Российская академия наук

Федеральное государственное бюджетное учреждение науки

Институт проблем передачи информации им. А.А. Харкевича

На правах рукописи

Суворова Инна Андреевна

Коэволюция транскрипционных факторов семейства GntR и их сайтов связывания

Диссертация на соискание ученой степени
кандидата биологических наук

Специальность 03.01.09 – математическая биология, биоинформатика

Научный руководитель:
к.ф.-м.н., д.б.н., профессор Михаил Сергеевич Гельфанд

Москва – 2016

Оглавление

Введение

Актуальность темы

Бактерии способны приспосабливаться к самым разным, меняющимся условиям окружающей среды [1,2]. Подобная адаптация осуществляется за счет изменения экспрессии генов, что позволяет клетке эффективно использовать имеющиеся ресурсы. Такая стратегия требует сложной системы регуляции, обеспечивающей адекватный ответ на внешние или внутриклеточные стимулы [1,2,3,4]. Регуляция экспрессии генов осуществляется на разных уровнях: транскрипции, трансляции, посттрансляционной модификации, однако наиболее эффективным и распространенным вариантом является регуляция на стадии инициации транскрипции [5]. Ключевой элемент такой регуляции – факторы транскрипции, специальные белки-регуляторы [1,4,6,7].

До недавнего времени исследование транскрипции проводилось исключительно экспериментальными методами, но в настоящее время развитие методов секвенирования и экспоненциальный рост количества данных о нуклеотидных и аминокислотных последовательностях привели к широкому и успешному использованию биоинформатических методов [4,7]. Подобные исследования часто применяются в качестве дополнения к эксперименту, однако изучение регуляции может осуществляться и исключительно методами сравнительной геномики [4,7,8]. Основной задачей биоинформатических исследований является выявление разнообразных регуляторных последовательностей, например, промоторов, сайтов связывания транскрипционных факторов и т.д. [4,9,10,11].

Роль регуляторных взаимодействий весьма велика, и сравнительный анализ регуляции экспрессии генов у различных бактерий позволяет делать выводы об эволюции функциональных систем и самих микроорганизмов, а также особенностях их взаимодействия с окружающей средой [3,7]. Таким образом, исследование ДНК-белковых взаимодействий и регуляции транскрипции является актуальной задачей современной молекулярной биологии и сравнительной геномики [7,8,9,12,13,14].

Цели и задачи исследования

Целью данной работы было исследование одного из наиболее распространенных среди бактерий семейств транскрипционных факторов, GntR, методами сравнительной геномики.

В работе решаются следующие общие и частные задачи:

1. Реконструкция регулонов транскрипционных факторов семейства GntR методами сравнительной геномики, построение распознающих правил для поиска их потенциальных сайтов связывания на основании результатов исследования 5’-регуляторных областей.

2. Исследование коэволюции мотивов связывания и аминокислотных последовательностей регуляторов транскрипции подсемейств FadR, HutC и YtrA семейства GntR путем анализа корреляций аминокислот ДНК-связывающих HTH-доменов транскрипционных факторов и нуклеотидов соответствующих сайтов связывания, предсказание вероятных ДНК-белковых взаимодействий.

3. Анализ особенностей структуры и расположения сайтов связывания регуляторов семейства GntR – исследование дивергонов, а также дополнительных боксов (полусайтов, симметричных элементов палиндромного мотива) у сайтов связывания.

4. Исследование регуляции метаболизма гексуронатов у Gammaproteobacteria родственными транскрипционными факторами UxuR и ExuR методами сравнительной геномики, разделение их мотивов связывания и построение распознающих правил для предсказания сайтов связывания, реконструкция гексуронатных регулонов, исследование оперонной структуры и идентификация новых членов регулонов, построение вероятных сценариев эволюции этой метаболической системы.

5. Исследование регуляции метаболизма малоната и пропионата у Proteobacteria транскрипционными факторами MatR/MdcY, MdcR, PrpR методами сравнительной геномики, выявление новых регуляторов метаболизма малоната и пропионата, построение распознающих правил для предсказания сайтов связывания и реконструкция соответствующих регулонов, исследование оперонной структуры и идентификация новых членов регулонов, построение возможной модели эволюции этих метаболических систем.

Научная новизна и практическое значение работы

В работе впервые исследован целый ряд транскрипционных факторов семейства GntR в различных таксономических группах, предсказаны их потенциальные сайты связывания и ДНК-белковые взаимодействия, реконструированы регулоны. Кроме того, обобщены сведения о расположении и структуре сайтов связывания.

Также было проведено детальное исследование регуляции метаболизма гексуронатов у Gammaproteobacteria и метаболизма малоната и пропионата у Proteobacteria. С помощью методов сравнительной геномики были обнаружены ранее неизвестные члены регулонов этих метаболических систем, показана вариабельность организации регулонов и их регуляции, в частности, были идентифицированы новые регуляторы метаболизма малоната и пропионата и

выявлены их потенциальные сайты связывания. Кроме того, в работе были предложены потенциальные сценарии эволюции регулонов метаболизма гексуронатов, а также малоната и пропионата.

Работа имеет теоретический характер, однако полученные данные потенциально могут найти применение в области биотехнологии и генной инженерии.

Апробация работы

Основные положения диссертации были представлены на российских и международных конференциях: Информационные технологии и системы ИТиС'09 (Бекасово, декабрь 2009), Ломоносов-2010 (Москва, апрель 2010), Информационные технологии и системы ИТиС'10 (Геленджик, сентябрь 2010), Постгеномные методы анализа в биологии, лабораторной и клинической медицине (Москва, ноябрь 2010), Информационные технологии и системы ИТиС'11 (Геленджик, октябрь 2011), Молекулярная и клеточная биология: прикладные аспекты (Москва, апрель 2012), Информационные технологии и системы ИТиС'12 (Петрозаводск, август 2012), Moscow Conference on Computational Molecular Biology MCCMB'13 (Москва, июль 2013), и на научных встречах международной учебно-научной группы «Regulation and Evolution of Cellular Systems (RECESS)» (Мюнхен, Германия, май 2011; Москва, июнь 2012; Венеция, Италия, май 2013).

Глава 1. Обзор литературы

1.1 Основные принципы регуляции экспрессии генов. Регуляция на уровне транскрипции

Несмотря на относительную простоту организации, прокариотическим организмам необходимо иметь немногим менее сложные, чем у эукариот, системы контроля и регуляции основных жизненных функций [1,2,3,4]. Бактерии способны адаптироваться к самым разнообразным, быстро меняющимся условиям среды, и это приспособление достигается за счет изменения экспрессии генов [1,2]. Существенная часть генов экспрессируется только в определенных условиях, позволяя эффективно использовать и экономить ресурсы организма, однако такая стратегия требует наличия сложной системы регуляции метаболизма, обеспечивающей быстрое изменение экспрессии генов определенных метаболических путей в ответ на соответствующее изменение условий среды [1,2,3,4]. Экономия ресурсов и повышение эффективности происходящих в клетке процессов при этом осуществляется, например, за счет выключения ряда энергозатратных реакций биосинтеза. Кроме того, избыточный синтез белка может приводить к негативным последствиям для клетки в результате его накопления и нарушения метаболических процессов: например, известно, что повышенный уровень экспрессии большинства трансмембранных белков и транскрипционных факторов, а также многих ферментов токсичен [15,16].

Таким образом, регуляция экспрессии генов позволяет бактериям приспосабливаться к самым различным условиям, эффективно использовать имеющиеся в геноме возможности в зависимости от текущих нужд клетки. Подобная регуляция возможна на любой стадии пути от ДНК к белку (транскрипция, трансляция и т.п.) [5,6]. Соответственно, существует несколько уровней регуляции экспрессии генов:

1. **дотранскрипционный** – может осуществляться за счет метилирования [17] или же изменения сверхспирализации ДНК [18,19];

2. **транскрипционный** – осуществляется преимущественно на стадии инициации транскрипции (например, за счет взаимодействия с регуляторными белками [3,4,5,12,20], участия альтернативных σ-факторов [5,21,22]), а также на стадиях элонгации и терминации (задержка за счет образования шпилек РНК, взаимодействие с белками (например, белками холодового шока, белками-антитерминаторами [6,23,24,25]), регуляция с помощью РНК-переключателей [25,26,27,28], Т-боксов [25,29], аттенюаторов транскрипции [6,25,30] и т.п.);

3. **посттранскрипционный/трансляционный** – заключается в изменении стабильности мРНК (например, усиление или защита от деградации за счет ассоциации с малыми некодирующими или антисмысловыми РНК [2,6,26,31]), или же может осуществляться через контроль инициации или элонгации трансляции (например, активация или ингибирование трансляции в результате формирования альтернативных вторичных структур РНК (РНК-переключатели, Т-боксы и т.п.) [25,26,28,29], связывания транскрипта с малыми некодирующими РНК [2,26,31] или белками (например, с факторами инициации трансляции [6,32], белками холодового шока, РНК-хеликазами [23,24,32], рибосомными белками [6,33] и т.п.), запрограммированный сдвиг рамки считывания [6]).

4. **посттрансляционный** – осуществляется за счет регуляции фолдинга, стабильности и активности белков [6,34] посредством белок-белковых взаимодействий [34,35,36], ковалентной модификации [4,5,37,38,39] и аллостерической регуляции при взаимодействии с низкомолекулярными веществами-эффекторами (как правило, такими способами регулируется активность ферментов, а также ДНК-связывающие свойства факторов транскрипции) [6,40,41,42,43].

Регуляция экспрессии на разных этапах осуществляется с различной скоростью и служит для разных целей. Так, аллостерическая регуляция представляет собой наиболее быстрый и лабильный вариант регуляции, который позволяет организму своевременно реагировать на изменяющиеся условия [6,42]. Регуляция при помощи альтернативных структур мРНК и малых некодирующих РНК также осуществляется с высокой скоростью, что обеспечивает быстрый ответ на внешние и внутренние стимулы [26]. В то же время, регуляция на уровне транскрипции является более медленной реакцией на изменение условий среды и служит в основном для оптимизации использования имеющихся ресурсов клетки и инактивации ненужных в данных условиях процессов [16]. При этом такой вариант регуляция является более тонким и точным, нежели глобальная регуляция экспрессии за счет изменения топологии ДНК [18,19]. Именно регуляция экспрессии на стадии инициации транскрипции наиболее эффективна и распространена, а также наиболее изучена [5].

Ключевыми составляющими регуляции транскрипции являются РНК-полимераза и факторы транскрипции – специальные регуляторные белки [1,4,6,7,12].

1.1.1 РНК-полимераза

ДНК-зависимая РНК-полимераза, ключевой фермент транскрипции, представляет собой крупный (~480 кДа) комплекс из нескольких субъединиц [6,22,44]. Главный компонент РНК-полимеразы (кор) включает две α-субъединицы, ββ' комплекс и ω-субъединицу [5,6,22,44,45].

Каждая α-субъединица состоит из двух доменов, соединенных гибким линкером: С-концевой домен участвует в связывании РНК-полимеразы с ДНК и необходим для правильного взаимодействия фермента с промоторами и транскрипционными факторами, N-концевой домен связывает остальные компоненты РНК-полимеразы и служит для димеризации α-субъединиц [5,6,46,47]. β-субъединица обладает полимеразной активностью, β'-субъединица неспецифически связывается с ДНК, вместе они формируют каталитический активный центр фермента [5,6]. Роль ω-субъединицы полностью не ясна, возможно, она выполняет структурную функцию и способствует поддержанию дееспособной формы РНК-полимеразы; так, обнаружено, что она защищает и стабилизирует конформацию β'-субъединицы [5,48,49].

Главный компонент РНК-полимеразы способен к элонгации, но не к инициации транскрипции, так как сам по себе не связывается с промотором [5,6,22,44,45]. Узнавание промотора у бактерий требует ассоциации с еще одной субъединицей – σ-фактором [6,22,44,45,46]. σ-фактор значительно снижает сродство РНК-полимеразы к неспецифическим областям ДНК, в то же время повышая ее сродство к определенным промоторам, и обеспечивает расплетание двойной спирали ДНК в области старта транскрипции [5,21,45]. σ-факторы – белки, в среднем размером 20-70 кДа, состоящие из четырех, реже трех доменов (у ряда альтерантивных σ-факторов 1-й домен отсутствует) [5,22,50]. Домены 2, 3 и 4 взаимодействуют с различными элементами промотора (см. далее), функция же домена 1 точно не выяснена, возможно, он играет роль в изменениях конформации σ-фактора, открывая или закрывая ДНК-связывающие домены, и влияет на эффективность инициации транскрипции в зависимости от последовательности промотора [5,22,50].

1.1.2 Строение промотора и механизм связывания РНК-полимеразы

Промотор – участок ДНК в цис-положении относительно регулируемых генов, с которым происходит связывание РНК-полимеразы и который определяет точку начала транскрипции. Он является минимальной последовательностью, которая специфически распознается РНК-полимеразой [6]. В зависимости от типа σ-фактора конкретная последовательность промотора варьирует (см. далее) [21,22,45].

Бактериальные промоторы содержат специфические элементы, узнаваемые РНК-полимеразой. Наиболее важными для распознавания элементами стандартного σ^{70}-промотора (см. далее) являются гексануклеотидные последовательности –35 (TTGACA) и –10 (TATAAT), расположенные на указанном расстоянии относительно старта транскрипции, взаимодействующие с 4-м и 2-м доменами σ-фактора, соответственно [5,6,21,22,44,50,51,52]. Расстояние между этими последовательностями оказывает влияние на активность промотора:

длина отрезка в 17 нуклеотидов является оптимальной и наиболее распространенной [6,52]. Кроме того, ряд промоторов имеет два дополнительных элемента: удлиненный –10 сайт, включающий короткую последовательность TGN непосредственно перед –10 элементом промотора и взаимодействующий с 3-м доменом σ-фактора [5,6,22,51,52]; а также UP-элемент, представляющий собой AT-богатую последовательность перед –35 сайтом, обычно с консенсусом NNAAAWWTWTTTTNNAAAANNN, которая взаимодействует с C-концевым доменом α-субъединицы РНК-полимеразы, существенно усиливая уровень транскрипции [5,6,46,47,22,52]. В целом, активность промотора коррелирует с числом и взаиморасположением различных его элементов, а также степенью их сходства с консенсусной последовательностью [5,6,52].

Часто узнавания вышеупомянутых промоторных последовательностей достаточно для полноценного связывания РНК-полимеразы, но в ряде случаев (например, при использовании некоторых альтернативных σ-факторов) для нормальной инициации транскрипции необходимо присутствие активатора [22].

1.1.3 Стадии транскрипции

Процесс транскрипции включает три основных стадии: инициацию, элонгацию и терминацию [6,44]. При этом инициация транскрипции – сложный стадийный процесс, что позволяет осуществлять ее точную и эффективную регуляцию и, как было отмечено выше, делает регуляцию на стадии инициации транскрипции наиболее распространенной [5,44]. После связывания с ДНК молекулы РНК-полимеразы перемещаются вдоль двойной спирали ДНК, осуществляя поиск промоторов, на которых происходит формирование инициационных комплексов [6,44]. При связывании РНК-полимеразы с ДНК происходит ее локальное расплетание приблизительно в области между –10 и +3 относительно старта, образуется открытый комплекс и инициируется транскрипция [5,6,44,53]. После синтеза РНК длиной более 9-11 нуклеотидов фермент покидает промотор, σ-фактор отделяется от инициационного комплекса, формируется стабильный элонгирующий комплекс, и далее элонгация осуществляется одним только главным компонентом РНК-полимеразы [5,6,22,44,53]. Элонгация заканчивается по достижении молекулами РНК-полимеразы специальных регуляторных участков ДНК (терминаторов транскрипции), после чего происходит освобождение новосинтезированных РНК из транскрипционных комплексов [6,44,54]. Типичные терминаторы, не требующие для своего распознавания РНК-полимеразой дополнительных белковых факторов, содержат симметричный GC-богатый участок, способный образовывать устойчивую шпильку, за которым располагается олиго(T)-последовательность, на

которой и завершается транскрипция [6,54,55].

Помимо описанных выше, у большинства бактерий существуют терминаторы транскрипции, распознаваемые РНК-полимеразой только в присутствии белкового фактора терминации Rho [6,54,55]. Этот фактор представляет собой гексамерную АТФ-зависимую ДНК/РНК хеликазу/транслоказу [6,54]. Фактор Rho связывается с новосинтезированной РНК в специфических участках, перемещается вдоль цепи и, взаимодействуя с РНК-полимеразой, вызывает прекращение элонгации и высвобождение РНК из транскрипционного комплекса [54,56].

1.1.4 Основной и альтернативные σ-факторы РНК-полимеразы

Большая часть промоторов у *E. coli* распознается основным сигма-фактором σ^{70} (RpoD) – это ключевой σ-фактор, при участии которого транскрибируются необходимые для поддержания основных жизненных функций гены («housekeeping»), обеспечивающие, например, репликацию ДНК, транскрипцию, трансляцию, центральный метаболизм и т.п. [5,6,22,57]. Кроме основного, у бактерий существует ряд альтернативных σ-факторов, которые активируются в определенных условиях, например, при окислительном, осмотическом или тепловом стрессе или в процессе морфогенеза [5,21,57,58]. Эти σ-факторы распознают меньшее число промоторов, контролируя экспрессию конкретных генов, необходимых в данных условиях [22,57,58]. Селективность связывания промоторов РНК-полимеразой, контролируемая сменой σ-факторов – один из ключевых механизмов глобального изменения и синхронной регуляции экспрессии множества генов у бактерий [21,22]. Инактивация основного σ-фактора летальна, альтернативных же – как правило, нет [22].

Число таких дополнительных σ-факторов варьирует в зависимости от вида бактерий, большинство имеет несколько альтернативных σ-факторов, помимо основного [5,22,44,57]. Так, у *Mycoplasma genitalium*, известной небольшим размером генома, всего 2 из примерно 500 генов кодируют σ-факторы [22], тогда как у бактерий с более крупным геномом и сложным жизненным циклом, например, *Streptomyces* spp., обычно присутствует несколько десятков различных альтернативных σ-факторов [5,58,59]. σ-факторы делятся на два основных неродственных семейства: σ^{70} и σ^{54} [4,21,22,45,58].

Семейство σ^{70}, распознающее стандартные –35/–10 промоторы, включает в себя ряд факторов, родственных главному σ-фактору (PA, «primary alternative»), а также содержит большое число разнообразных ECF σ-факторов («extracytoplasmic function») [4,21,22,57,58]. PA σ-факторы могут контролировать экспрессию генов, участвующих в процессе споруляции,

ответе на тепловой шок и общий стресс, а также генов жгутикового аппарата и т.п. [4,21,22,57]. ECF σ-факторы контролируют самые разные процессы, такие как синтез секретируемых веществ, импорт и экспорт ионов, ответ на внешние стимулы и стрессовые воздействия, а также регулируют гены, участвующие в проявлении патогенности [4,21,22,57,58]. Разнообразие регуляции и разница в числе σ-факторов у бактерий, особенно родственных, часто обусловлена именно ECF σ-факторами, содержание которых возрастает приблизительно пропорционально размеру генома и может составлять подавляющее большинство σ-факторов организма (у *Streptomyces coelicolor* – 51 ECF из 65), тогда как количество PA σ-факторов сравнительно постоянно (в среднем, 10-20) [57,58,59]. У *Bacillus subtilis*, модельного представителя грамположительных бактерий, найдено 17 альтенативных σ-факторов, среди которых 7 ECF [22,57,58]. У модельной грамотрицательной бактерии *E. coli* обнаружено шесть альтенативных σ-факторов, среди которых два ECF [22,58].

Семейство σ^{54} содержит факторы, которые контролируют различные процессы, включая азотный метаболизм, катаболизм аминокислот, формирование биопленок и проявления патогенности, узнают нетипичный высококонсервативный промотор –24 (TGGCACG)/–12 (TTGCW) и, в отличие от σ^{70}, требуют для работы присутствия белка-активатора [4,21,22,45,60].

Номенклатура σ-факторов неоднозначна и имеет целый ряд вариантов: многие σ-факторы обозначаются буквенным индексом (например, σ^{H} или σ^{E}), индексом, соответствующим их молекулярной массе (σ^{28}, σ^{32} и т.п.) или наименованию гена (например, σ^{FecI}), кроме того, используются также названия соответствующих генов (например, *rpoD* или *sigA*) [21,22,58].

σ-факторы, как правило, не детектируют изменение условий сами, а обычно являются конечным или (реже) промежуточным звеном какого-либо регуляторного каскада. В связи с этим σ-факторы обычно сами подвержены регуляции, причем осуществляться она может на транскрипционном, трансляционном и посттрансляционном уровнях (регуляция синтеза *de novo,* процессинг неактивных предшественников, ковалентная модификация, связывание с регуляторными белками, анти-σ-факторами) [5,22,57,59].

Каждый σ-фактор имеет собственный регулон (совокупность оперонов, находящихся под контролем одного фактора транскрипции) и, как правило, определенную функциональную роль [21], однако многие промоторы узнаются несколькими σ-факторами, функции которых частично перекрываются, так как разные процессы (например, окислительный стресс, тепловой шок) могут приводить к одинаковым последствиям для клетки [58,59].

1.1.5 Оперонная организация бактериальных генов

Первые работы по регуляции метаболизма были сделаны при изучении утилизации лактозы бактерией *E. coli*. Для описания лактозного метаболизма Ф. Жакоб и Ж. Моно в 1961 г. ввели термин «оперон» [6,42]. Оперон представляет собой группу из двух или более совместно транскрибируемых генов (иногда говорят и о моноцистронных, т.е. содержащих один ген, оперонах) [4,61]. Белки, кодируемые генами одного оперона, обычно тесно связаны друг с другом функционально и обеспечивают протекание какого-либо метаболического процесса (например, биосинтеза определенной аминокислоты или утилизацию углевода) [4,62]. Организация генов в виде оперонов облегчает координированную регуляцию их экспрессии на уровне транскрипции [4,42,61,62]. Такой контроль экспрессии обычно осуществляется с помощью регуляторных белков, которые действуют, связывая специальную последовательность – оператор, обычно находящийся в непосредственной близости от промотора [1,4,6,7,12].

В регуляции также участвуют, как правило, низкомолекулярные вещества-эффекторы, специфически взаимодействующие с регуляторным белком в качестве индукторов, анти-индукторов или ко-репрессоров; соответственно, в зависимости от действия молекул-эффекторов различают индуцибельные и репрессируемые опероны [5,6,20,43,63]. Эффектор влияет на ДНК-связывающие свойства регуляторного белка, изменяя его конформацию [5,6,20,43,63].

1.1.6 Регуляция экспрессии при помощи альтернативных структур РНК

Помимо специфического связывания ДНК с регуляторными белками, одним из важных способов регуляции экспрессии генов является формирование альтернативных (взаимоисключающих) вторичных структур мРНК – либо на стадии транскрипции за счет образования терминаторов/антитерминаторов, либо на стадии инициации трансляции в результате образования секвесторов (шпилек, перекрывающихся с последовательностью Шайна-Дальгарно или старт-кодоном) или антисеквесторов [25,26,27,29]. Выделяют ряд регуляторных механизмов такого типа, в частности, аттенюаторы транскрипции, РНК-переключатели (riboswitches) и тРНК-связывающие элементы (T-boxes) [25,26,27,29]. Подобные цис-регуляторные структурные элементы, как правило, располагаются в 5’-нетранслируемой области мРНК [25,29,27,64,65], что позволяет им быть синтезированными в первую очередь и взаимодействовать с лигандом-эффектором еще до синтеза полноразмерной мРНК [27].

Аттенюаторы представляют собой регулируемые терминаторы транскрипции, которые используются многими бактериями для изменения уровня экспрессии оперонов биосинтеза

аминокислот [6,25]. Такой способ регуляции основан на сопряженности транскрипции и трансляции у прокариот, при этом альтернативные шпильки формируются под влиянием рибосом [25,27,30]. Последовательность аттенюатора содержит один или несколько кодонов, кодирующих аминокислоту, которая синтезируется продуктами генов соответствующего оперона [25,30]. В условиях недостатка этой аминокислоты рибосома останавливает трансляцию на соответствующих кодонах, закрывая последовательность, необходимую для формирования терминаторной шпильки, и транскрипция не прерывается на аттенюаторе [6,25,27,30]. Если же концентрация данной аминокислоты и, соответственно, аминоацилированной тРНК достаточна, скорость трансляция высока, и рибосомы мешают образованию антитерминаторной шпильки, формируется альтернативная шпилька-терминатор и транскрипция прекращается [6,25]. Первым обнаруженным и одним из наиболее изученных аттенюаторов является аттенюатор триптофанового оперона [6,30].

Типичный РНК-переключатель состоит из двух доменов: сенсора-аптамера, который может напрямую связывать низкомолекулярные метаболиты-эффекторы, и регуляторного домена, способного образовывать альтернативные вторичные структуры и взаимодействовать с клеточной системой транскрипции или трансляции [26,27,28,66,67,68]. Подобные структуры, стабилизированные связыванием эффектора, могут функционировать либо как активаторы, либо как репрессоры, в зависимости от расположения сенсорных и регуляторных элементов [25,26]. РНК-переключатели характерны преимущественно для эубактерий (однако найдены также у архей и эукариот) [27,28,64,67,69] и, вероятно, являются одними из наиболее эволюционно древних регуляторных элементов [25,27,28]. В настоящее время известно множество РНК-переключателей, регулирующих самые разные процессы, лигандами которых служат, например, флавинмононуклеотид (рибофлавин-5-фосфат) [64,66], аденозилкобаламин [65,67], тиаминпирофосфат [70,71], азотистые основания аденин и гуанин, а также их производные [25,28,67,69], аминокислоты глицин [25,28,68], глутамин [68], лизин [25,28,67,68], и многие другие соединения [25,27,28]. Примечательно, что структура и механизм действия первого известного РНК-переключателя, регулирующего синтез рибофлавина, были сначала предсказаны биоинформатическими методами [66], что в дальнейшем получило экспериментальное подтверждение [64]. Аналогичным образом были предсказаны и впоследствии экспериментально подтверждены РНК-переключатели, контролирующие синтез тиамина [70,71] и кобаламина [65,67], метаболизм пуринов [25,27,69] и азотный метаболизм [68].

Еще одним вариантом подобных регуляторных элементов являются Т-боксы, которые регулируют экспрессию генов, кодирующих аминоацил-тРНК синтетазы, транспортеры и

ферменты биосинтеза аминокислот [25,29]. Т-боксы на 5'-конце лидерной мРНК напрямую взаимодействуют с неаминоацилированными тРНК, высокая концентрация которых является признаком недостатка соответствующих аминокислот [25,27,29]. Связывание неаминоацилированных тРНК способствует транскрипции (стабилизируя антитерминаторную шпильку и блокируя формирование терминаторной) или же инициации трансляции (препятствуя формированию секвестора и высвобождая последовательность Шайна-Дальгарно) соответствующих генов [25,29].

Кроме того, в антитерминации могут участвовать специальные РНК-связывающие белки [25]. Так, например, белок-антитерминатор GlpP в присутствии глицерол-3-фосфата связывается с инвертированным повтором лидерной мРНК гена глицерол-3-фосфат дегидрогеназы *glpD*, регулируя его экспрессию [72]. Аналогично, активируемый триптофаном белок TRAP (*trp* RNA-binding attenuation protein) у *Bacillus* spp. взаимодействует с последовательностью нуклеотидов на 5'-конце лидерного транскрипта оперона *trpEDCFBA*, блокируя формирование антитерминаторной шпильки и приводя к аттенюации транскрипции за счет образования альтернативной шпильки-терминатора, а также может ингибировать трансляцию, способствуя формированию секвестора [25,73,74].

1.2 Факторы транскрипции

Факторы транскрипции – это регуляторные белки, связывающиеся с ДНК и контролирующие экспрессию генов, активируя или репрессируя ее, в ответ на внешние или внутриклеточные стимулы [1,5,9]. Для осуществления регуляции транскрипционные факторы узнают и специфически связывают определенную последовательностью ДНК – оператор/сайт связывания [1,4,6,7,12,75]. Разные сайты, с которыми взаимодействует один и тот же белок, сходны, но не идентичны, и различия в последовательности обеспечивают различное сродство транскрипционного фактора к разным сайтам связывания внутри регулона; чем ближе последовательность сайта к консенсусу, тем, как правило, выше связывающая способность [4,75]. Очень часто транскрипционные факторы имеют в составе два домена: ДНК-связывающий домен и второй домен, отвечающий за димеризацию и/или связывание лиганда-эффектора [20,76].

В геноме *E. coli* имеется около 300 генов, кодирующих известные и потенциальные транскрипционные факторы [5,7,20,77]. При этом ключевую роль в интеграции регуляторных сетей и адаптации к текущим условиям среды у *E. coli* играет сравнительно небольшое число глобальных регуляторов: всего семь транскрипционных факторов (ArcA, Crp, FIS, Fnr, IHF, Lrp и NarL) контролируют экспрессию более половины всех генов [4,5,20,78]. Так, NarL активирует

транскрипцию генов, ответственных за нитрат/нитритное дыхание (например, *narGHJI, narK, narP, narQ, narXL, nirBDC*), а также ряда дегидрогеназ (*nuoA-N, fdnGHI*), и репрессирует экспрессию оперонов, продукты которых участвуют в иных вариантах анаэробного дыхания (например, *dmsABC, torCAD, frdABCD*) [79,80,81,82]. Большинство транскрипционных факторов, даже локальных, регулируют экспрессию не одного, а нескольких генов [4]. Кроме того, большая часть бактериальных факторов транскрипции (~55-70%) авторегулируема [7,20]. При этом негативная авторегуляция, т.е. репрессия транскрипционным фактором экспрессии своего собственного гена, представляет собой наиболее распространенный вариант (например, ~40% регуляторов у *E. coli*) [83,84].

Число факторов транскрипции в конкретном геноме варьирует и зависит от множества условий, в частности, от места обитания организма и размера генома [3,4,77,85,86]. К примеру, известно, что паразиты и эндосимбионты с небольшими геномами имеют существенно меньше транскрипционных факторов (около 1%), нежели свободноживущие организмы, имеющие геномы большего размера (в среднем 4-7%, до 9-12%) [3,4,77,85]. Вероятно, это является свидетельством того, что с усложнением организации организма требуется все более сложная регуляция для осуществления программ развития и роста, а также адекватной реакции на внешние стимулы [3,4,75]. Есть данные о том, что в среднем количество транскрипционных факторов возрастает квадратично относительно увеличения размера генома [3,75,86,87].

При этом увеличение числа и разнообразия транскрипционных факторов, принадлежащих к одному структурному семейству, является результатом множества дупликаций с последующей дивергенцией и специализацией [3,77]. Сайты связывания транскрипционных факторов, принадлежащих одному семейству, часто весьма схожи; кроме того, регуляторные белки, имеющие похожие мотивы связывания, обычно имеют сходные функции, участвуя в близких биологических процессах, что снижает эффект от возможного кросс-узнавания и позволяет родственным факторам частично дублировать функции друг друга [7,75].

1.2.1 Основные группы транскрипционных факторов

Транскрипционные факторы весьма разнообразны, в настоящее время известны десятки различных семейств регуляторных белков, встречающихся среди разных таксономических групп [3,4,12,88]. Самые крупные из известных семейств – LysR, а также AraC и TetR [4,77]. Содержание транскрипционных факторов различных семейств варьирует в зависимости от таксономической группы, например, у *B. subtilis* большая часть регуляторов принадлежит семейству MarR, тогда как у *Bordetella* spp. – семейству IclR [4].

Обычно транскрипционные факторы классифицируют на основании строения ДНК-связывающего домена – можно выделить преимущественно α-спиральные домены, домены, состоящие в основном из β-тяжей, а также смешанные по составу домены [12,14,89]. Наиболее распространенным элементом вторичной структуры, участвующим в распознавании ДНК, является α-спираль [14].

Примерами преимущественно α-спиральных доменов могут служить ДНК-связывающие домены типа «лейциновая застежка» (leucine zipper), домены, содержащие мотив «спираль-петля-спираль» (helix-loop-helix, HLH), а также «спираль-поворот-спираль» (helix-turn-helix, HTH) [12,14,89].

Типичные для эукариот домены типа «лейциновая застежка» состоят из α-спирали, участок которой содержит 4-5 периодически расположенных остатков лейцина, располагающихся на одной стороне α-спирали через каждые два витка и обеспечивающих димеризацию содержащих их факторов за счет взаимопроникновения лейциновых α-спиралей по типу молнии [6,12,89]. В результате происходит правильное ориентирование собственно ДНК-связывающего домена, расположенного непосредственно рядом с «застежкой», основные аминокислоты которого связываются с большой бороздкой ДНК [6,12,89].

Похожую структуру, а также способ димеризации и связывания с ДНК имеют и домены, содержащие HLH-мотив: две α-спирали, соединенные друг с другом петлями различной длины [12,14]. HLH-домены распространены среди эукариот и часто присутствуют в составе транскрипционных факторов, контролирующих развитие и дифференцировку организма [6,12].

ДНК-связывающий домен подавляющего большинства факторов транскрипции (например, таких семейств, как AraC, CRP, FIS, Fur, GntR, LacI, XRE, гомеобоксных белков и т.п.), регулирующих самые разные функции, особенно среди прокариот, устроен по типу HTH [1,3,6,14,77,89]. Его основу составляют две α-спирали, расположенные под углом друг к другу так, что одна из спиралей («распознающая») ложится в большую бороздку ДНК [1,12,14, 76,77,89]. Следует также отметить, что многие транскрипционные факторы, например, регуляторы устойчивости к антибиотикам из семейства MarR, имеют в составе вариацию данного мотива с дополнительным антипараллельным β-листом (winged helix-turn-helix, wHTH), который, как правило, образует дополнительные контакты с малой бороздкой ДНК [12,14].

Распространенными ДНК-связывающими доменами с разными комбинациями β-листов и α-спиральных структур в составе являются различные цинк-связывающие мотивы, например, «цинковые пальцы» (zinc fingers) [12,14,89]. Этот тип домена распространен преимущественно среди эукариотических организмов [12], однако встречается и у прокариот (например,

регуляторы Ros и MucR у представителей Alphaproteobacteria и т.д.) [4,90] и включает короткую α-спираль, которая взаимодействует с большой бороздкой ДНК, и β-лист из двух антипараллельных цепей [12,14,89]. Вторичная структура домена стабилизируется координационным связыванием иона цинка с помощью аминокислотных остатков гистидина и/или цистеина (мотивы типа Cys2His2, Cys4 и Cys6) [12,14,89]. В составе белка, как правило, присутствуют несколько периодически расположенных цинковых пальцев [12,14,89].

Кроме того, α-спирали и β-структуры также формируют ДНК-связывающий домен типа «лента-спираль-спираль» (ribbon-helix-helix, RHH) [14]. RHH-мотив входит в состав ряда бактериальных транскрипционных факторов, например, репрессоров MetJ и Arc [12,14,89]. Две α-спирали домена участвуют в процессе димеризации, тогда как узнавание ДНК обеспечивается антипараллельными β-лентами, взаимодействующими с большой бороздкой ДНК, или же, как в случае транскрипционных факторов IHF и HU, с малой бороздкой [12,14,89].

Примерами ДНК-связывающих доменов, преимущественно состоящих из β-структур, могут служить: TATA-бокс связывающие белки (TBP), универсальные компоненты мультибелковых факторов инициации транскрипции у эукариот, взаимодействующие с малой бороздкой ДНК при помощи β-листа из десяти антипараллельных цепей [6,12,14], домены со структурами типа β-сэндвич (иммуноглобулин-подобные домены, например, у p53-подобных транскрипционных факторов), β-трилистник (β-trefoil, например, ядерный эффектор CSL сигнального пути Notch) и β-β-β (десять β-цепей, организованных в три антипараллельных β-листа, например, у домена AgrA *Staphylococcus aureus*) [14].

1.2.2 Механизм работы транскрипционных факторов

Факторы транскрипции осуществляют изменение экспрессии регулируемых генов в соответствии с условями внешней или внутренней среды [1,5]. При этом изменение условий, воспринимаемое факторами транскрипции, приводит к модуляции их активности или экспрессии [1,5].

Одним из основных вариантов подобной модуляции является взаимодействие транскрипционного фактора с низкомолекулярным эффектором-лигандом, меняющим его сродство к ДНК [4,5,6,20]. Классическим примером подобного взаимодействия является репрессор лактозного оперона *lacZYA* у *E. coli* – LacI [5,42,43,63]. В отсутствие или при низкой концентрации индуктора (аллолактозы) в клетке белок-репрессор соединяется с оператором и препятствует транскрипции, блокируя синтез ферментов метаболизма лактозы [42,43,63]. В условиях достаточной концентрации, аллолактоза связывается с репрессором, вызывая изменение его конформации, которое приводит к диссоциации регулятора от

последовательности оператора, в результате чего индуцируется транскрипция генов оперона [42,43].

Вторым возможным механизмом, влияющим на связывание фактора транскрипции с ДНК, является ковалентная модификация регулятора, например, ацетилирование, гликозилирование, метилирование, фосфорилирование и т.п. [4,5,39]. Наиболее распространенной модификацией является фосфорилирование. У прокариот этот механизм, как правило, характерен для двукомпонентных систем, широко используемых для передачи сигнала и ответа на различные внешние и внутренние стимулы [4,91,92]. Двухкомпонентные системы состоят из сенсорной киназы и транскрипционного фактора; фосфорилирование киназой белка-регулятора меняет его сродство к ДНК [91,93]. Примерами могут служить система ответа на фосфатное голодание PhoBR у *E. coli* [92,93], двухкомпонентная система FixLJ у *Bradyrhizobium japonicum* и *Rhizobium meliloti,* регулирующая многочисленные гены, участвующие в процессе фиксации азота и анаэробном дыхании [91,94], а также системы NarXL and NarQP, ответственные за нитрат/нитритное дыхание у *E. coli* [5,79,82].

Третьим вариантом является регуляция гена фактора транскрипции другим регулятором. Подобные регуляторные каскады – типичный и широко распространенный механизм контроля экспрессии генов глобальными регуляторами, такими как FNR, ArcA, NarL, CRP и т.д [20, 79,82]. К примеру, CRP контролирует экспрессию генов регуляторов метаболизма L-идоната IdnR [95,96], пропионата PrpR [97], гексуронатов UxuR [96], D-галактозы GalS [98], а также ряд других транскрипционных факторов вместе с соответствующими катаболическими оперонами [20].

Среди факторов транскрипции встречаются белки-активаторы, репрессоры, а также регуляторы, способные действовать и как репрессоры, и как активаторы (в зависимости от места и способа связывания с ДНК, а также механизма взаимодействия с РНК-полимеразой) [4,5,6,77]. Примером последнего может служить транскрипционный фактор CRP [6,20]. CRP в комплексе с эффектором цАМФ активирует транскрипцию генов, участвующих в катаболизме различных органических соединений (преимущественно сахаров, используемых в качестве источника углерода) [6,20]. Концентрация цАМФ в клетке возрастает в случае роста бактерий на бедных питательных средах и снижается в условиях избытка глюкозы [6,20]. Поэтому CRP обеспечивает включение экспрессии специальных катаболических оперонов лишь в случае нехватки легко усваиваемых источников углерода и энергии [6,20]. В то же время, CRP может выступать и в роли репрессора множества разных генов, например, лактозного оперона *lacZYA* и галактозного оперона *galETK* у *E. coli* [6,99,100].

1.2.3 Репрессоры транскрипции

Репрессорное действие факторов транскрипции на регулируемые гены реализуется различными способами. Репрессор может мешать связыванию и продвижению РНК-полимеразы, формированию открытого комплекса или же первых фосфодиэфирных связей [101,102]. Наиболее простым и распространенным вариантом является связывание транскрипционного фактора с ДНК в области промотора, что создает стерическое затруднение для связывания РНК-полимеразы и мешает инициации транскрипции или же препятствует дальнейшему продвижению РНК-полимеразы в тех случаях, когда оператор расположен после промоторной последовательности [4,5,6,103]. Таким способом регулирует транскрипцию репрессор LacI [5,63,103].

Второй возможный механизм – взаимодействие репрессора с белком-активатором данного гена (анти-активация) [5]. Например, CytR, регулятор транспорта и утилизации нуклеозидов, взаимодействуя с CRP и связывая ДНК в области между двумя CRP-сайтами, блокирует активацию CRP соответствующих генов [5,101,104]. Примером также может являться ингибирование репрессором Fur активирующего действия CRP на гены пектинолиза *pelD* и *pelE* у *Erwinia chrysanthemi 3937* [105].

Кроме того, факторы транскрипции могут олигомеризоваться, связываясь с соседними операторами и осуществляя кооперативную регуляцию [103,106,107]. В этом случае репрессия обусловлена формированием петель ДНК в области промотора, мешающих связыванию РНК-полимеразы с промотором и инициации транскрипции, в результате олигомеризации молекул репрессора, связанных с несколькими сайтами [4,5,106,107]. Подобный вариант кооперативной репрессии показан, например, для регуляторов утилизации L-арабинозы AraR *B. subtilis* [106], метаболизма лактозы LacI [103,107] и D-галактозы GalR у *E. coli* [5,98,102,108], репрессора cI фага λ и т.д. [103].

1.2.4 Активаторы транскрипции

Активатор может быть необходим для усиления начального связывания РНК-полимеразы с промотором, для изомеризации из закрытого в открытый комплекс, либо для освобождения промотора (начала собственно транскрипции) [5,103,109,110]. Соответственно, активатор может либо вступать в непосредственный контакт с РНК-полимеразой, либо изменять конформацию ДНК в области промотора, вызывая расплетание и облегчая инициацию транскрипции [6,22,110].

Существует несколько основных типов белков-активаторов. Активаторы класса I взаимодействуют с С-концевым доменом α-субъединицы РНК-полимеразы, а их сайты

связывания располагаются на различных расстояниях перед –35 последовательностью промотора [5,109,110]. Это объясняется строением α-субъединицы РНК-полимеразы, которая состоит из двух доменов, соединенных между собой гибким линкером [5,47,110]. В отличие от N-концевого домена, положение которого фиксировано, так как он связан с другими субъединицами РНК-полимеразы, С-концевой домен α-субъединицы достаточно подвижен и может контактировать с активаторами на разном расстоянии [5,47]. Подобным образом осуществляется, например, активация экспрессии лактозного оперона *lacZYA* регулятором CRP в отсутствие глюкозы [5,103,110] или же активация экспрессии гена НАДН-дегидрогеназы *ndh* регулятором Fis у *E. coli* [111].

Активаторы класса II взаимодействуют одновременно как с α-субъединицей РНК-полимеразы, так и с σ-фактором, сайты связывания при этом перекрываются с –35 последовательностью промотора или находятся в непосредственной близости от него [5,92,109,110]. В отличие от предыдущего класса активаторов, позиция сайта связывания в данном случае определяется достаточно жестко, поскольку σ-фактор по отношению к промотору фиксирован [5]. Примером активации такого типа является активация экспрессии галактозного оперона *galETK* регулятором CRP [99,110] или же генов ответа на фосфатное голодание регулятором PhoB у *E. coli* [92,112].

Еще один механизм активации транскрипции характерен для белков семейства MerR [5,113], таких как, например, регулятор ответа на окислительный стресс SoxR [113,114], регуляторы транспортеров множественной лекарственной устойчивости BltR, BmrR, MtaN [113,115,116], а также белки CueR, MerR, ZntR, регулирующие гены устойчивости бактериальных клеток к ионам меди, ртути и цинка, соответственно [113,115,117]. Регулятор связывается с промоторной областью между –35 и –10 элементами (расстояние между которыми при этом, как правило, больше оптимального, равного 17 нуклеотидам) и изменяет конформацию ДНК, правильно ориентируя элементы промотора и способствуя его связыванию с σ-фактором РНК-полимеразы [5,113,114,115,116,117].

Как и для ряда репрессоров транскрипции, для активаторов также характерны олигомеризация, связывание с соседними операторами и кооперативная регуляция [5,103]. Кооперативная активация показана, например, для репрессора cI фага λ [103]. Кроме того, взаимодействующие транскрипционные факторы разной природы также могут осуществлять коактивацию, как это было показано, например, для регуляторов CRP и MelR, совместно активирующих экспрессию оперона *melAB* у *E. coli* [5,103].

1.3 Сравнительно-геномные методы исследования. Изучение регуляции транскрипции

До недавнего времени исследование транскрипции и ее регуляции осуществлялось только экспериментально. На данный момент существует масса хорошо разработанных и широко применяемых экспериментальных методов исследования транскрипции, например, задержка ДНК в геле (gel shift assay, EMSA) [118,119,120], ДНК-футпринтинг (DNA footprinting, выявление защищенных от расщепления последовательностей, связывающих регуляторные белки) [120,121], направленный мутагенез и исследование мутантных организмов [6,122,123,124] и т.п. Однако эти методы довольно трудоемки, а также не отражают полностью все многообразие регуляторных взаимодействий, так как используются для исследования транскрипции ограниченного числа генов [4,7,8]. В то же время, современные экспериментальные методы массового анализа экспрессии, например, метод иммунопреципитации хроматина (ChIP), комбинированный с массовым параллельным секвенированием (ChIP-Seq) или гибридизацией на микрочипах (ChiP-ChiP) [118,125,126,127,128], поиск функционально важных участков нуклеиновых кислот (аптамеров), взаимодействующих со специфическими лигандами (SELEX) [129] и т.д., высокопроизводительны (позволяют получать данные об экспрессии большого числа генов, от нескольких десятков до нескольких десятков тысяч), однако часто дорогостоящи и имеют высокий уровень шума, а также характеризуются сложностью разделения прямых и косвенных регуляторных эффектов (например, регуляторных каскадов от корегуляции) [4,7].

Подобные недостатки и сложности применения экспериментальных методов, вместе с лавинообразно нарастающими темпами накопления геномных данных (см. далее), сделали необходимым использование иных способов исследования регуляции экспрессии генов. Разработанные за последние годы методы биоинформатики представляют собой эффективный и недорогой подход к изучению регуляции, особенно у прокариот, и в настоящее время они широко используются в дополнение или же вместо экспериментальных исследований [4,7,10,11].

Основой изучения регуляции методами биоинформатики является идентификация различных регуляторных последовательностей путем сравнительного анализа: поиск потенциальных регуляторов и предсказание их сайтов связывания, поиск промоторов, аттенюаторов и терминаторов транскрипции и т.д. [4,9,10,11]. Кроме того, сравнительно-геномный анализ регуляции может также включать: уточнение границ генов и определение их оперонной структуры, предсказание функций генов и метаболическую реконструкцию [4,11,130].

1.3.1 Предсказание функций генов с помощью сравнения последовательностей

С 1990-х годов началось секвенирование геномов клеточных организмов. Впервые полная последовательность генома бактерии, облигатного паразита *Haemophilus influenzae* Rd KW20, была опубликована в 1995 году [131]. Первым секвенированным эукариотическим геномом в 1996 году стал геном дрожжей *Saccharomyces cerevisiae* [3,132].

В настоящее время, благодаря удешевлению и повышению производительности технологий секвенирования, биология превращается в науку, «богатую данными»: в базе данных KEGG [133] по состоянию на начало 2016 года насчитывается 3580 полных последовательностей геномов бактерий, 218 – архей, и 313 – эукариот, и накопление данных идет возрастающими темпами [103,134]. Следует отметить, что количество секвенированных эукариотических геномов возрастает заметно медленнее по сравнению с прокариотическими, что обусловлено как размерами геномов, так и сложностью сборки и аннотации в связи с большим числом повторов [3]. Ключевые проблемы, которые возникают в связи с накоплением огромного объема геномных данных, – проблемы хранения и обработки информации [103,134].

Следующим этапом исследований после определения полной последовательности генома является функциональная аннотация, т.е., определение границ генов, соотнесение последовательностей генов с функциями кодируемых ими белков, выявление регуляторных сайтов и т.п. Экспоненциальный рост числа секвенированных геномов делает практически невозможной экспериментальную аннотацию, поэтому аннотация большинства последовательностей в настоящее время осуществляется при помощи методов биоинформатики.

В процессе биоинформатической аннотации геномов определяются потенциальные белок-кодирующие участки (открытые рамки считывания, ORF) – последовательности между предполагаемыми старт- и стоп-кодонами, расположенными на расстоянии, кратном трем нуклеотидам и превышающем определенный порог (обычно ~300 нуклеотидов), не перекрывающиеся или минимально перекрывающиеся с соседними генами [135,136]. Существует целый ряд программ поиска потенциальных генов, основанных на статистическом анализе (GeneMark [137], GLIMMER [138]), а также анализе выравнивания родственных последовательностей (CRITICA [135] и ORPHEUS [136]).

Далее на основе сходства полученных последовательностей с какими-либо ранее экспериментально изученными генами/белками можно сделать предположения о вероятной функции этих ORF. Есть данные, что поиск ортологов (ортологичными называют гены в разных геномах, имеющие общего предшественника и, как правило, выполняющие одинаковые функции) позволяет определить точную функцию примерно 40-60% генов в новом геноме

[139,140]. Приблизительное представление о функции гена может также дать анализ паралогов, т.е., генов, образованных в результате дупликации, однако их точная аннотация затруднена, так как функция любого из паралогов может изменяться в ходе эволюции [139,140].

Наличие в составе определенных, хорошо идентифицируемых структурно-функциональных мотивов может дать дополнительную информацию о функции гена и кодируемого им белка. К примеру, для регуляторных белков характерно наличие ДНК-связывающих доменов [141], а для транспортеров типично наличие серии гидрофобных трансмембранных спиралей [142,143].

Существуют различные банки данных по нуклеотидным и аминокислотным последовательностям (например, GenBank, EMBL, UniProt [144,145,146]), а также структурным мотивам (например, Pfam, PROSITE, SMART [147,148,149]), которые можно использовать для функциональной аннотации. Для поиска гомологичных последовательностей и предсказания приблизительной функции генов и белков широко используются программы пакета BLAST [150]. Кроме того, для множественного выравнивания последовательностей и выделения функциональных фрагментов применяются такие программы, как CLUSTAL [151] и MUSCLE [152], а также TMHMM для выявления трансмембранных участков белков [142], SignalP для идентификации сигнальных пептидов [153], Mfold для предсказания вторичной структуры нуклеиновых кислот [154] и т.п.

Однако данных о сходстве последовательностей и анализа простых структурных мотивов недостаточно для аннотации всех новых найденных генов с неизвестной функцией. Для более точной и подробной аннотации необходимо использовать дополнительную информацию, например, о расположении генов на хромосоме, данные о регуляции и т.д. [155].

1.3.2 Кластеризация и слияние генов, профили встречаемости

Известно, что у прокариот функционально связанные гены (например, отвечающие за разные этапы переработки какого-либо вещества) часто бывают колокализованы и организованы в опероны [61,62,156]. Такая кластеризация может обеспечивать синхронность регуляции всего метаболического пути [62]. Тот факт, что два гена находятся рядом в каком-то одном геноме, не обязательно свидетельствует об их функциональной связи, однако если пара ортологичных генов колокализована в ряде геномов, причем у представителей разных таксономических групп, представляется вероятным, что они функционально связаны. Консервативная кластеризация генов на хромосоме часто позволяет делать выводы о функциях генов, не имеющих известных гомологов, и, например, реконструировать недостающие фрагменты метаболических путей [157].

Важным предположением для биоинформатического анализа является также то, что гены, принадлежащие к одному метаболическому пути, как правило, совместно наследуются в ходе эволюции и, таким образом, присутствуют в геноме по принципу «все или ничего» [61,62,130,157]. Кроме того, колокализация функционально связанных генов часто обусловлена и горизонтальным переносом генов, так как, если в новый геном переместится лишь один ген метаболического пути, он будет бесполезен [61,62]. Таким образом, наличие определенного набора генов в одних геномах и отсутствие в других (профили встречаемости) может свидетельствовать о том, что они функционально связаны, и можно довольно достоверно приписывать белки с ранее неизвестной функцией к определенной метаболической системе на основе общей встречаемости в разных геномах [130,157]. Профили встречаемости можно использовать также для анализа случаев, когда в одних геномах присутствует одна группа генов, а в других – другая; если при этом совместно эти гены никогда не встречаются, это может свидетельствовать об их функциональной взаимозаменяемости.

Исследование геномного контекста может осуществляться с помощью таких веб-серверов, как MicrobesOnline [158], SEED [159] и STRING [160].

Примером успешного использования анализа кластеризации генов на хромосоме и профилей встречаемости для предсказания функций генов может служить идентификация гена транс-2-цис-3-деценоил-ACP изомеразы *fabM* (фермента биосинтеза жирных кислот) у *Streptococcus pneumoniae* [161], а также гена транспортера тиамина *yuaJ* у *Bacillus* spp. и *Clostridium* spp. [71]. С помощью этих методов, а также с помощью анализа регуляции (см. далее), были предсказаны функции генов транспортера биотина *bioY* [162], транспортных систем кобальта и никеля *cbiMNQO* и *nikMNQO* [163], рибулокиназы *araK* у *Clostridium* spp. [164], а также N-ацетилгалактозамин-6-фосфат деацетилазы $agaA^{II}$, N-ацетилгалактозамин киназы *agaK*, галактозамин-6-фосфат изомеразы/деаминазы *agaS* у *Shewanella* spp. [165], L-лактальдегид редуктазы *rhaZ* у Gammaproteobacteria [166], ксилозоизомеразы $xylA^{II}$ у *Clostridia* spp. [167]. В дальнейшем предсказанные для этих генов функции были подтверждены экспериментально [163,164,165,166,167,168].

Известно также, что функционально связанные гены могут сливаться, формируя единый ген и приводя к образованию мультидоменных белков [157,169,170]. Подобное слияние может быть выгодным, например, обеспечивая близость ферментов и интермедиатов реакций одного метаболического пути, что облегчает и ускоряет протекание реакций, а также может упрощать регуляцию [169,170,171]. Таким образом, можно предсказывать функции генов, основываясь на слиянии их гомологов в других организмах с известными генами [169,170,171].

Например, было показано, что белки *S. cerevisiae* MXR1 (метионин сульфоксид редуктаза, компонент системы антиоксидантной защиты) и YCL033C (селенопротеин с неизвестной функцией), гомологичны, соответственно, N- и C-концевым доменам белков *Helicobacter pylori*, *Haemophilus influenzae* и *Treponema pallidum*, на основании чего для YCL033C было предсказано участие в защите клетки от окислительного стресса и нейтрализации активных форм кислорода [170]. Примером использования данных о слиянии генов (вместе с профилями встречаемости) может служить также предсказание функции гена *rhaEW* (*yuxG*), кодирующего два фермента пути утилизации рамнозы – L-рамнулозо-1-фосфат альдолазу (RhaE) и L-лактальдегид дегидрогеназу (RhaW), что далее было подтверждено экспериментально [166].

1.3.3 Поиск потенциальных сайтов связывания. Исследование регуляции транскрипции методами сравнительной геномики

Факторы транскрипции регулируют экспрессию генов, специфически связывая определенные последовательности ДНК [1,5,6,7,12]. Подобная специфичность транскрипционных факторов позволяет исследовать регуляцию экспрессии генов не только экспериментальными, но и теоретическими методами, основанными на анализе нуклеотидных и аминокислотных последовательностей.

Основным методом выявления потенциальных сайтов связывания в биоинформатических исследованиях является филогенетический футпринтинг (phylogenetic footprinting) [4]. Функционально важные участки ДНК, как правило, более консервативны. Таким образом, в основе метода филогенетического футпринтинга лежит поиск консервативных участков выравниваний регуляторных областей ортологичных генов, для которых предполагается регуляция [4,9]. Однако в ряде случаев метод филогенетического футпринтинга не применим: например, при анализе близкородственных видов выравнивание регуляторных областей может быть неинформативным в связи с высокой консервативностью как функциональных, так и нефункциональных участков; с другой стороны, при анализе далеких видов и в случае слабого сходства регуляторных последовательностей возникают трудности с построением корректного выравнивания [4,9]. Кроме того, метод филогенетического футпринтинга не подходит для исследования совместно регулируемых, но не ортологичных генов. В таких случаях могут быть использованы методы, основанные на поиске сходных мотивов в регуляторных областях исследуемых генов (программы AlignACE [172], MEME [173], SignalX [174], SOMBRERO [175] и т.п.).

На основе обучающей выборки из предсказанных/ранее известных сайтов строится распознающее правило (профиль), с помощью которого осуществляется поиск потенциальных

сайтов связывания в исследуемых геномах. Чаще всего для построения распознающих правил используется метод позиционных матриц весов (positional weight matrices, PWM), где для каждого нуклеотида учитывается его частота и консервативность позиции [10,11].

Одной из главных проблем поиска сайтов связывания является выбор порога: повышение его значения приводит к потере ряда верных, экспериментально подтвержденных сайтов, а при понижении появляется большое число ложно предсказанных сайтов. Метод проверки соответствия (consistency check), принцип которого состоит в поиске потенциальных сайтов перед ортологичными генами, позволяет отсеять неверные предсказания [4,11]. Если потенциальный регуляторный сайт обнаруживается перед геном только в каком-либо одном геноме, он, вероятнее всего, предсказан ошибочно. В том случае, если потенциальные сайты перед ортологичными генами присутствуют в нескольких геномах, они, скорее всего, предсказаны верно, и эти гены действительно являются регулируемыми. При этом выбор геномов зависит от ожидаемой консервативности регуляторного сигнала: в случае эволюционно близких геномов регуляторные области часто практически идентичны, что не позволяет делать выводы об истинности сайта, тогда как у эволюционно далеких организмов построение общего распознающего правила может быть невозможно в связи с большими отличиями между регуляторными последовательностями [4,9].

Впервые метод проверки соответствия был использован для анализа состава и структуры регулонов биосинтеза пуриновых нуклеотидов, аргинина и ароматических аминокислот у *E. coli* и *H. influenza* [10], и, показав свою эффективность, был применен в целом ряде работ по исследованию регуляции: например, при изучении регулонов утилизации различных углеводов (арабинозы, галактозы, ксилозы, маннозы, рамнозы, рибозы, фукозы и других моносахаридов, а также ряда дисахаридов и гликанов) [166,167,176,177,178,179,180], биосинтеза ароматических аминокислот [181,182], метаболизма метионина [183], биосинтеза НАД [184], ответа на тепловой шок [185], устойчивости к ионам тяжелых металлов у разных групп бактерий [186] и т.д.

В последнее время часто проводится изучение регуляции одной и той же группы генов, например, определенного метаболического пути, несколькими различными транскрипционными факторами в разных геномах и таксонах. Примерами подобных исследований могут служить исследования метаболизма жирных кислот и аминокислот с разветвленной боковой цепью, контролируемого регуляторами FadP, FadR, LiuQ, LiuR и PsrA [187], а также регуляции утилизации хитина и N-ацетилглюкозамина транскрипционными факторами NagC, NagQ и NagR [4,188].

Поиск потенциальных сайтов связывания и исследование регуляции методами сравнительной геномики позволяют делать весьма нетривиальные предсказания, которые, по всей видимости, было бы затруднительно получить исключительно экспериментальными методами. Примером подобного предсказания, в дальнейшем получившего экспериментальное подтверждение, является выявление цинк-зависимой регуляции генов рибосомных белков у бактерий [189,190,191,192]. В ходе исследования регуляции утилизации цинка было показано, что существует два паралогичных варианта белков рибосом – с мотивом связывания цинка в составе и без него, при этом в последнем случае перед генами рибосомных белков присутствуют сайты связывания цинкового репрессора Zur [189,190]. Рибосомы депонируют цинк в условиях его избытка, при этом выключен синтез белков рибосом, не содержащих мотив связывания цинка. При недостатке цинка они синтезируются, частично замещают в рибосомах паралоги, связывающие цинк, таким образом высвобождая его для жизненно важных цинк-зависимых ферментов [189].

1.4 ДНК-белковые взаимодействия

ДНК-белковые взаимодействия являются ключевыми для многих важнейших биологических процессов, включая репарацию и рекомбинацию ДНК, репликацию и транскрипцию [8,12,89]. Как было упомянуто выше, одним из основных механизмов регуляции экспрессии генов является специфическое связывание транскрипционных факторов с ДНК, и в среднем около 4-7% генов в геномах бактерий кодируют факторы транскрипции [3,4,13,77,85], однако их структура, специфичность и особенности связывания ДНК часто неизвестны [8,13]. Понимание механизмов ДНК-белковых взаимодействий – одна из важнейших задач молекулярной биологии и сравнительной геномики, необходимая для предсказания специфичности транскрипционных факторов [7,13,89].

Существуют определенные эмпирические принципы ДНК-белковых взаимодействий, основанные на физических и химических свойствах, таких как, например, возможность электростатического взаимодействия между радикалами аминокислот и азотистыми основаниями, гибкость боковой цепи аминокислоты и т.д. [193]. Считается, что вклад основной аминокислотной цепи в специфические взаимодействия с ДНК существенно меньше в сравнении с ролью боковой цепи (радикала) [194]. Наиболее важными и предпочтительными типами взаимодействия обычно являются водородные связи (в связи с их высокой специфичностью и направленным характером), а также кислотно-основные взаимодействия [193,195,196], однако прочие типы взаимодействий также вносят свой вклад [13]. В бороздках ДНК находится сравнительно небольшое количество неполярных атомов [193], а области ДНК-

белковых контактов богаты полярными аминокислотными остатками (заряженными и незаряженными), которые наиболее важны для связывания с ДНК, так как участвуют в формировании электростатических и водородных связей [197]; в связи с этим, гидрофобные взаимодействия не считаются ключевыми для связывания. Однако они тоже могут играть определенную роль в ДНК-белковых взаимодействиях: например, считается, что водородные связи более специфичны по отношению к пуриновым основаниям, тогда как гидрофобные контакты участвуют в распознавании пиримидинов (различении тимина и цитозина) [14].

Было показано, что наиболее часто взаимодействующими с основаниями ДНК аминокислотами являются Arg, Asn, Asp, Gln, Gly, Lys, Ser и Thr – их контакты составляют более 70% от общего числа взаимодействий, при этом контакты с одним только Arg составляют 23% [193,196]. В то же время, Cys, Ile, Leu, Met, Phe, Pro и Trp редко вступают во взаимодействие с ДНК (сумма всех их контактов составляет около 10% от общего числа) [196]. Arg-G, Asn-A, Asp-C, Gln-A, Glu-C, Lys-G и в меньшей степени His-G и Ser-G, по-видимому, представляют наиболее важные, сильные и специфичные контакты [193,196]. Предпочтительное взаимодействие также показано для пар Ala-C, Cys-G, Gly-G, Leu-A, Thr-G и Trp-C. Существуют также достоверно энергетически невыгодные взаимодействия: Asn-T, Asp-G, Gln-T, Glu-G, Ile-T, Leu-T, Met-T и Val-T [196].

В целом аминокислоты, являющиеся донорами при образовании водородных связей (Arg, Cys, His, Lys, Ser, Thr), предпочитают взаимодействие с G, аминокислоты с акцепторными свойствами (Asp и Glu) часто взаимодействуют с C, тогда как Asn и Gln, для которых возможна как донорная, так и акцепторная функция, предпочтительно образуют контакты с A [195,196]. Специфическое связывание аминокислот с G вносит основной энергетический вклад во взаимодействие ДНК с белками, что может быть обусловлено большим числом атомов этого азотистого основания, потенциально способных формировать многочисленные водородные связи [196]. Важность и частота встречаемости контактов с Arg, в особенности Arg-G, может быть объяснена тем, что длинный гибкий радикал этой аминокислоты дает возможность образовывать водородные связи в различных конформациях [193,196].

Однако эти тенденции не объясняют всего многообразия ДНК-белковых взаимодействий; универсального кода подобных контактов не существует, а контакты аминокислот и азотистых оснований могут сильно зависеть от окружения, и, таким образом, от особенностей структуры ДНК-связывающего белка [14,89,198].

Консервативность азотистого основания в определенной позиции сайта хорошо коррелирует с числом образуемых им контактов [13,194,199]. Нуклеотиды, которые формируют больше контактов с белком, более консервативны, так как такие взаимодействия стабилизируют

ДНК-белковый комплекс, следовательно, изменения в этих позициях будут оказывать существенно более негативный эффект, нежели в каких-либо других, и, таким образом, будут устраняться в ходе эволюции [13,194,199]. Анализ взаимной информации может служить хорошим средством предсказания контактов аминокислот с основаниями ДНК для различных семейств транскрипционных факторов, что позволяет получить определенную структурную информацию, используя одни только данные о последовательностях [200].

Принято считать, что контакты белка с сахаро-фосфатным остовом ДНК не играют важной роли в определении специфичности [14], однако они могут оказывать существенное влияние на связывание, стабилизируя и правильно ориентируя взаимодействующие элементы комплекса [14,199].

1.4.1 Семейство транскрипционных факторов GntR

GntR – широко распространенное среди различных таксономических групп бактерий семейство транскрипционных факторов, регулирующих самые разные биологические процессы [76,201,202]. Это семейство было впервые описано в 1991 году и названо по репрессору глюконатного оперона у *Bacillus subtilis* [76,201].

Как и многие другие транскрипционные факторы, регуляторы семейства GntR содержат два домена – ДНК-связывающий и эффекторный [76,202,203]. Для GntR-регуляторов характерен высоко консервативный внутри всего семейства N-концевой ДНК-связывающий HTH-домен, однако они различаются структурой С-концевых доменов, осуществляющих олигомеризацию и связывание эффекторов [76,201]. N-концевой ДНК-связывающий домен белков семейства GNTR включает центральный кластер β-листов и три α-спирали [76]. HTH-мотив (один из самых распространенных и хорошо изученных ДНК-связывающих мотивов у прокариот) состоит из α-спирали, петли и еще одной α-спирали, часто называемой «распознающей», так как она является тем структурным элементом, который взаимодействует с большой бороздкой ДНК при связывании [12,14,76,201,204]. Как правило, белки с HTH-мотивом связываются с симметричными палиндромными последовательностями/инвертированными повторами ДНК в виде димеров, где каждый мономер узнает половину сайта связывания [12,76,89,204].

Несмотря на то, что С-концевой домен не связывается с ДНК напрямую, он может играть важную роль в процессе регуляции, воздействуя на ДНК-связывающий домен [76,203]. Например, С-концевой домен может накладывать определенные стерические ограничения и снижать подвижность ДНК-связывающего домена относительно остальной части белка, мешая приспосабливаться к различным расстояниям между полусайтами палиндромного мотива

связывания [76,203]. Олигомеризация и конформационные изменения в связи со связыванием молекулы-эффектора позволяют корректно расположить HTH-мотив относительно ДНК и, таким образом, сделать возможным последующее связывание [76,203]. Влияние С-концевого домена на ДНК-связывающие свойства N-концевого домена было показано для многих белков [205,206,207]. Таким образом, несмотря на высокую консервативность собственно ДНК-связывающего домена, среди регуляторов семейства GntR наблюдаются различные консенсусные последовательности соответствующих мотивов связывания, что во многом обусловлено гетерогенностью С-концевого домена и его синергией с N-концевым доменом [76].

В соответствии со структурой С-концевого домена в семействе выделяют четыре основных (FadR, HutC, MocR и YtrA) и два минорных (AraR и PlmA) подсемейства [76,201, 203,208,209,210,211].

Подсемейство FadR наиболее многочисленное и включает около 40% всех регуляторов семейства GntR [76,201]. Характерной чертой транскрипционных факторов этого подсемейства является полностью α-спиральный С-концевой домен длиной 150-170 аминокислот, состоящий из шести или семи α-спиралей [76,201]. Эффекторами транскрипционных факторов подсемейства FadR являются небольшие органические лиганды, например, карбоновые кислоты [201]. В связывании молекул эффектора у регуляторов данной группы, предположительно, участвуют ионы металла (вероятнее всего Zn^{2+}) [201]. Большая часть белков подсемейства FadR контролирует метаболизм окисленных субстратов, родственных аминокислотам или связанных с центральным метаболизмом, и регулирует пересечения различных метаболических путей [76,201]. Например, типичными представителями транскрипционных факторов этой группы являются регуляторы метаболизма галактоната (DgoR), гликолата (GlcC), глюконата (GntR), лактата (LldR) и пирувата (PdhR) [76,201].

Второе по размеру подсемейство HutC включает в себя примерно 30% регуляторов семейства GntR [76]. Средняя длина С-концевого домена – около 170 аминокислот, он содержит как α-спирали, так и β-листы [76]. У регуляторов подсемейства HutC структура лиганд-связывающего домена сходна с таковой для хоризмат лиаз (UbiC *E. coli*), что позволяет предположить аналогичный механизм связывания небольших молекул-эффекторов, таких как гистидин (HutC), жирные кислоты (FarR), сахара (TreR) и алкилфосфонаты (PhnF) [202]. Некоторые регуляторы этой группы контролируют перенос конъюгативных плазмид у *Streptomyces* spp. (например, KorSA, KorA и TraR), а также участвуют в регуляции метаболизма N-ацетилглюкозамина (DasR, NagR, NagQ) и ряда других соединений [76,188,203,204].

Третье подсемейство, MocR, сильно отличается от прочих подсемейств структурой необычно длинного (около 350 аминокислот) С-концевого домена, для которого характерна

гомология с аминотрансферазами класса I (TyrB *E. coli*) [76,212,213,214]. Этот класс аминотрансфераз катализирует перенос аминогруппы от аминокислотного субстрата к акцептору – α-кетокислоте и использует пиридоксаль-5-фосфат (PLP) в качестве кофактора [76,212,213]. Аналогичная роль PLP была показана для ряда белков подсемейства MocR (GabR, TauR) [208,212,213,214], кроме того, известно, что транскрипционные факторы PdxR у ряда Actinobacteria напрямую участвуют в регуляции синтеза PLP [76,215,216]. Показано, что аминотрансферазы формируют димеры, связанные по типу «голова к хвосту», таким образом, аналогичная димеризация может происходить и в случае транскрипционных факторов подсемейства MocR [76,212,214].

Регуляторы из четвертого подсемейства, YtrA, которое является наименее многочисленным из основных подсемейств (около 6% от семейства GntR), обладают редуцированным С-концевым доменом, средней длиной около 50 аминокислот, содержащим только две α-спирали [76]. Это накладывает серьезные ограничения на возможность связывания эффектора и димеризации. Однако димеризация остается возможной, о чем свидетельствует наличие протяженных палиндромных сайтов связывания, предсказанных в 5'-области генов, регулируемых факторами транскрипции подсемейства YtrA [76]. Большая часть генов, кодирующих регуляторы данной группы, формирует опероны с генами АТФ-зависимых кассетных (ABC) транспортеров [76].

Небольшое подсемейство PlmA содержит исключительно цианобактериальные транскрипционные факторы [209]. Наибольшее сходство регуляторов этой группы наблюдается с белками подсемейств YtrA и MocR, одно из которых, вероятно, в результате дивергенции и изменения С-концевого домена дало начало PlmA [209]. PlmA (ген *all1076*) контролирует содержание плазмид у *Anabaena* (*Nostoc*) sp. PCC 7120, но остается неясным, является ли это общей функцией регуляторов данного подсемейства, так как у ряда цианобактерий, имеющих ортологи *plmA*, плазмид не обнаружено [209].

Транскрипционные факторы подсемейства AraR – химерные белки, состоящие из двух доменов разного происхождения: N-концевой ДНК-связывающий домен содержит аналогичный таковому у прочих представителей семейства GntR HTH-мотив, тогда как С-концевой домен гомологичен лиганд-связывающему домену семейства GalR/LacI [178,210,211]. AraR контролирует экспрессию генов, кодирующих транспортеры и ферменты метаболизма L-арабинозы и полисахаридов, содержащих арабинозу, а также ксилозы и галактозы у представителей Firmicutes [178,210,211].

1.4.2 Структура сайтов связывания регуляторов семейства GntR

ДНК-связывающие домены разного структурного типа распознают различные мотивы [217], тогда как ДНК-связывающие белки, принадлежащие к одному семейству, часто имеют сайты связывания одинаковой симметрии, длины, аналогичной структуры и специфичности [194]. Строение ДНК-связывающего домена и его способ взаимодействия с ДНК обычно довольно консервативны внутри всего семейства, что обуславливает характерные паттерны аминокислотных контактов с азотистыми основаниями ДНК [194]. Однако даже белки с высоким уровнем сходства аминокислотной последовательности (до 60-70%) могут иметь достаточно четко различающиеся мотивы связывания [217].

Несмотря на консервативность HTH-мотива внутри всего семейства GntR, консенсусная последовательность ДНК-связывающего домена немного различается для каждого из подсемейств [76]. Наибольший уровень сходства наблюдается между ДНК-связывающими доменами подсемейств MocR и YtrA. Одно из этих двух подсемейств могло эволюционировать на основе другого путем замещения С-концевого домена [76].

Большая часть экспериментально известных и предсказанных мотивов связывания различных регуляторов семейства GntR соответствует палиндромной последовательности с консенсусом $N_yGTN_xACN_y$ [76]. Эти мотивы различаются по числу (x,y) и природе (N) нуклеотидов, которые окружают консенсусные пары GT и AC [76]. Обычно это окружение включает преимущественно нуклеотиды A и T, и их число различается для регуляторов разных подсемейств [76]. Например, консенсусным мотивом для подсемейства FadR является последовательность $N_yGTM\text{-}N_{0–1}\text{-}KACN_y$, а для HutC – $N_yGTMTAKACN_y$ [76,203]. Центр палиндрома обычно консервативен, тогда как периферические части мотива вариабельны [76].

Расстояние между полусайтами имеет ключевое значение для правильного предъявления сайта связывания на поверхности ДНК [76]. Мотивы связывания регуляторов подсемейства YtrA сильно отличаются по этому параметру от мотивов подсемейств FadR и HutC: в этой группе консервативные пары GT и AC расположены на значительном расстоянии от центра палиндрома [76]. Подобное строение сайтов связывания регуляторов подсемейства YtrA может быть обусловлено недостаточной длиной их С-концевых доменов, что может приводить к специфическому способу димеризации и связывания с ДНК этих транскрипционных факторов, и, следовательно, нетипичной структуре мотива [76].

Сравнительный анализ факторов транскрипции подсемейства MocR не выявил никаких палиндромных последовательностей, удовлетворяющих консенсусной последовательности семейства GntR, или же общих для подсемейства в целом [76]. Предсказанные для некоторых регуляторов подсемейства MocR мотивы связывания, например, прямые повторы ATACCA у

GabR *B. subtilis* [213] и CTGGACYTAA у TauR *Rhodobacter capsulatus* [208], а также прямые и инвертированные повторы AAAGTGGW(−/T)CTA у PdxR *Corynebacterium glutamicum* [216] не имеют очевидного структурного сходства и, таким образом, не могут быть сопоставлены друг с другом. Отсутствие типичных для семейства GntR палиндромных мотивов связывания в случае MocR-регуляторов может быть связано с механизмом их димеризации – конфигурация «голова к хвосту» не адаптирована для связывания инвертированных повторов, однако подходит для прямых повторов с достаточно длинными спейсерами (что в целом характерно для подсемейства MocR) [76].

Следует отметить, что некоторые регуляторы подсемейств FadR и HutC также имеют мотивы связывания, которые не соответствуют общей консенсусной последовательности, характерной для данных групп семейства GntR, и/или не являются палиндромными [76]: например, FarR (прямые повторы TGTATTAWTT) [218], NagQ (прямые повторы TGGTATT) [188], BioR (палиндром TTATMKATAA) [219], NanR (прямые повторы TGGTATAW) [220].

В настоящее время известна кристаллическая структура ряда белков семейства GntR, например, в подсемействе FadR это FadR *Escherichia coli* (PDB – 1H9T, 1HW1, 1HW2), LldR *Corynebacterium glutamicum* (2DI3), TM0439 *Thermotoga maritime* (3SXK, 3SXY); в подсемействе HutC – YvoA (NagR) *Bacillus subtilis* (2WV0), HutC *Pseudomonas syringae pv. tomato* str. DC3000 (2PKH), AgaR *Enterococcus faecalis* V583 (3DDV); в подсемействе AraR – ДНК-связывающий домен AraR *Bacillus subtilis* (4EGY, 4EGZ, 4H0E). Однако только две из них (FadR *E. coli* и AraR *B. subtilis*) определены для комплекса с ДНК и, таким образом, могут служить основой для представлений о пространственной структуре и связывании регуляторов семейства GntR.

1.4.3 Пространственная структура FadR *E. coli* и AraR *B. subtilis* в комплексе с ДНК

Структурные данные показывают, что FadR *E. coli* формирует целый ряд неспецифических контактов с сахаро-фосфатным остовом ДНК, но только некоторые азотистые основания образуют специфические контакты [221,222]. Ключевой частью связывающего мотива FadR (ATCTGGTACGACCAGAT) являются три последовательно расположенных основания: T/A(4/14), G/C(5/13) и G/C(6/12), которые взаимодействуют с аминокислотными остатками His-65, Arg-35 и Arg-45, соответственно (здесь и далее через знак «/» обозначены симметричные нуклеотиды двух половин палиндрома) [221,222]. Arg-35 и Arg-45 образуют водородные связи с большой бороздкой ДНК, а His-65 встраивается в малую бороздку [221,222]. Центральные спейсерные нуклеотиды необходимы для задания правильного расстояния между двумя субъединицами димерного регулятора [76,221].

Специфические контакты образуются также между Thr-44 и 46 и центральной G/C парой азотистых оснований [222]. Кроме того, Thr-44, 46 и 47 взаимодействуют с сахаро-фосфатным остовом [221,222].

Аминокислоты Glu-34 и Glu-50 вступают в электростатические взаимодействия с Arg-35, 45 и 49, вероятно, стабилизируя этим их положение в комплексе; кроме того, Glu-34 и Arg-49 взаимодействуют с сахаро-фосфатным остовом [221,222]. Важным для связывания является также Gly-66, который способствует контакту His-65 с ДНК, так как любая другая аминокислота в этой позиции, в связи с наличием боковой цепи, привела бы к стерическим затруднениям [221,222].

Аминокислоты Ser-7, Pro-8 и Ala-9 образуют неспецифические контакты с сахаро-фосфатным остовом, вероятно, способствуя формированию специфических контактов другими аминокислотами [221,222]. Так, показано, что мутация Ala-9 в валин ухудшает способность FadR связывать ДНК [222]. Кроме того, неспецифические контакты с ДНК формируют также Ile-63, Lys-67 и Thr-69 [222].

Кристаллическая структура была определена также для ДНК-связывающего домена AraR *B. subtilis* в комплексе с двумя естественными сайтами (AATTGTTCGTACAAAT и ATTTGTCCGTATACAT) и одним искусственно сконструированным фрагментом ДНК (ATTTGTCCGTACATTT) [223]. Сайты различаются длиной спейсера между двумя ключевыми частями палиндромного мотива (TTG/CAW), что влияет на расположение двух мономеров регулятора друг относительно друга и их взаимодействие с ДНК [223].

Ключевые специфические контакты, присутствующие во всех трех структурах ДНК-связывающего домена AraR с ДНК, образованы Arg-63 с G(5/12’ или 14’) и Gln-83 с A(14/3’) (здесь и далее знаком «’» обозначено взаимодействие с нуклеотидами противоположной, комплементарной последовательности мотива, цепи ДНК) [210,223]. Gln-83 встраивается в малую бороздку ДНК, взаимодействуя с A; показано, что Gln-83 также может специфически связываться с T(16) [223]. Arg-63 в последовательности AraR соответствует Arg-45 регулятора FadR, в обоих случаях они входят в состав распознающей спирали и образуют Хугстиновские водородные связи с нуклеотидом G [210,223]. Таким образом, подобное взаимодействие может быть типичным для всего семейства GntR. Кроме того, специфические контакты азотистых оснований с аминокислотными остатками как FadR, так и AraR преимущественно формируются за счет нуклеотидов TKG/CMW в каждой из половин палиндромного сайта, которые, таким образом, представляют собой одну из наиболее важных частей мотива [221,223].

Специфическое взаимодействие также показано для Gly-84 и T(3). Gly-84 располагается внутри малой бороздки ДНК, и, как и в случае Gly-66 регулятора FadR *E. coli*, любая другая аминокислота в этой позиции привела бы к стерическим затруднениям [223].

Показано также, что Glu-52 образует водородные связи с Arg-63 и Arg-67, правильно ориентируя их для взаимодействия с ДНК [223]. Arg-67 не образует специфических контактов, однако взаимодействует с сахаро-фосфатным остовом [303]. Стабилизирующие водородные связи также формируются между сахаро-фосфатным остовом ДНК и аминокислотами Lys-26, Tyr-27 и Thr-65 [223].

Молекулы воды также могут вносить свой вклад в ДНК-белковые взаимодействия, опосредуя специфическое распознавание, например, играя роль связующего «мостика» между контактирующими нуклеотидами и радикалами аминокислот, или же уменьшая электростатическое отталкивание между образующими водородные связи донорными группами аминокислот и азотистых оснований [196]. Подобные специфические контакты (водородные связи, образованные с участием молекул воды) в комплексе AraR с ДНК образуют: Arg-63 с A(6'/11), His-64 с G(9/8') и T(10), а также Gly-84 с T(2, 16 и 15') [223]. Ион ацетата также может опосредовать взаимодействие Arg-63 с A(13) и Gly-84 с A(2) и T(1') [223].

1.5 Примеры метаболических систем, регулируемых транскрипционными факторами семейства GntR

В настоящей работе были подробно рассмотрены:

1. метаболизм гексуронатов, регулируемый гомологичными транскрипционными факторами UxuR и ExuR;

2. метаболизм малоната и пропионата, находящийся под контролем регуляторов MdcY и PrpR, соответственно.

1.5.1 Метаболизм гексуронатов у *E. coli*. Транскрипционные факторы UxuR и ExuR

Escherichia coli может использовать в качестве источника углерода два альдогексуроната: D-галактуронат и D-глюкуронат, которые метаболизируются по пути Ашвелла [224,225,226,227,228]. Гексуронаты поступают в клетку с помощью двух транспортных систем – GntP и ExuT (из семейств GntP and MFS, соответственно) [225,227]. GntP имеет специфичность к D-фруктуронату и D-тагатуронату, тогда как ExuT отвечает за импорт D-глюкуроната и D-галактуроната [225,228]. Кроме того, у *E. coli* присутствует транспортер β-D-глюкуронидов UidB и глюкуронидаза UidA, которые также могут обеспечивать поступление D-глюкуроната в клетку [96,229,230,231,232] (Рисунок 1).

И D-глюкуронат, и D-галактуронат метаболизируются до 2-кето-3-деокси-D-глюконата (KDG) – общего интермедиата двух параллельных катаболических путей [96,224,227,228,232]. Эти пути включают один общий фермент D-глюкуронат/D-галактуронат изомеразу (UxaC), а также две пары аналогичных по функциям ферментов – D-маннонат и D-альтронат гидролазы UxuA и UxaA и оксидоредуктазы UxuB и UxaB, соответственно (Рисунок 1) [224,225,226,227,228,232].

UxuB и UxaB *E. coli* гомологичны (идентичность – 26%; сходство последовательности – 42%), однако хорошо различимы, тогда как для UxuA и UxaA не отмечено существенного сходства.

Метаболизм гексуронатов у *E. coli* регулируется двумя родственными (46% – идентичность аминокислотной последовательности) белками подсемейства FadR семейства GntR – транскрипционными факторами UxuR и ExuR [96,225]. UxuR осуществляет контроль метаболизма D-глюкуроната, репрессируя экспрессию оперонов *uxuAB, uidABC, gntP*, а также собственного гена *uxuR* [96,224,225,229,230,228,231,232]. ExuR же осуществляет негативный контроль экспрессии генов, участвующих в метаболизме как D-галактуроната, так и D-глюкуроната, включая *exuT, uxaCA, uxaB, uxuAB* и *exuR* [96,224,225,226,227,228,231,232].

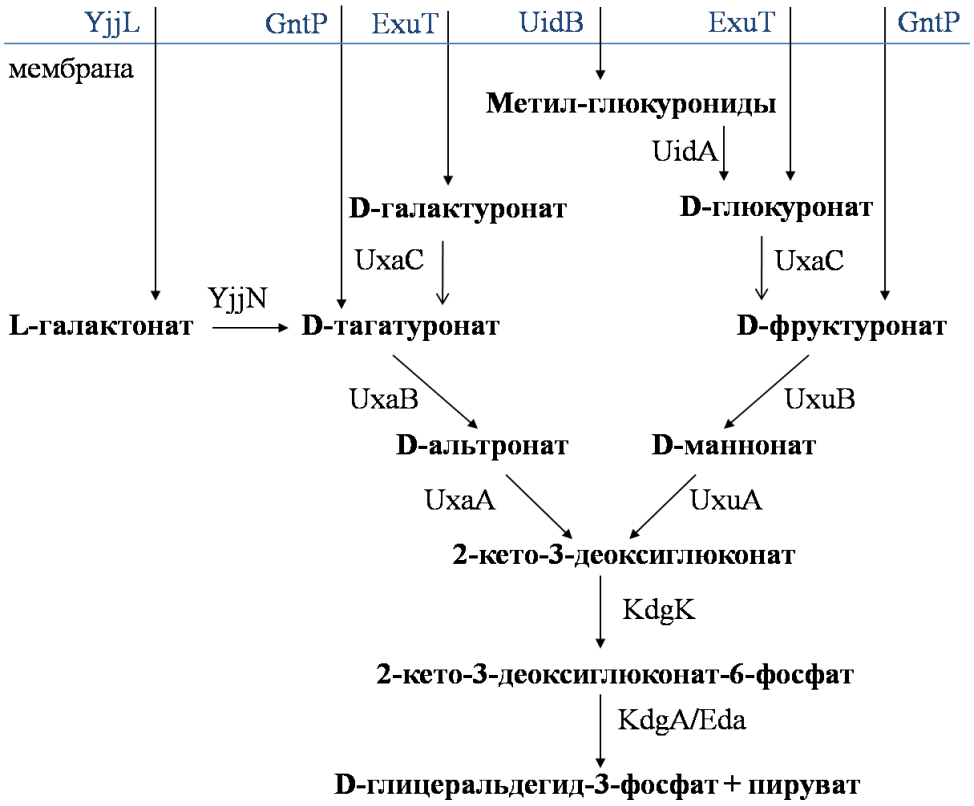


Рисунок 1. Схема метаболизма гексуронатов у *E. coli*

Транспортные белки показаны голубым цветом, ферменты – черным.

1.5.2 Метаболизм малоната и пропионата у Proteobacteria. Транскрипционные факторы MatR/MdcY, MdcR, PrpR

Малонат может использоваться как источник углерода многими бактериями, например, *Acinetobacter calcoaceticus, Klebsiella pneumoniae, Pseudomonas putida, Rhizobium leguminosarum* [233,234,235,236,237]. Известно, что малонат является конкурентным ингибитором сукцинат дегидрогеназы, играет важную роль в симбиозе и азотном метаболизме растений семейства бобовые и азотфиксирующих бактерий, а также используется в процессе синтеза антибиотиков промышленными штаммами *Streptomyces* spp. [234,235].

Описаны две группы структурных генов, участвующих в метаболизме малоната. Первая метаболическая система была охарактеризована у *Rhizobium leguminosarum* [234,235]. Она включает оперон из трех генов, *matA, matB* и *matC*, кодирующих малонил-CoA декарбоксилазу, малонил-CoA синтетазу и транспортер малоната, соответственно (Рисунок 2), а также дивергентно транскрибируемый ген малонатного регулятора *matR*. Транскрипционный фактор MatR принадлежит к подсемейству FadR семейства GntR.

Другая система метаболизма малоната была описана у *Acinetobacter calcoaceticus* [233,236]. Она включает структурный оперон *mdcLMACDEGBH*, кодирующий субъединицы транспортера малоната (MdcLM), малонат декарбоксилазы и вспомогательных белков (MdcABCDEGH) (Рисунок 2), и дивергентно расположенный ген регулятора *mdcY*. Как и MatR, MdcY принадлежит к подсемейству FadR семейства GntR. У некоторых Gammaproteobacteria в составе малонатного оперона присутствует другой ген, кодирующий транспортер малоната – *mdcF*, а не *mdcLM*, как у *A. calcoaceticus* [234,237]. Организация малонатного оперона у *K. pneumoniae* и *P. putida* аналогична таковой у *A. calcoaceticus*, однако у этих, а также многих других Gammaproteobacteria с генами транспорта и утилизации малоната колокализован регуляторный ген, кодирующий транскрипционный фактор MdcR из семейства LysR [236,237]. MdcR является активатором экспрессии генов *mdc*, а также репрессирует транскрипцию собственного гена [236,237].

Пропионат также может служить источником углерода для многих бактерий, например, *E. coli, Salmonella enterica, Ralstonia eutropha* и т.д. [238,239]. Метаболизм этого соединения тесно связан с метаболизмом малоната, а также с центральным метаболизмом, например, циклом трикарбоновых кислот (ЦТК). Превращение пропионата в пируват и сукцинат, входящий в ЦТК, осуществляют ферменты метилцитратного пути: пропионил-CoA синтетаза PrpE, 2-метилцитрат синтаза PrpC, 2-метилцитрат дегидратаза PrpD, 2-метилизоцитрат лиаза PrpB, 2-метилаконитат гидратаза AcnB(AcnM) или AcnD, и вспомогательный белок аконитазы PrpF

(Рисунок 2) [238,239]. AcnB представляет собой бифункциональный фермент и в качестве цитрат/изоцитрат изомеразы также входит в состав ЦТК и глиоксилатного шунта [239].

Пропионат может также включаться в цитрамалатный цикл, в состав которого входят следующие ферменты, осуществляющие превращение пропионил-CoA в сукцинил-CoA через интермедиат 2-метилмалонил-CoA (Рисунок 2): пропионил-CoA карбоксилаза (α- и β-субъединицы) PccAB, метилмалонил-CoA эпимераза Epi и метилмалонил-CoA мутаза MutB, а также вспомогательный белок MeaB (предположительно, металлошаперон, участвующий в защите и сборке MutB) [240,241].

Известно, что транскрипционным активатором генов пропионатного метаболизма *prp* у *E. coli* и *Ralstonia eutropha* HF39 является σ^{54}-зависимый транскрипционный фактор PrpR из семейства FIS [238]. Кроме того, у *Pseudomonas* spp. и *Vibrio cholerae* был описан иной регулятор метаболизма пропионата – транскрипционный фактор семейства GNTR, ген которого колокализован с кластером генов *prp* [238].

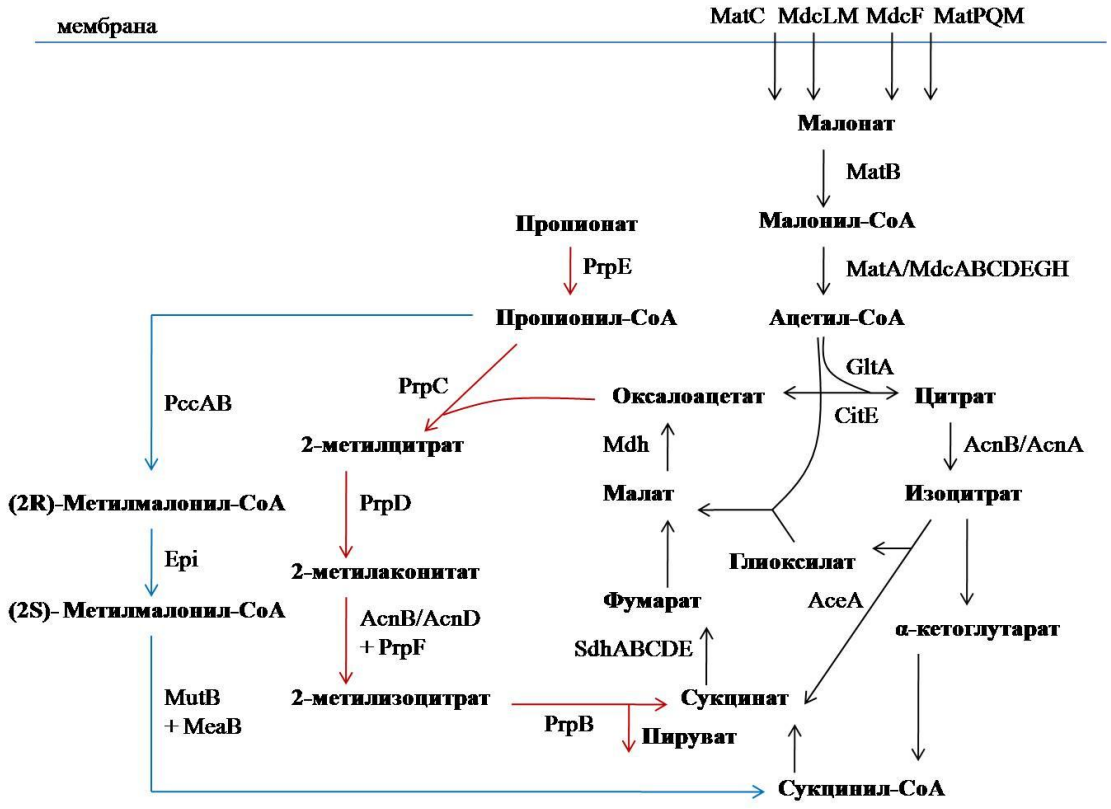


Рисунок 2. Схема метаболизма малоната и пропионата

Метилцитратный путь выделен красным цветом, часть цитрамалатного цикла – синим.

Глава 2. Материалы и методы

2.1 Программное обеспечение и методы биоинформатического анализа

Последовательности геномов исследованных бактерий были взяты из базы данных GenBank [144]; все геномы и соответствующие трехбуквенные обозначения приведены в Приложении А. Всего было исследовано 307 геномов.

Гомологи исследованных в данной работе транскрипционных факторов были идентифицированы с помощью программы PSI-BLAST [150] с заданными параметрами (пороговое значение e-value = 10^{-20}). Ортологи определялись при помощи построения филогенетических деревьев для найденных гомологов, а также с учетом геномного контекста (например, колокализации генов транскрипционных факторов с генами определенных метаболических путей). Как правило, в состав ортологической группы входило по одному фактору транскрипции из каждого генома. Однако в некоторых случаях, вероятно, являющихся результатом недавних дупликаций или близкородственных горизонтальных переносов, несколько паралогичных транскрипционных факторов включались в одну и ту же ортологическую группу.

Для выравнивания нуклеотидных и аминокислотных последовательностей использовалась программа MUSCLE (параметры по умолчанию) [152]. Для построения филогенетических деревьев использовались программы пакета PHYLIP (параметры по умолчанию) [242]. Визуализация филогенетических деревьев осуществлялась с помощью программы Dendroscope [243].

Для каждого из исследованных транскрипционных факторов семейства GntR была проведена реконструкция соответствующих регулонов: были идентифицированы потенциальные регулируемые гены и сайты связывания.

Потенциальные мотивы связывания идентифицировались методом филогенетического футпринтинга [4,9]. Множественные выравнивания 5'-областей ортологичных генов были использованы для идентификации групп консервативных позиций, основываясь на предположении, что сайты связывания являются более консервативными по сравнению с соседними нефункциональными участками межгенных областей.

Поиск потенциальных сайтов связывания в геномах осуществлялся при помощи матриц позиционных весов нуклеотидов (профилей, PWM) [10,11]. Построение профилей для мотивов связывания каждого из исследованных транскрипционных факторов проводилось с помощью программы SignalX, как было описано ранее [10,11,174], с использованием обучающей выборки

5’-областей генов, для которых известна или предполагается регуляция (как правило, это гены собственно факторов транскрипции, так как они часто авторегулируемы [7,20], а также колокализованные с ними гены, так как регулируемые гены часто имеют тенденцию располагаться в одном локусе с геном регулятора [7,11]).

Для поиска ортологов генов и потенциальных сайтов в геноме использовался пакет программ GenomeExplorer [174], а также веб-сервер RegPredict [244]. Поиск сайтов связывания транскрипционных факторов проводился в области от –400 до +50 нуклеотидов относительно старта трансляции. Диаграммы Logo, отображающие структуру мотивов связывания, были построены при помощи программы WebLogo [245].

Порог весов для идентифицированных сайтов выбирался так, чтобы количество генов, перед которыми предсказаны сайты связывания, не превышало 5% для данного генома (в ряде случаев для длинных консервативных мотивов число потенциальных сайтов не превышало 50 на геном), а также из расчета, что включенные в состав регулона гены функционально относятся к соответствующей метаболической системе. В большинстве случаев пороговым значением был минимальный вес сайта из обучающей выборки. Сайты с более слабым весом (на 10% ниже порога) также принимались в рассмотрение, если их позиция была аналогична таковой для сильных (с надпороговым весом) сайтов перед ортологичными генами, и не наблюдалось более сильных конкурирующих сайтов связывания в той же межгенной области.

Для подтверждения принадлежности определенного гена к регулону применялся метод проверки соответствия. Ген включался в состав регулона, если в его регуляторной области или же регуляторной области соответствующего оперона был обнаружен потенциальный сайт связывания транскрипционного фактора, сохраняющийся перед его ортологами в нескольких геномах (обычно, как минимум, в трех-четырех; конкретное число зависело от количества и эволюционной близости исследуемых геномов в данной группе, так как в близкородственных организмах консервативность участка межгенной области может определяться остаточным сходством последовательности) [10,11]. Эмпирически было установлено, что более строгий критерий может приводить к отсеиванию некоторых истинных членов регулона, тогда как более слабый – к большому количеству ложных предсказаний. Следует отметить, что в большинстве случаев количество ортологичных генов с потенциальными сайтами связывания было существенно больше четырех, что позволяло уверенно применять метод проверки соответствия.

При реконструкции регулонов проводилось также предсказание оперонной структуры генов, перед которыми найдены сайты связывания: гены относили к одному оперону, если они транскрибировались в одном направлении, межгенное расстояние не превышало 200

нуклеотидов, и подобная организация сохранялась в ряде геномов (конкретное число геномов варьировало, см. ранее).

Реконструированные регулоны размещены в базе данных RegPrecise [246] и доступны по ссылке http://regprecise.lbl.gov/RegPrecise/collection_tffam.jsp?tffamily_id=25.

Для осуществления анализа корреляций аминокислот ДНК-связывающих HTH-доменов транскрипционных факторов семейства GntR и нуклеотидов соответствующих сайтов связывания были выбраны только те транскрипционные факторы, предсказанные мотивы связывания которых соответствовали палиндромному консенсусу для семейства GntR. Для сравнения и верификации результатов корреляционного анализа были использованы данные кристаллической структуры FadR *E. coli* (PDB – 1H9T, 1HW1, 1HW2) и AraR *B. subtilis* (4EGY, 4EGZ, 4H0E) в комплексе с ДНК. Корреляции были определены для ДНК-связывающих HTH-доменов транскрипционных факторов, нумерация позиций аминокислот и нуклеотидов осуществлялась от нуля.

Анализ корреляций осуществлялся с помощью программы Prot-DNA-Korr (http://bioinf.fbb.msu.ru/Prot-DNA-Korr/main.html) отдельно для каждого из исследованных подсемейств (FadR, HutC и YtrA). Корреляции рассчитывались для каждой пары столбцов выравниваний аминокислотных последовательностей HTH-доменов транскрипционных факторов и нуклеотидных последовательностей сайтов связывания. Так как мотивы связывания различных регуляторов различались по длине, более короткие сайты были фланкированы так, чтобы соответствовать наиболее длинному мотиву в выборке. В качестве меры корреляции использовалась взаимная информация, статистическая значимость рассчитывалась как Z-score. Скоррелированные пары позиций представлены в виде карт интенсивности (где цвет ячейки соответствует статистической значимости корреляции для пары позиций), а также таблиц сопряженности (приведены ожидаемые и наблюдаемые значения вероятностей для статистически значимых корреляций, а также χ^2).

Статистический анализ данных для дивергонов и дополнительных боксов мотивов связывания проводился при помощи программы STATISTICA [247].

Глава 3. Транскрипционные факторы семейства GntR и их мотивы связывания: ДНК-белковые взаимодействия, особенности структуры и расположения сайтов

3.1 Общая статистика

В настоящей работе были исследованы транскрипционные факторы трех подсемейств семейства GntR: FadR, HutC и YtrA. Сайты связывания были предсказаны для 1252 транскрипционных факторов семейства GntR из 307 бактериальных геномов (Приложение А). Исследованные факторы транскрипции были классифицированы на 64 ортологические группы. Диаграммы Logo мотивов связывания для каждой из ортологических групп регуляторов приведены в Приложении Б. Содержание исследованных транскрипционных факторов в индивидуальных геномах и их распределение внутри разных таксономических группых варьировало (Таблица 1). Так, например, регуляторы подсемейства YtrA распространены преимущественно среди представителей Firmicutes, тогда как транскрипционные факторы подсемейства FadR типичны для Proteobacteria.

Таблица 1. Исследованные факторы транскрипции семейства GntR. Общие данные

Количество/Подсемейство		**FadR**	**HutC**	**YtrA**
Ортологические группы		36	16	12
Исследованные транскрипционные факторы		634	389	229
Регулируемые опероны		1740	975	283
Сайты (включая дивергентные и множественные)		2396	1341	294
Таксономическое распределение исследованных транскрипционных факторов				
Proteobacteria	Alpha	76	39	3
	Beta	151	64	0
	Gamma	308	112	25
	Delta	10	1	0
Firmicutes	Bacilli	18	97	89
	Clostridia	1	14	53
Actinobacteria		64	60	43
Thermotogae		0	0	14
Chloroflexi		6	0	1
Bacteroidetes		0	1	0
Cyanobacteria		0	1	0
Archaea		0	0	1

3.2 Анализ корреляций аминокислот HTH-доменов транскрипционных факторов семейства GntR и нуклеотидов соответствующих сайтов связывания

С целью предсказания вероятных ДНК-белковых взаимодействий, для трех подсемейств транскрипционных факторов семейства GntR (FadR, HutC и YtrA) был проведен анализ

корреляций аминокислотных последовательностей и нуклеотидов соответствующих сайтов связывания. Для сравнения и верификации результатов корреляционного анализа были использованы данные кристаллической структуры FadR *E. coli* и AraR *B. subtilis* в комплексе с ДНК. Так как корреляции были определены для ДНК-связывающих HTH-доменов транскрипционных факторов, для FadR и AraR были сопоставлены позиции в HTH-домене аминокислот, ключевых для взаимодействия этих регуляторов с ДНК (Таблица 2) [221,222,223].

3.2.1 Подсемейство FadR

Общая консенсусная последовательность сайтов связывания транскрипционных факторов подсемейства FadR представляет собой A/T-богатый палиндром с высоко консервативными группами TKGT/ACMA (Рисунок 3), вероятно, играющими ключевую роль в ДНК-белковом взаимодействии. Характерное расстояние между консервативными парами оснований GT и AC у большинства мотивов связывания транскрипционных факторов подсемейства FadR составляет 3 нуклеотида (например, DgoR, ExuR, FadR, GlcC, LldR, PdhR, и т.д.). Однако в ряде ортологических групп это расстояние равно 2 нуклеотидам (GntR, HpxS, HypR, MdcY, PrpR, UxuR и некоторые другие), соответственно, в ходе анализа корреляций такие сайты были включены в выборку после вставки однонуклеотидного пробела в середине мотива. Некоторые регуляторы подсемейства FadR, например, BioR (мотив связывания – палиндром TTATMKATAA) [219], NanR (прямые повторы TGGTATAW) [220], были исключены из корреляционного анализа, так как консенсусная последовательность их мотивов связывания не соответствовала общему консенсусу семейства GntR.

В связи с симметричной структурой анализируемых мотивов связывания и, следовательно, соответствующих карт интенсивности, корреляции, как правило, показаны для G/C или A/T пар, тогда как дальнейшее различение контактов с G или C, а также, соответственно, с A или T не всегда возможно и требует использования дополнительных соображений, например, сопоставления данных о корреляциях с контактами, известными для ДНК-белковых комплексов FadR и AraR, учета донорно-акцепторных свойств и т.п.

Анализ корреляций аминокислот HTH-доменов и нуклеотидов сайтов связывания показывает, что для подсемейства FadR в целом скоррелированные нуклеотидные и аминокислотные позиции, вероятно, определяющие специфичность связывания, хорошо соответствуют парам контактирующих позиций, известных для ДНК-белковых структур FadR *E. coli* и AraR *B. subtilis* (Рисунок 3, Таблица 2, Таблица 3).

Таблица 2. ДНК-белковые взаимодействия в комплексах FadR *E.coli* и AraR *B. subtilis* с ДНК

Позиция в HTH-домене	Аминокислота FadR *E.coli*	Функция в комплексе FadR-ДНК	Аминокислота AraR *B. subtilis*	Функция в комплексе AraR-ДНК
0	Ser-7	Неспецифический контакт с сахаро-фосфатным остовом	Pro-25	-
1	Pro-8	Неспецифический контакт с сахаро-фосфатным остовом	Lys-26	Неспецифический контакт с сахаро-фосфатным остовом
2	Ala-9	Неспецифический контакт с сахаро-фосфатным остовом	Tyr-27	Неспецифический контакт с сахаро-фосфатным остовом
27	Glu-34	Неспецифический контакт с сахаро-фосфатным остовом; ионная связь с Arg-35, Arg-45, Arg-49	Glu-52	Водородные связи с Arg-63, Arg-67
28	Arg-35	Специфический контакт Arg-G	Asn-53	-
37	Thr-44	Неспецифический контакт с сахаро-фосфатным остовом; специфические контакты Thr-C и Thr-G	Ser-62	-
38	Arg-45	Специфический контакт Arg-G	Arg-63	Специфический контакт Arg-G; специфический контакт Arg-A, опосредованный молекулами воды или ионами ацетата
39	Thr-46	Неспецифический контакт с сахаро-фосфатным остовом; специфические контакты Thr-C и Thr-G	His-64	Специфические контакты His-G и His-T, опосредованные молекулами воды
40	Thr-47	Неспецифический контакт с сахаро-фосфатным остовом	Thr-65	Неспецифический контакт с сахаро-фосфатным остовом
42	Arg-49	Неспецифический контакт с сахаро-фосфатным остовом	Arg-67	Неспецифический контакт с сахаро-фосфатным остовом
43	Glu-50	Ионные связи с Arg-35, Arg-45, Arg-49	Lys-68	-
56	Ile-63	Неспецифический контакт с сахаро-фосфатным остовом	Ser-81	-
58	His-65	Специфические контакты His-A и His-G	Gln-83	Специфические контакты Gln-A и Gln-T
59	Gly-66	Неспецифический контакт с сахаро-фосфатным остовом; предотвращение стерического затруднения	Gly-84	Специфический контакт Gly-T; специфические контакты Gly-T и Gly-A, опосредованные молекулами воды или ионами ацетата; предотвращение стерического затруднения
60	Lys-67	Неспецифический контакт с сахаро-фосфатным остовом	Gly-85	-
62	Thr-69	Неспецифический контакт с сахаро-фосфатным остовом	Gly-86	-

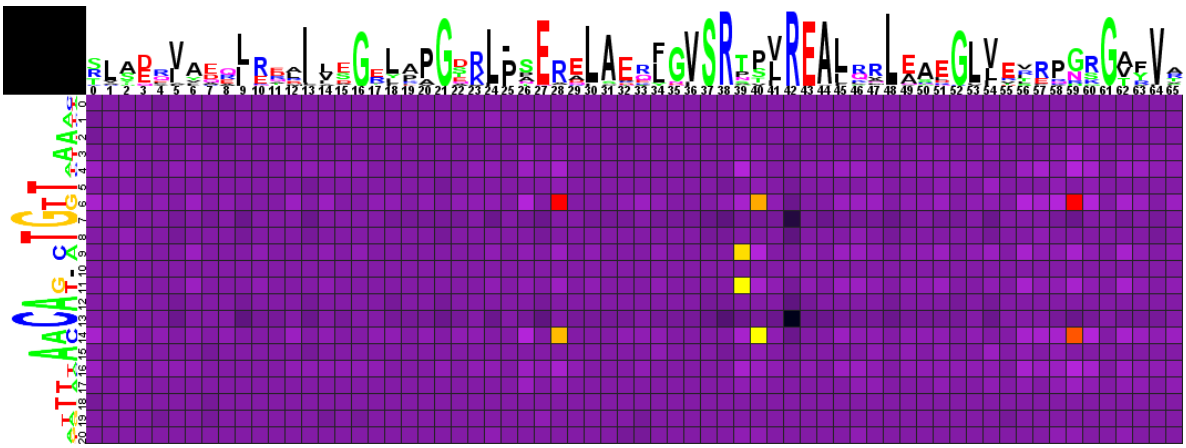


Рисунок 3. Карта интенсивности корреляций аминокислот HTH-доменов транскрипционных факторов подсемейства FadR и нуклеотидов соответствующих сайтов связывания

Диаграммы Logo ДНК-связывающих HTH-доменов и сайтов связывания показаны, соответственно, сверху и слева от карты интенсивности. Общая высота символов в каждой позиции соответствует информационному содержанию, тогда как высота конкретного символа пропорциональна частоте встречаемости аминокислоты/нуклеотида в данной позиции. Уровень корреляции показан цветом и изменяется по градиенту от желтого до красного для статистически значимо (выше автоматически определяемого порога) коррелирующих пар аминокислот и нуклеотидов; прочие пары показаны фиолетово-черным.

Так, для аминокислот в позиции 28 HTH-домена, формирующих один из специфических контактов FadR *E.coli* с ДНК [221,222], показана корреляция с нуклеотидами в позициях 6/14. В данной позиции наиболее часто встречается аргинин, показано его предпочтительное взаимодействие с парой G/C, тогда как контакт с парой A/T достоверно избегается. Более редкая в данной позиции аспарагиновая кислота также достоверно коррелирует с G/C парой. В соответствии с электрохимическими свойствами этих аминокислот, можно предположить, что вероятными контактами в этой позиции являются Arg-G и Asp-C.

Кроме того, с нуклеотидами 6/14 коррелируют также аминокислоты в позициях 40 и 59, которые являются важными для взаимодействия с ДНК у FadR *E. coli* и AraR *B. subtilis* (Таблица 2). Наиболее часто встречающиеся в позиции 40 аминокислоты – пролин и серин. Серин в данной позиции ассоциирован с наличием G/C пары (вероятно, образуя контакт с G), тогда как в случае пролина G/C пара достоверно избегается.

Наиболее частый в позиции 59 глицин коррелирует с наличием G/C пары, при этом A/T пара достоверно избегается, однако эта корреляция может не отражать непосредственное ДНК-белковое взаимодействие. Наличие в данной позиции глицина, не имеющего боковой цепи, может быть вызвано стерическими причинами, как это было показано для FadR *E. coli*

[221,222]. В позиции 59 также часто встречается аспарагин, для которого наблюдается предпочтительное взаимодействие с A/T парой, однако эта тенденция статистически недостоверна.

Кроме того, аминокислоты в позиции 39 HTH-домена, для которых показано участие во взаимодействии FadR *E. coli* и AraR *B. subtilis* с ДНК (Таблица 2), коррелируют с центральными нуклеотидами 9/11. Аспарагин в данной позиции достоверно коррелирует с A/T парой, вероятно, взаимодействуя с A, в соответствии с описанными ранее закономерностями. Треонин также часто встречается в позиции 39, и имеется тенденция к предпочтению им A/T пары, однако она статистически недостоверна.

3.2.2 Подсемейство HutC

Консенсусная последовательность мотивов связывания подсемейства HutC имеет высокое сходство с таковой для подсемейства FadR. Для подавляющего большинства сайтов подсемейства HutC расстояние между группами GT и AC мотива равно 4 нуклеотидам. Исключения составляют такие транскрипционные факторы как FarR (мотив связывания – прямые повторы TGTATTAWTT) [218], NagQ (прямые повторы TGGTATT) [188], SdhR (палиндром с внутренней симметрией TCTTATGTCTTATATAAGACATAAGA) [248]. Эти транскрипционные факторы не были включены в корреляционный анализ, так как соответствующие мотивы связывания не могли быть выровнены и сопоставлены с основной группой сайтов.

Анализ корреляций (Рисунок 4) показывает, что в подсемействе HutC позиции, определяющие специфичность связывания, сходны с таковыми для FadR *E.coli* и для подсемейства FadR в целом (Таблица 2, Таблица 3). В частности, для аминокислот в позиции 28 показана корреляция с нуклеотидами 8/17. Как и в подсемействе FadR, аргинин, наиболее часто встречающийся в этой позиции, значимо коррелирует с G/C парой (в соответствии с электрохимическими свойствами, вероятный контакт Arg-G), тогда как контакт с A/T парой достоверно избегается. Аспарагин также часто присутствует в данной позиции, и имеется тенденция к предпочтению им A/T пары, однако она статистически недостоверна.

С нуклеотидами 8/17 также коррелируют и аминокислоты в позициях 43 и 62, участвующие в связывании с ДНК у FadR *E.coli* (Таблица 2). Наиболее часто представленные здесь аминокислоты – это аргинин, глутамин и лизин в позиции 43, и треонин и серин в позиции 62, однако статистически значимого предпочтения какой-либо пары нуклеотидов для этих аминокислот не выявлено. В то же время, для более редкого в позиции 62 триптофана показана достоверная корреляция с G/C парой (вероятный контакт Trp-C).

Кроме того, аминокислоты в позиции 39 HTH-домена, как это было показано и для подсемейства FadR, коррелируют с центральными нуклеотидами 12/13. Наиболее часто встречающаяся в этой позиции аминокислота – метионин, однако тенденция к предпочтению им A/T пары статистически недостоверна. В то же время, более редкая в данной позиции аспарагиновая кислота достоверно коррелирует с G/C парой; в соответствии с электрохимическими свойствами этой аминокислоты, вероятным контактом является Asp-C.

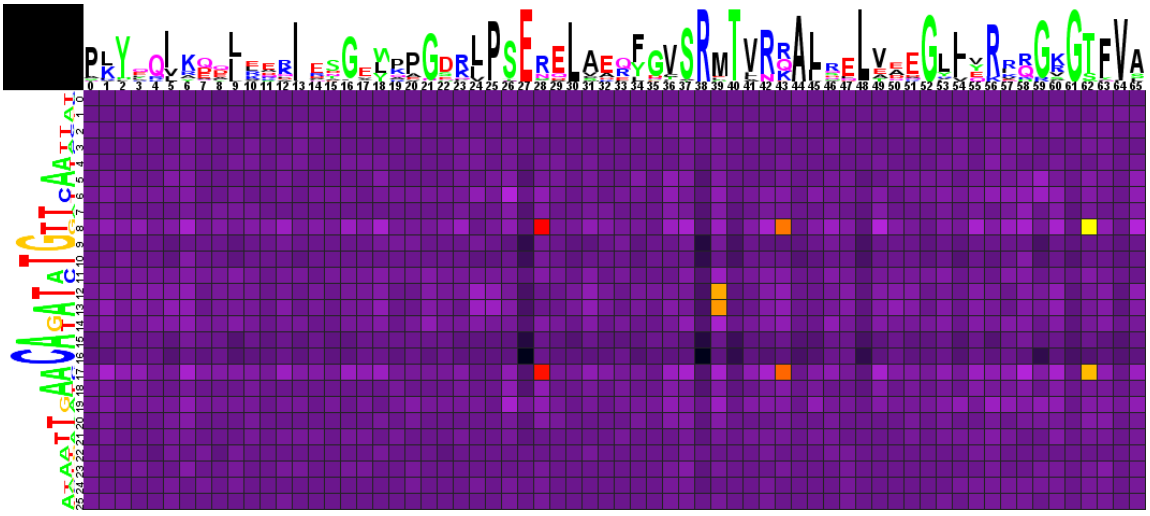


Рисунок 4. Карта интенсивности корреляций аминокислот HTH-доменов транскрипционных факторов подсемейства HutC и нуклеотидов соответствующих сайтов связывания

Обозначения как на Рисунке 3.

3.2.3 Подсемейство YtrA

Это подсемейство регуляторов имеет ряд отличий от остальных исследованных подсемейств семейства GntR. Типичная для транскрипционных факторов подсемейств FadR и HutC дивергентная организация регулируемых оперонов крайне редка в подсемействе YtrA. Кроме того, для YtrA подсемейства типичны одиночные сайты; двойные и тройные сайты связывания, довольно распространенные среди подсемейств FadR и HutC, были идентифицированы всего в нескольких случаях. Было показано, что подавляющее число регулонов подсемейства YtrA состоит из единственного оперона, включающего гены, кодирующие АТФ-зависимые кассетные (ABC) транспортеры, что согласуется с ранее опубликованными данными [76].

Мотивы связывания транскрипционных факторов подсемейства YtrA существенно длиннее, чем типичные мотивы регуляторов остальных подсемейств семейства GntR [76].

Несмотря на сильное отличие структуры сайтов, высокая консервативность HTH-доменов внутри всего семейства позволяет точно сопоставить их для транскрипционных факторов различных подсемейств, и проведенный корреляционный анализ показывает, что позиции аминокислот в составе HTH-домена, определяющие специфичность связывания с ДНК, сходны у всех исследованных подсемейств, в том числе и YtrA (Таблица 3).

Так, нуклеотиды 12-13/29-30, вероятно, участвуют в специфических взаимодействиях с аминокислотными остатками в позициях 27 и 28 (Рисунок 5). Как и в случае подсемейств FadR и HutC, наиболее частым в позиции 28 является аргинин, однако его корреляции с G/C парой (нуклеотиды 12/30) и A/T парой (нуклеотиды 13/29) статистически недостоверны. В то же время, более редкие в позиции 28 аспарагин и тирозин достоверно ассоциированы с наличием пары A/T в позициях 12/30; в соответствии с описанными выше закономерностями, вероятными контактами являются Asn-A и Tyr-A. В позиции 27 наиболее часто встречается валин, однако для него не выявлено значимых предпочтений каких-либо нуклеотидов, тогда как треонин, также часто представленный в этой позиции, достоверно коррелирует с A/T парой в позициях 12/30 и 13, предположительно формируя контакт Thr-A.

Корреляции также найдены для пар нуклеотидов 16-17/25-26 и аминокислотных остатков в позициях 37 и 39, важных для взаимодействия с ДНК у FadR *E.coli* (Таблица 2). В обеих позициях наиболее часто присутствует аспарагин, имеется тенденция к предпочтению им A/T пары в позициях 16-17/25-26, однако она статистически недостоверна. Более редкий в позиции 37 серин и изолейцин в позиции 39 достоверно коррелируют с A/T в позициях 16-17/25-26, тогда как гистидин в позиции 39 достоверно ассоциирован с G/C парой в позициях 25 и 26. В последнем случае, контактом, вероятно, является His-G, что согласуется с предпочтительными взаимодействиями для полярных положительно заряженных аминокислот, а также с наличием в этой позиции контакта His-G в комплексе AraR *B. subtilis* с ДНК (Таблица 2).

Кроме того, корреляции отмечены для аланина в позиции 39 с A/T в позициях 12,13 и G/C в позиции 24, а также для глицина в позиции 44 с A/T в позиции 13.

Таким образом, несмотря на значительные различия в структуре сайтов связывания, взаимодействие транскрипционных факторов подсемейства YtrA с ДНК организовано сходным образом с регуляторами подсемейств FadR и HutC.

Следует отметить, что карта интенсивности корреляций для подсемейства YtrA не точно симметрична, в отличие от таковых для подсемейств FadR и HutC. Это объясняется тем, что, несмотря на палиндромность мотивов связывания семейства GntR в целом, каждый индивидуальный сайт может отличаться от симметричного консенсуса. В случае подсемейств FadR и HutC большой размер выборки проанализированных сайтов (Таблица 1) сглаживает эти

отличия, тогда как в малочисленном подсемействе YtrA размер выборки на порядок меньше, что может приводить к некоторой асимметрии полученной карты интенсивности, поскольку вклад отдельного сайта в общую картину оказывается более выраженным. Кроме того, влияние на асимметричность также оказывает практически полное отсутствие дивергентно регулируемых оперонов в подсемействе YtrA.

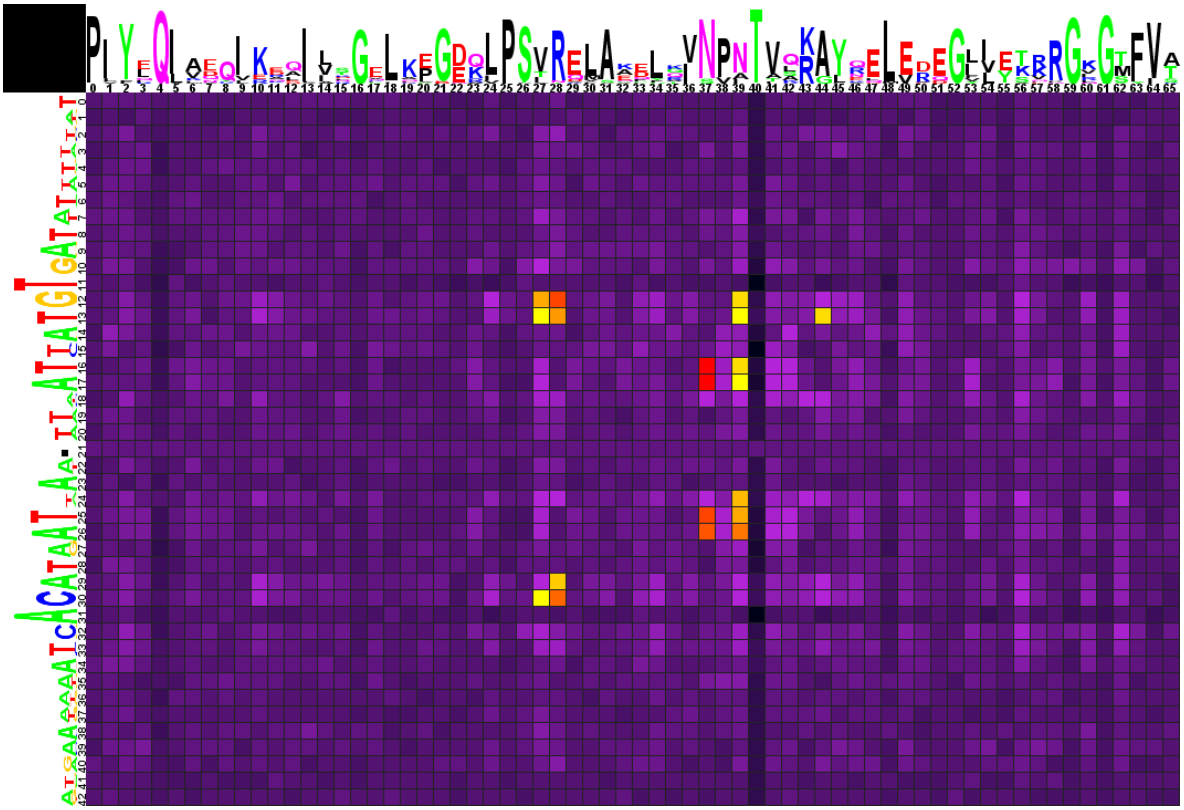


Рисунок 5. Карта интенсивности корреляций аминокислот HTH-доменов транскрипционных факторов подсемейства YtrA и нуклеотидов соответствующих сайтов связывания

Обозначения как на Рисунке 3.

3.2.4 Общие закономерности ДНК-белковых корреляций в семействе GntR

Как было упомянуто ранее, из литературных данных известно, что наиболее часто вступающими во взаимодействие с ДНК аминокислотами являются Arg, Asn, Asp, Gln, Gly, Lys, Ser и Thr, а наиболее предпочтительными контактами – Arg-G, Asn-A, Asp-C, Gln-A, Glu-C, His-G, Lys-G и Ser-G, а также в меньшей степени Ala-C, Cys-G, Gly-G, Leu-A, Thr-G и Trp-C [196]. Полученные данные корреляционного анализа согласуются с этими закономерностями: большинство коррелирующих пар включает именно перечисленные выше аминокислоты и нуклеотиды (Таблица 3).

Таблица 3. Скоррелированные пары аминокислот HTH-доменов транскрипционных факторов семейства GntR и нуклеотидов соответствующих сайтов связывания

Подсемейство	Позиция аминокислоты в HTH-домене	Позиция нуклеотида в сайте	Аминокислота	Нуклеотид											
				A			T			G			C		
				Наблюдаемое	Ожидаемое	χ^2	Наблюдаемое	Ожидаемое	χ^2	Наблюдаемое	Ожидаемое	χ^2	Наблюдаемое	Ожидаемое	χ^2
FadR	28	6	Arg	1,87	3,54	0,79	24,99 ↓	122,39	77,52	193,21 ↑	86,37	132,15	0,89	8,64	6,96
			Asp	0,61	0,34	0,21	11,22	11,71	0,02	0,14	8,26	7,98	9,17 ↑	0,83	84,23
		14’	Arg	27,76 ↓	121,66	72,48	4,92	4,47	0,05	1,84	9,49	6,17	186,43 ↑	85,33	119,81
			Asp	10,99	11,64	0,04	0,61	0,43	0,07	9,41 ↑	0,91	79,50	0,14	8,16	7,88
	39	9	Asn	8,20	25,99	12,18	50,41 ↑	12,74	111,42	2,33	2,19	0,01	8,29	28,31	14,16
		11’	Asn	48,79 ↑	11,18	126,49	8,44	27,52	13,24	10,09	28,62	11,99	1,91	1,91	0,00
	40	6	Pro	2,60	3,09	0,08	170,53	106,72	38,15	8,22 ↓	75,32	59,77	11,30	7,54	1,88
			Ser	1,09	1,83	0,30	11,35	63,41	42,74	101,07 ↑	44,75	70,89	0,96	4,48	2,77
		14’	Pro	167,67	106,09	35,75	3,13	3,89	0,15	10,43	8,28	0,56	11,43 ↓	74,40	53,30
			Ser	10,54	63,03	43,71	0,50	2,31	1,42	0,30	4,92	4,33	103,12 ↑	44,20	78,53
	59	6	Gly	3,85	3,65	0,01	40,44 ↓	126,01	58,11	182,22 ↑	88,92	97,89	0,97	8,90	7,07
		14’	Gly	43,08 ↓	125,26	53,91	5,14	4,60	0,06	1,41	9,77	7,15	177,84 ↑	87,85	92,19
HutC	28	8	Arg	4,64	7,32	0,98	9,38 ↓	70,72	53,21	107,61 ↑	40,36	112,09	1,85	5,09	2,07
		17’	Arg	22,09	76,57	38,76	1,31	4,79	2,53	2,43	5,72	1,89	97,65 ↑	36,40	103,03
	39	12	Asp	0,10	1,14	0,96	0,10	6,32	6,13	0,10	0,11	0,00	8,16 ↑	0,88	60,15
		13’	Asp	0,10	6,25	6,06	0,10	1,12	0,93	8,16 ↑	0,95	54,88	0,10	0,14	0,01
	62	17’	Trp	0,11	19,50	19,27	0,11	1,22	1,00	0,11	1,46	1,24	31,10 ↑	9,27	51,42
YtrA	27	12	Thr	2,25	1,74	0,15	34,42 ↑	7,07	105,89	0,14	26,65	26,37	0,14	1,50	1,23
		13	Thr	25,07 ↑	5,49	69,86	1,80	26,80	23,31	3,69	1,48	3,26	6,40	3,19	3,24
		30’	Thr	36,52 ↑	8,14	98,90	0,14	1,48	1,21	0,14	1,62	1,34	0,14	25,71	25,43
	28	12	Asn	2,23	1,09	1,18	20,77 ↑	4,44	59,96	0,12	16,76	16,51	0,12	0,94	0,71
			Tyr	2,74 ↑	0,13	50,65	0,04	0,55	0,47	0,04	2,06	1,98	0,04	0,12	0,05
		30	Asn	22,87 ↑	5,12	61,51	0,12	0,93	0,70	0,12	1,02	0,79	0,12	16,17	15,92

Подсемейство	Позиция аминокислоты в HTH-домене	Позиция нуклеотида в сайте	Аминокислота	Нуклеотид											
				A			T			G			C		
				Наблюдаемое	Ожидаемое	χ^2	Наблюдаемое	Ожидаемое	χ^2	Наблюдаемое	Ожидаемое	χ^2	Наблюдаемое	Ожидаемое	χ^2
YtrA	37	16	Ser	31,81 ↑	5,14	138,51	1,22	25,00	22,62	0,17	1,12	0,81	0,39	2,34	1,62
		17	Ser	0,17	26,15	25,82	33,09 ↑	5,58	135,50	0,17	1,19	0,88	0,17	0,66	0,37
		25’	Ser	30,36 ↑	5,18	122,26	0,17	26,05	25,72	0,17	0,86	0,56	2,90	1,50	1,31
		26’	Ser	3,95	26,12	18,81	29,31 ↑	4,72	127,96	0,17	1,22	0,91	0,17	1,53	1,22
	39	12	Ala	4,97	1,94	4,71	34,48 ↑	7,90	89,36	0,16	29,81	29,49	1,72	1,68	0,00
		13	Ala	27,06 ↑	6,14	71,31	0,16	29,97	29,65	5,69	1,66	9,76	8,43	3,57	6,63
		16	Ile	29,05 ↑	5,10	112,30	3,84	24,84	17,75	0,13	1,11	0,86	0,35	2,32	1,67
		17	Ile	2,33	25,99	21,54	30,32 ↑	5,55	110,65	0,59	1,18	0,30	0,13	0,66	0,42
		24’	Ala	0,16	11,08	10,76	0,16	12,39	12,07	33,87 ↑	9,91	57,93	7,14	7,95	0,08
		25’	His	1,50	1,03	0,22	1,19	5,15	3,05	3,89 ↑	0,17	81,49	0,06	0,30	0,19
			Ile	30,32 ↑	5,15	123,03	2,79	25,88	20,61	0,13	0,85	0,61	0,13	1,49	1,24
		26’	His	1,19	5,16	3,06	0,06	0,93	0,82	0,06	0,24	0,14	5,34 ↑	0,30	83,61
			Ile	3,84	25,95	18,84	29,27 ↑	4,69	128,74	0,13	1,21	0,96	0,13	1,52	1,27
	44	13	Gly	29,21 ↑	6,43	80,67	3,90	31,41	24,10	4,73	1,74	5,14	5,48	3,74	0,81

Стрелками показаны статистически значимые изменения вероятностей для скоррелированных пар аминокислот и нуклеотидов

Кроме того, результаты анализа корреляций показывают, что предсказанные ДНК-белковые взаимодействия для всех трех исследованных подсемейств регуляторов семейства GntR хорошо соотносятся с ДНК-белковыми контактами, известными для FadR *E. coli* и AraR *B. subtilis* [221,222,223].

Таким образом, несмотря на некоторые неоднозначные результаты (к примеру, корреляции Ser-A/T и Ser-G/C), большая часть контактов, предсказанных в результате анализа корреляций аминокислот HTH-доменов транскрипционных факторов семейства GntR и нуклеотидов соответствующих сайтов связывания (Arg-G, Asn-A, Asp-C, Gly-G, His-G, Trp-C), согласуется с ранее описанными закономерностями взаимодействия [195,196].

3.3 Дивергоны семейства GntR

Многие гены, регулируемые транскрипционными факторами семейства GntR, организованы в дивергентно транскрибируемые опероны (дивергоны). В настоящей работе были исследованы дивергоны подсемейств FadR и HutC. Подсемейство YtrA не представлено в этой части анализа в связи с практически полным отсутствием в данной группе дивергентно регулируемых оперонов.

Все найденные дивергоны были разделены на две группы: дивергоны, имеющие в составе ген транскрипционного фактора, и дивергоны, полностью состоящие из структурных генов (контрольная группа). Кроме того, дивергоны были разделены по числу сайтов связывания в межгенной области: были отдельно проанализированы дивергоны с одним и двумя сайтами.

Для дивергонов с единичным сайтом связывания были оценены: длина межгенной области и расстояние от центра сайта до старта каждого из дивергентных генов.

Для дивергонов с двойными сайтами были исследованы: длина межгенной области, расстояние от центра проксимального сайта до старта соответствующего гена, а также расстояние между центрами сайтов связывания.

В случае дивергонов с единичным сайтом в межгенной области целью было выяснить, принадлежит ли этот сайт обоим оперонам дивергона, или же только одному из них (например, такое может наблюдаться для дивергентно расположенных оперонов, один из которых содержит регулируемые структурные гены, а второй – неавторегулируемый ген транскрипционного фактора).

Для ряда транскрипционных факторов, в частности AraR, известно кооперативное связывание с несколькими близлежащими сайтами, что дает возможность более гибкой и точной регуляции экспрессии генов [106,223]. Таким образом, в случае дивергонов с двойными сайтами целью было определить, участвует ли пара сайтов в совместной регуляции обоих

оперонов дивергона (или же одного из них), представляя собой, по сути, единичный комплексный сайт, или же каждый из двух сайтов независимо регулирует свой оперон.

3.3.1 Дивергоны с единичным сайтом связывания

Полученные результаты показывают, что в случае дивергонов с геном транскрипционного фактора в составе как FadR (n = 96), так и HutC-подсемейств (n = 94) для обоих оперонов дивергона наблюдается общая тенденция к линейному росту расстояния от старта гена до середины сайта с увеличением размера межгенной области (Рисунок 6А, 6Б). Аналогичная тенденция наблюдается также и для дивергонов контрольной группы (FadR, n = 33; HutC, n = 23) (Рисунок 6В; в связи с полным совпадением отображена только одна линия регрессии).

А. Опероны с геном фактора транскрипции

Б. Структурные опероны

В. Контроль

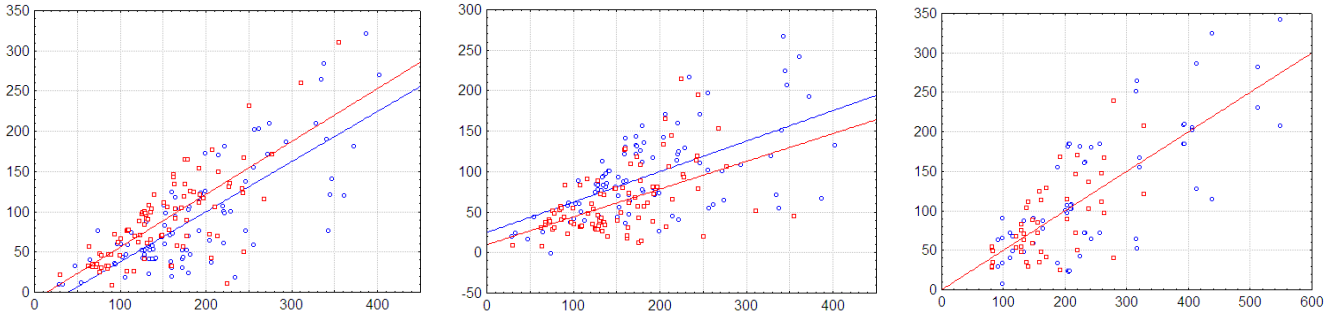


Рисунок 6. Зависимость расстояния между стартом гена и сайтом связывания от размера межгенной области в дивергонах с единичным сайтом

Межгенное расстояние показано на оси X. Расстояние между центром сайта и старт-кодоном гена – на оси Y. Показаны линии регрессии. Синим цветом обозначены данные для подсемейства FadR, красным – для подсемейства HutC.

Таким образом, единичные сайты, как правило, локализуются в центре межгенной области, и, вероятно, участвуют в регуляции обоих оперонов дивергона. Интересно отметить, что в случае дивергонов, содержащих ген транскрипционного фактора, с увеличением размера межгенной области расстояние между сайтом связывания и структурным опероном возрастает медленнее, чем расстояние между сайтом и опероном с геном регулятора в составе (Таблица 4, Рисунок 6А, 6Б). Это может быть следствием того, что авторегуляция генов транскрипционных факторов выражена слабее, чем регуляция соответствующих структурных генов.

Таблица 4. Зависимость расстояния между стартом гена и сайтом связывания от размера межгенной области в дивергонах с единичным сайтом

Коэффициент линейной регрессии (R^2)			
	Оба подсемейства	FadR	HutC
Опероны с геном фактора транскрипции	0,60 (0,55)	0,62 (0,60)	0,66 (0,56)
Структурные опероны	0,40 (0,35)	0,38 (0,35)	0,34 (0,26)
Контроль	0,50 (0,58)	0,50 (0,58)	0,50 (0,42)
Коэффициент корреляции Пирсона (p-value)			
	Оба подсемейства	FadR	HutC
Опероны с геном фактора транскрипции	0,74 ($p<1\cdot10^{-7}$)	0,77 ($p<1\cdot10^{-7}$)	0,75 ($p<1\cdot10^{-7}$)
Структурные опероны	0,59 ($p<1\cdot10^{-7}$)	0,59 ($p<1\cdot10^{-7}$)	0,51 ($p=2\cdot10^{-7}$)
Контроль	0,76 ($p<1\cdot10^{-7}$)	0,76 ($p<1\cdot10^{-7}$)	0,65 ($p=1\cdot10^{-6}$)

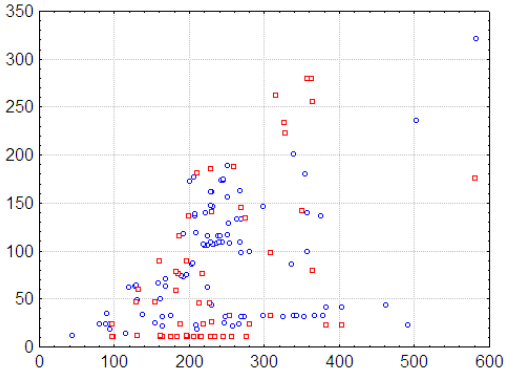
3.3.2 Дивергоны с двойными сайтами связывания

Как было отмечено ранее, в случае, если в дивергоне наблюдаются двойные сайты, существует три возможных варианта регуляции: либо каждый из сайтов принадлежит только одному из составляющих дивергон оперонов, либо пара сайтов совместно регулирует оба или один из оперонов дивергона. Логично предположить, что, если каждый из сайтов участвует в регуляции транскрипции лишь одного из оперонов, расстояние между сайтами будет положительно коррелировать с размерами межгенной области дивергона, так как каждый из сайтов будет располагаться ближе к регулируемому оперону.

Напротив, в случае кооперативной регуляции одного или обоих оперонов дивергона общей парой сайтов, расстояние между сайтами не будет зависеть от межгенного расстояния и, вероятно, будет приблизительно постоянным. При этом если общая пара сайтов участвует в регуляции транскрипции обоих дивергентных оперонов, вероятнее всего, данные сайты связывания будут располагаться в центральной части межгенной области дивергона, аналогично ситуации с единичным сайтом в дивергоне (см. выше). В противном случае, вероятно, что сдвоенные сайты будут располагаться в непосредственной близости от регулируемого оперона.

Полученные нами данные показывают, что в обоих подсемействах FadR (n = 100) и HutC (n = 60) наблюдаются две группы дивергонов, содержащих в составе ген транскрипционного фактора (Таблица 5, Рисунок 7А).

А. Дивергоны с геном фактора транскрипции



Б. Контроль

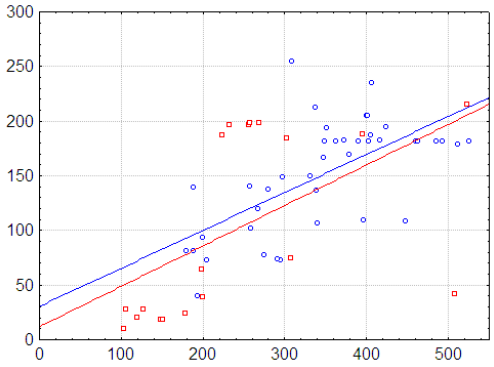
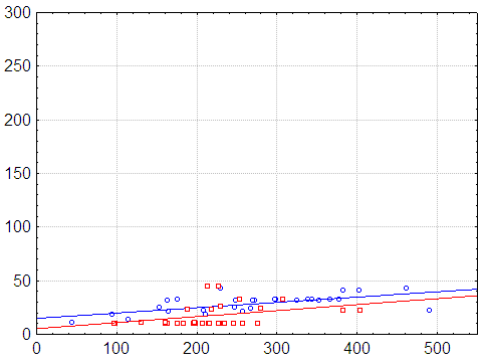


Рисунок 7. Расстояние между двойными сайтами связывания в дивергонах

Межгенное расстояние показано на оси X. Расстояние между центрами сайтов – на оси Y. Обозначения как на Рисунке 6.

Первая группа включает дивергоны (FADR, n = 29; HUTC, n = 32), в которых расстояние между двойными сайтами относительно неизменно вне зависимости от размеров дивергона (Рисунок 8А). В этой группе расстояние от старта гена до проксимального сайта возрастает с увеличением размера межгенной области как для структурных оперонов, так и для оперонов с геном транскрипционного фактора в составе (Таблица 6, Рисунок 9А, 9Б). Таким образом, двойные сайты в данной группе дивергонов обычно локализованы в центре межгенной области, и можно предполагать, что они образуют пару, кооперативно регулирующую транскрипцию обоих дивергентных оперонов.

А. Первая группа (постоянное расстояние между сайтами)



Б. Вторая группа (возрастающее расстояние между сайтами)

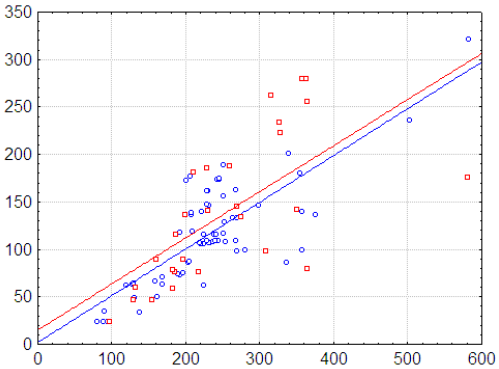


Рисунок 8. Две группы дивергонов с двойными сайтами

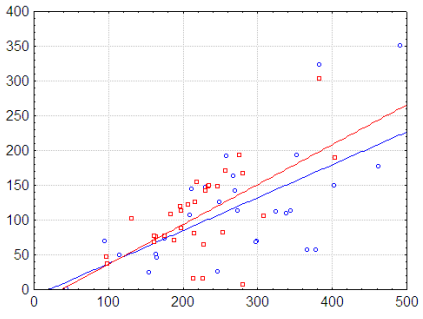
Межгенное расстояние показано на оси X. Расстояние между центрами сайтов – на оси Y. Обозначения как на Рисунке 6.

Таблица 5. Две группы дивергонов с двойными сайтами, зависимость расстояния между сайтами связывания от размера межгенной области

Коэффициент линейной регрессии (R^2)			
	Оба подсемейства	FADR	HUTC
Дивергоны с общей парой сайтов (постоянное расстояние между сайтами)	0,06 (0,26)	0,05 (0,41)	0,06 (0,11)
Дивергоны с независимыми сайтами (возрастающее расстояние между сайтами)	0,50 (0,53)	0,49 (0,61)	0,49 (0,43)
Коэффициент корреляции Пирсона (p-value)			
	Оба подсемейства	FADR	HUTC
Дивергоны с общей парой сайтов (постоянное расстояние между сайтами)	0,51 ($p=3\cdot10^{-5}$)	0,64 ($p=2\cdot10^{-4}$)	0,32 (p=0,07)
Дивергоны с независимыми сайтами (возрастающее расстояние между сайтами)	0,73 ($p<1\cdot10^{-7}$)	0,78 ($p<1\cdot10^{-7}$)	0,65 ($p=2\cdot10^{-4}$)

Вторая группа включает дивергоны (FADR, n = 71; HUTC, n = 28), где расстояние между парой сайтов линейно возрастает с увеличением размеров межгенной области дивергона (Рисунок 8Б). Для обоих оперонов дивергона наблюдается также тенденция к увеличению расстояния от старта гена до проксимального сайта с ростом межгенного расстояния, но эта зависимость существенно менее выражена, чем в случае дивергонов первой группы (Таблица 7, Рисунок 10А, 10Б). Следовательно, в данном случае мы не наблюдаем кооперативной регуляции, эти сайты предположительно независимо регулируют каждую из частей дивергона.

А. Опероны с геном фактора транскрипции



Б. Структурные опероны

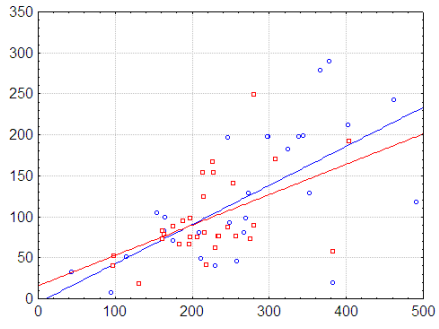


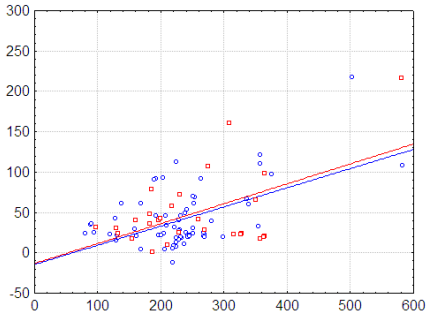
Рисунок 9. Зависимость расстояния между стартом гена и проксимальным сайтом связывания от размера межгенной области в дивергонах с общими двойными сайтами

Межгенное расстояние показано на оси X. Расстояние между центром проксимального сайта и старт-кодоном гена – на оси Y. Обозначения как на Рисунке 6.

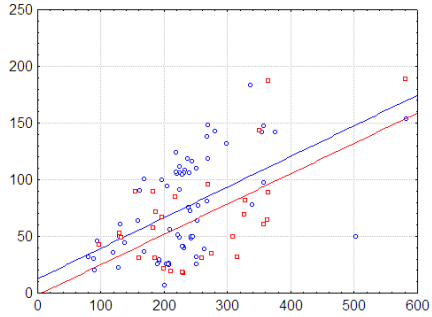
Таблица 6. Зависимость расстояния между стартом гена и проксимальным сайтом связывания от размера межгенной области в дивергонах с общими двойными сайтами

Коэффициент линейной регрессии (R^2)			
	Оба подсемейства	FadR	HutC
Опероны с геном фактора транскрипции	0,48 (0,38)	0,47 (0,40)	0,57 (0,37)
Структурные опероны	0,45 (0,38)	0,48 (0,42)	0,37 (0,24)
Коэффициент корреляции Пирсона (p-value)			
	Оба подсемейства	FadR	HutC
Опероны с геном фактора транскрипции	0,62 ($p=1\cdot10^{-7}$)	0,63 ($p=2\cdot10^{-4}$)	0,61 ($p=2\cdot10^{-4}$)
Структурные опероны	0,62 ($p=1\cdot10^{-7}$)	0,64 ($p=2\cdot10^{-4}$)	0,49 ($p=4\cdot10^{-3}$)

А. Опероны с геном фактора транскрипции



Б. Структурные опероны



В. Контроль

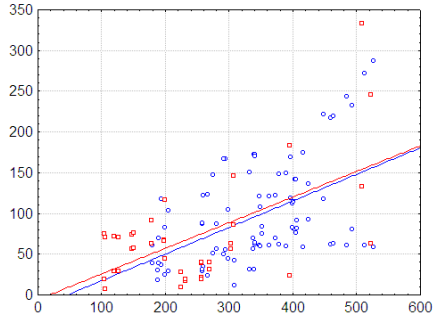


Рисунок 10. Зависимость расстояния между стартом гена и проксимальным сайтом связывания от размера межгенной области в дивергонах с независимыми двойными сайтами

Межгенное расстояние показано на оси X. Расстояние между центром проксимального сайта и старт-кодоном гена – на оси Y. Обозначения как на Рисунке 6.

Таблица 7. Зависимость расстояния между стартом гена и проксимальным сайтом связывания от размера межгенной области в дивергонах с независимыми двойными сайтами

Коэффициент линейной регрессии (R^2)			
	Оба подсемейства	FadR	HutC
Опероны с геном фактора транскрипции	0,24 (0,29)	0,24 (0,29)	0,25 (0,29)
Структурные опероны	0,26 (0,29)	0,27 (0,27)	0,27 (0,37)
Контроль	0,31 (0,31)	0,33 (0,25)	0,31 (0,34)
Коэффициент корреляции Пирсона (p-value)			
	Оба подсемейства	FadR	HutC
Опероны с геном фактора транскрипции	0,54 ($p=7\cdot10^{-9}$)	0,54 ($p=2\cdot10^{-6}$)	0,54 ($p=3\cdot10^{-3}$)
Структурные опероны	0,54 ($p=1\cdot10^{-8}$)	0,52 ($p=3\cdot10^{-6}$)	0,61 ($p=6\cdot10^{-4}$)
Контроль	0,56 ($p<1\cdot10^{-7}$)	0,50 ($p=3\cdot10^{-7}$)	0,58 ($p=1\cdot10^{-4}$)

Аналогичная тенденция была отмечена и для контрольной группы дивергонов (FadR, n = 46; HutC, n = 19) в обоих подсемействах (Таблица 7, Рисунок 7Б, Рисунок 10В). В контроле наблюдается только один тип дивергонов, где сайты независимы и принадлежат ближайшему из оперонов дивергона, осуществляя регуляцию его транскрипции.

Таким образом, двойные сайты в дивергонах могут быть классифицированы на кооперативные и оперон-специфические, однако функциональное отличие дивергонов обоих типов не было определено, так как между ними не наблюдается существенного различия генного состава.

3.4 Дополнительные полусайты мотивов связывания транскрипционных факторов семейства GntR

Типичный мотив связывания регуляторов семейства GntR представляет собой палиндром, однако было показано, что для заметного числа идентифицированных сайтов связывания характерно наличие более слабого дополнительного полусайта (бокса) в непосредственной близости. Для всех исследованных в настоящей работе транскрипционных факторов семейства GntR был проведен анализ областей, фланкирующих предсказанные сайты связывания. В 23 ортологических группах (13 групп, 170 транскрипционных факторов и 450 сайтов связывания в подсемействе FadR; 4 группы, 186 транскрипционных факторов и 514 сайтов в подсемействе HutC, а также 6 групп, 120 транскрипционных факторов и 167 сайтов в подсемействе YtrA) было отмечено наличие слабых дополнительных боксов, расположенных с одной или с обеих сторон от сильного палиндромного сайта, на расстоянии 7-12 нуклеотидов (нт) от центра сайта. Дополнительные боксы были идентифицированы при помощи визуального анализа диаграмм Logo, построенных для каждой из ортологических групп на основе всех выровненных сайтов связывания и их ближайшего окружения.

С целью оценить статистическую значимость этого наблюдения, было проведено сравнение дополнительных боксов с истинными сайтами, а также, в качестве контроля, со случайными последовательностями соответствующей длины (псевдобоксы). Для корректного статистического анализа (см. далее) каждый из сайтов сравнивался с двумя псевдобоксами в 5’-регулируемой области анализируемых генов, на расстоянии –20 и –21 нт от начала истинного сайта, соответственно (Рисунок 11).

Вес для каждой половины истинного палиндромного сайта был рассчитан при помощи соответствующих половин матриц позиционных весов нуклеотидов (PWM) для данного транскрипционного фактора (полученные значения – $W_{true\ left}$ и $W_{true\ right}$, соответственно). Соответствующие PWM были также использованы для расчета весов дополнительных боксов

($W_{near\ left}$ и $W_{near\ right}$, соответственно, слева и справа) и псевдобоксов, при этом вес для каждого из псевдобоксов был рассчитан дважды, с помощью каждой из половин PWM ($W_{random\ left1,2}$ и $W_{random\ right1,2}$).

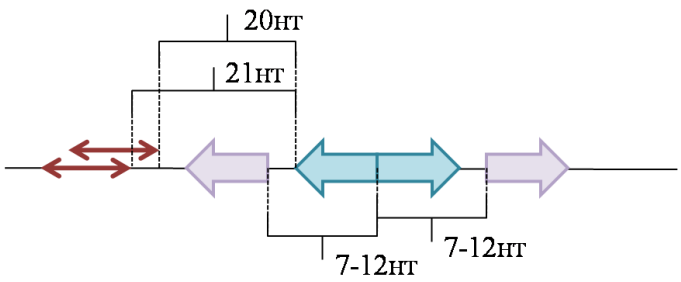


Рисунок 11. Схема расположения дополнительных боксов и контрольных псевдобоксов

Истинные полусайты показаны голубым, дополнительные боксы – фиолетовым, псевдобоксы (контроль) – красным.

После расчета весов осуществлялся отбор данных: из каждой пары дополнительных боксов для дальнейшего анализа выбирался один с наибольшим весом ($W_{near\ left}$ или $W_{near\ right}$). Каждый дополнительный бокс сравнивался с одним из контрольных псевдобоксов: из двух псевдобоксов, вес которых был рассчитан по той же части PWM, что и вес выбранного дополнительного бокса ($W_{random\ left1,2}$ или $W_{random\ right1,2}$), выбирали один с наибольшим весом (например, $W_{random\ left1}$ или $W_{random\ left2}$).

Мотивы связывания транскрипционных факторов из разных ортологических групп различались по длине и структуре, следовательно, различались также и соответствующие PWM. Поэтому для сравнения данных по каждому из подсемейств в целом рассчитанные веса дополнительных боксов и псевдобоксов были нормированы на значения весов соответствующих истинных полусайтов ($W_{true\ left}$ или $W_{true\ right}$), ориентированных так же, как и выбранный для анализа дополнительный бокс.

Данные для каждого из подсемейств GntR были нормированы по следующим формулам:

$$S_{near} = \frac{W_{true} - W_{near}}{W_{true}} \quad (1); \qquad S_{random} = \frac{W_{true} - W_{random}}{W_{true}} \quad (2)$$

– где S_{near} и S_{random} – нормированные веса для дополнительных боксов и псевдобоксов, соответственно; W_{true}, W_{near}, W_{random} – рассчитанные по PWM веса выбранных для анализа истинных полусайтов, дополнительных боксов и псевдобоксов, соответственно.

Распределение значений S_{near} и S_{random} (Рисунок 12) статистически значимо различалось для всех трех подсемейств FadR, HutC и YtrA (парный критерий Вилкоксона, $p < 0,001$). При этом среднее значение веса для дополнительных боксов W_{near} приблизительно соответствует половине веса истинного полусайта W_{true}, тогда как среднее значение W_{random} близко к нулю, что свидетельствует о правильности выбранного контроля.

Возможное участие данных дополнительных боксов в процессе регуляции транскрипции представляется интересным объектом дальнейших экспериментальных исследований. Эти дополнительные боксы, предположительно, могут участвовать в альтернативной димеризации транскрипционных факторов или же в связывании дополнительных субъединиц, позволяя кооперативную регуляцию и более тонкий контроль экспрессии.

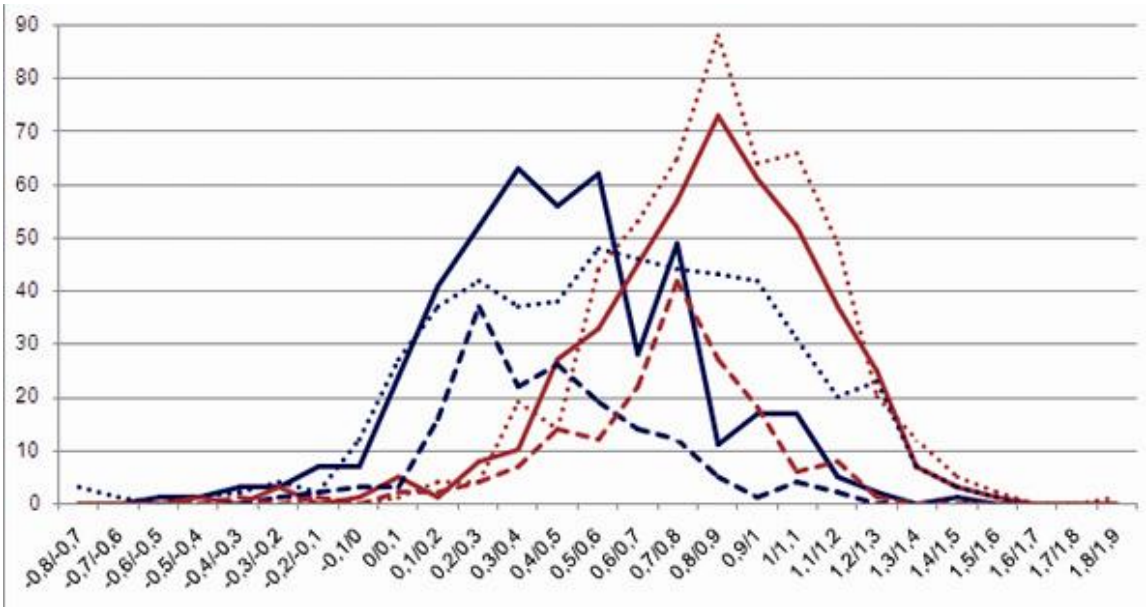


Рисунок 12. Распределение значений S_{near} и S_{random} для подсемейств FadR, HutC и YtrA

Интервалы значений S показаны на оси X. Количество значений S в соответствующем интервале – на оси Y. Синим цветом показаны распределения значений S_{near}, красным – S_{random}. Данные для подсемейства FadR обозначены непрерывными линиями, для подсемейства HutC– точечным пунктиром, для подсемейства YtrA – пунктирными линиями.

3.5 Заключение

В настоящей работе были исследованы 1252 транскрипционных фактора из 64 ортологических групп трех подсемейств семейства GntR (FadR, HutC, YtrA), для которых идентифицированы мотивы связывания и реконструированы соответствующие регулоны. Для каждого из подсемейств исследована коэволюция мотивов связывания ДНК и аминокислотных последовательностей регуляторов транскрипции путем анализа корреляций аминокислот HTH-доменов транскрипционных факторов и нуклеотидов соответствующих сайтов связывания, предсказаны вероятные ДНК-белковые контакты. Показано, что, несмотря на отличия структуры транскрипционных факторов разных подсемейств и соответствующих мотивов связывания, большая часть предсказанных ДНК-белковых контактов (Arg-G, Asn-A, Asp-C, Gly-G, His-G, Trp-C) сходны для всех трех подсемейств и хорошо соотносятся с данными, известными для комплексов FadR *E. coli* и AraR *B. subtilis* с ДНК, а также с общими закономерностями ДНК-белковых взаимодействий.

Также был проведен анализ структуры дивергонов, регулируемых транскрипционными факторами подсемейств FadR и HutC, выявлены основные тенденции расположения сайтов. В дивергонах с единичным сайтом, тот, как правило, располагается приблизительно в центре межгенной области и принадлежит обоим оперонам дивергона. При этом в дивергонах с геном транскрипционного фактора с увеличением размера межгенной области расстояние между сайтом связывания и структурным опероном возрастает медленнее, нежели расстояние между сайтом и опероном с геном регулятора. Это может отражать разницу в силе и устойчивости авторегуляции генов транскрипционных факторов и регуляции экспрессии структурных генов. В случае дивергонов с двойными сайтами предсказано два варианта регуляции: общая пара сайтов участвует в кооперативной регуляции обоих оперонов и располагается в центре межгенной области, или же каждый сайт из пары независимо регулирует собственный оперон и располагается рядом с ним. Таким образом, двойные сайты могут быть классифицированы на кооперативные и оперон-специфичные.

Для подсемейств FadR, HutC и YtrA семейства GntR проанализирована структура сайтов связывания и их ближайшее окружение, для 23 ортологических групп показано наличие более слабых дополнительных боксов на расстоянии 7-12 нуклеотидов от основного палиндромного сайта связывания, статистически значимо отличающихся от случайных последовательностей. Эти дополнительные боксы, предположительно, могут участвовать в альтернативной димеризации транскрипционных факторов или же в связывании дополнительных субъединиц.

Глава 4. Сравнительно-геномный анализ метаболизма гексуронатов у Gammaproteobacteria

4.1 Реконструкция регулонов UxuR и ExuR и эволюция метаболизма гексуронатов у Gammaproteobacteria

4.1.1 Таксономическое распределение и эволюция транскрипционных факторов UxuR и ExuR

Ортологи ExuR и UxuR были найдены только у представителей Gammaproteobacteria. При этом регулятор ExuR присутствует исключительно у представителей Enterobacteriales (а также *Photobacterium profundum* среди Vibrionales), тогда как регулятор UxuR характерен для целого ряда порядков Gammaproteobacteria (Alteromonadales, Enterobacteriales, Oceanospirillales, Pasteurellales и Vibrionales).

Филогенетическое дерево ортологов этих транскрипционных факторов (Рисунок 13) показывает, что дупликация предкового варианта регулятора на ветви UxuR и ExuR произошла до таксон-специфического разделения факторов UxuR. Таким образом, одним из возможных эволюционных сценариев является наличие у предковых форм регулонов с обоими транскрипционными факторами, с последующей потерей ExuR во всех группах, кроме Enterobacteriales и *P. profundum.* Альтернативный вариант предполагает дупликацию исходного регуляторного гена у общего предка Enterobacteriales, с последующей быстрой эволюцией ExuR и горизонтальным переносом этого регулятора к *P. profundum.* Оба эволюционных сценария встречают ряд сложностей: в первом случае предполагаются множественные независимые потери гена *exuR*, тогда как второй вариант противоречит позиции корня филогенетического дерева и требует необычно высокой скорости эволюции.

4.1.2 Идентификация мотивов связывания UxuR и ExuR

Специфические профили для UxuR и ExuR были построены по сайтам связывания перед известными членами соответствующих регулонов *E. coli* и их ортологами в геномах родственных Gammaproteobacteria. Обучающая выборка для профиля UxuR включала сайты перед *uxuR, uxuAB, gntP* и *uidABC E. coli* и их ортологами в *uxuR*-содержащих геномах (*Actinobacillus succinogenes* 130Z, *Haemophilus influenzae* 86-028NP, *Haemophilus somnus* 129PT, *Mannheimia succiniciproducens* MBEL55E, *Photobacterium profundum* SS9, *Photorhabdus luminescens* TTO1, *Salmonella enterica* Typhimurium LT2, *Vibrio parahaemolyticus* RIMD 2210633, *Vibrio vulnificus* CMCP6). Обучающая выборка для профиля ExuR включала сайты перед *exuR*,

exuT, *uxaCA* и *uxaB E. coli* и их ортологами в *exuR*-содержащих геномах (*Citrobacter koseri* ATCC BAA-895, *Enterobacter* sp. 638, *Klebsiella pneumoniae* MGH 78578, *Pectobacterium atrosepticum* SCRI1043, *Serratia proteamaculans* 568, *Yersinia pestis* KIM).

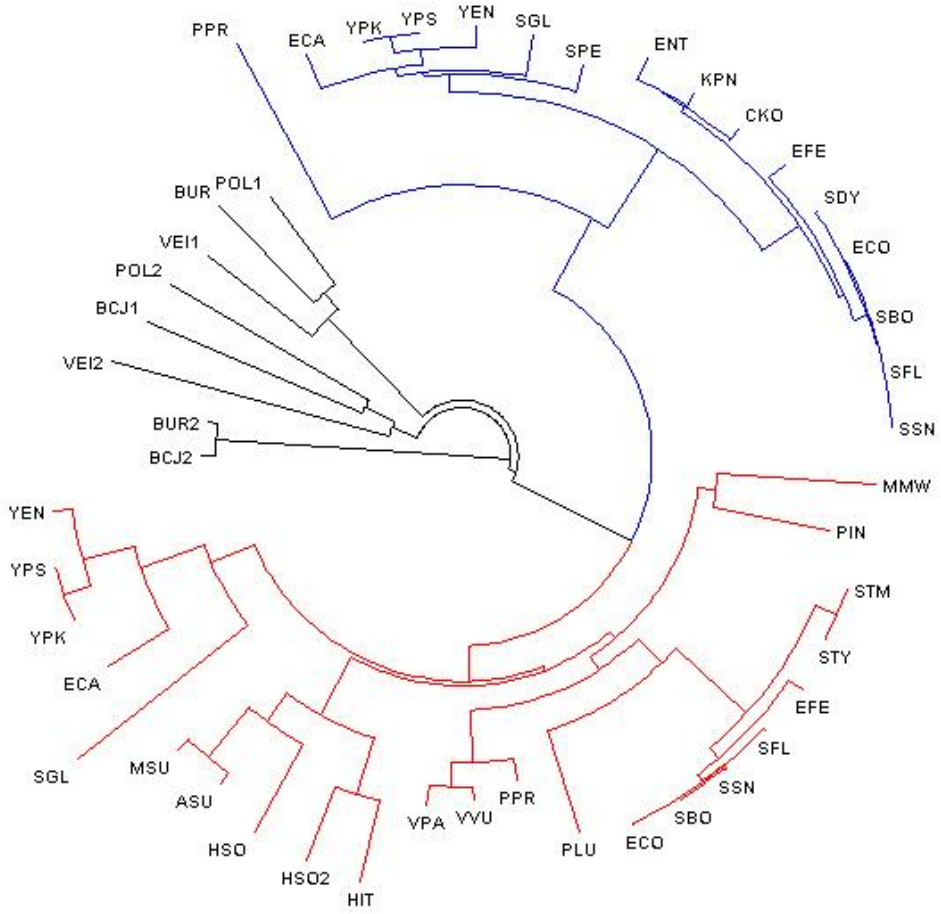


Рисунок 13. Филогенетическое дерево транскрипционных факторов UxuR и ExuR у Gammaproteobacteria

Красным отмечены ортологи UxuR; синим – ExuR; внешняя группа показана черным. Трехбуквенные обозначения геномов соответствуют аббревиатурам, приведенным в Приложении А.

Несмотря на то, что в ранее опубликованных работах делалось заключение о том, что сайты связывания как UxuR, так и ExuR соответствуют одному консенсусу (AAATTGGTATACCAATTT) и слишком близки, чтобы их различить [96], нами было показано, что, используя большую и более разнообразную выборку сайтов из различных геномов и относя их к определенному транскрипционному фактору в соответствии с функцией регулируемого гена, можно различить мотивы связывания для UxuR и ExuR. Эти два мотива (Рисунок 14)

отличаются структурой 3’-участка и длиной центрального спейсера между нуклеотидными группами GT/AC двух половин палиндромного сайта (2 и 3 нуклеотида, соответственно). Мотив связывания ExuR не точно палиндромный в позициях 6/14, 7/13 и 9/11. Подобное строение дает возможность кросс-узнавания одной и той же области обоими регуляторами ExuR и UxuR: последовательность RAYAA правого плеча палиндромного сайта может комплементарно соответствовать (с учетом одной замены) последовательности TGGT левого плеча как в 18-нуклеотидном сайте UxuR, так и в 19-нуклеотидном сайте связывания ExuR (см. пунктирные линии на Рисунке 14). Таким образом, в результате дупликации и последующей эволюции ExuR, вероятно, приобрел новую специфичность взаимодействия с ДНК.

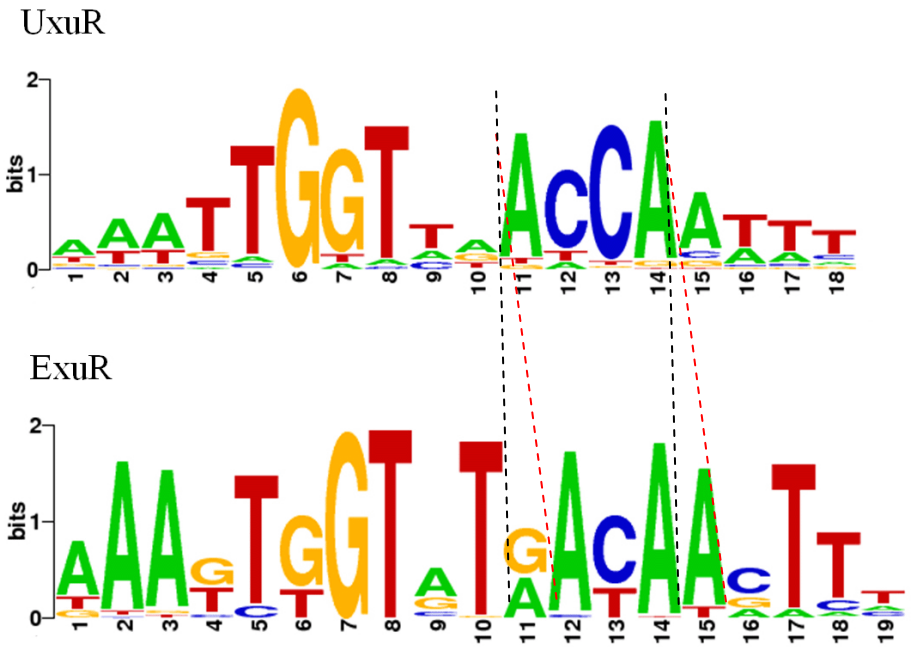


Рисунок 14. Диаграмма Logo мотивов связывания UxuR и ExuR

По горизонтальной оси отмечена позиция нуклеотида в сайте; по вертикальной оси – информационное содержание в битах. Общая высота символов в каждой позиции соответствует информационному содержанию; высота конкретного символа отражает частоту его встречаемости в данной позиции.

4.1.3 Строение гексуронатных регулонов

У большинства исследованных представителей Enterobacteriales в геноме присутствуют как *uxu*, так и *uxa* гены, таким образом, они способны метаболизировать и D-глюкуронат, и D-галактуронат. При этом у *Escherichia* spp., *P. atrosepticum*, *Shigella* spp. (за исключением *S. dysenteriae*, у которой отсутствуют гены утилизации глюкуроната *uxuABR* и, следовательно, регулятор UxuR), а также *Yersinia* spp. обнаружены оба транскрипционных фактора, UxuR и

ExuR, контролирующие метаболические пути утилизации D-глюкуроната и D-галактуроната, соответственно. В то же время, *C. koseri*, *Enterobacter* sp. 638, *K. pneumoniae* и *S. proteamaculans* имеют только транскрипционный фактор ExuR, таким образом, регулирующий оба гексуронатных катаболических пути (Таблица 8).

Кроме того, в этой таксономической группе *Photorhabdus luminescens* и *Salmonella* spp. имеют только путь утилизации D-глюкуроната, находящийся под негативным контролем UxuR. Интересно отметить, что структура регулона UxuR у вышеупомянутых бактерий существенно отличается от организации данного регулона у остальных исследованных Enterobacteriales: у *P. luminescens* гексуронатный регулон включает гены *uxuAB, uxuPQM, uxuR, uxaC, kdgK* и *eda* (аналогичный состав регулон UxuR имеет у Vibrionales, см. далее), а у *Salmonella* spp. – гены *uxuAB, uxuR, uxaC* и *exuT* (при этом ген регулятора UxuR располагается в отдельном локусе) (Таблица 8).

Для большинства исследованных представителей Enterobacteriaceae характерно наличие множества паралогов генов D-маннонат оксидоредуктазы, например, у *E. coli* это *uxuB, yeiQ* и *ydfI*. В 5'-регуляторной области *yeiQ* у *E. coli, Salmonella enterica* Typhimurium LT2 и *Shigella* spp. были идентифицированы предполагаемые сайты связывания UxuR. При этом у *S. dysenteriae*, несмотря на отсутствие UxuR, сильные потенциальные сайты связывания данного регулятора обнаруживаются перед генами *yeiQ* и *uidR* (Таблица 8), что свидетельствует о том, что потеря UxuR произошла недавно.

Следует также отметить, что среди Enterobacteriales гены *uid* находятся под регуляцией UxuR только у *E. coli* и *Shigella* spp.

В отличие от бактерий порядка Enterobacteriales, для прочих представителей Gammaproteobacteria из порядков Alteromonadales, Oceanospirillales, Pasteurellales и Vibrionales характерно наличие только репрессора UxuR (Рисунок 13), как правило, контролирующего метаболизм D-глюкуроната, но не D-галактуроната. Исключением является *P. profundum*, у которого, вероятно, в результате горизонтального переноса генов от представителей Pasteurellales (основываясь на филограмме ферментов UxaA, UxaB и UxaC, Приложение B1, B2, B3), в геноме присутствуют гены галактуронатного метаболизма (*uxaCBA*), а также как UxuR, так и ExuR, осуществляющие регуляцию катаболизма D-глюкуроната и D-галактуроната, соответственно. Характерной чертой гексуронатных регулонов у бактерий этих групп, а также *P. luminescens* среди Enterobacteriales, предположительно, получившего данные гены от общих предков с Vibrionales и Alteromonadales (в соответствии с филограммой субъединиц UxuP, UxuQ, UxuM, Приложение B4, B5, B6), является наличие в их составе генов АТФ-независимых

периплазматических TRAP транспортеров [249,250,251], вероятно, участвующих в транспорте гексуронатов вместо ExuT.

TRAP транспортеры, которые используют электрохимический градиент в качестве движущей силы, присутствуют у многих эубактерий и архей, но не у эукариот [249,250,251]. Некоторые из них содержат только одну подобную транспортную систему (например, *E. coli*), у других имеется несколько TRAP (*Bacillus halodurans*, *Pseudomonas aeruginosa*) [250]. Субстратами TRAP являются такие соединения, как глутамат, глюконат, малат, пируват, сукцинат, таурин, эктоин и т.д. [249,250,251]. Наиболее известные и хорошо охарактеризованные TRAP транспортеры – высокоаффинные C4-дикарбоксилат транспортеры Dct, состоящие из трех компонентов: периплазматической субстрат-связывающей DctP, а также малой DctQ и большой DctM трансмембранных субъединиц [249,250,251]. По аналогии с TRAP C4-дикарбоксилат транспортной системой DctPQM, UxuR-регулируемые TRAP транспортеры были названы нами UxuPQM.

Так как встречаемость генов *uxuPQM* среди Gammaproteobacteria коррелирует с отсутствием в геноме гена транспортера гексуронатов *exuT*, можно с уверенностью предполагать, что данные TRAP транспортеры обеспечивают импорт каких-либо гексуронатов в клетку. У большинства проанализированных в данной работе бактерий (*Haemophilus* spp., *Marinomonas* sp. MWYL1, *P. luminescens*, *Psychromonas ingrahamii* 37, *Vibrio* spp.) в геноме обнаруживается только один кластер *uxuPQM* в составе гексуронатных регулонов. В этом случае гены *uxuPQM* колокализуются с генами пути утилизации глюкуроната, следовательно, эти транспортеры вероятнее всего вовлечены в транспорт D-глюкуроната/D-фруктуроната, нежели участвуют в процессе переноса в клетку D-галактуроната/D-тагатуроната. В то же время у некоторых исследованных нами бактерий (*A. succinogenes*, *M. succiniciproducens*, *P. profundum*) в геноме присутствует несколько копий генов *uxuPQM* в составе гексуронатных регулонов. В вышеупомянутых геномах один из наборов генов *uxuPQM* колокализуется с генами, отвечающими за утилизацию D-галактуроната. Это дает возможность предположить, что некоторые из паралогов TRAP транспортеров в составе гексуронатных регулонов обеспечивают импорт D-галактуроната. Кроме того, было отмечено, что субъединицы этих транспортеров из *A. succinogenes*, *M. succiniciproducens* и *P. profundum* располагаются рядом на филогенетическом дереве (Приложение В4, В5, В6), что также может косвенно подтверждать их специализацию.

Общий интермедиат параллельных путей утилизации D-глюкуроната и D-галактуроната, KDG, далее последовательно подвергается действию киназы KdgK и альдолазы Eda/KdgA [252,253]. У *E. coli* эта часть метаболического пути находится под контролем репрессора KdgR

из семейства ICLR, который также регулирует процесс утилизации пектина у фитопатогенных бактерий порядка Enterobacteriales [96,252,253]. В нашей работе было показано, что гены *kdgK* и *eda* у представителей Pasteurellales, Vibrionales, а также у *P. luminescens* и *P. ingrahamii* колокализованы с генами регулона UxuR и, вероятно, входят в его состав. Все эти бактерии, за исключением *Vibrio* spp., не имеют регулятора KdgR, таким образом, их регулон UxuR расширяется, включая гены дальнейшего метаболического пути гексуронатов. У *Vibrio* spp. идентифицированы две паралогичные копии генов *kdgK* и *eda*, одна из которых находится под контролем UxuR, а вторая – под регуляцией KdgR [253].

У большинства исследованных Enterobacteriales (*Escherichia* spp., *K. pneumoniae, P. atrosepticum, Shigella* spp., *Yersinia* spp.) непосредственно после кластера генов утилизации галактуроната *uxaCBA/uxaCA* располагается ген *ygjV,* кодирующий транспортный белок [253]. Этот ген также колокализован с генами *kdgK* и *eda* в геномах *P. profundum, P. ingrahamii* и *Vibrio* spp. У всех Enterobacteriales за исключением *K. pneumoniae* в 5'-области данного гена обнаруживаются сильные сайты связывания KdgR, следовательно, *ygjV* является членом KdgR-регулона [253]. Предполагаемый сайт связывания KdgR локализуется между генами *uxaA* и *ygjV*, непосредственно сразу после предсказанной последовательности rho-независимого терминатора транскрипции оперона *uxa* [253]. Как и в случае генов *kdgK* и *eda,* отмечено наличие нескольких паралогов *ygjV* в геномах *Vibrio* spp., имеющих регулятор KdgR: одна из копий расположена в составе оперона *uxuQM-uxuB-uxaC-kdgK-eda-ygjV* (регулируемого UxuR и не регулируемого KdgR), в то время как второй паралог входит в состав оперона *kduD-ygjV-kdgF-spiX*, которому предшествуют два сайта связывания KdgR [253]. Подобная дупликация *kdgK, eda* и *ygjV* может отражать недавнюю специализацию данных паралогов в сторону катаболизма пектина (под регуляцией KdgR) или же гексуронатов. В последнем случае *ygjV,* вероятно, участвует в транспорте каких-либо интермедиатов гексуронатного катаболизма и входит в состав регулона UxuR, о чем свидетельствует колокализация с соответствующими метаболическими генами [253].

У *A. succinogenes* обнаружено два гексуронатных локуса, один из которых содержит *uxaABC, uxuA* и *uxuR*, две паралогичных копии *uxuPQM, kdgK* и *eda,* а также гены *lfaA, yjjM* и *yjjN* (см. далее). Второй кластер генов утилизации D-глюкуроната включает гены *uxuAB, uxuPQM, kdgK* и *eda*, а также гены, кодирующие алкоголь дегидрогеназу неизвестной специфичности и регулятор семейства GNTR, родственный UxuR, однако не являющийся его ортологом (*Asuc_0372,* не имеющий ортологов среди Gammaproteobacteria).

Haemophilus somnus имеет два паралога *uxuR* в геноме, один из которых кластеризован с генами *uxuAB, uxaC, uidB* и *eda*, а второй, соответственно, с *uxuPQM* и *gntP*.

Кроме того, у *Salmonella enterica* Typhimurium LT2 был идентифицирован еще один транскрипционный фактор семейства GntR, родственный UxuR и ExuR, однако не являющийся ортологом какого-либо из этих репрессоров и не имеющий ортологов среди исследованных в данной работе Gammaproteobacteria. Ген этого регулятора (*STM3084.S*) располагается в дивергоне с генами, кодирующими D-маннонат оксидоредуктазу *yeiQ*, L-галактонат оксидоредуктазу *yjjN*, дегидрогеназу неизвестной специфичности, а также маннитол дегидрогеназу.

Кластер генов *yjj* кодирует транспортер L-галактоната (YjjL), L-галактонат оксидоредуктазу (YjjN), окисляющую субстрат до D-тагатуроната, а также транскрипционный фактор семейства GntR (YjjM), предположительно участвующий в регуляции утилизации L-галактоната [254] (Рисунок 1). Было показано, что мутантные по генам *yjj* штаммы *E. coli* теряют способность расти на среде с L-галактонатом [254]. В исследованных геномах ген *yjjM* находится в дивергоне с *yjjN*, а кроме того, часто колокализован с геном *yjjL*, вероятно, формируя оперон *yjjML*. YjjM предположительно регулирует транскрипцию дивергона *yjjM(L)-yjjN*, однако сравнительный анализ межгенной области не выявил консервативного мотива, который мог бы соответствовать потенциальному сайту связывания YjjM. Ортологи *yjj* генов *E. coli* были идентифицированы у ряда представителей Enterobacteriales (*P. atrosepticum*, *Salmonella enterica* Typhimurium LT2, *S. proteamaculans*, *Shigella* spp.), а также двух бактерий порядка Pasteurellales (*A. succinogenes* и *M. succiniciproducens*) (Таблица 8). Слабые сайты связывания UxuR были предсказаны для дивергона *yjjM(L)-yjjN* у *E. coli*, *Shigella* spp. По результатам наших предсказаний лабораторией О. Озолинь Института биофизики клетки РАН экспериментально подтверждена UxuR-зависимая регуляция транскрипции *yjjM* и *yjjN* у *E. coli*.

Дивергентно транскрибируемые гены *lfaR-lfaTA* кодируют, соответственно, репрессор семейства LacI, транспортер семейства MFS (гомологичный ExuT – идентичность 40-45%, сходство последовательности 60-65%), а также α-глюкозидазу семейства GH31 гликозил-гидролаз [255]. Известно, что этот кластер генов участвует в проявлении фитопатогенных свойств *Erwinia chrysanthemi* 3937, и, вероятно, требуется для утилизации каких-либо углеводов растительного происхождения [255]. Гены *lfa* были идентифицированы и в прочих исследованных в настоящей работе патогенных и условно-патогенных Enterobacteriales, таких как *C. koseri*, *Enterobacter* sp. 638, *K. pneumonia*, *P. atrosepticum*, *S. proteamaculans*, *Shigella sonnei* и *Yersinia* spp., у которых этот кластер генов всегда находится под регуляцией UxuR или ExuR (Таблица 8).

Таблица 8. Состав гексуронатных регулонов у Gammaproteobacteria

	uxuR	uxuA	uxuB	gntP	**exuR**	uxaA	uxaC	exuT	uxaB	**uidR**	uidA	uidB	uidC	uxuPQM	kdgK	eda	yeiQ	ygjV	**lfaR**	lfaA	lfaT	**yjjM**	yjjN	yjjL
Enterobacteriales																								
CKO	0	+	+	0	+	+	+	+	+	0	0	0	0	0	+	+	+	0	+	+	+	0	0	0
ENT	0	+	+	0	+	+	+	+	+	0	0	0	0	0	+	+	+	0	+	+	+	0	0	0
ECA	+	+	+	0	+	+	+	+	+	0	0	0	0	0	+	+	+	+	+	+	+	+	+	+
ECO	+	+	+	+	+	+	+	+	+	+	+	+	+	0	+	+	+	+	0	0	0	+	+	+
EFE	+	+	+	0	+	+	+	+	+	0	0	0	0	0	+	+	+	+	0	0	0	0	0	0
KPN	0	+	+	0	+	+	+	+	+	0	0	0	0	0	+	+	0	+	+	+	+	0	0	+
PLU	+	+	+	0	0	0	+	0	0	0	0	0	0	+	+	+	0	0	0	0	0	0	0	0
STY	+	+	+	0	0	0	+	+	0	0	0	0	0	0	+	0	0	0	0	0	0	0	0	0
STM	+	+	+	0	0	0	+	+	0	0	0	0	0	0	+	+	+	0	0	0	0	0	+	0
SPE	0	+	+	0	+	+	+	+	0	0	0	0	0	0	+	+	0	0	+	+	+	+	+	+
SBO	+	+	+	0	+	+*	+	+	0	+	+*	+	+	0	+	+	+	+	0	0	0	+	+	0
SDY	0	0	0	0	+	+	+*	+	0	+	+*	+	0	0	0	+	+	+	0	0	0	0	0	0
SFL	+	+	+	+	+	+	+	+	+	+	+*	+	+	0	+	+	+	+	0	0	0	+	+*	+
SSN	+	+	+	0	+	+	+	+*	+	+	+*	+	+	0	+	+	+	+	+	+	+	+	+	0
SGL	+	+	+	0	+	0	0	0	0	0	0	0	0	0	+	0	0	+	0	0	0	0	0	0
YEN	+	+	+	+	+	+	+	+	+	0	0	0	0	0	+	+	0	+	+	+	+	0	0	0
YPK	+	+	+	+	+	+	+	+	+	0	0	0	0	0	+	+	+	+	+	+	+	0	0	0
YPS	+	+	+	+	+	+	+	+	+	0	0	0	0	0	+	+	0	+	+	+	+	0	0	0
Другие представители Gammaproteobacteria																								
ASU	+	+	+	0	0	+	+	0	+	0	0	0	0	+	+	+	0	0	0	+	0	+	+	0
HIT	+	+	+	+	0	0	+	0	0	0	0	0	0	+	+	+	0	0	0	0	0	0	0	0
HSO	+	+	+	+	0	0	+	0	0	0	+	+	0	+	0	+	0	0	0	0	0	0	0	0
MSU	+	+	+	+	0	+	+	0	+	0	0	0	0	+	+	+	0	0	0	+	0	+	+	0
MMW	+	+	+	+	0	+	+	0	0	0	+	0	0	+	0	+	0	0	0	+	0	0	0	0
PPR	+	+	+	+	+	+	+	0	+	0	0	0	0	+	+	+	0	+	0	0	0	0	0	0
PIN	+	+	+	0	0	0	+	0	0	0	0	0	0	+	+	+	0	+	0	0	0	0	0	0
VPA	+	+	+	0	0	0	+	0	0	0	0	0	0	+	+	+	0	+	0	0	0	0	0	0
VVU	+	+	+	0	0	0	+	0	0	0	0	0	0	+	+	+	0	+	0	0	0	0	0	0

Обозначения: + – ген присутствует в геноме, 0 – ген отсутствует, * отмечены псевдогены. Цветом показаны гены, расположенные в одном локусе; жирным шрифтом выделены гены транскрипционных факторов. Штриховкой обозначена регуляция: вертикальными полосами показаны гены, регулируемые UxuR, горизонтальной – ExuR, косой – находящиеся под двойной регуляцией. Трехбуквенные обозначения геномов соответствуют аббревиатурам, приведенным в Приложении А.

4.2 Заключение

В настоящей работе был осуществлен детальный анализ регуляции метаболизма гексуронатов у Gammaproteobacteria: построена модель эволюции транскрипционных факторов UxuR и ExuR, предсказаны их мотивы связывания, а также реконструированы соответствующие регулоны.

Показано, что ортологи ExuR присутствуют только у бактерий порядка Enterobacteriales, а также *P. profundum*, тогда как ортологи UxuR найдены у ряда представителей Alteromonadales, Enterobacteriales, Oceanospirillales, Pasteurellales и Vibrionales.

С помощью сравнительного анализа были разделены связывающие мотивы UxuR и ExuR. Мотив связывания UxuR представляет собой палиндром длиной 18 нуклеотидов; родственный и сходный по строению мотив связывания ExuR отличается от него длиной (19 нуклеотидов) и структурой 3'-участка. Предположительно, мотив связывания ExuR представляет собой эволюционный вариант мотива UxuR, и в настоящее время мы наблюдаем ранние стадии расхождения данных транскрипционных факторов.

В целом организация реконструированных регулонов UxuR и ExuR у исследованных Gammaproteobacteria разнообразна и лабильна. Среди представителей Enterobacteriales UxuR преимущественно репрессирует гены *uxuAB*, *uxuR*, а также (реже) *gntP*, *uidR*, *uidABC* и *yeiQ*, тогда как ExuR в основном осуществляет негативный контроль экспрессии *exuR*, *exuT* и *uxaABC*, а также генов *uxuAB* в организмах, где отсутствует регулятор UxuR.

Нами были также идентифицированы новые члены регулонов UxuR/ExuR: TRAP транспортеры гексуронатов (*uxuPQM*), гены углеводного катаболизма (*lfaR-lfaTA*), гены утилизации L-галактоната (*yjjN-yjjML*). Было также показано, что в ряде геномов регулон UxuR расширяется, включая гены дальнейших этапов утилизации гексуронатов (а именно, метаболизма KDG) – *kdgK*, *eda* и *ygjV*. Кроме того, UxuR-зависимая регуляция транскрипции *yjjM* и *yjjN* была экспериментально подтверждена у *E. coli* [256].

Глава 5. Регуляция и эволюция метаболизма малоната и пропионата у Proteobacteria

5.1 Реконструкция регулонов ранее описанных регуляторов метаболизма малоната и пропионата

5.1.1 Транскрипционные факторы MatR/MdcY из подсемейства FadR семейства GntR

MatR и MdcY были экспериментально описаны в разных бактериях (*R. leguminosarum* и *A. calcoaceticus*, соответственно) [233,235,236], что и послужило причиной различного наименования этих факторов транскрипции. Однако аминокислотные последовательности MatR и MdcY имеют высокую степень сходства (49%), кроме того, ранее предсказанные сайты связывания этих транскрипционных факторов идентичны [235,236]. Филогенетический анализ также не выявил явного разделения между регуляторами, гены которых колокализованы с генами *mat* или *mdc*. В связи с этим, в нашей работе этот транскрипционный фактор был обозначен MdcY, по названию, использованному в более ранней публикации.

Ортологи MdcY были обнаружены преимущественно среди Alphaproteobacteria (Rhizobiales, Rhodobacterales, Rhodospirillales, Sphingomonadales), а также у некоторых Beta- (Burkholderiales, Rhodocyclales) и Gammaproteobacteria (Alteromonadales, Pseudomonadales) (Приложение Г1).

Филогенетический футпринтинг 5’-регуляторных областей генов *mat* и *mdc* выявил мотив связывания MdcY с консенсусом TTGTATACAA, что согласуется с полученными ранее данными сравнительного анализа и ДНК-футпринтинга [235,236].

Бактерии различаются по наличию в геноме и организации генов *mdc* и *mat*. Гены *mdc* в составе MdcY-регулонов в большинстве случаев формируют оперон *mdcACDEGBH*, часто в дивергоне с *mdcLM*; гены *mat* обычно также организованы в оперон, расположение генов в котором варьирует. В ряде геномов присутствуют как *mat*, так и *mdc* гены, тогда как другие бактерии имеют только *mdc* или *mat* оперон. Как правило, ген регулятора *mdcY* колокализован с регулируемыми генами. У большинства бактерий в геноме присутствует только одна копия *mdcY*, тогда как у *Methylobacterium* spp. было выявлено два паралога данного транскрипционного фактора, один из которых кластеризован с опероном *mdc*, а второй – с генами TRAP транспортера или, как в случае *Methylobacterium* sp. 4-46, с генами TRAP транспортера и опероном *matAB*.

У многих Alphaproteobacteria, а также у ряда Beta- и Gammaproteobacteria (Таблица 9) гены метаболизма малоната колокализованы с генами, кодирующими С3-дикарбоксилат TRAP транспортеры, которые образуют с ними оперон или же имеют собственные сайты связывания MdcY в 5'-регулируемой области. Гены, кодирующие компоненты этих TRAP транспортеров у *Sinorhizobium meliloti*, были ранее названы *matPQM* [257]. Известно, что нуль-мутанты по какому-либо из генов *matPQM* не способны расти на минимальной среде, содержащей малонат в качестве единственного источника углерода [257]. Кроме того, в нашей работе было показано, что в большинстве случаев наличие генов TRAP транспортеров в составе малонатного регулона коррелирует с отсутствием генов иных известных транспортеров малоната – *mdcLM, mdcF* и *matC* (Таблица 9). Среди полностью секвенированных геномов, ортологи *matPQM* были идентифицированы только среди бактерий, в геноме которых присутствуют гены метаболизма малоната. Все это позволяет уверенно предполагать, что вышеупомянутые TRAP транспортеры участвуют в транспорте малоната и входят в состав регулона MdcY.

5.1.2 Активатор MdcR из семейства LysR

Ортологи MdcR были идентифицированы у представителей Gamma- (Alteromonadales, Enterobacteriales, Pseudomonadales) и Betaproteobacteria (Burkholderiales), но не у Alphaproteobacteria (Приложение Г2). В случае присутствия MdcR у Gammaproteobacteria это, как правило, единственный регулятор метаболизма малоната, тогда как у Betaproteobacteria MdcR часто встречается вместе с другими малонатными регуляторами.

У большинства представителей Gammaproteobacteria MdcR регулирует экспрессию единственного оперона, *mdcABCDEGHLM* или *mdcABCDEFGH*. У Betaproteobacteria гены *mdc* организованы по-разному, обычно также в виде единственного оперона (чаще всего это *mdcACDEGBH* или *mdcABCDE(LM)GH*), иногда двух, как, например, у *Burkholderia phytofirmans* PsJN, *Ralstonia eutropha* JMP134 (при этом гены *mdcLM* располагаются в другом локусе).

Транспортер MdcLM является наиболее типичным для регулона MdcR (Таблица 9). У некоторых Gammaproteobacteria (*C. koseri, Enterobacter sp.* 638) присутствует другой транспортер, кодируемый геном *mdcF*, а среди представителей Betaproteobacteria (*Delftia acidovorans* SPH-1, *Methylibium petroleiphilum* PM1) вместо *mdcLM* может встречаться *matC*.

Предполагаемый мотив связывания MdcR был идентифицирован при помощи филогенетического футпринтинга 5'-регуляторных областей *mdcA* и представляет собой палиндром длиной 23 нуклеотида с консенсусом ATCRTTACYYTGARSGTAAYGAT. У большинства Gamma- и Betaproteobacteria ген *mdcR* не авторегулируется, исключение

составляют *Ralstonia picketti* и *Psychromonas ingrahamii*, где *mdcR* располагается дивергентно к остальным *mdc* генам и, таким образом, делит с ними регуляторную область с предсказанным сайтом связывания.

5.1.3 Активатор PrpR из семейства FIS

PrpR является активатором транскрипции генов метаболизма пропионата у ряда Gamma- (Enterobacteriales и Xanthomonadales) и Betaproteobacteria (Burkholderiales) (Приложение Г3).

Ген регулятора *prpR* у всех представителей Enterobacteriales образует дивергон с генами утилизации пропионата, организованными в оперон *prpBCDE*. У большинства Xanthomonadales и Betaproteobacteria *prpR* располагается в дивергоне с *prpBC-acnD-prpF*. Ряд представителей Betaproteobacteria имеет иную структуру регулона PrpR, например, *acnD-prpF* у *Burkholderia phytofirmans* и *Burkholderia xenovorans* (Таблица 9). Кроме того, у многих Betaproteobacteria имеются также другие транскрипционные факторы, например, SdhR*, контролирующие метаболизм пропионата совместно с PrpR (см. далее). При этом интересно отметить, что у ряда представителей *Burkholderia* spp. и *Ralstonia* spp. в геноме обнаруживается два паралога генов *prpB* и/или *prpC*, один из которых находится под регуляцией PrpR, а другой – под регуляцией SdhR* (Таблица 9).

Предполагаемый мотив связывания PrpR был идентифицирован при помощи филогенетического футпринтинга 5'-регуляторных областей *prpB* и *prpR*. Предсказанный мотив представляет собой палиндром длиной 16 нуклеотидов с консенсусом RTTTCAWWWWTGAAAY. Так как PrpR является σ^{54}-зависимым активатором транскрипции [238], было проверено наличие соответствующих промоторов перед регулируемыми генами. Было показано, что вероятные σ^{54}-промоторные последовательности действительно обнаруживаются перед генами пропионатного метаболизма, принадлежащими PrpR регулону.

5.2 Новые регуляторы метаболизма малоната и пропионата, реконструкция регулонов

5.2.1 MlnR* – транскрипционный фактор из подсемейства FADR семейства GNTR

У ряда представителей Beta- (Burkholderiales, Rhodocyclales) и Gammaproteobacteria (Chromatiales, Xanthomonadales) был идентифицирован новый регуляторный ген, колокализованный с генами метаболизма малоната и кодирующий транскрипционный фактор из семейства GntR. Этот регулятор, названный нами MlnR* (здесь и далее звездочкой обозначены названия, присвоенные в данной работе), представляет собой родственный MdcY транскрипционный фактор, однако не является ортологом последнего (подтверждено

филогенетическим анализом, Приложение Г1). Для MlnR* было предсказано участие в регуляции метаболизма малоната, а также части цитрамалатного цикла.

У некоторых представителей Betaproteobacteria (*Acidovorax* sp. JS42, *Polaromonas* sp. JS666, *Verminephrobacter eiseniae*) в геноме было идентифицировано две паралогичных копии *mlnR**. В этом случае, один регуляторный ген образует оперон с *matAB*, тогда как второй паралог располагается в дивергоне с опероном, включающим гены ферментов цитрамалатного цикла, *mutB, meaB, pccBA* и *epi*. У бактерий, имеющих в геноме одну копию *mlnR**, регуляторный ген был кластеризован либо с генами метаболизма малоната *mat* (*Bordetella* spp., *Cupriavidus taiwanensis, Ralstonia eutropha, Ralstonia metallidurans*), либо с генами цитрамалатного цикла (*Delftia acidovorans, Leptothrix cholodnii, Polaromonas naphthalenivorans, Thauera* sp. MZ1T). У представителей Xanthomonadales *mlnR** формирует оперон с генами *mdcACDEGBH* и *matC*. Особенная организация регулона отмечена у *Alkalilimnicola ehrlichii* MLHE-1 (Chromatiales), где *mlnR** колокализован с генами *matAB* и *matPQM*; подобный состав регулона напоминает таковой для регулона MdcY многих Alphaproteobacteria (Таблица 9).

Филогенетический футпринтинг 5'-регуляторных областей *mlnR** выявил два типа предполагаемых палиндромных мотивов связывания. Первый из них (Тип 1) характерен для Betaproteobacteria, тогда как другой вариант (Тип 2) оператора MlnR* был идентифицирован у Gammaproteobacteria (Рисунок 15). Последовательность мотива второго типа частично совпадает после сдвига с мотивом связывания первого типа, в связи с чем ряд сайтов распознается обоими профилями (PWM). Оба варианта мотива связывания содержат общую структуру – короткий палиндром RTAATTAY, присутствующий в виде двух повторов в мотиве второго типа и в виде единственной последовательности в составе мотива первого типа.

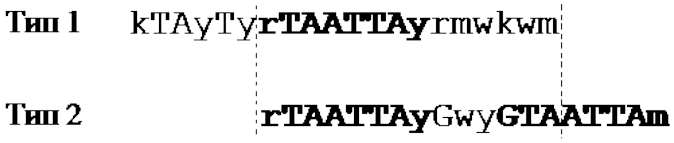


Рисунок 15. Структура двух типов мотивов связывания регулятора MlnR*

Общая часть мотивов выделена пунктиром; повторяющиеся элементы показаны жирным шрифтом.

5.2.2 Транскрипционные факторы из семейств GntR и LysR у *Burkholderia* spp.

В геноме представителей рода *Burkholderia* spp. присутствуют гены *mdc*, однако ортологов регуляторов MdcY не было обнаружено. Среди *Burkholderia* spp. ортологи MdcR присутствуют только у *B. multivorans, B. sp.* 383 и *B.phytofirmans*, более того, предполагаемые

сайты связывания MdcR были идентифицированы в 5’-регулируемой области генов метаболизма малоната только у *B. phytofirmans*. В ходе данной работы было проверено наличие каких-либо иных генов, колокализованных с генами *mdc* у бактерий рода *Burkholderia* spp., кодирующих возможные регуляторы метаболизма малоната. Оказалось, что у ряда представителей *Burkholderia* spp. (см. Таблица 9) рядом с генами *mdc* находится ген, кодирующий транскрипционный фактор семейства GntR. Этот регулятор – родственный MdcY транскрипционный фактор, однако не является ортологом последнего (подтверждено филогенетическим анализом, Приложение Г1). У некоторых других бактерий рода *Burkholderia* spp. оперон *mdc* был колокализован с геном транскрипционного фактора из семейства LysR (Таблица 9). Ортологи вышеупомянутых регуляторов отсутствовали у остальных исследованных бактерий. Филогенетический футпринтинг не выявил потенциальных мотивов связывания этих транскрипционных факторов в связи с высоким уровнем консервативности всей межгенной области исследованных генов у проанализированных близкородственных бактерий.

5.2.3 PrpR* – транскрипционный фактор из подсемейства FadR семейства GntR

У множества представителей Gammaproteobacteria (Alteromonadales, Oceanospirillales, Pseudomonadales, Vibrionales), а также у некоторых Beta- (Burkholderiales) и Deltaproteobacteria (*Geobacter metallireducens* GS-15) метаболизм пропионата находится под контролем регулятора из подсемейства FadR семейства GntR (Приложение Г4). Этот транскрипционный фактор был назван нами PrpR*.

Гены метаболизма пропионата у большинства представителей Alteromonadales и Pseudomonadales организованы в оперон *prpR*BC-acnD-prpFD*, или же более короткие опероны *prpR*BC-acnD-prpF* или *prpR*BC-acnD*. У большинства Vibrionales в геноме идентифицирован оперон *prpR*BC-acnD-prpFE*, тогда как у большинства представителей Oceanospirillales гены метаболизма пропионата образуют оперон *prpR*BCD*. Сходная организация оперона (*prpR*BDC*) наблюдается также у *G. metallireducens*.

Geobacter metallireducens – единственный представитель Deltaproteobacteria, в геноме которого присутствуют ортологи каких-либо регуляторов метаболизма малоната и пропионата, изученных в этой работе. Данные транскрипционные факторы отсутствовали в том числе и у других бактерий того же рода – *Geobacter sulfurreducens* и *Geobacter uraniireducens* – несмотря на наличие генов метаболизма пропионата *prp* в их геномах. Среди прочих представителей Delta- и Epsilonproteobacteria только у *Helicobacter hepaticus* было отмечено наличие в геноме полноценного пути утилизации пропионата (*prpEBCD* гены), однако не было

идентифицировано ортологов регуляторов метаболизма пропионата, проанализированных в данной работе.

Среди Betaproteobacteria PrpR*, вероятно, является достаточно редким и минорным регулятором метаболизма пропионата, тогда как большая часть пропионатного метаболизма контролируется транскрипционным фактором SdhR* из подсемейства HutC семейства GntR (см. далее). Из всех представителей Betaproteobacteria, PrpR* регулирует гены утилизации пропионата только у *Bordetella* spp. и *Verminephrobacter eiseniae* (Таблица 9).

Предсказанный мотив связывания PrpR*, идентифицированный с помощью филогенетического футпринтинга 5'-регулируемых областей *prpR**, представляет собой палиндромную последовательность длиной 12 нуклеотидов с консенсусом ATTGTCGACAAT.

5.2.4 PrpQ* – транскрипционный фактор из семейства XRE

У большинства исследованных Alphaproteobacteria (Caulobacterales, Rhizobiales, Rhodobacterales, Rhodospirillales, Sphingomonadales), а также у некоторых Betaproteobacteria (Burkholderiales) предполагаемым регулятором утилизации пропионата является транскрипционный фактор семейства XRE, названный нами PrpQ* (Приложение Г5).

Структура реконструированных регулонов PrpQ* варьирует у различных представителей Alpha- и Betaproteobacteria, у большинства в состав регулонов входят гены, кодирующие ферменты цитрамалатного и/или метилцитратного путей – *prpBCDF*, *acnD*, *pccBA* и *mutB* (Таблица 9).

С помощью филогенетического футпринтинга 5'-регулируемых областей *prpQ** и *pccB* был идентифицирован предполагаемый мотив связывания PrpQ*, представляющий собой короткий палиндром длиной 8 нуклеотидов с консенсусом TTTGCRAA. Подобная последовательность часто присутствует в 5'-регулируемой области во множестве копий.

5.2.5 SdhR* – транскрипционный фактор из подсемейства HutC семейства GntR

Многие Betaproteobacteria порядка Burkholderiales имеют транскрипционный фактор из подсемейства HutC семейства GntR, далее называемый SdhR* (Приложение Г6). Этот регулятор, в соответствии с колокализацией гена в геноме, вероятно, участвует в контроле экспрессии генов, кодирующих ферменты ЦТК, глиоксилатного шунта и метаболизма пропионата (Рисунок 2). У всех исследованных Burkholderiales в состав регулона входит оперон, включающий гены *sdhR**, *sdhABCDE* (сукцинат дегидрогеназа) [258] и *gltA* (цитрат синтаза), а также дивергентно расположенный ген *mdh* (малат дегидрогеназа), который часто находится в опероне с *citE* (цитрат лиаза), *tam* (транс-аконитат 2-метилтрансфераза) [259],

генами утилизации пропионата *prpB*, *prpC*, *prpD* и *prpF*, *acnA* и *acnB* (Таблица 9). Ген *acnB* может также располагаться в отдельном локусе и иметь собственный сайт связывания SdhR*. Кроме того, у многих Betaproteobacteria SdhR* также предположительно регулирует экспрессию гена еще одной аконитат гидратазы *acnD* и изоцитрат лиазы *aceA*.

С помощью филогенетического футпринтинга 5'-областей генов *sdhR**, *mdh* и *acnB* была выявлена высококонсервативная область, предположительно представляющая собой сайт связывания SdhR*. Предсказанный мотив связывания – палиндром длиной 26 нуклеотидов, с консенсусом TCTTATGTCTTATATAAGACATAAGA. Эта последовательность обладает внутренней симметрией: каждое плечо палиндрома включает два прямых повтора TCTTAT или ATAAGA, соответственно.

Таблица 9. Основные члены регулонов метаболизма малоната и пропионата у Proteobacteria.

	sdhABCDE	gltA	mdh	citE	aceA	acnB	acnA	acnD	prpC	prpB	prpD	prpF	prpE	pccB	pccA	mutB	mmcm	meaB	epi	matA	matB	matRQM	mdcABCDEGH	matC	mdcLM	mdcF
AJS	+	+	+	+	+	+	+	+	+	+	0	+	+	+	+	+	0	+	+	+	+	0	0	0	0	0
AZO	+	+	+	+	+	+	0	+	0	0	0	0	+	+	+	+	0	+	+	+	0	0	0	0	0	0
BAV	+	+	+	0	0	+	+	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
BBR	+	+	+	0	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	+	+	0	0	0	0	0
BPA	+	+	+	0	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	+	+	0	0	0	0	0
BPE	+	+	+	0	+	+	+	+	+	0	+	+	0	0	0	0	0	0	0	+	+	0	0	0	0	0
BPT	+	+	+	0	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	+	+	+	0	0	0	0
BAC	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	0	+	0	+	0
BCJ	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	0	+	0	+	0
BMU	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	0	+	0	+	0
BPH	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	+	0	0	0	0	0	0	+	0	0	0
BPY	+	+	+	+	+	+	+	+	+	0	+	+	+	0	0	0	0	0	0	0	0	0	+	0	+	0
BUR	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	0	+	0	+	0
BTE	+	+	+	+	+	0	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	0	+	0	+	0
BVI	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0	+	0	0	0	0	0	0	+	0	+	0
BXE	+	+	+	+	+	+	+	+	+	0	+	+	+	0	0	0	0	0	0	0	0	0	+	0	0	0
RME	+	+	+	+	+	+	+	+	+	+	0	+	+	0	0	+	0	0	0	+	+	0	0	0	0	0
CTI	+	+	+	+	+	+	+	+	+	+	0	+	+	0	0	+	0	0	0	+	+	0	0	0	0	0
DAR	+	+	+	+	+	+	0	0	0	0	0	0	+	+	+	+	0	+	+	+	0	+	+	0	0	0
DAC	+	+	+	+	+	+	+	+	+	+	0	+	+	+	+	+	0	+	+	+	+	0	+	+	0	0
LCH	+	+	+	+	+	+	0	0	0	0	0	0	+	+	+	+	0	+	+	0	+	0	+	0	0	0
MPT	+	+	+	+	+	+	0	+	0	0	0	0	+	+	+	+	0	+	+	0	+	0	+	+	0	0

	sdhABCDE	gltA	mdh	citE	aceA	acnB	acnA	acnD	prpC	prpB	prpD	prpF	prpE	pccB	pccA	mutB	mmcm	meaB	epi	matA	matB	matRQM	mdcABCDEGH	matC	mdcLM	mdcF
PNA	+	+	+	+	+	+	+	0	0	0	0	0	+	+	+	+	0	+	+	+	+	0	+	0	+	0
POL	+	+	+	+	+	+	+	+	0	0	0	0	+	+	+	+	0	+	+	+	+	0	0	0	0	0
REU	+	+	+	+	+	+	+	+	+	+	0	+	+	0	0	+	0	0	0	+	+	0	+	0	+	0
RPI	+	+	+	+	+	+	+	+	+	+	0	+	+	0	0	+	0	0	0	0	0	0	+	0	+	0
RSO	+	+	+	+	+	+	+	+	+	+	+	+	+	0	0	+	0	0	0	0	0	0	0	0	0	0
TMZ	+	+	+	+	+	+	0	+	0	0	0	0	+	+	+	+	0	+	+	+	0	0	0	0	0	0
VEI	+	+	+	0	+	+	+	0	+	+	+	0	+	+	+	+	0	+	+	+	+	+	+	0	0	0
ATC	+	+	+	0	+	+	0	+	0	0	0	+	0	+	+	+	0	0	0	0	0	0	0	0	0	0
AZC	+	+	+	0	0	0	0	+	0	0	0	+	+	+	+	+	+	+	0	+	+	+	0	0	0	0
BJA	+	+	+	0	0	0	+	0	0	0	0	0	+	+	+	+	0	+	0	+	+	0	+	0	+	+
BMB	+	+	+	0	+	0	+	+	0	0	0	0	+	+	+	+	0	0	0	0	0	0	0	0	0	0
BME	+	+	+	0	+	0	+	+	0	0	0	0	+	+	+	+	0	0	0	0	0	0	0	0	0	0
BOV	+	+	+	0	+	0	+	+	0	0	0	0	+	0	+	0	0	0	0	0	0	0	0	0	0	0
BMS	+	+	+	0	+	0	+	+	0	0	0	0	+	0	+	+	0	0	0	0	0	0	0	0	0	0
CAK	+	+	+	0	0	0	0	+	0	0	0	+	0	+	+	+	+	0	0	0	0	0	0	0	0	0
DSH	+	+	+	0	0	0	0	+	0	0	0	0	+	+	+	+	0	0	0	+	+	+	0	0	0	0
JAN	+	+	+	0	0	0	0	+	0	0	0	0	+	+	+	+	0	0	0	0	+	0	0	0	0	0
MAG	+	+	0	+	0	+	+	0	0	0	0	+	+	+	+	+	0	+	0	+	+	0	0	0	0	0
MCH	+	+	0	0	0	0	+	0	0	0	0	0	+	+	+	+	+	+	0	+	+	+	+	0	0	0
MEX	+	+	0	0	0	0	+	0	0	0	0	0	+	+	+	+	+	+	0	+	+	+	+	0	0	0
MPO	+	+	0	0	0	0	+	0	0	0	0	0	+	+	+	+	+	+	0	+	+	+	+	0	+	0
MRD	+	+	0	0	0	0	+	0	0	0	0	+	+	+	+	+	+	+	0	+	+	+	+	0	0	0
MET	+	+	0	0	0	0	+	0	0	+	0	0	+	+	+	+	+	+	0	+	+	+	+	0	+	0
MSL	+	+	0	0	+	0	+	+	0	0	0	0	0	+	+	+	+	0	0	0	0	0	+	0	+	0
OAN	+	+	0	0	+	0	+	+	0	0	0	0	+	+	+	+	0	0	0	0	0	0	0	0	0	0
PDE	+	+	0	0	0	0	+	+	+	+	+	0	+	+	+	+	0	+	0	+	+	0	0	+	0	0
PZU	+	+	0	0	0	0	+	+	0	0	0	0	0	+	+	+	+	0	0	0	+	0	0	0	0	0
RET	+	+	0	0	+	0	+	0	0	0	0	0	+	+	+	+	0	0	0	+	+	0	0	+	0	0
RLE	+	+	0	0	+	0	+	0	0	0	0	0	+	+	+	+	0	0	0	+	+	0	0	+	0	0
RSP	+	+	0	0	0	0	+	+	0	0	0	0	+	+	+	+	0	+	0	0	+	0	0	0	0	0
RPT	+	+	0	0	+	0	+	+	+	+	+	0	+	+	+	+	+	+	0	+	+	0	0	0	0	0
RCE	+	+	0	0	+	0	+	+	+	+	+	0	+	+	+	+	0	+	0	0	+	0	0	0	0	0
RRU	+	+	0	0	0	0	+	+	+	+	+	0	+	+	+	+	+	+	0	+	0	0	+	0	+	0
RDE	+	+	0	0	0	0	+	+	0	0	0	0	+	+	+	+	0	+	0	+	+	+	0	0	0	0
SMD	+	+	0	0	+	0	+	+	0	0	0	0	+	+	+	+	0	+	0	+	+	+	0	0	0	0
SME	+	+	0	0	+	0	+	0	0	0	0	0	+	+	+	+	0	0	0	+	+	+	0	0	0	0
SWI	+	+	0	0	0	0	+	+	+	+	0	+	0	+	+	+	0	0	0	0	+	+	+	0	+	0
ABY	+	+	0	0	+	+	+	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	+	0	+	0
ACI	+	+	0	0	+	+	+	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	+	0	+	0

	sdhABCDE	gltA	mdh	citE	aceA	acnB	acnA	acnD	prpC	prpB	prpD	prpF	prpE	pccB	pccA	mutB	mmcm	meaB	epi	matA	matB	matRQM	mdcABCDEGH	matC	mdcLM	mdcF
ABO	+	+	0	0	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
AEH	+	0	+	0	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	+	+	+	0	0	0	0
AMC	+	+	0	0	+	+	+	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
AVI	+	+	0	0	0	+	+	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	+	0	+	0
CSA	+	+	0	0	+	+	+	0	+	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CKO	+	+	0	0	+	+	+	0	+	+	+	0	+	0	0	0	0	0	0	0	0	0	+	0	0	+
CPS	+	+	0	0	0	+	0	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
ENT	+	+	0	0	+	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	0	0	+
ECO	+	+	0	0	+	+	+	0	+	+	+	0	+	0	0	+	0	0	0	0	0	0	0	0	0	0
EFE	+	+	0	0	+	+	+	0	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0	0
HCH	+	+	0	0	+	+	+	0	+	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ILO	+	+	0	0	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
KPN	+	+	0	0	+	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	0	0	0
MAQ	+	+	0	0	0	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
MMW	+	+	0	0	+	0	+	0	+	+	+	0	0	0	0	0	0	0	0	+	+	0	0	0	0	0
PPR	+	+	0	0	+	+	0	0	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0	0
PLU	+	+	0	0	+	+	+	0	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0	0
PAT	+	+	0	0	+	+	+	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PHA	+	+	0	0	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
PAE	+	+	0	0	0	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	+	0	+	0
PEN	+	+	0	0	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	+	0	+	0
PFL	+	+	0	0	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	0	+	0	+	0
PMY	+	+	0	0	0	+	+	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	+	0	+	0
PPW	+	+	0	0	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	+	0	+	0
PSA	+	+	0	0	0	+	+	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	+	0	+	0
PST	+	+	0	0	+	+	+	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	+	0	+	0
PAR	+	+	0	0	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	+	0	+	0
PIN	+	+	0	0	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	+	+	0	0	0
STY	+	+	0	0	+	+	+	0	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0	0
STM	+	+	0	0	+	+	+	0	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0	0
SAZ	+	+	0	0	+	+	0	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
SBM	+	+	0	0	+	+	0	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
SDN	+	+	0	0	+	+	0	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SFR	+	+	0	0	+	+	+	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
SHL	+	+	0	0	0	+	0	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
SLO	+	+	0	0	+	+	0	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
SON	+	+	0	0	+	+	0	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SPL	+	+	0	0	0	+	0	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
SWP	+	+	0	0	0	+	0	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0

	sdhABCDE	gltA	mdh	citE	aceA	acnB	acnA	acnD	prpC	prpB	prpD	prpF	prpE	pccB	pccA	mutB	mmcm	meaB	epi	matA	matB	matRQM	mdcABCDEGH	matC	mdcLM	mdcF
SPC	+	+	0	0	+	+	0	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
SSE	+	+	0	0	+	+	0	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
SHN	+	+	0	0	+	+	0	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
SMT	+	+	0	0	+	+	+	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	+	+	0	0
VCH	+	+	0	0	+	+	0	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
VHA	+	+	0	0	+	+	0	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
VPA	+	+	0	0	+	+	0	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
VSP	+	+	0	0	0	+	0	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
VVY	+	+	0	0	+	+	0	+	+	+	0	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0
XAC	+	+	0	0	+	+	+	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	+	+	0	0
XCC	+	+	0	0	+	+	+	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	+	+	0	0
XCV	+	+	0	0	+	+	+	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	+	+	0	0
XOM	+	+	0	0	0	+	+	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	+	+	0	0
XOP	+	+	0	0	0	+	+	+	+	+	0	+	0	0	0	0	0	0	0	0	0	0	+	+	0	0
GME	+	+	0	0	0	+	+	0	+	+	+	0	0	+	0	0	0	0	+	0	0	0	0	0	0	0

Обозначения: + – ген присутствует в геноме; 0 – ген отсутствует. Цветом обозначена регуляция: гены, регулируемые MdcY, выделены желтым, MdcR – оранжевым, MlnR* – розовым, SdhR* – голубым, PrpR – синим, PrpR* – зеленым, PrpQ* – фиолетовым, фактором транскрипции семейства LysR (Burkholderia) – коричневым, регулятором семейства GntR (Burkholderia) – темно-зеленым, гены, для которых неизвестна регуляция, показаны белым, двухцветные ячейки обозначают двойную регуляцию. Трехбуквенные обозначения геномов соответствуют аббревиатурам, приведенным в Приложении А.

5.3 Эволюция систем метаболизма малоната и пропионата у Proteobacteria

Наиболее однородная регуляция наблюдается среди Alphaproteobacteria: гены метаболизма малоната находятся под регуляцией MdcY, тогда как утилизация пропионата контролируется PrpQ* (Таблица 10). Оба транскрипционных фактора типичны для Alphaproteobacteria, тогда как у Beta- и Gammaproteobacteria они встречаются очень редко, вероятно, как результат горизонтального переноса генов (Рисунок 16).

Среди Gammaproteobacteria возможный перенос гена *mdcY* произошел у Alteromonadales и Pseudomonadales (Рисунок 17). В то же время, у представителей Chromatiales и Xanthomonadales присутствует иной регулятор метаболизма малоната, транскрипционный фактор MlnR* из

семейства GntR, типичный для Betaproteobacteria (см. далее) и, вероятно, унаследованный от общего предка с Betaproteobacteria (Приложение Г1). У прочих Gammaproteobacteria метаболизм малоната находится под контролем MdcR, и данный транскрипционный фактор представляется исходным и основным малонатным регулятором в этой таксономической группе (Рисунок 17). Метаболизм пропионата у Gammaproteobacteria регулируется преимущественно транскрипционным фактором PrpR* из семейства GntR, однако у некоторых представителей Enterobacteriales и всех исследованных Xanthomonadales пропионатным регулятором служит белок PrpR из семейства Fis. Этот транскрипционный фактор также обнаруживается у Betaproteobacteria и, согласно филогенетическому анализу (Приложение Г3), может быть унаследован Beta- и Gammaproteobacteria от их общего предка, или же получен общим предком Enterobacteriales и Xanthomonadales от Betaproteobacteria в результате горизонтального переноса генов. Оба эволюционных сценария включают множественные потери PrpR в различных линиях бактерий.

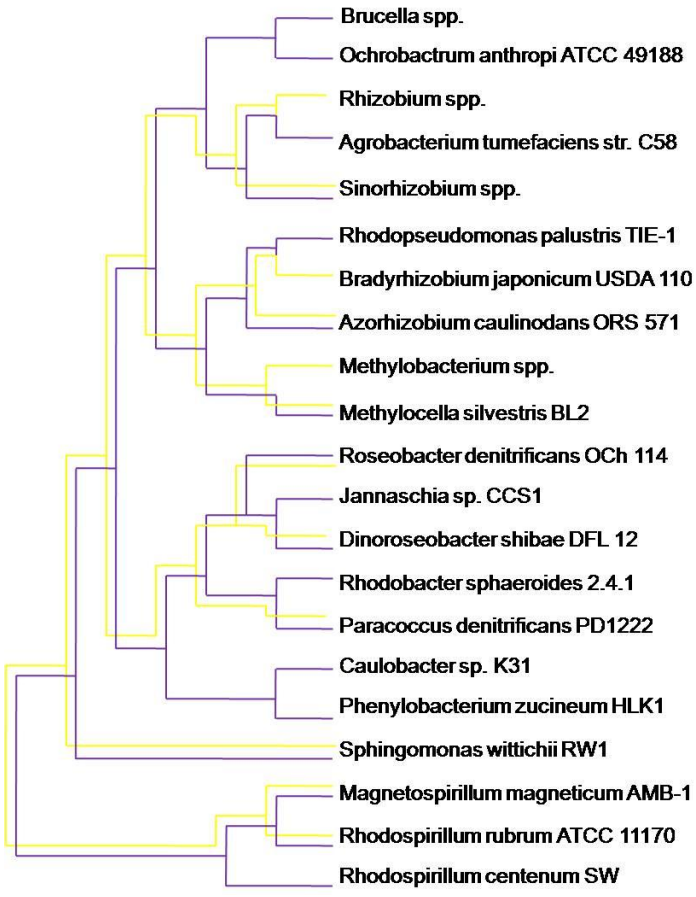


Рисунок 16. Регуляторы метаболизма малоната и пропионата у Alphaproteobacteria

Цвет линий обозначает наличие соответствующего регулятора в геноме: MdcY показан желтым, PrpQ* – фиолетовым.

Таблица 10. Распределение регуляторов метаболизма малоната и пропионата среди различных групп Proteobacteria

Таксон	Количество родов бактерий, имеющих данный транскрипционный фактор						
	MdcY	MlnR*	MdcR	PrpR	PrpR*	PrpQ*	SdhR*
Alpha	12	0	0	0	0	17	0
Beta	4	11	5	3	2	3	12
Gamma	2	3	7	6	16	0	0
Delta	0	0	0	0	1	0	0

Транскрипционный фактор PrpR* из семейства GntR широко распространен среди Gammaproteobacteria, а также присутствует у ряда Betaproteobacteria и *Geobacter metallireducens* (Deltaproteobacteria). В соответствии с расположением клады на филогенетическом дереве ортологов PrpR* (Приложение Г4), *G. metallireducens* предположительно получил PrpR* в результате горизонтального переноса генов от общего предка Alteromonadales. Филогенетический анализ дает основания предполагать, что часть Betaproteobacteria получила PrpR* в результате горизонтального переноса от Pseudomonadales (например, *Bordetella petrii*), тогда как ряд других представителей этого таксона (например, *Bordetella bronchiseptica, Verminephrobacter eiseniae*), вероятно, унаследовали этот транскрипционный фактор от общего предка с Gammaproteobacteria. Интересный вариант регуляции обнаруживается у *Bordetella avium*, имеющей две копии гена *prpR**, один из которых предположительно получен от представителей Pseudomonadales, тогда как второй унаследован от общего предка с Gammaproteobacteria. Возможно, в данный момент мы можем наблюдать промежуточную эволюционную стадию замещения транскрипционным фактором PrpR* регулятора PrpR у Gammaproteobacteria (Рисунок 17).

Наиболее разнообразная регуляция наблюдается у Betaproteobacteria (Рисунок 18, Таблица 9, Таблица 10). В целом среди бактерий этой группы идентифицировано пять регуляторов метаболизма малоната, включая: MdcY, предположительно полученный ими от Alphaproteobacteria в результате горизонтального переноса генов (Приложение Г1); MdcR, полученный от Gammaproteobacteria в результате горизонтального переноса или же унаследованный от общего предка Beta- и Gammaproteobacteria (Приложение Г2); специфические для бактерий рода *Burkholderia* spp. транскрипционные факторы из семейств

LysR и GntR. Однако у большинства Betaproteobacteria метаболизм малоната находится под контролем регулятора MlnR* из семейства GntR.

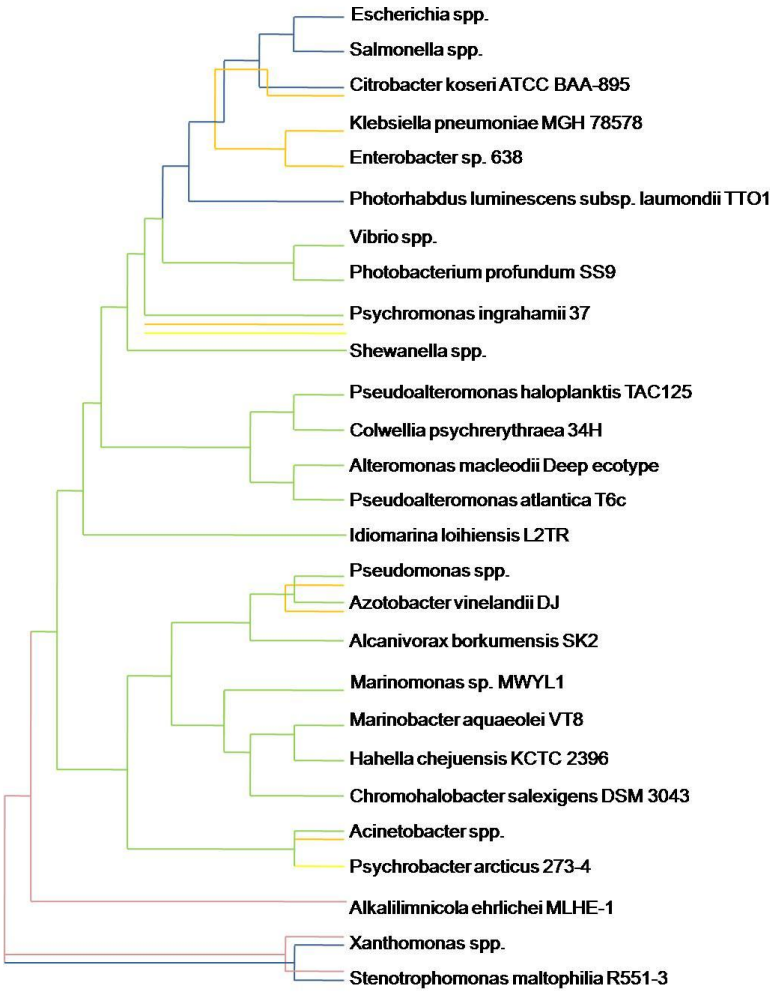


Рисунок 17. Регуляторы метаболизма малоната и пропионата у Gammaproteobacteria

Цвет линий обозначает наличие соответствующего регулятора в геноме: MdcY показан желтым, MdcR – оранжевым, MlnR* – розовым, PrpR – синим, PrpR* – зеленым.

Регуляция утилизации пропионата у Betaproteobacteria также разнообразна. PrpQ* является регулятором пропионатного метаболизма только у узкой группы Betaproteobacteria, трех близкородственных представителей семейства Comamonadaceae, вероятно, получивших этот транскрипционный фактор в результате горизонтального переноса генов от Alphaproteobacteria (Приложение Г5). У большинства Betaproteobacteria основным регулятором утилизации пропионата (либо единственным, либо совместно с PrpR или иногда PrpR*) служит SdhR* (Рисунок 18, Таблица 9). Следует отметить, что у многих Betaproteobacteria, имеющих

MlnR*, отсутствует регулятор PrpR из семейства FIS, что может быть обусловлено тем, что эти транскрипционные факторы контролируют альтернативные пути метаболизма пропионата.

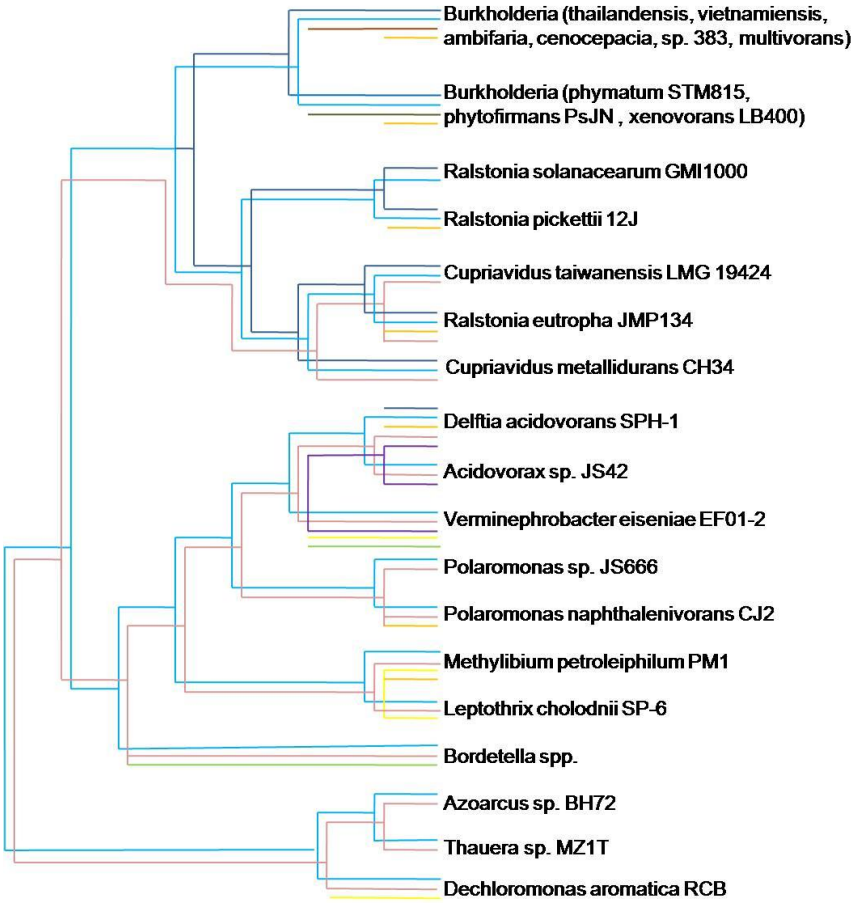


Рисунок 18. Регуляторы метаболизма малоната и пропионата у Betaproteobacteria

Цвет линий обозначает наличие соответствующего регулятора в геноме: MdcY показан желтым, MdcR – оранжевым, MlnR* – розовым, SdhR* – голубым, PrpR – синим, PrpR* – зеленым, PrpQ* – фиолетовым, регулятор семейства LYSR (*Burkholderia* spp.) – коричневым, регулятор семейства GNTR (*Burkholderia* spp.) – темно-зеленым.

5.4 Заключение

Нами была подробно исследована регуляция метаболизма малоната и пропионата у Proteobacteria, показана ее вариабельность и разнообразие в различных таксономических группах, а также построена модель эволюции данных метаболических систем.

В настоящей работе был идентифицирован ряд новых транскрипционных факторов, регулирующих метаболизм малоната и пропионата: MlnR*, PrpR*, PrpQ*, SdhR*. Для каждого из исследованных транскрипционных регуляторов метаболизма малоната и пропионата были определены мотивы связывания и реконструированы соответствующие регулоны, включающие, в частности: гены ЦТК (*sdhABCDE*, *mdh*, *gltA*, *citE*, *acnA* и *acnB*), глиоксилатного шунта (*aceA*), метилцитратного пути (*prpBCDEF*, *acnB* и *acnD*) и цитрамалатного цикла (*pccBA*, *epi*, *mutB* и *meaB*), а также TRAP транспортеры малоната (*matPQM*).

Выводы

1. Исследованы 1252 транскрипционных фактора семейства GntR из 64 ортологических групп, для которых идентифицированы мотивы связывания и реконструированы соответствующие регулоны.

2. Для подсемейств FadR, HutC и YtrA семейства GntR исследована коэволюция мотивов связывания ДНК и аминокислотных последовательностей регуляторов транскрипции, предсказаны вероятные ДНК-белковые контакты. Показано, что большая часть предсказанных контактов сходна для всех трех подсемейств и хорошо соотносится с известными данными о ДНК-белковых взаимодействиях.

3. Проведен анализ структуры дивергонов, регулируемых транскрипционными факторами подсемейств FadR и HutC семейства GntR, выявлены основные тенденции расположения сайтов связывания в дивергонах с единичными и двойными сайтами.

4. Для 23 ортологических групп регуляторов семейства GntR показано наличие слабых дополнительных боксов рядом с основным палиндромным сайтом связывания, которые, предположительно, могут участвовать в альтернативной димеризации транскрипционных факторов или же в связывании дополнительных субъединиц.

5. Детально исследована регуляция метаболизма гексуронатов у Gammaproteobacteria, построена модель эволюции родственных регуляторов UxuR и ExuR, разделены их мотивы связывания, реконструированы соответствующие регулоны, в том числе, идентифицирован ряд новых членов гексуронатных регулонов. По результатам предсказаний лабораторией Института биофизики клетки РАН экспериментально подтверждена UxuR-зависимая регуляция транскрипции *yjjM* и *yjjN* у *E. coli*.

6. Детально исследована регуляция метаболизма малоната и пропионата у Proteobacteria, показана ее вариабельность и разнообразие в различных таксономических группах, построена модель эволюции данных метаболических систем. Идентифицирован ряд новых транскрипционных регуляторов метаболизма малоната и пропионата. Для каждого из исследованных транскрипционных регуляторов метаболизма малоната и пропионата определены мотивы связывания и реконструированы соответствующие регулоны, в том числе описаны их новые члены.

Список используемых сокращений и обозначений

CoA – кофермент А

HLH – спираль-петля-спираль

HTH – спираль-поворот-спираль

KDG – 2-кето-3-деокси-D-глюконат

ORF – открытая рамка считывания

PLP – пиридоксаль-5-фосфат

PWM – матрица позиционных весов

RHH – лента-спираль-спираль

АТФ – аденозинтрифосфат

ДНК – дезоксирибонуклеиновая кислота

НАД – никотинамидадениндинуклеотид

п.н. – пары нуклеотидов

РНК – рибонуклеиновая кислота

цАМФ – циклический аденозинмонофосфат

ЦТК – цикл трикарбоновых кислот

Список работ, опубликованных по теме диссертации

Статьи в научных журналах:

1. Suvorova, I.A., Tutukina, M.N., Ravcheev, D.A., Rodionov, D.A., Ozoline, O.N., Gelfand, M.S. (2011). Comparative genomic analysis of the hexuronate metabolism genes and their regulation in gammaproteobacteria. *J Bacteriol*. **193(15)**:3956-63.

2. Suvorova, I.A., Ravcheev, D.A., Gelfand, M.S. (2012). Regulation and evolution of malonate and propionate catabolism in proteobacteria. *J Bacteriol*. **194(12)**:3234-40.

3. Suvorova, I.A., Korostelev, Y.D., Gelfand, M.S. (2015). GntR Family of Bacterial Transcription Factors and Their DNA Binding Motifs: Structure, Positioning and Co-Evolution. *PLoS One*. **10(7)**:e0132618.

Тезисы конференций:

1. Суворова И.А., Равчеев Д.А. Коэволюция белков семейства FadR и их сайтов связывания. Труды 32-й конференции молодых ученых и специалистов ИППИ РАН (ИТиС'09), Бекасово, 2009, с. 297-299.

2. Суворова И.А. Утилизация гексуронатов в гамма-протеобактериях. Исследование методами сравнительной геномики. Материалы XVII международной конференции студентов, аспирантов и молодых ученых «Ломоносов-2010», Москва, 2010.

3. Суворова И.А. Белок Prp2 подсемейства FadR – новый вариант регуляции метаболизма пропионата. Труды 33-й конференции молодых ученых и специалистов ИППИ РАН (ИТиС'10), Геленджик, 2010, с. 350-351.

4. Суворова И.А. Факторы транскрипции семейства GntR: эволюция регуляторных систем. Материалы международной научно-практической конференции «Постгеномные методы анализа в биологии, лабораторной и клинической медицине», Москва, 2010.

5. Suvorova I.A. Comparative genomic analysis of the hexuronate metabolism genes and their regulation in Gamma-Proteobacteria. Proceedings of Russian-German seminar «Regulation and Evolution of Cellular Systems» (RECESS, Регуляция и эволюция клеточных процессов), Мюнхен, 2011.

6. Суворова И.А. Регуляция метаболизма малоната и пропионата у Протеобактерий. Труды 34-й конференции молодых ученых и специалистов ИППИ РАН (ИТиС'11), Геленджик, 2011, с. 28-30.

7. Суворова И.А. Гельфанд М.С. Исследование сайтов связывания и их окружения для транскрипционных факторов семейства GntR. Конференция молодых ученых «Молекулярная и

клеточная биология: прикладные аспекты» в рамках отчетной конференции по программе фундаментальных исследований ПРАН «Молекулярная и клеточная биология», Москва, 2012, с. 29.

8. Suvorova I.A. Comparative analysis of the GntR family transcription factors and their binding sites. Proceedings of Russian-German seminar «Regulation and Evolution of Cellular Systems» (RECESS, Регуляция и эволюция клеточных процессов), Москва, 2012.

9. Суворова И.А. Исследование сайтов связывания и их окружения для транскрипционных факторов семейства GntR. Труды 35-й конференции молодых ученых и специалистов ИППИ РАН (ИТиС'12), Петрозаводск, 2012, с. 338-340.

10. Suvorova I.A. GntR family of bacterial transcription factors and their binding motifs: structure, positioning and co-evolution. Proceedings of Russian-German seminar «Regulation and Evolution of Cellular Systems» (RECESS, Регуляция и эволюция клеточных процессов), Венеция, 2013.

11. Suvorova, IA. GntR family of bacterial transcription factors and their binding motifs: structure and co-evolution. Proceedings of Moscow Conference on Computational Molecular Biology (MCCMB'13), Москва, 2013.

Благодарности

Выражаю благодарность Михаилу Сергеевичу Гельфанду за научное руководство при выполнении диссертации, Дмитрию Александровичу Родионову и Дмитрию Андреевичу Равчееву за ценное обсуждение работы, Семену Лейну и Ольге Цой за помощь в проверке текста диссертации, а также всем остальным коллегам из УНЦ «Биоинформатика» ИППИ РАН. Хочу также поблагодарить семью и друзей за терпение и моральную поддержку.

Список литературы

1. Santos, C.L., Tavares, F., Thioulouse, J., Normand, P. A phylogenomic analysis of bacterial helix-turn-helix transcription factors. (2009). *FEMS Microbiol Rev*. **33(2)**:411-29.

2. Gottesman, S., McCullen, C.A., Guillier, M., Vanderpool, C.K., Majdalani, N., Benhammou, J., Thompson, K.M., FitzGerald, P.C., Sowa, N.A., FitzGerald, D.J. (2006). Small RNA regulators and the bacterial response to stress. *Cold Spring Harb Symp Quant Biol.* **71**:1-11.

3. Charoensawan, V., Wilson, D., Teichmann, S.A. (2010). Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res*. **38(21)**:7364-77

4. Rodionov, D.A. (2007). Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chem Rev.* **107(8)**:3467-97.

5. Browning, D.F., Busby, S.J. (2004). The regulation of bacterial transcription initiation. *Nat Rev Microbiol*. **2(1)**:57-65.

6. Патрушев Л.И. Экспрессия генов. М.: «Наука», 2000, ISBN 5-02-001890-2.

7. Tan, K., McCue, L.A., Stormo, G.D. (2005). Making connections between novel transcription factors and their DNA motifs. *Genome Res.* **15(2)**:312-20.

8. Ofran, Y., Mysore, V., Rost, B. (2007). Prediction of DNA-binding residues from sequence. *Bioinformatics.* **23(13)**:i347-53.

9. Das, M.K., Dai H.K. (2007). A survey of DNA motif finding algorithms. BMC *Bioinformatics*. **8 Suppl 7**:S21.

10. Mironov, A.A., Koonin, E.V., Roytberg, M.A., Gelfand, M.S. (1999). Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res.* **27(14)**:2981-9.

11. Gelfand, M.S., Koonin, E.V., Mironov, A.A. (2000). Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Res.* **28(3)**:695-705.

12. Luscombe, N.M., Austin, S.E., Berman, H.M., Thornton, J.M. (2000). An overview of the structures of protein-DNA complexes. *Genome Biol.* **1(1)**:REVIEWS001.

13. Mirny, L.A., Gelfand, M.S. (2002). Structural analysis of conserved base pairs in protein-DNA complexes. *Nucleic Acids Res*. **30(7)**:1704-11.

14. Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., Mann, R.S. (2010). Origins of specificity in protein-DNA recognition. *Annu Rev Biochem*. **79**:233-69.

15. Subrahmanyam, S., Cronan, J.E. Jr. (1998). Overproduction of a functional fatty acid biosynthetic enzyme blocks fatty acid synthesis in Escherichia coli. *J Bacteriol*. **180(17)**:4596-602.

16. Singh, G.P., Dash, D. (2013). Electrostatic mis-interactions cause overexpression toxicity of proteins in E. coli. *PLoS One*. **8(5)**:e64893.

17. Erova, T.E., Kosykh, V.G., Sha, J., Chopra, A.K. (2012). DNA adenine methyltransferase (Dam) controls the expression of the cytotoxic enterotoxin (act) gene of Aeromonas hydrophila via tRNA modifying enzyme-glucose-inhibited division protein (GidA). *Gene*. **498(2)**:280-7.

18. Baranello, L., Levens, D., Gupta, A., Kouzine, F. (2012). The importance of being supercoiled: how DNA mechanics regulate dynamic processes. *Biochim Biophys Acta*. **1819(7)**:632-8.

19. Hatfield, G.W., Benham, C.J. (2002). DNA topology-mediated control of global gene expression in Escherichia coli. *Annu Rev Genet*. **36**:175-203.

20. Martínez-Antonio, A., Janga, S.C., Thieffry, D. (2008). Functional organisation of Escherichia coli transcriptional regulatory network. *J Mol Biol*. **381(1)**: 238–247.

21. Kazmierczak, M.J., Wiedmann, M., Boor, K.J. (2005). Alternative sigma factors and their roles in bacterial virulence. *Microbiol Mol Biol Rev*. **69(4)**:527-43.

22. Helmann, J. D. (2001). Sigma Factors in Gene Expression. eLS.

23. Phadtare, S., Severinov, K. (2009). Comparative analysis of changes in gene expression due to RNA melting activities of translation initiation factor IF1 and a cold shock protein of the CspA family. *Genes Cells*. **14(11)**:1227-39.

24. Hunger, K., Beckering, C.L., Wiegeshoff, F., Graumann, P.L., Marahiel, M.A. (2006). Cold-induced putative DEAD box RNA helicases CshA and CshB are essential for cold adaptation and interact with cold shock protein B in Bacillus subtilis. *J Bacteriol*. **188(1)**:240-8.

25. Gelfand, M.S. (2006). Bacterial cis-Regulatory RNA Structures. *Mol Biol*. **40(4)**:541–550.

26. Waters, L.S., Storz, G. (2009). Regulatory RNAs in bacteria. *Cell*. **136(4)**:615-28.

27. Breaker, R.R. (2011). Prospects for riboswitch discovery and analysis. *Mol Cell*. **43(6)**:867-79.

28. Kazanov, M.D., Vitreschak, A.G., Gelfand, M.S. (2007). Abundance and functional diversity of riboswitches in microbial communities. *BMC Genomics*. **8**:347.

29. Green, N.J., Grundy, F.J, Henkin, T.M. (2010). The T box mechanism: tRNA as a regulatory molecule. *FEBS Lett*. **584(2)**:318-24.

30. Yanofsky, C. (2000). Transcription attenuation: once viewed as a novel regulatory strategy. *J Bacteriol*. **182(1)**: 1–8.

31. McCullen, C.A., Benhammou, J.N., Majdalani, N., Gottesman, S. (2010). Mechanism of positive regulation by DsrA and RprA small noncoding RNAs: pairing increases translation and protects rpoS mRNA from degradation. *J Bacteriol*. **192(21)**:5559-71.

32. Giangrossi, M., Brandi, A., Giuliodori, A.M., Gualerzi, C.O., Pon, C.L. (2007). Cold-shock-induced de novo transcription and translation of infA and role of IF1 during cold adaptation. *Mol Microbiol*. **64(3)**:807-21.

33. Philippe, C., Bénard, L., Eyermann, F., Cachia, C., Kirillov, S.V., Portier, C., Ehresmann, B., Ehresmann, C. (1994). Structural elements of rps0 mRNA involved in the modulation of translational initiation and regulation of E. coli ribosomal protein S15. *Nucleic Acids Res*. **22(13)**:2538-46.

34. Mogk, A., Huber, D., Bukau, B. (2011). Integrating protein homeostasis strategies in prokaryotes. *Cold Spring Harb Perspect Biol*. **3(4)**. pii: a004366.

35. Parida, B.K., Douglas, T., Nino, C., Dhandayuthapani, S. (2005). Interactions of anti-sigma factor antagonists of Mycobacterium tuberculosis in the yeast two-hybrid system. *Tuberculosis (Edinb)*. **85(5-6)**:347-55.

36. López-Rubio, J.J., Elías-Arnanz, M., Padmanabhan, S., Murillo, F.J. (2002). A repressor-antirepressor pair links two loci controlling light-induced carotenogenesis in Myxococcus xanthus. *J Biol Chem*. **277(9)**:7262-70.

37. Okano, H., Hwa, T., Lenz, P., Yan, D. (2010). Reversible adenylylation of glutamine synthetase is dynamically counterbalanced during steady-state growth of Escherichia coli. *J.Mol Biol*. **404(3)**:522-36.

38. He, C., Hus, J.C., Sun, L.J., Zhou, P., Norman, D.P., Dötsch, V., Wei, H., Gross, J.D., Lane, W.S., Wagner, G., Verdine, G.L. (2005). A methylation-dependent electrostatic switch controls DNA repair and transcriptional activation by E. coli ada. *Mol Cell*. **20(1)**:117-29.

39. Filtz, T.M., Vogel, W.K., Leid, M. (2014). Regulation of transcription factor activity by interconnected post-translational modifications. *Trends Pharmacol Sci*. **35(2)**:76-85.

40. Grant, G.A. (2012). Contrasting catalytic and allosteric mechanisms for phosphoglycerate dehydrogenases. *Arch Biochem Biophys*. **519(2)**:175-85.

41. Helmstaedt, K., Krappmann, S., Braus, G.H. (2001). Allosteric regulation of catalytic activity: Escherichia coli aspartate transcarbamoylase versus yeast chorismate mutase. *Microbiol Mol Biol Rev*. **65(3)**:404-21.

42. Wilson, C.J., Zhan, H., Swint-Kruse, L., Matthews, K.S. (2007). The lactose repressor system: paradigms for regulation, allosteric behavior and protein folding. *Cell Mol Life Sci*. **64(1)**:3-16.

43. Daber, R., Stayrook, S., Rosenberg, A., Lewis, M. (2007). Structural analysis of lac repressor bound to allosteric effectors. *J Mol Biol*. **370(4)**:609-19.

44. Yang, X., Lewis, P.J. (2010). The interaction between bacterial transcription factors and RNA polymerase during the transition from initiation to elongation. *Transcription*. **1(2)**:66-9.

45. Hong, E., Doucleff, M., Wemmer, D.E. (2009). Structure of the RNA polymerase core-binding domain of sigma(54) reveals a likely conformational fracture point. *J Mol Biol*. **390(1)**: 70–82.

46. Gourse, R.L., Ross, W., Gaal, T. (2000). UPs and downs in bacterial transcription initiation: the role of the alpha subunit of RNA polymerase in promoter recognition. *Mol Microbiol*. **37(4)**:687-95.

47. Estrem, S.T., Gaal T., Ross, W., Gourse, R.L. (1998). Identification of an UP element consensus sequence for bacterial promoters. *Proc Natl Acad Sci U S A*. **95(17)**:9761-6.

48. Mathew, R., Ramakanth, M., Chatterji, D. (2005). Deletion of the gene rpoZ, encoding the omega subunit of RNA polymerase, in Mycobacterium smegmatis results in fragmentation of the beta' subunit in the enzyme assembly. *J Bacteriol*. **187(18)**:6565-70.

49. Mathew, R., Mukherjee, R., Balachandar, R., Chatterji, D. (2006). Deletion of the rpoZ gene, encoding the omega subunit of RNA polymerase, results in pleiotropic surface-related phenotypes in Mycobacterium smegmatis. *Microbiology*. **152(Pt 6)**:1741-50.

50. Vuthoori, S., Bowers, C.W., McCracken, A., Dombroski, A.J., Hinton, D.M. (2001). Domain 1.1 of the sigma(70) subunit of Escherichia coli RNA polymerase modulates the formation of stable polymerase/promoter complexes. *J Mol Biol*. **309(3)**:561-72.

51. Phadtare, S., Severinov, K. (2005). Extended -10 motif is critical for activity of the cspA promoter but does not contribute to low-temperature transcription. *J Bacteriol*. **187(18)**:6584-9.

52. Helmann, J.D. (1995). Compilation and analysis of Bacillus subtilis sigma A-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA. *Nucleic Acids Res*. **23(13)**: 2351–2360.

53. Margeat, E., Kapanidis, A.N., Tinnefeld, P., Wang, Y., Mukhopadhyay, J., Ebright, R.H., Weiss, S. (2006). Direct observation of abortive initiation and promoter escape within single immobilized transcription complexes. *Biophys J*. **90(4)**:1419-31.

54. Dutta, D., Chalissery, J., Sen, R. (2008). Transcription termination factor rho prefers catalytically active elongation complexes for releasing RNA. *J Biol Chem*. **283(29)**:20243-51.

55. Kingsford, C.L., Ayanbule, K., Salzberg, S.L. (2007). Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol*. **8(2)**:R22.

56. Geiselmann, J., Wang, Y., Seifried, S.E., von Hippel, P.H. (1993). A physical model for the translocation and helicase activities of Escherichia coli transcription termination protein Rho. *Proc Natl Acad Sci U S A*. **90(16)**:7754-8.

57. Schmidt, T.R., Scott, E.J. 2nd, Dyer, D.W. (2011). Whole-genome phylogenies of the family Bacillaceae and expansion of the sigma factor gene family in the Bacillus cereus species-group. *BMC Genomics*. **12**:430.

58. Helmann, J.D. (2002). The extracytoplasmic function (ECF) sigma factors. *Adv Microb Physiol*. **46**:47-110.

59. Paget, M.S., Hong, H.J., Bibb, M.J., Buttner, M.J. (2002). The ECF sigma factors of *Streptomyces coelicolor* A3(2). In: Hodgson, D.A. and Thomas, C.M. (eds.) Signals, switches, regulons, and cascades: control of bacterial gene expression. Cambridge University Press, pp. 105-126. ISBN 9780521813884.

60. Leang, C., Krushkal, J., Ueki, T., Puljic, M., Sun, J., Juárez, K., Núñez, C., Reguera, G., DiDonato, R., Postier, B., Adkins, R.M., Lovley, D.R. (2009). Genome-wide analysis of the RpoN regulon in Geobacter sulfurreducens. *BMC Genomics*. **10**:331.

61. Rogozin, I.B., Makarova, K.S., Murvai, J., Czabarka, E., Wolf, Y.I., Tatusov, R.L., Szekely, L.A., Koonin, E.V. (2002). Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res*. **30(10)**:2212-23.

62. Lawrence, J.G., Roth, J.R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*. **143(4)**:1843-60.

63. Daber, R., Sharp, K., Lewis, M. (2009). One is not enough. *J Mol Biol*. **392(5)**:1133-44.

64. Winkler, W.C, Cohen-Chalamish, S., Breaker, R.R. (2002). An mRNA structure that controls gene expression by binding FMN. *Proc Natl Acad Sci U S A*. **99(25)**:15908-13.

65. Vitreschak, A.G., Rodionov, D.A., Mironov, A.A., Gelfand, M.S. (2003). Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. *RNA*. **9(9)**:1084-97.

66. Vitreschak, A.G., Rodionov, D.A., Mironov, A.A., Gelfand, M.S. (2002). Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. *Nucleic Acids Res*. **30(14)**:3141-51.

67. Nahvi, A., Barrick, J.E., Breaker, R.R. (2004). Coenzyme B12 riboswitches are widespread genetic control elements in prokaryotes. *Nucleic Acids Res*. **32(1)**:143-50.

68. Ames, T.D., Breaker, R.R. (2011). Bacterial aptamers that selectively bind glutamine. *RNA Biol*. **8(1)**:82-9.

69. Kim, J.N., Breaker, R.R. (2008). Purine sensing by riboswitches. *Biol Cell*. **100(1)**:1-11.

70. Winkler, W., Nahvi, A., Breaker, R.R. (2002). Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*. **419(6910)**:952-6.

71. Rodionov, D.A., Vitreschak, A.G., Mironov, A.A., Gelfand, M.S. (2002). Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. *J Biol Chem*. **277(50)**:48949-59.

72. Glatz, E., Persson, M., Rutberg, B. (1998). Antiterminator protein GlpP of Bacillus subtilis binds to glpD leader mRNA. *Microbiology*. **144(Pt 2)**:449-56.

73. Szigeti, R., Milescu, M., Gollnick, P. (2004). Regulation of the tryptophan biosynthetic genes in Bacillus halodurans: common elements but different strategies than those used by Bacillus subtilis. *J Bacteriol*. **186(3)**:818-28.

74. Du, H., Yakhnin, A.V., Dharmaraj, S., Babitzke, P. (2000). trp RNA-binding attenuation protein-5' stem-loop RNA interaction is required for proper transcription attenuation control of the Bacillus subtilis trpEDCFBA operon. *J Bacteriol*. **182(7)**:1819-27.

75. Itzkovitz, S., Tlusty, T., Alon, U. (2006). Coding limits on the number of transcription factors. *BMC Genomics*. **7**:239.

76. Rigali, S., Derouaux, A., Giannotta, F., Dusart, J. (2002). Subdivision of the helix-turn-helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies. *J Biol Chem*. **277(15)**:12507-15.

77. Pérez-Rueda, E., Collado-Vides, J. (2000). The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12. *Nucleic Acids Res*. **28(8)**:1838-47.

78. Martínez-Antonio, A., Collado-Vides, J. (2003). Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol*. **6(5)**:482-9.

79. Ravcheev, D.A., Gerasimova, A.V., Mironov, A.A., Gelfand, M.S. (2007). Comparative genomic analysis of regulation of anaerobic respiration in ten genomes from three families of gamma-proteobacteria (Enterobacteriaceae, Pasteurellaceae, Vibrionaceae). *BMC Genomics*. **8**:54.

80. Darwin, A.J., Stewart, V. (1995). Expression of the narX, narL, narP, and narQ genes of Escherichia coli K-12: regulation of the regulators. *J Bacteriol*. **177(13)**:3865-9.

81. Favorov, A.V., Gelfand, M.S., Gerasimova, A.V., Ravcheev, D.A., Mironov, A.A., Makeev, V.J. (2005). A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics*. **21(10)**:2240-5.

82. Noriega, C.E., Lin, H.Y., Chen, L.L., Williams, S.B., Stewart, V. (2010). Asymmetric cross-regulation between the nitrate-responsive NarX-NarL and NarQ-NarP two-component regulatory systems from Escherichia coli K-12. *Mol Microbiol*. **75(2)**:394-412.

83. Thieffry, D., Huerta, A.M., Pérez-Rueda, E., Collado-Vides, J. (1998). From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. *Bioessays*. **20(5)**:433-40.

84. Madar, D., Dekel, E., Bren, A., Alon, U. (2011). Negative auto-regulation increases the input dynamic-range of the arabinose system of Escherichia coli. *BMC Syst Biol*. **5**:111.

85. Dobrindt, U., Hacker, J. (2001). Whole genome plasticity in pathogenic bacteria. *Curr Opin Microbiol*. **4(5)**:550-7.

86. van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. *Trends Genet*. **19(9)**:479-84.

87. Molina, N., van Nimwegen, E. (2009). Scaling laws in functional genome content across prokaryotic clades and lifestyles. *Trends Genet*. **25(6)**:243-7.

88. Pérez-Rueda, E., Collado-Vides, J., Segovia, L. (2004). Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. *Comput Biol Chem*. **28(5-6)**:341-50.

89. Freemont, P.S., Lane, A.N., Sanderson, M.R. (1991). Structural aspects of protein-DNA recognition. *Biochem J*. **278(Pt 1)**:1-23.

90. Janczarek, M., Skorupska, A. (2007). The Rhizobium leguminosarum bv. trifolii RosR: transcriptional regulator involved in exopolysaccharide production. *Mol Plant Microbe Interact*. **20(7)**:867-81.

91. Reyrat, J.M., David, M., Batut, J., Boistard, P. (1994). FixL of Rhizobium meliloti enhances the transcriptional activity of a mutant FixJD54N protein by phosphorylation of an alternate residue. *J Bacteriol*. **176(7)**:1969-76.

92. Gao, R., Stock, A.M. (2015). Temporal hierarchy of gene expression mediated by transcription factor binding affinity and activation dynamics. *MBio*. **6(3)**:e00686-15.

93. Yang, C., Huang, T.W., Wen, S.Y., Chang, C.Y., Tsai, S.F., Wu, W.F., Chang, C.H. (2012). Genome-wide PhoB binding and gene expression profiles reveal the hierarchical gene regulatory network of phosphate starvation in Escherichia coli. *PLoS One*. **7(10)**:e47314.

94. Nellen-Anthamatten, D., Rossi, P., Preisig, O., Kullik, I., Babst, M., Fischer, H.M., Hennecke, H. (1998). Bradyrhizobium japonicum FixK2, a crucial distributor in the FixLJ-dependent regulatory cascade for control of genes inducible by low oxygen levels. *J Bacteriol*. **180(19)**:5251-5.

95. Bausch, C., Ramsey, M., Conway, T. (2004). Transcriptional organization and regulation of the L-idonic acid pathway (GntII system) in Escherichia coli. *J Bacteriol*. **186(5)**:1388-97.

96. Rodionov, D.A., Mironov, A.A., Rakhmaninova, A.B., Gelfand, M.S. (2000). Transcriptional regulation of transport and utilization systems for hexuronides, hexuronates and hexonates in gamma purple bacteria. *Mol Microbiol*. **38(4)**:673-83.

97. Lee, S.K., Newman, J.D., Keasling, J.D. (2005). Catabolite repression of the propionate catabolic genes in Escherichia coli and Salmonella enterica: evidence for involvement of the cyclic AMP receptor protein. *J Bacteriol*. **187(8)**:2793-800.

98. Weickert, M.J., Adhya, S. (1993). The galactose regulon of Escherichia coli. *Mol Microbiol*. **10(2)**:245-51.

99. Jin, D.J. (1994). Slippage synthesis at the galP2 promoter of Escherichia coli and its regulation by UTP concentration and cAMP.cAMP receptor protein. *J Biol Chem*. **269(25)**:17221-7.

100. Liu, M., Garges, S., Adhya, S. (2004). lacP1 promoter with an extended -10 motif. Pleiotropic effects of cyclic AMP protein at different steps of transcription initiation. *J Biol Chem*. **279(52)**:54552-7.

101. Shin, M., Kang, S., Hyun, S.J., Fujita, N., Ishihama, A., Valentin-Hansen, P., Choy, H.E. (2001). Repression of deoP2 in Escherichia coli by CytR: conversion of a transcription activator into a repressor. *EMBO J*. **20(19)**:5392-9.

102. Choy, H.E., Adhya, S. (1992). Control of gal transcription through DNA looping: inhibition of the initial transcribing complex. *Proc Natl Acad Sci U S A*. **89(23)**:11264-8.

103. Bintu, L., Buchler, N.E., Garcia, H.G., Gerland, U., Hwa, T., Kondev, J., Kuhlman, T., Phillips, R. (2005). Transcriptional regulation by the numbers: applications. *Curr Opin Genet Dev*. **15(2)**:125-35.

104. Sernova, N.V., Gelfand, M.S. (2012). Comparative genomics of CytR, an unusual member of the LacI family of transcription factors. *PLoS One*. **7(9)**:e44194.

105. Franza, T., Michaud-Soret, I., Piquerel, P., Expert, D. (2002). Coupling of iron assimilation and pectinolysis in Erwinia chrysanthemi 3937. *Mol Plant Microbe Interact*. **15(11)**:1181-91.

106. Mota, L.J., Sarmento, L.M., de Sá-Nogueira, I. (2001). Control of the arabinose regulon in Bacillus subtilis by AraR in vivo: crucial roles of operators, cooperativity, and DNA looping. *J Bacteriol*. **183(14)**:4190-201.

107. Oehler, S., Eismann, E.R., Krämer, H., Müller-Hill, B. (1990). The three operators of the lac operon cooperate in repression. *EMBO J*. **9(4)**:973-9.

108. Choy, H.E., Park, S.W., Aki, T., Parrack, P., Fujita, N., Ishihama, A., Adhya, S. (1995). Repression and activation of transcription by Gal and Lac repressors: involvement of alpha subunit of RNA polymerase. *EMBO J*. **14(18)**:4523-9.

109. Busby, S., Ebright, R.H. (1999). Transcription activation by catabolite activator protein (CAP). *J Mol Biol*. **293(2)**:199-213.

110. Lawson, C.L., Swigon, D., Murakami, K.S., Darst, S.A., Berman, H.M., Ebright, R.H. (2004). Catabolite activator protein: DNA binding and transcription activation. *Curr Opin Struct Biol*. **14(1)**:10-20.

111. Jackson, L., Blake, T., Green, J. (2004). Regulation of ndh expression in Escherichia coli by Fis. *Microbiology*. **150(Pt 2)**:407-13.

112. Makino, K., Amemura, M., Kim, S.K., Nakata, A., Shinagawa, H. (1993). Role of the sigma 70 subunit of RNA polymerase in transcriptional activation by activator protein PhoB in Escherichia coli. *Genes Dev*. **7(1)**:149-60.

113. Brown, N.L., Stoyanov, J.V., Kidd, S.P., Hobman, J.L. (2003). The MerR family of transcriptional regulators. *FEMS Microbiol Rev*. **27(2-3)**:145-63.

114. Watanabe, S., Kita, A., Kobayashi, K., Miki, K. (2008). Crystal structure of the [2Fe-2S] oxidative-stress sensor SoxR bound to DNA. *Proc Natl Acad Sci U S A*. **105(11)**:4121-6.

115. Newberry, K.J., Brennan, R.G. (2004). The structural mechanism for transcription activation by MerR family member multidrug transporter activation, N terminus. *J Biol Chem*. **279(19)**:20356-62.

116. Zharov, I.A., Gel'fand, M.S., Kazakov, A.E. (2011). Regulation of multidrug resistance genes by transcriptional factors from the BltR subfamily. *Mol Biol (Mosk)*. **45(4)**:715-23.

117. Outten, C.E., Outten, F.W., O'Halloran, T.V. (1999). DNA distortion mechanism for transcriptional activation by ZntR, a Zn(II)-responsive MerR homologue in Escherichia coli. *J Biol Chem.* **274(53)**:37517-24.

118. Jiang, D., Jarrett, H.W., Haskins, W.E. (2009). Methods for proteomic analysis of transcription factors. *J Chromatogr A*. **1216(41)**:6881-9.

119. Pagano, J.M., Clingman, C.C., Ryder, S.P. (2011). Quantitative approaches to monitor protein-nucleic acid interactions using fluorescent probes. *RNA*. **17(1)**:14-20.

120. Carey, M.F., Peterson, C.L., Smale, S.T. (2012). Experimental strategies for the identification of DNA-binding proteins. *Cold Spring Harb Protoc*. **2012(1)**:18-33.

121. Hampshire, A.J., Rusling, D.A., Broughton-Head, V.J., Fox, K.R. (2007). Footprinting: a method for determining the sequence selectivity, affinity and kinetics of DNA-binding ligands. *Methods*. **42(2)**:128-40.

122. Shortle, D., DiMaio, D., Nathans, D. (1981). Directed mutagenesis. *Annu Rev Genet*. **15**:265-94.

123. Yan, Z., Sun, X., Engelhardt, J.F. (2009). Progress and prospects: techniques for site-directed mutagenesis in animal models. *Gene Ther*. **16(5)**:581-8.

124. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol*. **2**:2006.0008.

125. Northrup, D.L., Zhao, K. (2011). Application of ChIP-Seq and related techniques to the study of immune function. *Immunity*. **34(6)**:830-42.

126. Kim, H., Kim, J., Selby, H., Gao, D., Tong, T., Phang, T.L., Tan, A.C. (2011). A short survey of computational analysis methods in analysing ChIP-seq data. *Hum Genomics*. **5(2)**:117-23.

127. Gilchrist, D.A., Fargo, D.C., Adelman, K. (2009). Using ChIP-chip and ChIP-seq to study the regulation of gene expression: genome-wide localization studies reveal widespread regulation of transcription elongation. *Methods*. **48(4)**:398-408.

128. Wu, J., Smith, L.T., Plass, C., Huang, T.H. (2006). ChIP-chip comes of age for genome-wide functional analysis. *Cancer Res*. **66(14)**:6899-902.

129. Zimmermann, B., Bilusic, I., Lorenz, C., Schroeder, R. (2010). Genomic SELEX: a discovery tool for genomic aptamers. *Methods*. **52(2)**:125-32.

130. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*. **96(8)**:4285-8.

131. Fleischmann, R.D, Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*. **269(5223)**:496-512.

132. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S.G. (1996). Life with 6000 genes. *Science*. **274(5287)**:546, 563-7.

133. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*. **40(Database issue)**:D109-14.

134. Kahn, S.D. (2011). On the future of genomic data. *Science*. **331(6018)**:728-9.

135. Badger, J.H., Olsen, G.J. (1999). CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol*. **16(4)**:512-24.

136. Frishman, D., Mironov, A., Mewes, H.W., Gelfand, M. (1998). Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res*. **26(12)**:2941-7.

137. Besemer, J., Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res*. **33(Web Server issue)**:W451-4.

138. Salzberg, S.L., Delcher, A.L., Kasif, S., White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*. **26(2)**:544-8.

139. Koonin, E.V., Galperin, M.Y. (1997). Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Curr Opin Genet Dev*. **7(6)**:757-63.

140. Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*. **39**:309-38.

141. Kummerfeld, S.K., Teichmann, S.A. (2006). DBD: a transcription factor prediction database. *Nucleic Acids Res*. **34(Database issue)**:D74-81.

142. Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*. **305(3)**:567-80.

143. Sadovskaya, N.S., Sutormin, R.A., Gelfand, M.S. (2006). Recognition of transmembrane segments in proteins: review and consistency-based benchmarking of internet servers. *J Bioinform Comput Biol*. **4(5)**:1033-56.

144. Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J., Sayers, E.W. (2012). GenBank. *Nucleic Acids Res*. **40(Database issue)**: D48–D53.

145. Emmert, D.B., Stoehr, P.J., Stoesser, G., Cameron, G.N. (1994). The European Bioinformatics Institute (EBI) databases. *Nucleic Acids Res*. **22(17)**:3445-9.

146. Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S. (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Res*. **33(Database issue)**:D154-9.

147. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B.A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P.S., Sigrist, C.J. (2008). The 20 years of PROSITE. *Nucleic Acids Res*. **36(Database issue)**:D245-9.

148. Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L., Eddy, S.R., Bateman, A., Finn, R.D. (2012). The Pfam protein families database. *Nucleic Acids Res*. **40(Database issue)**:D290-301.

149. Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., Bork P. (2000). SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res*. **28(1)**:231-4.

150. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25(17)**:3389-402.

151. Thompson, J.D., Gibson, T.J, Plewniak, F., Jeanmougin, F., Higgins, D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. **25(24)**:4876-82.

152. Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. **5**:113.

153. Bendtsen, J.D., Nielsen, H., von Heijne, G., Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*. **340(4)**:783-95.

154. Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*. **31(13)**:3406-15.

155. Osterman, A., Overbeek, R. (2003). Missing genes in metabolic pathways: a comparative genomics approach. *Curr Opin Chem Biol*. **7(2)**:238-51.

156. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*. **96(6)**:2896-901.

157. Huynen, M., Snel, B., Lathe, W. 3rd, Bork, P. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res*. **10(8)**:1204-10.

158. Dehal, P.S., Joachimiak, M.P., Price, M.N., Bates, J.T., Baumohl, J.K., Chivian, D., Friedland, G.D., Huang, K.H., Keller, K., Novichkov, P.S., Dubchak, I.L., Alm, E.J., Arkin, A.P. (2010). MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res*. **38(Database issue)**:D396-400.

159. Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Rückert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., Vonstein, V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*. **33(17)**:5691-702.

160. Snel, B., Lehmann, G., Bork, P., Huynen, M.A. (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res*. **28(18)**:3442-4.

161. Marrakchi, H., Choi, K.H., Rock, C.O. (2002). A new mechanism for anaerobic unsaturated fatty acid formation in Streptococcus pneumoniae. *J Biol Chem*. **277(47)**:44809-16.

162. Rodionov, D.A., Mironov, A.A., Gelfand, M.S. (2002). Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea. *Genome Res*. **12(10)**:1507-16.

163. Rodionov, D.A., Hebbeln, P., Gelfand, M.S., Eitinger, T. (2006). Comparative and functional genomic analysis of prokaryotic nickel and cobalt uptake transporters: evidence for a novel group of ATP-binding cassette transporters. *J Bacteriol*. **188(1)**:317-27.

164. Zhang, L., Leyn, S.A., Gu, Y., Jiang, W., Rodionov, D.A., Yang, C. (2012). Ribulokinase and transcriptional regulation of arabinose metabolism in Clostridium acetobutylicum. *J Bacteriol*. **194(5)**:1055-64.

165. Leyn, S.A., Gao, F., Yang, C., Rodionov, D.A. (2012). N-acetylgalactosamine utilization pathway and regulon in proteobacteria: genomic reconstruction and experimental characterization in Shewanella. *J Biol Chem*. **287(33)**:28047-56.

166. Rodionova, I.A., Li, X., Thiel, V., Stolyar, S., Stanton, K., Fredrickson, J.K., Bryant, D.A., Osterman, A.L., Best, A.A., Rodionov, D.A. (2013). Comparative genomics and functional analysis of rhamnose catabolic pathways and regulons in bacteria. *Front Microbiol*. **4**:407.

167. Gu, Y., Ding, Y., Ren, C., Sun, Z., Rodionov, D.A., Zhang, W., Yang, S., Yang, C., Jiang, W. (2010). Reconstruction of xylose utilization pathway and regulons in Firmicutes. *BMC Genomics*. **11**:255.

168. Hebbeln, P., Rodionov, D.A., Alfandega, A., Eitinger, T. (2007). Biotin uptake in prokaryotes by solute transporters with an optional ATP-binding cassette-containing module. *Proc Natl Acad Sci U S A*. **104(8)**:2909-14.

169. Yanai, I., Derti, A., DeLisi, C. (2001). Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc Natl Acad Sci U S A*. **98(14)**:7940-5.

170. Enright, A.J., Ouzounis, C.A. (2001). Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol*. **2(9)**:RESEARCH0034.

171. Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice D.W., Yeates T.O., Eisenberg D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*. **285(5428)**:751-3.

172. Chen, X., Guo, L., Fan, Z., Jiang, T. (2008). W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data. *Bioinformatics*. **24(9)**:1121-8.

173. Bailey, T.L., Williams, N., Misleh, C., Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*. **34(Web Server issue)**:W369-73.

174. Mironov, A.A., Vinokurova, N.P., Gel'fand, M.S. (2000). Software for analyzing bacterial genomes. *Mol Biol (Mosk)*. **34(2)**:253-62.

175. Mahony, S., Hendrix, D., Golden, A., Smith, T.J., Rokhsar, D.S. (2005). Transcription factor binding site identification using the self-organizing map. *Bioinformatics*. **21(9)**:1807-14.

176. Laikova, O.N., Mironov, A.A., Gelfand, M.S. (2001). Computational analysis of the transcriptional regulation of pentose utilization systems in the gamma subdivision of Proteobacteria. *FEMS Microbiol Lett*. **205(2)**:315-22.

177. Rodionov, D.A., Mironov, A.A., Gelfand, M.S. (2001). Transcriptional regulation of pentose utilisation systems in the Bacillus/Clostridium group of bacteria. *FEMS Microbiol Lett*. **205(2)**:305-14.

178. Zhang, L., Leyn, S.A., Gu, Y., Jiang, W., Rodionov, D.A., Yang, C. (2012). Ribulokinase and transcriptional regulation of arabinose metabolism in Clostridium acetobutylicum. *J Bacteriol*. **194(5)**:1055-64.

179. Rodionov, D.A., Rodionova, I.A., Li, X., Ravcheev, D.A., Tarasova, Y., Portnoy, V.A., Zengler, K., Osterman, A.L. (2013). Transcriptional regulation of the carbohydrate utilization network in Thermotoga maritima. *Front Microbiol*. **4**:244.

180. Ravcheev, D.A., Godzik, A., Osterman, A.L., Rodionov, D.A. (2013). Polysaccharides utilization in human gut bacterium Bacteroides thetaiotaomicron: comparative genomics reconstruction of metabolic and regulatory networks. *BMC Genomics*. **14**:873.

181. Panina, E.M., Vitreschak, A.G., Mironov, A.A., Gelfand, M.S. (2001). Regulation of aromatic amino acid biosynthesis in gamma-proteobacteria. *J Mol Microbiol Biotechnol*. **3(4)**:529-43.

182. Panina, E.M., Vitreschak, A.G., Mironov, A.A., Gelfand, M.S. (2003). Regulation of biosynthesis and transport of aromatic amino acids in low-GC Gram-positive bacteria. *FEMS Microbiol Lett*. **222(2)**:211-20.

183. Leyn, S.A., Suvorova, I.A., Kholina, T.D., Sherstneva, S.S., Novichkov, P.S., Gelfand, M.S., Rodionov, D.A. (2014). Comparative genomics of transcriptional regulation of methionine metabolism in Proteobacteria. *PLoS One*. **9(11)**:e113714.

184. Rodionov, D.A., Li, X., Rodionova, I.A., Yang, C., Sorci, L., Dervyn, E., Martynowski, D., Zhang, H., Gelfand, M.S., Osterman A.L. (2008). Transcriptional regulation of NAD metabolism in bacteria: genomic reconstruction of NiaR (YrxA) regulon. *Nucleic Acids Res*. **36(6)**:2032-46.

185. Permina, E.A., Gelfand, M.S. (2003). Heat shock (sigma32 and HrcA/CIRCE) regulons in beta-, gamma- and epsilon-proteobacteria. *J Mol Microbiol Biotechnol*. **6(3-4)**:174-81.

186. Permina, E.A., Kazakov, A.E., Kalinina, O.V., Gelfand, M.S. (2006). Comparative genomics of regulation of heavy metal resistance in Eubacteria. *BMC Microbiol*. **6**:49.

187. Kazakov, A.E., Rodionov, D.A., Alm, E., Arkin, A.P., Dubchak, I., Gelfand, M.S. (2009). Comparative genomics of regulation of fatty acid and branched-chain amino acid utilization in proteobacteria. *J Bacteriol*. **191(1)**:52-64.

188. Yang, C., Rodionov, D.A., Li, X., Laikova, O.N., Gelfand, M.S., Zagnitko, O.P., Romine, M.F., Obraztsova, A.Y., Nealson, K.H., Osterman, A.L. (2006). Comparative genomics and experimental characterization of N-acetylglucosamine utilization pathway of Shewanella oneidensis. *J Biol Chem*. **281(40)**:29872-85.

189. Panina, E.M., Mironov, A.A., Gelfand, M.S. (2003). Comparative genomics of bacterial zinc regulons: enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins. *Proc Natl Acad Sci U S A.* **100(17)**:9912-7.

190. Makarova, K.S., Ponomarev, V.A., Koonin, E.V. (2001). Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biol*. **2(9)**:RESEARCH 0033.

191. Nanamiya, H., Akanuma, G., Natori, Y., Murayama, R., Kosono, S., Kudo, T., Kobayashi, K., Ogasawara, N., Park, S.M., Ochi, K., Kawamura, F. (2004). Zinc is a key factor in controlling alternation of two types of L31 protein in the Bacillus subtilis ribosome. *Mol Microbiol*. **52(1)**:273-83.

192. Shin, J.H., Oh, S.Y., Kim, S.J., Roe, J.H. (2007). The zinc-responsive regulator Zur controls a zinc uptake system and some ribosomal proteins in Streptomyces coelicolor A3(2). *J Bacteriol*. **189(11)**:4070-7.

193. Lustig, B., Jernigan, R.L. (1995). Consistencies of individual DNA base amino acid interactions in structures and sequences. *Nucleic Acids Res*. **23(22)**:4707-11.

194. Morozov, A.V., Siggia, E.D. (2007). Connecting protein structure with predictions of regulatory sites. *Proc Natl Acad Sci U S A.* **104(17)**:7068-73.

195. Seeman, N.C., Rosenberg, J.M., Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci U S A.* **73(3)**:804-8.

196. Marabotti, A., Spyrakis, F., Facchiano, A., Cozzini, P., Alberti, S., Kellogg, G.E., Mozzarelli, A. (2008). Energy-based prediction of amino acid-nucleotide base recognition. *J Comput Chem*. **29(12)**:1955-69.

197. Gromiha, M.M., Fukui, K. (2011). Scoring function based approach for locating binding sites and understanding recognition mechanism of protein-DNA complexes. *J Chem Inf Model*. **51(3)**:721-9.

198. Morozov, A.V., Havranek, J.J., Baker, D., Siggia, E.D. (2005). Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res*. **33(18)**:5781-98.

199. Luscombe, N.M., Thornton, J.M. (2002). Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J Mol Biol*. **320(5)**:991-1009.

200. Mahony, S., Auron, P.E., Benos, P.V. (2007). Inferring protein-DNA dependencies using motif alignments and mutual information. *Bioinformatics*. **23(13)**:i297-304.

201. Zheng, M., Cooper, D.R., Grossoehme, N.E., Yu, M., Hung, L.W., Cieslik, M., Derewenda, U., Lesley, S.A., Wilson, I.A., Giedroc, D.P., Derewenda, Z.S. (2009). Structure of Thermotoga maritima TM0439: implications for the mechanism of bacterial GntR transcription regulators with Zn2+-binding FCD domains. *Acta Crystallogr D Biol Crystallogr*. **65(Pt 4)**:356-65.

202. Aravind, L., Anantharaman, V. (2003). HutC/FarR-like bacterial transcription factors of the GntR family contain a small molecule-binding domain of the chorismate lyase fold. *FEMS Microbiol Lett*. **222(1)**:17-23.

203. Rigali, S., Schlicht, M., Hoskisson, P., Nothaft, H., Merzbacher, M., Joris, B., Titgemeyer, F. (2004). Extending the classification of bacterial transcription factors beyond the helix–turn–helix motif as an alternative approach to discover new cis/trans relationships. *Nucleic Acids Res*. **32(11)**:3418-26.

204. König, B., Müller, J.J., Lanka, E., Heinemann, U. (2009). Crystal structure of KorA bound to operator DNA: insight into repressor cooperation in RP4 gene regulation. *Nucleic Acids Res*. **37(6)**:1915-24.

205. Brinkman, A.B., Ettema, T.J, de Vos, W.M., van der Oost, J. (2003). The Lrp family of transcriptional regulators. *Mol Microbiol*. **48(2)**:287-94.

206. Sharadamma, N., Khan, K., Kumar, S., Patil, K.N., Hasnain, S.E., Muniyappa, K. (2011). Synergy between the N-terminal and C-terminal domains of Mycobacterium tuberculosis HupB is essential for high-affinity binding, DNA supercoiling and inhibition of RecA-promoted strand exchange. *FEBS J*. **278(18)**:3447-62.

207. Wray, L.V. Jr, Fisher, S.H. (2008). Bacillus subtilis GlnR contains an autoinhibitory C-terminal domain required for the interaction with glutamine synthetase. *Mol Microbiol*. **68(2)**:277-85.

208. Wiethaus, J., Schubert, B., Pfänder, Y., Narberhaus, F., Masepohl, B. (2008). The GntR-like regulator TauR activates expression of taurine utilization genes in Rhodobacter capsulatus. *J Bacteriol*. **190(2)**:487-93.

209. Lee, M.H., Scherer, M., Rigali, S., Golden, J.W. (2003). PlmA, a new member of the GntR family, has plasmid maintenance functions in Anabaena sp. strain PCC 7120. *J Bacteriol*. **185(15)**:4315-25.

210. Franco, I.S., Mota, L.J., Soares, C.M., de Sá-Nogueira, I. (2006). Functional domains of the Bacillus subtilis transcription factor AraR and identification of amino acids important for nucleoprotein complex assembly and effector binding. *J Bacteriol*. **188(8)**:3024-36.

211. Franco, I.S., Mota, L.J., Soares, C.M., de Sá-Nogueira, I. (2007). Probing key DNA contacts in AraR-mediated transcriptional repression of the Bacillus subtilis arabinose regulon. *Nucleic Acids Res*. **35(14)**:4755-66.

212. Bramucci, E., Milano, T., Pascarella, S. (2011). Genomic distribution and heterogeneity of MocR-like transcriptional factors containing a domain belonging to the superfamily of the pyridoxal-5'-phosphate dependent enzymes of fold type I. *Biochem Biophys Res Commun*. **415(1)**:88-93.

213. Belitsky, B.R. (2004). Bacillus subtilis GabR, a protein with DNA-binding and aminotransferase domains, is a PLP-dependent transcriptional regulator. *J Mol Biol*. **340(4)**:655-64.

214. Edayathumangalam, R., Wu, R., Garcia, R., Wang, Y., Wang, W., Kreinbring, C.A., Bach, A., Liao, J., Stone, T.A., Terwilliger, T.C., Hoang, Q.Q., Belitsky, B.R., Petsko, G.A., Ringe, D., Liu, D. (2013). Crystal structure of Bacillus subtilis GabR, an autorepressor and transcriptional activator of gabT. *Proc Natl Acad Sci U S A*. **110(44)**:17820-5.

215. Magarvey, N., He, J., Aidoo, K.A., Vining, L.C. (2001). The pdx genetic marker adjacent to the chloramphenicol biosynthesis gene cluster in Streptomyces venezuelae ISP5230: functional characterization. *Microbiology*. **147(Pt 8)**:2103-12.

216. Jochmann, N., Götker, S., Tauch, A. (2011). Positive transcriptional control of the pyridoxal phosphate biosynthesis genes pdxST by the MocR-type regulator PdxR of Corynebacterium glutamicum ATCC 13032. *Microbiology*. **157(Pt 1)**:77-88.

217. Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C.F., Coburn, D., Newburger, D.E., Morris, Q., Hughes, T.R., Bulyk, M.L. (2009). Diversity and complexity in DNA recognition by transcription factors. *Science*. **324(5935)**:1720-3.

218. Quail, M.A., Dempsey, C.E., Guest, J.R. (1994). Identification of a fatty acyl responsive regulator (FarR) in Escherichia coli. *FEBS Lett*. **356(2-3)**:183-7.

219. Rodionov, D.A., Gelfand, M.S. (2006). Computational identification of BioR, a transcriptional regulator of biotin metabolism in Alphaproteobacteria, and of its binding signal. *FEMS Microbiol Lett*. **255(1)**:102-7.

220. Condemine, G., Berrier, C., Plumbridge, J., Ghazi, A. (2005). Function and expression of an N-acetylneuraminic acid-inducible outer membrane channel in Escherichia coli. *J Bacteriol*. **187(6)**:1959-65.

221. Xu, Y., Heath, R.J., Li, Z., Rock, C.O., White, S.W. (2001). The FadR.DNA complex. Transcriptional control of fatty acid metabolism in Escherichia coli. *J Biol Chem*. **276(20)**:17373-9.

222. van Aalten, D.M., DiRusso, C.C., Knudsen, J. (2001). The structural basis of acyl coenzyme A-dependent regulation of the transcription factor FadR. *EMBO J*. **20(8)**:2041-50.

223. Jain, D., Nair, D.T. (2013). Spacing between core recognition motifs determines relative orientation of AraR monomers on bipartite operators. *Nucleic Acids Res*. **41(1)**:639-47.

224. Hugouvieux-Cotte-Pattat, N., Robert-Baudouy, J. (1983). Regulation of expression of the uxu operon and of the uxuR regulatory gene in Escherichia coli K12. *J Gen Microbiol*. **129(11)**:3345-53.

225. Bates Utz, C., Nguyen, A.B., Smalley, D.J., Anderson, A.B., Conway, T. (2004). GntP is the Escherichia coli fructuronic acid transporter and belongs to the UxuR regulon. *J Bacteriol*. **186(22)**:7690-6.

226. Blanco, C., Mata-Gilsinger, M., Ritzenthaler, P. (1983). Construction of hybrid plasmids containing the Escherichia coli uxaB gene: analysis of its regulation and direction of transcription. *J Bacteriol*. **153(2)**:747-55.

227. Portalier, R., Robert-Baudouy, J., Stoeber, F. (1980). Regulation of Escherichia coli K-12 hexuronate system genes: exu regulon. *J Bacteriol*. **143(3)**:1095-107.

228. Mata-Gilsinger, M., Ritzenthaler, P. (1983). Physical mapping of the exuT and uxaC operators by use of exu plasmids and generation of deletion mutants in vitro. *J Bacteriol*. **155(3)**:973-82.

229. Blanco, C., Ritzenthaler, P., Mata-Gilsinger, M. (1986). Negative dominant mutations of the uidR gene in Escherichia coli: genetic proof for a cooperative regulation of uidA expression. *Genetics*. **112(2)**:173-82.

230. Novel, M., Novel, G. (1976). Regulation of beta-glucuronidase synthesis in Escherichia coli K-12: pleiotropic constitutive mutations affecting uxu and uidA expression. *J Bacteriol*. **127(1)**:418-32.

231. Ritzenthaler, P., Mata-Gilsinger, M. (1982). Use of in vitro gene fusions to study the *uxuR* regulatory gene in *Escherichia coli K-12*: direction of transcription and regulation of its expression. *J Bacteriol*. **150(3)**:1040-7.

232. Robert-Baudouy, J., Portalier, R., Stoeber, F. (1981). Regulation of hexuronate system genes in Escherichia coli K-12: multiple regulation of the uxu operon by exuR and uxuR gene products. *J Bacteriol*. **145(1)**:211-20.

233. Koo, J.H., Kim, Y.S. (1999). Functional evaluation of the genes involved in malonate decarboxylation by Acinetobacter calcoaceticus. *Eur J Biochem*. **266(2)**:683-90.

234. Kim, Y.S. (2002). Malonate metabolism: biochemistry, molecular biology, physiology, and industrial application. *J Biochem Mol Biol*. **35(5)**:443-51.

235. Lee, H.Y., An, J.H., Kim, Y.S. (2000). Identification and characterization of a novel transcriptional regulator, MatR, for malonate metabolism in Rhizobium leguminosarum bv. trifolii. *Eur J Biochem*. **267(24)**:7224-30.

236. Koo, J.H., Cho, I.H., Kim, Y.S. (2000). The malonate decarboxylase operon of Acinetobacter calcoaceticus KCCM 40902 is regulated by malonate and the transcriptional repressor MdcY. *J Bacteriol*. **182(22)**:6382-90.

237. Peng, H.L., Shiou, S.R., Chang, H.Y. (1999). Characterization of mdcR, a regulatory gene of the malonate catabolic system in Klebsiella pneumoniae. *J Bacteriol*. **181(7)**:2302-6.

238. Brämer, C.O., Steinbüchel, A. (2001). The methylcitric acid pathway in Ralstonia eutropha: new genes identified involved in propionate metabolism. *Microbiology*. **147(Pt 8)**:2203-14.

239. Brock, M., Maerker, C., Schütz, A., Völker, U., Buckel, W. (2002). Oxidation of propionate to pyruvate in Escherichia coli. Involvement of methylcitrate dehydratase and aconitase. *Eur J Biochem*. **269(24)**:6184-94.

240. Hubbard, P.A., Padovani, D., Labunska, T., Mahlstedt, S.A., Banerjee, R., Drennan, C.L. (2007). Crystal structure and mutagenesis of the metallochaperone MeaB. Insight into the causes of methylmalonic aciduria. *J Biol Chem*. **282(43)**:31308-16.

241. Meister, M., Saum, S., Alber, B.E., Fuchs, G. (2005). L-malyl-coenzyme A/beta-methylmalyl-coenzyme A lyase is involved in acetate assimilation of the isocitrate lyase-negative bacterium Rhodobacter capsulatus. *J Bacteriol*. **187(4)**:1415-25.

242. Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol*. **266**:418-27.

243. Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M., Rupp, R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*. **8**:460.

244. Novichkov, P.S., Rodionov, D.A., Stavrovskaya, E.D., Novichkova, E.S., Kazakov, A.E., Gelfand, M.S., Arkin, A.P., Mironov, A.A., Dubchak, I. (2010). RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucleic Acids Res*. **38(Web Server issue)**:W299-307.

245. Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E. (2004). WebLogo: a sequence logo generator. *Genome Res*. **14(6)**:1188-90.

246. Novichkov, P.S., Laikova, O.N., Novichkova, E.S., Gelfand, M.S., Arkin, A.P., Dubchak, I., Rodionov, D.A. (2010). RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucleic Acids Res*. **38(Database issue)**:D111-8.

247. Hill, T., Lewicki, P. (2007). STATISTICS: Methods and Applications. Tulsa, OK: StatSoft

248. Suvorova, I.A., Ravcheev, D.A., Gelfand, M.S. (2012). Regulation and evolution of malonate and propionate catabolism in proteobacteria. *J Bacteriol*. **194(12)**:3234-40.

249. Forward, J.A., Behrendt, M.C., Wyborn, N.R., Cross, R., Kelly, D.J. (1997). TRAP transporters: a new family of periplasmic solute transport systems encoded by the dctPQM genes of Rhodobacter capsulatus and by homologs in diverse gram-negative bacteria. *J Bacteriol*. **179(17)**:5482-93.

250. Kelly, D.J., Thomas, G.H. (2001). The tripartite ATP-independent periplasmic (TRAP) transporters of bacteria and archaea. *FEMS Microbiol Rev*. **25(4)**:405-24.

251. Thomas, G.H., Southworth, T., León-Kempis, M.R., Leech, A., Kelly, D.J. (2006). Novel ligands for the extracellular solute receptors of two bacterial TRAP transporters. *Microbiology*. **152(Pt 1)**:187-98.

252. Murray, E.L., Conway, T. (2005). Multiple regulators control expression of the Entner-Doudoroff aldolase (Eda) of Escherichia coli. *J Bacteriol*. **187(3)**:991-1000.

253. Rodionov, D.A., Gelfand, M.S., Hugouvieux-Cotte-Pattat, N. (2004). Comparative genomics of the KdgR regulon in Erwinia chrysanthemi 3937 and other gamma-proteobacteria. *Microbiology*. **150(Pt 11)**:3571-90.

254. Reed, J.L., Patel, T.R., Chen, K.H., Joyce, A.R., Applebee, M.K., Herring, C.D., Bui O.T., Knight E.M., Fong S.S., Palsson B.O. (2006). Systems approach to refining genome annotation. *Proc Natl Acad Sci U S A*. **103(46)**:17480-4.

255. Van Gijsegem, F., Wlodarczyk, A., Cornu, A., Reverchon, S., Hugouvieux-Cotte-Pattat, N. (2008). Analysis of the LacI family regulators of *Erwinia chrysanthemi* 3937, involvement in the bacterial phytopathogenicity. *Mol Plant Microbe Interact*. **21(11)**:1471-81.

256. Suvorova, I.A., Tutukina, M.N., Ravcheev, D.A., Rodionov, D.A., Ozoline, O.N., Gelfand, M.S. (2011). Comparative genomic analysis of the hexuronate metabolism genes and their regulation in gammaproteobacteria. *J Bacteriol*. **193(15)**:3956-63.

257. Chen, A.M., Wang, Y.B., Jie, S., Yu, A.Y., Luo, L., Yu, G.Q., Zhu, J.B., Wang, Y.Z. (2010). Identification of a TRAP transporter for malonate transport and its expression regulated by GtrA from Sinorhizobium meliloti. *Res Microbiol*. **161(7)**:556-64.

258. McNeil, M.B., Clulow, J.S., Wilf, N.M., Salmond, G.P., Fineran, P.C. (2012). SdhE is a conserved protein required for flavinylation of succinate dehydrogenase in bacteria. *J Biol Chem*. **287(22)**:18418-28.

259. Cai, H., Clarke, S. (1999). A novel methyltransferase catalyzes the methyl esterification of trans-aconitate in Escherichia coli. *J Biol Chem*. **274(19)**:13470-9.

Приложения

Приложение А. Список исследованных геномов и соответствующих аббревиатур

Геном	Аббревиатура
Acidiphilium cryptum JF-5	ACR
Acidothermus cellulolyticus 11B	ACE
Acidovorax avenae subsp. citrulli AAC00-1	AAV
Acinetobacter baumannii AYE	ABY
Acinetobacter sp. ADP1	ACI
Actinobacillus pleuropneumoniae AP76	APA
Actinobacillus succinogenes 130Z	ASU
Actinosynnema mirum DSM 43827	AMR
Aeromonas hydrophila subsp. hydrophila ATCC 7966	AHA
Aeromonas salmonicida subsp. salmonicida A449	ASA
Agrobacterium tumefaciens C58	ATU
Alcanivorax borkumensis SK2	ABO
Aliivibrio salmonicida LFI1238	VSA
Alkalilimnicola ehrlichei MLHE-1	AEH
Alkaliphilus metalliredigens QYMF	AMT
Alkaliphilus oremlandii OhILAs	AOE
Alteromonas macleodii 'Deep ecotype'	AMC
Anoxybacillus flavithermus WK1	AFL
Arthrobacter aurescens TC1	AAU
Arthrobacter sp. FB24	ART
Azoarcus sp. BH72	AZO
Azorhizobium caulinodans ORS 571	AZC
Azotobacter vinelandii AvOP	AVI
Bacillus amyloliquefaciens FZB42	BAY
Bacillus anthracis Ames	BAN
Bacillus cereus ATCC 14579	BCE
Bacillus cereus E33L	BCZ
Bacillus clausii KSM-K16	BCL
Bacillus halodurans C-125	BHA
Bacillus licheniformis ATCC 14580	BLI
Bacillus pumilus SAFR-032	BPU
Bacillus subtilis subtilis 168	BSU
Bacillus thuringiensis Al Hakam	BTL
Bacteroides thetaiotaomicron VPI-5482	BTH
Beutenbergia cavernae DSM 12333	BCV
Bordetella avium 197N	BAV
Bordetella bronchiseptica RB50	BBR
Bordetella parapertussis 12822	BPA
Bordetella petrii DSM 12804	BPT

Геном	Аббревиатура
Bradyrhizobium japonicum USDA110	BJA
Bradyrhizobium sp. BTAi1	BBT
Brucella abortus 9-941 (biovar 1)	BMB
Brucella canis ATCC 23365	BCS
Brucella melitensis 16M	BME
Brucella ovis ATCC 25840	BOV
Brucella suis 1330	BMS
Burkholderia ambifaria MC40-6	BAC
Burkholderia cenocepacia J2315	BCJ
Burkholderia cepacia AMMD	BAM
Burkholderia mallei ATCC 23344	BMA
Burkholderia multivorans ATCC 17616 (JGI)	BMU
Burkholderia phymatum STM815	BPH
Burkholderia phytofirmans PsJN	BPY
Burkholderia pseudomallei 1710b	BPM
Burkholderia sp. 383	BUR
Burkholderia thailandensis E264	BTE
Burkholderia vietnamiensis G4	BVI
Burkholderia xenovorans LB400	BXE
Caldicellulosiruptor saccharolyticus DSM 8903	CSC
Carboxydothermus hydrogenoformans Z-2901	CHY
Catenulispora acidiphila DSM 44928	CAF
Caulobacter crescentus CB15	CCR
Caulobacter sp. K31	CAK
Cellvibrio japonicus Ueda107	CJA
Chloroflexus aggregans DSM 9485	CAG
Chloroflexus aurantiacus J-10-fl	CAU
Chloroflexus sp. Y-400-fl	CHS
Chromobacterium violaceum ATCC 12472	CVI
Chromohalobacter salexigens DSM 3043	CSA
Citrobacter koseri ATCC BAA-895	CKO
Clavibacter michiganensis michiganensis NCPPB 382	CMI
Clostridium acetobutylicum ATCC 824	CAC
Clostridium beijerinckii NCIMB 8052	CBE
Clostridium botulinum F	CBF
Clostridium difficile 630	CDF
Clostridium kluyveri DSM 555	CKL
Clostridium novyi NT	CNO
Clostridium perfringens 13	CPE
Clostridium phytofermentans ISDg	CPY
Clostridium tetani E88	CTC
Clostridium thermocellum ATCC 27405	CTH
Colwellia psychrerythraea 34H	CPS
Conexibacter woesei DSM 14684	CWO

Геном	**Аббревиатура**
Corynebacterium aurimucosum ATCC 700975	CAR
Corynebacterium diphtheriae gravis NCTC13129	CDI
Corynebacterium efficiens YS-314	CEF
Corynebacterium glutamicum ATCC 13032	CGB
Corynebacterium glutamicum R	CGT
Corynebacterium jeikeium K411	CJK
Corynebacterium urealyticum DSM 7109	CUR
Cupriavidus taiwanensis LMG 19424	CTI
Dechloromonas aromatica RCB	DAR
Delftia acidovorans SPH-1	DAC
Desulfitobacterium hafniense Y51	DSY
Desulfotomaculum reducens MI-1	DRM
Desulfovibrio desulfuricans G20	DDE
Desulfovibrio vulgaris vulgaris Hildenborough	DVU
Dinoroseobacter shibae DFL 12	DSH
Edwardsiella tarda EIB202	ETD
Enterobacter sakazakii ATCC BAA-894	ESA
Enterobacter sp. 638	ENT
Enterococcus faecalis V583	EFA
Erwinia amylovora ATCC 49946	EAM
Erwinia carotovora atroseptica SCRI1043	ECA
Erwinia pyrifoliae Ep1/96	EPY
Erwinia tasmaniensis Et1/99	ETA
Escherichia coli K-12 MG1655	ECO
Escherichia fergusonii ATCC 35469	EFE
Fervidobacterium nodosum Rt17-B1	FNO
Finegoldia magna ATCC 29328	FMA
Frankia alni ACN14a	FAL
Frankia sp. EAN1pec	FRE
Geobacillus kaustophilus HTA426	GKA
Geobacillus thermodenitrificans NG80-2	GTN
Geobacter metallireducens GS-15	GME
Geobacter sulfurreducens PCA	GSU
Geobacter uraniumreducens Rf4	GUR
Gluconacetobacter diazotrophicus PAl 5	GDI
Gluconobacter oxydans 621H	GOX
Haemophilus ducreyi 35000HP	HDU
Haemophilus influenzae 86-028NP	HIT
Haemophilus somnus 129PT	HSO
Hahella chejuensis KCTC 2396	HCH
Halothermothrix orenii H 168	HOR
Hyphomonas neptunium ATCC 15444	HNE
Idiomarina loihiensis L2TR	ILO
Jannaschia sp. CCS1	JAN

Геном	Аббревиатура
Jonesia denitrificans DSM 20603	JDE
Kineococcus radiotolerans SRS30216	KRA
Klebsiella pneumoniae subsp. pneumoniae MGH 78578	KPN
Kosmotoga olearia TBF 19.5.1	KOL
Kribbella flavida DSM 17836	KFL
Lactobacillus acidophilus NCFM	LAC
Lactobacillus brevis ATCC 367	LBR
Lactobacillus casei ATCC 334	LCA
Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842	LDB
Lactobacillus gasseri ATCC 33323	LGA
Lactobacillus johnsonii NCC 533	LJO
Lactobacillus plantarum WCFS1	LPL
Lactobacillus sakei 23K	LSA
Lactococcus lactis subsp. cremoris SK11	LLC
Leifsonia xyli subsp. xyli CTCB07	LXX
Leptothrix cholodnii SP-6	LCH
Leuconostoc mesenteroides ATCC 8293	LME
Listeria innocua Clip11262	LIN
Listeria monocytogenes EGD-e	LMO
Listeria welshimeri SLCC5334	LWE
Lysinibacillus sphaericus C3-41	LSP
Macrococcus caseolyticus JCSC5402	MCC
Magnetospirillum magneticum AMB-1	MAG
Mannheimia succiniciproducens MBEL55E	MSU
Maricaulis maris MCS10	MMR
Marinobacter aquaeolei VT8	MAQ
Marinomonas sp. MWYL1	MMW
Mesorhizobium loti MAFF303099	MLO
Mesorhizobium sp. BNC1	MES
Methanocorpusculum labreanum Z	MLA
Methylibium petroleiphilum PM1	MPT
Methylobacterium chloromethanicum CM4	MCH
Methylobacterium extorquens PA1	MEX
Methylobacterium populi BJ001	MPO
Methylobacterium radiotolerans JCM2831	MRD
Methylobacterium sp. 4-46	MET
Methylocella silvestris BL2	MSL
Moorella thermoacetica ATCC 39073	MTA
Mycobacterium avium subsp. paratuberculosis k10	MPA
Mycobacterium bovis BCG Pasteur 1173P2	MBB
Mycobacterium smegmatis MC2 155	MSM
Mycobacterium sp. JLS	MJL
Mycobacterium tuberculosis H37Rv	MTU
Mycobacterium vanbaalenii PYR-1	MVA

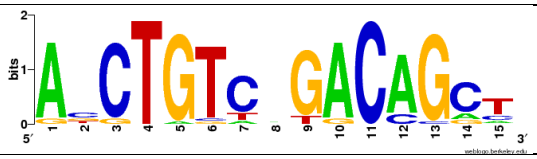
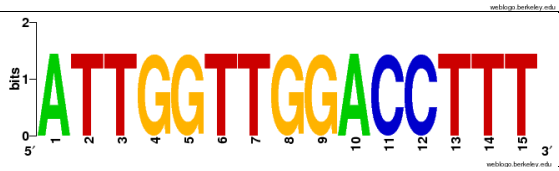
Геном	**Аббревиатура**
Nakamurella multipartita DSM 44233	NMU
Natranaerobius thermophilus JW/NM-WN-LF	NTH
Neisseria meningitidis MC58	NME
Nocardia farcinica IFM 10152	NFA
Nocardioides sp. JS614	NCA
Nostoc sp. PCC 7120	ANA
Novosphingobium aromaticivorans DSM 12444	NAR
Oceanobacillus iheyensis HTE831	OIH
Ochrobactrum anthropi ATCC 49188	OAN
Oenococcus oeni PSU-1	OOE
Oligotropha carboxidovorans OM5	OCA
Paenibacillus sp. JDR-2	PJR
Paracoccus denitrificans PD1222	PDE
Pasteurella multocida PM70	PMU
Pediococcus pentosaceus ATCC 25745	PPE
Pelotomaculum thermopropionicum SI	PTH
Petrotoga mobilis SJ95	PMO
Phenylobacterium zucineum HLK1	PZU
Photobacterium profundum SS9	PPR
Photorhabdus luminescens subsp. laumondii TTO1	PLU
Polaromonas naphthalenivorans CJ2	PNA
Polaromonas sp. JS666	POL
Propionibacterium acnes KPA171202	PAC
Proteus mirabilis HI4320	PMR
Pseudoalteromonas atlantica T6c	PAT
Pseudoalteromonas haloplanktis TAC125	PHA
Pseudomonas aeruginosa PA01	PAE
Pseudomonas entomophila L48	PEN
Pseudomonas fluorescens Pf-5	PFL
Pseudomonas fluorescens PfO-1	PFO
Pseudomonas mendocina ymp	PMY
Pseudomonas putida GB-1	PPG
Pseudomonas putida KT2440	PPU
Pseudomonas putida W619	PPW
Pseudomonas stutzeri A1501	PSA
Pseudomonas syringae pv. tomato DC3000	PST
Psychrobacter arcticum 273-4	PAR
Psychromonas ingrahamii 37	PIN
Ralstonia eutropha H16	REH
Ralstonia eutropha JMP134	REU
Ralstonia metallidurans CH34	RME
Ralstonia pickettii 12J	RPI
Ralstonia solanacearum GMI1000	RSO
Rhizobium etli CFN 42	RET

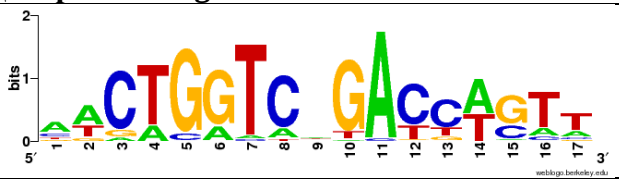
Геном	Аббревиатура
Rhizobium etli CIAT 652	REC
Rhizobium leguminosarum bv. viciae 3841	RLE
Rhodobacter sphaeroides 2.4.1	RSP
Rhodococcus sp. RHA1	RHA
Rhodoferax ferrireducens DSM 15236	RFR
Rhodospirillum centenum SW	RCE
Rhodospirillum rubrum ATCC 11170	RRU
Roseiflexus castenholzii DSM13941	RCA
Roseiflexus sp. RS-1	RRS
Roseobacter denitrificans OCh 114	RDE
Rubrobacter xylanophilus DSM 9941	RXY
Saccharophagus degradans 2-40	SDE
Saccharopolyspora erythraea NRRL 2338	SER
Salinispora tropica CNB-440	STP
Salmonella enterica serovar Typhi CT18	STY
Salmonella typhimurium LT2	STM
Serratia proteamaculans 568	SPE
Shewanella amazonensis SB2B	SAZ
Shewanella baltica OS155	SBL
Shewanella denitrificans OS217	SDN
Shewanella frigidimarina NCIMB 400	SFR
Shewanella halifaxensis HAW-EB4	SHL
Shewanella loihica PV-4	SLO
Shewanella oneidensis MR-1	SON
Shewanella pealeana ATCC 700345	SPL
Shewanella piezotolerans WP3	SWP
Shewanella putrefaciens CN-32	SPC
Shewanella sediminis HAW-EB3	SSE
Shewanella sp. ANA-3	SHN
Shewanella woodyi ATCC51908	SWD
Shigella boydii Sb227	SBO
Shigella dysenteriae Sd197	SDY
Shigella flexneri 2a str. 301	SFL
Shigella sonnei Ss046	SSN
Sinorhizobium medicae WSM419	SMD
Sinorhizobium meliloti 1021	SME
Sodalis glossinidius morsitans	SGL
Sphaerobacter thermophilus DSM 20745	SPT
Sphingomonas wittichii RW1	SWI
Sphingopyxis alaskensis RB2256	SAL
Staphylococcus aureus subsp. aureus Mu50	SAV
Staphylococcus capitis SK14	SCP
Staphylococcus carnosus subsp. carnosus TM300	SCA
Staphylococcus epidermidis ATCC 12228	SEP

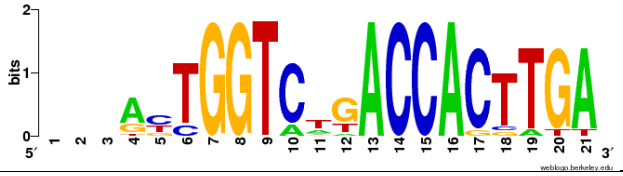
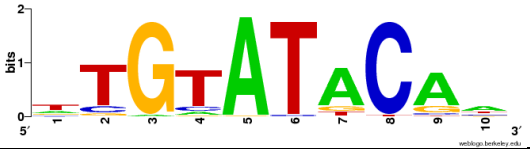
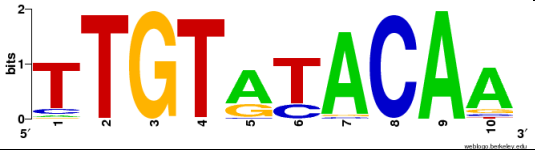
Геном	Аббревиатура
Staphylococcus haemolyticus JCSC1435	SHA
Staphylococcus lugdunensis HKU09-01	SLG
Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305	SSP
Stenotrophomonas maltophilia R551-3	SMT
Streptococcus agalactiae 2603V/R	SAG
Streptococcus equi subsp. zooepidemicus MGCS10565	SEZ
Streptococcus mutans UA159	SMU
Streptococcus pneumoniae TIGR4	SPN
Streptococcus pyogenes M1 GAS	SPY
Streptococcus sanguinis SK36	SSA
Streptococcus suis 05ZYH33	SSU
Streptococcus thermophilus LMD-9	STE
Streptococcus uberis 0140J	SUB
Streptomyces avermitilis MA-4680	SMA
Streptomyces coelicolor A3(2)	SCO
Streptomyces griseus subsp. griseus NBRC 13350	SGR
Symbiobacterium thermophilum IAM 14863	STH
Syntrophomonas wolfei subsp. wolfei str. Goettingen	SWO
Thauera sp. MZ1T	TMZ
Thermoanaerobacter pseudethanolicus ATCC 33223	TPD
Thermoanaerobacter tengcongensis MB4	TTE
Thermobifida fusca YX	TFU
Thermosipho africanus TCF52B	TAF
Thermosipho melanesiensis BI429	TME
Thermotoga lettingae TMO	TLE
Thermotoga maritima MSB8	TMA
Thermotoga neapolitana DSM 4359	TNE
Thermotoga petrophila RKU-1	TPT
Thiomicrospira crunogena XCL-2	TCX
Verminephrobacter eiseniae EF01-2	VEI
Vibrio cholerae O1 eltor N16961	VCH
Vibrio fischeri ES114	VFI
Vibrio fischeri MJ11	VFM
Vibrio harveyi ATCC BAA-1116	VHA
Vibrio parahaemolyticus RIMD 2210633	VPA
Vibrio splendidus LGP32	VSP
Vibrio vulnificus CMCP6	VVU
Vibrio vulnificus YJ016	VVY
Xanthomonas axonopodis pv. citri 306	XAC
Xanthomonas campestris pv. campestris ATCC 33913	XCC
Xanthomonas campestris pv. vesicatoria 85-10	XCV
Xanthomonas oryzae pv. oryzae KACC10331	XOO
Xanthomonas oryzae pv. oryzae MAFF 311018	XOM
Xanthomonas oryzae PXO99A	XOP

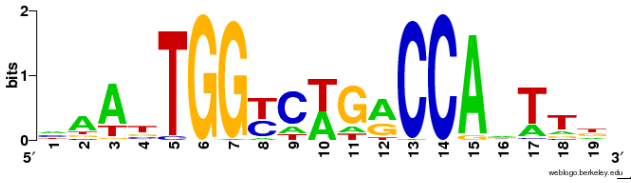
Геном	Аббревиатура
Xylanimonas cellulosilytica DSM 15894	XCE
Yersinia enterocolitica subsp. enterocolitica 8081	YEN
Yersinia pestis KIM	YPK
Yersinia pseudotuberculosis IP 32953	YPS

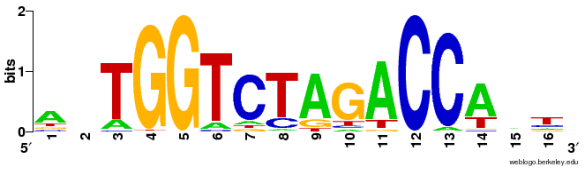
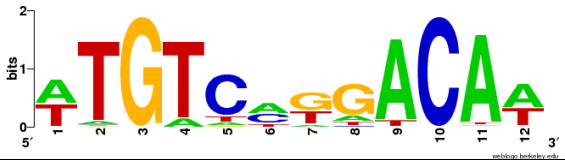
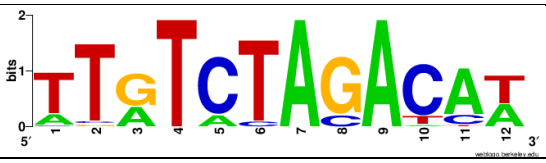
Приложение Б. Диаграммы Logo мотивов связывания исследованных транскрипционных факторов семейства GntR

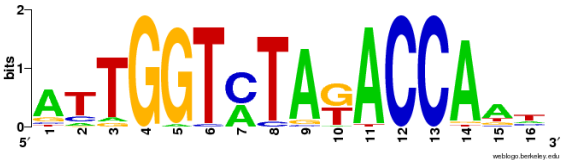
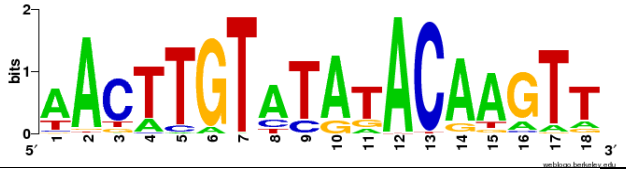
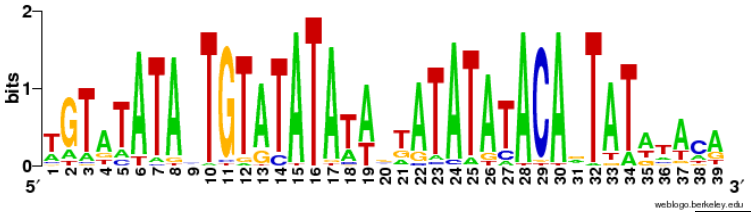
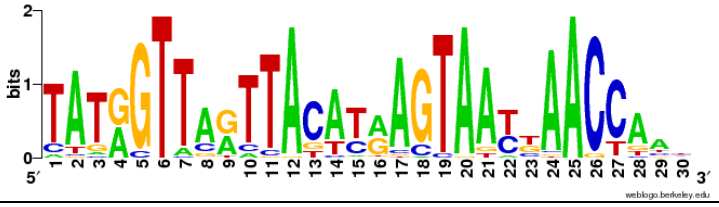
Ортологическая группа	Диаграмма Logo мотива связывания
Подсемейство FadR	
AnsR	
Arth_3541	
BCAM2278	
Bpet4349	
Bxe_A4089	
DgoR	
DVU2644	
DVU2802	
ExuR	

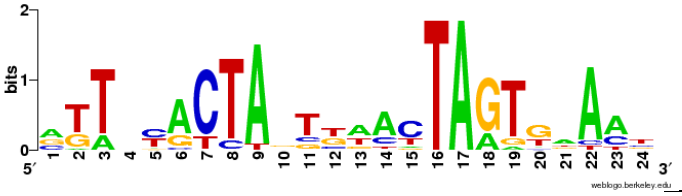
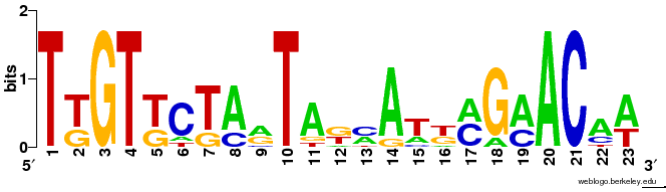
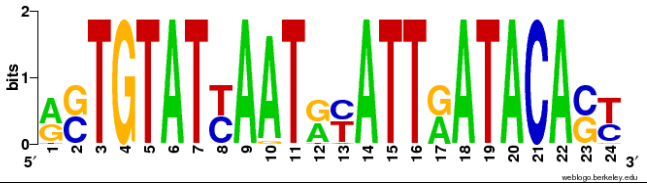
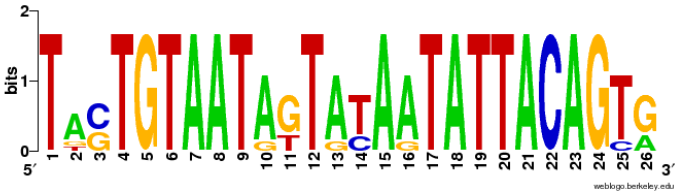
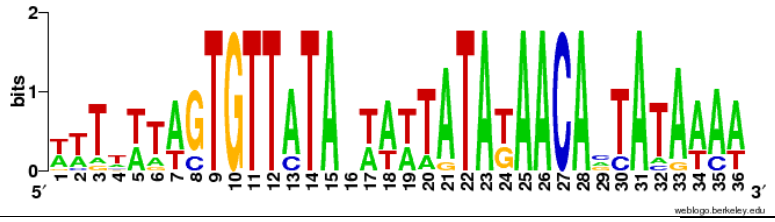
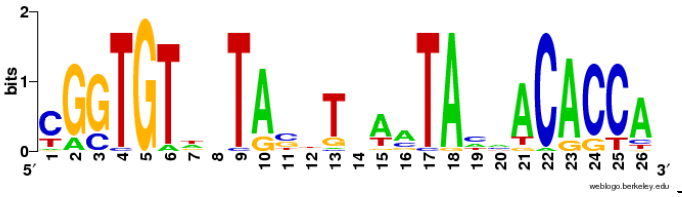
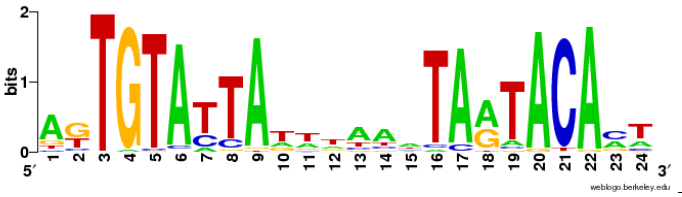
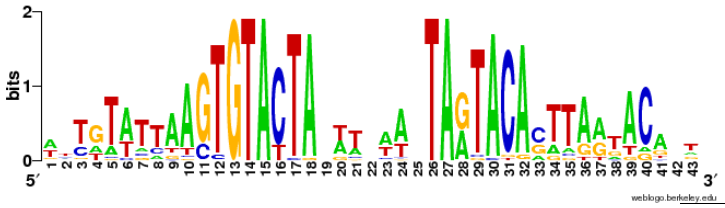
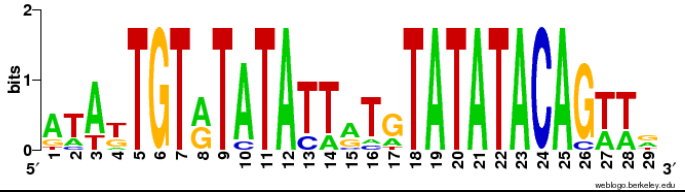
Ортологическая группа	Диаграмма Logo мотива связывания
FadR	
GlcC	
GlcC2	
GntR	
GntR2	
GudR	
HpxS	
HypR	
HypR2	
HypR3	

Ортологическая группа	Диаграмма Logo мотива связывания
HyuR	
LldR	
LldR2	
Mce2R	
MdcY	
NtaR	
PA1520	
PdhR	
PhnR	
PrpR	

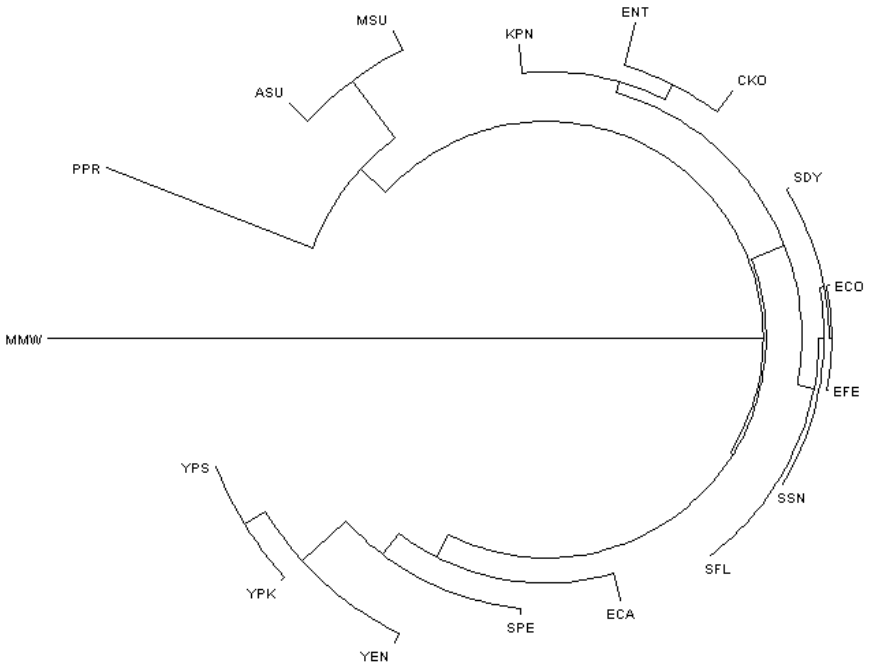
Ортологическая группа	Диаграмма Logo мотива связывания
SCO1410	
SdaR	
UraR	
UreR	
UxuR	
UxuR2	
UxuR3	
Подсемейство HutC	
AgaR	
AgaR2	
AgaR3	

Ортологическая группа	Диаграмма Logo мотива связывания
DasR	
GamR	
GmuR	
HutC Alphaproteobacteria	
HutC Betaproteobacteria	
HutC Gammaproteobacteria + Actinobacteria	
IolR	
ManR	
NagR	
PhnF	

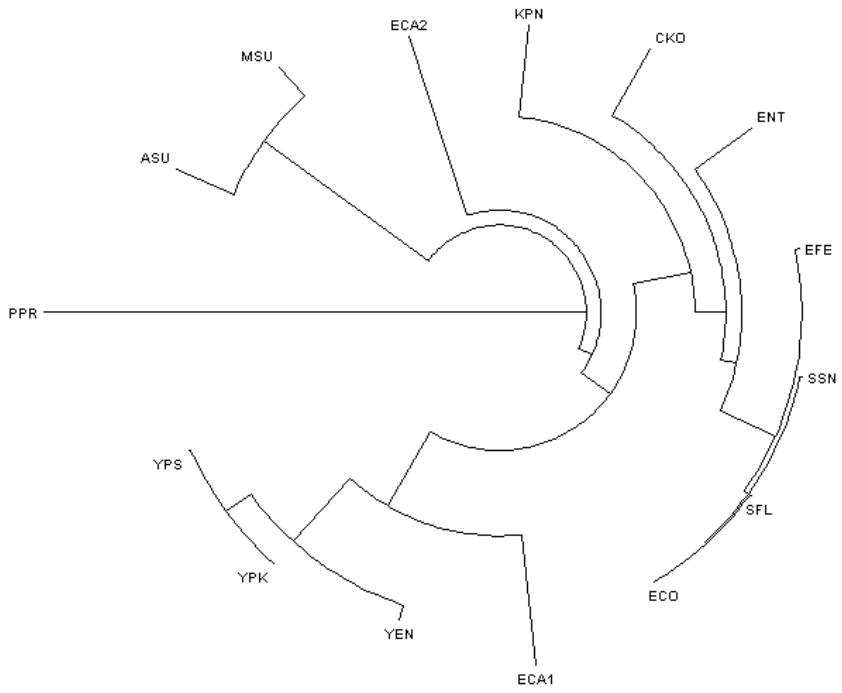
Ортологическая группа	Диаграмма Logo мотива связывания
PhnF Betaproteobacteria	
PhnF2	
PhnR	
PhnR2	
TreR	
YihL	
Подсемейство YtrA	
BH0651	
BH1164	
EF1676	

Ортологическая группа	Диаграмма Logo мотива связывания
LndYR	
SCO3812	
SO0072	
TM0766	
TTE0438	
XAC1548	
YhcF	
YtrA	
YtrA2 (Staphylococcus)	

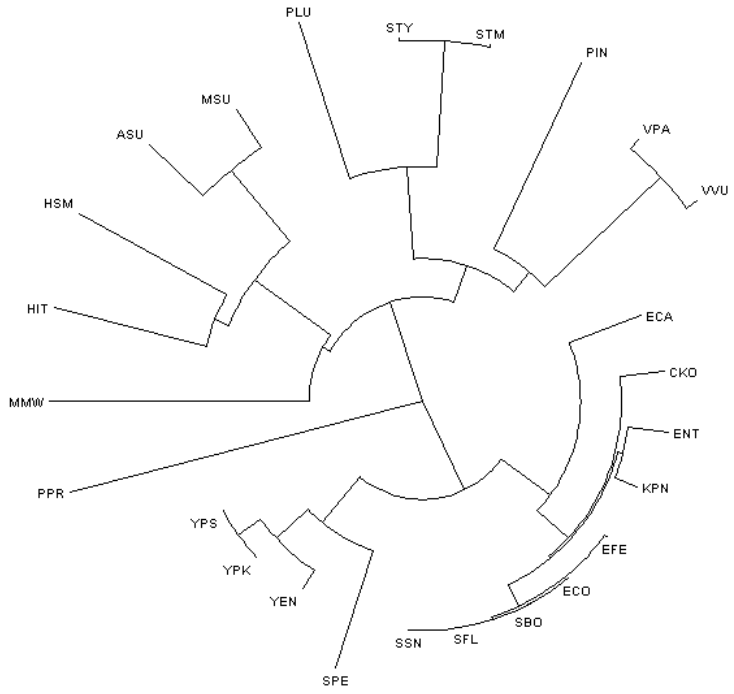
Приложение В. Филогенетические деревья транспортеров и ферментов метаболизма гексуронатов у Gammaproteobacteria. Трехбуквенные обозначения геномов соответствуют аббревиатурам, приведенным в Приложении А.



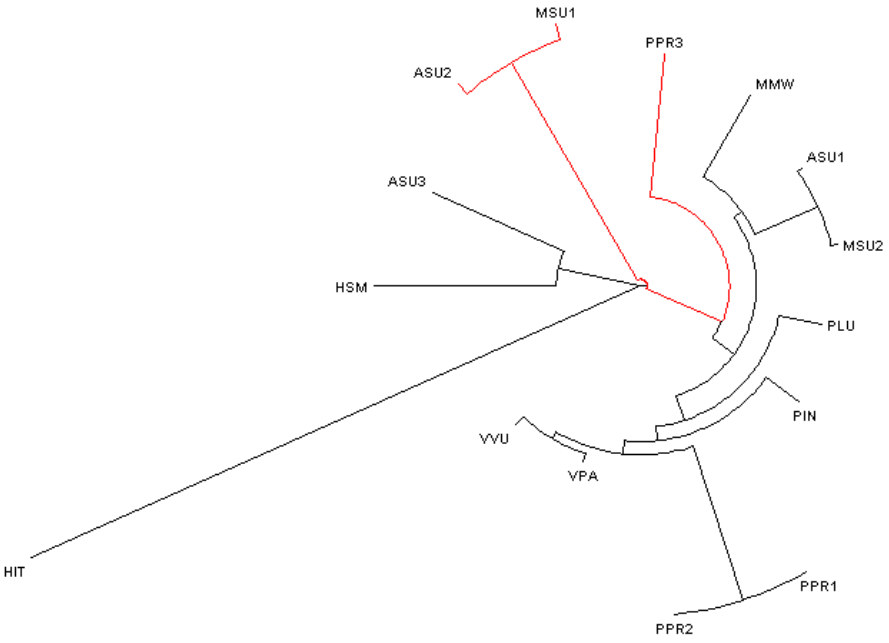
В1. Филогенетическое дерево D-альтронат гидролаз UxaA



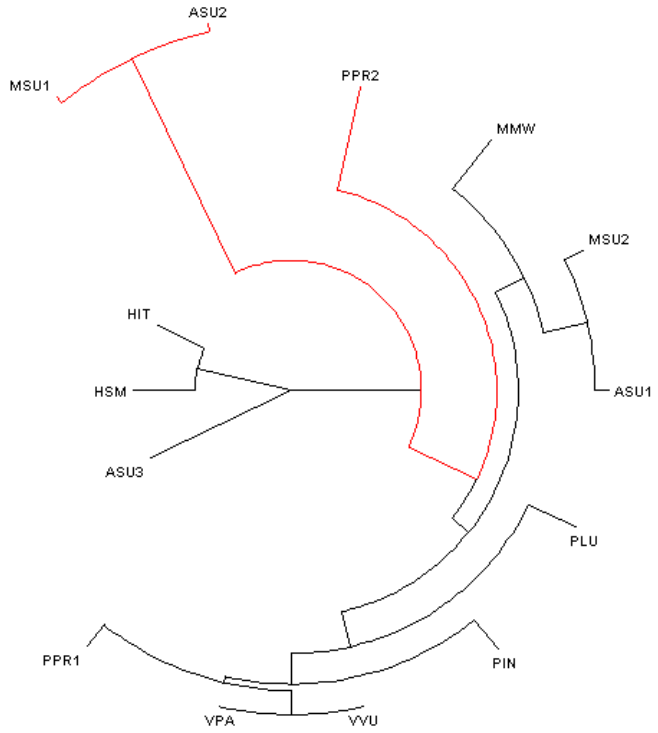
В2. Филогенетическое дерево D-альтронат оксидоредуктаз UxaB



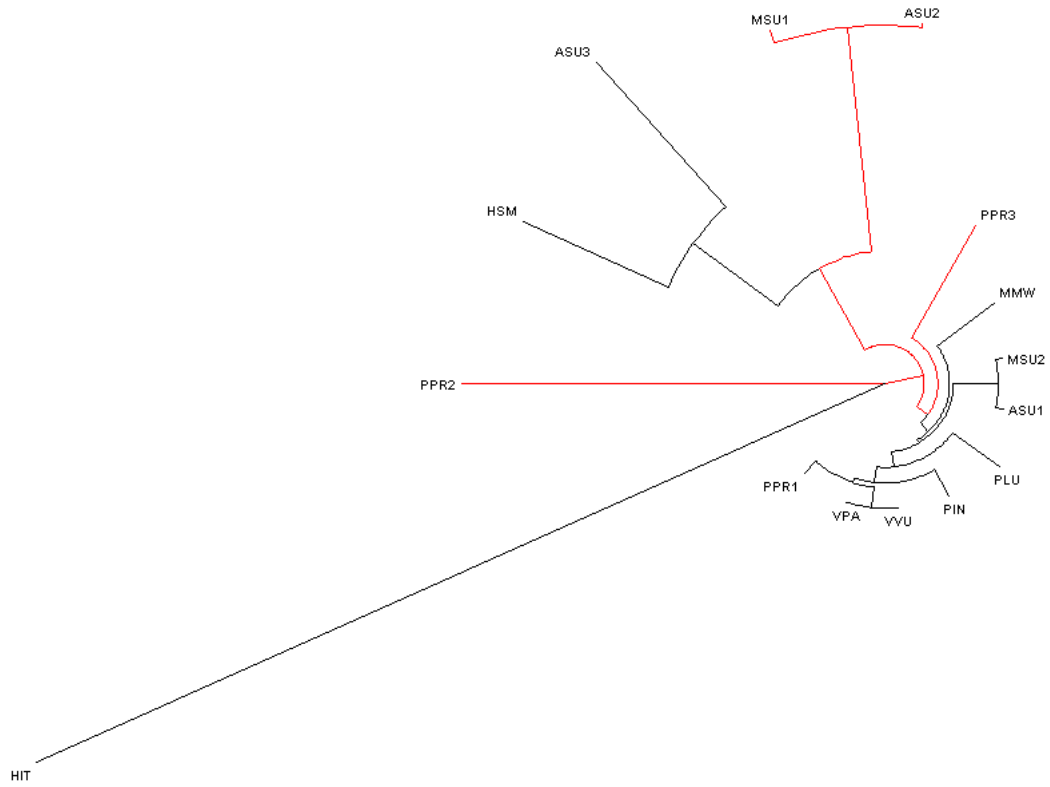
B3. Филогенетическое дерево D-глюкуронат/D-галактуронат изомераз UxaC



B4. Филогенетическое дерево субъединиц UxuP TRAP транспортера гексуронатов

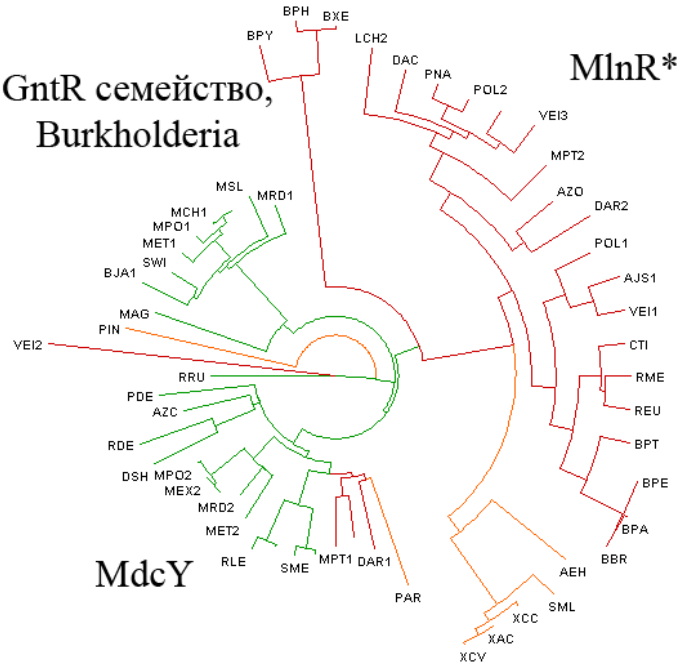


B5. Филогенетическое дерево субъединиц UxuQ TRAP транспортера гексуронатов

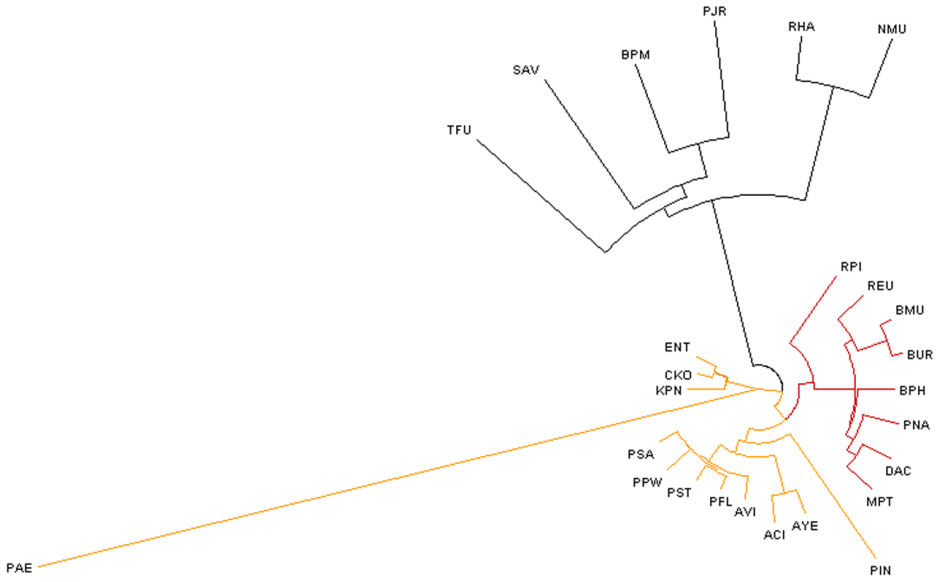


B6. Филогенетическое дерево субъединиц UxuM TRAP транспортера гексуронатов

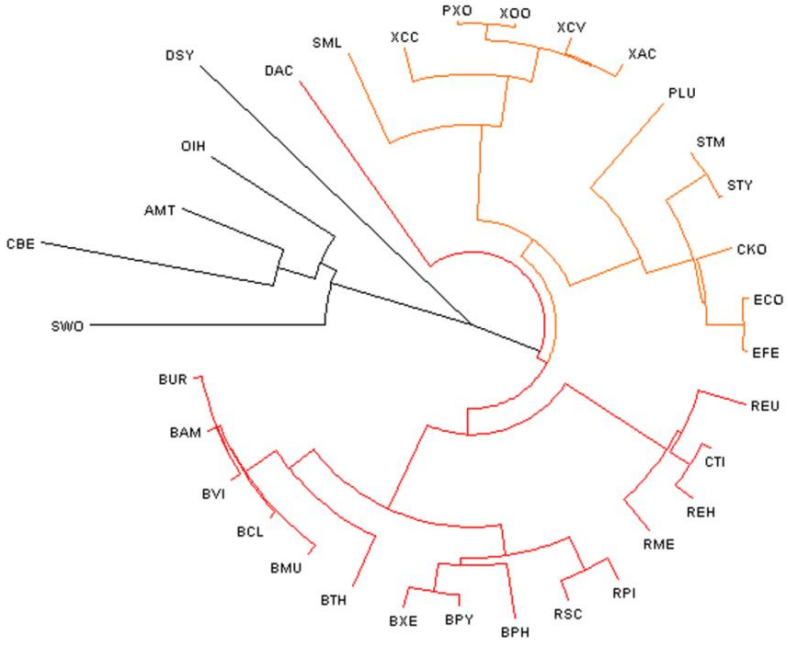
Приложение Г. Филогенетические деревья транскрипционных регуляторов метаболизма малоната и пропионата у Proteobacteria. Представители Alphaproteobacteria показаны зеленым, Betaproteobacteria – красным, Gammaproteobacteria – оранжевым, Deltaproteobacteria – фиолетовым. Внешняя группа показана черным. Трехбуквенные обозначения геномов соответствуют аббревиатурам, приведенным в Приложении А.



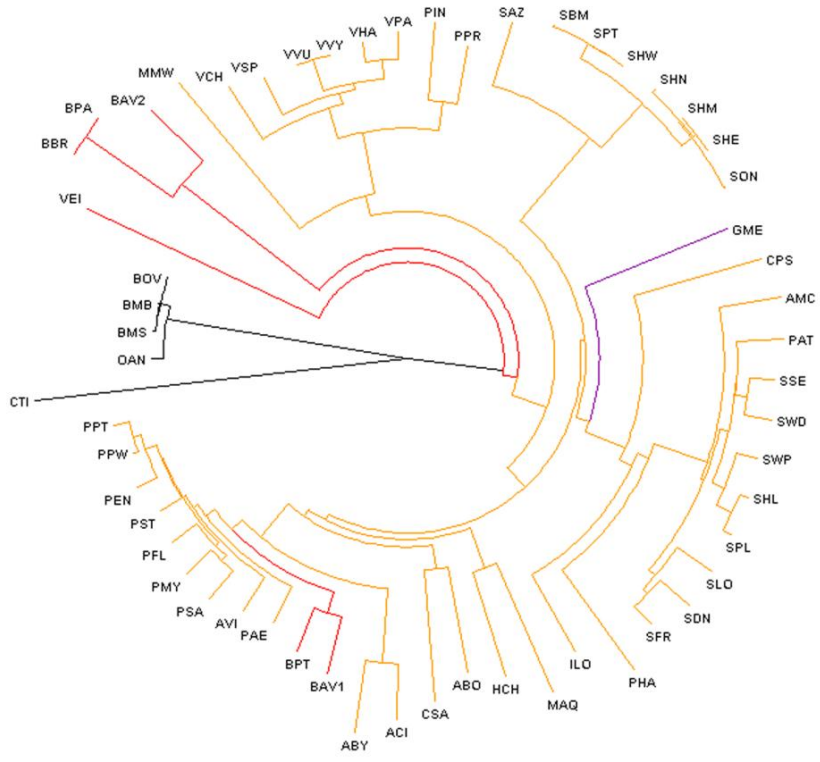
Г1. Филогенетическое дерево транскрипционных факторов MdcY, MlnR* и специфических для Burkholderia транскрипционных факторов семейства GntR



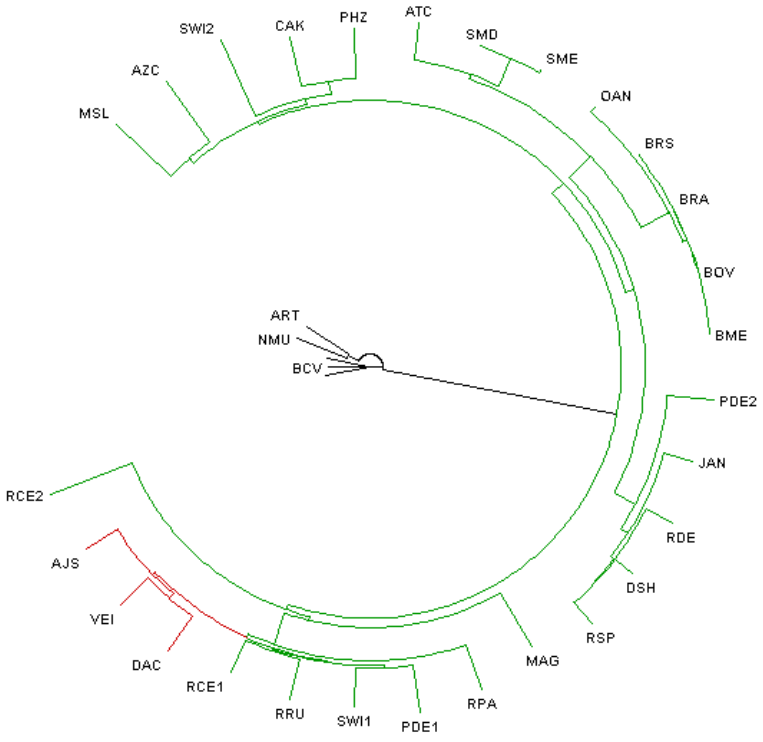
Г2. Филогенетическое дерево транскрипционных факторов MdcR



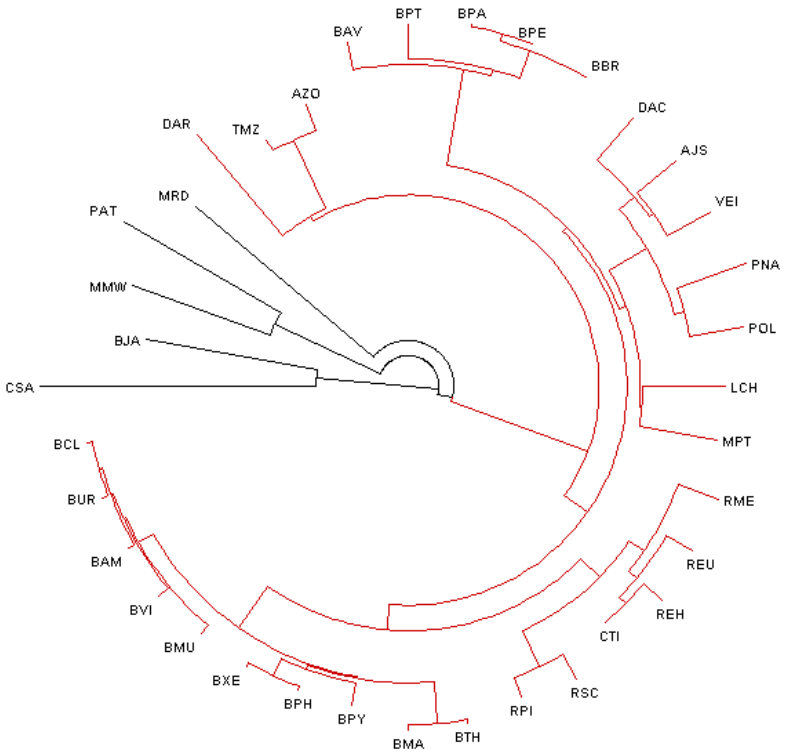
Г3. Филогенетическое дерево транскрипционных факторов PrpR



Г4. Филогенетическое дерево транскрипционных факторов PrpR*



Г5. Филогенетическое дерево транскрипционных факторов PrpQ*



Г6. Филогенетическое дерево транскрипционных факторов SdhR*