

На правах рукописи

Лейн Семен Александрович

**Эволюция транскрипционной регуляции метаболизма
углеводов в бактериях**

03.01.09 - математическая биология, биоинформатика

Автореферат

диссертации на соискание ученой степени
кандидата биологических наук

Москва – 2012

Работа выполнена в Учебно-Научном Центре Биоинформатики Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук.

Научный руководитель:

Родионов Дмитрий Александрович

кандидат биологических наук

заведующий сектором №6 Функциональной и сравнительной геномики прокариот.

Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук.

Официальные оппоненты:

Озолин Ольга Николаевна

доктор биологических наук, профессор,

заведующая лабораторией Функциональной геномики и клеточного стресса.

Федеральное государственное бюджетное учреждение науки Институт биофизики клетки Российской академии наук.

Мясникова Екатерина Марковна

кандидат физико-математических наук,

ведущий научный сотрудник лаборатории Молекулярной вирусологии и онкологии.

Центр перспективных исследований Федерального государственного автономного образовательного учреждения высшего образования Санкт-Петербургского государственного политехнического университета.

Ведущее Учреждение:

Научно-исследовательский институт физико-химической биологии имени

А.Н.Белозерского Московского государственного университета имени М.В.Ломоносова

Защита диссертации состоится «16» октября 2014 г в 14 часов на заседании Диссертационного Совета Д.002.077.04 при Федеральном государственном бюджетном учреждении науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук по адресу: 127994, г. Москва, ГСП-4, Большой Каретный переулок, 19, стр. 1.

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук и на сайте www.iitp.ru

Автореферат диссертации разослан « ____ » сентября 2014 г.

Ученый секретарь диссертационного совета Д.002.077.04,

доктор биологических наук, профессор

Рожкова Г.И.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы

Углеводы и их производные являются основными источниками углерода и энергии для большинства гетеротрофных бактерий, а также важным строительным материалом в живой природе. Они образуют различные по составу и структуре моно-, олиго- и полисахариды. В ходе эволюции бактерии выработали множество путей утилизации углеводов, подстраиваясь под это многообразие субстратов. Геномы многих свободноживущих и патогенных бактерий кодируют до нескольких десятков метаболических путей катаболизма сахаров.

Экспрессия генов метаболизма углеводов у бактерий регулируется в зависимости от доступности тех или иных субстратов. Ключевым элементом в регуляции экспрессии генов выступают факторы транскрипции – специальные белки-регуляторы, которые контролируют инициацию транскрипции генов. Большой интерес микробиологов представляет эволюция регулонов утилизации различных сахаров, что помимо регуляции включает возникновение альтернативных биохимических путей, неортологичное замещение генов, образование функционально гетерогенных семейств паралогов. Причем, этот интерес исходит как от фундаментальной науки, так и от промышленной биотехнологии, перед которой стоит задача эффективной переработки органических отходов, большинство из которых является растительной биомассой и, следовательно, содержит большое разнообразие углеводсодержащих соединений.

До недавнего времени основным средством получения новых знаний о регуляции экспрессии генов являлись кропотливые эксперименты с модельными микроорганизмами. Однако, несмотря на достаточно большую надежность данных методов, они не дают представления о полной картине регуляторных взаимодействий, сосредоточившись лишь на небольшом фрагменте системы. Значительно улучшили ситуацию методы высокопроизводительных экспериментов. И хотя в них прослеживается изменение экспрессии тысяч генов, они характеризуются высоким уровнем шума и сложностью разделения прямых регуляторных эффектов от косвенных, например таких как регуляторные каскады или ко-регуляция генов.

В связи с резким увеличением числа полностью секвенированных геномов в последнее время у исследователей появился мощный инструмент для биоинформатического анализа бактериальных геномов – методы сравнительной геномики. К настоящему моменту полностью секвенированы и общедоступны почти 2000 бактериальных геномов, которые широко покрывают разнообразные таксономические группы. При этом удешевление технологий секвенирования позволяет получать геномные

данные с нарастающей скоростью. Изучение регуляции транскрипции генов методами биоинформатики ставит своей задачей выявление регуляторных участков ДНК (промоторов, терминаторов, РНК-переключателей, сайтов связывания транскрипционных факторов), а также предсказание новых ДНК-связывающих белков для обнаруженных сайтов ДНК.

Быстрое увеличение числа геномных последовательностей ставит необходимым решение другой важной проблемы геномики – получение достоверной функциональной аннотации генов. Основным способом аннотации генов до недавнего времени являлся перенос функции экспериментально охарактеризованных белков на другие с помощью поиска сходства последовательностей. Для осуществления поиска по базам данных последовательностей были разработаны специализированные программы, наибольшую популярность среди которых получил алгоритм BLAST и его варианты.

Несмотря на большой успех методов основанных на поиске сходства в последовательностях, они не позволяют аннотировать многие гены и часто создают неточные или ошибочные аннотации. Существенно меняют ситуацию подходы сравнительной геномики, одним из которых является анализ эволюции структуры регулона в различных геномах [Gelfand et.al., 2000; Osterman and Overbeek, 2003]. Данный подход предполагает возможную функциональную связь между ко-регулируемыми генами, что дает существенную помощь в предсказании аннотации.

Таким образом, исследование регуляции с применением методов сравнительной геномики позволяет существенно улучшить понимание процессов в бактериальных клетках и их способностей к взаимодействию с окружающей средой.

Цели и задачи исследования

Целью данной работы был анализ эволюции регулонов метаболизма углеводов в геномах бактерий с помощью методов сравнительной геномики. В качестве объектов исследования были выбраны регулятор катаболизма арабинозы AraR, репрессор катаболизма N-ацетилгалактозамина в протеобактериях AgaR, регулятор центрального метаболизма углеводов в протеобактериях HexR и 43 регулятора метаболизма сахаров в семействе бактерий *Bacillaceae*.

В работе решаются следующие задачи:

1. Реконструкция регулонов метаболизма сахаров в экспериментально неизученных бактериях методами сравнительной геномики, построение распознающих правил для предсказания сайтов связывания факторов транскрипции, описание состава регулона и оперонной структуры.

2. Предсказание функций новых генов и уточнение существующих аннотаций генов вовлеченных в катаболизм различных углеводов.
3. Анализ таксон-специфических особенностей регуляции транскрипции генов катаболизма углеводов и построение вероятных сценариев эволюции регуляторных взаимодействий.

Научная новизна и практическое значение работы

В работе впервые полностью описаны локальные регулоны метаболических путей утилизации арабинозы и N-ацетилгалактозамина (AgaR и AgaR), а также глобальный регулон HexR контролирующей центральный метаболизм углерода у бактерий из двух таксономических групп. Для семейства *Bacillaceae* реконструировано 11 новых регуляторных систем для генов катаболизма арил-бета-глюкозидов, инозитола, рамногалактуронана, фруктозы, глюкозамина, хитина, мальтодекстрина, N-ацетилмурамата и альфа-галактозидов. При этом, для каждого нового белка регулятора были не только описаны наборы регулируемых генов, но и предсказаны их потенциальные ДНК сайты связывания и молекулы-эффекторы. Для 11 других факторов транскрипции контролирующих пути катаболизма углеводов, были впервые обнаружены их мотивы ДНК сайтов.

Метаболическая реконструкция путей утилизации арабинозы позволила обнаружить новые ферменты L-рибулокиназу AgaB-II и L-арабиноизомеразу AgaA-II, а также большое разнообразие систем транспорта и деградации арабинозосодержащих полисахаридов. Также было показано, что обнаруженная ранее в биоинформатическом анализе *Clostridium acetobutylicum* L-рибулокиназа AgaK является наиболее распространенным вариантом пути в бактериях типа *Firmicutes*.

Обнаружено множество вариаций в начальных стадиях метаболических путей катаболизма и транспорта N-ацетилгалактозамина у протеобактерий. Анализ геномного контекста *aga* генов позволил предположить специфичность новых транспортных систем PTS типа к различным аminosахарам. Было впервые показано, что функция галактозамин-6-фосфат диаминазы/изомеразы, ранее приписываемая ферменту AgaI, принадлежит белку AgaS.

Также в работе были предложены потенциальные сценарии эволюции изученных регулонов метаболизма сахаров. В результате анализа филогении регуляторных систем было выдвинуто предположение о сильном влиянии горизонтальных переносов в периферических путях метаболизма сахаров.

Работа имеет преимущественно теоретический характер, однако полученные данные могут применяться и в геномной инженерии. Обнаруженные новые варианты метаболических путей могут использоваться для построения сахаро-специфичных генетических модулей и получения высокопродуктивных геномно-инженерных штаммов микроорганизмов с заданными фенотипами утилизации сахаров, которые могут быть использованы, например, для производства биотоплива.

Апробация работы

Основные положения диссертации были представлены на международных конференциях: Информационные технологии и системы'09 (Бекасово, декабрь 2009), 2010 Genomic Science Contractor-Grantee and Knowledgebase Workshop (Кристал Сити, шт. Вирджиния, США), American Society For Microbiology 110th General Meeting 2010 (Сан-Диего, шт. Калифорния США, май 2010), Информационные технологии и системы'10 (Геленджик, сентябрь 2010), 3rd Annual Joint Conference on Systems Biology, Regulatory Genomics, and Reverse Engineering Challenges 2010 (Нью-Йорк, шт. Нью-Йорк, США, ноябрь 2010), Genomic Science Awardee Meeting IX and USDA-DOE Plant Feedstock Genomics for Bioenergy Awardee Meeting (Кристал Сити, шт. Вирджиния, США), Moscow Conference on Computational Molecular Biology 2011 (Москва, июль 2011), Информационные технологии и системы'11 (Геленджик, октябрь 2011), 2012 Department of Energy (DOE) Genomic Science Program Awardee Meeting (Бетезда, шт. Мэриленд, США, февраль 2012).

Объем и структура диссертации

Диссертационная работа изложена на 173 страницах и состоит из введения, шести глав, выводов и списка цитированной литературы. Глава 1 содержит обзор литературы по теме диссертации. Глава 2 содержит описание методик и программ, используемых для реконструкции регуляции транскрипции и функциональной аннотации генов. Главы с 3 по 6 содержат описание собственных исследований. Список литературы включает 247 наименований. Работа содержит 24 рисунка, 9 таблиц и 4 приложения.

СОДЕРЖАНИЕ РАБОТЫ

В первой главе приводится обзор литературы по теме диссертации. Обзор начинается с рассмотрения общих принципов инициации транскрипции в бактериях. Приводятся основные положения о РНК-полимеразе и белковых факторах транскрипции. Показаны механизмы взаимодействия факторов транскрипции с сайтами связывания на

ДНК и РНК-полимеразой для активации и репрессии транскрипции. Далее приводится описание методов сравнительной геномики для реконструкции регуляторных взаимодействий и аннотации генов. Указаны основные способы описания ДНК мотивов и поиска потенциальных сайтов связывания белков регуляторов. Рассмотрены также и методы предсказания функций генов. Показаны как стандартные способы, основанные на сходстве аминокислотных последовательностей, так и подходы сравнительной геномики, в частности, методы анализа геномного контекста.

Дальнейший обзор литературы посвящен отдельным системам регуляции путей утилизации арабинозы, N-ацетилгалактозамина и пути Энтнера-Дудорова у различных бактерий. А также всевозможным известным сахарным регулонам в грамположительной бактерии *B. subtilis*. Было показано, что большинство путей и систем регуляции было изучено лишь в нескольких модельных организмах.

Во **второй главе** приводятся методы, используемые для реконструкции регуляторных взаимодействий и метаболических путей. Реконструкция начиналась с определения набора геномов, в которых присутствует ортолог исследуемого фактора транскрипции. Для реконструкции AgaR, AgaR и HexR регулонов была взята выборка геномов, в которых присутствует ортологичный ген регулятора. Для реконструкции регуляции в семействе *Bacillaceae* была сделана репрезентативная выборка из 10 геномов семейства *Bacillaceae* исходя из филогенетического дерева видов. Затем проводился поиск известных или потенциальных регуляторов генов сахарного метаболизма.

Для отобранных регуляторов строились филогенетические деревья, которые позволяли определить группы ортологов для построения общего распознающего правила для поиска сайтов связывания.

Поиск потенциальных сайтов связывания регуляторных белков производился с помощью двух методов: филогенетического футпринтинга и метода матриц позиционных весов [Gelfand et.al.,2000; Rodionov, 2007]. В первом методе строилось множественное выравнивание 5'-некодирующих областей ортологичных генов, для которых предполагается сохранение структуры сайтов связывания. Найденные последовательности предполагаемых сайтов связывания использовались для построения матриц позиционных весов ДНК связывающих мотивов соответствующих транскрипционных факторов.

В качестве основного инструмента для реконструкции транскрипционной регуляции и описания регулонов использовалась программа RegPredict [Novichkov et.al., 2010a]. Все описанные в данной работе регулоны были загружены в базу данных RegPrecise, доступную в сети Интернет по адресу <http://regprecise.lbl.gov/RegPrecise/> [Novichkov et.al., 2010b].

Третья глава посвящена исследованию эволюции арабинозного регулона, контролируемого фактором транскрипции AraR, который обладает химерной структуры поскольку N-концевой ДНК-связывающий домен принадлежит GntR семейству, а C-концевой субстрат-связывающий домен относится к LacI семейству регуляторов. У модельной бактерии *B. subtilis* было установлено, что AraR контролирует гены утилизации арабинозы и арабинозосодержащих полисахаридов.

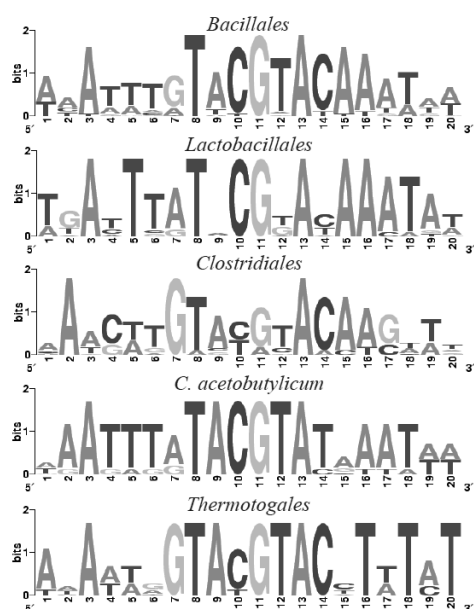


Рисунок 1. Диаграммы LOGO, построенные с использованием всех сайтов связывания AraR для каждой группы геномов.

Ортологи белка AraR были обнаружены в 28 бактериях, относящихся четырех порядкам: *Bacillales*, *Lactobacillales*, *Clostridiales* и *Thermotogales*.

При построении матриц позиционных весов для распознавания сайтов связывания AraR была выбрана следующая стратегия. В начале была построена матрица по потенциальным сайтам из в 5'-некодирующих областей известных регулируемых оперонов в *B. subtilis* и их ортологов в ближайших геномах. Затем

производился поиск потенциальных сайтов связывания AraR в остальных группах геномов. Если сайты находились перед генами, принадлежащими системе утилизации

арабинозы, то они добавлялись в обучающую выборку. Если новая матрица давала большую предсказательную силу, то она сохранялась для следующей итерации. Если же поиск с помощью новой матрицы повышал количество перепредсказаний, не подтверждаемых методами проверки соответствия или функциональной причастностью потенциальных регулируемых генов к пути утилизации арабинозы, то либо сохранялась старая матрица, либо новые геномы относились к другой группе и для них строилась отдельная матрица. Для повышения специфичности поиска в некоторых случаях были использованы регуляторные последовательности из дополнительных геномов, не вошедших в список исследуемых. В результате было получено пять матриц для каждой из групп: *Bacillales*, *Lactobacillales*, *Clostridiales* и *Thermotogales* и *Clostridium acetobutylicum*.

Во всех изучаемых геномах консервативной оказалась регуляция гена *araD*, кодирующего фермент L-рибулозо-5-фосфат-4-эпимеразу, который соединяет путь утилизации арабинозы с пентозофосфатным путем. Для двух других ключевых ферментов пути утилизации арабинозы были найдены неортологичные замены. Так, кроме изученной

в *B. subtilis* L-рибулокиназы AraB была обнаружена широкая распространенность рибулокиназы AraK во всех порядках, кроме *Thermotogales*, где функция рибулокиназы была предсказана для белка AraB-II. Также было обнаружено неортологичное замещение L-арабинозоизомеразы на AraA-II в геномах *Clostridium cellulolyticum*, *Clostridium* sp. SS2/1 и *Thermotoga lettingae*.

Было показано широкое разнообразие генов новых транспортных систем под регуляцией AraR – как однокомпонентных пермеаз (*araE* и *araT*), так и ABC транспортеров (*araFGH*, *araN-II-araP-II-araQ-II*, *araN-III-araP-III-araQ-III* и *araT1T2T3*). К сожалению, точно определить специфичность к арабинозе или арабинозидам не предоставляется возможным. Также были обнаружены новые гены гидролаз, находящихся под AraR регуляцией (*abf3*, *abf4*, *abf5*, *xylX*, *xynB*, *glsA*). В *C. acetobutylicum* и *Clostridium beijerinckii* под регуляцией AraR были обнаружены гены ферментов пентозофосфатного пути, что, вероятно, позволяет быстрее метаболизировать арабинозу.

При изучении эволюции AraR регулона было обнаружено несколько ключевых особенностей:

- Наиболее часто под AraR регуляцией встречаются гены *araA*, *araK*, *araD* и *araE*, составляющие минимальный путь утилизации L-арабинозы.
- Именно в этом составе регулон сохраняется в порядке *Lactobacillales*, а также в геномах *Bacillus licheniformis*, *Bacillus pumilus* и *Oceanobacillus iheyensis*. При этом соответствующие белки AraR образуют на дереве отдельные далекие группы (Рис. 2).
- Структура оперонов и регулонов сильно различается даже внутри одного таксона.
- Единственным таксоном вне типа *Firmicutes*, где был найден ортолог AraR является порядок *Thermotogales*.



Рисунок 2. Филогенетическое дерево белков AraR

Белки указаны с помощью идентификаторов *locus_tag*. В скобках указаны трехбуквенные аббревиатуры геномов.

Таким образом, из анализа реконструированных регулонов и филогенетических деревьев можно сделать предположение о возможном пути эволюции AraR регулона. Вероятно, что первоначально AraR регуляция возникла в общем предке *Firmicutes* в виде простого регулона, состоящего из генов *araA*, *araB*, *araD* и *araE*, в каком виде он остался в *Lactobacillales*. Похоже, что в *C. acetobutylicum* произошел горизонтальный перенос одного из ранних вариантов регулона, где претерпел существенное расширение, захватив также гены пентозофосфатного пути *tkt*, *tal* и *ptk*. Об этом можно судить во-первых, по относительно недалекому расположению AraR из *C. acetobutylicum* от группы белков AraR из *Lactobacillales* на филогенетическом дереве (Рис. 2), во-вторых, по структуре сайтов связывания (Рис. 1). Также в отличие от других групп, у *Lactobacillales* и *C. acetobutylicum* обнаружено существенное снижение консервативности симметричных G и C в 7 и 14 позициях в сайтах связывания.

В общем предке *Bacillales* также присутствовал минимальный регулон в таком виде, как он сохранился в *B. licheniformis*, *B. pumilus* и *O. iheyensis*. Однако, затем произошла неортологичная замена AraK на AraB, и в состав регулона вошли системы деградации и транспорта поли- и олигоарабинозидов. Интересно, что в *B. licheniformis* под AraR регуляцией присутствуют два оперона, содержащие полные наборы генов для деградации арабинозы, *araKDAE* и *araABDMNPQ-abfA*. Похоже, что второй оперон необходим для утилизации олигоарабинозидов.

В *Clostridiales* AraR регулон претерпел существенные оперонные перестройки и приобрел новую двухкомпонентную систему регуляции AraI/AraJ, которая судя по консервативности расположения на хромосоме, регулирует транспорт арабинозы в клетку.

Наконец, в *Thermotogales*, наблюдаются последствия возможного горизонтального переноса из *Bacillales*. Данный вывод сделан на основе анализа генов, входящих в регулон. Так, AraB-II на филогенетическом дереве FGGY семейства киназ стоит достаточно близко к AraB, чтобы предположить расхождение от общего предка. К тому же, *Thermotogales* и *Bacillales* – это единственные таксоны, содержащие под AraR регуляцией ген *araM*.

Обнаруженная с помощью биоинформатического подхода функция L-рибулокиназы AraK из *C. acetobutylicum* была экспериментально проверена в совместной работе с лабораторией доктора Янг из Шанхайского Института биологических наук Китайской академии наук.

В четвертой главе описывается исследование AgaR регулона в протеобактериях. Фактор транскрипции AgaR, принадлежащий к семейству ДНК-связывающих белков DeoR, был впервые охарактеризован в *E. coli* в качестве репрессора генов утилизации N-ацетилгалактозамина (НАГА) и галактозамина (ГА). Для данного регулона была проведена как реконструкция регуляции транскрипции, так и подробная метаболическая реконструкция для каждого генома (Рис. 3).

Исследование AgaR регулона ограничивалось только протеобактериями. Ортологи фактора транскрипции AgaR были найдены в 19 геномах гамма-протеобактерий и двух геномах бета- и альфа-протеобактерий. Все найденные ортологи гена *agaR* кластеризуются на хромосоме с генами утилизации НАГА, что доказывает консервативность функции ортологов AgaR. Филогенетический анализ найденных в протеобактериях гомологов AgaR показал наличие пяти групп регуляторов (Рис. 4). В каждой из пяти групп 5'-некодирующие участки генов утилизации НАГА/ГА были использованы для проведения филогенетического футпринтинга. В результате были получены консервативные сайты ДНК, которые использовались для построения матриц позиционных весов и последующего сканирования геномов с целью поиска дополнительных сайтов.

В первых трех группах предполагаемые мотивы AgaR сайтов имеют одинаковую структуру прямых повторов с консенсусом СТТТС-5пн-СТТТС, тогда как число таких сайтов и их ориентация для промоторной области каждого регулируемого гена может различаться. В четвертой группе предсказанный регуляторный мотив является инвертированным повтором с консенсусом СТТТС-15пн-GAAAG. Наконец,

предсказанный мотив для пятой группы имеет структуру прямого повтора с консенсусом GAAAG-16-18нт-GAAAG.

Реконструкция AgaR регулонов в протеобактериях показала различные наборы генов, предположительно участвующих в утилизации НАГА/ГА. Значительная доля регулируемых AgaR генов кодирует ранее неизвестные ферменты и транспортные системы. Наиболее консервативным членом AgaR регулона является ген, кодирующий изомеразу AgaS. Также было показано, что белок AgaI, которому приписывали функцию ГА-6-фосфат деаминазы/изомеразы, встречается всего в двух организмах и тем самым не может выполнять данную роль. Это позволило предположить, что именно AgaS является ГА-6-фосфат деаминазой/изомеразой.

В организмах рода *Shewanella* кластер *aga* генов содержит ген, кодирующий НАГА-6-фосфат деацетилазу, имеющую наибольшее сходство с N-ацетилглюкозамин-6-фосфат деацетилазой NagA из бактерий рода *Shewanella*. Этот новый вариант НАГА-6-фосфат деацетилазы был назван AgaA-II. Все три фермента: AgaA из *E. coli*, AgaA-II и NagA из *Shewanella* принадлежат к одному семейству амидогидролаз (COG1820). Тем не менее, филогенетический анализ этого семейства подтвердил, что AgaA-II является паралогом NagA, что свидетельствует о недавней дупликации генов.

Гены, кодирующие гомологичные PTS системы были обнаружены под регуляцией AgaR в бактериях порядков *Enterobacteriales* и *Vibrionales*. Филогенетический анализ ПС компонент PTS, являющихся мембранными белками и определяющими их специфичность, показал, что на дереве хорошо различаются клады, соответствующие генам, лежащим в одном локусе с геном *agaA* и лежащим на хромосоме отдельно от последнего. Можно предположить, что гены PTS системы, сцепленные с *agaA* участвуют в транспорте НАГА, тогда как PTS системы, не связанные с *agaA*, специфичны к ГА.

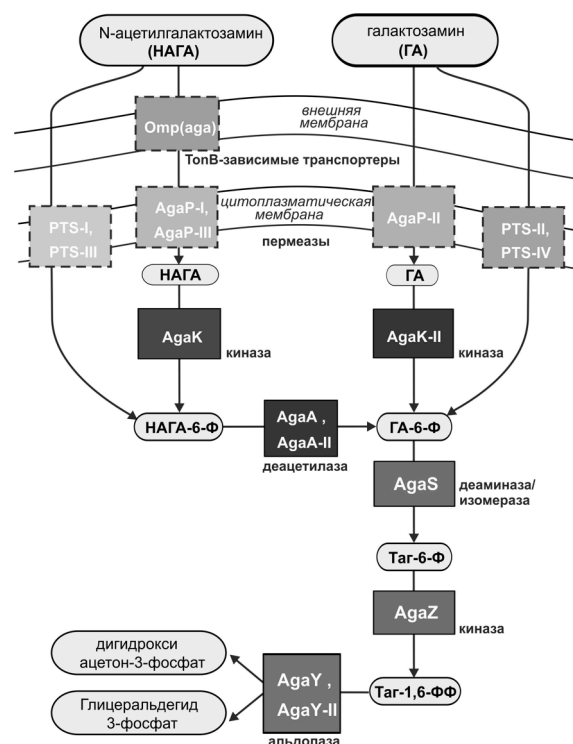


Рисунок 3. Реконструкция путей утилизации НАГА и ГА в протеобактериях. Метаболиты показаны в серых овалах. Белки показаны в прямоугольниках.

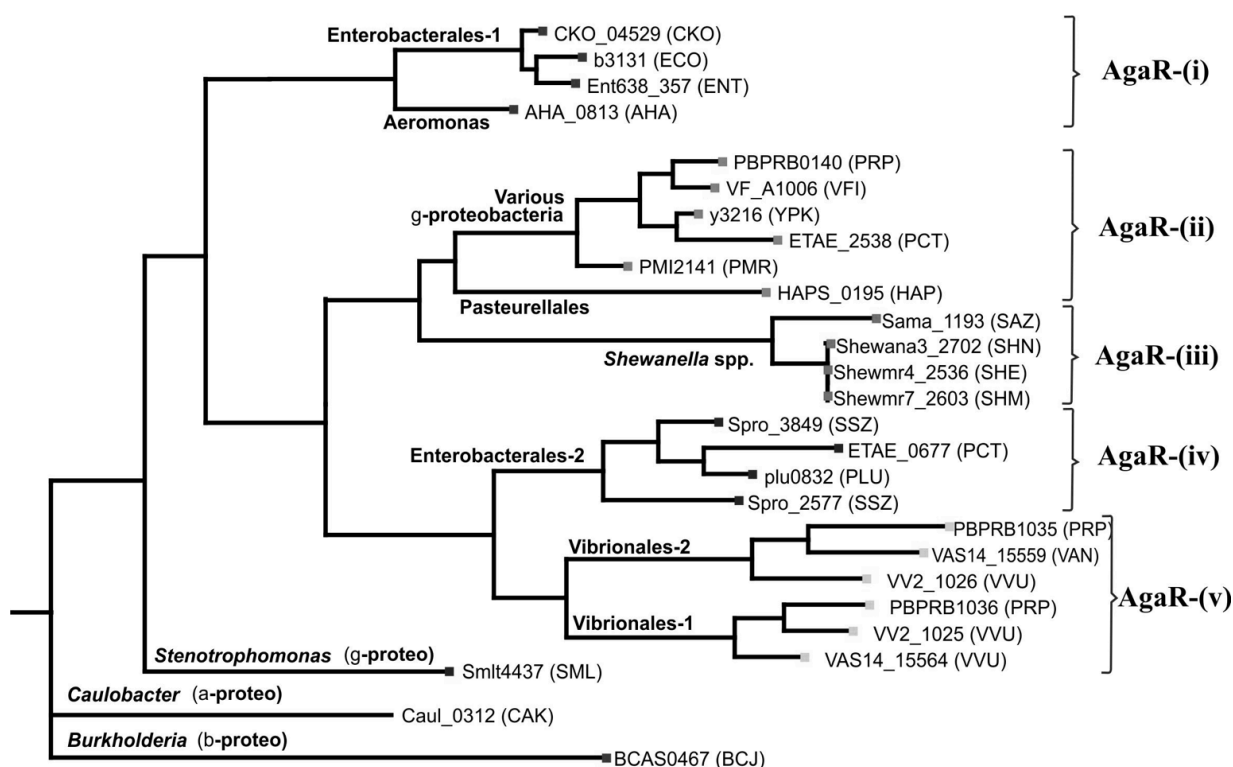


Рисунок 4. Филогенетическое дерево белков AgaR из протеобактерий.

В бактериях рода *Shewanella* и в некоторых других протеобактериях под регуляцией AgaR были обнаружены системы транспорта и последующего фосфорилирования субстратов отличные от PTS систем. Например, в бактериях рода *Shewanella* AgaR регулоны содержат новые гены, кодирующие предсказанные НАГА пермеазу AgaP, киназу AgaK и TonB-зависимый транспортер внешней мембраны Omp^(aga). Предсказанная пермеаза AgaP принадлежит к GGP семейству сахарных транспортеров и является ближайшим паралогом N-ацетилглюкозамин пермеазы из бактерий рода *Shewanella*. В AgaR регулоне патогенной бактерии *Stenotrophomonas maltophilia* также был найден ген *agaK*, однако, закодированная в этом же регулоне пермеаза AgaP-III принадлежит к другому семейству, называемому Sugar_tr. В *Caulobacter* sp. K31 и *Burkholderia cepacia* в *aga* кластере закодированы сахарная киназа AgaK-II из семейства VcgAD_BadFG и транспортер AgaP-II из семейства EamA. Поскольку в данных геномах не было обнаружено НАГА-6-фосфат деацетилазы, было предположено, что AgaP-II и AgaK-II участвуют в транспорте и фосфорилировании ГА, соответственно.

В реконструированных AgaR регулонах были обнаружены новые гликозил гидролазы, которые, как предполагается, участвуют в метаболизме НАГА/ГА. Так, в группе геномов *Shewanella* была найдена секретлируемая альфа-N-ацетилгалактозаминидаза AgaO, принадлежащая к семейству GH109 гликозил гидролаз. Геномы *Photobacterium luminescens* и *Aeromonas hydrophila* содержат ген гидролазы AgaH

из семейства GH36, которая скорее всего также является альфа-N-ацетилгалактозаминидазой.

Филогенетический анализ белков, входящих в путь утилизации НАГА позволяет предположить наиболее вероятные эволюционные сценарии появления AgaR регулонов в различных таксонах протеобактерий.

Наиболее интересным представляется происхождение оперона утилизации НАГА в бактериях рода *Shewanella*. Пермеаза AgaP и деацетилаза AgaA-II скорее всего образовались путем дупликации генов, кодирующих N-ацетилглюкозамин пермеазу NagP и N-ацетилглюкозаминдеацетилазу NagA у одной из предковых бактерий рода *Shewanella*. При этом гены *agaR*, *agaS* и *agaZ* были приобретены в результате горизонтального переноса.

Уникальный вариант пути утилизации ГА был обнаружен в *Haemophilus parasuis*. Оперон *agaRS-PTS-V-bgaZ-agaY-II* также кодирует две группы белков с различными эволюционными происхождением. К первой группе относятся белки AgaR и AgaS, наиболее похожие на аналогичные белки из *Enterobacteriales*. Ко второй группе относятся компоненты транспортера PTS-V и цитоплазматическая бета-галактозидаза BgaZ, для которых наиболее близкие гомологи находятся в бактериях типа Firmicutes, например в *Streptococcus pneumoniae*. Данные обстоятельства позволяют предположить, что часть AgaR регулона *H. parasuis* была перенесена горизонтальным переносом из Firmicutes.

На филогенетических деревьях различных белков из AgaR регулона обнаруживается парафилетическая группа из белков, принадлежащих бактериям двух таксонов – *Enterobacteriales* и *Vibrionales*. Сам кластер *aga* генов сохраняет свою структуру в этих видах. Из этого можно предположить, что в *Yersinia pestis*, *Edwardsiella tarda*, *Proteus mirabilis*, *Vibrio fisheri* и *Photobacterium profundum* локус генов утилизации НАГА был привнесён в результате недавних горизонтальных переносов из одного источника.

Наконец, в *Serratia proteamaculans* присутствуют два *aga* кластера. В обоих кластерах обнаружен паралог *araR* гена, но при этом остальные гены не дублированы. Скорее всего это является результатом дупликации изначального *aga* кластера в предковом организме с последующей потерей дублирующийся генов.

Таким образом было показано, что основными событиями в эволюционной истории AgaR регулона являются горизонтальный перенос и дупликации генов.

Сделанные предсказания новых метаболических ферментов были проверены на примере бактерии *Shewanella* sp. ANA-3 в совместной работе с лабораторией доктора Янг из Шанхайского института биологических наук Китайской академии наук.

Пятая глава посвящена изучению регуляции метаболизма сахаров фактором транскрипции HexR, принадлежащему RpiR семейству. Данный фактор транскрипции был ранее экспериментально изучен в бактериях рода *Pseudomonas*, где он регулирует гены пути Энтнера-Дудорова [Daddaoua et.al., 2009]. Также ранее была проведена реконструкция HexR регулона в бактериях рода *Shewanella* [Rodionov et.al., 2011].

В настоящей работе впервые была проведена реконструкция HexR регулона в геномах 62 гамма-протеобактерий и 25 бета-протеобактерий из 11 различных таксонов. В большинстве таксонов мотив сайтов связывания HexR можно записать в виде палиндромного консенсуса TGRAR-5нт-YTACA, где R – A или G, Y – C или T. Однако, обнаружилось и расхождение с этим мотивом. В бактериях семейства *Pseudomonadaceae* были найдены два паралога HexR. Было предсказано, что один паралог распознает стандартный консенсус, а другой – инвертированный повтор со спейсером переменной длины TGTTGT-4-8нт-ACAACAT. В бактериях родов *Nahella* и *Marinobacter* был обнаружен укороченный мотив с консенсусом GWAGTATACTWC, где W – A или T.

Анализ методами сравнительной геномики показал, что HexR регулоны значительно различаются между таксонами, как по размеру, так и по функциональному составу (Рис. 5). Так, если в группах *Enterobacteriales*, *Ralstonia* и *Burkholderia* HexR регулон является локальным и контролирует 1-2 оперона, то в группах *Aeromonadales*, *Vibrionales* и *Shewanella* число потенциальных регулируемых оперонов достигает до 20. Наиболее консервативные гены кодируют ферменты гликолиза (*gpmM*, *tpiA*, *pgl*, *gapA*, *pykA*, *glk* и *pgi*) и пути Энтнера-Дудорова (*zwf*, *edd* и *eda*). Также в HexR регулоне часто встречаются гены гликолиза, глюконеогенеза (*ppsA*, *gapB* и *pckA*), пентозофосфатного пути (*tal*), метаболизма пирувата (*aceEF* и *ppc*), ферментации (*adhE*, *pflBA* и *grcA*), глиоксилатного шунта (*aceBA*), биосинтеза аминокислот (*gltBD*) и окисления NADPH (*pntAB*).

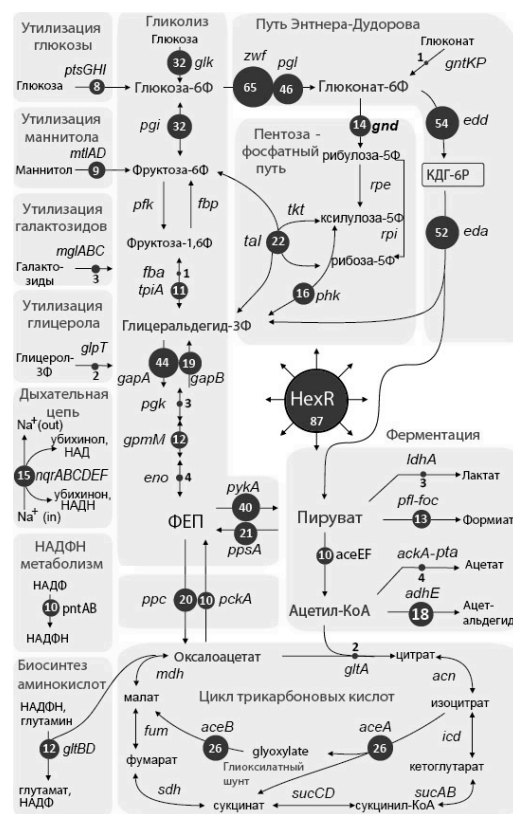


Рисунок 5. Метаболический контекст реконструированного HexR регулона в протеобактериях. Цифры в кружках обозначают количество регулонов, в которых присутствует данный ген.

Таксон-специфическая регуляция в HexR регулоне достаточно разнообразна, однако, в ней также преобладают гены метаболизма углерода.

Способность белка HexR связываться со всеми пятнадцатью предсказанными сайтами *S. oneidensis* была экспериментально проверена в лабораториях доктора Остермана из Института медицинских исследований Сэнфорд и Бернэма и доктора Янг из Шанхайского института биологических наук Китайской академии наук.

В шестой главе рассматривается изучение регуляции катаболизма сахаров в бактериях семейства *Bacillaceae*. Для анализа регуляторных систем, отвечающих за метаболизм углеводов среди бактерий семейства *Bacillaceae*, были отобраны десять геномов: *B. subtilis*, *Bacillus amyloliquefaciens*, *B. pumilus*, *B. licheniformis*, *Anoxybacillus flavithermus*, *Geobacillus kaustophilus*, *Bacillus cereus*, *Bacillus halodurans*, *Bacillus clausii* и *O. iheyensis*. Анализ литературных данных и базы данных регуляции транскрипции DBTBS показал, что в *B. subtilis* ранее было изучено 32 регулятора метаболизма углеводов и их производных, при этом для 11 из них не было показано мотивов сайтов связывания. В данной работе, с помощью анализа геномного контекста и изучения возможной авторегуляции генов факторов транскрипции было предсказано дополнительно новых 11 регулонов, участвующих в метаболизме сахаров. Таким образом, была реконструирована сеть транскрипционной регуляции, состоящая из 43 регулонов (Табл. 1).

Изученные регуляторы принадлежат к пятнадцати семействам факторов транскрипции. Однако, наибольшее число относится к двум семействам – LacI (11 регулонов) и GntR (10 регулонов). Всего 6 регулогов присутствуют во всех изучаемых геномах: глобальный регулог метаболизма углеводов CspA, регулог гликолиза CggR, регулог глюконеогенеза CspN, а также регулоги утилизации фруктозы (FruR), рибозы (RbsR) и лактата (LutR). При этом больше половины реконструированных регулогов содержатся в пяти или менее геномах.

В сеть регуляции метаболизма углеводов в бактериях семейства *Bacillaceae* входит более 300 ортологичных рядов регулируемых генов. *B. subtilis* предсказано более 60 новых регуляторных взаимодействий, из которых 42 сделано для CspA регулона. Среди новых членов CspA регулона в *B. subtilis* присутствуют гены утилизации различных сахаров: рамногалактуронана (*yesOPQRSTUVWXYZ-yetA-lplABCD*), галактуроната (*kduID*), фруктозы (*fruRKA*), сахарозы (*sacPA*), мальтозы (*cycB-ganPQA*), лактата (*lutABC*) и других. Также интересно, что сайты CspA были найдены перед генами, кодирующими ферменты цикла трикарбоновых кислот, таких как сукцинатдегидрогеназа (*sdhCAB-ysmA*), сукцинил-КоА лигаза (*sucCD*) и 2-оксоглутарат дегидрогеназа (*odhAB*).

Среди 11 новых реконструированных регулонов были найдены два регулона утилизации арил-бета-глюкозидов: BglR и YkvZ. Также было обнаружено, что регулятор DegA из контролирует экспрессию сцилло-инозитол дегидрогеназы IolX и предсказанной дегидрогеназы неизвестного производного инозитола YrbE в *B. subtilis*. Интересно, что в *G. kaustophilus* DegA регулирует экспрессию генов основного пути утилизации инозитола, заменив собой IolR. Не обнаружен IolR и в *B. halodurans*. Однако, там за регуляцию генов утилизации инозитола, как было предсказано, отвечает фактор транскрипции из LacI семейства IolR2.

Еще восемь новых регулонов отвечает за утилизацию различных сахаров и их производных: RhgR (раннее YesS) и RmgR (раннее YtdP) контролируют гены утилизации рамногалактуронана; FruR – гены утилизации фруктозы; GamR (ранее YbgA) – гены утилизации глюкозамина; MdxR – гены катаболизма мальтодекстрина; MsmR – гены утилизации альфа-галактозидов и MurR – гены утилизации N-ацетилмурамата.

Среди 43 реконструированных регулонов лишь 14 полностью сохраняют свой состав во всех геномах. При этом, среди них всего два присутствуют во всех анализируемых организмах – регулятор генов утилизации фруктозы FruR и генов глюконеогенеза CcpN. Также к группе абсолютно консервативных регулонов, присутствующих во всех геномах, можно было бы причислить CggR регулон, однако в *B. cereus* произошел переход генов *tpi* и *pgk* в псевдогены. В шести и пяти геномах, соответственно, были найдены регулоны утилизации ацетоина AcoR и рибулозо-монофосфатного пути HxlR. Остальные регулоны, входящие в эту группу (AlsR, CitR, GlvR, GntR, GutR, LevR, ManR, MdxR, MurR, NtdR), мало распространены среди изучаемых организмов и встречаются в двух-четырёх гномах.

При анализе консервативности реконструированных регулонов было отмечено, что для локальных регулонов основные отличия между организмами заключаются в изменении регуляции генов, ответственных за ранние шаги метаболических путей. Так, частым изменением в составе регулона является смена регуляции транспортеров субстратов, что было замечено в 12 регулонах (KdgR, IolR, LutR, MtlR, RbsR, NagR, MalR, CitT, GudR, RmgR, XylR, RhgR). Наблюдались как потери сайтов связывания перед генами транспортеров, так и потеря самого гена или появление новых генов транспортеров. Также вариабельны гены гидролаз, что показано, например, в регулогах MsmR, YkvZ, RmgR.

Часто различия состоят в неортологичном замещении генов с близкими или одинаковыми функциями. Регулон GamR интересен тем, что в *B. subtilis* состав регулона полностью отличается от GamR регулонов в других бактериях. Если в *B. subtilis* под регуляцией находятся гены, кодирующие путь утилизации глюкозамина, то в других

бактериях наиболее вероятным представляется путь деградации хитина до N-ацетилглюкозамина.

Можно резюмировать, что в семействе *Bacillaceae* наблюдалась большая пластичность транскрипционной регуляции, отвечающей за экспрессию генов метаболизма сахаров. Это также отражается в низкой консервативности CsrA регулона, объясняемой скорее всего необходимостью клетки быстро подстраивать новые периферические пути в общий метаболизм.

Таблица 1. Обобщенная таблица реконструированных регулонов. В колонке «Инф.» показана информация о литературных данных о реконструированных регулонов: «С» - были показаны сайты связывания белка-регулятора в изучаемых организмах, «С(др.)» - сайты связывания были показаны на других организмах, «Р» - были показаны только регулируемые гены, «0» - информации о регулоне в литературе не обнаружено.

Имя	Функция регулируемых генов	Семейство	Инф.	Количество регулируемых генов											
				BSU	BAY	BLI	BPU	BCE	AFL	GKA	BNA	BCL	OIN		
RhgR* (YesS)	Утилизация рамногалактуронана	AraC	0	17	0	19	0	0	0	0	0	0	0	0	0
RmgR* (YtdP)	Утилизация рамногалактуронана	AraC	0	5	0	5	5	0	0	0	0	11	11	7	7
LicR	Утилизация бета-глюкозидов	BglG	C	4	4	0	4	0	0	0	0	0	5	5	5
ManR	Утилизация маннозы	BglG	P	3	3	3	0	0	0	0	0	0	0	0	0
MtlR	Утилизация маннитола	BglG	C(др.), P	4	4	3	4	0	3	4	0	0	0	4	4
CcpN	Глюконеогенез	CcpN	C	2	2	2	2	2	1	2	2	2	2	2	2
CitT	Утилизация цитрата	CitB	C	1	1	1	1	1	0	0	1	1	1	3	3
MalR	Утилизация малата	CitB	C	4	3	3	3	0	0	0	0	0	0	0	0
FruR	Утилизация фруктозы	DeoR	0	3	3	3	3	3	3	3	3	3	3	3	3
IolR	Утилизация инозитола	DeoR	C	12	12	12	0	0	0	0	0	0	10	0	0
RhaR* (YuiB)	Утилизация рамнозы	DeoR	0	5	0	5	0	0	0	0	0	4	0	2	2
AcoR	Утилизация ацетона	Fis	C	4	4	1	4	4	0	0	0	4	0	0	0
AraR	Утилизация арабинозы	GntR	C	13	13	13	5	0	6	7	11	0	0	5	5
BglR* (YudK)	Утилизация бета-глюкозидов	GntR	0	3	2	0	2	0	0	0	0	0	0	0	0
FriR	Утилизация фруктозолизина	GntR	P	6	7	0	0	0	0	0	0	0	0	0	0
GamR* (YbgA)	Утилизация глюкозамина	GntR	0	3	0	7	0	0	0	0	0	6	6	6	6
GmuR	Утилизация глюкоманнана	GntR	P	8	7	6	5	0	0	0	5	5	0	0	0
GntR	Утилизация глюконата	GntR	C	4	0	4	0	0	0	0	0	0	0	3	3
GudR	Утилизация глюкората; Утилизация галактората	GntR	P	6	0	6	0	0	0	0	0	0	6	7	7
LutR	Утилизация лактата	GntR	P	5	5	5	3	3	5	4	4	4	4	5	5

Имя	Функция регулируемых генов	Семейство	Инф.	BSU	BAY	BLI	BPU	BCE	AFL	GKA	BNA	BCL	OIN
NagR	Утилизация N-ацетилглюкозамина	GntR	C	4	4	5	4	3	0	3	4	2	1
TreR	Утилизация трехалозы	GntR	C	3	3	3	3	2	3	3	3	2	0
GutR	Утилизация сорбитола	GutR	C	3	3	0	0	0	0	0	0	0	0
HxiR	Рибулоза монофосфатный путь	HxiR	C	2	2	2	2	0	0	0	0	0	2
SspA	Метаболизм углерода	LacI	C	217	171	195	144	44	50	44	119	119	81
SspB	Метаболизм углерода	LacI	P	6	2	0	0	0	0	0	0	0	0
DegA	Утилизация инозитола	LacI	0	2	0	1	0	0	0	13	6	0	0
ExuR	Утилизация глюкороната; Утилизация галактуроната	LacI	C	10	5	5	4	0	0	0	0	0	0
GanR	Утилизация галактана	LacI	P	6	0	6	6	0	0	0	5	5	0
KdgR	Утилизация 2-кето-3-деоксиглюконата	LacI	C	6	0	6	5	0	0	0	0	0	0
MdxR	Утилизация мальтодекстрина	LacI	0	7	0	7	0	0	0	0	0	0	0
MsmR	Утилизация альфа-галактозидов	LacI	0	5	5	5	5	0	0	0	5	0	0
NtdR	Утилизация неотрехалозамина	LacI	P	4	0	4	4	0	0	0	0	0	0
RbsR	Утилизация рибозы	LacI	C(др.)	7	6	6	6	5	6	6	6	6	6
YkvZ	Утилизация бета-глюкозидов	LacI	0	2	2	1	1	2	0	0	0	0	0
LevR	Утилизация левана	LevR	C	5	0	5	0	0	0	0	0	0	0
AlsR	Биосинтез ацетоина	LysR	P	3	3	3	3	0	0	0	0	0	0
SspC	Метаболизм цитрата	LysR	C	5	4	2	2	1	0	0	0	0	2
CitR	Метаболизм цитрата	LysR	P	2	2	2	0	0	0	0	0	2	0
XylR	Утилизация ксиллозы	ROK	C	5	5	10	5	0	0	8	12	6	7
GivR	Утилизация мальтозы	RpiR	P	3	3	3	0	0	0	0	0	0	0
MurR	Утилизация N-ацетилмурамата	RpiR	0	6	0	6	6	0	0	0	0	0	0
SggR	Гликолиз	SogC	C	6	6	6	6	2	6	6	6	6	6
Количество регулонов:				43	30	38	28	12	9	13	19	17	19

* - Новое имя регулона, данное в этой работе. Старое название дано в скобках.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. Проведена реконструкция регулонов утилизации L-арабинозы, включающая в себя определение потенциальных сайтов связывания белка AraR, оперонной структуры и функциональной аннотации регулируемых генов в четырех порядках бактерий: *Bacillales*, *Lactobacillales*, *Clostridiales* и *Thermotogales*. Предсказаны неортологичные замещения ферментов L-рибулокиназы и L-арабиноизомеразы у ряда бактерий. Показано широкое разнообразие ферментов деградации арабинозосодержащих полисахаридов и систем транспорта внутрь клетки продуктов деградации. Предложена модель эволюции AraR регулона, включающая таксон-специфичное расширение регулона в определенных группах бактерий.
2. Проведена реконструкция регулонов утилизации N-ацетилгалактозамина (НАГА) у протеобактерий, включающая в себя определение потенциальных сайтов связывания белка AgaR, оперонной структуры и функциональной аннотации регулируемых генов. Показано, что в эволюции AgaR регулона основными движущими силами являются горизонтальный перенос и дубликации генов. Филогенетический анализ семейств белков участвующих в транспорте и деацетилировании аминокислот позволил предложить эволюционную модель их возникновения путем дубликации и специализации кодирующих их генов. Реконструкция метаболических путей утилизации НАГА и производных сахаров у протеобактерий позволила обнаружить ряд новых систем транспорта данных сахаров в клетку и ферментов осуществляющих начальные стадии катаболизма аминокислот в цитоплазме.
3. Проведена реконструкция регулонов транскрипционного фактора HexR контролирующего центральный метаболизм углерода в геномах гамма- и бета-протеобактерий. Каждый регулон включает в себя определение потенциальных сайтов связывания белка HexR, оперонной структуры и функциональной аннотации регулируемых генов. Впервые показано, что HexR является глобальным регулятором генов центрального метаболизма углерода в бактериях порядков: *Vibrionales*, *Aeromonadales*, *Psychromonadales* и *Alteromonadales*. Сравнение функционального состава HexR регулонов позволило определить группы консервативных и изменчивых генов регулона.
4. Проведен широкомасштабный анализ регуляторных систем, отвечающих за метаболизм углеводов и их производных в семействе бактерий *Bacillaceae*. Проведена реконструкция 43 регулонов, для каждого из которых были

определены ДНК мотивы сайтов связывания, оперонная структура и функциональная аннотация регулируемых генов. Для 22 регулонов мотивы сайтов связывания белка-регулятора были обнаружены впервые. Показана высокая вариабельность данной регуляторной сети у различных бактерий из семейства *Bacillaceae*.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в научных журналах:

1. Leyn, S.A., Li, X., Zheng, Q., Novichkov, P.S., Reed, S., Romine, M.F., Fredrickson, J.K., Yang, C., Osterman, A.L., Rodionov, D.A. Control of Proteobacterial Central Carbon Metabolism by the HexR Transcriptional Regulator. *Journal of Biological Chemistry* (2011); **286**(41): 35782-35794.
2. Zhang, L., Leyn, S.A., Gu, Y., Jiang, W., Rodionov, D.A., Yang, C. Ribulokinase and transcriptional regulation of arabinose metabolism in *Clostridium acetobutylicum*. *Journal of bacteriology* (2012); **194**(5): 1055-1064.
3. Leyn, S.A., Gao, F., Yang, C., Rodionov, D.A. N-Acetylgalactosamine Utilization Pathway and Regulon in Proteobacteria: Genomic Reconstruction And Experimental Characterization In *Shewanella*. *Journal of Biological Chemistry* (2012); **287**(33): 28047-28056.

Тезисы конференций:

1. Лейн С.А., Равчеев Д.А. Эволюция химерного белка: исследование AraR зависимой регуляции методами сравнительной геномики. Труды 32-й конференции молодых ученых и специалистов ИППИ РАН ИТиС'09. 2009, с. 303-304.
2. Dmitry Rodionov, Xiaoqing Li, Samantha Reed, Margaret Romine, James Fredrickson, Pavel Novichkov, Semen Leyn, Andrei Osterman. Transcription regulation of *Shewanella* Central Carbon Metabolism by HexR. Proceedings of 2010 Genomic Science Contractor-Grantee and Knowledgebase Workshop. 2010, pp. 139-140.
3. D.A. Rodionov, X. Li, A. Osterman, S. Leyn. Comparative and Functional Genomics of the Central Carbon Metabolism Regulon HexR in Proteobacteria. Proceedings of American Society for Microbiology. 110th General Meeting. 2010, R-2867.

4. Semen Leyn, Dmitry Rodionov. HexR – new central carbohydrate metabolism transcription regulator. Comparative approach study. Труды 33-й конференции молодых ученых и специалистов ИППИ РАН ИТиС'10. 2010, с. 345-349.
5. Semen Leyn, Xiaoqing Li, Andrei Osterman, Dmitry Rodionov. Comparative and functional genomics of the central carbon metabolism regulon HexR in Proteobacteria. Proceedings of 3rd Annual Joint Conference on Systems Biology, Regulatory Genomics, and Reverse Engineering Challenges 2010 – DREAM5. 2010, p. 211.
6. Pavel Novichkov, Alexey Kazakov, Semen Leyn, Dmitry Ravcheev, Andrei Osterman, Inna Dubchak, Adam Arkin, Dmitry Rodionov. Large-Scale Genomic Reconstruction of Transcriptional Regulatory Networks in Bacteria. Genomic Science Awardee Meeting IX and USDA-DOE Plant Feedstock Genomics for Bioenergy Awardee Meeting. 2011, pp. 199-200.
7. Semen Leyn, Marat Kazanov, Pavel Novichkov, Dmitry Rodionov. Reference collection of transcriptional regulons in Bacillales family of bacteria. Proceedings of Moscow Conference on Computational Molecular Biology 2011. 2011, pp. 202-203.
8. Semen Leyn, Fang Gao, Chen Yang, Dmitry Rodionov. Comparative genomic reconstruction of N - acetylgalactosamine catabolic pathways and transcriptional regulons in Proteobacteria. Proceedings of Moscow Conference on Computational Molecular Biology 2011. 2011, pp. 204-205.
9. Semen Leyn, Dmitry Rodionov. Comparative genomic reconstruction of N-acetylgalactosamine catabolic pathways and transcriptional regulons in Proteobacteria. Труды 34-й конференции молодых ученых и специалистов ИППИ РАН ИТиС'11. 2011, с. 45-47.
10. Pavel Novichkov, Dmitry Ravcheev, Alexey Kazakov, Semen Leyn, Adam Arkin, Inna Dubchak, Dmitry Rodionov. Toward System Biology KnowledgeBase on Transcriptional Regulation in Bacteria. Proceedings of 2012 Genomic Science Awardee Meeting X. 2012, pp. 187-188.
11. Semen Leyn, Marat Kazanov, Pavel Novichkov, Dmitry Rodionov. Reference Collection of Transcriptional Regulons in Bacillales. Proceedings of 2012 Genomic Science Awardee Meeting X. 2012, p. 188.