На правах рукописи

Казнадзей Анна Денисовна

Геномная ко-локализация генов углеводного метаболизма бактерий

03.01.09 – математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

Москва – 2019

Работа выполнена в Учебно-научном центре «Биоинформатика» Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук.

Научный руководитель:

Гельфанд Михаил Сергеевич
кандидат физико-математических наук, доктор биологических наук, профессор
(Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации имени А.А. Харкевича Российской академии наук)

Официальные оппоненты:

Озолинь Ольга Николаевна
доктор биологических наук, профессор, зав. лабораторией
(Институт биофизики клетки Российской академии наук – обособленное подразделение Федерального государственного бюджетного учреждения науки «Федеральный исследовательский центр «Пущинский научный центр биологических исследований Российской академии наук», лаборатория функциональной геномики и клеточного стресса)

Мошковский Сергей Александрович
доктор биологических наук, профессор, зав. кафедрой
(Федеральное государственное бюджетное образовательное учреждение высшего образования "Российский национальный исследовательский медицинский университет имени Н.И. Пирогова" Министерства здравоохранения Российской Федерации, кафедра биохимии медико-биологического факультета)

Ведущая организация:

Федеральное государственное бюджетное учреждение науки Институт общей генетики им. Н.И. Вавилова Российской академии наук

Защита диссертации состоится «___» _____ 2019 года в_____на заседании диссертационного совета Д 002.077.04 при Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации имени А.А. Харкевича Российской академии наук (ИППИ РАН) по адресу: 127051, г. Москва, Большой Каретный переулок, д. 19, стр. 1.

С диссертацией можно ознакомиться в библиотеке ИППИ РАН, а также на сайте ИППИ РАН по адресу: **http://iitp.ru/upload/content/.....**

Автореферат разослан «___» _____ 2019 г.

Ученый секретарь диссертационного совета
доктор биологических наук, профессор Рожкова Г.И.

Общая характеристика работы

Актуальность темы исследования.

Одним из важнейших объектов современных исследований являются бактериальные геномы. Бактерии способны приспосабливаться к самым разным условиям среды и, в частности, катаболизировать широкий спектр углеводов. Белки, участвующие в соответствующих процессах, закодированы в бактериальных генах; исследования, касающиеся структуры, функций и регуляции работы таких генов, а также их сочетаний, ведутся уже несколько десятков лет. Так, лактозный оперон (*lac*-оперон) кишечной палочки, состоящий из трех генов, стал первым описанным опероном прокариот; за эту работу исследователи Ф. Жакоб и Ж. Моно получили в 1965 году Нобелевскую премию. До сих пор, однако, не было проведено масштабных исследований, касающихся общих тенденций взаиморасположения генов углеводного метаболизма в бактериальных геномах и факторов, влияющих на эти тенденции. Такие исследования позволят лучше понимать механизмы адаптации бактерий к разнообразным и меняющимся условиям среды. В целом, бактериальная геномика является важным направлением современной биологии, так как бактерии обладают большим генетическим и метаболическим разнообразием, а также имеют существенное прикладное значение в медицине, сельском хозяйстве и биотехнологии.

Изучение консервативных сочетаний генов в бактериальных геномах методами сравнительной геномики может позволять делать успешные предсказания о свойствах кодируемых ими белков. Экспериментальная проверка подобных предсказаний важна с точки зрения соотношения теоретических и практических знаний и, помимо получения новых данных о функциях белков, вносит существенный вклад в понимание эволюционного значения геномного окружения генов. Кроме того, выявление новых, ранее неизвестных функций в дополнение к уже описанным позволяет затронуть малоизученный вопрос о мультифункциональных особенностях бактериальных белков.

Одним из распространенных методов сравнительной геномики является сравнение нуклеотидных последовательностей. С развитием технологий секвенирования в последние годы количество данных о последовательностях ДНК растет с огромной скоростью. При этом задачи, связанные со сравнением нуклеотидных последовательностей, не характеризующихся очень высоким уровнем сходства (филогенетический анализ, осуществление структурных и функциональных предсказаний и т.п.), по-прежнему решаются либо с помощью чувствительных и медленных, либо с помощью быстрых и малочувствительных алгоритмов. В результате либо время работы соответствующих инструментов оказывается неприемлемо долгим, либо в ходе поиска теряется значительная часть результатов. Таким образом, актуальной на данный момент является разработка быстрых, но при этом точных и чувствительных методов сравнения последовательностей ДНК удаленного сходства; в рамках исследования геномных локусов углеводного метаболизма такие методы

необходимы, в частности, для выявления причин ко-локализации генов.

Цели и задачи исследования. Целью работы было выяснить, как организованы геномные локусы бактерий, содержащие гены углеводного метаболизма, какие факторы влияют на эту организацию, какие эволюционные механизмы стоят в ее основе, и как можно использовать данные о ко-локализации этих генов для предсказания их функций.

Были поставлены следующие задачи:

1. Оценить, как часто гены углеводного метаболизма располагаются на бактериальных хромосомах рядом, т.е. формируют в геномах кассеты, и как часто они располагаются по отдельности, а также описать разнообразие кассет.

2. Выяснить, как функциональные и структурные характеристики кодируемого белка влияют на склонность соответствующего гена к формированию кассет, а также как склонность к формированию кассет варьирует среди разных таксонов бактерий.

3. Оценить тенденции к ко-локализации генов разных функций и тенденции к ко-локализации генов сходных функций.

4. Разработать инструмент, позволяющий эффективно оценивать уровень сходства нуклеотидных последовательностей, различающихся на 10% и более, и применить этот инструмент для оценки вклада событий локальной дупликации в ко-локализацию генов сходных функций.

5. Применить результаты анализа тенденций ко-локализации генов углеводного метаболизма для конкретного предсказания функций генов с последующей проверкой.

Научная новизна и практическая ценность.

В работе рассмотрены актуальные вопросы и решен ряд задач современной сравнительной геномики. Полученные в данной работе результаты в первую очередь имеют фундаментальное значение, поскольку они позволяют понять эволюционное значение геномного окружения генов углеводного метаболизма в широком спектре бактериальных видов. Выявлены основные факторы, влияющие на формирование кассет этих генов. Исследованы тенденции попарных сочетаний генов разных функциональных классов и разных ортологических кластеров, а также тенденции ко-локализации генов сходных функций. Выявлен вклад в такие случаи событий локальной дупликации.

Выдвинута гипотеза о том, что сравнительный анализ сочетаний функций генов углеводного метаболизма внутри кассет может позволять предсказывать общую функцию кассеты и ее участие в соответствующем метаболическом пути. Гипотеза подтверждена для кассеты генов *Escherichia coli*, совпадающей по общему функциональному составу с консервативной кассетой, участвующей в катаболизме

лактозы у бактерий класса Bacilli. Впервые, таким образом, описан альтернативный путь катаболизма лактозы у кишечной палочки (в дополнение к известному пути, ферменты которого закодированы в *lac*-опероне, описанном Ф. Жакобом и Ж. Моно). Предсказаны мультифункциональные характеристики соответствующих белков. Также впервые были картированы промоторы генов указанной кассеты кишечной палочки и описан механизм переключения регуляции их экспрессии при росте на разных источниках углерода.

Разработан и программно реализован биоинформатический инструмент, позволяющий проводить поиск заданных нуклеотидных последовательностей удаленного сходства в больших базах данных ДНК, который по сочетанию основных параметров работы превосходит инструменты, считающиеся индустриальным стандартом.

Положения, выносимые на защиту

Разработан инструмент NSimScan для поиска нуклеотидных последовательностей удаленного сходства; наилучшим образом он подходит для поиска последовательностей, различающихся на 60-90%. По совокупности таких параметров как чувствительность, частота ошибок и скорость он превосходит все стандартные инструменты в своей области.

Описана сеть эволюционных связей 148 тысяч генов углеводного метаболизма 665 видов бактерий, выраженная в форме их ко-локализационных тенденций. 53% таких генов находятся в составе кассет, то есть ко-локализованы, остальные располагаются на бактериальных геномах по отдельности. Склонность к формированию кассет различается у разных генов; ключевыми факторами, влияющими на их ко-локализационные тенденции, являются функциональные и структурные характеристики гена и таксономическая принадлежность соответствующей бактерии. Склонность к формированию кассет у разных функциональных классов составляет от 23 до 93%; у разных кластеров ортологических групп генов – 0 до 100%, у разных бактериальных классов – от 40 до 76%.

Функциональные классы могут формировать консервативные и, по всей видимости, эволюционно значимые ко-локализационные связи; всего описано 45 таких связей для 19 исследуемых классов. Количество связей для каждого класса сильно варьирует, что указывает на существенное различие в предпочтениях к непосредственному геномному окружению у генов разных функций. Для 11 классов гены одного и того же класса также демонстрируют выраженное предпочтение к ко-локализации, причем большинство таких случаев, по-видимому, не являются результатом событий локальной дупликации.

Исследование консервативных комбинаций внутри кассет генов углеводного метаболизма позволяет успешно предсказывать их функции. На основании сходства консервативной кассеты генов бактерий семейства Enterobacteriaceae, отвечающей за

катаболизм серосодержащих сахаров, с консервативной кассетой бактерий класса Bacilli, участвующей в катаболизме лактозы, предсказана и экспериментально подтверждена роль кассеты *Escherichia coli* в утилизации лактозы. Описан, таким образом, ранее неизвестный путь катаболизма лактозы у кишечной палочки и предсказаны мультифункциональные характеристики соответствующих белков. В переключении механизмов экспрессии генов этой кассеты при смене источника углерода в среде участвуют локальный регулятор YihW и глобальный регулятор CRP.

Степень достоверности и апробация результатов. По материалам диссертации опубликовано три статьи в рецензируемых научных журналах, индексируемых Web of Science. Результаты работы были представлены на международных конференциях по вычислительной молекулярной биологии (Moscow Conference on Computational Molecular Biology – MCCMB’11, MCCMB’13 – Москва) и по системной биологии (Bioinformatics of Genome Regulation and Structure\Systems Biology BGRS\SB-2012 – Новосибирск); а также на конференциях «Информационные технологии и системы» (ИТиС'10, ИТиС'11 - Геленджик, ИТиС'12 – Петрозаводск, ИТиС'15 - Сочи, ИТиС'16 – Санкт-Петербург).

Структура и объем работы. Диссертация изложена на 145 страницах. Она состоит из 4 глав: "Литературный обзор", "Инструмент NSimScan для сравнения последовательностей ДНК удаленного сходства", "Организация генов углеводного метаболизма бактерий", и "Участие *yih*-кассеты *Escherichia coli* в катаболизме лактозы". Работа содержит 21 рисунок и 3 таблицы. В конце приведен список литературы, содержащий 117 ссылок, и приложение, содержащее 4 таблицы.

Содержание работы

Глава 1 содержит обзор литературы по теме диссертации, разделенный на три части.

Первая часть включает описание ранних и современных методов для сравнения нуклеотидных последовательностей. В целом они подразделяются на алгоритмы, направленные на поиск очень сходных последовательностей, необходимые, например, для картирования прочтений, полученных в результате секвенирования - такие, как BowTie, BWA и другие, и на алгоритмы, направленные на поиск последовательностей более удаленного сходства, необходимые, например, для проведения филогенетического анализа, а также для осуществления структурных и функциональных предсказаний. Последние, в свою очередь, подразделяются на чувствительные и точные алгоритмы, которые находят большую часть или все нуклеотидные последовательности по заданному порогу, но работают медленно (например, SSearch), и на быстрые алгоритмы, которые имеют при этом существенные

ограничения по чувствительности и точности (например, BLAT или MegaBLAST).

Вторая часть литературного обзора посвящена описанию известных тенденций и причин ко-локализации бактериальных генов, а также общим свойствам генов углеводного метаболизма бактерий. Последний отличается большим разнообразием, поскольку самые разные углеводы служат бактериям источниками энергии и углерода. Углеводы также участвуют во множестве ключевых клеточных процессов и являются важным структурным элементом клетки. Белки, участвующие в соответствующих метаболических путях, закодированы в бактериальных генах. В обзоре описаны предыдущие работы, связанные с исследованием ко-локализации генов в бактериальных геномах, в том числе, касающиеся оперонных структур.

Другие работы описывали соотношение эволюционных и функциональных модулей генов на больших выборках бактериальных геномов. С одной стороны, известно, что белки, входящие в общий метаболический путь или физически взаимодействующие друг с другом, часто закодированы поблизости друг от друга. С другой стороны, это не является обязательным правилом, и, более того, гены внутри консервативных, эволюционно закрепленных комбинаций не всегда связаны функционально. Анализ таких комбинаций, тем не менее, может стать важным шагом предсказания функций генов.

Заключительная часть диссертационной работы посвящена конкретному такому предсказанию и его экспериментальной проверке у *Escherichia coli*, и в третьей части литературного обзора описываются соответственные метаболические пути, а именно, способы утилизации лактозы у представителей класса Bacilli и семейства Enterobacteriaceae.

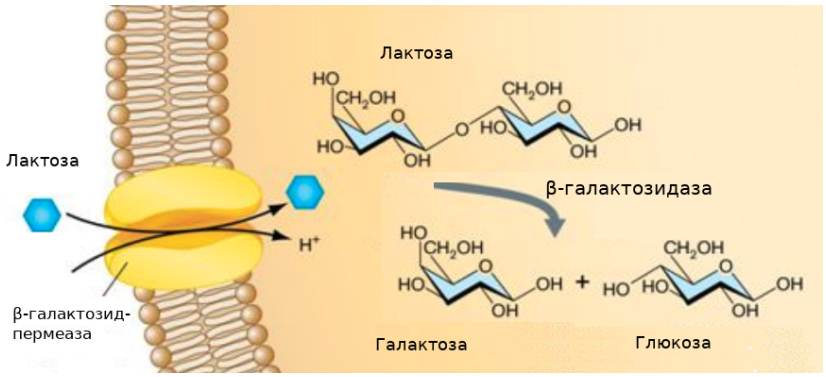

Рис. 1. Лактоза переносится через клеточную мембрану *E. coli* β-галактозидпермеазой одновременно с протоном, после чего расщепляется на глюкозу и галактозу с помощью β-галактозидазы.

У кишечной палочки и родственных ей бактерий известен только один путь катаболизма лактозы, а именно, происходящий с помощью β-галактозидпермеазы и β-галактозидазы, закодированных в первом описанном и хорошо на данный момент изученном бактериальном опероне (Рис. 1)

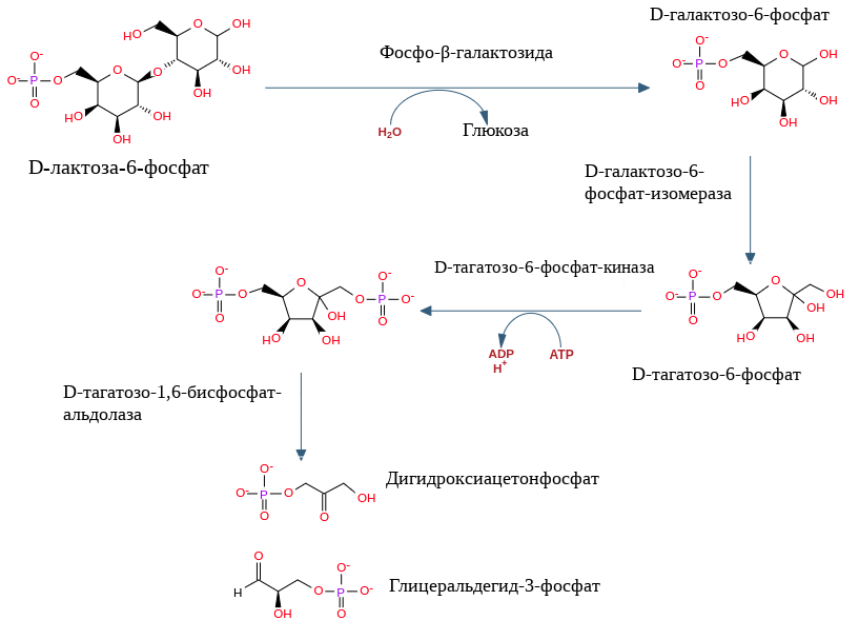

Рис. 2. Схема пути катаболизма лактозы бактерий класса Bacilli

В случае класса Bacilli путь устроен несколько сложнее: лактоза доставляется в клетку и сразу фосфорилируется с помощью транспортной системы PTS, после она подвергается последовательной трансформации под воздействием фосфо-β-галактозидазы, D-галактозо-6-фосфат-изомеразы, D-тагатозо-6-фосфат-киназы и D-тагатозо-1,6-бисфосфат-альдолазы (Рис. 2). По общему функциональному составу (транспортер, гидролаза, изомераза, киназа и альдолаза) кассета, отвечающая за катаболизм лактозы у Bacilli имеет пересечение с кассетой, участвующей в катаболизме сульфосодержащих сахаров у *Escherichia coli*. (Рис. 3)

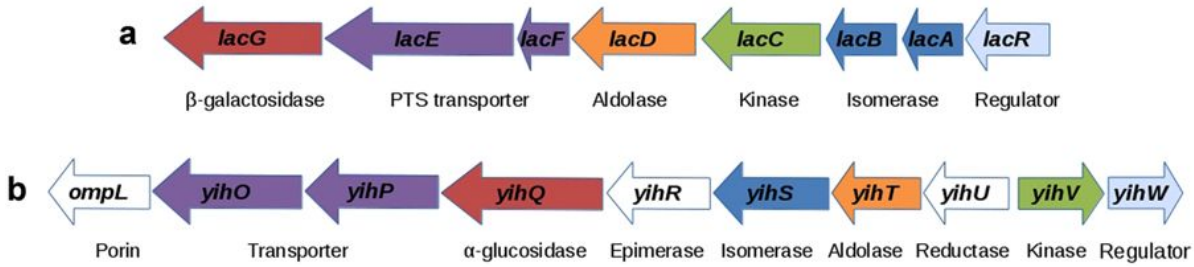

Рис. 3. Кассета бактерий класса Bacilli, участвующая в катаболизме лактозы (a) и кассета семейства Enterobacteriaceae, участвующая в сульфогликолизе (b). Одинаковые цвета генов обозначают пересечение функций кодируемых белков. Белым отмечены гены, кодирующие функции, не представленные в другой кассете.

В последней части обзора также описаны механизмы регуляции работы генов, связанных с метаболизмом углеводов, в том числе, механизм катаболитной репрессии. Регуляция экспрессии генов углеводного метаболизма часто осуществляется за счет

локального и глобального факторов транскрипции, которые могут обладать как противоположными, так и взаимодополняющими функциями.

Глава 2 содержит описание алгоритма работы и оценки эффективности программы NSimScan (Nucleotide Similarity Scanner), разработанной для поиска сходных нуклеотидных последовательностей в больших базах данных ДНК. Он предназначается, в том числе, для проведения филогенетического анализа, предсказания функций генов и для других сравнительных исследований, а также для исследования некодирующих последовательностей и детекции загрязнения образцов ДНК.

NSimScan представляет собой генератор предполагаемых участков сходства (первичных совпадений), объединенных с серией фильтров с увеличивающейся вычислительной нагрузкой. Ниже приведен сокращенный алгоритм его работы.

При запуске NSimScan сначала прочитывает все искомые последовательности (queries) и составляет индексную таблицу, в которой хранит координаты всех k-меров (оптимальная длина – от 8 до 12 нуклеотидов) каждой последовательности. Таблица адресуется непосредственно двоично упакованным представлением последовательности k-мера. Это весьма существенный момент, отличающий NSimScan от других программ и позволяющий значительно ускорять процедуру поиска в таблице.

В памяти выделяется место для списка диагоналей матрицы совпадений. Последовательно прочитываются все нуклеотидные последовательности ("записи") из базы данных, в которой осуществляется поиск. Для каждой позиции из каждой записи проверяется наличие соответствующего k-мера в индексной таблице. Первичные вхождения используются для обновления значений весов диагоналей матрицы сходства у каждой позиции из записи.

Когда вес диагонали превышает заданный порог, диагональ успешно проходит фильтр оценки выравнивания. Для прошедших первичный фильтр диагоналей далее строится субоптимальное выравнивание. Для этого используется жадный эвристический алгоритм, который составляет выравнивание за одно прохождение по диагонали путем последовательного расширения зоны сходства по текущей и нескольким соседним диагоналям в обоих направлениях, пока вес выравнивания остается положительным. Это очень быстрая процедура, поскольку она является линейной – ее скорость зависит только от длины выравнивания. Высокая эффективность процедуры также достигается благодаря тому, что выравнивание осуществляется с помощью битовых операций над упакованными последовательностями (искомые последовательности и записи из базы данных представлены в бинарном виде).

Полученные выравнивания пропускаются через фильтр соотношения длины и процента сходства. Поскольку оценка параметров проводится каждый раз, когда вес диагонали оказывается выше порога, для некоторых позиций выстраивается серия

длинных и относительно хороших выравниваний, которые могут несколько различаться в силу эвристического алгоритма их построения. Из таких перекрывающихся выравниваний выбирается одно, обладающее наибольшим весом.

Основными задаваемыми параметрами программы, таким образом, являются размер k-мера (первичного вхождения), весовой порог диагонали (первичный фильтр) и параметры вторичного фильтра для выравнивания: минимальная длина выравнивания, минимальная доля сходства на минимальной длине и минимальная доля сходства на полной длине.

Программа NSimScan доступна для скачивания на сайте https://github.com/abadona/qsimscan. Более подробные алгоритм работы, руководство по применению и примеры параметров для командной строки находятся по адресу https://github.com/abadona/qsimscan/blob/master/nsimscan_users_guide.txt.

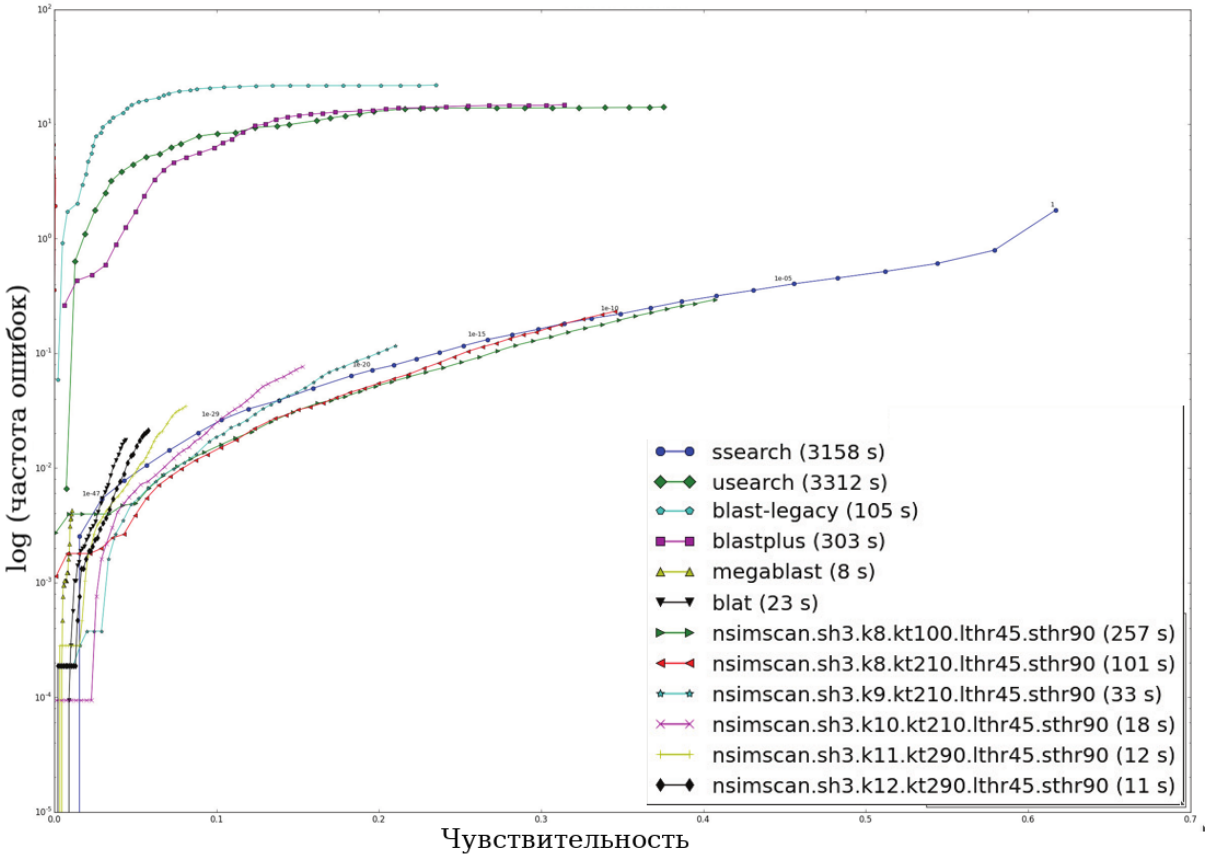

Рис. 5. Эффективность работы NSimScan, представленная в виде частоты ошибок поиска (по вертикали, log) относительно его чувствительности (по горизонтали) при разных параметрах и в сравнении с SSearch, USEARCH, BLAT, BLAST (BLAST plus и Legacy BLAST) и MegaBLAST. Параметры NSimScan: *sh* – максимальный сдвиг по диагонали; *k* – размер k-мера; *kt* – порог веса диагонали; *lthr* – наименьшее совпадение на полной длине выравнивания; (*sthr* – наименьшее совпадение на мин. длине выравнивания, 40 нуклеотидах). В скобках указано время работы инструмента в секундах.

Для проверки эффективности работы мы провели два исследования. Первое включало 10600 последовательностей 53 семейства генов, кодирующих рибосомные

белки. Каждого представителя искали против остальной выборки с помощью стандартных в области инструментов BLAST, MegaBLAST, BLAT, USEARCH и SSearch. Совпадение с представителями своего семейства считалось истинно-положительными результатом (TP), совпадение с представителями чужих семейств – ложно-положительным (FP), отсутствие совпадения между членами одного семейства – ложно-отрицательным (FN), а отсутствие совпадения между членами разных семейств – истинно-отрицательным (TN).

Для полученных совпадений, отсортированных по ожидаемым значениям, мы вычислили частоту ошибок на каждый поиск – FP/(количество искомых последовательностей) и чувствительность, которая также называется покрытием – TP/(TP + FN). Соответствующие данные представлены в виде графика на Рис. 5.

Во всех вариантах заданных условий NSimScan по соотношению чувствительности и частоты ошибок оказался на уровне SSearch (который считается самым чувствительным инструментом среди указанных), обогнав при этом все остальные инструменты на два порядка. Даже с наименее жесткими первичными параметрами поиска NSimScan работает с несколько большей чувствительностью, чем USEARCH (который демонстрирует наилучшие показатели по чувствительности среди остальных инструментов, кроме SSearch), при этом частота ошибок у NSimScan почти на два порядка ниже, а скорость выше на порядок. По скорости NSimScan сопоставим с MegaBLAST при средних порогах чувствительности и с BLAST при высоких порогах. Он не уступает уровню BLAST по чувствительности, при этом скорость работы NSimScan при соответствующих параметрах оказывается в три раза выше.

Мы также протестировали скорость работы NSimScan на другой модельной задаче – в рамках филогенетического анализа большого набора метагеномных данных. Для этого мы провели поиск репрезентативных последовательностей 16S РНК для 749 таксонов против образца метагенома корней огурца (Таблица 1)

Tool	Time	MemUse	Detected	Tx#	MissTx#	ExtraTx#
MegaBLAST	5 h 18 min	24.6 Gb	252956	310	n/a	n/a
NSimScan	26 min	5.7 Gb	240934	360	4.7%	21.6%

Таблица 1. Сравнение работы NSimScan и MegaBLAST в рамках филогенетического анализа. В колонках: Tool – инструмент, Time – время работы инструмента; MemUse – количество использованной оперативной памяти; Detected – количество обнаруженных фрагментов 16S РНК; Tx# – количество выявленных бактериальных таксонов, MissTx# – количество таксонов, которые обнаружены с помощью MegaBLAST, но не обнаружены с помощью NSimScan, ExtraTx# – количество таксонов, которые обнаружены с помощью NSimScan, но не обнаружены с помощью MegaBLAST.

NSimScan работал в 10 раз быстрее, чем MegaBLAST, использовал в 4 раза меньше

оперативной памяти и обнаружил практически все искомые фрагменты – более 95% таксонов, которые нашел MegaBLAST и существенное количество таксонов (21,6%), которые MegaBLAST не обнаружил.

Мы показали, что NSimScan превосходит по производительности программы, считающиеся индустриальным стандартом, по совокупности таких параметров, как чувствительность, частота ошибок и скорость. В целом, по чувствительности NSimScan сравним с BLAST и USEARCH, по частоте ошибок – с SSearch, а по скорости – с MegaBLAST. Наилучшим образом он подходит для поиска последовательностей, отличающихся друг от друга на 10-40 процентов.

NsimScan был успешно использован исследовательской группой из калифорнийского института Joint Genomic Institute для вычисления УНР (усредненного нуклеотидного расстояния) в опубликованном широкомасштабном филогенетическом исследовании, включающем данные геномов 3032 видов прокариот [Varghese *et al*, 2015]. Мы рассчитываем, что эта программа будет полезна и в других, самых разнообразных проектах, требующих эффективного обнаружения нуклеотидных последовательностей удаленного сходства.

Третья глава посвящена анализу ко-локализационных тенденций генов углеводного метаболизма бактерий.

Склонность генов к ко-локализации и разнообразие кассет генов

Всего было изучено 665 бактериальных геномов разных видов, общее количество генов углеводного метаболизма в которых составило 148 тысяч. Мы использовали двухуровневую классификацию генов. Первый уровень, классы, соответствовал глобальной функции гена и учитывал реакционную и общую субстратную специфичность соответствующих ферментов. В отдельные два класса были выделены траспортеры и транскрипционные факторы. Второй уровень соответствовал структурно-эволюционным характеристикам гена, отраженным в его принадлежности к определенному COG (кластеру групп ортологических генов). Кассеты генов определялись на основании ко-локализации генов на бактериальных хромосомах. Считалось, что гены углеводного метаболизма формируют кассету, если они располагаются на хромосоме подряд, причем расстояние между каждой парой не превышает 200 нуклеотидов; допускался один пропуск длиной 1500 нуклеотидов.

Выяснилось, что только 53% генов углеводного метаболизма входили в состав кассет. Изначально мы ожидали увидеть более сильную тенденцию к ко-локализации у генов, белки которых потенциально выполняют взаимосвязанные функции [Ogata *et al*, 2000; Glazko *et al*, 2004]. Известно, однако, что эволюционные модули, состоящие из групп генов, всегда одновременно присутствующих или отсутствующих в геномах или располагающихся непосредственно рядом друг с другом, не обязательно тождественны функциональным модулям (то есть функции соответствующих генов не всегда взаимосвязаны) [Spirin *et al*, 2006]. Кроме того, в исследовании большой выборки

прокариотических генов всевозможных функций было показано, что менее двух третей из них формируют консервативные кассеты, т.е. имеют хоть сколько-нибудь заметную склонность к эволюционной устойчивости своего геномного окружения [Mavromatis *et al*, 2009].

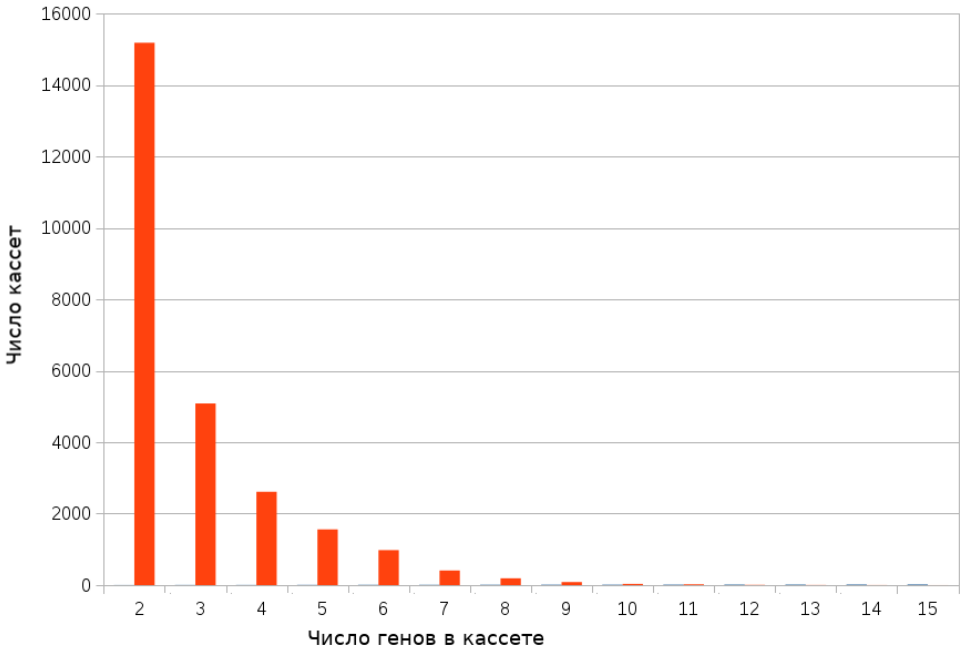

Рис. 6. Распределение количества кассет в зависимости от их размера (числа генов в кассете)

Всего исследуемые гены вошли в состав 26 тысяч кассет. Большая часть этих кассет были короткими; 55% состояли из двух генов, 20% – из трех (Рис. 6).

Распределение кассет по размеру среди разных классов бактерий, а также распределение функциональных классов в кассетах разных размеров показаны на Рис. 7 и Рис. 8, соответственно.

Большинство представленных на этих графиках кривых соответствует тенденции, отраженной на Рис. 6, однако есть несколько исключений. Так, гены, кодирующие транспортные белки, встречаются в 2-генных кассетах почти так же часто, как и в 3-генных, что можно объяснить широким распространением крупных белковых транспортных комплексов, таких как ABC-транспортеры, которые состоят не менее, чем из 3 субъединиц. У Fusobacteria, Thermotogae и Firmicutes 5- и 6-генные кассеты встречались практически не реже, чем 4-генные.

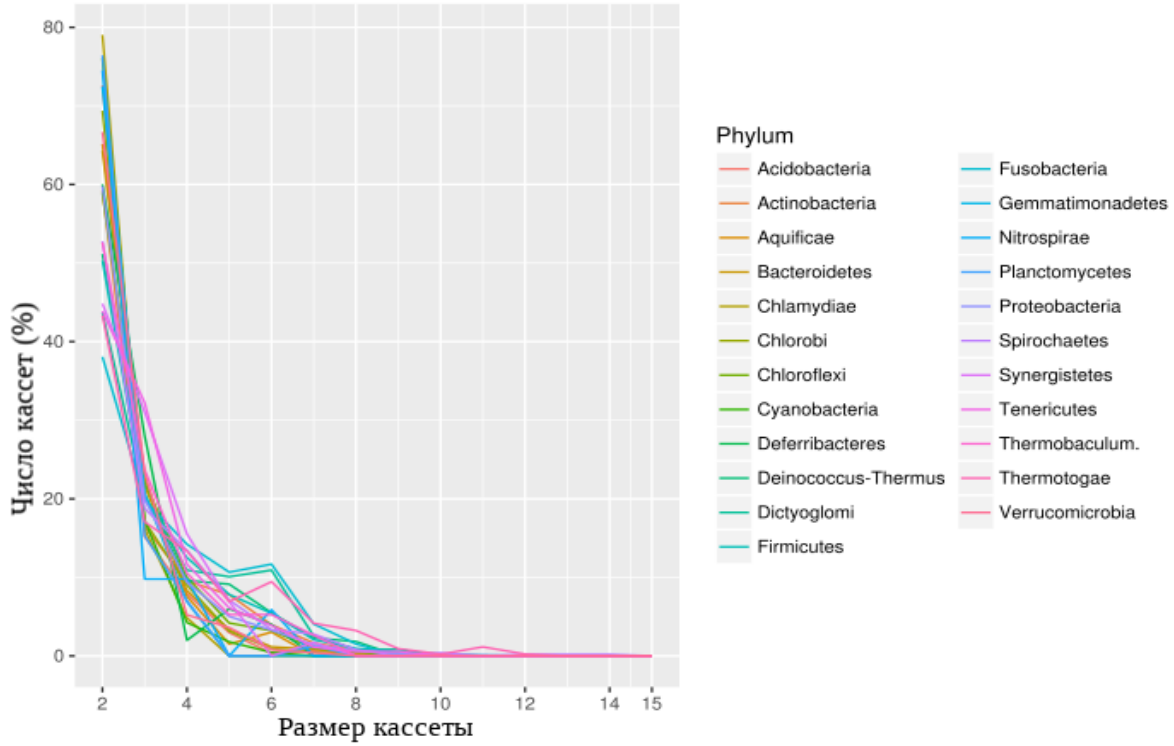

Рис. 7. Распределение кассет по размеру среди разных типов бактерий.

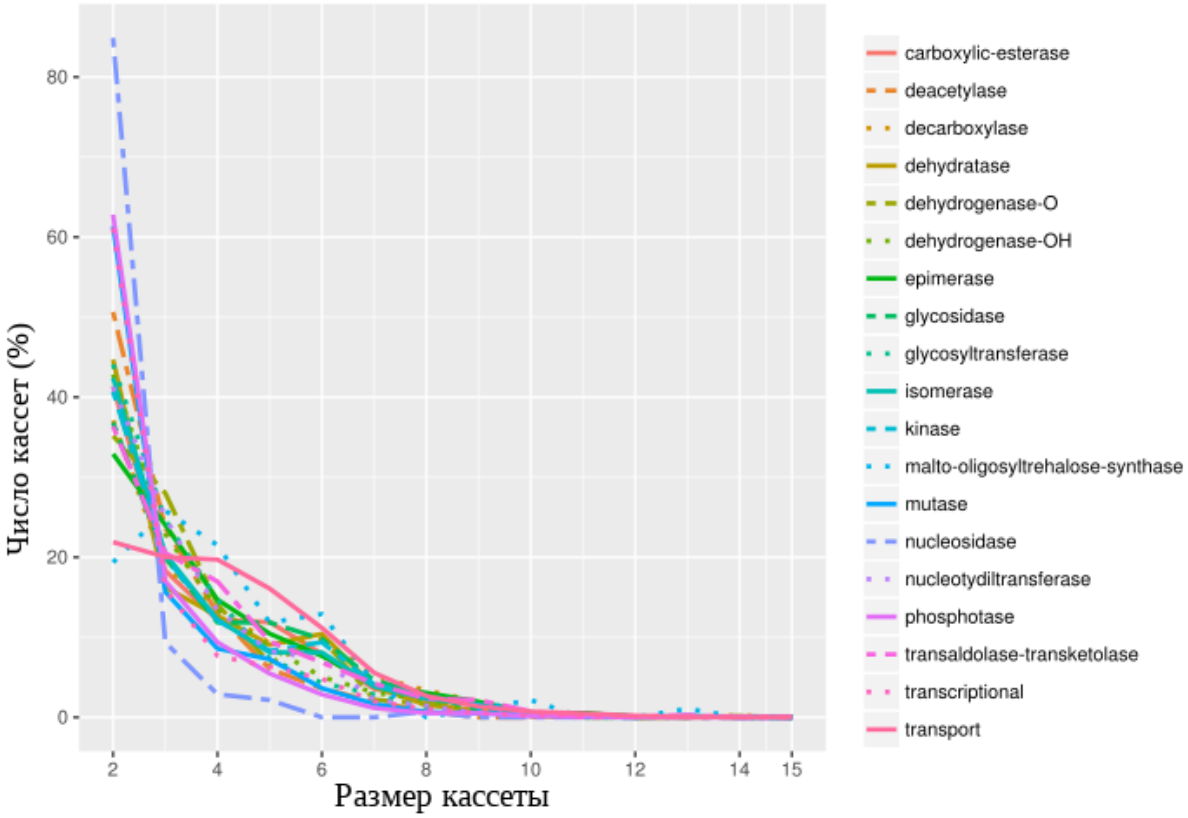

Рис. 8. Распределение по размеру кассет, содержащих гены разных функциональных классов.

Всего в кассетах встречалось около 10,4 тысяч разных комбинаций кластеров COG и около 2,5 тысяч разных комбинаций функциональных классов генов. По своему функциональному составу 45% кассет были уникальными.

Более того, только 43% всех исследованных нами генов входили в состав консервативных по составу кластеров COG кассет (кассета считалась консервативной, если встречалась в исследованных геномах по крайней мере дважды), тогда как в упомянутых выше исследованиях для всех бактериальных белок-кодирующих генов эта доля составляла 69%. Такие наблюдения подтверждают гипотезу о том, что значительная часть прокариотических генов не формирует очевидных эволюционно-устойчивых комбинаций на бактериальных хромосомах, причем оказывается, что внутри сегмента углеводного метаболизма эта доля еще больше.

Склонность генов разных функциональных классов и кластеров COG к формированию кассет

Функциональный класс	Количество генов	Склонность к образованию кассет	Идентификатор Enzyme Nomenclature
транскрипционные факторы (transcriptional)	39136	35,29%	-
транспортные белки (transport)	29701	70,83%	-
гликозилтрансферазы (glycosyltransferase)	14579	62,30%	2.4.1.
гликозидазы (glycosidase)	11475	64,74%	3.2.1.
киназы (kinase)	9250	57,95%	2.7.1.; 2.7.9
изомеразы (isomerase)	6458	55,20%	5.3.1.
дегидрогеназы-OH (dehydrogenase-OH)	5518	57,67%	1.1.
декарбоксилазы (decarboxylase)	2788	58,97%	4.1.
нуклеотидилтрансферазы (nucleotydiltransferase)	2125	70,96%	2.7.7.; 2.7.8
дегидратазы (dehydratase)	2091	52,75%	4.2.
фосфотазы (phosphotase)	2036	37,77%	3.1.3.
эпимеразы (epimerase)	1753	61,78%	5.1.3.
деацетилазы (deacetylase)	1525	51,02%	3.5.1.
трансальдолазы/транскетолазы (transaldolase/transketolase)	1514	70,54%	2.2.1.
мутазы (mutase)	1502	40,35%	5.4.2.
карбоксил-эстеразы (carboxylic-esterase)	1153	63,49%	3.1.1.
дегидрогеназы-O (dehydrogenase-O)	781	69,78%	1.2.
нуклеозидазы (nucleosidase)	597	23,28%	3.2.2.
мальто-олигозилтрегалоз-синтазы (malto-oligosyltrehalose-synthase)	100	93,00%	5.4.99

Таблица 2. Функциональные классы генов и их склонность к формированию кассет.

Долю генов, входящую в состав кассет, мы будем дальше называть склонностью к

образованию кассет для данной группы генов. Функциональные классы значительно различались по этому параметру – он варьировал от 23% до 93% (см. Таблицу 2). Наименьшей склонностью к образованию кассет обладали нуклеозидазы, фосфатазы и мутазы. Это можно объяснить участием продуктов таких генов в других типах метаболических путей, традиционно не относимых к углеводному метаболизму. Наибольшая склонность к образованию кассет наблюдалась у небольшого класса мальтоолигозилтрегалозсинтаз, на втором месте оказались трансальдолазы и транскетолазы, а на третьем – транспортеры.

Склонность к образованию кассет у разных кластеров COG различалась еще сильнее, чем у функциональных классов, варьируя между 0% и 100% (Рис. 9). Большинство крупных кластеров, содержащих более 4 тысяч генов, имели большую долю генов без соседей, относящихся к углеводному метаболизму, и склонность к образованию кассет для таких кластеров, в том числе для вторичных транспортеров суперсемейства MFS и многих транскрипционных регуляторов, составляла менее 40%. При этом склонность к образованию кассет у некоторых кластеров среднего размера, включающих от двух до четырех тысяч генов, составляла более 90% (здесь представлены, в том числе, транспортеры систем ABC). Самые маленькие кластеры, включающих менее двух тысяч генов, с наиболее высокой склонностью к образованию кассет, принадлежали к классам дегидрогеназ, изомераз, киназ, эпимераз и трансальдолаз/транскетолаз.

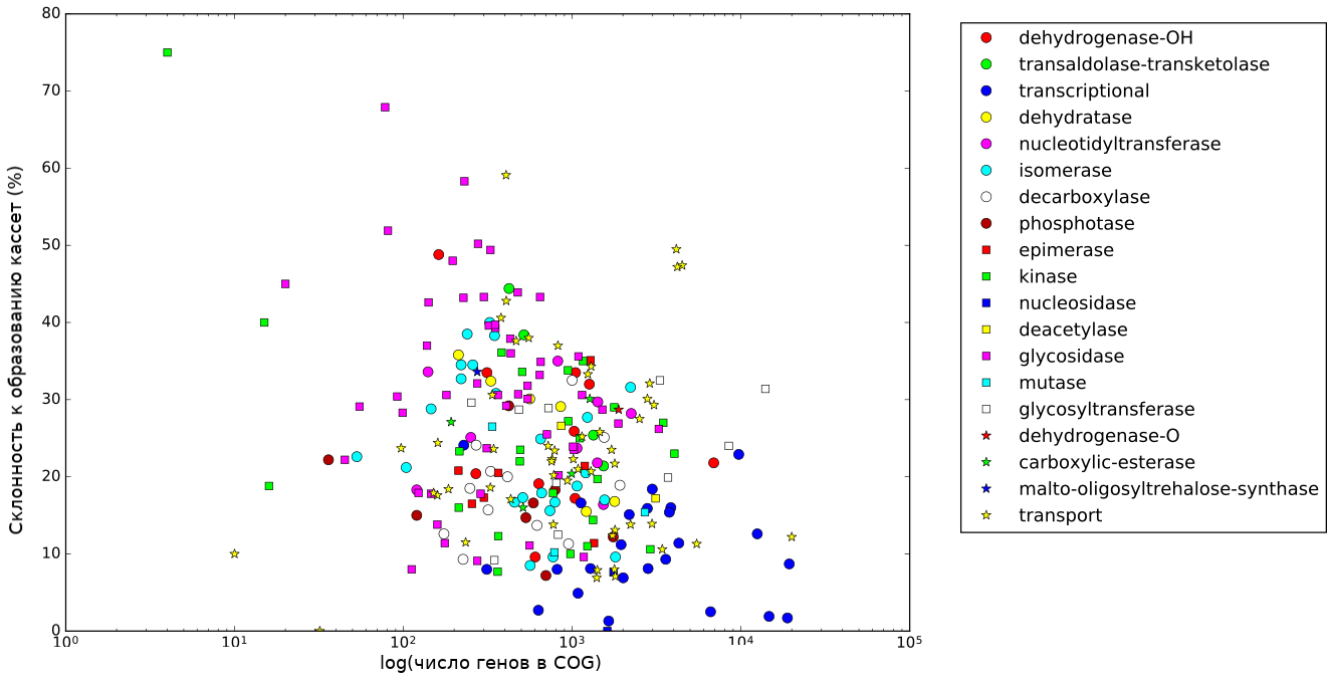

Рис. 9. Склонность к образованию кассет (по вертикали) у разных кластеров COG (размер кластера, т.е. число генов в COG, указан по горизонтали). Форма и цвет значка каждого кластера указывают на функциональный класс, к которому он принадлежит.

Склонность генов разных бактериальных классов к формированию кассет

Филогенетические факторы также играли важную роль в склонности генов к формированию кассет. Для разных бактериальных классов она варьировала между 37% и 76% (Рис. 10).

Наибольшей склонностью к образованию кассет обладали представители классов Dictyoglomi и Fusobacteria (76%), Thermotogae (72%) и Bacilli (65%). Это соответствует опубликованным данным о том, что гены представителей этих классов (например, рода *Streptococcus*) часто лежат в составе длинных оперонов [Dehal *et al*, 2010; Gama-Castro *et al*, 2016]. Наименьшей склонностью к образованию кассет обладали представители классов Planctomycetia (37%) и Chlamydiae (37%) и Chlorobia (40%).

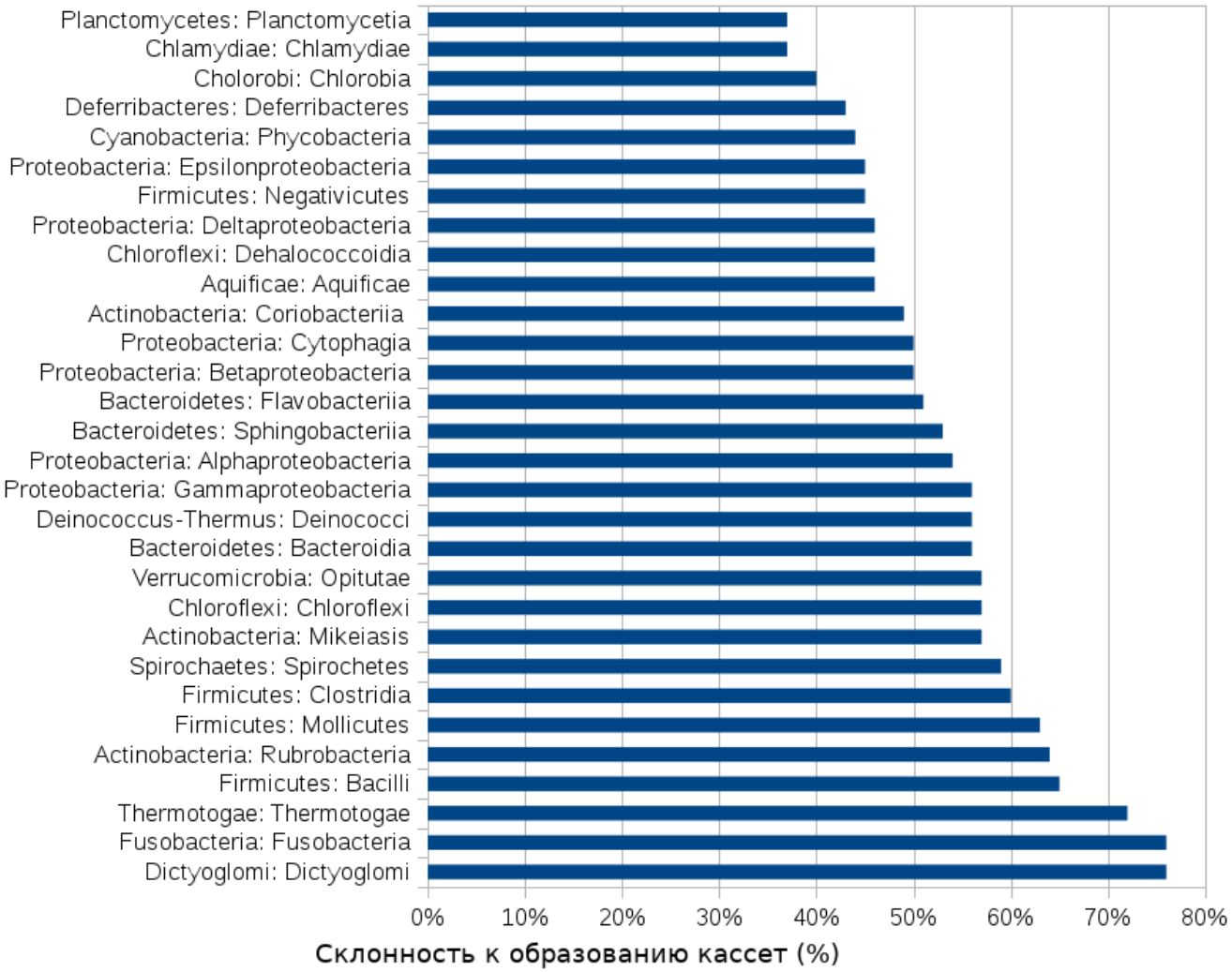

Рис. 10. Склонность к образованию кассет у генов, принадлежащих геномам бактерий разных классов. По вертикали указаны тип и класс бактерий. По горизонтали - склонность к образованию кассет в процентах.

Среди крупных классов, в каждом из которых было аннотировано не менее восьми тысяч генов углеводного метаболизма, представители класса Deltaproteobacteria обладали наименьшей склонностью к образованию кассет (46%), а наибольшая склонность к образованию кассет оказалась у классов Clostridia (60%) и Bacilli (65%).

Функциональный состав кассет генов углеводного метаболизма

Наиболее распространенным участником кассет среди функциональных классов оказались гены, кодирующие транспортеры, гликозидазы и гликозилтрансферазы. Самая длинная кассета, обнаруженная в геноме *Stackebrandtia nassauensis* DSM 44728, включала 15 генов, среди которых было 11 транспортеров, 2 изомеразы, одна гликозидаза и одна гликозилтрансфераза.

Транспортеры встречались в 18% кассет, причем 10% кассет содержали не меньше двух транспортеров. Гликозидазы встречались в 19% кассет, причем 5,8% кассет имели не меньше двух гликозидаз. Гликозилтрансферазы также встречались в 19% кассет, причем в 9,4% кассет они встречались не менее двух раз.

Ни один из функциональных классов не оказался представлен одновременно более, чем в пятой части изученных кассет, что подчеркивает существенное разнообразие ко-локализационных тенденций генов, относящихся к углеводному метаболизму бактерий.

Попарные ко-локализационные тенденции представителей разных функциональных классов

Для того, чтобы оценить значимость событий попарной ко-локализации генов разных функциональных классов, мы сравнивали соответствующие события ко-локализации в кассетах с событиями случайной модели. Ожидалось, что разнообразие эволюционно значимых связей между классами будет достаточно высоким. Однако из 190 возможных пар функциональных классов только у 45 (24%) число событий ко-локализации оказалось значительно выше, чем в случайной модели.

Количество связей варьировало для каждого класса от 0 до 8 (Рис. 11). Размер класса, то есть число входящих в него генов, напрямую не влиял на это значение. Так, несмотря на большие размеры класса транспортеров, включающего более 21 тысячи генов в составе кассет, он не имел ни одной значимой связи с другими классами. Класс трансальдолаз/транскетолаз, включающий около тысячи генов, входящих в кассеты, продемонстрировал четыре значимых связи с другими классами, тогда как сходный по размеру класс деацетилаз обладал всего двумя. Склонность к формированию кассет представителей класса, как таковая, по-видимому, также не влияла напрямую на число ко-локализационных связей этого класса с другими. Класс декарбоксилаз, склонность к формированию кассет которого составляла 60%, участвовал в восьми связях, а класс гликозилтрансфераз с аналогичной склонностью участвовал всего в четырех.

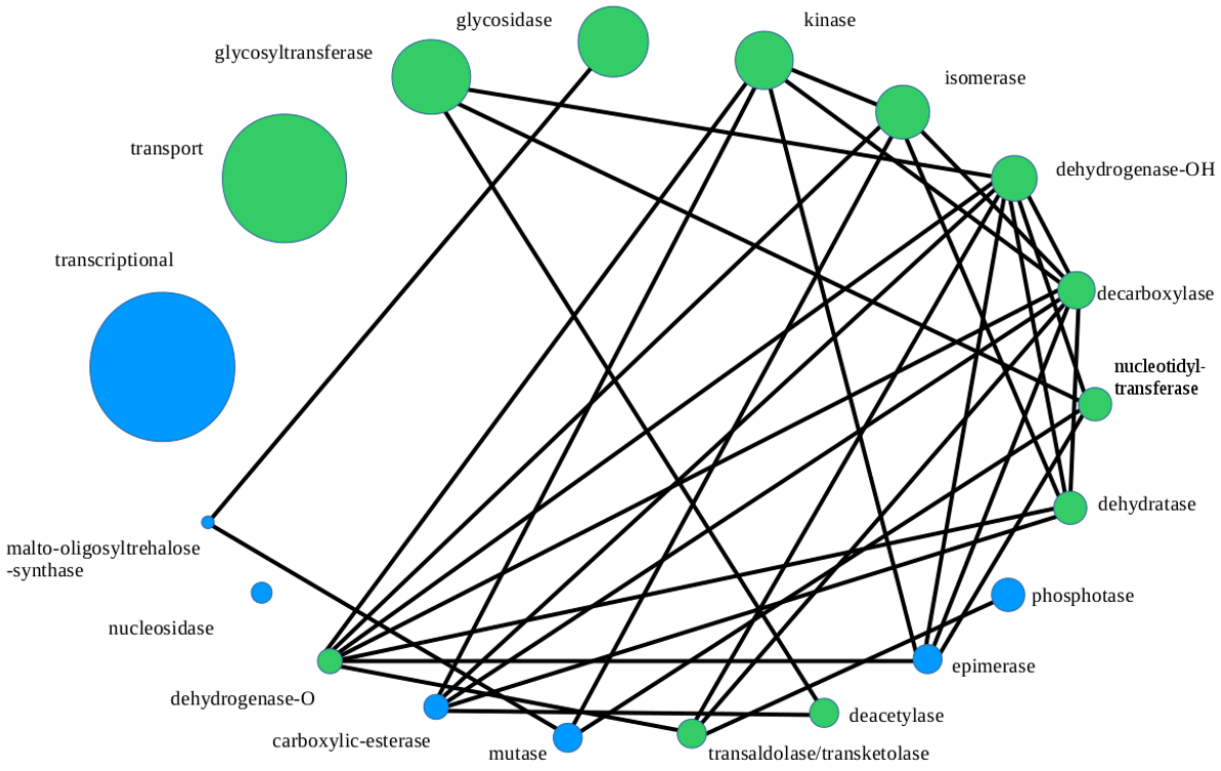

Рис. 11. Ко-локализационные связи между функциональными классами генов углеводного метаболизма. Кругами представлены разные классы, размер круга соответствует относительному размеру класса. Линиями соединены классы, имеющие значимую ко-локализационную связь. Зеленым цветом отмечены классы, представители которых имеют тенденцию к ко-локализации друг с другом.

Значительная часть связей была сформирована классами декарбоксилаз, дегидрогеназ-OH и дегидрогеназ-O (у них оказалось 8, 8 и 7 связей, соответственно). Таким образом, именно эти классы обладали наиболее сложными и при этом неслучайными предпочтениями по отношению к своему геномному окружению.

Большая часть связей была образована парами функций, встречающихся в распространенных и хорошо изученных метаболических путях. Так, например, изомераза и киназа одновременно присутствуют в путях, связанных с деградацией лактозы, галактозы, хитина и арабинозы. Декарбоксилаза и киназа присутствуют во всех вариациях пути Энтнера-Дудорова. Эпимераза и мутаза встречаются в путях гликолиза и глюконеогенеза, а также, например, в пути деградации маннана. Дегидрогеназа и карбоксил-эстераза участвуют в путях деградации галактозы.

Это наблюдение соответствует представлениям о том, что белки, участвующие в одном и том же метаболическом пути, неслучайно закодированы в ко-локализованных генах или даже расположены в составе единого оперона. Однако ко-локализационные

события многих пар функций, присутствующих в известных метаболических путях, в рамках данного анализа не преодолели порога значимости – например, гликозидаза и киназа, совместно участвующие в гликолизе и других метаболических путях (гены гликозидаз и киназ недостаточно часто встречались в кассетах вместе, чтобы эти события можно было отличить от случайных событий ко-локализации). Такой результат опровергает гипотезу о том, что ко-локализация является почти обязательным условием для генов, кодирующих белки с взаимосвязанными функциями.

Попарные ко-локализационные тенденции представителей одних и тех же функциональных классов

Из 45 выявленных ко-локализационных связей 12 были сформированы благодаря ко-локализации представителей одного и того же класса. Это означает, что в составе общих кассет часто присутствовали два или несколько генов, принадлежащих к одному и тому же функциональному классу, и такая тенденция оказалась неслучайной.

Больше всего ко-локализованных генов одного класса оказались среди транспортеров, гликозидаз, транскетолаз/трансальдолаз и гликозилтрансфераз. Стоит отметить, что класс гликозилтрансфераз и класс трансальдолаз/транскетолаз были представлены в кассетах несколькими генами чаще, чем одним.

Гены одного и того же класса, ко-локализованные в кассетах, делились на две группы – гены, кодирующие разные субъединицы белковых комплексов, и гены, кодирующие отдельные белки. Наиболее распространенным примером из первой группы являлись гены, кодирующие субъединицы транспортных комплексов. Остальные ко-локализованные гены одного класса, обычно, кодировали самостоятельные белки; в части случаев они могли быть участниками последовательных этапов метаболических путей.

Так, например, известно, что несколько гликозидаз могут участвовать в последовательных этапах деградации сложных полисахаридов [Kabisch *et al*, 2014]. Несколько гликозилтрансфераз могут участвовать в последовательных этапах путей биосинтеза клеточной стенки бактерий [Lamothe *et al*, 2002]. Две или три киназы также могут одновременно участвовать в последовательных этапах одного метаболического пути – например, гликолиза или деградации лактозы [Caspi *et al*, 2016]. Для других случаев, например, для декарбоксилаз, причины частой ко-локализации генов одного и того же функционального класса не столь очевидны.

Роль событий локальной дупликации и образования ксенологов и псевдопаралогов в ко-локализации генов сходных функций

Нашей задачей было выяснить, как часто ко-локализованные гены, имеющие сходные функции, являются результатом событий локальной дупликации, поскольку в противном случае их ко-локализация не имеет очевидного объяснения и имеет смысл обсуждать более глубокие эволюционные или функциональные ее причины.

В 44% случаев ко-локализованные гены одного и того же функционального класса также принадлежали к одному и тому же кластеру COG, а следовательно, обладали определенным структурным сходством. Среди 264 кластеров COG нашей базы данных 189 кластеров были ко-локализованы в исследуемых геномах хотя бы однажды. Для того, чтобы оценить вклад в такую ко-локализацию событий локальных дупликаций, мы использовали разработанный нами инструмент NSimScan.

Только в 3,6% случаев гены в паре продемонстрировали наибольшее сходство друг с другом. Во всех остальных случаях для одного или обоих участников пары среди других представителей COG отыскивалось более близкое совпадение, причем в 62% случаев соответствующий ген располагался не только в другой кассете, но и в другом геноме.

В целом, мы можем сказать, что ко-локализация генов близких функций в преобладающем большинстве случаев, по-видимому, не является результатом события локальной дупликации. Кроме того, менее 10% пар ко-локализованных генов из одного COG оказывались более похожими на один и тот же ген в другом геноме, чем друг на друга, в большинстве случаев они были похожи на два разных гена – мы также исключаем существенный вклад в ко-локализацию генов сходных функций псевдопаралогов и ксенологов.

Эволюционное значение попарной ко-локализации представителей одного функционального класса

Мы предполагаем, что гены сходных функций, расположенные на бактериальной хромосоме рядом друг с другом (особенно многочисленными группами, как это происходит, например, у гликозилтрансфераз и гликозидаз), могут использоваться бактерией одновременно в ситуациях определенного типа. Такой набор может иметь общий механизм регуляции транскрипции, и в определенных условиях его гены могут экспрессироваться одновременно, например, когда клетке необходимо включение целого ряда ферментов, участвующих в деградации или биосинтезе углеводов. Это происходит, например, при утилизации или биосинтезе сложных полисахаридов, где разные гликозилтрансферазы или гликозидазы задействованы в рамках общих или тесно переплетенных метаболических путей.

Кроме того, известно, что, оказавшись в неоптимальных для роста условиях среды, клетка может активировать экспрессию сразу целой группы генов. Так, известно, что некоторые бактерии условиях голодания одновременно активируют экспрессию множества генов, ответственных за катаболизм и транспорт альтернативных источников углерода. Это касается, в частности, транспортеров и гликозидаз. Ко-локализация генов, активирующихся в подобных стрессовых условиях, может также объясняться удобством одновременной регуляции их транскрипции. Кроме того, ко-локализованные гены будет чаще совместно передаваться в другие геномы при событиях горизонтального переноса, и соответствующие комбинации могут

эволюционно закрепляться как в родственных, так и в других видах бактерий.

Четвертая глава описывает участие *yih*-кассеты *Escherichia coli* в катаболизме лактозы.

В некоторых случаях геномное окружение позволяет успешно предсказывать роль генов в тех или иных процессах. Мы проанализировали наиболее распространенные комбинации функциональных классов внутри кассет, исходя из предположения, что консервативность будет указывать на эволюционные преимущества такой колокализации. Большинство консервативных сочетаний оказались короткими (двух- или трех-генными). Среди более длинных сочетаний в качестве кандидата для экспериментальной проверки предсказания функции мы выбрали комбинацию из шести функциональных классов – транспортера, регулятора транскрипции, гликозидазы, альдолазы, киназы и изомеразы. Она встречалась в кассетах не близкородственных бактерий, что указывало на неслучайность подобной колокализации. Кассета с таким составом оказалась распространена как у гаммапротеобактерий семейства Enterobacteriaceae, среди которых наиболее известным представителем является кишечная палочка *Escherichia coli* (кассета *ompL-yihOPQRSTUVW,* далее - *yih*-кассета), так и у бактерий класса Bacilli, среди родов *Streptococcus* и *Staphylococcus* (кассета *lacGEFDCBAR)* (Рис. 12).

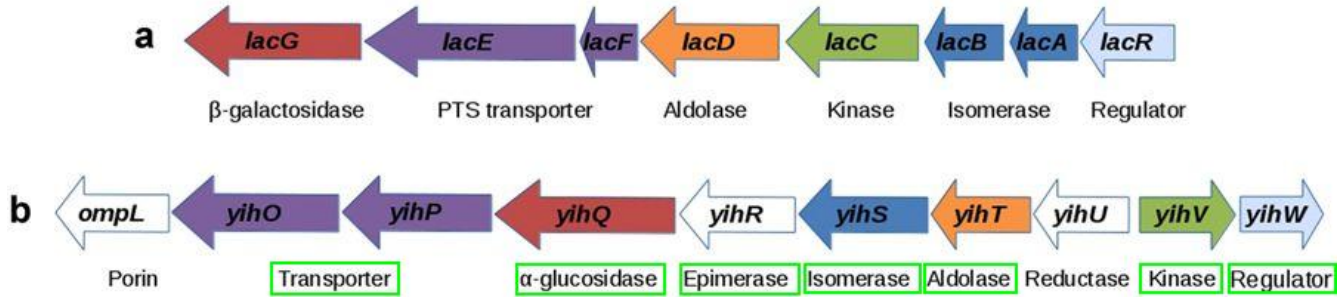

Рис. 12. Кассеты бактерий класса Bacilli (a) и семейства Enterobacteriaceae (b). Одинаковые цвета генов указывают на сходство функций кодируемых ими белков. Белым отмечены гены, кодирующие функции, не представленные в другой кассете. Зеленым в кассете Bacilli обведены функции белков, которые могут быть задействованы в катаболизме лактозы.

В исследовании Denger *et al* в 2014 году впервые было описано участие белков *yih*-кассеты в метаболизме серосодержащих углеводных соединений. У представителей класса Bacilli кассета *lacGEFDCBAR* кодирует белки, участвующие в катаболизме лактозы. Мы предположили, что помимо редкой функции утилизации серосодержащих углеводов, кассета кишечной палочки может также участвовать в катаболизме лактозы, а ее гены могут, таким образом, кодировать мультифункциональные белки.

Предположение было подтверждено с помощью экспериментального исследования. Эта часть работы выполнялась в лаборатории функциональной геномики и клеточного стресса Института биофизики клетки РАН г. Пущино под руководством М.Н. Тутукиной.

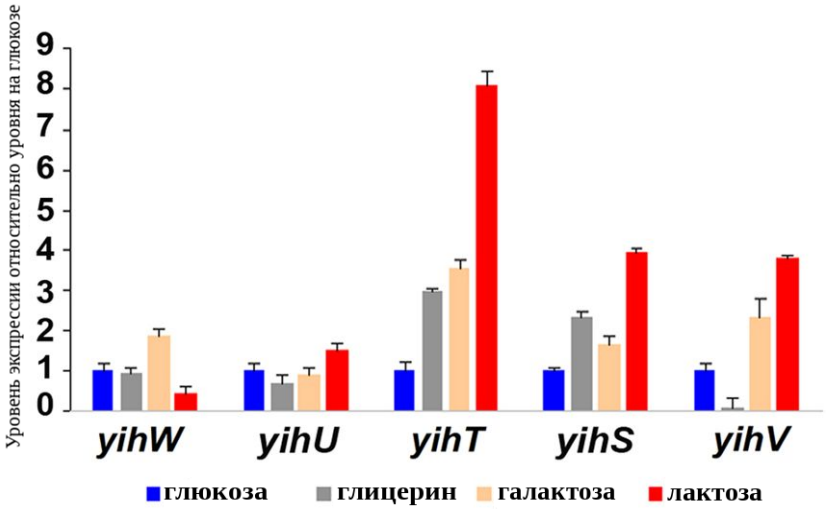

Рис. 12. Уровни мРНК сравнивали с помощью метода количественной ПЦР с детекцией в реальном времени (qRT-PCR). Уровень экспрессии генов при росте на глюкозе взят за единицу.

Выяснилось, что экспрессия генов, кодирующих в кишечной палочке альдолазу (*yihT*), изомеразу (*yihS*) и киназу (*yihV*) значительно повышалась во время роста клеток штамма *E. coli* K12 MG1655 на лактозе (Рис. 12). В ходе работы были также впервые идентифицированы точки начала транскрипции для всех этих генов *in silico*, *in vitro* and *in vivo* (в предыдущих работах большая часть кассеты рассматривалась, скорее, как единый оперон [Denger *et al*, 2014]) и показано, что из трех промоторов гена альдолазы один активировался именно при росте клеток на лактозе.

Мы также проанализировали механизм переключения регуляции транскрипции данной кассеты. Было показано, что в этом процессе участвует локальный регулятор YihW, принадлежащего к семейству регуляторов DeoR и глобальный регулятор углеводного метаболизма cAMP-CRP, и в зависимости от условий среды, они могут оказывать либо комплементарное, либо противоположное воздействие на экспрессию генов кассеты. С помощью метода филогенетического футпринта были предсказаны, а затем экспериментально подтверждены потенциальные сайты связывания CRP с межгенными участками кассеты.

Для основной части работы мы использовали штамм *E. coli* M182 с выключенным *lac*-опероном. Его клетки не могли катаболизировать лактозу с помощью своего стандартного, хорошо известного пути. Всего работа проводилась с тремя типами культур M182 – диким типом (wt), мутантом по *yihW* и мутантом по *crp* (Рис. 13).

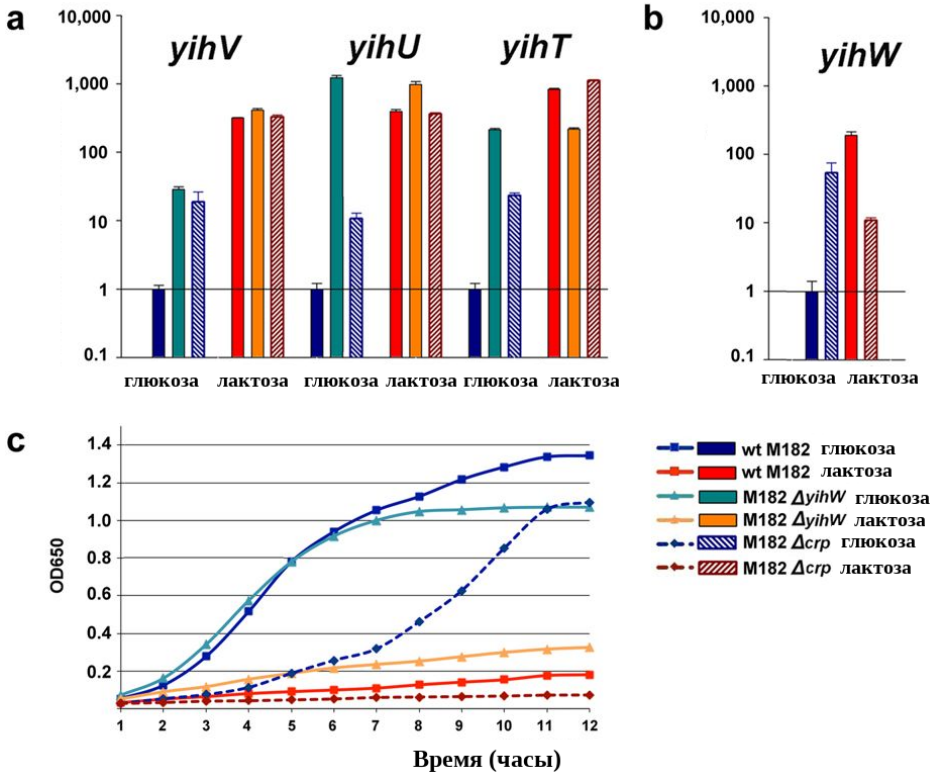

Рис. 13. Влияние делеции генов *yihW* и *crp* на уровень мРНК в генах *yih*-кассеты (a,b) и рост клеток на глюкозе и лактозе (c). Уровни мРНК указаны относительно уровня в родительском штамме при росте культуры на глюкозе.

Профиль транскрипции *yihS* оказался практически идентичен *yihT*, что, по-видимому, свидетельствует об их совместной транскрипции. Выяснилось, что экспрессия *yihTS* как на глюкозе, так и на лактозе контролируется фактором YihW, который выполняет роль углевод-зависимого двойного переключателя (Рис. 13, a). Во время роста на глюкозе экспрессия *yihTS* подавляется фактором CRP. Экспрессия самого *yihW* активируется с помощью CRP на лактозе и подавляется на глюкозе (Рис. 13, b). Оба фактора YihW и CRP работают как репрессоры транскрипции гена *yihV* на глюкозе (Рис. 13, a). Наконец, общий рост культуры *E. coli* M182, мутантной по *yihW*, существенно снижен на лактозе относительно роста на глюкозе (Рис 13, c).

В целом, YihW, по-видимому, играет в *yih*-кассете роль двойного переключателя, активируя некоторые ее гены (*yihT*, *yihS*) во время фазы экспоненциального роста культуры на лактозе, и репрессируя некоторые ее гены (*yihT*, *yihS*, *yihV*) во время роста на глюкозе. Фактор CRP при росте на лактозе активирует транскрипцию гена *yihW*, то есть выполняет роль, комплементарную YihW, а при росте на глюкозе репрессирует его транскрипцию (Рис. 14).

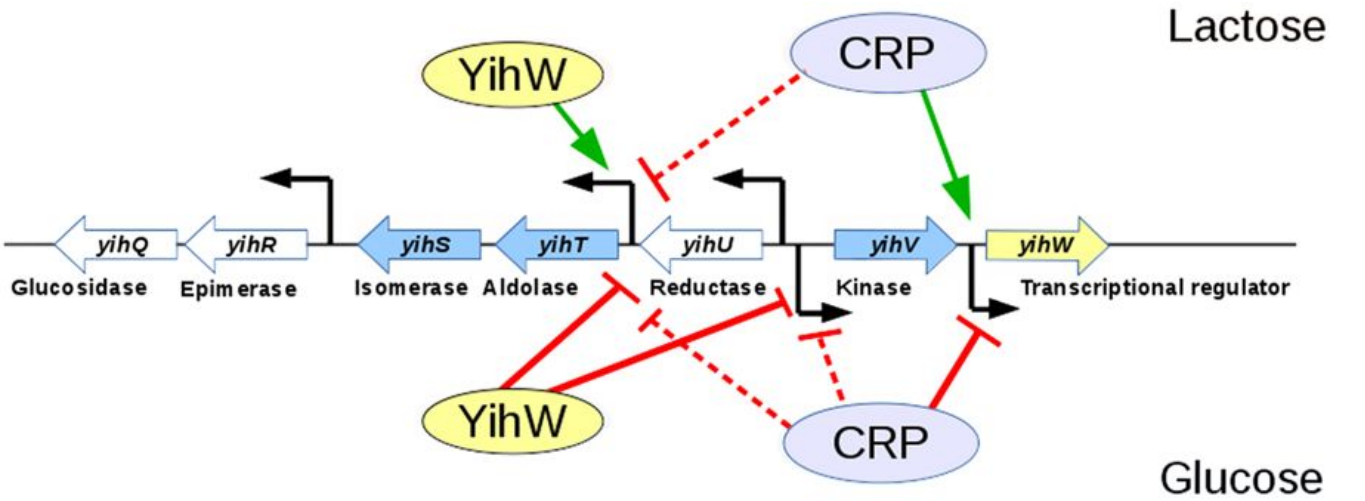

Рис. 14. Регуляция генов *yih*-кассеты при росте культуры на лактозе (сверху) и глюкозе (снизу), осуществляемая факторами транскрипции CRP и YihW. Зелеными стрелками отмечена активация транскрипции, красными линиями – подавление транскрипции. Прерывистыми линиями отмечены процессы, где происходит лишь умеренное подавление, которое, возможно, осуществляется не напрямую, а через дополнительные транскрипционные факторы.

Описанный случай является примером успешного предсказания функций генов на основе их ко-локализационных тенденций. Механизм регуляции транскрипции генов *yih*-кассеты оказался достаточно сложным и зависящим от условий среды. Пользуясь этой тонко налаженной системой, бактерия, по всей видимости, может использовать один и тот же набор белков для разных задач.

Мы показали, что кассета генов *Escherichia coli*, участвующая в деградации серосодержащих углеводов, также связана с катаболизмом лактозы. Таким образом впервые со времен классической работы по описанию *lac*-оперона мы можем говорить о наличии у кишечной палочки альтернативного пути утилизации лактозы, включающего в себя все этапы после первичного гидролиза.

Выводы

1. Разработан инструмент NSimScan для поиска нуклеотидных последовательностей удаленного сходства; по совокупности таких параметров как чувствительность, частота ошибок и скорость он превосходит все стандартные инструменты в своей области. Наилучшим образом он подходит для поиска последовательностей, различающихся на 60-90%.

2. Описана сеть эволюционных связей 148 тысяч генов углеводного метаболизма 665 видов бактерий, выраженная в форме их ко-локализационных тенденций. 53% таких генов оказались ко-локализованы, остальные располагаются на бактериальных геномах по отдельности.

3. Склонность к ко-локализации, т.е. к формированию кассет различается у разных генов; ключевыми ее факторами являются функциональные и структурные характеристики гена и таксономическая принадлежность бактерии. Склонность к формированию кассет у разных функциональных классов составляет от 23 до 93%; у разных кластеров ортологических групп генов – 0 до 100%, у разных бактериальных классов – от 40 до 76%.

4. Среди 19 исследуемых функциональных классов 45 пар формируют консервативные и, по всей видимости, эволюционно значимые ко-локализационные связи. Количество таких связей для каждого класса сильно варьирует, подчеркивая существенную разницу в предпочтениях к хромосомному окружению у генов разных функций. Для 11 классов гены одного и того же класса также демонстрируют выраженное предпочтение к ко-локализации, причем большинство таких случаев, по-видимому, не являются результатом событий локальной дупликации.

5. Анализ консервативных сочетаний внутри кассет генов позволяет успешно предсказывать их функции. С его помощью экспериментально подтверждено участие *yih*-кассеты генов *Escherichia coli* в катаболизме лактозы; предложен, таким образом, новый путь утилизации лактозы у кишечной палочки и предсказаны мультифункциональные характеристики соответствующих белков. В переключении механизмов экспрессии генов этой кассеты в разных условиях среды участвуют локальный регулятор YihW и глобальный регулятор CRP.

Список публикаций по теме диссертации

По теме диссертации опубликовано три статьи в рецензируемых международных научных журналах, входящих в основные библиометрические базы данных (PubMed, WoS и Scopus):

1. V. Novichkov, A. Kaznadzey, N. Alexandrova, D. Kaznadzey (2016) NSimScan: DNA comparison tool with increased speed, sensitivity and accuracy. Bioinformatics 32(15):2380-1

2. A. Kaznadzey, P. Shelyakin, M. Gelfand (2017) Sugar Lego: gene composition of bacterial carbohydrate metabolism genomic loci. Biology Direct 12(1):28.

3. A. Kaznadzey, P. Shelyakin, E. Belousova, A. Eremina, U. Shvyreva, D. Bykova, V. Emelianenko, A. Korosteleva, M. Tutukina, M. Gelfand (2018) The genes of the sulphoquinovose catabolism in Escherichia coli are also associated with a previously unknown pathway of lactose degradation. Scientific Reports 8(1):3177.

Результаты работы опубликованы в сборниках тезисов международных и российских конференций:

1. A. Kaznadzey (2010) Evolutional study of carbohydrate metabolism loci in bacterial genomes, Interdisciplinary School and Conference of Information Technology and Systems (ITaS'10), Геленджик.

2. A. Kaznadzey, P. Shelyakin (2011) Study of evolution and classification of genome loci of carbohydrate metabolism of bacteria. Interdisciplinary School and Conference of Information Technology and Systems (ITaS'11), Геленджик.

3. A. Kaznadzey, P. Shelyakin (2011) Evolution study and classification of carbohydrate metabolism genome loci in bacteria. International Moscow Conference on Computational Molecular Biology (MCCMB'11), Москва.

4. A. Kaznadzey, P. Shelyakin (2012) Diversity of genome loci and co-localization patterns study of the protein families from different functional classes of the bacterial carbohydrate metabolism. 8th International Conference on the Bioinformatics of Genome Regulation and Structure – Systems Biology (BGRS\SB-2012), Новосибирск.

5. A. Kaznadzey, P. Shelyakin (2012) Diversity of genome loci and co-localization patterns study of the protein families from different functional classes of the bacterial carbohydrate metabolism. Interdisciplinary School and Conference of Information Technology and Systems (ITaS'12), Петрозаводск.

6. A. Kaznadzey, P. Shelyakin (2013) Structure, classification, evolution and phylogenetics of carbohydrate metabolism gene loci in bacteria. Moscow Conference on Computational Molecular Biology (MCCMB'13), Москва.

7. A. Kaznadzey, P. Shelyakin (2015) Co-evolution of carbohydrate metabolism genes of same and different functional classes in bacteria' (ITaS'15), Сочи.

8. A. Kaznadzey, M. Tutukina, A. Eremina, E. Belousova, P. Shelyakin, M. Gelfand (2016) Escherichia coli gene cassette previously described as an operon responsible for sulphoglycolipide degradation: not an operon and has other functions as well. Interdisciplinary School and Conference of Information Technology and Systems (ITaS'16), Санкт-Петербург.