUC1-2

Согласно работе (Agata et al., 2008), пептид MUC1-CD из района последовательности 1–20 (MUC1-1, CQCRR KNYGQLDIFPARDTY) связывается с участком p18 белка каспазы-8, в то время как пептид из участка 46–72 (MUC1-2, YEKVSAGNGGSSLSYTNPAVAATSANL) способен связываться с доменом DED белка FADD. Одна из гипотез связывания MUC1-2 с доменом DED белка FADD может заключаться в том, что пептид способен принимать конформацию участка связывания белка каспазы-8 и, таким образом, ингибировать взаимодействие

между доменами DED белков FADD и каспазы-8.

Отметим, что MUC1-CD представляет домен белка MUC1, имеющий неупорядоченную третичную структуру (Raina et al., 2015). Можно ожидать, что пептиды MUC1-1 и MUC1-2 также не имеют единой третичной структуры и могут принимать множество конформаций. В связи с этим для изучения конформационных особенностей таких пептидов наиболее адекватными могут оказаться методы *de novo* предсказания структур белков из развернутой цепи, включая моделирование молекулярной динамики в неявной воде (Nguyen et al., 2013).

Как видно из рис. 2, в ходе МД пептид MUC1-1 большую часть времени находился в неупорядоченном состоянии либо в состоянии с наличием бета-складок. В то же время С-конец пептида MUC1-2 часто оставался в состоянии альфа-спирали. Таким образом, основываясь на данных моделирования МД в неявной воде, можно предположить, что для связывания с участком p18 белка каспазы-8 не требуется образования альфа-спиральных конформаций пептидов.

Проведенный кластерный анализ конформаций пептида MUC1-2 позволил выявить ряд конформаций, наблюдаемых в ходе молекулярной динамики. На рис. 4 показаны центры наиболее крупных, различных по структуре кластеров конформеров. Отметим, что кластеры с фиксированной упаковкой пептида наблюдались лишь в коротком промежутке времени моделирования. Среди них можно выделить три конформации со следующими типами вторичной структуры: антипараллельные бета-складки, альфа-спираль – петля – альфа-спираль – петля, а также альфа-спираль – бета-складка (рис. 4). Наиболее часто наблюдаемая в ходе молекулярной динамики конформация пептида представляет собой стабильную структуру альфа-спирали на С-конце и неупорядоченный участок на N-конце. Принимая это во внимание, мы рассматривали пространственные структуры пептида с данным типом упаковки как основного кандидата на связывание с белком FADD.

Следующим шагом нашего анализа было структурное выравнивание всех

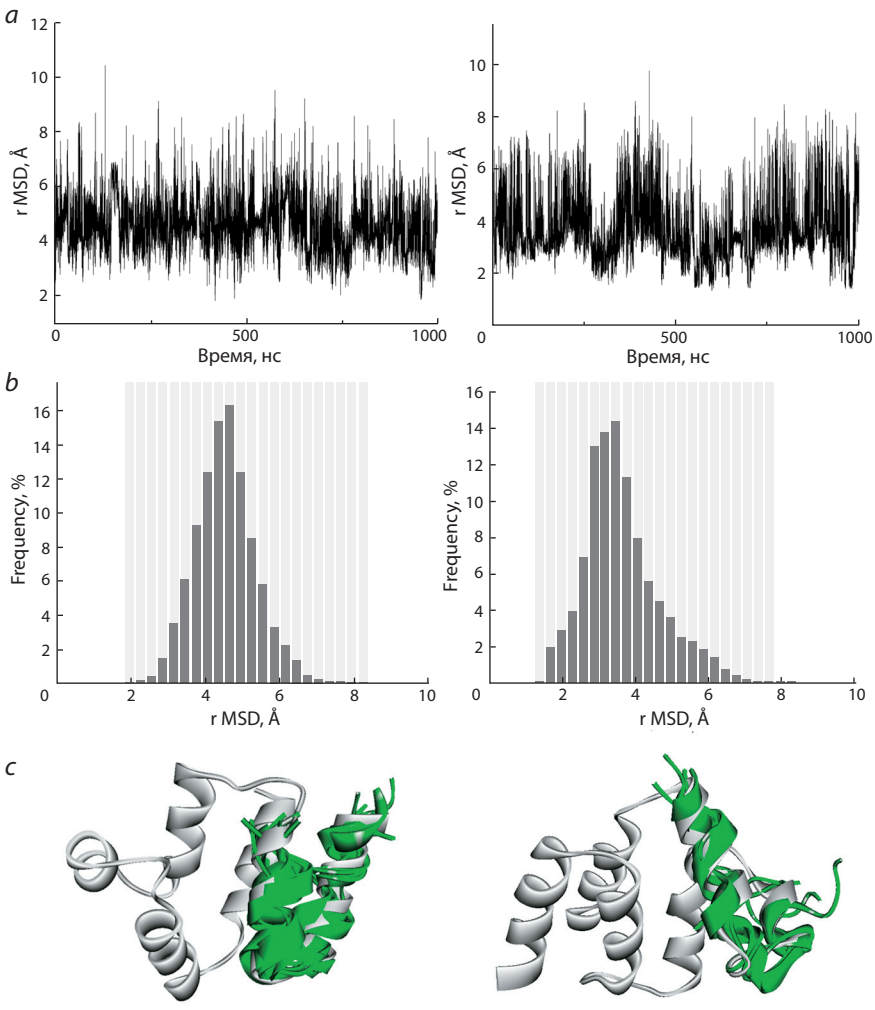


Fig. 3. identific tion of FADD-1 and FADD-2 conformations corresponding to those observed in the full-length FADD protein: (*a*), root-mean square deviation of c α atoms of a peptide from the conformation in full-length FADD; (*b*), the frequency of occurrence of conformations with a given r MSD throughout 1-μs molecular dynamics trajectory; (*c*), superimposition of peptide conformations with r MSD < 2 Å on the DED domain of FADD. The first and last th ee amino acid residues of peptide were omitted from the calculation of r MSD, as they were assumed to be disordered.

конформаций пептида MUC1-2 с доменом DED2 каспазы-8 (pdb: 4ZBW) (Shen et al., 2015) для выяснения того, какие из них обладают большей способностью имитировать конформацию белка каспазы-8. Поскольку неупорядоченный домен пептида MUC1-2 мог принимать большой набор конформаций, то при сравнении, помимо рассмотрения отклонений между структурами пептида и участками белка по координатам Cα атомов, также учитывали и сходство аминокислотных последовательностей. В результате структурного выравнивания были отобраны конформации пептида, удовлетворяющие следующим условиям: пространственные структуры пептида и какого-либо участка белка имеют отличия RMSD < 2 Å, а в выравнивании их последовательностей имеются не менее трех идентичных остатков. Таким образом, было отобрано около 1 % всех конформаций. Неожиданным для нас оказалось, что практически все конформации пептида, удовлетворяющие заданным условиям, принадлежат кластеру 1, содержащему наиболее стабильные из них. Примеры суперпозиции пространственных структур пептидов MUC1-2 из кластера 1 и домена DED2 каспазы-8 показаны на рис. 5. Следует заметить, что большинство пептидов имели структурное сходство с участками домена DED2 каспазы-8, что согла-

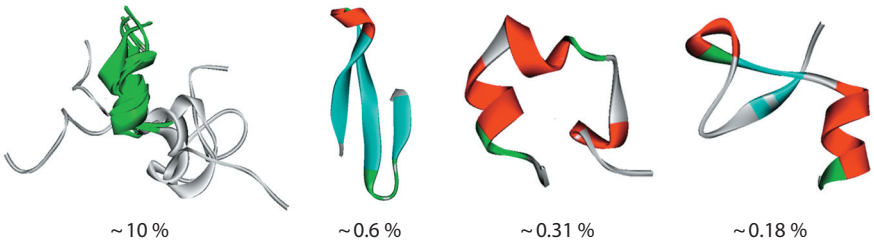


Fig. 4. clustering of the MUC1-2 peptide conformations obtained in the course of MD. conformations representing centers of the four largest clusters and their occurrence frequencies are shown.

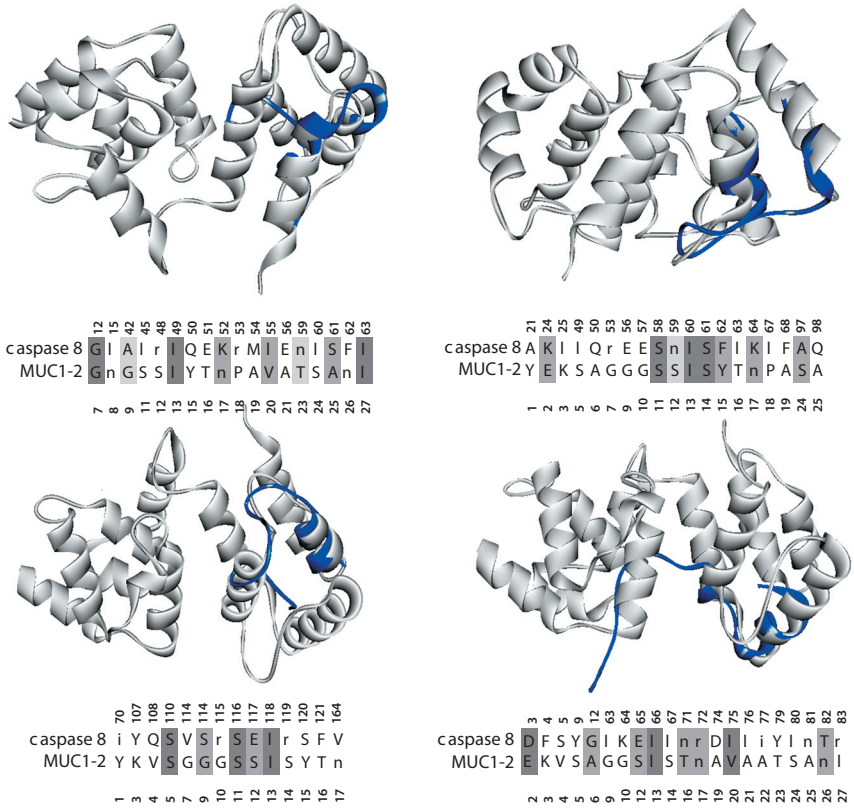


Fig. 5. identified onformations of the MUC1-2 peptide able to mimic conformations of portions of the DED2 domain of caspase 8.

MUC1-2 is shown in blue. The alignments of sequences used in spatial superimposition by MultiProt program and positions corresponding to the alignments are shown.

суется с работами (Yang et al., 2005; Dickens et al., 2012; Schleich et al., 2012), в которых каспаза-8 связывается с FADD через домен DED2.

Предсказание третичной структуры белков и пептидов является чрезвычайно важной и актуальной задачей, решение которой необходимо при изучении молекулярных взаимодействий в клетке и тканях, а также критическим шагом при рациональном конструировании лекарственных препаратов. В последнее время методы МД заняли прочные позиции среди других методов, используемых для решения данных задач. В частности, широкое развитие получили методы *in silico* предсказания третичной структуры белков и пептидов с использованием методов МД в неявной воде (в приближении обобщенной модели Борна). Эти методы позволяют проводить моделирование МД в микросекундном диапазоне и решать задачу фолдинга белков небольших размеров. В работе с помощью данного метода были проведены моделирование и анализ конформаций пептидов, соответствующих фрагментам последовательности MUC1, которые, согласно экспериментальным данным, способны ингибировать взаимодействие FADD с каспазой-8 (Agata et al., 2008).

На первом этапе анализа была проведена оценка применимости данного подхода для предсказания функционально важных конформаций пептидов, участвующих во взаимодействии с DED, на примере белка FADD. Было показано, что моделирование структуры пептидов с использованием метода МД в неявной воде, последовательности которых были взяты из различных фрагментов последовательности белка FADD, позволило выявить стабильные конформации этих пептидов, имеющих сходство с конформациями соответствующих участков полноразмерного белка. В частности, для пептида FADD-2 данная конформация превалировала в ходе моделирования МД.

Анализ траектории МД пептида MUC1-1, который, согласно экспериментальным данным, связывается с субъединицей p18 каспазы-8, показал, что N- и C-концы пептида имеют предрасположенность образовывать бета-складки с неупорядоченным линкером между ними. Для пептида MUC1-2, способного, согласно данным тех же авторов, связываться с доменом DED белка FADD, характерно образование альфа-спирали на C-конце с неупорядоченным N-концом. С применением подхода структурного выравнивания всех конформаций пептида MUC1-2 на структуру доменов DED2 каспазы-8 было сделано предсказание конформаций, потенциально участвующих во взаимодействии с доменом DED белка FADD. Таким образом, полученные результаты являются отправной точкой для планирования сайт-специфических экспериментов по мутагенезу и выявлению структуры функционально важных конформаций пептидов, что позволит в будущем провести рациональный дизайн более эффективных пептидов и низкомолекулярных пептидов-ми-

метиков. Для такой экспериментальной работы особый интерес могут представлять позиции MUC1-2, содержащие сходные аминокислотные остатки каспазы-8, такие как S5, G7, L13, L27, S11, L13, S14 (рис. 5).

Acknowledgments

The modeling of the spatial structure and analysis of apoptosis-inhibiting peptides were supported by the Russian Science Foundation, project 14-44-00011 “Programmed cell death induced via death receptors: identification of molecular mechanisms initiating apoptosis by molecular simulation”. Implicit validation of the molecular dynamics method was supported by project VI.61.1.2. Supercomputation was done at the Bioinformatics Shared Access Center.

Conflict of interest

The authors declare no conflict of interest.

References

Agata N., Ahmad R., Kawano T., Raina D., Kharbanda S., Kufe D. MUC1 oncoprotein blocks death receptor-mediated apoptosis by inhibiting recruitment of caspase-8. Cancer Res. 2008;68(15):6136-6144. DOI 10.1158/0008-5472.CAN-08-0464

Carrington P.E., Sandu C., Wei Y., Hill J.M., Morisawa G., Huang T., Gavathiotis E., Wei Y., Werner M.H. The structure of FADD and its mode of interaction with procaspase-8. Mol. Cell. 2006;22(5):599-610. DOI 10.1016/j.molcel.2006.04.018

Case D.A., Berryman J.T., Betz R.M., Cerutti D.S., Cheatham T.E., Darden T.A., Duke R.E., Giese T.J., Gohlke H., Goetz A.W., Homeyer N., Izadi S., Janowski P., Kaus J., Kovalenko A., Lee T.S., LeGrand S., Li P., Luchko T., Luo R., Madej B., Merz K.M., Monard G., Needham P., Nguyen H., Nguyen H.T., Omelyan I., Onufriev A., Roe D.R., Roitberg A., Salomon-Ferrer R., Simmerling C.L., Smith W., Swails J., Walker R.C., Wang J., Wolf R.M., Wu X., York D.M., Kollman P.A. AMBER 2015. University of California, San Francisco, 2015.

Dickens L.S., Boyd R.S., Jukes-Jones R., Hughes M.A., Robinson G.L., Fairall L., Schwabe J.W.R., Cain K., MacFarlane M. A death effector domain chain DISC model reveals a crucial role for caspase-8 chain assembly in mediating apoptotic cell death. Mol. Cell. 2012; 47(2):291-305. DOI 10.1016/j.molcel.2012.05.004

Herbert A., Sternberg M. J. E. MaxCluster – A tool for Protein Structure Comparison and Clustering. 2014. URL: http://www.sbg.bio.ic.ac.uk/~maxcluster/

Huang L., Chen D., Liu D., Yin L., Kharbanda S., Kufe D. MUC1 oncoprotein blocks glycogen synthase kinase 3β-mediated phosphorylation and degradation of β-catenin. Cancer Res. 2005;65(22):10413-10422. DOI 10.1158/0008-5472.CAN-05-2474

Kabsch W., Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577-2637.

Kufe D., Inghirami G., Abe M., Hayes D., Justi-Wheeker H., Schlom J. Differential reactivity of a novel monoclonal antibody (DF3) with human malignant versus benign breast tumors. Hybridoma. 1984; 3:223-32. DOI 10.1089/hyb.1984.3.223.

Levitin F., Stern O., Weiss M., Gil-Henn C., Ziv R., Prokocimer Z., Smorodinsky N.I., Rubinstein D.B., Wreschner D.H. The MUC1 SEA module is a self-cleaving domain. J. Biol. Chem. 2005;280(39): 33374-33386. DOI 10.1074/jbc.M506047200

Li Y., Kuwahara H., Ren J., Wen G., Kufe D. The c-Src tyrosine kinase regulates signaling of the human DF3/MUC1 carcinoma-associated antigen with GSK3β and β-catenin. J. Biol. Chem. 2001;276(9): 6061-6064. DOI 10.1074/jbc.C000754200

Ligtenberg M.J., Kruijshaar L., Buijs F., Van Meijer M., Litvinov S.V., Hilkens J. Cell-associated episialin is a complex containing two proteins derived from a common precursor. J. Biol. Chem. 1992;267(9): 6171-6177.

Macao B., Johansson D.G., Hansson G.C., Härd T. Autoproteolysis coupled to protein folding in the SEA domain of the membrane-bound MUC1 mucin. Nat. Struct. Mol. Biol. 2006;13(1):71-76. DOI 10.1038/nsmb1035

Nguyen H., Roe D.R., Simmerling C. Improved generalized born solvent model parameters for protein simulations. J. Chem. Theory Comput. 2013;9(4):2020-2034. DOI 10.1021/ct3010485

Raina D., Agarwal P., Lee J., Bharti A., McKnight C.J., Sharma P., Kharbanda S., Kufe D. Characterization of the MUC1-C cytoplasmic domain as a cancer target. PLOS One. 2015;10(8):e0135156. DOI 10.1371/journal.pone.0135156

Raina D., Ahmad R., Kumar S., Ren J., Yoshida K., Kharbanda S., Kufe D. MUC1 oncoprotein blocks nuclear targeting of c-Abl in the apoptotic response to DNA damage. EMBO J. 2006;25(16):3774-3783. DOI 10.1038/sj.emboj.7601263

Ren J., Li Y., Kufe D. Protein kinase C δ regulates function of the DF3/MUC1 carcinoma antigen in β-catenin signaling. J. Biol. Chem. 2002;277(20):17616-17622. DOI 10.1074/jbc.M200436200

Schleich K., Warnken U., Fricker N., Öztürk S., Richter P., Kammerer K., Schnölzer M., Karmmer P.H., Lavrik I.N. Stoichiometry of the CD95 death-inducing signaling complex: experimental and modeling evidence for a death effector domain chain model. Mol. Cell. 2012;47(2):306-319. DOI 10.1016/j.molcel.2012.05.006

Shatsky M., Nussinov R., Wolfson H.J. A method for simultaneous alignment of multiple protein structures. Proteins- Structure, Function, and Bioinformatics. 2004;56(1):143-156. DOI 10.1002/prot.10628

Shen C., Yue H., Pei J., Guo X., Wang T., Quan J.M. Crystal structure of the death effector domains of caspase-8. Biochem. Bioph. Res. Co. 2015;463(3):297-302. DOI 10.1016/j.bbrc.2015.05.054

Wei X., Xu H., Kufe D. Human MUC1 oncoprotein regulates p53-responsive gene transcription in the genotoxic stress response. Cancer Cell. 2005;7(2):167-178. DOI 10.1016/j.ccr.2005.01.008

Yang J.K., Wang L., Zheng L., Wan F., Ahmed M., Lenardo M.J., Wu H. Crystal structure of MC159 reveals molecular mechanism of DISC assembly and FLIP inhibition. Mol. Cell. 2005;20(6):939-949. DOI 10.1016/j.ccr.2005.01.008

Zagrovic B., Pande V. Solvent viscosity dependence of the folding rate of a small protein: distributed computing study. J. Comput. Chem. 2003;24(12):1432-1436. DOI 10.1002/jcc.10297

Вавиловский журнал генетики и селекции. 2015;19(6):738-744
Doi 10.18699/VJ15.097

Идентифицируемость математических моделей медицинской биологии

С.И. Кабанихин[1, 2], Д.А. Воронов[1, 2], А.А. Гродзь[2], О.И. Криворотько[1, 2]

1 Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия
2 Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия

Анализ биологических данных является важнейшим вопро-сом в биоинформатике, вычислительной геномике, молеку-лярном моделировании и системной биологии. Рассматриваемые в статье подходы позволяют сократить затраты на проведение экспериментов по получению биологических данных. В статье рассмотрен вопрос идентифицируемости математических моделей физиологии, фармакокинетики и эпидемиологии. Рассматриваемые процессы моделируются с помощью нелинейных систем обыкновенных дифференциальных уравнений. Математическое моделирование дина-мических процессов основано на использовании закона сохранения масс. В процессе решения задачи по оценке параметров, характеризующих исследуемый процесс, нередко возникает вопрос неединственности решения. В случае, когда известны результаты эксперимента (данные на выходе) и данные на входе, целесообразно проводить априорный анализ информативности этих данных. В статье рассмотрено понятие идентифицируемости математических моделей. Представлен обзор методов анализа идентифицируемости динамических систем. В работе приведен обзор следующих подходов: метод передаточной функции, применяемый для линейных моделей (удобен для анализа фарма-кокинетических данных, так как большой класс препаратов характеризуется линейной кинетикой); метод разложения в ряды Тейлора, применяемый для нелинейных моделей; метод, основанный на теории дифференциальной алгебры (структура данного алгоритма допускает его реализацию на ЭВМ); метод, основанный на теории графов (данный метод не только определяет идентифицируемость модели, но и позволяет найти замену переменных специального вида, приводящую исходную модель к идентифицируемой). На конкретных примерах продемонстрирована необходимость проводить априорный анализ идентифицируемости модели перед проведением численных расчетов по определению параметров, характеризующих тот или иной процесс. Рассмотрены примеры анализа идентифицируемости математических моделей медицинской биологии.

Ключевые слова: идентифицируемость; математические модели медицинской биологии; система обыкновенных дифференциальных уравнений; фармакокинетика; эпидемиология; физиология; дифференциальная алгебра.

HOW TO CITE THIS ARTICLE?
Kabanikhin S.i., Voronov D.A., Grodz A.A., Krivorotko o.i. identifiability of mathematical models in medical biology. Vavilovskii Zhurnal Genetiki i Selektsii = **Vavilov Journal of Genetics and Breeding.** 2015;19(6):738-744. Doi 10.18699/VJ15.097

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?
Кабанихин С.И., Воронов Д.А., Гродзь А.А., Криворотько О.И. Идентифицируемость математических моделей медицинской биологии. Вавиловский журнал генетики и селекции. 2015;19(6):738-744. Doi 10.18699/VJ15.097

Received 20.09.2015
Accepted for publication 29.10.2015

e-mail: dmitriy.voronov.89@gmail.com

Identifiability of mathematical models in medical biology

S.I. Kabanikhin[1, 2], D.A. Voronov[1, 2], A.A. Grodz[2], O.I. Krivorotko[1, 2]

1 Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia
2 Novosibirsk State University, Novosibirsk, Russia

Analysis of biological data is a key topic in bioinformatics, computational genomics, molecular modeling and systems biology. The methods covered in this article could reduce the cost of experiments for biological data. The problem of identifiabili y of mathematical models in physiology, pharmacokinetics and epidemiology is considered. The processes considered are modeled using nonlinear systems of ordinary differential equations. Math modeling of dynamic processes is based on the use of the mass conservation law. w hile addressing the problem of estimation of the parameters characterizing the process under the study, the question of nonuniqueness arises. w hen the input and output data are known, it is useful to perform an a priori analysis of the relevance of these data. The definition of ide tifiabili y of mathematical models is considered. Methods for analysis of identifiabili y of dynamic models are reviewed. in this review article, the following approaches are considered: the transfer function method applied to linear models (useful for analysis of pharmacokinetic data, since a large class of drugs is characterized by linear kinetics); the Taylor series expansion method applied to nonlinear models; a method based on differential algebra theory (the structure of this algorithm allows this to be run on a computer); a method based on graph theory (this method allows for analysis of the identifiabili y of the model as well as finding a p oper reparametrization reducing the initial model to an identifiable one). The need to perform a priory identifiabili y analysis before estimating parameters characterizing any process is demonstrated with several examples. The examples of identifiabili y analysis of mathematical models in medical biology are presented.

Key words: identifiabili y; mathematical models in medical biology; system of ordinary differential equations; pharmacokinetic; epidemiology; physiology; differential algebra.

Динамические процессы фармакокинетики, эпидемиологии и физиологии моделируются (Bellman, Astrom, 1970; Brown, 1980; Walter et al., 1984; Tunali, Tarn, 1987; Bellu et al., 2007) с помощью систем обыкновенных дифференциальных уравнений (СОДУ) вида:

$$\begin{cases} \dot{x} = f(x(t),p) + \sum_{i=1}^{n} h(x(t),p)u_i(t) \\ y(t) = g(u(t),x(t),p) \end{cases}, \quad (1)$$

где $x(t)$ – n-мерная функция состояния (в фармакокинетике – доза препарата, в эпидемиологии – количество больных разного типа), $y(t)$ – k-мерная функция экспериментальных данных (в фармакокинетике – концентрация препарата в крови и/или моче, в эпидемиологии – количество больных по годам), $u(t)$ – функция входных данных (в фармакокинетике – способ введения препарата в организм) (Goodwin, Payne, 1977), g – функция, связывающая модель с измерениями, h – функция, определяющая структуру входных данных, $f(x(t), p)$ – функция, определяющая структуру модели, p – s-мерный вектор параметров, характеризирующий рассматриваемый процесс (в фармакокинетике – скорость перехода препарата между органами, в эпидемиологии – смертность, приток индивидов, скорость развития болезни и др.).

При изучении динамических процессов (1) возникает вопрос неединственности набора параметров $p = [p_1, \ldots, p_s]$, удовлетворяющего имеющимся экспериментальным данным. С другой стороны, зачастую по измеренным данным невозможно определить набор параметров p. Нередко в физиологических и клинических исследованиях проведение экспериментов по получению данных $y(t)$ является финансово затратной процедурой. Кроме того, получение экспериментальных данных – не всегда приятная процедура для пациента, и избежание «неинформативных» измерений – важный момент и с этической стороны.

На практике у исследователя есть два типа информации: входные данные – функция $u(t)$ и данные на выходе – $y(t)$. Априорный анализ идентифицируемости математической модели отвечает на вопрос: можно ли по информации на входе и на выходе утверждать, что искомый набор параметров определяется единственным образом? Если ответ на данный вопрос положительный, то можно проводить серию измерений для получения экспериментальных данных $y(t)$, не опасаясь иметь, с одной стороны, бесполезные, а с другой – затратные данные.

Определение идентифицируемости (Brown, 1982; Carson, Cobelli, 2001, 2008; Bellu et al., 2007) формулируется следующим образом:

Определение 1. Модель (1) называется априорно идентифицируемой, если ее параметры $p = [p_1, \ldots, p_n]$ можно однозначно определить по входным данным $u(t)$ и данным измерений $y(t) = g(t,p)$.

Информация, содержащаяся в измерениях, является функцией от времени и вектора параметров $p = [p_1, \ldots, p_n]$ вида:

$$y(t) = g(t, p), \quad (2)$$

где $g(t, p)$ – функция, связывающая модель с измерениями.

На практике *точные* экспериментальные данные $y(t)$ недоступны из-за ошибок измерений (например, настроек прибора по измерению данных). Обычно в наборе дискретных точек по времени $[t_1, \ldots, t_N]$ экспериментальные данные имеют вид:

$$z(t_i) = y_i + v_i = g(t_i,p) + v_i, \quad (3)$$

где v_i – погрешность (ошибка) i-го измерения, которую можно рассматривать как случайную величину, имеющую нулевое математическое ожидание (Carson, Cobelli, 2008). Зачастую некоторая статистическая информация о погрешности известна заранее.

На рис. 1 приведен пример фармакокинетических данных: крестиками отмечены значения измеренных концентраций препарата, внутривенно введенного в момент времени $t = 0$. Данные $z_1, \ldots, z_n$ могут быть представлены в виде (3) с помощью функции $g(t,p) = A_1 e^{-\lambda_1 t} + A_2 e^{-\lambda_2 t}$ следующим образом:

$$z(t_i) = y_i + v_i = g(t_i,p) + v_i = A_1 e^{-\lambda_1 t_i} + A_2 e^{-\lambda_2 t_i} + v_i.$$

Известно (Carson, Cobelli, 2008), что если для точных данных $y(t)$ модель является неидентифицируемой, то набор $p = [p_1, \ldots, p_n]$ невозможно определить единственным образом. А при использовании «зашумленных» данных вида (3) возможно получение данных, не имеющих никакого смысла.

В первом разделе статьи рассмотрены подходы к анализу идентифицируемости. Так как в определении идентифицируемости фигурируют функции $y(t)$ и $u(t)$, и отсутствует $x(t)$, то большинство методов основано на «исключении» $x(t)$ и получении соотношений между $y(t)$ и $u(t)$. Во втором разделе рассмотрены примеры анализа идентифицируемости математических моделей.

1. Обзор методов

1.1. Метод передаточной функции

Для динамических процессов, которые могут быть представлены с помощью линейной системы обыкновенных дифференциальных уравнений вида (4), чаще всего применим метод передаточной функции.

$$\begin{cases} \dot{x}(t) = A(p)x(t) + B(p)u(t) \\ y(t,p) = g(x(t,p);p) \end{cases}, \quad (4)$$

где $A(p)$ и $B(p)$ – постоянные матрицы. Функция измерений чаще всего имеет вид: $y(t,p) = C(p)x(t,p)$, где $C(p)$ – постоянная матрица.

Метод передаточной функции основан на использовании преобразования Лапласа и так называемой передаточной функции $H(s) = \frac{Y(s)}{U(s)}$, которая выражает связь между параметрами модели и наблюдаемыми параметрами (Cobelli, DiStefano, 1980; Audoly, D'Angio, 1983; Jacquez, Greif, 1985; Ben-Zvi et al., 2004; Carson, Cobelli, 2008). Здесь $U(s)$ и $Y(s)$ – преобразования Лапласа функций $u(t)$ и $y(t)$ соответственно. Метод заключается в анализе коэффициентов матрицы передаточной функции вида:

$$H(s,p) = [H_{ij}(s,p)] = \frac{Y_i(s,p)}{U_j(s)}, \quad (5)$$

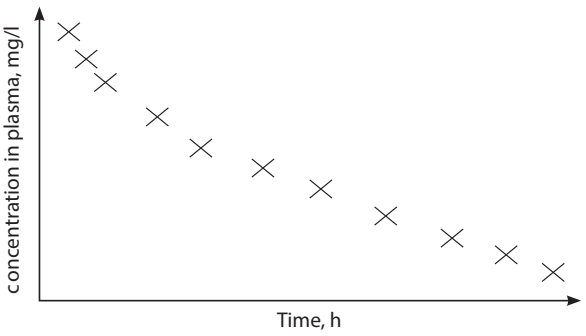


Fig. 1. Drug concentration in plasma.

где каждый элемент H_{ij} матрицы H является преобразованием Лапласа функций выхода $y_i(t,p)$ и входа $u_j(t)$. $H_{ij}(s,p)$ представляют собой рациональные полиномы с коэффициентами: $\beta_1^{ij}(p)$, …, $\beta_n^{ij}(p)$ в числителе и $\alpha_1^{ij}(p)$, …, $\alpha_n^{ij}(p)$ в знаменателе. Коэффициенты перед мономами представляются в виде комбинаций параметров $p = [p_1, \ldots, p_n]$. Приравняем эти комбинации к неопределенным параметрам φ_l^{ij}, которые назовем наблюдаемыми величинами, и составим систему алгебраических уравнений вида:

$$\begin{cases} \beta_1^{11}(p) = \varphi_1^{11} \\ \cdots \\ \alpha_n^{11}(p) = \varphi_{2n}^{11} \\ \cdots \\ \beta_1^{rm}(p) = \varphi_1^{rm} \\ \cdots \\ \alpha_n^{rm}(p) = \varphi_{2n}^{rm} \end{cases}. \qquad (6)$$

Если система (6) разрешима единственным образом относительно вектора параметров p, значит исходная математическая модель (4) является идентифицируемой. Таким образом, анализ идентифицируемости СОДУ сводится к вопросу разрешимости системы алгебраических уравнений вида (6). Во втором разделе рассмотрен фармакокинетический пример такого анализа.

1.2. Метод, основанный на разложении в ряд тейлора

При анализе идентифицируемости можно использовать разложение функции выходных данных (2) в ряд Тейлора. Данный метод в отличие от метода передаточной функции применим и к системам нелинейных дифференциальных уравнений. Подробное описание данного метода можно посмотреть в работе (Carson, Cobelli, 2008).

Разложение в ряд Тейлора функции $y_i(t,p)$ в точке $t_0 = 0$ выглядит следующим образом: $y_i(t) = y_i(t_0) + t\dot{y}_i(t_0) + \frac{t^2}{2!}\ddot{y}_i(t_0) + \frac{t^3}{3!}\dddot{y}_i(t_0)$. Обозначим:

$$y_i^{(k)}(t_0,p) = \varphi_k^i;\ k = 0, 1, 2 \ldots \qquad (7)$$

Данный метод основан на замене в равенстве (7) левых частей на выражения, получаемые из исходной СОДУ (1). После такой замены получается система нелинейных алгебраических уравнений. Как и в методе передаточной функции, вопрос идентифицируемости математической модели сводится к анализу разрешимости алгебраической системы. В разделе 2 рассмотрен пример такого анализа.

1.3. Метод, основанный на теории дифференциальной алгебры

Перепишем систему (1) в виде:

$$\begin{cases} \dot{x} - f(x(t),p) - \sum_{i=1}^{n} h(x(t),p)u_i(t) = 0 \\ y(t) - g(u(t),x(t),p) = 0 \end{cases}. \qquad (8)$$

Рассмотрим ее как систему алгебраических уравнений в переменных состояния $x(t)$, измерения $y(t)$, входных данных $u(t)$, а также соответствующих производных этих функций. Ниже приведена краткая схема анализа идентифицируемости.

1. Введем формальное упорядочение: будем считать, что $u < \dot{u} < \ddot{u} < \ldots < y < \dot{y} < \ddot{y} < \ldots < x < \dot{x} < \ddot{x} < \ldots$. Если дифференциальные полиномы являются рациональными функциями, то необходимо привести их к общему знаменателю.

2. Проводится упорядочение всех уравнений системы (8) в соответствии со старшим слагаемым в каждом дифференциальном полиноме.

3. Проводится специальная процедура псевдоделения, исключающая из рассматриваемых уравнений функции $x, \dot{x}, \ddot{x}, \ldots$. Подробно данная процедура изложена в работах (Carson, Cobelli, 2001; Bellu et al., 2007; Meshkat et al., 2012). В результате такого алгоритма остается столько дифференциальных многочленов относительно $y(t)$, $u(t)$, их производных, а также всевозможных комбинаций, сколько в исходной системе (8) выходных данных $y_i(t)$.

4. Составляем нелинейную систему алгебраических уравнений, приравнивая коэффициенты при каждом слагаемом дифференциальных полиномов к наблюдаемым параметрам (аналогично предыдущим методам). Анализ разрешимости системы проводится с использованием теории базисов Гребнера.

5. Вопрос идентифицируемости системы (8) сводится к вопросу разрешимости нелинейной алгебраической системы.

Данный метод привлекателен тем, что структура алгоритма позволяет автоматизировать его и проводить анализ идентифицируемости нелинейных динамических систем с использованием ЭВМ (Bellu et al., 2007; Meshkat et al., 2014). Подробное описание данного подхода можно посмотреть, например, в работах (Carson, Cobelli, 2001; Bellu et al., 2007; Meshkat et al., 2014).

1.4. Метод, основанный на теории графов

Иногда неидентифицируемую модель с помощью специальной замены можно привести к идентифицируемой. Ниже приведен алгоритм поиска так называемой рациональной репараметризации для линейных динамических систем (4). Замена такого типа крайне полезна в фармакокинетических приложениях (Meshkat, 2014). Этот метод позволяет найти такую рациональную замену, которая, во-первых, позволяет сохранить вид функции $x_1(t,p)$, так как чаще измерения доступны в виде $y_1(t) = x_1(t,p)$. Во-вторых, после такой замены переменных в качестве

искомых параметров выступают комбинации исходных параметров, обладающие физическим смыслом.

Метод заключается в сопоставлении линейной СОДУ графа G (Cobelli et al., 1976). По данному графу G с n вершинами и $m \le 2n-2$ дугами мы либо указываем конкретную замену, которая сведет нашу модель к идентифицируемой, либо говорим, что такой замены не существует. Пусть пространство параметров $P \subseteq \mathbb{R}^{2n-1}$ состоит из таких матриц, структура которых индуцируется графом G. Отображение $c: P \to \mathbb{R}^{2n-1}$ сопоставляет матрице $A \subseteq P$ вектор $(c_1(A), \ldots, c_n(A), d_1(A), \ldots, d_{n-1}(A))$, составленный из коэффициентов при характеристических многочленах. Отображение $c: P \to \mathbb{R}^{2n-1}$ называется двойным характеристическим полиномиальным отображением. На первом этапе алгоритма мы вычисляем $d = \dim(Imc)$ – размерность образа так называемого двойного характеристического полиномиального отображения. Если $d \neq m+1$, то не существует замены переменных в необходимом для нас виде. Иначе:

а) ищем остовное дерево T из графа G с вершинами $j_1 \to i_1, \ldots, j_{n-1} \to i_{n-1}$;

б) формируем матрицу инцидентности $E(T)$ таким образом, что первые $n-1$ столбцов отвечают ребрам T. Далее вычеркиваем первую строку и получаем матрицу E_1. Другими словами, получили E_1 – матрицу размера $(n-1)\times(n-1)$, отвечающую ребрам T;

в) определяем мономиальное масштабирование: $X_i = f_i(A)x_i$. Положим $f_1(A) = 1$.

Пусть также $r_i = (r_{1,i}, \ldots, r_{n-1,i})^T$ – i-ый столбец матрицы $C_1 = E_1^{-1}$. Тогда $f_{i+1}(A) = a_{i_1 j_1}^{r_{1,i}} \ldots a_{i_{n-1} j_{n-1}}^{r_{n-1,i}}$;

г) заменяем элементы a_{ij} матрицы A на элементы $a_{ij} f_i(A)/f_j(A)$.

Сделаем некоторые замечания по описанному алгоритму. Во-первых, d может быть вычислено как с помощью определения ранга матрицы Якоби отображения c в общей точке, так и с помощью поиска обнуляющего идеала образа отображения c, при использовании базисов Грёбнера. Также следует отметить, что в некоторых случаях этот шаг можно пропустить: если G является индуктивно сильно связным графом, то условие $d = m+1$ автоматически выполняется и нет необходимости вычислять d трудоемкими методами.

В случае существования идентифицируемой масштабируемой замены переменных, новая матрица будет иметь $n-1$ элементов, отвечающих остовному дереву T, равных единице, и оставшиеся внедиагональные элементы (их $m-n+1$) понимаются как новые параметры в репараметризованной системе. Получаемая с помощью такой замены модель является идентифицируемой относительно рациональных комбинаций исходных параметров. Более подробно этот алгоритм изложен в работе (Meshkat, 2014).

2. Примеры анализа идентифицируемости

2.1. Анализ идентифицируемости фармакокинетической модели контроля глюкозы-инсулина методом передаточной функции

Рассмотрим модель контроля глюкозы-инсулина (9) (Carson, Cobelli, 2008). Следует отметить, что в рамках камерного подхода к моделированию фармакокинетических процессов в качестве камер могут выступать не только органы, но и различные вещества внутри камеры (например, глюкозы и инсулина в крови).

$$\begin{cases} \dfrac{dx_1(t)}{dt} = -q_1 x_1(t) - q_2 x_2(t) + u(t), \\ \dfrac{dx_2(t)}{dt} = -q_3 x_2(t) + q_4 x_1(t), \\ y_1(t) = \dfrac{x_1(t)}{V_1}, \\ x_1(0) = x_2(0) = 0, \end{cases} \tag{9}$$

где $x_1(t)$, $x_2(t)$ – массы глюкозы (назовем глюкозу – 1-я камера) и инсулина (2-я камера), $u(t) = D \cdot \delta(t)$ – доза. $p = [q_1, q_2, q_3, q_4, V_1]$ – вектор неизвестных параметров. Пусть V_2 – объем второй камеры. Предположим, что в качестве данных измерений нам известна только концентрация глюкозы $y_1(t)$. Следует отметить, что на практике начальные данные системы (9) не могут быть равными нулю. В этом примере такие значения выбраны для удобства и никак не влияют на схему анализа.

Для определения идентифицируемости данной модели воспользуемся методом передаточной функции. Запишем передаточную функцию в виде: $H_1(s) = \dfrac{Y_1(s)}{U(s)}$.

Применяя преобразования Лапласа для уравнений системы (9), получим:

$$\begin{cases} sX_1(s) - x_1(0) = -q_1 X_1(s) - q_2 X_2(s) + D \\ sX_2(s) - x_2(0) = -q_3 X_2(s) + q_4 X_1(s) \\ Y_1(s) = \dfrac{X_1(s)}{V_1} \end{cases}.$$

Прописными буквами обозначены преобразования Лапласа для соответствующих функций. Находим выражение для $Y_1(s)$ путем выражения $X_1(s)$ из первых двух уравнений в последней системе. Получим следующее выражение:

$$Y_1(s) = \frac{X_1(s)}{V_1} = \frac{s^2 + (q_1+q_3)s + q_1 q_3 + q_2 q_4}{DV_1(s+q_3)}.$$

Теперь подставляем значения $Y_1(s)$ и $U(s)$ в формулу передаточной функции:

$$H_1(s) = \frac{Y_1(s)}{U(s)} = \frac{\dfrac{s+q_3}{V_1}}{s^2 + (q_1+q_3)s + q_1 q_3 + q_2 q_4} = \frac{\varphi_2 s + \varphi_1}{s^2 + \varphi_4 s + \varphi_3}.$$

$H_1(s)$ – рациональный многочлен с коэффициентами, зависящими от неизвестных параметров. Записываем систему алгебраических уравнений (10), приравнивая коэффициенты при мономах к наблюдаемым параметрам φ_i:

$$\begin{cases} \dfrac{q_3}{V_1} = \varphi_1 \\ \dfrac{1}{V_1} = \varphi_2 \\ q_1 q_3 + q_2 q_4 = \varphi_3 \\ q_1 + q_3 = \varphi_4 \end{cases}. \tag{10}$$

Система (10) не имеет единственного решения, следовательно, модель априорно неидентифицируема. Если рассмотреть вместо параметров q_2 и q_4 – их произведение,

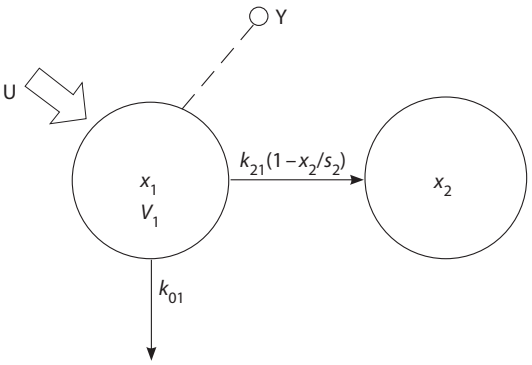


Fig. 2. A nonlinear two-compartmental pharmacokinetic model.

x_1, drug amount in blood; x_2, drug amount in the peripheral compartment; V_1, volume of the first ompartment.

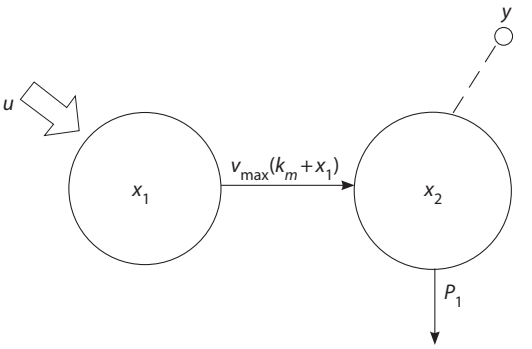


Fig. 3. A nonlinear two-compartmental model with Michaelis – Menten's kinetics.

x_1, x_2, drug amounts in the blood and tissues; u, the drug loading function; y, measurements in the second compartment.

считая их одним параметром, то система разрешится, и набор параметров, определяемых единственным образом, имеет вид: $p = [q_1, q_3, V_1, q_2q_4]$.

Предположим, что в качестве данных измерений известна также концентрация инсулина, т.е. $y_2(t) = \frac{x_2(t)}{V_2}$. Для таких данных измерений также выпишем передаточную функцию. Применяя преобразование Лапласа для соответствующих уравнений системы, получим выражение для $Y_2(s)$, передаточная функция $H_2(s)$ имеет следующий вид:

$$H_2(s) = \frac{Y_2(s)}{U(s)} = \frac{\frac{q_4}{V_2}}{s^2+(q_1+q_3)s+q_1q_3+q_2q_4}.$$

К системе алгебраических уравнений (10) добавляется еще одно уравнение, связывающее наблюдаемые данные с данными измерений: $\frac{q_4}{V_2} = \varphi_5$. Тогда набор идентифицируемых параметров имеет вид: $p = \left[q_1, q_3, V_1, q_2V_2, \frac{q_4}{V_2}\right]$.

Следует отметить, что если известно значение V_2 (или, например, что $V_1 = V_2$), модель станет идентифицируемой. Таким образом, можно заключить, что добавление информации об исследуемом процессе делает вычисление параметров более надежным, что согласуется с интуитивным представлением.

2.2. Анализ идентифицируемости двухкамерной фармакокинетической модели методом разложения в ряд тейлора

Рассмотрим нелинейную двухкамерную фармакокинетическую модель (Carson, Cobelli, 2008) (рис. 2).

Предположим, что препарат вводится в момент времени $t = 0$ ($U(t) = D\delta(t)$) в плазму крови (камера 1). В этой же камере измеряется концентрация $y(t) = \frac{x_1(t)}{V_1}$. Элиминация препарата происходит линейно и характеризуется параметром k_{01}, переход вещества в недоступную для измерений камеру 2 (группа органов с низким кровоснабжением) осуществляется «ленгмюровской кинетикой насыщения» – со скоростью $k_{21}(1 - \frac{x_2(t)}{s_2})$. Математическая модель имеет следующий вид:

$$\begin{cases} \dot{x}_1(t) = -k_{01}x_1(t) - k_{21}\left(1 - \frac{x_2(t)}{s_2}\right)x_1(t) + U(t) \\ \dot{x}_2(t) = k_{21}\left(1 - \frac{x_2(t)}{s_2}\right)x_1(t) \\ y(t) = \frac{x_1(t)}{V_1} \end{cases}. \quad (11)$$

Для определения идентифицируемости данной модели применим метод разложения в ряд Тейлора. Для этого нужно определить значения $y(0)$, $\dot{y}(0)$, $\ddot{y}(0)$ и $\dddot{y}(0)$.

Из описания модели имеем $x_1(0) = D$. Для $\dot{x}_1(0)$, учитывая структуру системы (11) и тот факт, что $x_2(0) = 0$, получаем $\dot{x}_1(0) = -(k_{01}+k_{21})x_1(0)$. Аналогично $\ddot{x}_1(0) = -(k_{01}+k_{21})\dot{x}_1(0) + \frac{k_{21}}{s_2}k_{21}\dot{x}_1^2(0)$ и $\dddot{x}_1(0) = -\frac{k_{21}^2}{s_2}\left[\frac{k_{21}}{s_2}x_1^2(0) - 3\dot{x}_1(0)\right]x_1(0) - (k_{01}+k_{21})\ddot{x}_1(0)$.

Далее легко найти $y(0)$, $\dot{y}(0)$, $\ddot{y}(0)$ и $\dddot{y}(0)$ из полученных соотношений, разделив каждое из уравнений на V_1. Согласно алгоритму выражения для $y(0)$, $\dot{y}(0)$, $\ddot{y}(0)$ и $\dddot{y}(0)$, приравняем к наблюдаемым параметрам φ_j для получения системы алгебраических уравнений относительно искомых параметров:

$$\begin{cases} \frac{D}{V_1} = \varphi_1 \\ \frac{1}{V_1}[-(k_{01}+k_{21})x_1(0)] = \varphi_2 \\ \frac{1}{V_1}\left[-(k_{01}+k_{21})\dot{x}_1(0) + \frac{k_{21}}{s_2}k_{21}\dot{x}_1^2(0)\right] = \varphi_3 \\ \frac{1}{V_1}\left[-\frac{k_{21}^2}{s_2}\left(\frac{k_{21}}{s_2}x_1^2(0) - 3\dot{x}_1(0)\right)x_1(0) - (k_{01}+k_{21})\ddot{x}_1(0)\right] = \varphi_4 \end{cases}.$$

Система имеет единственное решение, а значит, рассматриваемая модель является идентифицируемой.

2.3. Анализ идентифицируемости модели методом, основанным на теории дифференциальной алгебры

Рассмотрим двухкамерную фармакокинетическую модель, описывающую распространение препарата с нелинейной

кинетикой в крови и тканях (Bellu et al., 2007) (рис. 3). Система (1) для такой модели имеет вид:

$$\begin{cases} \dot{x}_1 = -\dfrac{v_{\max}}{k_m + x_1} x_1 + u \\ \dot{x}_2 = \dfrac{v_{\max}}{k_m + x_1} x_1 - p_1 x_2 \\ y = x_2 \end{cases} . \qquad (12)$$

Данные измерений известны в тканях, т. е. функция выходных данных имеет вид: $y = x_2$, p_1 – константа скорости, $v_{\max}$ и k_m – классические параметры кинетики Михаэлиса-Ментен. Неизвестный набор параметров $p = [p_1, v_{\max}, k_m]$.

Проведем анализ идентифицируемости такой модели методом из раздела 2.3.

1. Формальное упорядочение слагаемых имеет вид: $u < y < \dot{u} < \ddot{u} < \dot{y} < \ddot{y} < x_1 < x_2 < \dot{x}_1 < \dot{x}_2$.

2. Упорядочение полиномов имеет следующий вид:

$$\begin{cases} y - x_2 = 0 \\ \dot{x}_1(k_m + x_1) + v_{\max} x_1 - u(k_m + x_1) = 0 \\ \dot{x}_2(k_m + x_1) - v_{\max} x_1 + p_1 x_2 (k_m + x_1) = 0 \end{cases} .$$

3. После применения процедуры псевдоделения получаем одно соотношение на данные входа и выхода:

$$-\ddot{y} k_m^2 v_{\max}^2 - \dot{y}^3 k_m v_{\max} + \dot{y}^2\left(-3p_1 k_m v_{\max} y + k_m(u v_{\max} + 2v_{\max}^2)\right) +$$
$$- \dot{y}\left(-3p_1^2 k_m v_{\max} y^2 + p_1(-k_m^2 v_{\max}^2 + k_m)(2v_{\max} u y + 4v_{\max}^2 y)\right) +$$
$$+ k_m(-2v_{\max}^2 u - v_{\max}^3) - p_1^3 k_m v_{\max} y^3 + p_1^2 k_m (u v_{\max} y^2 + 2v_{\max}^2 y^2) +$$
$$+ p_1 k_m(-2u v_{\max}^2 y - v_{\max}^3 y) + k_m v_{\max}^3 u = 0.$$

4. Нелинейная система алгебраических уравнений имеет вид:

$$\begin{cases} k_m v_{\max} = \gamma\beta \\ -p_1 = -\alpha \\ v_{\max} = \beta \\ -p_1^2 = -\alpha^2 \\ -p_1 v_{\max} = -\alpha\beta \\ v_{\max}^2 = \beta^2 \\ -p_1^3 = -\alpha^3 \\ -p_1^2 v_{\max} = -\alpha^2\beta \\ -p_1 v_{\max}^2 = -\alpha\beta^2 \\ -k_m p_1 v_{\max} = -\gamma\alpha\beta \end{cases} . \qquad (13)$$

5. Система алгебраических уравнений (13) разрешима единственным образом относительно искомого набора параметров $p = [p_1, v_{\max}, k_m]$, а значит, исходная фармакокинетическая модель (12) является идентифицируемой.

2.4. Пример замены переменных, приводящей неидентифицируемую модель к идентифицируемой

Рассмотрим фармакокинетическую трехкамерную модель (Meshkat, 2014) с введением препарата и взятием проб из первой камеры ($y = x_1$) и элиминацией из всех трех камер

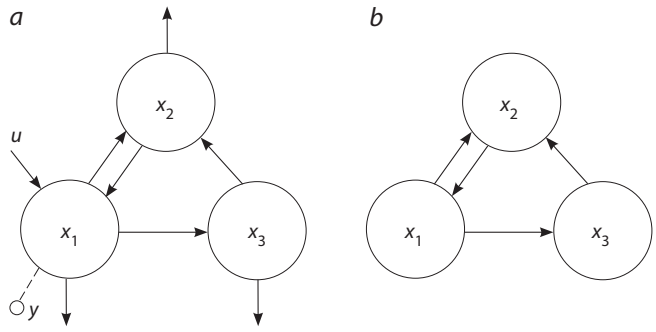


Fig. 4. (*a*) A three-compartmental pharmacokinetic linear model; *u*, the drug loading function; *y*, measurements in the first ompartment. (*b*) A graph with three vertices.

(рис. 4). Соответствующая система дифференциальных уравнений имеет вид:

$$\begin{cases} \dot{x}_1 = a_{11} x_1 + a_{12} x_2 \\ \dot{x}_2 = a_{21} x_1 + a_{22} x_2 + a_{23} x_3 \\ \dot{x}_3 = a_{31} x_1 + a_{33} x_3 \\ y = x_1 \end{cases} .$$

Согласно алгоритму из 1.4 первым шагом является построение остовного дерева, в нашем случае остовное дерево соответствует ребрам a_{23} и a_{12}. Второй шаг состоит в построении матрицы инцидентности $E(T) = \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{pmatrix}$. Вычеркиваем 1-ю строку матрицы $E(T)$, получим E_1 и находим $E_1^{-1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Делаем замену согласно формуле в пункте (в) раздела 1.4. Получаем явный вид рациональной репараметризации для этой модели: $X_1 = x_1, X_2 = a_{12} x_2$, $X_3 = a_{12} a_{23} x_3$.

Итак, полученная модель является идентифицируемой и имеет следующий вид:

$$\begin{cases} \dot{X}_1 = a_{11} X_1 + X_2 \\ \dot{X}_2 = a_{12} a_{21} X_1 + a_{22} X_2 + X_3 \\ \dot{X}_3 = a_{12} a_{31} a_{23} X_1 + a_{33} X_3 \\ y = X_1 \end{cases} .$$

При определенной структуре модели и данных на входе и выходе анализ априорной идентифицируемости позволяет избежать лишних экспериментов, если искомые параметры являются неидентифицируемыми. Проведение такого анализа позволяет определить комбинации параметров рассматриваемой модели, относительно которых модель является идентифицируемой.

В случае неидентифицируемости модели следует предпринять один из следующих шагов:

1) добавить данные о характере эксперимента, привлекать дополнительные данные измерений;

2) упростить исходную модель, меняя ее структуру. Добавить априорную информацию о некоторых параметрах модели;

3) искать замену переменных, приводящую исходную модель к идентифицируемой.

Acknowledgments

This work was supported by the Ministry of Education and Science of the Russian Federation, contract No. 107 "Design of program tools for the study of numerical solution of direct and inverse problems in pharmacokinetics and epidemiology" and by the Ministry of Education and Science of the Kazakhstan Republic, contract No. 1746/GF4 "Theory and numerical methods of the solution of inverse and ill-posed problems in natural science".

Conflict of interest

The authors declare no conflict of interest.

References

Audoly S., D'Angio L. On the identifiability of linear compartmental system: a revisited transfer function approach based on topological properties. Mathematical Biosciences. 1983;66:201-228.

Bellman R., Astrom K. On strucural identifiability. Mathematical Biosciences. 1970;7(3):329-339.

Bellu G., Saccomani P., Audoly S., D'Angio L. Daisy: a new software tool to test global identifiablity of biological and physiological system. Comput Methods Programs Biomed. 2007;88(1):52-61. DOI 10.1016/j.cmpb.2007.07.002

Ben-Zvi A., McLellan P.J., McAuley B.K. Identifability of linear time-invariant diferential-algebraic systems. Industrial Engineering Chemistry Res. 2004;43(8):1251-1259.

Brown R. Compartmental system analysis: state of the Art. IEEE Transactions on Biomedical Engineering. 1980;27(1):1-38.

Brown R.F. Identifiability: role in design of pharmacokinetic experiments. IEEE Transactions on Biomedical Engineering. 1982;29:49-54.

Carson E., Cobelli C. Modelling Methodology for Physiology and Medicine. San Diego: Academic Press, 2001.

Carson E., Cobelli C. Introduction to Modelling in Physiology and Medicine. San Diego: Academic Press, 2008.

Cobelli C., DiStefano III J. Parameter and structural identifiability concepts and ambiguities: a Critical review and analysis. Amer. J. Physiology-Regulatory, Integrative and Comparative Physiology. 1980;23(9):7-24.

Cobelli C., Lepschy A., Jacur G.R. Identifiability of compartmental systems and related structural properties. Mathematical Biosciences. 1976;48:1-18.

Goodwin G.C., Payne R.L. Dynamic system identification: experiment design and data analysis. N.Y.: Academic Press, 1977.

Jacquez J.A., Greif P. Numerical parameter identifiability and estimability: integrating identifiability, estimability and optimal design. Mathematical Biosciences. 1985;77:201-227.

Meshkat N. Identifiable reparametrizations of linear compartment models. J. Symbolic Computation. 2014;63:46-67. DOI 10.1016/j.jsc.2013.11.002

Meshkat N., Anderson C., DiStefano III J. Alternative to Ritt's Pseudodivision for finding the input-output equations of multi-output models. Mathemat. Biosciences. 2012;239(1):117-123. DOI 10.1016/j.mbs.2012.04.008

Meshkat N., Eisenberg M., DiStefano III J. On finding and using identifiable parameter combinations in nonlinear dynamic systems biology models and COMBOS: A novel web implementation. Plos One. 2014;9(10). DOI 10.1371/journal.pone.0110261

Tunali T., Tarn T.J. New Results for Identifiability of Nonlinear Systems. IEEE Transactions on Automatic Control. 1987;32(2):146-154.

Walter E., Lecourtier Y., Happel J., On the structural output distinguishability of parametric models, and its relation with structural identifiability. IEEE Trans. Aut. Control. 1984;29:56-57.

Вавиловский журнал генетики и селекции. 2015;19(6):745-752
Doi 10.18699/VJ15.095

Современные подходы к математическому и компьютерному моделированию в микробиологии

А.И. Клименко[1, 2], З.С. Мустафин[1], А.Д. Чеканцев[1, 2], Р.К. Зудин[1, 2], Ю.Г. Матушкин[1], С.А. Лашин[1, 2]

1 Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия 2 Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия

Бактериальные сообщества являются тесно взаимосвязанными системами, состоящими из большого числа видов, что значительно осложняет анализ их структуры и взаимоотношений. В настоящий момент существует ряд экспериментальных методов, предоставляющих гетерогенные данные, касающиеся различных аспектов этого объекта исследования. Произошедшее за последнее время резкое увеличение объема доступных метагеномных данных представляет интерес не только для биостатистиков, но и для специалистов в области моделирования биосистем, поскольку эти данные позволяют повысить качество моделей. В то же время методы математического и компьютерного моделирования оказываются полезны для понимания эволюции микробных сообществ и их функции в экосистеме. В статье представлен обзор существующих методов и средств математического и компьютерного моделирования, использующихся в области экологии микробных сообществ и опирающихся на различные типы экспериментальных данных. Рассмотрены подходы, фокусирующиеся на описании таких аспектов микробного сообщества, как его трофическая структура, метаболическая и популяционная динамика, генетическое разнообразие, а также пространственная гетерогенность и динамика распространения. В работе также приведена классификация существующих программных средств моделирования микробных сообществ. Показано, что несмотря на преобладание тенденции к использованию гибридных подходов к моделированию, остаются актуальными проблемы интеграции между моделями, описывающими различные уровни биологической организации сообществ. Многоаспектность интеграционных подходов, используемых для моделирования микробных сообществ, основана на необходимости учитывать гетерогенные данные, полученные из различных источников с помощью высокопроизводительных экспериментальных методов исследования генома.

Ключевые слова: микробные сообщества; экологическое моделирование; эволюционное моделирование; прокариоты.

HOW TO CITE THIS ARTICLE?
Klimenko A.I., Mustafin Z.S., chek antsev A.D., Zudin r.K ., Matushkin Yu.G., lashin S.A. A review of simulation and modeling approaches in microbiology. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2015;19(6):**745-752.** Doi 10.18699/VJ15.095

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ?
Клименко А.И., Мустафин З.С., Чеканцев А.Д., Зудин Р.К., Матушкин Ю.Г., лашин С.А. Современные подходы к математическому и компьютерному моделированию в микробиологии. Вавиловский журнал генетики и селекции. 2015;19(6):745-752. Doi 10.18699/VJ15.095

Received 25.09.2015
Accepted for publication 29.10.2015

A review of simulation and modeling approaches in microbiology

A.I. Klimenko[1, 2], Z.S. Mustafin[1], A.D. Chekantsev[1, 2], R.K. Zudin[1, 2], Yu.G. Matushkin[1], S.A. Lashin[1, 2]

1 Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
2 Novosibirsk State University, Novosibirsk, Russia

Bacterial communities are tightly interconnected systems consisting of numerous species making it challenging to analyze their structure and relations. There are several experimental techniques providing heterogeneous data concerning various aspects of this object. A recent avalanche of metagenomic data challenges not only biostatisticians but also biomodelers, since these data are essential to improve the modeling quality while simulation methods are useful to understand the evolution of microbial communities and their function in the ecosystem. An outlook on the existing modeling and simulation approaches based on different types of experimental data in the field of mic obial ecology and environmental microbiology is presented. A number of approaches focusing on a description of such microbial community aspects as its trophic structure, metabolic and population dynamics, genetic diversity as well as spatial heterogeneity and expansion dynamics is considered. we also propose a classific tion of the existing software designed for simulation of microbial communities. it is shown that although the trend for using multiscale/hybrid models prevails, the integration between models concerning different levels of biological organization of communities still remains a problem to be solved. The multiaspect nature of integration approaches used to model microbial communities is based on the need to take into account heterogeneous data obtained from various sources by applying high-throughput genome investigation methods.

Key words: microbial communities; ecological simulation; evolutionary modeling; prokaryotes.

e-mail: klimenko@bionet.nsc.ru

Микроорганизмы образуют разнообразные сообщества, которые динамически изменяются по структуре и функции в ответ на изменения окружающей среды. Примерами таких сообществ являются биопленки и бактериальные маты (Karunakaran et al., 2011), а также сообщества, населяющие, например, кишечник (Chewapreecha, 2013) или ротовую полость (Salli, Ouwehand, 2015) человека. Являясь сложной адаптивной системой, микробное сообщество демонстрирует свойства более высокого порядка, которые не присутствуют в отдельных микробах, но возникают из их взаимодействий. Как было отмечено в статье Comolli (2014), взаимодействия комплексной природы, включающие трофические, физические и даже информационные (например, кворум-чувствительность) факторы, возникающие между клетками микробного сообщества, в том числе и между клетками разных видов, играют важную роль в функционировании этого сообщества в качестве единого целого, голобионта.

В последние годы в научной литературе опубликовано много работ по моделированию различных аспектов жизнедеятельности бактерий. В одних статьях рассматривали биологические аспекты моделирования, такие как связь между индивидуальным и популяционным ростом бактериальных клеток (Kutalik et al., 2005), способность поддержания биологического разнообразия системы при разных ландшафтах приспособленности и частотах мутаций (Beardmore et al., 2011). В других рассматривали различные методики компьютерного моделирования (Song et al., 2014), анализировали целесообразность, а также плюсы и минусы использования индивидуально-ориентированного моделирования, вместо классических методов (DeAngelis, Mooij, 2005; Grimm et al., 2006) или же клеточных автоматов (Esteban, Rodríguez-Patón, 2011).

Предсказательные математические и компьютерные модели не только помогли бы понять фундаментальные законы, лежащие в основе динамики и синергетических свойств естественных и синтетических микробных сообществ, но также представили бы практический интерес для их применения в задачах генной инженерии. Отметим особо тот факт, что сразу несколько биологических особенностей микробных сообществ делают их весьма сложным объектом для изучения *in vitro*: это и наличие некультивируемых видов, и физические размеры сообществ, и сложности в воспроизведении в лаборатории пространственной структуры и других физических параметров среды обитания сообщества. Соответственно, верификация математических и компьютерных моделей природных сообществ сопряжена с проблемами поиска качественных экспериментальных данных, в ряде случаев принципиально неразрешимых. Для решения подобных проблем предлагается создавать серии искусственных микробных сообществ, для каждого из которых одновременно строилась бы математическая модель, которая затем верифицировалась бы по экспериментальным данным, полученным при исследовании этих сообществ (De Roy et al., 2013; Wolfe, Dutton, 2015). При этом отмечается широкий спектр экспериментальных техник, которые могли бы использоваться при таком подходе (Колмакова, 2013), в частности *in vitro* культивирование, микроскопия, *in situ* мониторинг и сэмплинг, высокопроизводительное секвенирование и метагеномика, метатранскриптомика, метапротеомика, метаболомика. Отметим, что одним из средств дизайна подобных синтетических сообществ являются методы математического и компьютерного моделирования (Wolfe, Dutton, 2015). В работе Ларсена с коллегами (Larsen et al., 2012) были рассмотрены различные подходы к моделированию с точки зрения исследования микробного средового интерактома (МСИ). Показано, что МСИ может быть описан с помощью трех параметрических пространств: параметров среды, структуры микробного сообщества и средового метаболома. При этом для описания взаимоотношений между различными парами этих пространств подходят соответствующие методики.

Помимо большого числа разработанных моделей, посвященных описанию тех или иных аспектов функционирования микробных сообществ, в настоящее время существует ряд компьютерных средств, предназначенных для моделирования пространственно распределенных бактериальных сообществ. Большинство из них, такие как клеточные автоматы UMCCA (Laspidou, Rittmann, 2004), компьютерные гибридные системы моделирования AQUASIM (Wanner, Morgenroth, 2004; Mburu et al., 2014), INDISIM (Ginovart et al., 2002), делают серьезный упор на детальное описание пространственной структуры сообществ. Другие концентрируются на описании того, как процессы генетической изменчивости могут влиять на пространственную структуру сообщества, такие как AEvol (Knibbe et al., 2008; Beslon et al., 2010), однако они описывают пространственную организацию сообществ недостаточно детально.

В статье приведен обзор методов моделирования микробов и микробных сообществ. В моделях описаны как отдельные уровни их биологической организации, так и одновременно несколько таких уровней. Последние помогают выявить закономерности эволюции микробных сообществ, возникающие на генетическом уровне и распространяющиеся в дальнейшем на все прочие уровни функционирования микробного сообщества.

Методы моделирования разных уровней биологической организации микробного сообщества

В настоящее время существует ряд методов и программных средств моделирования микробных сообществ, фокусирующихся на тех или иных аспектах их жизнедеятельности. Хотя эти аспекты в живых организмах так тесно переплетены, что иногда бывает трудно их разделить, это приходится делать потому, что для разных процессов следует применять разные методы описания и моделирования. Рассмотрим эти аспекты в том виде, в каком они обычно выделяются при моделировании:

Экологическая структура сообщества. Под экологической структурой сообщества понимаются, прежде всего, взаимоотношения между видами. Для ее описания применимы любые виды реконструкции биологических сетей: нелинейная регрессия, продукционные методы и т. д.

Метаболическая и популяционная динамика. Широко используемыми методами в этой сфере являются обык-

новенные дифференциальные уравнения (ОДУ), алгебраические и разностные уравнения, булевы функции, матричные модели, термодинамическое стохастическое моделирование и т. д.

Генетическое разнообразие. Для описания генетического разнообразия используются дискретные (уравнения динамики аллелей) и стохастические модели, а также индивидуально-ориентированные подходы.

Пространственные гетерогенность и динамика. К этой категории относятся гетерогенные распределения клеток, субстратов, метаболитов и т. д., паттерны этих распределений, пространственно специфичное взаимодействие между видами и средой обитания, подвижность клеток, миграции и т. д. К используемым для моделирования пространственной гетерогенности и динамики методам относят уравнения в частных производных (УЧП), клеточные автоматы, агентное моделирование, модели сетевые, популяционного баланса (интегро-дифференциальные уравнения в частных производных) и т. д.

реконструкция экологической структуры сообщества

Реконструкция экологических взаимоотношений в сообществе, устанавливающая его трофическую организацию (сеть метаболических связей между видами), является одним из первых этапов анализа этого сообщества. Методы метагеномики и биоинформатики позволяют идентифицировать виды членов сообщества и оценить их относительные плотности, а также функциональные способности (Wooley et al., 2010). Появление большого количества метагеномных данных привело к развитию методов реконструкции трофических сетей сообществ на основе этой информации (Faust, Raes, 2012). Как правило, это регрессионные и продукционные методы, а также динамическое моделирование (Zomorrodi et al., 2014) и стехиометрические подходы расчета обмена метаболитами (Klitgord, Segre, 2010). Эти методы позволяют оценивать экологические отношения в сообществе, в том числе и в зависимости от параметров среды обитания. Микробные отношения могут быть реконструированы из данных о плотностях популяций. Основываясь на традиционном восприятии, мы можем называть отношение пары организмов конкурентным (или отрицательным), если их плотности по всем образцам антикоррелированы, несмотря на то что они обладают общей экологической нишей, и, напротив, мы можем называть отношение пары организмов кооперативным (или положительным), если они демонстрируют схожее распределение плотностей. Сеть микробных взаимосвязей может быть предсказана с использованием методов, называемых реконструкцией сетей. Парные отношения выводятся с помощью основанных на сходстве методов путем анализа распределения взаимных встречаемости/исключения двух видов исходя из суммы баллов похожести. Сложные взаимодействия между более чем двумя видами могут быть зафиксированы с использованием других техник, таких как регрессионные и продукционные методы. Регрессионные методы представляют плотность определенного вида как функцию от плотностей других видов. Продукционные методы изначально перечисляют все логически возможные правила сосуществования/исключения видов, которые поддерживаются набором данных об их наличии или отсутствии. В ходе последовательного фильтрационного процесса сохраняются только значимые правила. В работе (Faust, Raes, 2012) представлен исчерпывающий обзор по этому вопросу.

Установленные отношения между видами-членами могут быть представлены как сеть микробных взаимосвязей, состоящая из вершин (виды или таксоны) и ребер (межвидовые взаимодействия). Поскольку отношения между видами часто асимметричны, т. е. наличие одного вида может влиять на популяцию другого, но не наоборот, то данная сеть представляет собой ориентированный граф. Направление и сила микробного взаимодействия могут быть представлены в виде стрелки соответствующей толщины. Переменные окружающей среды также могут быть встроены в сеть путем трактовки их как дополнительных видов-вершин. Эта расширенная сеть описывает взаимоотношения между видами и признаками окружающей среды. Например, согласованная совместная встречаемость между определенными видами и питательными веществами (например, нитритами и нитратами) свидетельствует о вовлеченности особых микробов в биогеохимические циклы (Fuhrman, 2009).

Итак, микробные отношения могут быть систематически реконструированы из данных о плотности видов. Полученные таким образом взаимосвязи между микробами специфичны по отношению к условиям. Это означает, что информация об отношениях между видами микробов, полученная при одних условиях, может быть недействительной в других условиях, поскольку структура и свойства сетей взаимосвязей между микробами могут значительно видоизменяться в зависимости от условий окружающей среды. Также эти методы ничего не говорят о биологических причинах того, почему определенные виды взаимодействуют особым образом, в то время как другие – нет. Чтобы получить более механистичное понимание, требуются методы, основанные на физиологии, такие как стехиометрическое моделирование.

Моделирование метаболизма и генетической регуляции

Для моделирования микробного метаболизма используется широкий круг математических методов, включающий дифференциальные уравнения (обыкновенные и в частных производных), булевы сети и сети Петри, алгебраические линейные и нелинейные уравнения, клеточные автоматы и др. Моделирование метаболизма часто сопряжено с моделированием генетической регуляции (De Jong, 2002; Hecker et al., 2009; Лихошвай и др., 2010). Интегрирующая роль здесь отводится концепции генных сетей. Как правило, в подобных моделях описывалась отдельная метаболическая подсистема микробной клетки, возможно, с сопутствующей ей генетической регуляцией (Covert et al., 2001; Likhoshvai, Ratushny, 2007; Oberhardt et al., 2009). Однако с конца XX–начала XXI столетия начинаются попытки создания полной модели метаболизма клетки, так называемые модели электронной клетки (Tomita et al., 1999; Tomita, 2001; Ishii et al., 2004; Price

et al., 2004; Sauer et al., 2007; Durot et al., 2009). В 2012 г. Карр с коллегами сообщили о том, что построенная ими модель электронной клетки предсказывает фенотип по генотипу (Karr et al., 2012).

Недавно один из широко используемых методов моделирования клеточного метаболизма – динамический анализ стационарных потоков – был расширен на случай моделирования микробных сообществ (Mahadevan, Henson, 2012; Henson, Hanly, 2014). Кроме того, широко используются оптимизационные методы, такие как метод минимизации метаболического регулирования (MOMA) (Segrè et al., 2002), а также методы, включающие многокритериальную оптимизацию (см. Zomorrodi, Maranas, 2012; Zomorrodi et al., 2014), которая позволяет исследователю использовать критерии приспособленности уровня всего сообщества. Помимо этого, для моделирования метаболизма также используются анализ элементарных режимов (EM) (Schuster et al., 2000) и эволюционная теория игр (EGT) (Pfeiffer, Schuster, 2005; Frey, 2010) .

Пространственная гетерогенность и популяционная динамика микробных сообществ

Другой аспект жизнедеятельности микробного сообщества, являющийся предметом как экспериментальных, так и теоретических исследований, – это популяционная динамика, т. е. изменение численностей составляющих сообщество популяций во времени или в ряду поколений. В простейшем случае рассматриваются модели однородных сред с равномерным перемешиванием. Математические модели микробных популяций берут свое начало с работ Ж. Моно, предложившего теорию хемостата (культиватора) и модель микробной популяции в культиваторе с одним субстратом, от концентрации которого зависит скорость роста клетки (Monod, 1950; Ризниченко, Рубин, 1993). На основе принципа лимитирующего фактора в ферментативных процессах, сформулированного Н.Д. Иерусалимским (Чернавский, Иерусалимский, 1965), для функции «скорость роста клетки» была предложена более реалистичная формула, учитывающая, кроме концентрации субстрата, ингибирующее влияние продуктов метаболизма микробных клеток, известная как формула Моно – Иерусалимского (Ризниченко, Рубин, 1993). Непрерывная модель возрастной структуры микробной популяции, оперирующая не с численностями отдельных групп популяции, а с непрерывной функцией плотности распределения организмов по возрастам, была получена Мак-Кендриком в 1926 г., а затем переоткрыта фон Ферстером в 1959 г. (Ризниченко, 2003). Матричные модели динамики структуры популяций (например, возрастной), впервые предложенные П. Лесли (Leslie, 1945), детально исследованы в работах (Гимельфарб и др., 1974) (Логофет, Белова, 2007).

Однако, несмотря на то что предположение о равномерном перемешивании удобно с точки зрения проведения численного исследования и широко используется, оно слабо согласуется с большинством наблюдений реальных биологических систем, в которых градиенты питательных веществ, света и метаболитов играют важную роль в структуризации сообщества (Wimpenny et al., 2000). Поэтому пространственная неоднородность и динамика являются еще одной областью исследований, где оказываются полезными методы математического и компьютерного моделирования.

Использование моделей в формализме уравнений в частных производных является одним из традиционных подходов к описанию пространственной гетерогенности и исследованию образующихся в системе паттернов распределения. Одним из первых случаев применения данной методики к исследованиям в области теоретической биологии была знаменитая работа А.М. Тьюринга (Turing, 1952), в которой он предложил модель «реакция – диффузия», способную в несложных системах синтеза веществ, связанных отношениями активации и ингибирования и распространяющихся в пространстве посредством диффузии, воспроизводить нетривиальные паттерны пространственного распределения. При описании проточных систем используются классические для гидродинамики уравнения Навье – Стокса (Lencastre Fernandes et al., 2011), также являющиеся уравнениями в частных производных.

Существует ряд подходов, позволяющих описать не только пространственную гетерогенность, но и изменчивость внутри популяций. Одним из таких подходов является использование моделей популяционного баланса (Ramkrishna, 2000). С математической точки зрения, данные модели представляют собой интегро-дифференциальные уравнения в частных производных, описывающие как пространственные координаты, так и внутренние характеристики объекта, такие как, например, масса клетки, ее возраст и морфология. Индивидуально-ориентированные модели также позволяют сочетать описание пространственного распределения с внутренними характеристиками моделируемых объектов. В моделях данного типа пространственная гетерогенность описывается с помощью «лоскутов» (Stauffer et al., 2005) квадратной решетки или ячеек сетки соответствующей размерности (Klimenko et al., 2015). Среди других методов для описания пространственной неоднородности используются клеточные автоматы (Wimpenny, Colasanti, 1997), методы теории графов (O'Donnell et al., 2007) и др. С проблемой пространственной динамики микробных сообществ тесно связано описание подвижности организмов. Известно, что значительное число видов бактерий способны активно передвигаться в окружающей среде по направлению к питательным веществам или же лучшим условиям обитания (Adler, 1976). Как правило, микроорганизмы используют для своего передвижения жгутики (Henrichsen, 1972) или другие механизмы, такие как, например, специальные белки, расположенные на мембране (например, *Flavobacterium johnsoniae*) (Shrout, 2015), реснички, позволяющие скользить клеткам *Oscillatoria princeps* (Halfen, Castenholz, 1971), или же микроорганизмы передвигаются, изменяя поверхностное натяжение путем выделения поверхностно-активных веществ (как это делают представители вида *M. xanthus*) и т. д. Способность передвигаться в соответствии с градиентами определенных экологических факторов называется таксисом (например, хемотаксис, фототаксис и т. д.) (Нетрусов, Котова, 2007). Подробный обзор математических подходов, используемых для мо-

делирования бактериального хемотаксиса, представлен в работах (Tindall et al., 2008a, b). Для микробных сообществ, распределенных вдоль одномерного пространства, популяционное уравнение, описывающее как случайные, так и хемотаксические движения, может быть дано в форме, известной как модель хемотаксиса Келлера – Сигела (Tindall et al., 2008b).

В работе (Emonet et al., 2005) было представлено программное средство, позволяющее изучить влияние стохастических флуктуаций в межклеточных взаимодействиях на поведение отдельных клеток. Была разработана мультиагентная программная система AgentCell, с помощью которой авторы смоделировали хемотаксический ответ свободных клеток *E. coli* на градиент хемоаттрактантов в трехмерной среде. В данной модели каждая клетка бактерии является самостоятельным агентом, имеющим собственную генную сеть хемотаксиса, молекулярные моторы и жгутик. Использовалась модель генной сети хемотаксиса Мортон-Фирт и Коробковой. На вход модели сети поступал параметр занятости рецептора (вероятность того, что рецептор связан с лигандом), что соответствует концентрации питательного вещества в среде. Выходным параметром сети является количество молекул регулятора хемотаксического ответа CheY-P внутри клетки. Для проверки был смоделирован хемотаксический ответ свободно плавающих бактерий на линейный градиент концентрации. Результаты моделирования согласуются с экспериментальными данными, полученными для отдельных клеток и клеток, взятых из бактериальной популяции (Emonet et al., 2005).

Другим примером использования индивидуально-ориентированного подхода для моделирования подвижности бактериальных клеток является работа Б. Ниу с коллегами (Niu et al., 2013). Они смоделировали процессы хемотаксиса, сравнив результаты поведения бактерий в 3D среде с чувством кворума и без него. Авторы рассмотрели различные стратегии обмена информацией между клетками бактериальной популяции и оценили их эффективность в деле достижения глобального оптимума. Согласно их результатам, клетки популяции достигают наиболее благоприятных условий при наиболее интенсивных коммуникациях, задействующих как индивидуальные, так и межгрупповые механизмы обмена информацией.

Проблемы интеграции и многоуровневые подходы к моделированию микробных сообществ

Одной из острых и актуальных проблем в сфере моделирования микробных сообществ является проблема интеграции различных методик моделирования в рамках единого исследования. В обзоре (Song et al., 2014) приводится следующая классификация стратегий по интегрированию моделей: 1) информационная обратная связь; 2) непрямое сцепление; 3) прямое сцепление. Информационная обратная связь является самой слабой формой интеграции: в данном случае результаты «верхнего» модельного слоя используются, чтобы настроить предположения, лежащие в основе независимого «нижнего» модельного уровня. Непрямое сцепление представляет собой конвейер, при котором результаты одной модели передаются на вход другой независимой модели (Scheibe et al., 2009). Прямое сцепление подразумевает степень интеграции, в которой различные методики моделирования сливаются в единую систему. При многоуровневом моделировании (с прямым сцеплением, по терминологии Сонга) индивидуально-ориентированные методики выигрывают по сравнению со своими аналогами, благодаря своей гибкости и способности интегрировать различные методики в качестве подмоделей единой системы моделирования. Например, данная концепция была успешно применена в работе (Rudge et al., 2012) для комбинированного моделирования внутриклеточной динамики, межклеточного сигналинга и клеточной биофизики бактериальных клеток, образующих биопленку. Авторы решают вопрос *in silico* предсказания поведения синтетических биопленок до их создания *in vitro*. При этом особый упор делается на эмерджентные свойства, проявляемые тысячами растущих и обменивающихся сигналами бактериальных клеток, поскольку эти свойства имеют решающее значение для проектирования синтетических биопленок.

С развитием компьютерных технологий широкое распространение получили методы индивидуально-ориентированного моделирования (DeAngelis, Mooij, 2005). В рамках данного подхода популяции моделируются как системы, состоящие из агентов, представляющих собой индивидуальные организмы или группы похожих организмов, обладающих набором признаков, варьирующих среди агентов. При этом каждый агент обладает своей уникальной историей взаимодействий со средой и другими агентами. Индивидуально-ориентированное моделирование широко применяется в экологическом моделировании, социальной динамике и моделировании эволюционных процессов. В рамках данных моделей изучается то, как поведение отдельных индивидуумов, следующих локальным правилам, приводит к формированию сложных паттернов, в том числе и пространственно распределенных, например, косяки рыб, стаи птиц, рой насекомых и т. п. (DeAngelis, Mooij, 2005).

Преимуществом данного подхода является то, что он позволяет максимально гибко отобразить разнообразие характеристик индивидуальной особи и в то же время явно описывает взаимодействия между отдельными организмами на микроуровне.

Основные недостатки метода индивидуально-ориентированного моделирования – необходимость в большом количестве экспериментальных данных для детального описания биологических объектов и высокая вычислительная сложность. Трудоемкость вычислений накладывает определенные ограничения на размеры моделируемых сообществ. Есть два основных подхода к снижению вычислительной нагрузки: 1) ограничение вычислительной области небольшим представительным пространством и 2) использование понятия супериндивидов. Например, можно снизить число моделируемых клеток, сосредоточившись на маленькой области биопленки или озера. Масштабирование к большему пространству на основе этого подхода становится трудным в случае, когда значима пространственная неоднородность в системах. В настоящее время индивидуально-ориентированное моделирование микробных сообществ ограничено масштабами

Comparison of software tools designed for the simulation of bacterial communities

Software name (reference)	Simulation unit	Population size	Genetic diversity	Spatial distribution	The range of target problems	Availability of documentation	Supported platforms
AgentCell (v. 2.0) (Emonet et al., 2005)	Cell	Thousands of cells	*Escherichia coli* only	Cells move through a 3D space with a predefined attractant gradient	Simulation of the chemotactic response of cells to a gradient of an attractant in 3D environment	installation guide	Linux
AQUASiM (Wanner, Morgenroth, 2004)	Compartment	Depends on biofilm thickness	Allows modeling multispecies batch fermenters but without dynamic genetic variation during the simulation	Takes into account various compartment and membrane connectivity topologies	Bacterial biofilm simulation in aquatic ecosystems. Provides tools for sensitivity analysis and parameter estimation	Manual and support mailing list	Windows, Linux, MacOS
inDiSiM (Ginovart et al., 2002)	Cell, super-individual	Millions of cells	Supports individual diversity among cells but does not allow for dynamic genetic variation during the simulation	Square lattice	investigation of: – biomass distribution throughout the colony; – dependency of colony growth rate on nutrient concentrations and temperature; – metabolite concentration variations in batch fermenters	–	Windows
Haploid Evolutionary Constructor (Klimenko et al., 2015)	Metabolically homogeneous population	Over one billion cells	Supports the whole range of genetic diversity and its dynamic variation (up to speciation)	Square lattice	– Tasks of evolutionary and ecological modelling; – investigation of interactions between population genetics, spatial factors and trophic structure of the ecosystem and their impact on the microbial community evolution	Documentation available on the website	Windows, Linux

от микрометров до сантиметров (Tang, Valocchi, 2013). В качестве альтернативы можно моделировать на основе супериндивидов, представляющих группу отдельных клеток (Scheffer et al., 1995). В таком случае возникает проблема, как согласованно определить супериндивидов для данной изучаемой системы, поскольку определение супериндивидов таким образом, чтобы те содержали большое число клеток, в конечном счете ослабляет присущую индивидуально-ориентированному моделированию силу, которая способна объяснить динамику каждой отдельной клетки. Классификация существующих программных средств моделирования микробных сообществ представлена в таблице.

Другой особенностью, проистекающей из многоаспектности этих интеграционных подходов, является необходимость учитывать гетерогенные данные, полученные из различных источников с помощью высокопроизводительных экспериментальных методов исследования генома, транскриптома, протеома и метаболома сообщества. Секвенирование нового поколения, масс-спектрометрия и другие высокопроизводительные методы генерируют огромные массивы экспериментальных данных, предоставляющих информацию о генетической структуре сообщества, представленности видов, экспрессии тех или иных функциональных групп генов и т. д. Поэтому учет этого информационного пласта в существующих методиках моделирования микробных сообществ является актуальной задачей.

В последнее время большое количество метагеномных данных послужило базой для обширного биостатистического анализа с использованием методов оценки биоразнообразия и видового богатства сообщества, методов понижения размерности, дисперсионного анализа, линейной регрессии и др. Хотя эти методы помогают исследователю прийти к определенным гипотезам касательно внутренней структуры объекта исследования, они не способны раскрыть механизм и установить причинно-следственные связи, лежащие в основе наблюдений о микробном сообществе. Для этих целей могут быть использованы методы математического и компьютерного моделирования, кото-

рые способны не только давать предсказания о поведении биологических систем, но и обнаруживать наши пробелы в знаниях, необходимых для их реконструкции *in silico*.

В данной работе показано, что одним из трендов развития данной области является объединение возможностей различных подходов к моделированию в рамках гибридных или многоуровневых моделей, что позволяет получить более полное знание о такой биологической системе, как микробное сообщество. Однако на этом пути существует ряд проблем, связанных как с межуровневой интеграцией моделей, так и с интеграцией данных из гетерогенных источников. Несмотря на все эти сложности, нет никаких сомнений, что исследователям удастся их преодолеть и вывести моделирование прокариотических сообществ на новый уровень.

Acknowledgments

This work was supported in part by Budgeted Project VI.61.1.2 and the Russian Foundation for Basic Research, project 15-07-03879.

Conflict of interest

The authors declare no conflict of interest.

References

Adler J. Chemotaxis in bacteria. J. Supramol. Struct. 1976;4:305-317. DOI 10.1146/annurev.bi.44.070175.002013

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;472:342-346. DOI 10.1038/nature09905

Beslon G., Parsons D.P., Sanchez-Dehesa Y., Peca J.-M., Knibbe C. Scaling laws in bacterial genomes: a side-effect of selection of mutational robustness? Biosystems. 2010;102:32-40. DOI 10.1016/j.biosystems.2010.07.009

Chernavskiy D.S., Ierusalimskiy N.D. On the determinant link in the system of enzyme reactions. Izvestiya AN SSSR, seriya biologiya = Proceedings of USSR Academy of Sciences. Biology. 1965;5:665-672.

Chewapreecha C. Your gut microbiota are what you eat. Nat. Rev. Microbiol. 2013;12:8. DOI 10.1038/nrmicro3186

Comolli L.R. Intra- and inter-species interactions in microbial communities. Front. Microbiol. 2014;5:1-3. DOI 10.3389/fmicb.2014.00629

Covert M.W., Schilling C.H., Famili I., Edwards J.S., Goryanin I.I., Selkov E., Palsson B.O. Metabolic modeling of microbial strains in silico. Trends Biochem. Sci. 2001;26:179-186. DOI 10.1016/S0968-0004(00)01754-0

De Jong H. Modeling and simulation of genetic regulatory systems: a literature review. J. Comput. Biol. 2002;9:67-103. DOI 10.1089/10665270252833208

De Roy K., Marzorati M., Van den Abbeele P., Van de Wiele T., Boon N. Synthetic microbial ecosystems: An exciting tool to understand and apply microbial communities. Environ. Microbiol. 2013;16:1472-1481. DOI 10.1111/1462-2920.12343

DeAngelis D.L., Mooij W.M. Individual-based modeling of ecological and evolutionary processes 1. Annu. Rev. Ecol. Evol. Syst. 2005;36:147-168. DOI 10.1146/annurev.ecolsys.36.102003.152644

Durot M., Bourguignon P.-Y., Schachter V. Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS Microbiol. Rev. 2009;33:164-190. DOI 10.1111/j.1574-6976.2008.00146.x

Emonet T., Macal C.M., North M.J., Wickersham C.E., Cluzel P. AgentCell: a digital single-cell assay for bacterial chemotaxis. Bioinformatics. 2005;21:2714-2721. DOI 10.1093/bioinformatics/bti391

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 J.M. Ferràndez, J.R. Álvarez Sànchez, F. de la Paz, F.J. Toledo. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. DOI 10.1007/978-3-642-21326-7

Faust K., Raes J. Microbial interactions: from networks to models. Nat. Rev. Microbiol. 2012;10:538-550. DOI 10.1038/nrmicro2832

Frey E. Evolutionary game theory: Theoretical concepts and applications to microbial communities. Phys. A Stat. Mech. its Appl. 2010; 389:4265-4298. DOI 10.1016/j.physa.2010.02.047

Fuhrman J.A. Microbial community structure and its functional implications. Nature. 2009;459:193-199. DOI nature08058 [pii]\n10.1038/nature08058 [doi]

Gimel'farb A.A., Ginzburg L.R., Poluektov R.A., Pykh Yu.A., Ratner V.A. Dinamicheskaya teoriya biologicheskikh populyatsiy [Dynamic theory of biological populations]. Moscow, Nauka, 1974.

Ginovart M., López D., Valls J. INDISIM, an individual-based discrete simulation model to study bacterial cultures. J. Theor. Biol. 2002; 214:305-319. DOI 10.1006/jtbi.2001.2466

Grimm V., Berger U., Bastiansen F., Eliassen S., Ginot V., Giske J., Goss-Custard J., Grand T., Heinz S.K., Huse G., Huth A., Jepsen J.U., Jørgensen C., Mooij W.M., Müller B., Pe'er G., Piou C., Railsback S.F., Robbins A.M., Robbins M.M., Rossmanith E., Rüger N., Strand E., Souissi S., Stillman R. a., Vabø R., Visser U., DeAngelis D.L. A standard protocol for describing individual-based and agent-based models. Ecol. Modell. 2006;198:115-126. DOI 10.1016/j.ecolmodel.2006.04.023

Halfen L.N., Castenholz R.W. Gliding motility in the blue-green alga oscillatoria princeps. 1971.

Hecker M., Lambeck S., Toepfer S., van Someren E., Guthke R. Gene regulatory network inference: Data integration in dynamic models–A review. Biosystems. 2009;96:86-103. DOI 10.1016/j.biosystems.2008.12.004

Henrichsen J. Bacterial Surface Translocation: a Survey and a Classification. Bacteriol. Rev. 1972;36:478-503.

Henson M.A., Hanly T.J. Dynamic flux balance analysis for synthetic microbial communities. IET Syst. Biol. 2014;8:214-229. DOI 10.1049/iet-syb.2013.0021

Ishii N., Robert M., Nakayama Y., Kanai A., Tomita M. Toward large-scale modeling of the microbial cell for computer simulation. J. Biotechnol. 2004;113:281-294. DOI 10.1016/j.jbiotec.2004.04.038

Karr J.R., Sanghvi J.C., MacKlin D.N., Gutschow M.V., Jacobs J.M., Bolival B., Assad-Garcia N., Glass J.I., Covert M.W. A whole-cell computational model predicts phenotype from genotype. Cell. 2012;150:389-401. DOI 10.1016/j.cell.2012.05.044

Karunakaran E., Mukherjee J., Ramalingam B., Biggs C.A. «Biofilmology»: a multidisciplinary review of the study of microbial biofilms. Appl. Microbiol. Biotechnol. 2011;90:1869-1881. DOI 10.1007/s00253-011-3293-4

Klimenko A.I., Matushkin Y.G., Kolchanov N.A., Lashin S.A. Modeling evolution of spatially distributed bacterial communities: a simulation with the haploid evolutionary constructor. BMC Evol. Biol. 2015;15:S3. DOI 10.1186/1471-2148-15-S1-S3

Klitgord N., Segre D. Environments that induce synthetic microbial ecosystems. PLoS Comput. Biol. 2010;101:1435-1439. DOI 10.1371/Citation

Knibbe C., Fayard J.-M., Beslon G. The topology of the protein network influences the dynamics of gene order: from systems biology to a systemic understanding of evolution. Artif. Life. 2008;14:149-156. DOI 10.1162/artl.2008.14.1.149

Kolmakova O.V. Contemporary Methods for Identification of Bacterioplankton Species-Specific Biogeochemical Functions. Zhurnal sibirskogo federalnogo universiteta. Biologiya = Journal of Siberian Federal University. Biology. 2013;6(1):73-95.

Kutalik Z., Razaz M., Baranyi J. Connection between stochastic and deterministic modelling of microbial growth. J. Theor. Biol. 2005;232:285-299. DOI 10.1016/j.jtbi.2004.08.013

Larsen P., Hamada Y., Gilbert J. Modeling microbial communities: Current, developing, and future technologies for predicting microbial community interaction. J. Biotechnol. 2012;160:17-24. DOI 10.1016/j.jbiotec.2012.03.009

Laspidou C.S., Rittmann B.E. Evaluating trends in biofilm density using the UMCCA model. Water Res. 2004;38:3362-33672. DOI 10.1016/j.watres.2004.04.051

Lencastre Fernandes R., Nierychlo M., Lundin L., Pedersen A.E., Puentes Tellez P.E., Dutta A., Carlquist M., Bolic A., Schäpper D., Brunetti A.C., Helmark S., Heins A.L., Jensen A.D., Nopens I., Rottwitt K., Szita N., van Elsas J.D., Nielsen P.H., Martinussen J., Sørensen S.J., Lantz A.E., Gernaey K.V. Experimental methods and modeling techniques for description of cell population heterogeneity. Biotechnol. Adv. 2011;29:575-599. DOI 10.1016/j.biotechadv.2011.03.007

Leslie P.H. On the use of matrices in certain population mathematics. Biometrika. 1945. DOI 10.2307/2332297

Likhoshvay V.A., Khlebodarova T.M., Ratushnyy A.V., Lashin S.A., Turnaev I.I., Podkolodnaya O.A., Anan'ko E.A., Smirnova O.G., Ibragimova S.S., Kolchanov N.A. Kompyuternyy geneticheskiy konstruktor: matematicheskoe modelirovanie geneticheskikh i metabolicheskikh podsistem *E. soli* [Computer-based genetic constructor: mathematical modelling of genetic and metabolic subsystems of *E. coli*]. Rol mikroorganizmov v funktsionirovanii zhivykh sistem: fundamentalnye problemy i bioinzhenernye prilozheniya [Role of Microorganisms in Living Systems Functioning: Fundamental Problems and Bioengineering Applications]. Eds: Vlasov V.V., Degermendzhi A.G., Kolchanov N.A., Parmon V.N., Repin E.A. Novosibirsk, Izdatel'stvo SO RAN, 2010:376-391.

Likhoshvai V.A., Ratushny A.V. Generalized Hill function method for modeling molecular processes. J. Bioinform. Comput. Biol. 2007;05: 521-531. DOI 10.1142/S0219720007002837

Logofet, D.O., Belova, I.N., 2007. Nonnegative matrices as a tool to model population dynamics: Classical models and contemporary expansions. Fundamentalnaya i prikladnaya matematika=Fundamental and Applied Mathematics. 2007;13:145-164. DOI:10.1007/s10958-008-9249-2

Mahadevan R., Henson M.A. Genome-based modeling and design of metabolic interactions in microbial communities. Comput. Struct. Biotechnol. J. 2012;3:1-7. DOI 10.5936/csbj.201210008

Mburu N., Rousseau D.P.L., Stein O.R., Lens P.N.L. Simulation of batch-operated experimental wetland mesocosms in AQUASIM biofilm reactor compartment. J. Environ. Manage. 2014;134:100-108. DOI 10.1016/j.jenvman.2014.01.005

Monod J. La technique de culture continue. Theorie et applications. Ann. Inst. Pasteur. 1950;79:391-410.

Netrusov A.I., Kotova I.B., 2007. Mikrobiologiya [Microbiology]. Moscow, Akademiya, 2007.

Niu B., Wang H., Duan Q., Li L. Biomimicry of quorum sensing using bacterial lifecycle model. BMC Bioinformatics. 2013;14(Suppl. 8): S8. DOI 10.1186/1471-2105-14-S8-S8

O'Donnell A.G., Young I.M., Rushton S.P., Shirley M.D., Crawford J.W. Visualization, modelling and prediction in soil microbiology. Nat. Rev. Microbiol. 2007;5:689-699. DOI 10.1038/nrmicro1714

Oberhardt M.A., Palsson B.Ø., Papin J.A. Applications of genome-scale metabolic reconstructions. Mol. Syst. Biol. 2009;5. DOI 10.1038/msb.2009.77

Pfeiffer T., Schuster S. Game-theoretical approaches to studying the evolution of biochemical systems. Trends Biochem. Sci. 2005;30: 20-25. DOI 10.1016/j.tibs.2004.11.006

Price N.D., Reed J.L., Palsson B.Ø. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat. Rev. Microbiol. 2004;2:886-897. DOI 10.1038/nrmicro1023

Ramkrishna D. Population Balances: Theory and Applications to Particulate Systems in Engineering, Chemical Engineering. 2000.

Riznichenko G.Yu. Matematicheskie modeli v biofizike i ekologii [Mathematical models in biophysics and ecology]. Moscow; Izhevsk, Institut kompyuternykh issledovaniy, 2003.

Riznichenko G.Yu., Rubin A.B. Matematicheskie modeli biologicheskikh produktsionnykh protsessov [Mathematical models of biological production processes]. Moscow, MGU Publ., 1993.

Rudge T.J., Steiner P.J., Phillips A., Haseloff J. Computational modeling of synthetic microbial biofilms. ACS Synthetic Biology 2012;1(8): 345-352. DOI 10.1021/sb300031n

Salli K.M., Ouwehand A.C. The use of in vitro model systems to study dental biofilms associated with caries: a short review. J. Oral Microbiol. 2015;7. DOI 10.3402/jom.v7.26149

Sauer U., Heinemann M., Zamboni N. GENETICS: getting closer to the whole picture. Science. 2007;316:550-551. DOI 10.1126/science.1142502

Scheffer M., Baveco J.M., DeAngelis D.L., Rose K.A., van Nes E.H. Super-individuals a simple solution for modelling large populations on an individual basis. Ecol. Modell. 1995;80:161-170. DOI 10.1016/0304-3800(94)00055-M

Scheibe T.D., Mahadevan R., Fang Y., Garg S., Long P.E., Lovley D.R. Coupling a genome-scale metabolic model with a reactive transport model to describe in situ uranium bioremediation. Microb. Biotechnol. 2009;2:274-286. DOI 10.1111/j.1751-7915.2009.00087.x

Schuster S., Fell D.A., Dandekar T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat. Biotechnol. 2000;18:326-332. DOI 10.1038/73786

Segrè D., Vitkup D., Church G.M. Analysis of optimality in natural and perturbed metabolic networks. Proc. Natl Acad. Sci. USA. 2002;99: 15112-15117. DOI 10.1073/pnas.232349399

Shrout J.D. A fantastic voyage for sliding bacteria. Trends Microbiol. 2015;23:244-246. DOI 10.1016/j.tim.2015.03.001

Song H.-S., Cannon W., Beliaev A., Konopka A. Mathematical modeling of microbial community dynamics: a methodological review. Processes. 2014;2:711-752. DOI 10.3390/pr2040711

Stauffer D., Kunwar A., Chowdhury D. Evolutionary ecology in silico: Evolving food webs, migrating population and speciation. Physica A. 2005;352:202-215. DOI 10.1016/j.physa.2004.12.036

Tang Y., Valocchi A.J. An improved cellular automaton method to model multispecies biofilms. Water Res. 2013;47:5729-5742. DOI 10.1016/j.watres.2013.06.055

Tindall M.J., Maini P.K., Porter S.L., Armitage J.P. Overview of mathematical approaches used to model bacterial chemotaxis II: Bacterial populations. Bull. Math. Biol. 2008a. DOI 10.1007/s11538-008-9322-5

Tindall M.J., Porter S.L., Maini P.K., Gaglia G., Armitage J.P. Overview of mathematical approaches used to model bacterial chemotaxis I: The single cell. Bull. Math. Biol. 2008b. DOI 10.1007/s11538-008-9321-6

Tomita M. Whole-cell simulation: a grand challenge of the 21st century. Trends Biotechnol. 2001;19:205-210. DOI 10.1016/S0167-7799(01)01636-5

Tomita M., Hashimoto K., Takahashi K., Shimizu T., Matsuzaki Y., Miyoshi F., Saito K., Tanida S., Yugi K., Venter J., Hutchison C. E-CELL: software environment for whole-cell simulation. Bioinformatics. 1999;15:72-84. DOI 10.1093/bioinformatics/15.1.72

Turing A.M. The chemical theory of morphogenesis. Phil. Trans. Roy. Soc. 1952;13:1.

Wanner O., Morgenroth E. Biofilm modeling with AQUASIM. Water Sci. Technol. 2004;49:137-144.

Wimpenny J., Manz W., Szewzyk U. Heterogeneity in biofilms. FEMS Microbiol. Rev. 2000. DOI 10.1016/S0168-6445(00)00052-8

Wimpenny J.W.T., Colasanti R. A unifying hypothesis for the structure of microbial biofilms based on cellular automaton models. FEMS Microbiol. Ecol. 1997. DOI 10.1016/S0168-6496(96)00078-5

Wolfe B.E., Dutton R.J. Review fermented foods as experimentally tractable microbial ecosystems. Cell. 2015;161:49-55. DOI 10.1016/j.cell.2015.02.034

Wooley J.C., Godzik A., Friedberg I. A primer on metagenomics. PLoS Comput. Biol. 2010. DOI 10.1371/journal.pcbi.1000667

Zomorrodi A.R., Islam M.M., Maranas C.D. D-OptCom: dynamic multi-level and multi-objective metabolic modeling of microbial communities. ACS Synth. Biol. 2014;3:247-257. DOI 10.1021/sb4001307

Zomorrodi A.R., Maranas C.D. OptCom: A multi-level optimization framework for the metabolic modeling and analysis of microbial communities. PLoS Comput. Biol. 2012;8. DOI 10.1371/journal.pcbi.1002363