На правах рукописи

ЦОЙ ОЛЬГА ВЛАДИСЛАВОВНА

Эволюция систем регуляции транскрипции в геномах бактерий

03.01.09 - математическая биология, биоинформатика

Автореферат
диссертации на соискание ученой степени
кандидата биологических наук

Москва, 2014

Работа выполнена на факультете биоинженерии и биоинформатики Московского государственного университета им. М.В. Ломоносова и в Учебно-научном центре „Биоинформатика“ Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук.

Научный руководитель – Гельфанд Михаил Сергеевич
доктор биологических наук,
кандидат физико-математических наук, профессор,
заведующий Учебно-научным центром «Биоинформатика»,
заместитель директора ИППИ РАН по научным вопросам

Официальные оппоненты:

Самсонова Мария Георгиевна
доктор биологических наук, профессор
начальник отдела компьютерной биологии

Кафедра прикладной математики и Центр перспективных исследований Федерального государственного автономного образовательного учреждения высшего образования «Санкт-Петербургский государственный политехнический университет»

Кулаковский Иван Владимирович
кандидат физико-математических наук,
старший научный сотрудник лаборатории вычислительных методов системной биологии

Федеральное государственное бюджетное учреждение науки Институт молекулярной биологии имени В.А. Энгельгардта Российской академии наук

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

Защита диссертации состоится __ _____ 201_ года в 14 часов на заседании диссертационного совета Д 002.077.04 при Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук по адресу: 127994, г. Москва, ГСП-4, Большой Каретный переулок, д. 19, стр. 1.

С текстом автореферата и диссертации можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук, а также на сайте ИППИ РАН по адресу www.iitp.ru

Автореферат разослан __ _____ 201_ года

Ученый секретарь диссертационного совета

доктор биологических наук, профессор Г.И. Рожкова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Введение

Актуальность темы исследования

Многие системы в живых организмах образуют биологические сети. Примером такой сети является и система регуляции транскрипции. Самый нижний уровень этой сети составляют наборы взаимодействующих базовых элементов, состоящие из транскрипционного фактора, участка ДНК, с которым он связывается, и регулируемого гена. Объединяясь, на следующем уровне эти наборы образуют регуляторные блоки. Один или несколько регуляторных блоков образуют функциональный модуль, а несколько функциональных модулей составляют всю транскрипционную сеть. Растущее количество данных и экспериментальных методик позволяют изучать организмы в контексте целых сетей, модулей и блоков, а не отдельных функциональных систем. В настоящей работе мы анализировали эту систему на уровне блоков и наборов базовых элементов методами сравнительной геномики.

Методы сравнительной геномики позволяют реконструировать транскрипционную сеть в наборе родственных геномов. Поиск новых регуляторных элементов в бактериальных геномах является одним из наиболее сложных этапов ее изучения. В самом простом случае существуют экспериментальные данные хотя бы для одного генома, опираясь на которые становится возможным изучение других, родственных организмов. Но более частыми бывают ситуации, когда ни локализация, ни структура регуляторных элементов не известна.

Постановка и решение вопросов, связанных с транскрипционной сетью бактериальных генов, необходима для детальной функциональной аннотации генов; реконструкции метаболических путей – не только универсальных, но, в большей степени, таксон-специфичных; изучения роли этих путей в организме; а также исследования эволюции регуляторных систем и организмов в целом. Изучение принципов, лежащих в основе организации регуляторных сетей – неотъемлемая часть понимания молекулярных механизмов жизнедеятельности, в том числе, патогенных бактерий, биотехнологических штаммов, организмов, применяемых в системах биоочистки и.т.п.

Степень разработанности темы

В настоящий момент изучению систем регуляции транскрипции посвящено много работ, но они ограничены изучением нескольких модельных организмов (например, *Escherichia* coli и *Bacillus subtilis*) или проведены только на уровне отдельных функциональных систем. В то же время в базе данных Genbank содержится несколько тысяч полных последовательностей бактериальных геномов разной степени родства, и еще больше находится в процессе определения последовательности или аннотации. Рост числа геномов делает изучение индивидуальных геномов слишком трудоемким, но, с другой стороны, такое количество информации позволяет реконструировать биологические системы (в том числе, транскрипционные сети) в немодельных организмах, и их изучение представляется интересным уже не просто в рамках отдельных организмов, а в ряду близко родственных геномов, что позволяет проследить за особенностями их эволюции. Для подобного анализа требуется применение компьютерных, а не экспериментальных, методов.

Цели и задачи исследования

Целью настоящего исследования была реконструкция и анализ сетей транскрипционных регуляторных элементов в близкородственных геномах с помощью методов компьютерной биологии для немодельных организмов. В рамках поставленной цели решались следующие общие и частные задачи:

Предсказание регуляторных элементов транскрипционной сети *de novo* на примере пути утилизации этаноламина.

Изучение эволюции регуляторных блоков на примере блока треугольник.

Научная новизна

В настоящей работе были обобщены сведения о закономерностях эволюции транскрипционной регуляторной сети, а также проведено детальное исследование на материале обширной группы близкородственных геномов. Мы впервые показали, что различные регуляторные взаимодействия изменяются по-разному: как в зависимости от положения в транскрипционной сети, так и от свойств транскрипционного фактора.

С помощью методов сравнительной геномики мы обнаружили ранее неизвестные регуляторные элементы в пути утилизации этаноламина, а именно мотив участка связывания транскрипционного фактора AraC-семейства EutR. Кроме того, нам удалось

показать регуляторную связь этого пути с метаболизмом кобаламина – кофактора основного фермента пути (этаноламин лиазы), а также описать эволюцию генов этого пути.

Теоретическая и практическая значимость

На сегодняшний день доступно несколько тысяч полных бактериальных геномов, включая штаммы, и еще больше находится в процессе определения последовательности или функциональной аннотации. Но, несмотря на экспоненциально растущее количество геномов, экспериментальное изучение регуляторных сетей, вследствие его трудоемкости, ограничено несколькими модельными организмами либо отдельными функциональными системами. Таким образом, следствием большого количества данных является необходимость применения методов компьютерной биологии и статистики для их анализа, позволяющих описывать свойства немодельных организмов, основываясь на экспериментально полученных данных по родственным модельным организмам. Более того, применение сравнительно-геномных методов позволяет описывать регуляторные взаимодействия в таксономических группах, не содержащих хорошо изученных модельных организмов.

Методология и методы исследования

Для решения поставленных задач использовались разные методы сравнительной геномики. Применялись методы поиска гомологов на основе критерия двустороннего лучшего сходства, поддержанного анализом филогенетических деревьев и анализом структуры оперона, методы анализа аминокислотных, нуклеотидных последовательностей, а также построения множественных выравниваний и филогенетических деревьев методом наилучших соседей с применением статистического размножения выборки. Для проверки статистической значимости использовался критерий χ^2 и гипергеометрический тест.

Положения, выносимые на защиту

1. Анализ таксономического распределения генов утилизации этаноламина показал, что существует два возможных пути катаболизма использования этаноламина, связанных с типом оперона. Первый, короткий, позволяет использовать этаноламин только в качестве источника азота. Второй, длинный, дает возможность использовать его и как источник азота, и как источник углерода.

2. Исследование эволюции генов утилизации этаноламина показало, что короткий оперон является предковым, из которого путем добавления новых генов образовался длинный тип. В ходе эволюции генов утилизации этаноламина произошло минимум три события горизонтального переноса.
3. В Enterobacteriales и Burkholderiales предсказан мотив участка связывания транскрипционного фактора EutR. В Enterobacteriales участки связывания EutR обнаружены в промоторной области семи оперонов, среди которых есть непосредственно гены утилизации этаноламина, а также гены синтеза кофактора основного фермента пути этаноламинлиазы – кобаламина. Это наблюдение устанавливает связь между путями утилизации этаноламина и синтезом кобаламина за счет EutR-зависимой регуляции. В Burkholderiales участки связывания EutR обнаружены только непосредственно в промоторной области генов утилизации этаноламина.
4. Локальная регуляция является эволюционно подвижной, что согласуется с необходимостью быстрой адаптации к условиям окружающей среды. В штаммах *Escherichia coli* взаимодействия типа Л→ген (локальный транскрипционный фактор →ген) чаще сохраняются в регуляторных блоках, чем вне их. На уровне порядка Enterobacteriales эти взаимодействия консервативнее в парных взаимодействиях, по сравнению с регуляторными блоками, что может быть результатом изменчивости транскрипционной сети.
5. Разные типы треугольников на уровне порядка Enterobacteriales эволюционируют по-разному. Регуляция как глобальными, так и локальными транскрипционными факторами в несогласованном треугольнике оказывается консервативнее, чем в согласованном.

Степень достоверности и апробация результатов

По материалам диссертации опубликовано 2 статьи в международных рецензируемых научных журналах. Результаты работы были представлены на международной конференции МССМВ’09, российских конференциях ИТИС’08, ИТИС’10, а также на международных семинарах RECESS’10, RECESS’11 и Chemical, Synthetic And Systems Biology: New Directions Of Biochemistry In The 21st Century’11.

Структура и объем работы

Диссертация изложена на 100 страницах машинописного текста и содержит следующий разделы: обзор литературы, материалы и методы, результаты в двух главах, обсуждение и выводы. В конце приведен список литературы. Материал включает 21 рисунок, 14 таблиц и список литературы, содержащий 166 ссылок.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

1. Предсказание регуляторных элементов транскрипционной сети на примере пути деградации этаноламина

Путь утилизации этаноламина

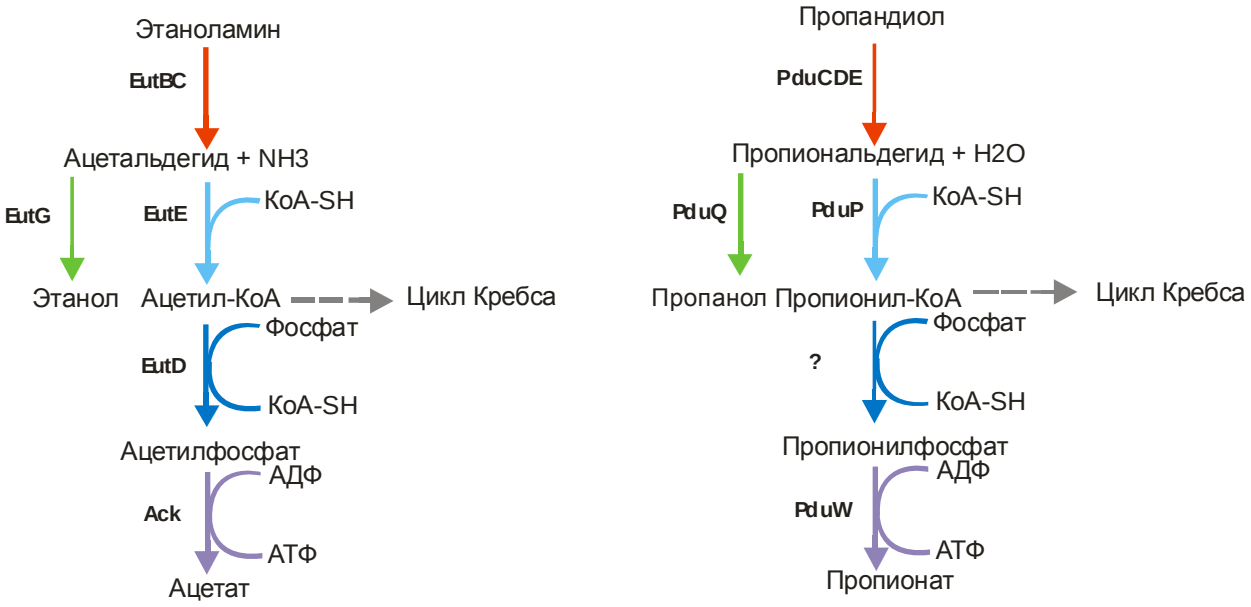
Одним из источников углерода, азота и энергии для бактерий может служить этаноламин. Путь утилизации этаноламина является частным случаем реакций утилизации диолов с образованием альдегидов (Рисунок 1). Все известные подобные пути требуют наличия кофактора – кобаламина (витамина B12).

Путь утилизации этаноламина экспериментально изучен у *Salmonella typhimurium* LT2. Основным ферментом пути является этаноламин лиаза (EC 4.3.1.7) – белковый комплекс, состоящий из двух субъединиц: большой EutB и малой EutC. Этаноламин лиаза осуществляет главную реакцию пути – расщепление до ацетальдегида и аммиака, который используется как источник азота. Дальнейшее превращение ацетальдегида осуществляется оксидоредуктазой EutE до ацетил-КоА или алкогольдегидрогеназой EutG до этанола. Ацетил-КоА в свою очередь может быть превращен в ацетат с получением одной молекулы АТФ с помощью ферментов ацетат киназы Ack и фосфоацетилтрансферазы EutD либо отправлен в цикл Кребса (Kofoid et al, 1999) (Рисунок 1а).

Кофактором этаноламин лиазы является аденозилкобаламин, который может быть получен извне или из цианокобаламина или гидроксикобаламина, синтезируемых в клетке. Превращение в аденозилкобаламин происходит с помощью аденозилкобаламинтрансферазы EutT (Sheppard et al, 2004). Цианокобаламин и гидроксикобаламин являются ингибиторами этаноламин лиазы. EutA, компонент ферментативного комплекса который обнаружен у многих бактерий, растущих на этаноламине, защищает этаноламин лиазу от их воздействия.

У некоторых бактерий, способных к деградации диолов, обнаружен особый компартмент, так называемая метаболосома (Kofoid et al, 1999, Brinsmade et al, 2005). У таких организмов превращения этаноламина до ацетил-коА происходят внутри этого компартмента. Структурные белки метаболосомы – EutS, EutM, EutN, EutL, EutK – являются гомологами структурных белков карбоксисомы – органеллы цианобактерий,

ответственной за накопление CO_2 для фиксации ферментом рибулозобисфосфаткарбоксилазой (Orus et al, 1995) Предположительная роль метаболосом в клетках бактерий – предотвращение потерь альдегидов во время утилизации диолов (Penrod, Roth, 2006).



Одинаковым цветом показаны химически сходные реакции.
Рисунок 1 - (а) Путь утилизации этаноламина; (б) Путь утилизации пропандиола.

Ранее предпринимались попытки экспериментального изучения регуляции транскрипции *eut*-оперона (Roof, Roth, 1992). Для *S. thyphimurium* была экспериментально показана роль EutR, ген которого также находится в составе *eut*-оперона, в регуляции экспрессии генов утилизации этаноламина (Roof, Roth, 1992). Участки связывания EutR с ДНК не установлены.

Транскрипционный фактор EutR обнаружен далеко не у всех бактерий, содержащих ферменты утилизации этаноламина. Для таких организмов возникает отдельный вопрос о том, как происходит регуляция этого пути. Для таксономической группы Firmicutes ранее замечено, что рядом с генами утилизации этаноламина часто

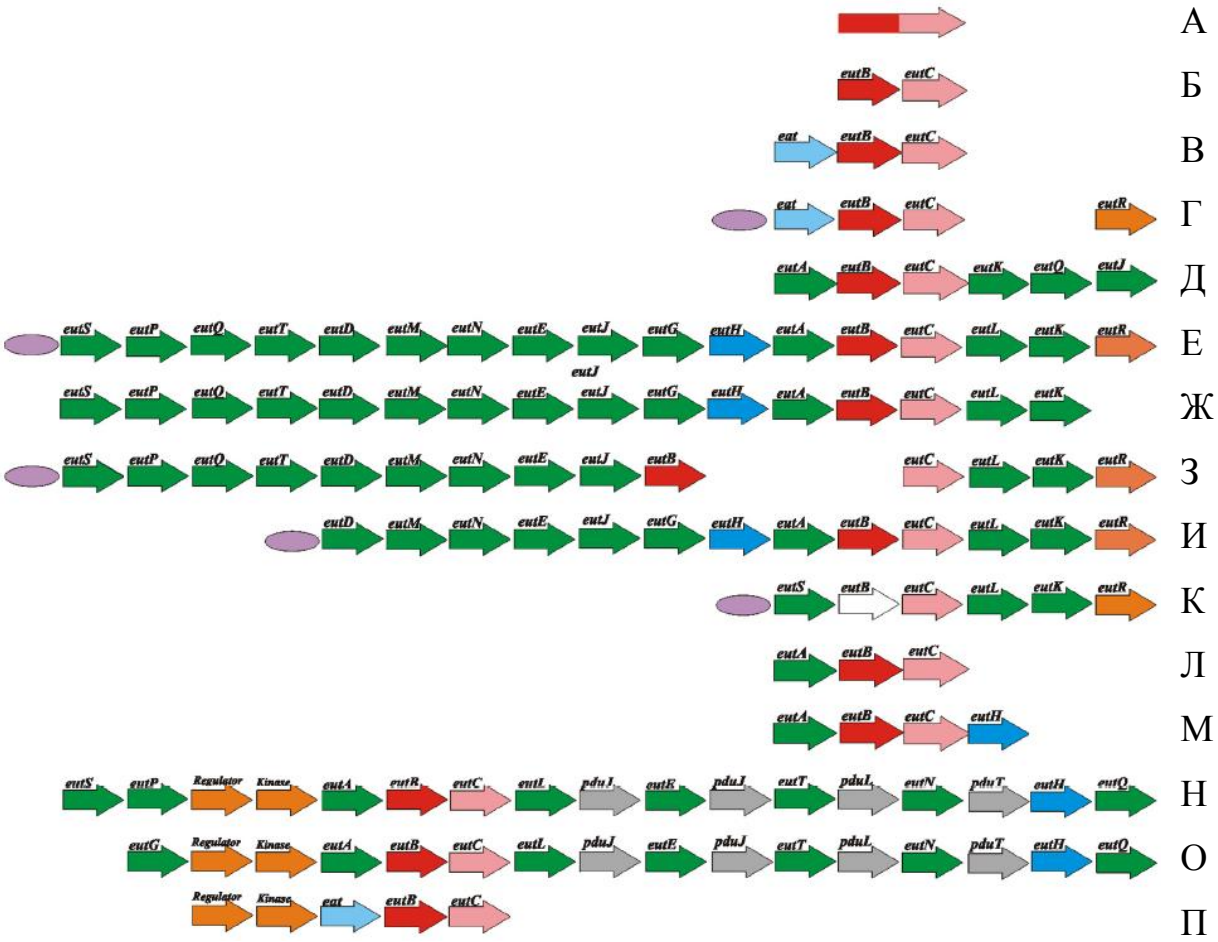
обнаруживаются гены двухкомпонентной регуляторной системы, например, гены *lin1136* и *lin1137* в *Listeria innocua* (А. Мушегян, частное сообщение).

Происхождение этой сложной метаболической системы неизвестно.

Путь утилизации этаноламина интересен еще и тем, что была показана его связь с пищевым отравлением. Так, обнаружено, что большинство бактерий, вызывающих пищевые отравления, имеют ферменты утилизации этаноламина, а многие также и пропандиола (Korbel et al, 2005).

Поиск ортологов и эволюция генов утилизации этаноламина

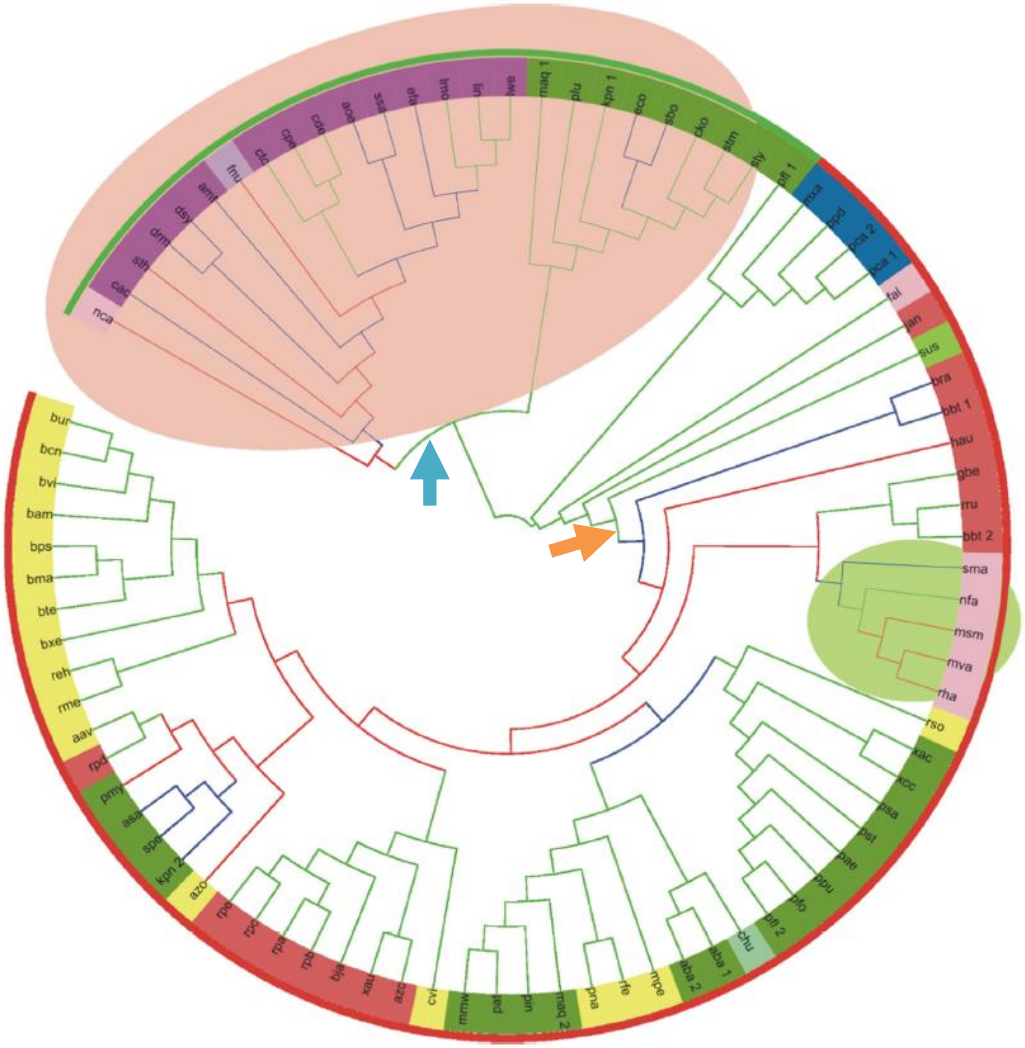
Ортологи EutB и EutC были обнаружены в более чем 100 видах бактерий из различных таксономических групп: Proteobacteria, Firmicutes, Actinobacteria, Acidobacteria, Bacteroidetes, Chloroflexi, Flavobacteria, Lentisphaerae, Planctomycetacia и Fusobacteria. Структура *eut*-оперона оказалась различной в разных группах бактерий. Мы выделили два основных типа подобных оперонов – *короткий*, содержащий минимальный набор генов утилизации этаноламина, и *длинный* (Рисунок 2).



Короткие опероны: А, *Deltaproteobacteria*; Б, *Proteobacteria*, *Chlorophlexi* и *Bacteroidetes*; В, некоторые *Proteobacteria* и *Acidobacteria*; Г, *Betaproteobacteria* (*eutR* обнаружен отдельно от *eutBC* и *eat*). Длинные опероны: Д, *Nocardioides* sp.; Е, *Enterobacteriaceae*; Ж, *M. aquaeolei*; З, *S. boydii* Sb227; И, *S. sonnei* Ss046; К, *S. dysenteriae* Sd197; Л, *Symbiobacterium thermophilum* and *P. luminescens*; М, *P. fluorescens* Pf-5; Н, *Clostridiaceae* и *F. nucleatum*; О, *Listeriaceae* и *Enterococcaceae*; П, *C.acetobutylicum*.
Белым цветом обозначен псевдоген *eutB*. Предсказанные участки связывания EutR обозначены фиолетовыми овалами.
Рисунок 2 – Разнообразие оперонных структур, содержащих гены *eutBC*.

Для изучения эволюции этой сложной метаболической системы на основе аминокислотных последовательности ферментов EutB и EutC из полных геномов были построены филогенетические деревья методом объединения соседних пар («neighbour-joining»). Для оценки достоверности наблюдаемого расположения ветвей применялось статистическое размножение выборки («бутстреп-анализ»). Мы обнаружили, что гены из разных типов оперонов, короткого и длинного. Первая ветвь содержит белки из

Proteobacteria и Actinobacteria, вторая ветвь – из Enterobacteriales, относящихся к Proteobacteria, и Firmicutes (Рисунок 3). В случае геномов, содержащих два *eut*-оперона (*K. pnemoniae, M aquaeolei, P. fluorescens*), eutB из оперонов разного типа лежат на разных ветвях.



Цвет ветвей соответствует бутстреп-значениям: зеленый – > 70%, синий – от 50 до 70%, красный – <50%.

Цвет внешнего круга соответствует длине оперона: зеленый – длинный оперон, красный – короткий оперон.

Цвет внутреннего круга соответствует таксономической принадлежности: красный, Alphaproteobacteria; желтый, Betaproteobacteria; зеленый, Gammaproteobacteria; синий, Deltaproteobacteria; фиолетовый, Firmicutes; розовый, Actinobacteria; зелено-желтый, Acidobacteria; светло-фиолетовый, Fusobacteria; голубой, Chlorophlexi.

Овалами обозначены гены, расположение которых на дереве противоречит таксономии, т.е. события горизонтального переноса генов.
Стрелками обозначены ветви, бутстреп-значения которых поддерживают гипотезу горизонтального переноса генов: оранжевая стрелка – бутстреп-значение$_{eutB}$1, голубая стрелка – бутстреп-значение$_{eutB}$2,
Обозначения геномов взяты из базы данных KEGG. Длины ветвей не отражают расстояния между видами.
Рисунок 3 – Сводное филогенетическое дерево EutB, построенное методом объединения соседних пар с бутстреп-анализом.

Случаи, когда расположение ветвей на филогенетическом дереве не согласуется с таксономией, представляют особый интерес (Рисунок 3). Такая топология дерева свидетельствует о возможном событии горизонтального переноса генов.

Первый случай отклонения от таксономии был обнаружен для белков бактерий типа Actinobacteria. Ветвь Actinobacteria оказывается внутри ветви, соответствующей белкам из типа Proteobacteria (бутстреп-значение$_{eutB}$1 > 70%, Рисунок 3), в то время как на дереве видов таксоны Actinobacteria и Proteobacteria являются равнозначными. Кроме того, и Actinobacteria, и Proteobacteria имеют короткий тип оперона. Такую топологию можно объяснить следующим образом: вначале произошел горизонтальный перенос короткого оперона генов утилизации этаноламина из Proteobacteria к общему предку Actinobacteria, а далее он распространился по некоторым организмам этого таксона. Таким образом, гены утилизации этаноламина из Actinobacteria являются потомками генов из Proteobacteria.

Второй случай отклонения от таксономии наблюден в *Fusobacterium nucleatum*. Ветвь, соответствующая белкам из *F. nucleatum* (тип Fusobacteria), находится внутри ветви Firmicutes (бустреп-значение$_{eutB}$2>70%, Рисунок 3). Fusobacteria и Firmicutes являются равнозначными таксонами, поэтому и здесь, вероятно, также имел место недавний горизонтальный перенос из Firmicutes в Fusobacteria. Дополнительным свидетельством недавнего переноса является то, что в остальных четырех представителях Fusobacteria, для которых доступна полногеномная последовательность (*Ilyobacter polytropus, Leptotrichia buccalis, Sebaldella termitidis, Streptobacillus moniliformis*), генов утилизации этаноламина не найдено.

Самый яркий пример отличия топологии дерева EutB от дерева таксонов – расщепление ветви Gammaproteobacteria, относящейся к Proteobacteria, на две (Рисунок 3). Расположение одной из ветвей соответствует таксономии (Рисунок 3), а другая ветвь (порядок Enterobacteriales) становится сестринской к Firmicutes. Интересно, что для организмов из первой ветви характерен короткий тип оперона, в то время как организмы второй ветви, как и Firmicutes, содержат длинный тип оперона. Можно предположить две возможные причины возникновения такой топологии. Во-первых, стоит отметить, что только для этих организмов характерно наличие специального компартмента для ферментов утилизации этаноламина – метаболосомы. По этой причине можно предположить, что эволюция ферментов в условиях макромолекулярного комплекса с метаболосомой могла проходить конвергентно в двух таксономически далеких ветвях. Но более вероятной является возможность еще одного события горизонтального перенос генов от Proteobacteria к Firmicutes, или наоборот.

Для детального исследования происхождения такой топологии мы реконструировали вероятное строение оперона *eut* у предка Firmicutes, предка Proteobacteria и предка Enterobacteriales, используя метод максимальной экономии (maximal parsimony), реализованный в программе MESQUITE. На основании реконструкции составов предковых оперонов можно выдвинуть следующую гипотезу об эволюции генов утилизации этаноламина. Из всех генов утилизации этаноламина только *eutBC* были у предков всех трех исследумых таксономических групп. Таким образом, предковый оперон с высокой вероятностью состоял только из генов *eutBC*. Опираясь на тот факт, что похожий тип оперона у ныне живущих бактерий масштабно представлен только у Proteobacteria, мы выдвинули первую гипотезу. Она состоит в том, что гены утилизации этаноламина *eutBC* были у общего предка всех Proteobacteria. Позднее к ним добавился ген транспортера *eat*. Далее этот основной набор генов был дополнен геном, кодирующим транскрипционный фактор EutR. В таком виде – *eat-eutBC* и соседний ген *eutR* – состав генов утилизации этаноламина сохранился у ныне живущих представителей Proteobacteria. В ветви Enterobacteriales предковый оперон был дополнен структурными генами метаболосомы, геном синтеза аденозилкобаламина *eutT* и генами расщепления этаноламина до этанола и ацетата с получением АТФ (Рисунок 1).

Что касается Firmicutes, то их *eut*-опероны могли возникнуть в результуте горизонтальных переносов от Enterobacteriales с заменой транскрипционного фактора EutR на двухкомпонентную регуляторную систему. В результате горизонтальных переносов от Enterobacteriales к другим Proteobacteria в геномах *M. aquaeolei* VT8, *K. pneumoniae subsp. pneumoniae* MGH 78578, *P. fluorescens* Pf-5 возникло два независимых *eut*-оперона.

Вторая гипотеза состоит в том, что путь утилизации этаноламина был еще раньше – у гипотетического общего предка Firmicutes и Proteobacteria. Дальнейшая эволюция *eut*-оперонов в этих ветвях привела к независимому образованию короткого и длинного типов. Короткий тип развился в Proteobacteria, а длинный – в Firmicutes.

Горизонтальный перенос генов от Firmicutes к древней бактерии, предку порядка Enterobacteriales, привел к тому, что у большинства Enterobacteriales возник длинный *eut*-оперон. Геномы Proteobacteria с двумя *eut*-оперонами (*M. aquaeolei* VT8, *K. pneumoniae subsp. pneumoniae* MGH 78578, *P. fluorescens* Pf-5) являются результатами горизонтальных переносов либо от Firmicutes, либо от Enterobacteriales, при которых собственные гены утилизации этаноламина еще не успели деградировать.

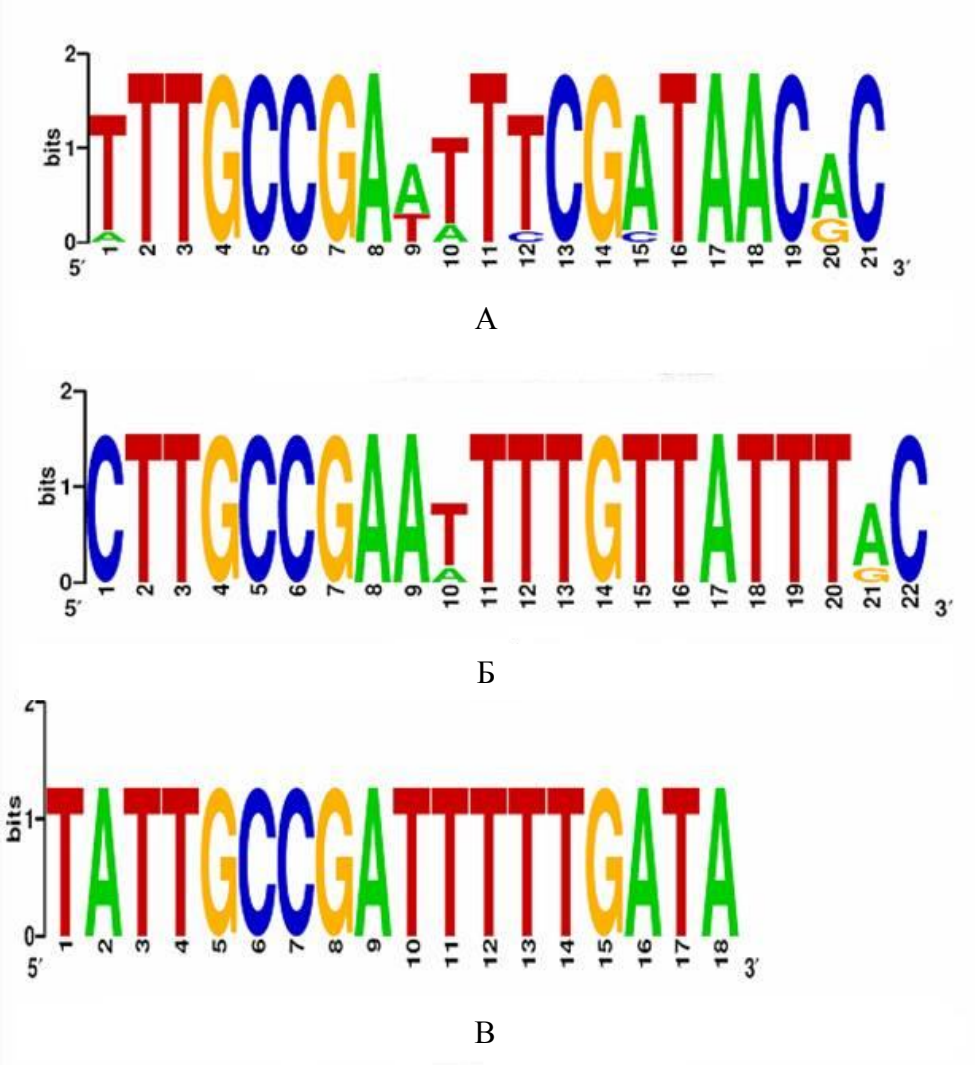
Регуляция утилизации этаноламина

Для поиска регуляторных элементов были отобраны группы близкородственных геномов, содержащих ортологи EutR: первая группа содержала бактерии порядка Enterobacteriales (*E.coli* K12 MG1655, *E.coli* O157, *C.koseri, K.pneumoniae, S.enterica, S.boydii, S.dysenteriae*), вторая – Burkholderiales (*B.cepacia, B.cenocepacia, B.mallei, B.pseudomallei, B.vietnamiensis, B.thailandensis, B.xenovorans, B.thailandensis, P.naphthalenivorans, A.avenae, M.petroleiphilum*).

У Enterobacteriales *eut*-оперон содержит 17 генов, первым из которых является ген *eutS*. Методом филогенетического футпринтинга в промоторной области гена *eutS* найдены два консервативных участка. Один из них при детальном рассмотрении обнаружил сходство с участком связывания CRP – wwwTGTGAtyyrgwTCACTtWt. Мы предположили, что вторая консервативная область является участком связывания EutR. Косвенным подтверждением этого предположения является тот факт, что наличие транскрипционного фактора EutR коррелирует с наличием найденного участка связывания

в промоторной области оперона *eut* (в геномах, не содержащих гена транскрипционного фактора EutR, подобного участка найдено не было).

У представителей Burkholderiales оперон содержит ген транспортера этаноламина *eat* и гены этаноламин лиазы *eutBC*. В промоторной области гена *eat* была обнаружена только одна консервативная область. При сравнении обнаруженных участков (Рисунок 4) выявилось их сходство, что также подтверждает наши предположения о том, что это участок связывания именно транскрипционного фактора EutR.



А

Б

В

Рисунок 4 – Сравнение диаграмм Лого для предполагаемых участков связывания EutR в регуляторных областях *eut*- и *cob*- оперонов: А) Участок связывания EutR в регуляторной области *eut*-оперона у Burkholderiales; Б) Участок связывания EutR в регуляторной области *eut*-оперона у Enterobacteriales; В) Участок связывания EutR в регуляторной области *cob*-оперона у Enterobacteriales.

По горизонтальной оси указан номер позиции нуклеотида, по вертикальной – информационное содержание позиции в битах. Относительная высота каждой буквы соответствует частоте нуклеотида в данной позиции.

На основании полученных последовательностей участков связывания EutR была построена матрица позиционных весов. Эту матрицу использовали для поиска генов, экспрессия которых может регулироваться EutR. Поиск охватывал область −400…+100 от старта трансляции.

Сканирование при помощи матрицы позиционных весов обнаружило около 60 генов из 7 регулонов, относящихся к разным метаболическим путям (Таблица 1). Похожая консервативная область обнаружена перед геном *cbiA* – первым геном оперона синтеза кобаламина. Кобаламин является кофактором основного фермента утилизации этаноламина, поэтому этот случай представляет особый интерес (Toraya, 2000) (Таблица 1).

Таблица 1 Участки связывания EutR перед потенциальными членами регулона. Условные обозначения: «+» – обнаружен потенциальный участок связывания перед опероном; «−» – потенциальный участок связывания перед опероном не найден;
«0» – ортологов данных генов не обнаружено.
Условные обозначения геномов взяты из базы данных KEGG.

Гены	**Геномы**					**Функция**
	kpn	**sty**	**cko**	**eco**	**sdy**	
yabB-mraZW-ftsLI-murEF-mraY-murD-ftsW-murGC-ddlB-ftsQAZ-lpxC	+	+	−	+	+	Деление клетки/синтез клеточной стенки
cbiA-cbiBCDETFGHJKLMNQOP-cobUST	+	+	+	0	0	Синтез кобаламина
yciW	−	+	+	+	+	Неизвестный белок
yhhM	−	+	+	+	+	Белок внутренней мембраны
xylFGHR	+	0	+	+	−	ABC транспортер ксилозы
eutSPQTDMNEJGHAB	+	+	+	+	+	Утилизация этаноламина

CKLR						
stpA	+	+	+	+	+	ДНК-связывающий белок

Обнаруженный участок связывания EutR перекрывается с одним из участков связывания PocR. Видимо, как и в случае утилизации пропандиола, транскрипционный фактор EutR также может одновременно контролировать утилизацию этаноламина и синтез необходимого для этого кофактора – кобаламина.

2. Регуляторные блоки в сетях регуляции транскрипции

Транскрипционную сеть можно рассматривать как направленный граф, в вершинах которого находятся транскрипционные факторы и регулируемые гены, а ребра отображают связь факторов с регулируемыми генами. Особенностью графа на основе биологических сетей является наличие внутренних структур, повторяющихся чаще, чем в случайном графе. Такие подграфы, частота которых в биологическом графе значимо больше, чем в случайном графе, получили название «структурных мотивов» («network motif») (Shen-Orr et al, 2002). В данной работе мы будем называть такие подграфы «регуляторные блоки», чтобы отличать этот объект от мотива участков связывания транскрипционных факторов.

Для различных биологических сетей характерны разные регуляторные блоки. Было показано, что в транскрипционных сетях самыми частыми регуляторными блоками являются «треугольник» (feed-forward loop) и «веер» («bi-fan») (Shen-Orr et al, 2002). В настоящей работе мы рассматривали регуляторный блок из трех элементов – «треугольник». Это структура их трех генов, два из которых являются транскрипционными факторами. Один из этих транскрипционных факторов (X, так называемый, общий) регулирует другой (Y, специальный), а вместе они регулируют ген (Z) (Рисунок 5). Несмотря на то, что всего существует 13 возможных способов связать три вершины в направленном графе (Рисунок 6), показано, что в транскрипционных сетях среди блоков с тремя вершинами перепредставлен только «треугольник» (Shen-Orr et al, 2002).

Так как транскрипционный фактор может быть активатором или репрессором, существует восемь различных подтипов подобного регуляторного блока треугольник

(Mangan, Alon, 2003). В зависимости от того, какие транскрипционные факторы находятся в вершинах треугольника, различают согласованные и несогласованные треугольники (Рисунок 7). Мотив является согласованным, если прямой эффект общего транскрипционного фактора имеет тот же знак (положительный или отрицательный), что и непрямой эффект, опосредованный специальным фактором. Если знаки не совпадают, такой треугольник называется несогласованным.

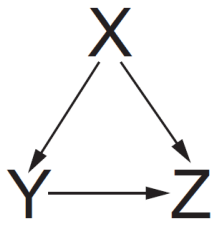


Рисунок 5 - Общая схема регуляторного блока типа «треугольник».

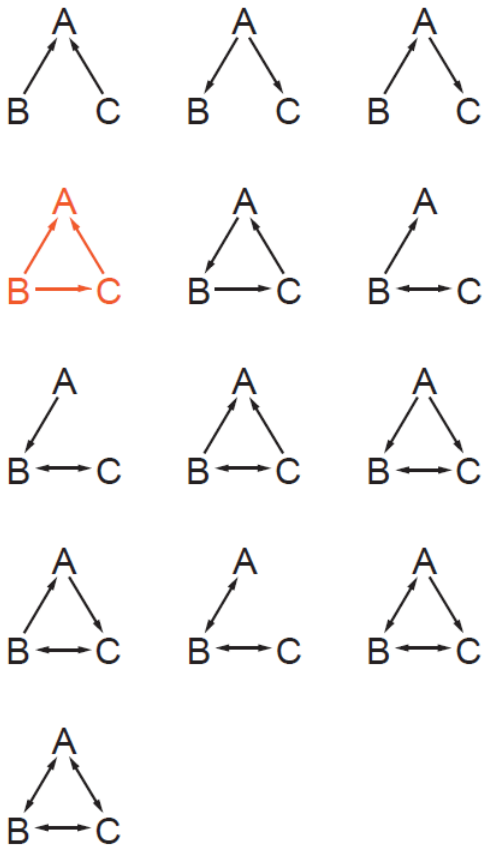


Рисунок 6 – Возможные регуляторные блоки с тремя элементами. Оранжевым выделен «треугольник».

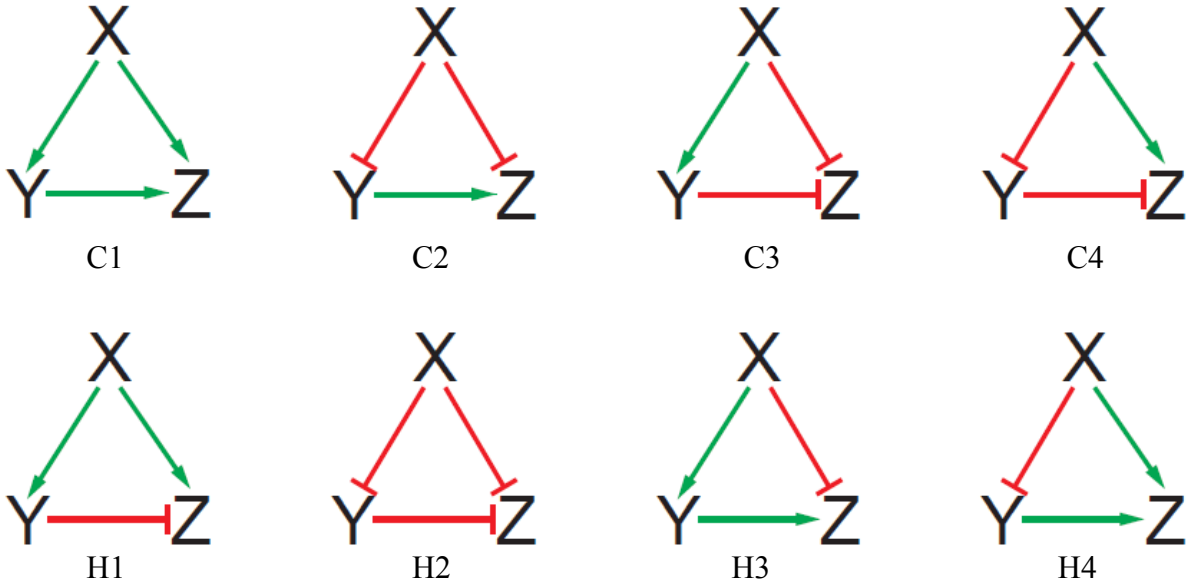


Рисунок 7 - Типы согласованных (C) и несогласованных (H) треугольников.

Взаимодействия внутри треугольников можно разделить по природе образующего его транскрипционного фактора. В научной литературе принято разделять транскрипционный факторы на так называемые локальные и глобальные, хотя единого формального критерия разделения не существует. Одно из первых определений: глобальный транскрипционный фактор – это такой фактор, который участвует в регуляции нескольких метаболических путей (Gottesman, 1984, Martinez-Antonio, Collado-Vides, 2003). Другой подход основан на количестве регулируемых генов или оперонов (Shen-Orr et al, 2002, Madan Babu, Teichmann, 2003). Недавно был предложен структурный подход, в котором транскрипционная регуляция рассматривалась как сеть, а глобальным считался фактор, который регулирует несколько модулей в этой сети (Ma, Buer, Zeng, 2004). Тип транскрипционных факторов моет быть важен потому, что глобальные и локальны транскрипционные факторы по-разному сохраняются в ходе эволюции.

Определение локальных и глобальных транскрипционных факторов

Мы объединили несколько подходов для определения глобальных транскрипционных факторов. Транскрипционный фактор считался глобальным, если он, во-первых, принимает участие в регуляции более 20 оперонов, и, во-вторых, эти опероны принадлежат разным метаболическим путям. Этим критериям соответствуют восемь транскрипционных факторов: CRP, IHF, FNR, Fis, ArcA, Lrp, H-NS и FUR.

Функциональная аннотация треугольников

Регуляторные блоки могут быть неравномерно представлены в метаболических путях. Анализ функциональных категорий по базе данных COG (Tatusov et al, 2003) генов, входящих в состав треугольников, показал, что, в треугольниках перепредставлены гены, участвующие в производстве энергии (p-значение 0), транспорте и метаболизме сахаров (p-значение $7{,}3\times10^{-7}$). Для оценки статистической значимости использован гипергеометрический тест. Этот результат объясняется тем, что большинство треугольников образовано глобальными транскрипционными факторами дыхания Fnr и ArcA или катаболизма сахаров CRP (Gelfand, 2006).

Эволюция регуляторного блока типа треугольник в близкородственных организмах на примере штаммов *Escherichia coli* и в Enterobacteriales

Эволюция мотива треугольник в близкородственных организмах была изучена на примере 25 штаммов *E. coli*. Для анализа мы отобрали такие межгенные участки, которые сохраняются по крайней мере в 10 штаммах. Мы считали, что межгенный участок сохраняется, если, во-первых, сохраняются оба гена, между которыми он расположен, во-вторых, ориентация этих генов осталась неизменной. Кроме того, здесь мы полагали, что участок связывания исчезает, если в нем возникает хотя бы одна замена. Для очень близких организмов со степенью схожести последовательности более 95%, событие замены происходит очень редко, поэтому необходимо использовать такие жесткие критерии, чтобы таким образом фиксировать редкие события изменения последовательности участка связывания.

Информация о регуляторных взаимодействиях была получена из базы данных RegulonDB (Gama-Castro et al, 2011). В 25 штаммах *E.coli* содержится 335 регуляторных взаимодействий, принимающих участие в образовании треугольников, и 367 прочих взаимодействий, которые мы назвали *парными*. В 335 взаимодействиях из треугольников у 194 в вершине находится глобальный транскрипционный фактора у 161 – локальный. В парных взаимодействиях в вершинах 105 взаимодействий находится глобальный транскрипционный фактор, а в вершинах 262 – локальный (Таблица 2).

Таблица 2. Количество различных типов регуляторных взаимодействий в 25 штаммах *E.coli*.

	Треугольники (всего)	Треугольники (консервативные)	Парные (всего)	Парные (консервативные)
Глобальные	194	102	105	63
Локальные	161	109	262	152
Всего	335	211	367	215

Анализ консервативности участков связывания транскрипционных факторов показал, что из 199 глобальных взаимодействий сохранилось 165: 63 парных и 102 из треугольников. Из 421 локального взаимодействия сохранилось только 261 взаимодействие: 152 парных и 109 из треугольников. Оказалось, что локальные взаимодействия более консервативны в треугольниках, чем в парных взаимодействиях (статистическая значимость 0.047 согласно тесту χ^2).

Анализ был сделан на 25 штаммах *E. coli*. Так как их число постоянно растет, возникает вопрос, насколько может измениться полученный результат, если анализировать большее число организмов. Мы проанализировали, как меняется доля консервативных регуляторных взаимодействий в зависимости от числа штаммов и обнаружили, что она практически не меняется, начиная с 15 ± 2 штаммов, в зависимости от типа взаимодействия.

В Enterobacteriales мы проанализировали частоту всех возможных механизмов изменения регуляторного взаимодействия:

- сохранение всех трех элементов – транскрипционного фактора, гена и участка связывания (обозначено в таблицах 3-6 как «Консервативная регуляция»)
- исчезновение транскрипционного фактора («Нет транскрипционного фактора»)
- исчезновение регулируемого гена («Нет регулируемого гена»)
- исчезновение участка связывания: хотя бы одного или всех в случае множественного регуляции («Нет участка связывания»); исчезновение участка связывания определялось как сильное уменьшение его веса, определяемого с помощью матрицы позиционных весов.

Информация по регуляторным взаимодействиям, как и при анализе штаммов *E. coli*, была использована из базы данных RegulonDB.

Сила связывания с транскрипционным фактором может быть примерно оценена при помощи матрицы позиционных весов – чем выше вес участка связывания, тем выше может быть его сила связывания с транскрипционным фактором. В настоящей работе вес участка связывания рассчитывался с помощью матрицы позиционных весов. Числа в ячейках такой матрицы – это позиционные веса (вес каждого нуклеотида в каждой позиции участка связывания), которые вычисляются по формуле:

$$\mathbf{W}(b,k) = \log(N(b,k) + 0{,}5) - 0{,}25 \sum_{i=A,C,G,T} \log(N(i,k) + 0{,}5),$$

где N (*b, k*) – количество нуклеотидов *b* в позиции *k* в обучающей выборке. Первый член в формуле зависит от того, сколько раз данный нуклеотид встретился в данной позиции, второй – от консервативности самой позиции. Вес участка связывания определяется суммой соответствующих позиционных весов, и измеряется в единицах стандартного отклонения распределения весов на случайных последовательностях.

Чтобы принять во внимание изменения в силе связывания (т.е. веса) участка связывания, мы анализировали только такие транскрипционные факторы, для которыз матрица позиционных весов есть в базах данных RegPrecise (Novichkov et al, 2012) или ее можно создать на основе последовательностей участков связывания в RegulonDB. Таким образом, для анализа было использовано 96 транскрипционных фактора. Участок связывания считался консервативным в данном организме, если его вес уменьшился не более чем на две единицы по сравнению с весом аналогичного участка в *Escherichia coli* K12.

На основе данных из RegulonDB для Enterobacteriales мы получили 473 парных и 418 взаимодействий, образующих треугольник. Разделив их в зависимости от типа транскрипционного фактора, мы получили 6 видов взаимодействий: Г (глобальный транскрипционный фактор)→регулируемый ген, Л (локальный транскрипционный фактор)→регулируемый ген, Г→Г, Г→Л, Л→Л, Л→Г. Оказалось, что взаимодействия типа Л→Г встречаются очень редко, а взаимодействия типа Г→Г характерны только для треугольников и не встречаются в парах. Таким образом, мы изучили оставшиеся четыре типа регуляторных взаимодействий. В таблицах указано количество определенных событий, произошедших с каждым типом взаимодействия (Таблицы 3-4).

Таблица 3. Количество событий, произошедших с парными взаимодействиями в порядке Enterobacteriales

	Г→Л	Л→Л	Г→ген	Л→ген	Л→ген (в *S.enterica*)
Нет транскрипционного фактора	0	73	0	240	45
Нет регулируемого гена	31	8	607	674	26
Нет участка связывания (хотя бы одного)	7	62	708	1201	14
Нет участка связывания (всех)	1	25	596	746	11
Консервативная регуляция	51	321	840	2522	250
Всего	89	452	2155	4637	335

Таблица 4. Количество событий, произошедших с треугольниками в порядке Enterobacteriales

	Г→Л	Л→Л	Г→ген	Л→ген	Л→ген (в *S.enterica*)
Нет транскрипционного фактора	0	189	0	208	44
Нет регулируемого гена	281	82	572	193	15
Нет участка связывания (хотя бы одного)	186	117	828	395	17
Нет участка связывания (всех)	105	44	536	186	8
Консервативная регуляция	290	483	1310	596	110
Всего	757	871	2710	1371	186

Для всех проанализированных организмов взаимодействие с локальным транскрипционным фактором оказалось более консервативным в парных взаимодействиях. Статистическая значимость теста χ^2 для *Salmonella enterica* равна 0.006.

Для других организмов результат не значим, но наблюденное отклонение сохраняется. Консервативность регуляторных взаимодействий, образованных глобальными транскрипционными факторами, не различалась между треугольниками и парами. Таким образом, для подобных регуляторных взаимодействий наличие структурного мотива не влияет на их консервативность.

Далее мы изучили поведение самых частых треугольников согласованного (C1) и несогласованного (H1) типа. Согласованный треугольник оказался менее консервативным, чем несогласованный: участки связывания в таком треугольнике исчезают быстрее (Таблицы 5-6). Статистическая значимость теста χ^2 равна 0,021 для глобальных взаимодействий, и $6{,}7\times10^{-7}$ для локальных.

Таблица 5. Количество событий, произошедших с согласованными треугольниками типа C1

	Г→Л	Л→Л	Г→ген	Л→ген
Нет транскрипционного фактора	0	72	0	138
Нет регулируемого гена	157	15	131	32
Нет участка связывания (хотя бы одного)	50	40	282	184
Нет участка связывания (всех)	10	14	178	77
Консервативная регуляция	87	107	483	147
Всего	294	234	896	501

Таблица 6. Количество событий, произошедших с несогласованными треугольниками типа H1

	Г→Л	Л→Л	Г→ген	Л→ген
Нет транскрипционного фактора	0	193	0	100
Нет регулируемого гена	135	50	245	92
Нет участка связывания (хотя бы одного)	154	58	367	131
Нет участка связывания (всех)	57	30	166	26
Консервативная регуляция	271	258	712	276
Всего	560	559	1324	599

Выводы

1. Анализ таксономического распределения генов утилизации этаноламина показал, что существует два возможных пути катаболизма использования этаноламина, связанных с типом оперона. Первый, короткий, позволяет использовать этаноламин только в качестве источника азота. Второй, длинный, дает возможность использовать его и как источник азота, и как источник углерода.

2. Исследование эволюции генов утилизации этаноламина показало, что короткий оперон является предковым, из которого путем добавления новых генов образовался длинный тип. В ходе эволюции генов утилизации этаноламина произошло минимум три события горизонтального переноса.

3. В Enterobacteriales и Burkholderiales предсказан мотив участка связывания транскрипционного фактора EutR. В Enterobacteriales участки связывания EutR обнаружены в промоторной области семи оперонов, среди которых есть непосредственно гены утилизации этаноламина, а также гены синтеза кофактора основного фермента пути этаноламинлиазы – кобаламина. Это наблюдение устанавливает связь между путями утилизации этаноламина и синтезом кобаламина за счет EutR-зависимой регуляции. В Burkholderiales участки связывания EutR обнаружены только непосредственно в промоторной области генов утилизации этаноламина.

4. Локальная регуляция является эволюционно подвижной, что согласуется с необходимостью быстрой адаптации к условиям окружающей среды. В штаммах *Escherichia coli* взаимодействия типа Л → ген (локальный транскрипционный фактор → ген) чаще сохраняются в регуляторных блоках, чем вне их. На уровне порядка Enterobacteriales эти взаимодействия консервативнее в парных взаимодействиях, по сравнению с регуляторными блоками, что может быть результатом изменчивости транскрипционной сети.

5. Разные типы треугольников на уровне порядка Enterobacteriales эволюционируют по-разному. Регуляция как глобальным, так и локальными транскрипционными факторами в несогласованном треугольнике оказывается консервативнее, чем в

согласованном.

Список публикаций по теме диссертации

Статьи в научных журналах

1. Tsoy O., Ravcheev D., Mushegian A. «Comparative genomics of ethanolamine utilization». // Journal of Bacteriology. - 2009 – V. 191(23) – P. 7157-7164.
2. Tsoy O.V., Pyatnitskiy M.A., Kazanov M.D., Gelfand M.S. «Evolution of transcriptional regulation in closely related bacteria». // BMC Evol Biol. – 2012 – V. 12(1) – P. 200.

Тезисы конференций

1. Tsoy O., Ravcheyev D., Mushegian A. «Comparative genomics of the ethanolamine utilization pathway». // 4–th Moscow Conference on Computational Molecular Biology'09. Book of abstracts. – 2008 – P.352
2. Tsoy O., Mushegian A. «Ethanolamine utilization: comparative genomics of spoiled food». // Regulation and Evolution of Cellular Systems (RECESS, 2010), Germany, June 20-23.
3. Цой О., Остерман И. Эволюция регуляторных взаимодействий в бактериях. // Информационные технологии и системы (ИТиС'10). Сборник тезисов. – 2010 – с.343-344.
4. Tsoy O. «The evolution of transcriptional regulation in bacteria». // Regulation and Evolution of Cellular Systems (RECESS, 2010), Germany, May 10-13.
5. Tsoy O., Gelfand M. «Evolution of transcriptional regulation in Enterobacteriales». // ASBMB Symposia. Chemical, Synthetic and Systems Biology: New Directions of Biochemistry in the 21st Century, Utah, USA, October 12-16.