На правах рукописи

Герасимова Анна Викторовна

Анализ регуляции транскрипции генов дыхания в гамма-протеобактериях методами сравнительной геномики

03.00.03 Молекулярная биология

А В Т О Р Е Ф Е Р А Т

диссертации на соискание ученой степени
кандидата биологических наук

Москва - 2006

Работа выполнена в лаборатории биоинформатики Государственного научно-исследовательского института генетики и селекции промышленных микроорганизмов.

Научный руководитель:

кандидат физико-математических наук, доктор биологических наук

Михаил Сергеевич Гельфанд

Официальные оппоненты:

доктор биологических наук, профессор **В. П. Вейко**

кандидат физико-математических наук **В. Е. Раменский**

Ведущая организация:

Институт физико-химической биологии им. А.Н. Белозерского, МГУ.

Защита состоится 25 апреля 2006 года в 14 часов на заседании Диссертационного совета Д217.013.01 при Государственном научно-исследовательском институте генетики и селекции промышленных микроорганизмов по адресу 117545, Москва, 1-ый Дорожный проезд, д. 1.

С диссертацией можно ознакомиться в библиотеке ГосНИИГенетика

Автореферат разослан « 24 » марта 2006 г.

Ученый секретарь

Диссертационного совета,

кандидат биологических наук **В.И. Щербакова**

Общая характеристика работы

Актуальность темы

Изучение транскрипционной регуляции генов, кодирующих дыхательные белки важно для понимания процессов, происходящих в бактериальной клетке, при изменении среды её обитания. Например, большинство кишечных патогенов являются факультативными анаэробами, то есть, вне носителя они обитают в среде, богатой кислородом, а, попадая в организм человека, или животного, они оказываются в анаэробных условиях.

В процессе биологической очистки загрязненных почв, например, при помощи *Shewanella* spp. (Martin-Gil, 2004) анализ регуляции дыхания является одним из главных направлений изучения соответствующих штаммов.

В ответе на изменение концентрации кислорода участвуют несколько сотен генов, образующих сложные регуляторные каскады. Сами дыхательные регулоны эволюционно подвижны, и анализ бактериального дыхания даже в близкородственных геномах представляет собой объемную и сложную задачу.

Одним из современных подходов к решению этой задачи является биоинформатический, который использует тот факт, что с каждым годом секвенируется всё больше бактериальных геномов. На данный момент известны нуклеотидные последовательности для 350 полных бактериальных геномов и 570 незаконченных. Такое количество геномов позволяет проводить сравнительный анализ как на далеких, так и на близких эволюционных расстояниях, находить таксоноспецифичные особенности регулона и оценивать консервативность ядра регулона внутри семейства или вида.

В настоящей работе новые методы сравнительной геномики и большое число геномных последовательностей прокариот были использованы для анализа регуляции транскрипции генов дыхания в гамма-протеобактериях.

Цель и задачи исследования

Целью настоящей работы является применение современных методов сравнительной геномики к исследованию регуляции дыхания в гамма-протеобактериях. В работе решаются следующие задачи:

- Построение распознающего правила для участков связывания изучаемых белков-регуляторов.
- Поиск в бактериальных геномах новых генов, зависящих от переключения аэробного/анаэробного метаболизма.
- Анализ таксон-специфичной регуляции различными дыхательными регуляторами.

- Совместный анализ различных регулонов, участвующих в одном процессе – дыхании.

Научная новизна и практическое значение

В работе впервые получены следующие результаты:

- Впервые полностью проанализирована транскрипционная регуляция дыхания в геномах трех различных семейств Enterobacteriaceae, Pasteurellaceae и Vibrionaceae.
- Построены матрицы для поиска потенциальных сигналов для регуляторов дыхания FNR, ArcA, нитрат-нитритного переключателя NarP, регулятора транспорта молибдата ModE и регулятора биосинтеза НАД – NadR.
- Разработан и применен способ анализа обобщенного регулона в группе близкородственных малоизученных геномов, основанный на полном попарном сравнении регулонов.
- С помощью подробного анализа найдены несколько десятков новых индивидуальных членов регулонов, а также найдены члены обощенных регулонов, такие как опероны *atp*, *torYZ*, *nqr* и ген *b1674*.
- Показана авторегуляция генов *arcA* и *nadR* в нескольких рассмотренных геномах.

Апробация работы

Материалы исследований по теме диссертации докладывались на межлабораторных семинарах ГосНИИГенетика (2001-2006); на международных конференциях "Artifical Intelligence and Heuristic Methods for Bioinformatics" (Сан-Миниато, Италия, 2001); BGRS'2002 (Conference on Bioinformatics of Genome Regulation and Structure), BGRS'2004, Новосибирск, Россия, MCCMB'03 (Moscow Conference on Computational and Molecular Biology), MCCMB'05, Москва, Россия, ISMB'04, (Intellectual Systems for Molecular Biology) Глазго, Великобритания. Апробация диссертации состоялась на заседании секции Ученого совета по молекулярной биологии Государственного научно-исследовательского института генетики и селекции промышленных микроорганизмов 10 марта 2006 года.

Объем и структура диссертации

Диссертационная работа изложена на 116 страницах машинописного текста и состоит из трех глав – введение, обзор литературы, результаты и обсуждение – и выводов. Глава с результатами состоит из трех частей, каждая из которых начинается с краткого резюме и содержит описание выполненных автором исследований, изложение полученных результатов и их обсуждение.

Первая часть посвящена построению распознающих правил для поиска потенциальных сигналов связывания изучаемых регуляторных факторов.

Вторая часть посвящена разработке и применению методов анализа таксоноспецифичной регуляции в геномах бактерий.

Третья часть посвящена детальному анализу дыхательных регулонов в геномах гамма-протеобактерий.

Список литературы, приведенный в конце диссертации, содержит 203 наименования. Работа содержит 31 рисунок и 8 таблиц.

Содержание работы

Часть 1. Построение распознающих правил для поиска потенциальных сайтов связывания регуляторов дыхания

Для анализа и поиска новых членов регулонов необходимо создать распознающие правила, или матрицы для поиска сигнала. Если цель работы – изучить регуляцию на больших эволюционных расстояниях, то для каждой группы (семейства или вида) геномов, создаются собственные матрицы для поиска сигнала. В настоящей работе изучалась регуляция в близкородственных организмах, поэтому создание индивидуальных матриц было нецелесообразно.

Для построения матрицы поиска потенциальных сайтов связывания FNR были взяты 5'-некодирующие области перед теми генами из генома *E. coli*, для которых FNR регуляция была показана экспериментально. В качестве таких генов использовалась выборка, доступная в банке данных DPInteract (http://arep.med.harvard.edu/dpinteract/). Матрица графически представлена на рисунке 1.

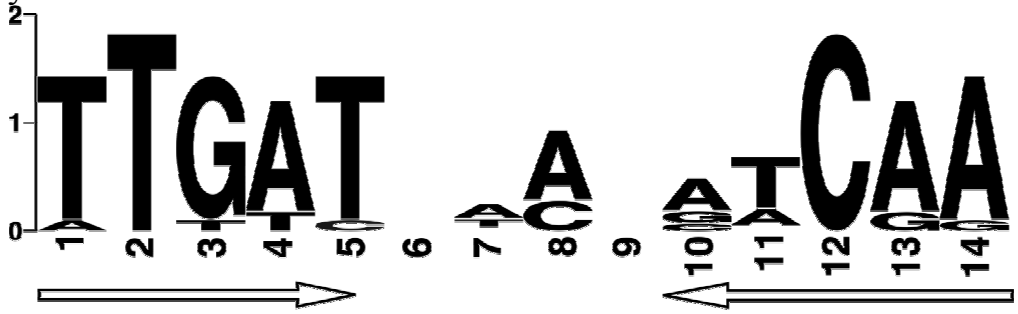

Рисунок 1. Диаграмма ЛОГО для сайтов связывания FNR.

Матрицей позиционных весов, которая и использовалась в данной работе для поиска потенциальных сайтов, называлась матрица, построенная на основе выборки операторных участков, каждый из которых, в общем случае, имеет длину *L*. Вес *w(a,i)* нуклеотида "*a*" в позиции "*i*" рассчитывался по формуле:

$$w(a,i) = log\,[N(a,i) + 0.5\,] - 0.25\,\Sigma_{b=A,C,G,T}\,\log\,[\,N(b,i) + 0.5\,],$$

где *N(a,i)* - количество нуклеотидов "*a*" из нашей выборки, которые находятся в позиции "*i*". Таким образом, используя данную матрицу, любому участку

нуклеотидной последовательности длинн L можно поставить в соответствие вес S:

$$S = \Sigma_{i=1...L}\, w(a_i,i),$$

где a_i – нуклеотид в позиции i. Далее можно ввести некоторое пороговое значение S_p, при котором, если вес S больше или равен этому значению, то данная последовательность является потенциальным операторным участком, а в противном случае – нет. Используя такой подход, можно проанализировать области, находящиеся непосредственно перед началами генов для всего генома, и определить потенциальные сайты связывания белков-регуляторов (Миронов, 1999).

Полученные правила для поиска сайтов связывания также могут быть продемонстрированы наглядно при помощи диаграмм Лого. В Лого высота каждой буквы показывает степень её консервативности, а общая высота каждой колонки – статистическую значимость данной позиции.

Чтобы построить матрицу для поиска потенциальных сайтов связывания ArcA, была использована программа SeSiMCMC, в качестве обучающей выборки были взяты 5'-некодирующие области перед следующими генами и оперонами: цитохром b убихинол оксидоредуктаза (*cydAB*), транспортер формиата (*focA*), изоцитрат дегидрогеназа (*icdA*), альдегид дегидрогеназа (*aldA*), L-лактат пермеаза (*lldP*), цитрат синтаза (*gltA*), сукцинат дегидрогеназа (*sdhCD*), супероксид дисмутаза (*sodA*), липоамид дегидрогеназа (*lpdA*) и регулятор распада малых молекул (*glcCD*). Полученный сигнал показан на рисунке 2.

Рисунок 2. Лого для поиска ArcA сигнала.

На рисунке 2 отчетливо видно, что сайт связывания белка ArcA представляет собой пятнадцатинуклеотидный неполный прямой повтор. Такая структура сайта подтверждает предположение, о том, что характерной нуклеотидной последовательностью, узнаваемой белками-регуляторами подсемейства OmpR/PhoB, является тандемный повтор. ЛОГО для остальных изученных в данной работе сигналов приведены в тексте диссертации.

Часть 2. Разработка и применение нового метода сравнительно-геномикного анализа регуляции

2.1. Анализ таксоноспецифичной регуляции

С каждым годом количество полностью секвенированных геномов растет, только в группе Enterobacteriaceae, с учетом различных штаммов, их доступно более двадцати. Такое многообразие позволило нам разработать и применить метод полного попарного сравнения геномов.

Известно, что самый хорошо изученный на сегодняшний день организм – это энтеробактерия *Escherichia coli*. Именно его обычно и использовали в качестве образца для сравнения с остальными, малоизученными бактериями. Но нельзя не считаться с тем фактом, что далеко не все гены, присутствующие, например в *Yersinia pestis*, имеют ортологов в геноме *E. coli*.

В предыдущих компьютерных исследованиях транскрипционной регуляции использовали следующую процедуру: строили распознающие правило (матрицу для поиска сигнала), при помощи этой матрицы проводили поиск потенциально регулируемых генов в геноме *E. coli*, а затем проверяли сохранение сигнала перед ортологичными генами в близкородственных геномах. Эта процедура позволяла с большой долей вероятности предсказывать регуляцию генов, имеющих ортологов в *E. coli*, однако найти новых членов регулона, свойственных другим бактериям, было невозможно.

Поэтому мы разработали новый подход для анализа таксон-специфичной регуляции. Мы провели полное попарное сравнение геномов организмов, относящихся к одной таксономической группе: в каждом из геномов проводился поиск потенциальных сайтов одного белка-регулятора. Далее на основании сохранения сайтов перед ортологичными генами определялись потенциальные члены обобщенного регулона. Если сайт перед геном сохранялся минимум в трех геномах организмов из одной группы, то ген считался членом обобщенного регулона. Для таких генов также проверялось наличие сайтов перед его ортологами в геномах организмов из других групп.

2.2. Применение метода таксоноспецифичного анализа для изучения эволюции регулона NadR в семействе Enterobacteriaceae

Убедившись в сохранении всех доменов белка NadR в геномах *E. coli* K-12 MG1655 (EC), *Shigella flexneri* 2457T (SF), *Salmonella typhi* CT18 (ST), *Y. pestis* CO92 (YP), *Yersinia enterocolitica* 8081 (YE), *Photorhabdus luminescens* subsp. *laumondii* TT01 (PHL), *Erwinia carotovora* subsp. *atroseptica* SCRI1043 (ERW), *Klebsiella pneumoniae* MGH78578 (KP) и *Serratia marcescens* Db11 (SM), мы применили процедуру попарного сравнения геномов и изучили эволюцию регулона NadR в Enterobacteriaceae.

Мы считали, что ген может быть потенциальным членом регулона NadR, если сайт сохраняется как минимум перед четырьмя ортологичными генами из рассмотренных геномов. Результаты приведены в таблице 1.

	nadB	nadA	pncB	nadR	ynfL/M	rpsP
геном	имя гена	имя гена	имя гена	имя гена	имя гена	имя гена
EC	**nadB**	**nadA**	**pncB**	nadR	**ynfL/M**	rpsP
SF	**вне рамки**	**nadA**	**pncB**	nadR	**ynfL/M**	rpsP
ST	**STY2834**	**STY0797**	**STY1010**	nadR	STY1578/79	STY2863
KP	**nadB**	**nadA**	**pncB**	nadR	**ynfL/M**	вне рамки
ERW	nadB	**ECA1378**	pncB	**ECA0463**	**ECA2259/60**	**ECA3359**
SM	nadB	**nadA**	**pncB**	**nadR**	**ynfL/M**	**вне рамки**
YP	nadB	**nadA**	**pncB**	**nadR**	**вне рамки /YPO2266**	**rpsP**
YE	RYE01420	**RYE03344**	RYE02025	**RYE00967**	**RYE00573/74**	**RYE01243**
PHL	nadB	**plu1468**	pcnB	nadR	**plu2225/24**	**rpsP**

Таблица 1. Гены потенциального регулона NadR. Поиск проводился в интервалах от (–300) пар нуклеотидов до (+10) пар нуклеотидов от старта трансляции. Пороговое значение веса сайта 4.6. Жирным шрифтом показаны гены с сайтом выше порогового. Фраза “Вне рамки” означает, что нам удалось найти ортологичный ген только по нуклеотидной последовательности.

Видно, что кроме уже известных членов NadR регулона, появились новые кандидаты. Был проведен детальный анализ этих кандидатов с применением метода филогенетического футпринтинга всех 5’-некодирующих областей генов из таблицы 1. В основе метода филогенетического футпринтинга лежит поиск консервативных последовательностей перед ортологичными генами из родственных геномов с помощью выравнивания (McCue, 2001).

Как было отмечено в таблице 1, правилу сохранения сайта как минимум в четырех из рассмотренных девяти геномов, удовлетворяет также и ген *nadR*, кодирующий белок-регулятор NadR. Потенциальный сайт связывания NadR был найден в *E. carotovora, S. marcescens, Y. pestis* и *Y. enterocolitica.* Выравнивание 5’-некодирующих областей *nadR* приведено на рисунке 3.

```
                                                          TGTTTA      TAAACA
YE|     -TTTAGTCGTGTTCATCGCGC--AGCATACTGTGTTATTTTTGTTTAGTGCGCTAAACAA
YP|     -TTTAATAGTGCCAATCCCGC--AGCATACTGTGTTATTTTTGTTTAGTACGCTAAACAA
SM|     TTTTATTAATGTCGCTGACGGCGAGCAAGATATG-CTATTTTGTTTAGCATATTAAACGG
ERW|    ---GAGTAGCGGACGTTGCCT----TATCTTGTG-GTATTTTGTTTAGTAGGCTAAACAA
            * *   *    *  *       *   * ** ************     *****

YE|     GGAGGCGG---TATGCTGCAGTTCGACTATCTCAAAACAGCGATTAAGCAAAAAGGCT
YP|     GGAGACCG---CATGCTGCAGTTCGACTATCTCAAAACGGCAATTAAGCAAAAAGGCT
SM|     AGGTGGC----CATGCCGCAATTTGATTACCTGAAGACGTCCATCAAGCAAAAGGGTT
ERW|    GAGTGTTGCACCATGTCATCATTTGATTACCTGAAATCCGCTATCCGCCAGAAGGGTT
                   ***      ** ** ** ** **  *  * **    ** ** ** *
```

Рисунок 3. Сохранение сайта связывания NadR в 5’-некодирующей области гена *nadR*. Сайт отмечен серым, жирным шрифтом выделены позиции, совпадающие с консенсусом и стартовый кодон ATG.

На рисунке 3 видно, что область потенциального связывания NadR весьма консервативна по сравнению со всем остальным межгенным промежутком.

Выравниваниях 5’-некодирующей областей генов *nadA, nadB* и *pncB* также продемонстрировали высокую консервативность сайта связывания NadR, тогда как анализ 5’-некодирующих областей генов *ynfL*, *ynfM* и *rpsP* показал большую степень сохранения всего некодирующего участка, а не только потенциальных сайтов и эти гены не были отнесены к потенциальному регулону.

Подобный анализ показал, что даже очень простой регулон, отвечающий за необходимый метаболический путь, может заметно различаться в весьма близкородственных организмах. Варьировать может не только набор регулируемых генов, но даже авторегуляция. Впервые предсказанная в этой работе авторегуляция *nadR* является особенностью нескольких, но далеко не всех организмов из семейства энтеробактерий.

Одним из возможных объяснений этого факта может быть предположение, что регулон NadR относительно молод, поскольку он присутствует только в одном семействе (Enterobacteriaceae) из всех гамма-протеобактерий.

Часть 3. Анализ дыхательных регулонов в геномах гамма-протеобактерий

3.1. Анализ регулона FNR

С использованием матрицы для поиска FNR сигнала, описанной выше, был проведен поиск потенциальных FNR-сайтов, расположенных перед другими генами *E. coli*. Получилось, что при пороге 4.0 потенциальные FNR-сайты обнаруживаются перед 121 геном, в то время как, при построении весовой матрицы использовались сайты для 9 из 121 отобранных генов (полученные из базы данных регуляторных сайтов DPInteract). Используя стандартную процедуру сравнения близкородственных геномов, мы обнаружили, что геномы всех рассмотренных бактерий, (*S. typhi, K. pneumoniae, Y. pestis, H. influenzae, V. cholerae* и *P. aeruginosa)* содержат гены, ортологичные *fnr*, что указывает на консервативность FNR-регулона у этих гамма-протеобактерий.

Ортологи 121 гена *E. coli,* перед которыми найдены потенциальные FNR-боксы, идентифицированы и в других геномах. Построенную весовую матрицу применили для анализа 5'-областей ортологичных генов. В работе получены 39 генов, 5'-области которых содержат потенциальные FNR-сайты в геномах как минимум трех разных бактерий, одна из которых – *E. coli*. Эти гены были разделены на три группы. В первую группу вошли 9 генов, сайты которых использовали для построения распознающей матрицы. Вторая группа состоит из 12 генов, регуляция которых белком FNR показана экспериментально, но сами FNR-сайты не выявлены. В третью группу вошли 18 генов, 5'-области которых содержат потенциальные FNR-боксы как минимум в двух геномах, кроме *E. coli*, при том, что FNR-регуляция этих генов не изучена.

Хорошо известно, что белок FNR гомологичен регулятору CRP и что сигналы связывания этих двух регуляторов похожи. Поэтому нельзя было исключить, что часть из предсказанных сайтов на самом деле являются сайтами связывания CRP. С использованием весовой матрицы для поиска CRP-сайтов (Gelfand, 2000) были получены веса потенциальных сайтов связывания CRP для отобранных генов из FNR-регулона. Видно, что многие известные гены из FNR-регулона (первая и вторая группы таблицы 2.2 из текста диссертации) имеют потенциальный CRP-сайт. Кроме того из литературы известно, что гены *ansB* и *tdcA* находятся под двойной регуляцией CRP и FNR (Green, 1996), (Chattopadhyay, 1997**).** Наконец, CRP регулирует ген *mtlA*, входящий в третью группу (Ramseier, 1995). Нам удалось обнаружить перед этим геном несколько потенциальных сайтов связывания CRP, и их вес выше, чем вес потенциального FNR-сайта. Гипотетический ген *b2503* мы отнесли к генам, регулируемым FNR, так как выравнивание 5'-областей ортологичных ему генов выявило консервативность именно области потенциального FNR-сайта.

К потенциальному FNR-регулону относится, по-видимому, и ген *aldA*. Этот ген не имеет ортологов в рассматриваемых нами геномах, и поэтому на основании формального критерия не может быть отнесен к регулону. Однако в геноме *E. coli* перед ним находится потенциальный FNR-сайт, и известно (Pellicer, 1999), что этот ген регулируется ArcA и CRP (сайт связывания CRP не совпадает с найденным нами FNR-сайтом), то есть принадлежит к известным регулонам центрального метаболизма. Всё это даёт возмодность предсказывать потенциальную FNR-регуляцию гена *aldA*.

Таким образом, были идентифицированы сайты связывания FNR в 5’-областях 12 генов *E. coli*, регулируемых FNR, и найдены еще 17 генов, которые предположительно принадлежат к FNR-регулону. Описаны FNR-регулоны *S. typhi, K. pneumoniae, Y. pestis, H. influenzae, V. cholerae* и *P. aeruginosa*.

3.2. Анализ регулона ArcA

Поиск по обеим цепям генома *E.coli* потенциальных сайтов связывания (сайтов) ArcA, соответствующих матрице, описанной в части 1, при пороге 4.25 обнаружил потенциальные сайты связывания ArcA в 5’-областях 257 генов.

С помощью стандартной процедуры поиска ортологичных генов в геномах гамма-протеобактерий, родственных *E.coli* (*Y. pestis, P. multocida* и *V. vulnificus*), были выявлены гены, ортологичные гену белка-регулятора ArcA. Это позволило предположить наличие регулона ArcA в этих бактериях и с помощью методов сравнительной геномики существенно сузить результаты поиска в *E.coli*. Для этого с помощью комплекса программ "Genome Explorer" были проанализированы полные геномы четырех родственных гамма-протеобактерий *E. coli*, *Y. pestis*, *P. multocida* и *V. vulnificus*. Для каждого из отобранных ранее 257 генов из *E.coli* с потенциальными сайтами связывания ArcA, искали ортологи в остальных геномах. В случае, если хотя бы два гена-

ортолога из трех рассмотренных геномов содержали в своей 5'-некодирующей области потенциальный сайт ArcA с весом выше порогового (4.00), ген считался имеющим консервативный сайт ArcA.

Мы обнаружили в геноме *E.coli* 23 гена с консервативным сайтом ArcA (таблица приведена в тексте диссертации). Из них 14 упомянуты в литературе как регулируемые кислородно-зависимыми механизмами. Следовательно, описанная процедура как минимум позволяет распознавать гены, регулируемые одним (или несколькими) из кислородно-зависимых регуляторов.

Обучающая выборка состояла из ArcA-регулируемых генов. Известные сайты узнавания других кислородных регуляторов отличаются от использованного нами мотива ArcA, который был эволюционно консервативен, и, следовательно, биологически функционален в проанализированных геномах. Интересно, что одним из найденных 23 генов оказался ген самого белка ArcA. Это позволяет предполагать наличие обратной связи в механизме его регуляции (автоторегуляции).

3.3. Анализ регулона NarP

В общей сложности было рассмотрено 13 геномов различных организмов, относящихся к группе гамма-протеобактерий: *E. coli* K12 (EC), *S. typhi* Ty2 (ST), *Erwinia carotovora* subsp. *atroseptica* (EO), *Y. pestis* KIM (YP) *Haemophilus ducreyi* 35000HP (HD), *Haemophilus influenzae* Rd (HI), *P. multocida* Pm70 (PM), *V. cholerae* O1 (VC), *Vibrio parahaemolyticus* RIMD 2210633 (VP), *V. vulnificus* CMCP6 (VV), *Y. enterocolitica* (YE), *Actinobacillus actinomycetemcomitans* HK1651 (AA) и *Vibrio fischeri* ES114 (VF).

В исследуемых геномах проводился поиск сайтов с весом выше порогового значения 3,50. Однако в геноме *A. actinomycetemcomitans* перед генами-потенциальными членами регулона зачастую удавалось обнаружить лишь сайты с весом ниже принятого порогового значения. Поэтому для данного генома было установлено пороговое значение 3,25. При данных пороговых значениях потенциальные сайты связывания NarP обнаруживались в каждом геноме перед приблизительно 400 генами. Ясно, что отдельные предсказания таких сайтов недостоверны.

Применив метод таксон-специфичного анализа регуляции, описанный выше, и проанализировав NarP-регуляцию внутри трёх групп, в общей сложности к обобщенному регулону было отнесено 77 генов, организованных как минимум в 29 оперонов. В исследованных геномах обобщенный NarP-регулон включает в себя почти все гены, входящие в объединенный NarL-NarP-регулон в *E. coli.*

3.4. Анализ регулона ModE

В семействе Enterobacteriaceae был проведен детальный анализ ModE регулона. С помощью матрицы для поиска ModE сигнала, описанной выше, в областях [-400; 0] нуклеотидов от старта трансляции при пороге 4.4 был

проведен поиск потенциальных сайтов в геномах *E. coli* K-12 MG1655, *S. typhi* CT18, *Y. pestis* CO92, *Y. enterocolitica* 8081, *P. luminescens* subsp. *laumondii* TT01 (PHL), *E. carotovora* subsp. *atroseptica* SCRI1043 (ERW), *K. pneumoniae* MGH78578 (KP) и *S. marcescens* Db11 (SM).

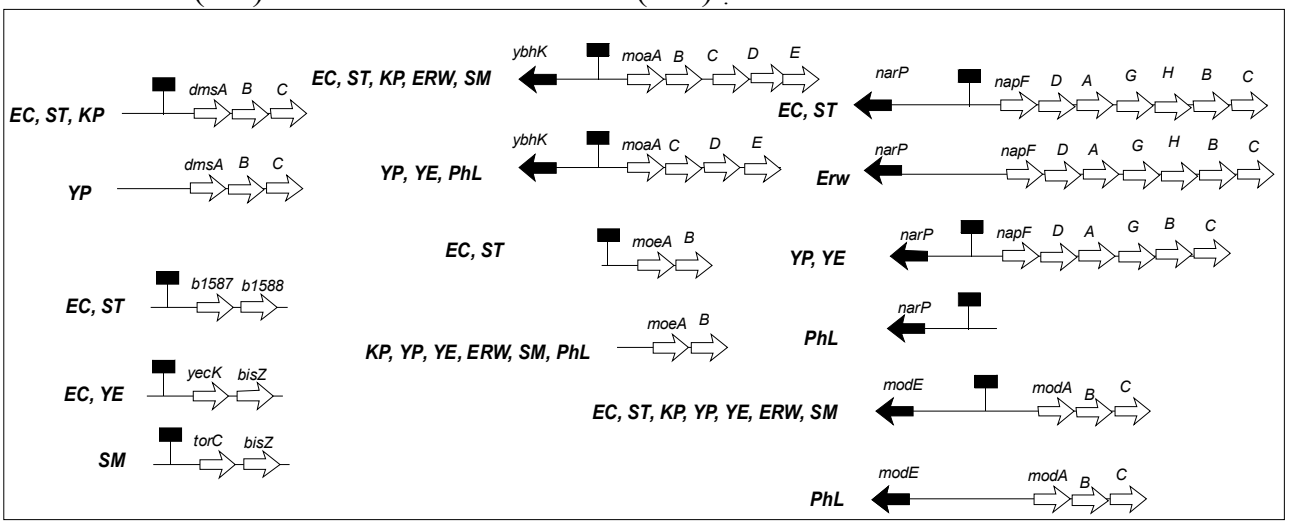

Рисунок 4. Известные ModE-регулируемые опероны.
Обозначения. Черным прямоугольником обозначен ModE сигнал, белыми стрелками – опероны, черными стрелками – дивергентные гены.

На рисунке 4 схематически изображены опероны, для которых в геноме *E. coli* была показана регуляция ModE. Здесь же можно видеть, что произошло с опероном, дивергоном и сайтами в близкородственных организмах.

На рисунке 5 приведены новые потенциальные члены регулона ModE, которых нам удалось идентифицировать. Это гены *STY3313* и *STY3312*, входящие в один оперон из *S. typhi* и их ортологи в геномах *Erwinia carotovora, Klebsiella pneumoniae* и *Serratia marcescens.* Эти гены кодируют потенциальный транспортер молибдена. Регуляция потребителей – частое свойство кофакторных регулонов, поэтому обнаружение сохраняющегося сайта связывания ModE перед этим опероном делает регуляцию весьма вероятной.

Кроме того, консервативный сайт был найден перед геном оксидоредуктазы, содержащей молибдоптеприн (*STY0659*) и опероном, кодирующим анаэробную сульфит редукатзу *asrABC* из генома *S. typhi.*

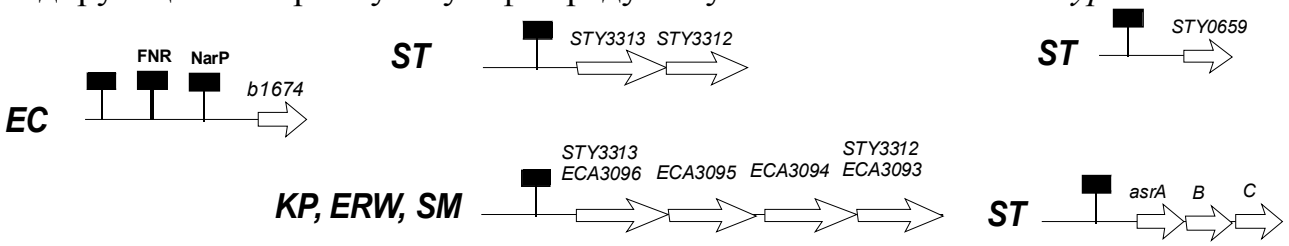

Рисунок 5. Новые ModE-регулируемые опероны. Черным прямоугольником обозначен ModE сигнал, белыми стрелками – опероны.

Ген *b1674* кодирует уникальную оксидоредуктазу, присутствующую только в геноме *Escherichia coli*, потому стандартные методы сравнительной геномики не позводили бы нам определить этот ген, как участник какого либо регулона. Но в этом случае мы обнаружили в 5'-области оксидоредуктазы (*b1674*) потенциальные сайты связывания не только регулятора ModE, но также NarP и FNR. Как известно, многие участники ModE-регулона входят также и в FNR-регулон, так что ген *b1674* был отнесен нами к этим дыхательным регуонам. Следует отметить, что наше предположение было впоследствие подтверждено экспериментально, и сайт ModE перед *b1674* оказался функциональным (Rob Gunsalus, Университет Калифорнии, Лос-Анжелес, частное сообщение).

3.5. Комплексный анализ дыхательных регулонов

Изучение дыхательной регуляции предоставило возможность для комплексного анализа нескольких регуляторных систем, реагирующих на сходные изменения окружающей среды.

Для трёх групп бактерий: Enterobacteriaceae, Pasteurellaceae и Vibrionaceae, был проведен детальный анализ регулонов FNR, ArcA и NarP. Результаты приведены в таблице 3.5.1 в тексте диссертации, а история оперонных перестроек отражена в таблице 3.5.2 из текста диссертации.

Гены, попавшие в эти таблицы, могут быть разделены на несколько функциональных групп.

Дыхательные ферменты

Для оперонов *cydAB*, *nap*, *ccm*, *nrf*, *dms*, *torCAD*, *frd*, *fdn*, *ndh*, *glpABC* и *glpD* из *E. coli* регуляция как минимум одним из рассматриваемых регуляторов была показана экспериментально. Мы провели детальный анализ таксон-специфичной регуляции этих генов и предсказали регуляцию некоторых других генов, участвующих в дыхании.

Одним из наиболее важных результатов была идентификация консервативных сайтов в 5'-некодирующих областях оперонов *atpIBEFHAGCD*. Этот оперон кодирует все субъединицы АТФ-синтетазного комплекса, который является важным компонентом дыхательного пути. Экспериментальные данные о регуляции этого оперона в *E. coli* противоречивы. В частности, в работе (Kasimoglu, 1996) утверждалось, что экспрессия *atp* оперона не зависит от FNR или ArcA. С другой стороны, было показано увеличение экспрессии различных *atp* генов в 1.5 – 8 раз в *fnr*-мутантном штамме *E. coli* по сравнению с диким штаммом (Salmon, 2003). В нашей работе потенциальные ArcA-сайты были найдены в 5'-некодирующих областях *atp* оперона в двух геномах *Yersinia*, а FNR-сайты во всех геномах Vibrionaceae. Таким образом, по крайней мере в нескольких геномах гамма-протеобактерий экспрессия генов АТФ синтетазы, по-видимому, контролируется глобальными регуляторами дыхания.

Перед Na^+-экспортирующей NADH-дегидрогеназой (*nqrABCDEF*) также были найдены потенциальные регуляторные сайты. Потенциальные сайты

связывания FNR сохранились во всех рассмотренных геномах, тогда как сайты связывания ArcA были найдены во всех геномах Vibrionaceae и Pasteurellaceae, но не в *Yersinia* spp., а NarP-регуляция вообще специфична для Pasteurellaceae.

В геноме *E. coli* есть два оперона, кодирующих редуктазы три-метиламин-оксид (TMAO) азота, *torCAD* и *torYZ* . Экспрессия *torCAD* репрессируется регулятором NarL и активируется TorR (Iuchi, 1987. Simon, 1994), тогда как экспрессия *torYZ* конститутивна (Gon, 2000). В тех трёх геномах Pasteurellaceae, где гены *torCAD* отсутствуют, перед *torYZ* опероном были найдены сайты для всех трех дыхательных регуляторов. В геномах Vibrionaceae транскрипция the *torCAD* контролируется FNR, а *torYZ* – NarP.

Аналогичная ситуация имеет место и в случае с генами формиат редуктаз. В *E. coli* оперон *fdn* регулируется факторами FNR, NarL и NarP (Li, 1992, Wang, 2003), тогда как экспрессия оперона *fdo* постоянна (Abaibou, 1992). Но в геномах *Y. pestis* и *Y. entercolitica* гены *fdn* отсутствуют, и потенциальные сайты белков FNR, ArcA и NarP были найдены перед опероном *fdo*.

Потенциальные сайты связывания регулятора ArcA были обнаружены перед геном *fdhD*, кодирующим белок, участвующий в формировании белковых комплексов Fdn и Fdo в геномах Pasteurellaceae и *Yersinia* spp. Более того, перед геном *fdhD* мы обнаружили сайты связывания NarP в двух геномах *Yersinia* spp. Интересно, что потенциальные регуляторные сайты были найдены только в геномах, содержащих гены *fdo* или *fdn*.

В 5’-некодирующей области оперона *dadAX* в геномах *Yersinia* spp. и Vibrionaceae мы обнаружили потенциальные сайты ArcA. Первый ген этого оперона кодирует малую субъединицу дыхательной D-аминокислотной дегидрогеназы (Lobocka, 1994). Таким образом, работа D-аминокислот, как электронных доноров, в некоторых геномах может контролироваться ArcA.

Синтез молибденового кофактора

Молибденовый кофактор – необходимый компонент некоторых дыхательных дегидрогеназ и редуктаз, например, комплексов Fdn, Fdo, Nap, Dms и Tor (Gennis *et al*., 1996). Как было сказано выше, регуляция экспрессии оперонов *moa* и *moe* в геноме *E. coli* осуществляется разными дыхательными регуляторами, см. раздел 3.2. Во всех геномах за исключением *V. fischeri* в 5’-некодирующей области оперона *moa* были найдены сайты хотя бы для одного регулятора дыхания. Сохранение потенциальных сайтов FNR и ArcA перед опероном *moe* было отмечено только в геномах Pasteurellaceae.

Центральный метаболизм и брожение

В различных экспериментальных работах было показано, что экспрессия оперонов *pdhR-aceEF-lpdA*, *pflB*, *yfiD*, *gltA*, *sucABCD*, *sdhCDAB*, *aspA*, *fumC*, *mdh*, *ldhA* и *adhE* в *E. coli* регулируется хотя бы одним глобальным регулятором дыхания, см. ссылки в таблице 3.5.1 из текста диссертации. Перед генами, участвующими в гликолизе, глюконеогенезе, пентозо-фосфатном пути,

метаболизме пирувата и лактатата также были найдены искомые потенциальные сайты.

Ген *sfcA* кодирует малатдегидрогеназу, участвующую в пути глюконеогенеза. В геномах *Yersinia* spp. перед ним были найдены сайты связывания NarP, а FNR-сайты были обнаружены в группе Vibrionaceae.

Фермент глюконеогенеза, фосфоэнол пируват синтаза кодируется геном *ppsA*. Перед ним были найдены потенциальные сайты связывания FNR в *Yersinia* spp. и ArcA в Vibrionaceae.

Ещё один ген из этого пути, *pckA*, кодирующий фосфоэнол-пируват карбоксигеназу, имеет потенциальные сайты ArcA во всех геномах Vibrionaceae и Pasteurellaceae, за исключением *P. multocida*.

Перед геном *eno*, кодирующим энолазу и участвующим в гликолизе/глюконеогенезе в Pasteurellaceae были найдены потенциальные сайты связывания FNR и NarP. Мы также предсказали дыхательную регуляцию для гена *pgk,* кодирующего фосфоглицерат-киназу: консервативные сайты связывания NarP были обнаружены в Pasteurellaceae, а FNR – в Vibrionaceae.

Ген *talB* кодирует транс-альдолазу В, фермент пентозно-фосфатного пути. В геномах Pasteurellaceae мы обнаружили потенциальные сайты NarP в 5'-некодирующей области гена *talB*.

Кроме того, потенциальные сайты FNR и ArcA были найдены во всех Vibrionaceae перед геном *aldB*. Этот ген кодирует альдегид дегидрогеназу, фермент, участвующий в метаболизме пирувата.

Метаболизм углеводов

Регуляция генов сахарного метаболизма белками-регуляторами дыхания не была изучена экспериментально. Однако мы обнаружили потенциальные сайты связывания FNR, ArcA и NarP перед некоторыми оперонами сахарного метаболизма. Нельзя назвать это неожиданным результатом, поскольку метаболизм сахаров близок центральному метаболизму, и сахара поставляют субстраты для производства энергии.

Гены *ptsH* и *ptsI* кодируют белки HPr и EI общего компонента всех фосфоэнол-пируват: карбогидрат фосфотрансферазных систем PTS. В *E. coli* гены *ptsHI* контранскрибируются с *crr*, кодирующим EIIA компонент глюкозо-специфичной PTS (Ryu, 1995). Структура этого оперона сохраняется во всех рассмотренных геномах. В Pasteurellaceae этот оперон предсказан как ArcA-регулируемый.

Оперон *mtlADR* кодирует компонент EIIABC маннитол-специфичной PTS, манитол-1-фосфат дегидрогеназу и репрессор этого оперона. ArcA-регуляция была предсказана для геномов *Yersinia* spp., FNR-регуляция – для Vibrionaceae.

Перед опероном *deoCABD* в геномах Vibrionaceae нами были найдены потенциальные сайты связывания NarP. Гены этого оперона отвечают за деградацию пиримидиновых дезоксинуклоизидов до ацетальдегида и глицерол-3-фосфата (Neughard, 1990).

Во всех бактериях Vibrionaceae перед генами *nagB*, кодирующими глюкозамин-6-деаминазу, были найдены потенциальные сайты связывания FNR.

Кластер генов *glgBXCAP* содержит гены, отвечающие за биосинтез гликогена и катаболизм. Эти гены образуют один оперон в геномах *Yersinia* spp. В Pasteurellaceae генам *glg* предшествует ген *malQ,* который мы тоже относим к потенциальному оперону *malQ-glgBXCAP*. В обеих *Yersinia* spp. потенциальные сайты связывания FNR были найдены перед геном *glgB*, тогда как в Pasteurellaceae сайты связывания FNR и NarP обнаружились перед геном *malQ*. Сохранение этих сайтов, несмотря на оперонные перестройки, говорит о том, что гены *glg* действительно являются членами дыхательных регулонов.

Метаболизм жирных кислот

Регуляция оперонов, участвующих в метаболизме жирных кислот, не была показана экспериментально, однако имелись некоторые косвенные подтверждения (Iuchi and Lin, 1988). Перед некоторыми из этих оперонов есть консервативные сайты связывания, и поэтому, скорее всего, они действительно входят в дыхательные регулоны.

Опероны *fabBA* и *fadIJ* гомологичны и кодируют белковый комплекс для бета-окисления жирных кислот. Белковый комплекс FadBA принимает участие в деградации жирных кислот и в аэробных, и в анаэробных условиях, тогда как FadIJ в основном работает в анаэробных условиях (Clark and Cronan, 1996, Campbell *et al*., 1993). Эффект от мутации гена *arcA* на эспрессию *fadBA* оперона был отмечен в работе (Iuchi and Lin, 1988), но прямая регуляция показана не была. В *Yersinia* spp. сохранение потенциальных сайтов связывания ArcA обнаружено перед опероном *fadIJ*, а сайтов связывания FNR и ArcA – перед опероном *fabBA*. В геномах Vibrionaceae наблюдалась обратная ситуация: сайты связывания FNR и ArcA располагались перед опероном *fadIJ*, а одиночные сайты ArcA перед опероном *fabBA*.

ArcA-регуляция гена *fadD* предсказана в геномах *Yersinia* spp. и Vibrionaceae. Ген *fadD* кодирует ацил-коА синтазу, фермент, участвующий в бета-окислении жирных кислот (Weimar, 2002).

Ещё один оперон, кодирующий белки синтеза жирных кислот (*acpP-fabF*), может участвовать в регуляции дыхания. Оба белка AcpP и FabF вовлечены в биосинтез жирных кислот (Cronan and Rock, 1996). Консервативные сайты связывания ArcA перед этим опероном были найдены в геномах *Yersinia* spp., а сайты связывания FNR в Vibrionaceae.

Можно предпололжить несколько возможных объяснений регуляции метаболизма жирных кислот в зависимости от способа дыхания. Во-первых, метаболизм жирных кислот близок к центральному метаболизму, например через ацетил-КоА (Lin, 1996, Hassan, 1992) . Во-вторых, одни ферменты окисения жирных кислот предпочитают анаэробные условия, а другие – аэробные (Campbell, 2003).

Кислородный стресс

Было показано, что экспрессия гена супероксид-дисмутазы (*sodA)* в геноме *E. coli* регулируется белками FNR и ArcA (Hassan and Sun, 1992). Подобная регуляция сохраняется в геномах Pasteurelaceae, а в *Yersinia* spp. были обнаружены только потенциальные сайты ArcA.

Нуклеотид редуктазы

Известно, что FNR активирует оперон *nrdDG* в геноме *E. coli* в анаэробных условиях (Boston, 2003), а эспрессия еще одного оперона нуклеотид редуктазы, *nrdAB*, не зависит от FNR в *E. coli* (Boston, 2003). Сохранение сайтов связывания FNR в 5'-некодирующих областях *nrdDG* оперона наблюдалось в двух геномах *Yersinia* spp. и во всех Vibrionaceae. В геномах Pasteurellaceae, наоборот, потенциальные сайты FNR найдены перед *nrdAB* опероном.

Транспорт

Регуляция хотя бы одним из изучаемых регуляторов в геноме *E. coli* была показана экспериментально для следующих оперонов: *focA*, *dcuA*, *dcuB*, *dcuC*, *glpFK* и *feoAB*, см. ссылки в таблице 3.5.1 в тексте диссертации. Регуляция некоторых других транспортных генов предсказана сейчас.

Потенциальные сайты связывания ArcA обнаружены перед геном *gltP* в геномах *Yersinia* spp. и Vibrionaceae. Продукт этого гена – траспортер глутамата и аспартата (Wallace, 1994). Транспорт дикарбоксилатов контролируется дыхательными регуляторами, например, экспрессия генов *dcuA*, *dcuB* и *dcuC*, кодирующих транспортеры дикарбоксилата, в геноме *E. coli* регулируется белком FNR (Golby, 1998 и Zientz, 1999) .

Оперон *gntXY* кодирует белок из системы транспорта глюконата (Porco, 1998). Потенциальные сайты связывания ArcA и FNR найдены перед *gntXY* опероном во всех геномах Pasteurellaceae.

Наконец, регуляция была предсказана для гена *fadL,* кодирующего транспортер жирных кислот (Black, 1991). Потенциальные сайты связывания FNR и ArcA были найдены в геномах *Yersinia* spp., а сайты связывания ArcA – в Vibrionaceae.

T-пептидаза

Экспрессия гена *pepT*, кодирующего аминотрипептидазу T (Lombardo, 1997), активируется регулятором FNR в *S. typhimurium* (Strauch, 1985). Это предсказание подтверждается высокой консервативностью сайтов связывания FNR, обнаруженных во всех рассмотренных геномах, таблица 3.5.1 из текста диссертации.

Регуляторы транскрипции

Регуляция генов *fnr*, *arcA*, *narP* и *narQ* различными дыхательными регуляторами в геноме *E. coli* была показана экспериментально (ссылки в таблице 3.5.1 в тексте диссертации). Были показаны существенные перестройки этого регуляторного каскада в других бактериях. В качестве примера на рисунке 6 показаны регуляторные каскады, для геномов семейства

Vibrionaceae. Каскады для остальных семейств приведены в тексте диссертации.

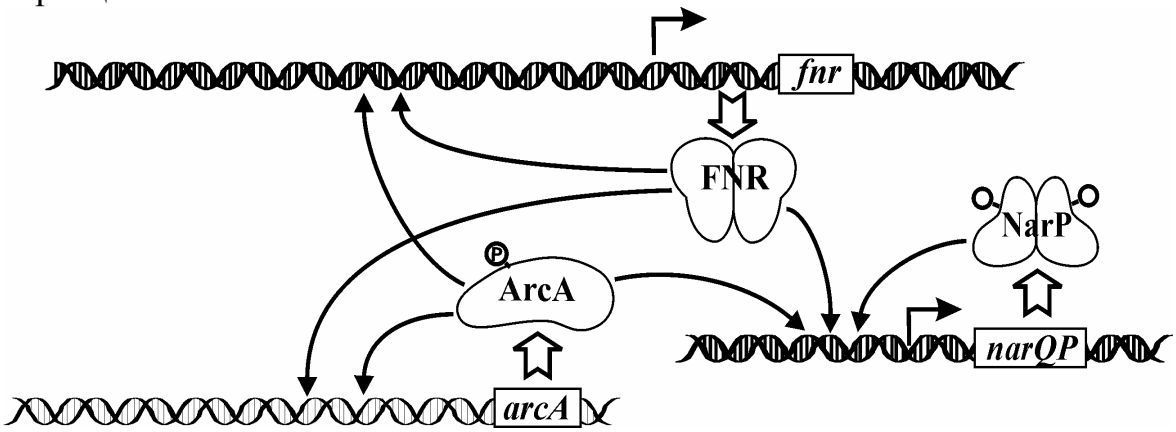

Рисунок 6. Регуляторный каскад в геномах Vibrionaceae.

Fis – это широко представленный белок, отвечающий за архитектуру ДНК. Он регулирует сверхспирализацию бактериальной ДНК (Browning, 2005). Экспериментально было показано, что в *E. coli* Fis вместе с FNR и/или ArcA регулирует экспрессию *nrf*, *nir*, *ndh, adhE*, *yfiD*, *sdhCDAB-sucABCD*, *acnB* и *narK* оперонов. В геноме *E. coli* ген *fis* ко-транскрибируется с геном *dusB* (Ninnemann, 1992), и структура кластера *dusB-fis* сохраняется во всех рассмотренных геномах. В группе Pasteurellaceae потенциальные сайты связывания FNR и ArcA были найдены перед геном *dusB*. Таким образом, можно предположить, что в рассмотренных геномах транскрипция гена *fis* контролируется дыхательными регуляторами. Так как белок Fis контролирует экспрессию дыхательных генов, то, возможно, FNR и ArcA, регулируя экспрессию гена *fis*, таким образом влияют и на экспрессию собственных генов.

Оперон *cpxRA* кодирует двухкомпонентную регуляторную систему, участвующую в стресс-ответе (Raivio, 2005). Некоторые данные указывают на перекрестное взаимодействие между системами CpxR-CpxA и ArcA-ArcB в *E. coli* (Iuchi, 1989). В геномах Pasteurellaceae были обнаружены консервативные сайты связывания FNR и ArcA перед опероном *cpx*.

OxyR – фактор транскрипции ответа на дыхательный стресс (Christman, 1989). Во всех геномах Pasteurellaceae, были обнаружены потенциальные сайты связывания ArcA перед генами *oxyR*. Так как ArcA проявляет активность как фактор транскрипции только в условиях отсутствия кислорода, возможно, что в этих условиях он репрессирует экспрессию сенсора *oxyR*.

Ген *fur*, кодирующий регулятор транспорта железа, в 5’-некодирующей области содержит потенциальные сайты FNR в геномах Pasteurellaceae, и ArcA в Vibrionaceae. Нельзя сказать, что наши выводы о дыхательной регуляции Fur неожиданны, потому как железо – неотъемлемый компонент большинства дыхательных комплексов (Gennis and Stewart, 1996) и самого фактора FNR (Sutton, 2004).

Таким образом, в данной работе три дыхательных регулона в десяти геномах, принадлежащих к трем семействам были изучены сначала независимо, потом данные были суммированы и найдены возможные дополнительные члены обобщенных регулонов. Потенциальные члены регулонов были разбиты на десять функциональных групп, и для каждого оперона был проведен анализ литературных данных, дающий нашим предсказаниям больший или меньший вес. К обобщенному потенциальному дыхательному регулону мы относим 68 оперонов из группы гамма-протеобактерий.

Выводы

1. Впервые полностью проанализирована транскрипционная регуляция дыхания в геномах трех различных семейств Enterobacteriaceae, Pasteurellaceae и Vibrionaceae.
2. Были построены матрицы для поиска потенциальных сигналов для регуляторов дыхания FNR, ArcA, нитрат-нитритного переключателя NarP, регулятора транспорта молибдата ModE и регулятора биосинтеза НАД – NadR.
3. Разработан и применен способ анализа обобщенного регулона в группе близкородственных малоизученных геномов, основанный на полном попарном сравнении регулонов.
4. С помощью подробного анализа были найдены несколько десятков новых индивидуальных членов регулонов, в том числе, была предсказана потенциальная FNR-регуляция оперона *narQP* в геномах Vibrionaceae, ArcA-регуляция оперона *adhE* в группе геномов Vibrionaceae и *frdA* в Pasterellaceae и Yersinia, регуляция оперона *moa* белком NarP в Pasterellaceae Vibrionaceae.
5. Также были найдены члены обощенных регулонов, такие как опероны *atp*, *torYZ*, *nqr* и ген *b1674*.
6. Впервые показана потенциальная авторегуляция гена *arcA* в геномах *Yersinia* spp., Vibrionaceae и *H. ducreyi*.
7. Детально проанализирована регуляция транскрипции биосинтеза НАД в девяти геномах из семейства Enterobacteriaceae. Впервые показана авторегуляция гена *nadR* в геномах *E. carotovora, S. marcescens, Y. pestis* и *Y. enterocolitica.*

Список работ, опубликованных по теме диссертации

Публикации в научных журналах:

1. A.V. Gerasimova, M.S. Gelfand. Evolution of the NadR regulon in Enterobacteriaceae // Journal of Bioinformatics and Computational Biology, 2005, V.3 (4), pp.1007-1019.
2. A.V. Favorov, M.S. Gelfand, A.V. Gerasimova, D.A. Ravcheev, A.A. Mironov, V.J. Makeev. Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length // Bioinformatics, 2005, V.21(10), pp.2240–2245.
3. А.В. Герасимова, М.С. Гельфанд, В. Ю. Макеев, А.А. Миронов, А.В. Фаворов. Регулятор-ArcA в гамма-протеобактериях. Определение сайтов-связывания и описание регулона // Биофизика, 2003, т.48(1), стр. 21-25.
4. А.В. Герасимова, Д.А. Родионов, А.А. Миронов, М.С. Гельфанд. Компьютерный анализ регуляторных сигналов в бактериальных геномах. Участки связывания Fnr // Молекулярная Биология, 2001, т.35(6), стр. 1001-1010.

Тезисы международных научных конференций:

1. A.V. Gerasimova, M.S. Gelfand. Comparative analysis of transcriptional regulation in E. coli and relative genomes: FNR, ArcA, NarP and ModE regulons // Conference on New Frontiers in Microbiology and Infection, 2005, pp.26-27, Villars-sur-Ollon, Switzerland, September 4-8, 2005
2. A.V. Gerasimova, D.A. Ravcheev. How gamma-proteobacteria switch the mode of respiration: a comparative genomic analysis // Moscow Conference on Computational and Molecular Biology, pp. 297-299, Moscow, Russia, July 18-21, 2005.
3. A.V. Gerasimova, D.A. Ravcheev, A.B. Rakhmaninova, M.S. Gelfand. Computer analyses of aerobic-anaerobic regulation in gamma-proteobacteria // Intellectual Systems for Molecular Biology pp.111,Glasgow, United Kingdom, July 31–August 4, 2004.
4. A.V. Favorov, M.S. Gelfand, A.V. Gerasimova, A.A. Mironov, V.J. Makeev. Gibbs Sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length and its validation on the ArcA binding sites // Conference on Bioinformatics of Genome Regulation and Structure, 2004, V.2, pp.269–272, Novosibirsk, Russia, July 25-30, 2004.
5. А.В. Герасимова, Д.А. Равчеев. Компьютерное предсказание потенциальных участков связывания FNR и ArcA // Международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов-2004», стр. 10, Москва, Россия, Апрель 2004.
6. A.V. Gerasmova. Comparative genomics of FNR,- DNR,- ANR- and EtrA regulons of gamma-proteobacteria // Moscow Conference on Computational and Molecular Biology, 2003, pp.78–79, Moscow, Russia, July 22–25, 2003.

7. A.V. Favorov, A.V. Gerasimova. Yet Another Digging-for-DNA-Motifs Gibbs Sampler // Moscow Conference on Computational and Molecular Biology, 2003, pp.67–68, Moscow, Russia, July 22-25, 2003.
8. A.V. Gerasimova, D.A. Rodionov, A.A. Mironov, M.S. Gelfand. FNR/DNR/ANR–regulon in Gamma–Proteobacteria // Conference on Bioinformatics of Genome Regulation and Structure, 2002, V.2, pp.19–20, Novosibirsk, Russia, July 14–20, 2002.
9. A.V. Gerasimova. Computational analysis of regulatory sites in bacterial genomes: FNR– and ANR–binding sites // NATO Advanced Studies: Intelligent Systems for Molecular Biology. pp. 25, San Miniato, Italy, October 1–12, 2001.