

На правах рукописи

Садовская Наталья Сергеевна

**Сравнительный геномный анализ систем
метаболизма длинноцепочечных жирных кислот и
мембранных белков γ -протеобактерий**

03.01.09 Математическая биология, биоинформатика

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата биологических наук

Москва – 2012

Работа выполнена в Государственном научном центре Российской Федерации
Федеральном государственном унитарном предприятии Государственного
научно-исследовательского института генетики и селекции промышленных
микроорганизмов

Научный руководитель:

доктор биологических наук, профессор
Гельфанд Михаил Сергеевич

Официальные оппоненты:

Туманян Владимир Гайевич
доктор физико-математических наук, профессор,
Федеральное государственное бюджетное учреждение науки
Институт молекулярной биологии им. В.А. Энгельгардта
Российской академии наук, заведующий лабораторией

Опарина Нина Юрьевна
кандидат биологических наук
Федеральное государственное бюджетное учреждение
“Научно-исследовательский институт биомедицинской химии
им. В.Н. Ореховича” Российской академии медицинских наук,
старший научный сотрудник

Ведущая организация:

Федеральное государственное бюджетное учреждение науки
Институт Общей Генетики им. Вавилова Российской академии наук

Защита диссертации состоится 30 мая 2012 года в 14-00 часов на заседании
диссертационного совета Д 002.077.04 на базе Федерального государственного
бюджетного учреждения науки Института 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 диссертацией можно ознакомиться в библиотеке Федерального
государственного бюджетного учреждения науки Института проблем передачи
информации им. А.А. Харкевича Российской академии наук.

Автореферат разослан апреля 2012 года

Ученый секретарь диссертационного совета
доктор биологических наук, профессор

Рожкова Г.И.

Общая характеристика работы

Актуальность темы

Длинноцепочечные жирные кислоты являются ключевым компонентом всех липидов и таким образом представляют собой важнейшие компоненты мембран. Мембрана в свою очередь ограничивает содержимое клетки и выполняет роль барьера между цитоплазмой и окружающей средой. Согласно жидкостно-мозаичной модели, мембраны рассматривают как динамическую систему, основными составляющими которой являются липидный бислой и различные белки, обладающие широким спектром функциональной активности. Мембранные белки участвуют во всех основных функциях клетки и играют значительную роль в ее жизнедеятельности. Соответственно, полная характеристика как длинноцепочечных жирных кислот, так и мембранных белков является актуальной проблемой современной биологии, в том числе, биоинформатики.

В последнее время у исследователей появились новые возможности компьютерного анализа, обусловленные стремительным ростом количества полностью секвенированных геномов. Объем опубликованных нуклеотидных и аминокислотных последовательностей многократно превышает экспериментальные возможности их изучения. Таким образом, аннотация новых последовательностей зачастую осуществляется исключительно биоинформатическими методами и нередко определяет дальнейшие экспериментальные исследования. При этом необходимо отметить, что методы *in silico* требуют меньших временных и материальных затрат и, следовательно, имеют преимущества перед другими подходами.

Одним из таких методов является сравнительный анализ геномных последовательностей. Он позволяет выявить новые члены метаболического пути и предсказать их функции, что особенно существенно при проведении поиска недостающих членов исследуемого пути. Кроме того, этот подход дает возможность переносить уже имеющуюся информацию о регуляции от одного хорошо изученного генома на другие, менее изученные экспериментально.

Алгоритмы, позволяющие идентифицировать трансмембранные белки на основании аминокислотной последовательности, а также дающие возможность предсказать положение трансмембранных сегментов в трансмембранных белках, представляют собой другой подход, широко применяемый в современной биоинформатике. При этом качество предсказаний алгоритмов составляет около 80%, а результаты сопоставления алгоритмов, полученные различными группами исследователей, заметно различаются. Следует отметить, что работ по сравнительному анализу, выполненных исследователями, которые не разрабатывали тот или иной алгоритм, крайне мало.

Цели и задачи исследования

Цель работы – описание регуляции транскрипции, кодирующей ферменты метаболизма генов длинноцепочечных жирных кислот в γ -протеобактериях, и

разработка метода тестирования алгоритмов предсказания трансмембранных сегментов в условиях отсутствия экспериментальной тестовой выборки.

В соответствии с этим были поставлены следующие задачи:

1. провести поиск выборки известных сайтов связывания факторов транскрипции FadR и FabR, регулирующих гены метаболизма длинноцепочечных жирных кислот;
2. построить распознающее правило для поиска потенциальных сайтов связывания FadR и FabR;
3. построить ортологические ряды генов, вовлеченных в метаболизм длинноцепочечных жирных кислот и определить их регуляцию в родственных организмах;
4. провести поиск новых членов регулонов FadR и FabR;
5. определить разметку потенциальных трансмембранных сегментов для α -спиральных белков и белков типа β -бочонок с использованием доступных алгоритмов, реализованных в виде интернет-серверов;
6. разработать и апробировать метод оценки качества предсказания алгоритмов на основе критерия самосогласованности в условиях дефицита данных трансмембранных белков с известной мембранной разметкой.

Научная новизна и практическая значимость

Впервые исследована регуляция белком FadR в четырех геномах γ -протеобактерий. Благодаря проведенному анализу выявлены три новых гена, кодирующие ферменты катаболизма генов длинноцепочечных жирных кислот в γ -протеобактериях, и показана регуляция одного нового гена:

- ген *uafH*, кодирующий ацил-CoA-дегидрогеназу, идентифицирован как ген, описанный ранее в литературе как *fadE* без привязки к геному;
- гены, входящие в состав оперона *b2342-b2341*, кодирующие β -кетоацил-CoA тилазу и 3-гидроксиацил-CoA дегидрогеназу, соответственно, которые впоследствии получили название *fadIJ*.
- показана регуляция гена *fadH*.

Исследована регуляция белком FabR в шести группах γ -протеобактерий. Благодаря проведенному анализу выявлен один новый регулируемый ген *lcfH*, кодирующий CoA-лигазу длинноцепочечных жирных кислот.

Разработан метод оценки качества предсказания алгоритмов на основе критерия самосогласованности в условиях дефицита трансмембранных белков с известной мембранной разметкой.

Выявлены наиболее надежные алгоритмы, реализованные в виде интернет-серверов, предсказывающие положение трансмембранных сегментов в α -спиральных белках и в белках типа β -бочонок.

Апробация работы

Основные результаты диссертации были представлены на следующих конференциях: The Third International Conference on Bioinformatics of Genome Regulation and Structure (Novosibirsk, July 2002); The First International Moscow

Conference on Computational Molecular Biology (Moscow, July 2003); The Fourth International Conference on Bioinformatics of Genome Regulation and Structure (Новосибирск, июль 2004); Школа молодых ученых "Сравнительная геномика", (Севастополь, Украина, июнь 2005); The Second International Moscow Conference on Computational Molecular Biology (Moscow, July 2005); The Third International Moscow Conference on Computational Molecular Biology (Moscow, July 2007); 30-я конференции молодых ученых и специалистов ИППИ РАН "Информационные технологии и системы" Россия, сентябрь 2007; The Fourth International Moscow Conference on Computational Molecular Biology (Moscow, July 2009).

Структура и объем диссертации

Диссертационная работа состоит из введения, трех глав, выводов и списка цитируемой литературы. Работа изложена на 154 страницах машинописного текста и включает в себя 116 рисунков и 19 таблиц. Список цитируемых литературных источников содержит 215 наименований.

Содержание работы

Глава 1. Обзор литературы. Содержит критический обзор литературы по рассмотренным в диссертации задачам.

Глава 2. Материалы и методы. Содержит описание основных методов и программ, как известных, так и оригинальных, разработанных непосредственно для решения поставленных задач, а также интернет-ресурсов, использованных в работе.

Подход, основанный на сравнении геномов

Первичные данные, полученные с использованием матриц позиционных весов (МПВ), разделяют на потенциально функциональные и перепредсказанные сайты, применяя метод проверки соответствия. Суть данного метода состоит в том, что группа генов, образующих регулон (т.е. группу генов, совместно регулируемых определенным транскрипционным фактором) в одном геноме, будет также образовывать регулон в другом родственном геноме в тех случаях, когда в геноме присутствуют ортологи соответствующего транскрипционного фактора.

Для обнаружения группы ортологичных генов, имеющих потенциальный консервативный сайт связывания в 5'-некодирующей области, первичные данные из ряда родственных геномов сопоставляют друг с другом (рис. 1). Как правило, ген считают принадлежащим к обобщенному регулону, когда потенциальные сайты связывания встречаются в трех и более геномах перед соответствующими генами. Тем не менее, это число не является однозначно определенным и зависит от количества анализируемых геномов и степени их родства. За базовый геном обычно принимается тот, из которого были взяты сайты связывания для построения матрицы позиционных весов.

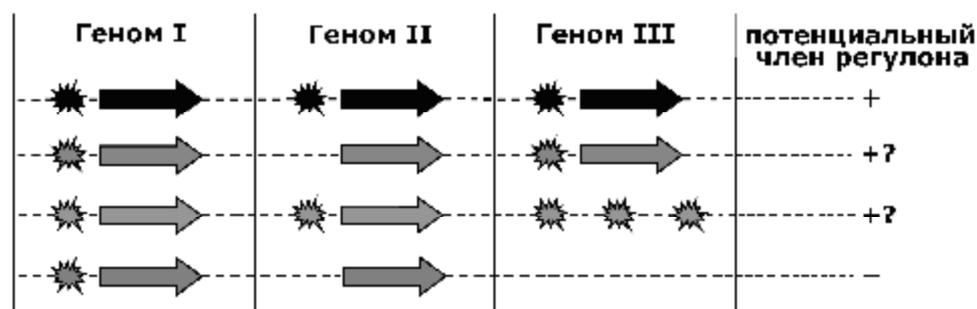


Рисунок 1. Схема подхода, основанного на сравнении геномов. На схеме гены изображены стрелками, пунктирной линией соединены ортологичные гены, потенциальные сайты связывания отмечены звездочками. Знаком “+” обозначены случаи, когда гены рассматривают как потенциальные члены регулона; знаком “-” – когда гены считают перепредсказанными членами регулона; знаками “+?” – случаи, свидетельствующие о неоднозначности ответа, и как следствие, о необходимости дополнительных исследований.

В качестве недостатка метода проверки соответствия можно указать возможность того, что ген утратил регуляцию или отсутствует непосредственно в базовом геноме по сравнению с ортологичными генами из других родственных геномов. Для выявления подобных случаев применяют модифицированный метод проверки соответствия. При этом первичные данные из всех возможных пар геномов, принадлежащих одной таксономической группе, сравнивают друг с другом. В случае выявления нового потенциального члена обобщенного регулона проверяют наличие консервативного сайта связывания в 5'-некодирующих областях его ортологов из других таксономических групп.

Оценка предсказания алгоритмов: коэффициент Жаккарда Q и коэффициент перекрытия сегментов C

В основе исследования лежат следующие предположения:

- в группе близкородственных белков должна сохраняться структура белка и, следовательно, положение трансмембранных сегментов (ТМ-сегментов);

- в паре выравненных близкородственных белков ТМ-сегменты с хорошей точностью будут проецироваться один на другой.

Для оценки предсказания алгоритмов выбрали две величины: коэффициент Жаккарда (Q) и коэффициент перекрытия сегментов (C) (рис. 2).

Сравнение предсказаний по аминокислотно проводили с помощью коэффициента Жаккарда. Для каждой пары выравненных белков его определяют как размер пересечения предсказанных ТМ-сегментов, поделенный на размер их объединения. Точнее, пусть K – число выравненных аминокислотных остатков, предсказанных как входящие в ТМ-сегменты в обоих белках, U – число аминокислотных остатков, входящих в ТМ-сегмент хотя бы в одном из белков, тогда:

$$Q = K / U.$$

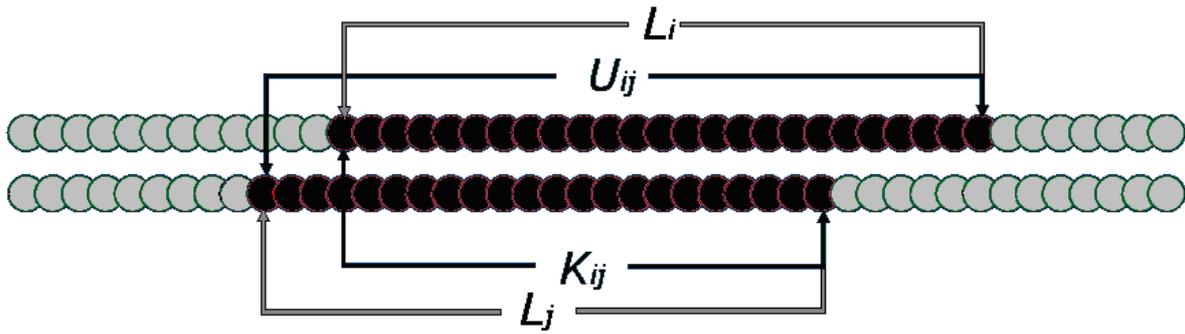


Рисунок 2. Подсчет величин Q и C . Черным цветом обозначены ТМ-сегменты, серым цветом обозначены петли двух выравненных ортологичных белков. L_i – длина сегмента i , L_j – длина сегмента j , K_{ij} – размер перекрытия сегментов i и j , U_{ij} – размер объединения сегментов i и j .

Коэффициент перекрытия сегментов C определяли как долю ТМ-сегментов, присутствующих в паре сравниваемых белков. Обозначим через n_1 и n_2 число ТМ-сегментов, предсказанных для первого и второго белка соответственно. Пусть $i = 1, \dots, n_1$, $j = 1, \dots, n_2$ – номер ТМ-сегмента в первом и втором белке, соответственно. Рассмотрим все пары ТМ-сегментов ij , проекции которых перекрываются хотя бы по одному аминокислотному остатку. Введем величину V_{ij} , показывающую частичное перекрытие сегмента i по отношению к сегменту j . Будем принимать $V_{ij} = 1$, если хотя бы половина сегмента i перекрывается с сегментом j , в противном случае будем принимать $V_{ij} = 0$. Иначе говоря, пусть L_i – длина сегмента i , L_j – длина сегмента j , и пусть K_{ij} – размер перекрытия сегментов i и j . Тогда:

$$V_{ij} = 1, \text{ если } K_{ij}/L_i \geq 0,5,$$

$$V_{ij} = 0, \text{ если } K_{ij}/L_i < 0,5.$$

Аналогично определяется частичное перекрытие сегмента j по отношению к сегменту i . Поскольку, вообще говоря, $L_i \neq L_j$, то $V_{ij} \neq V_{ji}$.

После этого для пары белков коэффициент перекрытия сегментов C вычисляется как сумма локальных перекрытий для всех пар ТМ-сегментов, поделенная на общее предсказанное число ТМ-сегментов:

$$C = \sum_{ij}(V_{ij} + V_{ji}) / (n_1 + n_2).$$

Если предсказания для двух родственных белков схожи, то величины Q и C должны быть близки к 1.

Глава 3. Результаты и обсуждение

Построение матрицы позиционных весов для регулона FadR

Первоначально, опираясь на литературные данные, были выбраны гены участники метаболизма жирных кислот (ЖК): *fabA*, *fabB*, *fadB*, *fadD*, *fadL* – а также репрессор глиоксилатного пути *iclR*, экспрессия которых контролируется регулятором FadR. Для построения МПВ выбрали 5'-некодирующие области этих генов из *E. coli*. Матрица графически представлена в виде диаграммы ЛОГО на рис. 3.

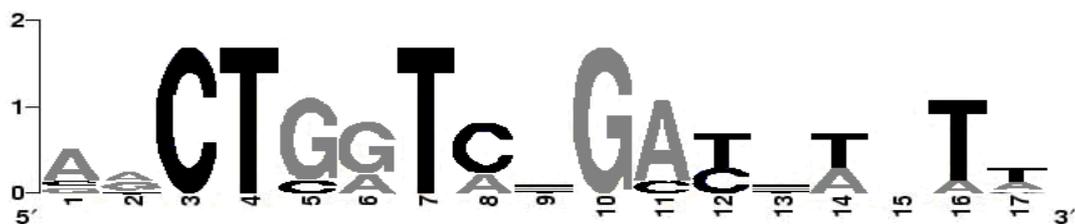


Рисунок 3. ЛОГО для поиска сигнала FadR

Сайт связывания сигнала FadR представляет собой семнадцатibuквенную нестрогую палиндромную последовательность со спейсером в 1 нуклеотид.

Построение матрицы позиционных весов для регулона FabR

Основой построения МПВ послужили 10 сайтов связывания FabR перед генами *fabA*, *fabB* и *yqfA* из семи бактериальных геномов (*E. coli*, *H. influenzae*, *A. actinomycetemcomitans*, *P. aeruginosa*, *V. cholerae*, *S. putrefaciens*, *Y. pestis*), описанных в литературе. Матрица графически представлена в виде диаграммы ЛОГО на рис. 4.



Рисунок 4. ЛОГО для поиска сигнала FabR.

Сайт связывания сигнала FabR представляет собой восемнадцатibuквенную нестрогую палиндромную последовательность со спейсером в 2 нуклеотида.

Далее был проведен анализ 35 полноразмерных геномов из следующих групп γ -протеобактерий, в которых сохраняются ортологи FabR: Aeromonadales, Enterobacteriales, Pasteurellales, Pseudomonadales, Vibrionales, Xanthomonadales. После этого с помощью полученной МПВ провели поиск генов, ортологичных генам *fabA*, *fabB*, *yqfA*, и потенциальных сайтов связывания, расположенных в 5'-некодирующих областях этих генов в геномах каждой из вышеперечисленных групп γ -протеобактерий. На основе полученных сайтов связывания построили МПВ отдельно для каждой группы γ -протеобактерий.

В связи с тем, что для группы Xanthomonadales не было выявлено потенциальных сайтов связывания, расположенных в 5'-некодирующих областях генов *fabA*, *fabB*, *yqfA*, для ее исследования использовали МПВ, полученную ранее на основании 10 сайтов связывания FabR описанных в литературе. Поскольку группа Aeromonadales была представлена только двумя геномами, для построения МПВ к полученным предсказанным сайтам связывания добавили 10 сайтов связывания, полученных на основе литературных данных. Дубликаты потенциальных сайтов связывания удаляли с целью улучшения МПВ.

Анализ регулонов метаболизма жирных кислот

Анализ регулона FadR

Для исследования были выбраны следующие геномы γ -протеобактерий, полная или предварительная нуклеотидная последовательность которых была известна на момент проведения исследования: *E. coli* (Eco), *H. influenzae* (Hin), *V. cholerae* (Vch), *Y. pestis* (Ype).

Прежде всего, убедились в том, что в этих геномах сохраняются ортологи FadR. Затем, используя процедуру попарного сравнения бактериальных геномов, провели сравнительный анализ регулона FadR. В качестве базового генома использовали геном *E. coli* как наиболее хорошо изученный.

В силу того, что на момент проведения исследования число секвенированных геномов было ограничено, ген рассматривали как потенциальный член регулона FadR, если сильный сайт сохранялся по крайней мере перед двумя ортологичными генами из четырех рассмотренных геномов.

Помимо уже известных генов, использованных при построении МПВ, выявлены новые потенциальные члены регулона FadR (таблица 1).

Таблица 1. Новые потенциальные члены регулона FadR

	Геном	Ген	Позиция	Вес	Сайт
<i>fadIJ</i>	Eco	<i>fadIJ</i> (b2342-41)	-42	4.46	AtCaGGTCaGACCAcTT
	Hin	0	–	–	–
	Vch	<i>fadIJ</i>	–	–	–
	Ype	<i>fadIJ</i>	-67	4.72	AtCaGGTCaGACcTgTT
<i>fadE</i>	Eco	<i>fadE</i> (<i>yafH</i>)	-37	4.09	AAgTGGTCaGACcTccT
	Hin	0	–	–	–
	Vch	<i>fadE</i>	-83	3.95	AACTGGTtaGACCAcTa
	Ype	<i>fadE</i>	-37	4.35	AACaGGTCaGACcTccT
<i>fadH</i>	Eco	<i>fadH</i>	-47	4.53	AACTcaTCCGACCAcaT
	Hin	0	–	–	–
	Vch	VC1993	-41	4.08	ttCTGGTCaGACCAtaT
	Ype	<i>fadH</i>	-83	4.53	AtCTcaTCCGACCAcTT

Примечание: сайты с порогом выше 3.90 рассматривали как значимые

Так, проведенный анализ оперона *b2342-b2341* выявил, что его гены являются паралогами генов, образующих оперон *fadBA*, который кодирует мультиферментный комплекс FadAB, обладающий пятью различными ферментативными активностями. Нами было показано, что оба эти оперона содержат сильный сайт связывания в регуляторной области. Впоследствии эти данные были подтверждены экспериментально, и гены, составляющие оперон *b2342-b2341*, получили название *fadIJ*. Следует отметить, что мультиферментный комплекс FadAB принимает участие в катаболизме длинноцепочечных ЖК как в аэробных, так и в анаэробных условиях, а мультиферментный комплекс FadIJ – в основном в анаэробных условиях.

Ген *yafH*, кодирующий ацил-СоА-дегидрогеназу, также имеет сильный сайт в 5'-некодирующей области в геноме *E. coli*. Кроме того, в геномах *Y. pestis*

и *V. cholerae* выявлены ортологи *yafH*, сохраняющие в 5'-некодирующей области сильный и слабый потенциальные сайты, соответственно. Известно, что в β -окислении длинноцепочечных ЖК ген *fadE* кодирует ацил-СоА-дегидрогеназу. Следовательно, ген *yafH* описан ранее в литературе как *fadE*. Эти данные также впоследствии были подтверждены экспериментально.

Ген *fadH* кодирует 2,4-диеноил-редуктазу, катализирующую деградацию ненасыщенных ЖК, у которых двойная связь приходится на четное число атомов углерода. Ортологи этого гена обнаружены во всех геномах рассматриваемой группы. Кроме того, в геномах *E. coli* и *Y. pestis* выявлен сильный потенциальный сайт связывания FadR в 5'-некодирующей области исследуемого гена, что позволяет сделать вывод о регуляции *fadH* этим репрессором. Наличие потенциального сайта связывания хорошо согласуется с функцией FadR как репрессора деградации длинноцепочечных ЖК. Полученные данные о регуляции *fadH* репрессором FadR впоследствии были также подтверждены экспериментально.

Проведенное исследование позволяет предположить, что FadR регулирует все стадии окисления длинноцепочечных ЖК и частично биосинтез ЖК.

Анализ регулона FabR

Прежде всего, из следующих групп γ -протеобактерий: Aeromonadales, Enterobacteriales, Pasteurellales, Pseudomonadales, Vibrionales, Xanthomonadales – выбрали полноразмерные геномы, в которых сохраняются ортологи FabR. Далее, используя специфические МПВ, провели сравнительный анализ регулона FabR, применяя метод попарных сравнений. Исследование проводили отдельно для каждой группы γ -протеобактерий.

Ген рассматривали как потенциальный член регулона, если сайт сохранялся по крайней мере перед тремя ортологичными генами внутри исследуемой группы. Исключение составляла группа Aeromonadales: в силу того, что она была представлена только двумя геномами, ген относили к потенциальному члену регулона, если сайт сохранялся перед обоими ортологичными генами внутри этой группы. Кроме того, в связи с тем, что *A. hydrophila*, *A. salmonicida* и *E. coli* представляют собой близкородственные организмы и число геномов в группе Aeromonadales сильно ограничено, то при исследовании данной группы дополнительно проводили сопоставление с результатами из *E. coli*.

В γ -протеобактериях потенциальные сайты связывания FabR перед исследуемыми генами *fabA*, *fabB* и *yqfA* в целом сохраняются, за исключением группы Xanthomonadales. Следует отметить, что гены, ортологичные генам *fabA*, *fabB*, *yqfA* в группе Xanthomonadales, сохраняются, при этом гены *fabA* и *fabB* образуют оперон *fabAB*, потенциальный сайт связывания ни перед геном *yqfA*, ни перед опероном *fabAB* не выявлен, новых потенциальных членов регулона не обнаружено.

Потенциальный сайт связывания FabR, расположенный в 5'-некодирующей области гена *fabA*, в геномах оставшихся пяти групп γ -протеобактерий в целом сохраняется.

Потенциальный сайт связывания FabR, расположенный в 5'-некодирующей области гена *fabB*, в группах Enterobacteriales (таблица 2) и Pasteurellales в основном сохраняется. В группе Pseudomonadales гены *fabA* и *fabB* образуют оперон *fabAB*, перед которым сохраняется потенциальный сайт связывания FabR. В группах Aeromonadales и Vibrionales ген *fabB* сохраняется, но потенциальный сайт связывания перед ним не обнаружен.

Для группы Enterobacteriales показали расположение сайтов связывания FabR и FadR в 5'-некодирующих областях генов *fabA* и *fabB* (таблица 2, рис. 5А, 5Б). Результаты множественного выравнивания свидетельствуют о том, что участки, совпадающие с сайтами связывания FabR и FadR, характеризуются хорошей консервативностью и расположены на расстоянии 1 нуклеотид друг от друга, как и описано в литературе.

Таблица 2. Потенциальные сайты связывания FadR и FabR для генов *fabA* и *fabB* (группа Enterobacteriales)

	Геном	Ген	Поз.	Вес	Сайт FadR	Поз.	Вес	Сайт FabR
<i>fabA</i>	Eco	<i>fabA</i>	-73	5.00	AACTGaTCGGACTtGTT	-55	6.14	AGCGTACAcGTGTtaGCT
	Cko	<i>CKO_02114</i>	-75	5.00	AACTGaTCGGACTtGTT	-57	6.14	AGCGTACAcGTGTtaGCT
	Esp	<i>Ent638_1466</i>	-73	5.00	AACTGaTCGGACTtGTT	-55	6.14	AGCGTACAcGTGTtaGCT
	Eam	<i>fabA</i>	-74	4.35	AgtTGaTCGGACTtGTT	-56	5.87	tGctTACAActGTaaGCT
	Kpn	<i>fabA</i>	-74	5.00	AACTGaTCGGACTtGTT	-56	6.05	AGctTACAcGTGTtaGCT
	Plu	<i>fabA</i>	-	-	-	-56	6.24	AGCGTACAAGTGTACTcT
	Sen	<i>SPAB_02491</i>	-75	5.00	AACTGaTCGGACTtGTT	-57	6.14	AGCGTACAcGTGTtaGCT
	Sty	<i>fabA</i>	-73	5.00	AACTGaTCGGACTtGTT	-55	6.14	AGCGTACAcGTGTtaGCT
	Spr	<i>Spro_1751</i>	-73	4.51	ggCTGaTCGGACTtGTT	-55	6.32	AGCGTACAcGTGTaaGCT
	Sfl	<i>fabA</i>	-73	5.00	AACTGaTCGGACTtGTT	-55	6.14	AGCGTACAcGTGTtaGCT
	Sso	<i>fabA</i>	-73	5.00	AACTGaTCGGACTtGTT	-55	6.14	AGCGTACAcGTGTtaGCT
	Sgl	<i>SG1026</i>	-	-	-	-	-	-
	Yen	<i>fabA</i>	-73	4.30	ggCTGaTCGGACTtGcT	-55	6.28	AGCGTACAGcTGTACGCT
Ype	<i>fabA</i>	-73	4.02	ggCTaaTCGGACTtGcT	-55	6.28	AGCGTACAGcTGTACGCT	
<i>fabB</i>	Eco	<i>fabB</i>	-82	4.51	ggCTGaTCGGACTtGTT	-64	6.32	gGCGTACAAGTGTACGCT
	Cko	<i>CKO_00462</i>	-84	4.51	ggCTGaTCGGACTtGTT	-66	6.32	gGCGTACAAGTGTACGCT
	Esp	<i>Ent638_2871</i>	-81	4.51	ggCTGaTCGGACTtGTT	-63	6.49	AGCGTACAAGTGTACGCT
	Eam	<i>fabB</i>	-81	4.79	AgCTGaTCGGACTtGTT	-63	6.14	gGCGTACAActTGTAaGCT
	Kpn	<i>fabB</i>	-82	4.44	gtCTGaTCGGACTtGTT	-64	6.32	gGCGTACAAGTGTACGCT
	Plu	<i>fabB</i>	-95	4.09	tcCTGaTCGGACTtGTT	-77	6.49	AGCGTACAAGTGTACGCT
	Sen	<i>SPAB_00592</i>	-85	4.51	ggCTGaTCGGACTtGTT	-67	6.32	gGCGTACAAGTGTACGCT
	Sty	<i>fabB</i>	-83	4.51	ggCTGaTCGGACTtGTT	-65	6.32	gGCGTACAAGTGTACGCT
	Spr	<i>Spro_3368</i>	-83	4.40	cgCTGaTCGGACTtGTT	-65	6.11	gGCGTACAtcTGTACGCT
	Sfl	<i>fabB</i>	-82	4.51	ggCTGaTCGGACTtGTT	-64	6.32	gGCGTACAAGTGTACGCT
	Sso	<i>fabB</i>	-82	4.51	ggCTGaTCGGACTtGTT	-64	6.32	gGCGTACAAGTGTACGCT
	Sgl	<i>SG1623</i>	-79	4.33	ctCTGaTCGGACTtGTT	-61	6.10	AGctTACAtcTGTAAgCT
	Yen	<i>fabB</i>	-86	4.40	cgCTGaTCGGACTtGTT	-68	6.23	gGCGTACAActTGTAaGCT
Ype	<i>fabB</i>	-86	4.40	cgCTGaTCGGACTtGTT	-68	6.23	gGCGTACAActTGTAaGCT	

Примечание: сайты FadR с порогом выше 3.90 и сайты FabR с порогом выше 4.90 считали значимыми

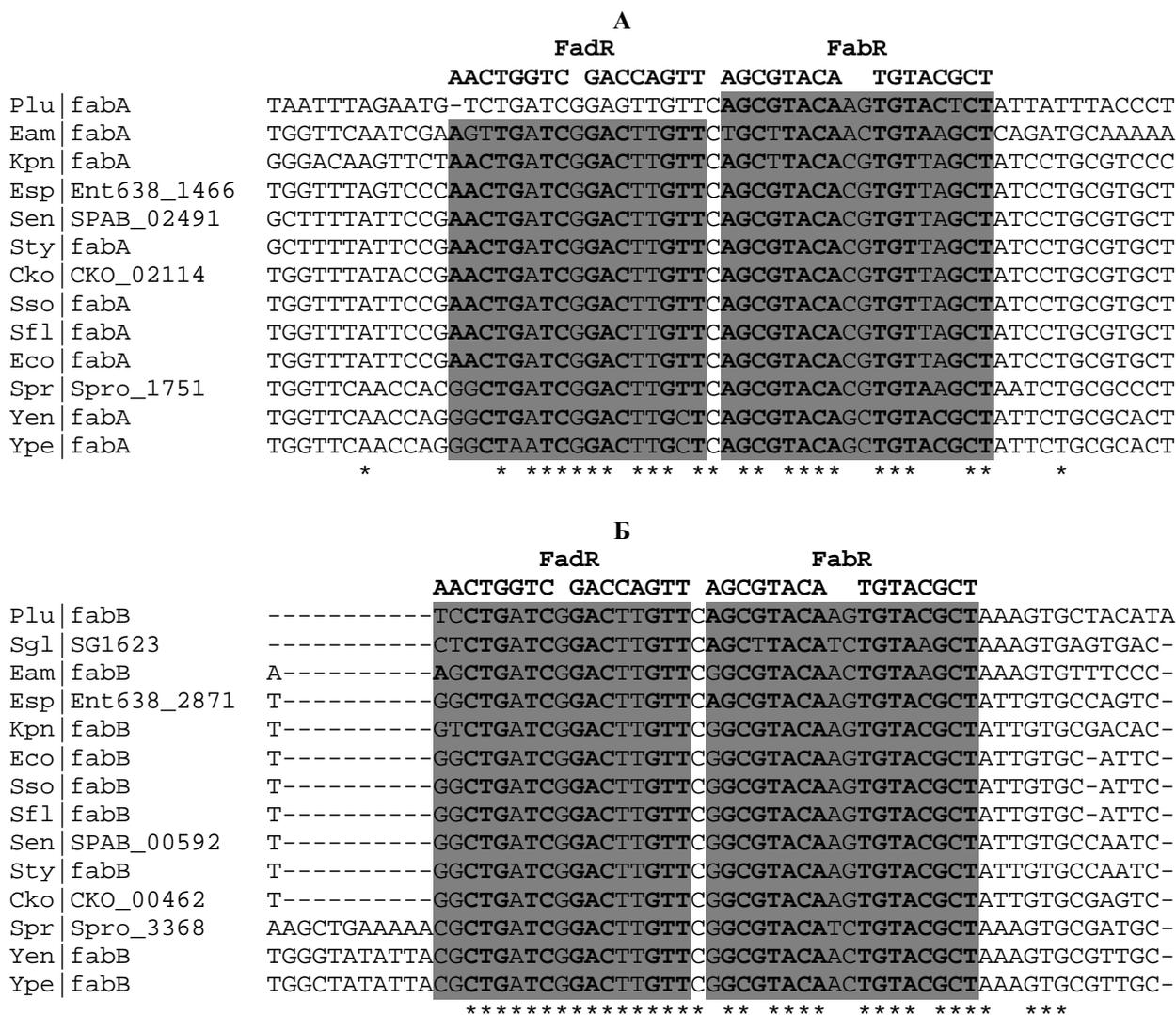


Рисунок 5. Множественное выравнивание 5'-некодирующих областей генов *fabA* (А) и *fabB* (Б) для группы Enterobacteriales. Показано сохранение сайтов связывания FabR и FadR. Сайты выделены серым цветом, полужирным шрифтом обозначены позиции, совпадающие с консенсусом.

Потенциальный сайт связывания FabR, расположенный в 5'-некодирующей области гена *yqfA*, в геномах группы Aeromonadales, Enterobacteriales, Pseudomonadales и Vibrionales в целом сохраняется, в группе Pasteurellales ген *yqfA* сохраняется, но потенциальный сайт связывания перед ним не выявлен. Следует отметить, что дополнительное сопоставление результатов, полученных для двух исследуемых геномов группы Aeromonadales, с результатами из *E. coli*, показало хорошую консервативность сайта связывания FabR, расположенного в 5'-некодирующих областях генов *fabA* и *yqfA*.

Кроме того, в четырех группах γ -протеобактерий (Aeromonadales, Enterobacteriales, Pasteurellales, Vibrionales) выявлен новый потенциальный член регулона FabR ген *lcfH* (long-chain-fatty-acid-CoA ligase), кодирующий CoA-лигазу длинноцепочечных ЖК, которая принимает участие в биосинтезе ЖК (таблица 3). При этом следует подчеркнуть, что ген *lcfH* отсутствует в геноме *E. coli*.

Таблица 3. Новый потенциальный член регулона FabR, ген *lcfH*

Геном	Ген <i>lcfH</i>	Позиция	Вес	Сайт
Aeromonadales				
<i>A. hydrophila</i>	<i>AHA_0722</i>	-97	5.64	gGCGAACActTGTaaGCT
<i>A. salmonicida</i>	<i>ASA_0719</i>	-96	5.83	AGCGAACActTGTaaGCT
Enterobacteriales				
<i>P. luminescens</i>	<i>plu3671</i>	-113	5.84	AGCtaACAccTGTAaGCT
<i>S. proteamaculans</i>	<i>Spro_0747</i>	-83	4.92	AGtaaACAcGTGTAaGCT
<i>Y. enterocolitica</i>	<i>YE0656</i>	-	-	-
<i>Y. pestis</i>	<i>YPO0537</i>	-85	5.43	AGtGaACAccTGTAaGCT
Pasteurellales				
<i>A. pleuropneumoniae</i>	<i>APL_1413</i>	-40	6.07	AGCGAACAgTGTaAGCT
<i>H. influenzae</i>	<i>HI0002</i>	-21	5.37	AGCGcACAAcTGTTcGCT
<i>M. succiniciproducens</i>	<i>MS2265</i>	-43	5.71	AGCGtACAAaTGTaAaCT
<i>P. multocida</i>	<i>PM0925</i>	-57	6.58	AGCTAACAAAGTGTaAGCT
Vibrionales				
<i>P. profundum</i>	<i>PBPRA0424</i>	-36	6.15	AGCTTACAaGTGTAaGCT
<i>V. cholerae</i>	<i>VC2484</i>	-30	5.52	gGCTTACAgGTGTAaGCT
<i>V. fischeri</i>	<i>VF2264</i>	-32	5.68	AGCTTACACcTGTAaGCc
<i>V. harveyi</i>	<i>VIBHAR_00824</i>	-36	5.11	AGCTTACgCcTGTAaGCc
<i>V. parahaemolyticus</i>	<i>VP0351</i>	-59	5.27	tGaGTACAcTGTtCGCT
		-36	5.11	AGCTTACgCcTGTAaGCc
<i>V. vulnificus</i>	<i>VV1_0649</i>	-30	5.28	AGCTTACAtGTGTAaGCc

Примечание: в таблице приведены данные только для геномов, в которых выявлены ортологи гена *lcfH*. Сайты с порогом выше 4.90 считали значимыми.

Ортологи этого гена обнаружены в геномах *A. hydrophila*, *A. salmonicida* (Aeromonadales), *P. luminescens*, *S. proteamaculans*, *Y. enterocolitica*, *Y. pestis* (Enterobacteriales), *P. profundum*, *V. cholerae*, *V. fischeri*, *V. harveyi*, *V. parahaemolyticus*, *V. vulnificus* (Vibrionales). Потенциальный сайт связывания FabR, расположенный в 5'-некодирующей области гена *lcfH*, сохраняется во всех перечисленных геномах, за исключением *Y. enterocolitica*, а в геноме *V. parahaemolyticus* выявлен дополнительный сайт связывания FabR. Следует отметить, что в группе Vibrionales гены, ортологичные гену *lcfH*, обнаружены во всех исследуемых геномах.

Обсуждение результатов анализа регулонов метаболизма жирных кислот

Обсуждение результатов анализа регулона FadR

В ходе проведенного исследования было установлено, что в состав регулона FadR входят гены, описанные ранее в литературе: *fabA*, *fabB*, *fadD*, *fadL*, *iclR* – и гены, образующие оперон *fadBA*. Кроме того, выявлены новые члены регулона FadR: ген *fadE* (*yafH*) и гены, входящие в состав опрона *fadIJ*, а также впервые было показано наличие потенциального сайта связывания FadR в 5'-некодирующей области гена *fadH*.

Впоследствии в работе А.Е. Казакова и соавторов было показано, что регулон FadR сохраняется в следующих четырех группах γ -протеобактерий:

Enterobacteriales, Pasteurellales, Vibrionales и Alteromonadales. При этом анализ выявил высокий уровень консервативности регулона FadR в группе Enterobacteriales, в то время как его состав в других группах γ-протеобактерий в некоторой степени отличался.

Кроме того, в группе Enterobacteriales дополнительно выявлен ген *fadM* (*ybaW*) и обнаружен потенциальный сайт связывания FadR в 5'-некодирующей области оперона *aceBAK* в геноме *Y. pestis*, а в группе Vibrionales выявлен ген *plsB*. В нашем исследовании эти случаи были пропущены из-за ограниченного числа геномов. Таким образом, увеличение числа отсеквенированных геномов позволило выявить дополнительных участников данного метаболического пути.

Обсуждение результатов анализа регулона FabR

В ходе проведенной работы было установлено, что в состав регулона FabR входят три ключевых гена *fabA*, *fabB*, *yqfA*, известных ранее. Следует отметить, что именно перед этими генами сайты связывания FabR подтверждены экспериментально в геноме *E. coli*. Также выявлен новый потенциальный член регулона, ген *lcfH*, кодирующий CoA-лигазу длинноцепочечных ЖК (таблица 3). Других генов в составе исследуемого регулона не обнаружено.

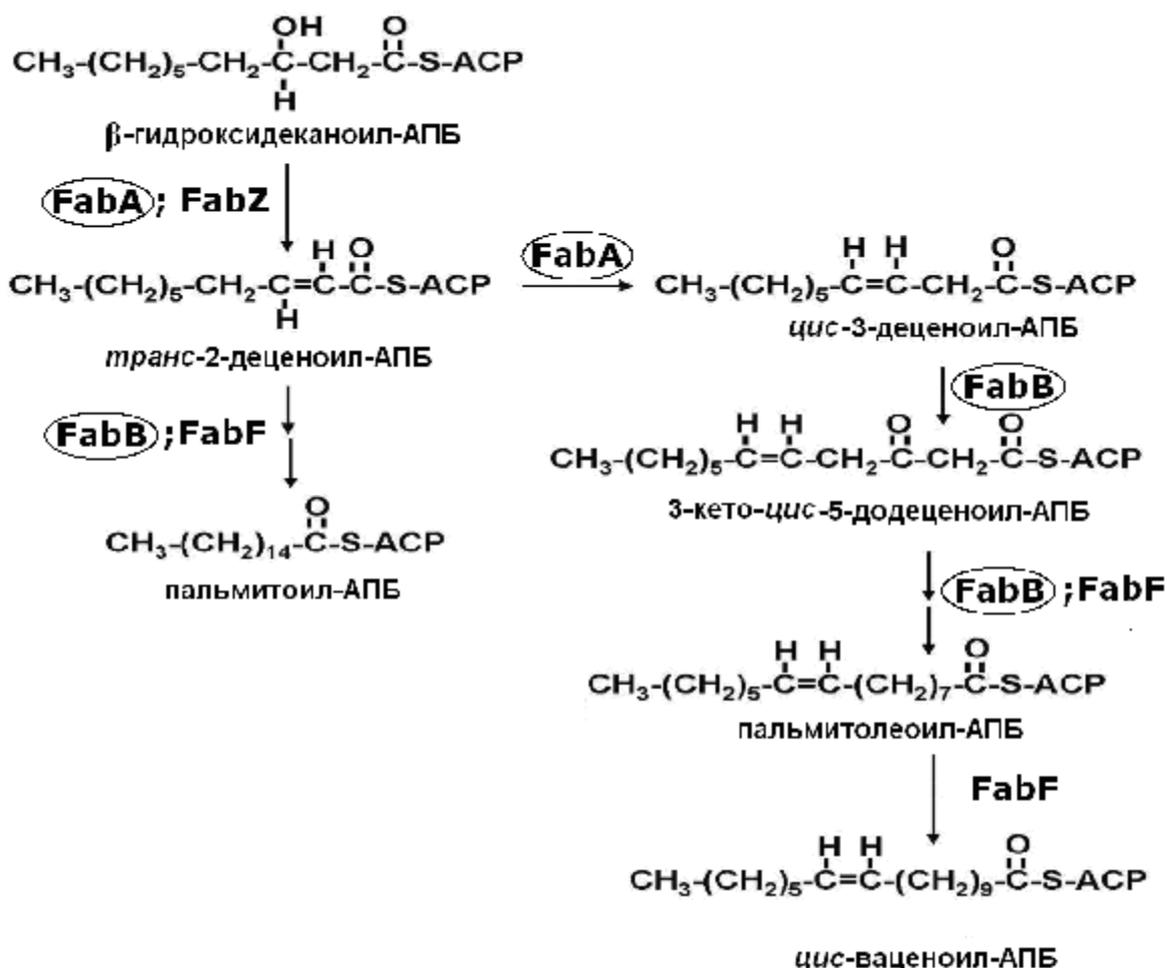


Рисунок 6. Схема ответвления биосинтеза ненасыщенных жирных кислот в *E. coli*. Овалами выделены ферменты, кодируемые генами регулона FabR.

Кроме того, для группы Enterobacteriales показано, что сайты связывания FabR и FadR в 5'-некодирующих областях генов *fabA* и *fabB* расположены вплотную друг к другу. Столь близкое взаимное расположение сайтов связывания FabR и FadR указывает то, что два фактора транскрипции, по всей видимости, не могут одновременно связываться с промотором каждого из этих генов. Соответственно, уровень экспрессии генов *fabA* и *fabB* может контролироваться комбинационным переключением между транскрипционными факторами FabR и FadR.

Следует отметить, что гены *fabA* и *fabB* являются ключевыми генами биосинтеза ненасыщенных ЖК (рис. 6). Таким образом, можно предположить, что регулон с небольшим количеством генов необходим для осуществления баланса между насыщенными и ненасыщенными ЖК, соотношение которых определяет физические свойства мембраны, и зависит от уровня ферментов FabA и FabB в клетке.

Построение тестовой выборки

Тестовая выборки α -спиральных трансмембранных белков

Для построения тестовой выборки α -спиральных ТМ-белков из баз данных TCDB и TransportDB взяли всех представителей бактериальных транспортеров класса TC.2.A по классификации Сайера. Начальная выборка составила 1312 белков из 101 семейства. С целью обогащения начальной выборки родственными, но неклассифицированными белками, для каждой аминокислотной последовательности провели поиск гомологов в бактериальных геномах базы данных ERGO с использованием программы BLAST.

Поиск гомологов проводили только в относительно завершенных геномах при наложении следующих условий: E-value $\leq 10^{-10}$, identity (величина идентичности) $\geq 30\%$. За счет этой процедуры начальная выборка белков расширилась на 860 гомологичных белков и суммарная выборка, полученная на текущем этапе, составила 2172 белка.

Построение кластеров

Первоначально набор полученных аминокислотных последовательностей разделяли на кластеры согласно их эволюционному родству с нижним порогом идентичности 30%. За меру эволюционного расстояния принимали значения идентичности, полученные с использованием программы BLAST. Для этого были проведены все попарные сравнения аминокислотных последовательностей из суммарной выборки (2172 белка). Далее набор последовательностей разделяли на кластеры методом ближайшего соседа (т.е. методом выделения в графе компонент связности) с нижним порогом идентичности 30%. За нижний порог идентичности принимали величину, с которой сравнивают величины идентичности между каждыми двумя последовательностями. Далее были рассмотрены последовательности, процент идентичности которых лежал в двух рассматриваемых диапазонах: 40–49% и 50–59%. Суммарная выборка

составляла 2356 пар белков для первой группы (40–49%) и 909 пар белков для второй группы (50–59%).

Тестовая выборка белков типа β -бочонки

При построения тестовой выборки белков типа β -бочонки из базы данных TCDB для каждого семейства взяли по одному представителю бактериальных транспортеров класса TC.1.B по классификации Сайера. Первоначально выбирали белки только из *E. coli*, белки из других геномов не рассматривали. Для полученных белков провели поиск кластеров ортологичных генов COG из грам-отрицательных бактерий. Белки, длина аминокислотной последовательности которых составляла менее 80% от длины аминокислотной последовательности белка, по которому искали кластер, исключали из выборки. Итоговая выборка составляла 274 белка из 15 COG, которые образовывали 5831 пару гомологичных белков. В связи с тем, что количество белков типа β -бочонки сильно ограничено, паралогичные белки не удаляли и не устанавливали ограничения на порог идентичности. Исследования проводили для трех диапазонов идентичности: 1–50%, 51–100% и 1–100%.

Сравнительный анализ алгоритмов

Как в случае с α -спиральными ТМ-белками, так и в случае с белками типа β -бочонки сравнение проводили для всех пар белков, принадлежащих одному кластеру. Множественное выравнивание строили для аминокислотных последовательностей, принадлежащих каждому кластеру.

Для каждого алгоритма использовали значения, предложенные интернет-сервером по умолчанию. ТМ-сегменты предсказывали независимо для каждой пары белков. Пересекающиеся и соприкасающиеся ТМ-сегменты принимали за один сегмент. В случае, когда по результатам работы алгоритма PRED-TMBV следовало, что данный белок не относится к белкам типа β -бочонок, запрос повторяли с указанием отнести данный белок к белкам типа β -бочонок.

Сравнительный анализ алгоритмов, предсказывающих положение трансмембранных сегментов в α -спиральных белках

На этом этапе работы проводили анализ 10 алгоритмов, реализованных в виде интернет-серверов, предсказывающих положение ТМ-сегментов в α -спиральных белках: DAS, HMMTOP 2.0, PRED-TMR, TMAP, ТМНММ 2.0, ТМpred, TopPred II, PHDhtm, PSORT, SOSUI. Среднеарифметические значения коэффициента Жаккарда Q , коэффициента перекрытия сегментов C и величины стандартного отклонения σ для каждого алгоритма представлены в таблице 4.

Таблица 4. Среднеарифметические значения коэффициента Жаккарда Q , коэффициента перекрытия сегментов C и величины стандартного отклонения σ для различных интервалов идентичности белков ID

ID	40–49%	40–49%	50–59%	50–59%
Алгоритм	$Q \pm \sigma$	$C \pm \sigma$	$Q \pm \sigma$	$C \pm \sigma$
PHDhtm	$0,88 \pm 0,12$	$0,94 \pm 0,11$	$0,89 \pm 0,12$	$0,94 \pm 0,11$

HMMTOP 2.0	0,73 ± 0,12	0,93 ± 0,12	0,76 ± 0,13	0,94 ± 0,13
TMHMM 2.0	0,72 ± 0,13	0,92 ± 0,12	0,76 ± 0,11	0,93 ± 0,12
TMpred	0,70 ± 0,10	0,91 ± 0,08	0,74 ± 0,10	0,93 ± 0,08
TopPred II	0,71 ± 0,12	0,89 ± 0,10	0,76 ± 0,13	0,91 ± 0,09
PRED-TMR	0,69 ± 0,13	0,89 ± 0,12	0,73 ± 0,13	0,90 ± 0,12
SOSUI	0,69 ± 0,11	0,88 ± 0,13	0,72 ± 0,11	0,89 ± 0,13
TMAP	0,64 ± 0,10	0,85 ± 0,11	0,67 ± 0,11	0,87 ± 0,10
DAS	0,64 ± 0,11	0,83 ± 0,10	0,69 ± 0,10	0,87 ± 0,09
PSORT	0,63 ± 0,14	0,84 ± 0,14	0,69 ± 0,14	0,86 ± 0,14

Как видно из таблицы 4, наиболее самосогласованные результаты предсказаний TM-сегментов получены с использованием алгоритма PHDhtm, за которым следуют HMMTOP и TMHMM. Сопоставление результатов нашей работы с результатами, полученными другими группами исследователей, позволило установить следующее.

Arweiler и соавт. получили лучшие результаты для алгоритма TMHMM, предсказания алгоритмов PHDhtm и HMMTOP в этом исследовании показали менее хорошие результаты. В работе Rost и соавт. лучшие результаты получены для PHDhtm и HMMTOP, в то время как алгоритм TMHMM показал средние результаты предсказаний.

Как в нашей работе, так и в работе Arweiler и соавт. алгоритмы TMpred и TopPred II попадают в середину списка.

Наименее согласованные результаты предсказаний в нашем исследовании получены для группы алгоритмов TMAP, DAS и PSORT. При этом DAS продемонстрировал средний уровень предсказания в работах Arweiler и соавт. и Rost и соавт.; TMAP попадает в середину списка в работе Arweiler и соавт. Оставшиеся алгоритмы не были рассмотрены в цитируемых работах.

Sansom и соавт. оценивают предсказания алгоритмов TMHMM, HMMTOP, TMAP и SPLIT как достаточно хорошие, при этом ни одна из рассмотренных программ не была выделена как лучшая.

Таким образом, несмотря на использование различных критериев оценки качества работы алгоритмов, результаты сравнительного анализа, продемонстрированного в данной работе, в целом согласуются с результатами работ, полученными тремя независимыми группами исследователей: Arweiler, Rost и Sansom с соавторами.

Сравнительный анализ алгоритмов, предсказывающих положение трансмембранных сегментов в белках типа β -бочонки

На следующем этапе работы провели анализ восьми алгоритмов, реализованных в виде интернет-серверов: B2TMPRED, B2TMR, HMM-B2TMR, PRED-TMBB (N-best method), PRED-TMBB (posterior decoding method), PRED-TMBB (the Viterbi method), PROFtmb и TMBETA-NET.

В таблице 5 представлены среднеарифметические значения коэффициента Жаккарда Q , коэффициента перекрытия сегментов C и величины стандартного отклонения σ для каждого алгоритма.

Таблица 5. Среднеарифметические значения коэффициента Жаккарда Q , коэффициента перекрытия сегментов C и величины стандартного отклонения σ для различных диапазонов идентичности белков ID

ID	1–50%		51–100%		1–100%	
	$Q \pm \sigma$	$C \pm \sigma$	$Q \pm \sigma$	$C \pm \sigma$	$Q \pm \sigma$	$C \pm \sigma$
Алгоритм						
B2TMR	$0,67 \pm 0,15$	$0,84 \pm 0,11$	$0,85 \pm 0,18$	$0,93 \pm 0,11$	$0,68 \pm 0,15$	$0,84 \pm 0,11$
PROFtmb	$0,64 \pm 0,15$	$0,82 \pm 0,13$	$0,83 \pm 0,19$	$0,92 \pm 0,13$	$0,64 \pm 0,16$	$0,82 \pm 0,13$
HMM-B2TMR	$0,64 \pm 0,16$	$0,82 \pm 0,13$	$0,83 \pm 0,19$	$0,93 \pm 0,12$	$0,64 \pm 0,16$	$0,83 \pm 0,13$
B2TMPRED	$0,49 \pm 0,15$	$0,68 \pm 0,14$	$0,79 \pm 0,23$	$0,88 \pm 0,17$	$0,50 \pm 0,16$	$0,68 \pm 0,14$
HMM-B2TMR	$0,55 \pm 0,26$	$0,71 \pm 0,31$	$0,70 \pm 0,35$	$0,78 \pm 0,35$	$0,55 \pm 0,27$	$0,71 \pm 0,31$
PROFtmb	$0,54 \pm 0,26$	$0,70 \pm 0,31$	$0,70 \pm 0,35$	$0,78 \pm 0,35$	$0,55 \pm 0,27$	$0,71 \pm 0,31$
PRED-TMBB_N	$0,37 \pm 0,18$	$0,56 \pm 0,24$	$0,67 \pm 0,27$	$0,78 \pm 0,25$	$0,38 \pm 0,19$	$0,57 \pm 0,25$
PRED-TMBB_V	$0,37 \pm 0,18$	$0,56 \pm 0,24$	$0,67 \pm 0,27$	$0,78 \pm 0,25$	$0,38 \pm 0,19$	$0,57 \pm 0,25$
PRED-TMBB_P	$0,37 \pm 0,17$	$0,56 \pm 0,23$	$0,66 \pm 0,28$	$0,78 \pm 0,24$	$0,37 \pm 0,18$	$0,57 \pm 0,24$
TMBETA-NET	$0,36 \pm 0,08$	$0,54 \pm 0,10$	$0,66 \pm 0,24$	$0,79 \pm 0,20$	$0,37 \pm 0,10$	$0,54 \pm 0,11$

Примечание: PRED-TMBB_N (N-best method), PRED-TMBB_V (the Viterbi method), PRED-TMBB_P (posterior decoding method). Полужирным шрифтом отмечены результаты для алгоритмов PROFtmb и HMM-B2TMR, если не учитывать белки, отнесенные к “нетрансмембранному” классу.

Наиболее согласованные предсказания получены для алгоритма B2TMR, за ним с небольшим отрывом следует B2TMPRED. Следует отметить, что по результатам работы алгоритмов HMM-B2TMR и PROFtmb рассматриваемые белки довольно часто попадали в “нетрансмембранный” класс. Таким образом, эти алгоритмы попадают в середину списка, несмотря на результаты, сопоставимые с B2TMPRED. Когда из тестовой выборки были удалены белки, отнесенные алгоритмом PROFtmb и алгоритмом HMM-B2TMR к “нетрансмембранному” классу, анализируемая выборка составила 4997 пар белков и HMM-B2TMR – 5018 пар белков, соответственно. При этом среднеарифметические значения коэффициентов Q и C улучшились и стали сопоставимы с результатами, полученными для алгоритма B2TMR, что отражено в таблице 5.

Серия алгоритмов PRED-TMBB и алгоритм TMBETA-NET завершают список. Следует подчеркнуть, что в результате работы алгоритмов группы PRED-TMBB рассматриваемые белки нередко соотносились с “нетрансмембранным” классом, а внутри самой группы алгоритмов PRED-TMBB получились результаты, схожие между собой.

В работе Namodrakas и соавт. лучшие результаты предсказания получены для PRED-TMBB, HMM-B2TMR и PROFtmb, несколько хуже предсказания у B2TMPRED и TMBETA-NET. Следует отметить, что эти авторы не рассматривали алгоритм B2TMR.

Таким образом, результаты оценки работы алгоритмов, полученные нами и независимой группой исследователей в значительной степени схожи, несмотря на различные критерии оценки. Однако следует отметить, что для алгоритма B2TMPRED наши данные не согласуются с данными этой группы исследователей, а для алгоритмов PRED-TMBB и PROFtmb только результаты,

полученные нами после удаления белков, отнесенных алгоритмом к “нетрансмембранному” классу, сопоставимы с результатами, полученными этой группой исследователей. Вероятно, это связано с тем, что в работе Namodrakas и соавт. тестовая выборка состояла всего из 20 белков с известной структурой, в то время как наша выборка включала в себя 5673 пары белков.

Обсуждение результатов сравнительного анализа алгоритмов, предсказывающих положение трансмембранных сегментов в α -спиральных белках и в белках типа β -бочонки

На основании проведенного нами исследования наиболее согласованные результаты предсказания ТМ-сегментов в α -спиральных белках получены с использованием алгоритма PHDhtm. За ним с небольшим отрывом следуют НММТОР и ТМНММ. В середину списка попадают алгоритмы ТМpred и TopPred. Оставшиеся алгоритмы PRED-TMR, SOSUI, TMAP, DAS и PSORT завершают список.

Наиболее согласованные результаты для предсказания ТМ-сегментов в белках типа β -бочонок получены для алгоритма B2TMR, а также для алгоритмов B2TMPRED, НММ-B2TMR, PROFtmb, которые несколько уступают ему. В результате работы последних двух алгоритмов белки типа β -бочонок часто попадали в “нетрансмембранный” класс. В подобных случаях полезно отправить запрос другому алгоритму. Завершает список группа алгоритмов PRED-TMBB, которые также нередко относили белки типа β -бочонок к “нетрансмембранному” классу.

Таким образом, на сегодняшний день предсказание ТМ-сегментов как в α -спиральных белках, так и в белках типа β -бочонок по-прежнему остается сложной задачей. В важных случаях желательно использовать результаты предсказания нескольких алгоритмов, а для получения более качественного предсказания рекомендуется проанализировать не один белок, а все семейство его гомологов.

Выводы

1. Впервые выявлены три новых гена, регулируемые белком FadR:
 - ген *yafH*, кодирующий ацил-СоА-дегидрогеназу, идентифицирован как ген, описанный ранее в литературе как *fadE* без привязки к геному;
 - гены, входящие в состав оперона *b2342-b2341*, кодирующие β -кетоацил-СоА тиолазу и 3-гидроксиацил-СоА дегидрогеназу, соответственно, которые впоследствии получили название *fadI*.
2. Впервые было показано наличие потенциального сайта связывания FadR в 5'-некодирующей области гена *fadH*, кодирующего 2,4-диеноил-СоА-редуктазу.
3. Показано, что белок FabR является регулятором генов *fabA*, *fabB* и *yqfA*, и выявлен новый член регулона ген *lcfH*, кодирующий СоА-лигазу длинноцепочечных жирных кислот.
4. Разработана методика тестирования алгоритмов, предсказывающих положение трансмембранных сегментов в трансмембранных белках, с использованием соображения консервативности вторичной структуры белков, и

на ее основе проведен сравнительный анализ алгоритмов, предсказывающих положение трансмембранных сегментов в трансмембранных белках.

5. Показано, что среди алгоритмов, предсказывающих положение трансмембранных сегментов в α -спиральных белках, наиболее самосогласованными являются алгоритмы PHDhtm, HMMTOP и TMHMM.

6. Показано, что среди алгоритмов, предсказывающих положение трансмембранных сегментов в белках типа β -бочонок, наиболее самосогласованными являются алгоритмы B2TMR, B2TMPRED и HMM-B2TMR.

Список работ, опубликованных по теме диссертации

Публикации в научных журналах

1. Садовская Н.С., Лайкова О.Н., Миронов А.А., Гельфанд М.С. Изучение регуляции метаболизма длинноцепочечных жирных кислот с использованием компьютерного анализа полных бактериальных геномов // Молекулярная биология. – 2001 – Т. 35 – №. 6 – С. 1010-1014.

2. Садовская Н.С., Сутормин Р.А., Рахманинова А.Б., Гельфанд М.С. Сравнительный анализ программ, предсказывающих трансмембранные сегменты в трансмембранных белках // Информационные процессы. – 2002 – Т. 2 – №. 1 – С. 96-99.

3. Sadovskaya N.S., Sutormin R.A., Gelfand M.S. Recognition of transmembrane segments in proteins: review and consistency-based benchmarking of internet servers // J. Bioinform. Comput. Biol. – 2006 – V. 4 – N. 5 – P. 1033-1056.

4. Sadovskaya N.S., Gelfand M.S. Benchmarking of programs that predict the position of transmembrane segments in beta-barrel proteins // Biophysics. – 2008 – V. 53 – N. 2 – P. 134-139.

Публикации в сборниках трудов конференций

1. Sadovskaya N.S., Sutormin R.A., Rakhmaninova A.B., Gelfand M.S. Benchmarking of programs for recognition of transmembrane segments in transporter proteins // Proc. of The Third International Conference on Bioinformatics of Genome Regulation and Structure. – 2002 – V. 3 – P. 116-117.

2. Sadovskaya N.S. Comparative analysis of servers for prediction transmembrane domains // Proc. of the International Moscow Conference on Computational Molecular Biology. – 2003 – P. 206-207.

3. Sadovskaya N.S. Benchmarking of transmembrane helix prediction servers // Proc. of The Fourth International Conference on Bioinformatics of Genome Regulation and Structure. – 2004 – V. 1 – P. 358-360.

4. Садовская Н.С. Анализ программы HMM-B2TMR на примере белка FadL и его ортологов // Школа молодых ученых "Сравнительная геномика". – 2005 – Т. 1 – С. 16-17.

5. Sadovskaya N.S. Analysis of a method HMM-B2TMR using a protein FadL and its orthologs. Comparative genomics of the fatty acids biosynthesis in gamma-

proteobacteria // Proc. of the International Moscow Conference on Computational Molecular Biology. – 2005 – P. 324-327.

6. Садовская Н.С. Сравнительный анализ программ, предсказывающих положение трансмембранных сегментов в белках типа бета-бочонок // Сборник трудов 30-й конференции молодых ученых и специалистов ИППИ РАН “Информационные технологии и системы ИТИС’07”. – 2007 – С. 316-320.

7. Sadovskaya N.S. Benchmarking of internet servers for recognition of transmembrane segments in beta-barrel proteins from gram-negative bacteria // Proc. of the International Moscow Conference on Computational Molecular Biology. – 2007 – P. 268-270.

8. Sadovskaya N.S. Comparative genomics of the fatty acids biosynthesis in gamma-proteobacteria // Proc. of the International Moscow Conference on Computational Molecular Biology. – 2009 – P. 319.

Благодарности

Автор выражает глубокую благодарность своему научному руководителю профессору, доктору биологических наук Михаилу Сергеевичу Гельфанду за чуткое научное руководство, помощь и поддержку в ходе выполнения научной работы; Всеволоду Юрьевичу Макееву за предоставленную возможность выполнить работу в лаборатории “Биоинформатика” Государственного научного центра “ГосНИИгенетика”; Андрею Александровичу Миронову, Александре Борисовне Рахманиновой, Роману Александровичу Сутормину, Алексею Евгеньевичу Казакову, Сергею Владимировичу Ковниру, Александру Владимировичу Фаворову, Дмитрию Александровичу Родионову, Илье Алексеевичу Жарову, Алексею Сергеевичу Шарыкину и Ольге Александровне Шарыкиной за участие, неоценимую помощь в работе, ценные советы и продуктивное обсуждение; всем сотрудникам, аспирантам, стажерам и студентам УНЦ “Биоинформатика” ИППИ РАН за поддержку и дружеское понимание. Автор выражает огромную благодарность своей семье и друзьям за любовь, поддержку, терпение, понимание и доброту при выполнении диссертации.