На правах рукописи

ДАНИЛОВА Людмила Владимировна

КОМПЬЮТЕРНЫЙ ПОИСК РЕГУЛЯТОРНЫХ САЙТОВ БЕЛОК-ДЕЗОКСИРИБОНУКЛЕИНОВОГО ВЗАИМОДЕЙСТВИЯ В ГЕНОМАХ БАКТЕРИЙ И ЕГО ПРИЛОЖЕНИЯ

05.13.17 – Теоретические основы информатики,

03.00.28 – Биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата физико-математических наук

Москва – 2004

Работа выполнена в

Институте проблем передачи информации РАН

Научный руководитель: доктор физико-математических наук,
профессор В.А. ЛЮБЕЦКИЙ.

Официальные оппоненты: доктор физико-математических наук,
профессор В.Г. ТУМАНЯН,
доктор физико-математических наук,
профессор А.В. ЧЕРНАВСКИЙ.

Ведущая организация: Федеральное государственное унитарное предприятие Государственный научный центр Государственный научно-исследовательский институт генетики и селекции промышленных микроорганизмов.

Защита диссертации состоится «__» _____ 2004 г. на заседании диссертационного совета Д.002.077.01 в Институте проблем передачи информации РАН по адресу: 127994, Москва, Б. Каретный, 19.

С диссертацией можно ознакомиться в библиотеке Института проблем передачи информации РАН.

Автореферат разослан «__» _____ 2004 г.

Ученый секретарь диссертационного совета:

доктор тех. наук., профессор С.Н. Степанов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Биоинформатика как самостоятельное научное направление появилась сравнительно недавно, благодаря созданию быстрых методов секвенирования последовательностей ДНК. Открылась возможность сравнительного изучения многих полных геномных последовательностей, прежде всего, у родственных организмов на основе компьютерного анализа, использующего современные алгоритмы. Секвенирование геномов стало рутинным процессом, ежемесячно публикуются по несколько геномов, и стало ясно, что все возрастающая доля геномов может быть исследована только компьютерно, по крайней мере, на стадии предсказания в исходных данных эффектов, требующих дополнительного экспериментального изучения. В последние годы появилось много новых методик, алгоритмов и компьютерных программ для изучения геномов, начиная от определения генов, предсказания их функций, поиска родственных генов в других организмах и вплоть до предсказания механизмов регуляции различных метаболических путей, эволюции геномов и т.д.

Одна из важных задач биоинформатики состоит в распознавании различных регуляторных сигналов, и, в частности, в поиске потенциальных *сайтов связывания транскрипционных факторов*. Эта задача представляется вычислительно и биологически весьма сложной. Поставленная более 15 лет тому назад, она до сих пор далека от эффективного решения. Часто недостаточный объем исходной выборки и низкая степень консервативности сигнала мешают надежному предсказанию сигнала. Но даже и в выборке большего объема не всегда удается найти достоверный сигнал. Поскольку механизм белок-дезоксирибонуклеинового взаимодействия плохо изучен, не всегда можно заранее указать длину искомого сигнала и его структуру, а также исходная выборка часто включает последовательности, не содержащие искомого сигнала, – все это значительно затрудняет исследование.

Цель работы. Создание быстрой и эффективной программы для выделения регуляторных сигналов белок-дезоксирибонуклеинового взаимодействия в геномах и использование ее для поиска новых сигналов связывания транскрипционных факторов в различных таксономических группах организмов и для разных регуляторных систем.

Методика исследования. Создание программного приложения на языке Object Pascal в среде программирование Delphi. Тестирование эффективности алгоритма на различных искусственных и биологических данных и затем его применение к биологическим задачам поиска регуляторных сигналов[1].

[1] Алгоритм реализован также на языке ANSI C для параллельной вычислительной архитектуры, этот результат не включается в диссертационную работу.

Научная новизна. Предложенный алгоритм был реализован в виде программного приложения, разнообразно тестирован и применен для поиска консервативных сигналов в геномах гамма-протеобактерий и грам-положительных бактерий из группы бацилл/клостридии, а также – для исследования регуляции метаболизма глицерол-3-фосфата. При этом обнаружены новые потенциальные сайты связывания белка GlpR, которые имеют различные структуры (палиндромы или повторы) для разных групп организмов.

Основные результаты. В диссертации получены следующие основные результаты:

- Предложен и реализован в виде компьютерной программы алгоритм выделения регуляторных сигналов белок-ДНКового взаимодействия.
- Показана практическая эффективность и актуальность созданной программы на основе ее детального тестирования.
- Проведен поиск потенциальных сигналов белок-ДНКового взаимодействия в регуляторных областях генов гамма-протеобактерий и грам-положительных бактерий.
- Найдены новые потенциальные сайты связывания регулятора GlpR, которые имеют своеобразные структуры (палиндромы или повторы) для разных групп организмов.

Теоретическая и практическая ценность. Полученная программа может применяться для исследования как отдельных геномов организмов, так и их ортологичных рядов с целью поиска новых регуляторных сигналов указанного типа и других функционально-значимых участков. В программе предусмотрено задание различных вариантов функции качества сигнала, что позволяет искать сигналы с наперед заданными структурными особенностями (палиндромность, неравномерный буквенный состав и т.д.).

Апробация работы. Результаты диссертации докладывались на:

- 3-ей международной конференции «Проблемы управления и моделирования в сложных системах», Самара, РАН, 4-9 сентября 2001;
- 3d International Conference on Bioinformatics of Genome Regulation and Structure, BGRS'2002, 14-20 July 2002, Novosibirsk, Russia.
- Moscow Conference on Computational Molecular Biology (MCCMB'03), 22-25 July 2003, Moscow, Russia.
- Научном семинаре по биоинформатике Института проблем передачи информации РАН под руководством профессора, члена-корреспондента РАН Л.М. Чайлахяна.

- Научном семинаре по алгоритмам в геномике Московского государственного университета им. Ломоносова (механико-математический факультет) под руководством профессора В.А. Любецкого.
- Московском семинаре по компьютерной генетике Института молекулярной биологии им. В.А. Энгельгардта РАН.

Публикации. По теме диссертации опубликовано 8 печатных работ.

Структура и объем работы. Диссертация состоит из введения и четырех глав. Библиографический список использованной литературы включает 86 наименований. Объем работы . : страниц машинописного текста, в том числе 14 таблиц и 12 рисунков.

СОДЕРЖАНИЕ РАБОТЫ

Введение

Исходные понятия.

Потребность клетки в некоторых белках значительно изменяется во времени, поэтому имеются механизмы регуляции, обеспечивающие изменение уровня синтеза белков в соответствии с потребностью в них. В частности, специальная группа белков контролирует синтез мРНК на основе белок-ДНКового взаимодействия, регулируя таким образом концентрацию соответствующих ферментов. Такая регуляция может быть как положительной (тогда регулирующий белок называется *активатором*), так и отрицательной (*репрессором*). Аминокислотная последовательность самого *белка-регулятора*, как и любого другого белка, также кодируется в ДНК; определяющий ее ген называется *геном-регулятором*. Регуляторы специфичны, то есть каждый из них влияет на синтез какого-либо одного или нескольких определенных белков. В работе исследуется *случай прокариотов*, хотя предлагаемый нами алгоритм, естественно, не зависит от класса организмов. В простейших (прокариотических организмах) эта специфичность достигается специфичностью связывания молекулы белка-регулятора с определенными некодирующими участками молекул ДНК, расположенными непосредственно перед участком, кодирующим мРНК регулируемого набора ферментов. Специфические участки нуклеотидной последовательности, с которыми связываются регуляторные белки, называются *сайтами*. Много более длинный участок в ДНК, включающий эти сайты и расположенный перед кодирующим участком, называется *лидерной областью* или *апстримом*.

Общепринятое и подтвержденное предположением состоит в том, что все сайты связывания одного белка достаточно сходны между собой. Это предположение позволяет поставить задачу поиска набора сайтов связывания одного белка-регулятора в исходном наборе родственных (относительно него) лидерных областей как задачу нахождения набора

наиболее сходных фрагментов в этой выборке лидерных областей. Сам такой набор называется *сигналом*, а слова, входящие в него, естественно называются *сайтами* (или иногда – *потенциальными сайтами*). Обычно структура и некоторые численные характеристики (например, длина) искомого сигнала заранее фиксируются или подбираются в ходе вычислений. Сигналу приписывается некоторое качество, которое тем выше, чем более похожи попарно друг на друга входящие в него сайты. Возможны разные точные определения *качества сигнала*. Саму задачу нахождения оптимального (наилучшего) по качеству сигнала в данном исходном наборе (выборке) родственных регуляторных областей называют задачей *поиска оптимального сигнала.* Она имеет некоторую связь с задачей множественного локального выравнивания, но, конечно, никак не сводится к ней. Существует подход, использующий метод поиска сигнала для построения выравнивания нескольких последовательностей [15].

Оценки качества сигнала и способы его описания.

Существует несколько способов оценки качества полученного сигнала. Один из них – использование *матрицы позиционных весов,* элементы которой вычисляются по формуле:

$$W(i,\alpha) = A \cdot \left(\ln \frac{C(i,\alpha)}{n} - B_\alpha \right),$$

где $\alpha \in \{A, C, T, G\}$, $C(i,\alpha)$ – количество появлений нуклеотида α в позиции i, n – число последовательностей. Коэффициенты A и B_i подбираются так, чтобы выполнялись условия $\sum_{i=1}^{l} W(i,\alpha) \cdot p_\alpha = 0$ и $\frac{1}{4} \sum_{\alpha \in \{A,C,T,G\}} \sum_{i=1}^{l} W^2(i,\alpha) \cdot p_\alpha = 1$, где p_α - фоновая вероятность нуклеотида α, а l – число позиций в сайте. Фоновые вероятности p_α букв исходного алфавита $\{A, C, T, G\}$ определяются как частоты вхождений букв в полный геном рассматриваемого организма или в исходную выборку регуляторных областей геномов; иногда в этом качестве используются априорные частоты, как-то характеризующие исходный генетический материал. Данная матрица использовалась в диссертационной работе при исследовании метаболизма глицерол-3-фосфта (см. главу 4).

Другой способ состоит в том, что сигнал описывается *матрицей выравнивания*, каждый элемент $n_{\alpha,i}$ которой показывает число появлений каждой буквы α (из того же алфавита) в i-ой позиции сигнала (рис. 1). По ней строится *вероятностная позиционная матрица* сигнала:

$$f(\alpha,i)=\frac{n_{\alpha,i}+c_\alpha}{\sum_{\alpha\in\{A,C,T,G\}}\left(n_{\alpha,i}+c_\alpha\right)}$$

Значения поправок c_α обычно выбираются так, чтобы выполнялось $\sum_{\alpha\in\{A,C,T,G\}} c_\alpha=\sqrt{n}$, где n - число последовательностей в выравнивании ($n=4$ на рис. 1), а сами эти поправки были пропорциональны фоновым вероятностям p_α появления букв в том материале, где ищется регуляторный сигнал. Заметим, что $\sum_{\alpha\in\{A,T,C,G\}} f(\alpha,i)=1$ в любом столбце (позиции) i.

С помощью этой матрицы вычисляется *информационное содержание* сигнала по формуле:

$$I_{seq}=\sum_{i=1}^{l}\sum_{\alpha\in\{A,T,C,G\}} f(\alpha,i)\cdot\ln\frac{f(\alpha,i)}{p_\alpha}.$$

	A	A	T	T	G	A
	A	G	G	T	C	C
	A	G	G	A	T	G
	A	G	G	C	G	T
	1	2	3	4	5	6
A	4	1	0	1	0	1
C	0	0	0	1	1	1
G	0	3	3	0	2	1
T	0	0	1	2	1	1
consensus:	A	G	G	T	G	N

Рисунок 1. Матрица выравнивания для 4 слов длины 6.

Величина информационного содержания иногда используется как характеристика качества найденного сигнала, а вероятностная позиционная матрица – как решающее правило для поиска новых сайтов в исходных и новых регуляторных областях. Для описания и оценки качества сигнала применяются и другие, более сложные методы, например, марковские статистические модели.

Также для оценки сигнала в диссертационной работе используются сумма попарных сходств всех сайтов, входящих в сигнал, и среднее этих сходств.

$F(x,y)$ - функция, отражающая степень сходства для двух слов x и y длины l, в данном случае, количество совпадающих букв в них.

S – сигнал длины l, $s_1,\ldots,s_k (k\le n)$ - сайты, входящие в сигнал S, n - количество последовательностей.

Качество слова s_j $q(s_j)=\sum_{\substack{i=1\\ i\neq j}}^{k} F(s_i,s_j)$.

Среднее качество слова s_j в сигнале: $p(s_j) = \frac{1}{k-1} q(s_j)$. Если в найденном сигнале все слова одинаковы, то эта величина равна *l*.

Качество сигнала S - $Q(S) = \sum_{i=1}^{k} q(s_i)$

Среднее качество сигнала S - $P(S) = \frac{1}{k} \sum_{i=1}^{k} p(s_i)$

Лучшим считается сигнал с наибольшим значением $P(S)$, а все сайты сигнала имеют качество $p(s_j)$.

Основные алгоритмы поиска регуляторных сигналов

Известные подходы и алгоритмы выделения потенциальных регуляторных сигналов в исходном наборе (предполагаемых) регуляторных областей условно делятся на две группы: оптимизационные и комбинаторные. Некоторые алгоритмы сочетают в себе черты обеих групп. Оптимизационные алгоритмы основаны на некоторой характеристике качества сигнала (например, его информационного содержания). Далее производится построение цепочки сигналов, так чтобы их качество (иногда говорят: значение *функционала*) постепенно возрастало. Таким образом, процедура сводится к поиску экстремума некоторого функционала в пространстве всех допустимых сигналов. Таковы алгоритмы максимизации ожидания MEME [1], стохастические и жадные алгоритмы: Gibbs sampler [3] и ряд других (например, имитация теплового отжига и DMS [5]).

Комбинаторные алгоритмы работают также с пространством сигналов, однако в этом случае цель состоит в построении специального слова (*консенсуса*), представленного в каждой или во многих последовательностях из исходной выборки в том смысле, что искомые сайты отклонялись бы от него (и в этом смысле друг от друга) наименьшим образом (т.е. здесь также присутствует некоторая функция качества – какая-то мера компактности полученного набора сайтов, например его диаметр). К их числу относятся CONSENSUS [4], PROJECTION [7], WINNOWER, SP-STAR [11], MITRA [2] и другие (например, ConsInd и MatInd, ITB, WORDUP [10]).

В диссертационной работе предложен и тестирован новый алгоритм для выделения сигнала в исходной выборке невыровненных нуклеотидных последовательностей (в наборе предполагаемых родственных регуляторных областей). Этот алгоритм является промежуточным с точки зрения приведенной выше классификации: в нем происходит оптимизация некоторой функции качества, которая определяется через суммарное попарное сходство сайтов, а не как информационное содержание набора.

Глава 1. Алгоритм поиска выделения регуляторных сигналов белок-ДНКового взаимодействия

Глава содержит подробное описание этапов работы предлагаемого нами алгоритма, начиная с входных данных и заканчивая обработкой и интерпретацией результатов. Затем приводится описание его реализации в виде компьютерной программы, названной здесь IRSA.

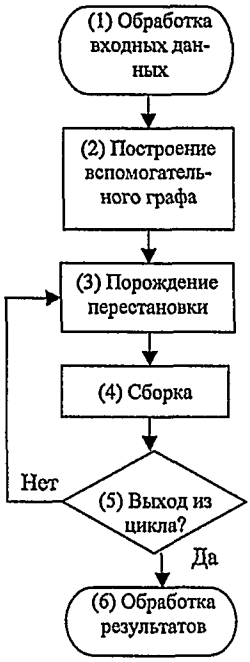

Рисунок 2. Схема программы

. *Постановка задачи.* Дан набор из n нуклеотидных последовательностей (выборка), длины которых m_i, где $i = 1$ до n, если все m_i одинаковы, то $m_i=m$.

Сигнал (иногда говорят: *система*) понимается как набор слов (*сайтов*) фиксированной длины l, по одному слову из одной последовательности; в сигнал включаются слова из какой-то *заранее не фиксированной части исходных последовательностей*; сигнал должен состоять из как можно более попарно похожих друг на друга слов (по возможности из большего числа последовательностей). Похожесть двух слов понимается, например, в смысле расстояния Хэмминга или в смысле какого-то другого фиксированного «расстояния» между словами. Или, наконец, похожесть прямо задается некоторой фиксированной функцией сходства $F(x, y)$, которая для двух слов x и y длины l отражает степень их сходства между собой (например, количество совпадающих букв в них), а также отражает и количественную оценку присутствия в них каких-то других желательных свойств (например, палиндромности). Ищется сигнал (система), который является потенциальным биологическим сигналом, а входящие в него слова – биологическими сайтами из соответствующих регуляторных областей. Наш алгоритм позволяет искать сигнал и в более общем случае, когда из каждой последовательности разрешается выбирать по несколько слов, что, вообще говоря, в большей мере соответствует биологическому пониманию сигнала (так как в последовательности может быть по несколько сайтов от одного сигнала).

Предложенный алгоритм решает исходную задачу за время, квадратичное от числа n исходных последовательностей и кубичное от длины m каждой из них

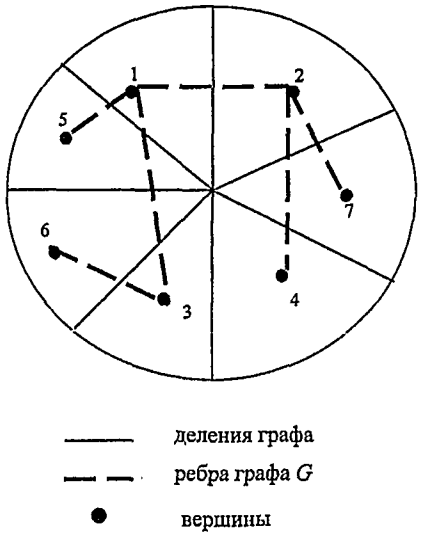

Рисунок 3. Порождение вспомогательного графа *G*.

Общая схема алгоритма представлена на рис. 2. На **этапе 1** входные последовательности разбиваются на подслова длины *l* и для них вычисляются и запоминаются значения функции сходства *F (x, y)*. На **этапе 2** образуется вспомогательный граф *G*, который остается фиксированным в процессе работы алгоритма (он задает определенный порядок при просмотре всех исходных последовательностей). Граф *G* состоит из *n* вершин и всех ребер, возникающих в процессе выполнения следующей процедуры (в примере на рис. 3 *n* = 7). На первом шаге все вершины графа *G* разбиваются на две равные (с точностью до единицы, если *n* нечетное) части и между этими частями произвольным образом проводится ребро *(A, B)* (на рис. 3 это *(A, B) = (1, 2))*. Далее такое разбиение итеративно повторяется «вглубь» графа *G*. А именно, каждую из двух его полученных частей снова разбиваем на две (в том же смысле) равные части. Относительно этих разбиений один конец ребра уже определен: это *A* в одной и *B* в другой частях, а второй конец ребра выбирается произвольно (но без совпадения вершин). На рис. 3 ребра второго уровня деления это (1, 3) и (2, 4). И так далее: каждую появившуюся в этой процедуре не одновершинную часть *P* разбиваем на две равные части P_1 и P_2 так, чтобы ребра новых частей выходили из концов ребра предыдущей части *P*, рис. 4. Конечно, при этом каждое *P* равно объединению его частей P_1 и P_2. Можно остановиться и когда эти части станут просто мелкими (из 1-3 вершин). **Этап 3.** Здесь выполняется цикл, содержание которого состоит в приписывании каждой вершине графа *G* одной из исходных последовательностей без их повторений. Такое приписывание назовем *расстановкой* последовательностей по вершинам графа и обозначим *r* (далее будем писать *A*, понимая под этим последовательность *r(A)*, приписанную вершине *A* в при какой-то фиксированной расстановке *r*).

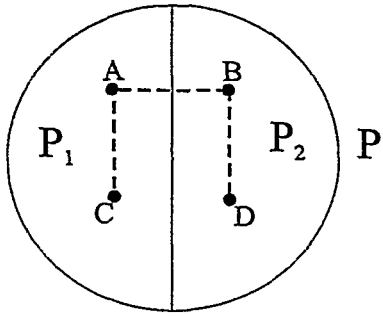

Рисунок 4. Индуктивный шаг сборки.

Каждая следующая расстановка выбирается таким образом, чтобы как можно больше пар последовательностей, не соединенных на предыдущих итерациях этого цикла, теперь соединились. Из вершин ребра, полученного при первом делении, выходит больше всего ребер, поэтому на следующей итерации в эти вершины ставятся последовательности, из которых на текущий момент меньше всего выходило ребер. Этот цикл прекращается, когда любая пара последовательностей хотя бы один раз была соединена в графе G каким-то ребром. Такого типа условие окончания этого (внешнего) цикла обеспечивает разумное количество итераций (далекое от полного перебора) при достаточном разнообразии обработанных пар расстановок (порядка n). На **этапе 4** выполняется цикл внутри цикла из этапа 3: для текущей расстановки r ищется один определенный сигнал (определенная система слов), соответствующий данной расстановке r. Этот цикл называется *сборкой*, а граф G организует сборку; которая является процессом обратным к процессу деления графа G на этапе 2; теперь мы объединяем части графа и по двум уже найденным сигналам, соответствующим двум более мелким частям (выполняя индуктивный шаг) находим сигнал, соответствующий их объединению (еще раз рис. 4). И так по индукции пока не найдем сигнал, соответствующий всем n последовательностям, т.е. всему исходному графу G при данной расстановке r на нем. Полученный сигнал, конечно, зависит от r. Поэтому на следующем **этапе 6** происходит статистическая обработка сигналов, полученных по многим разным расстановкам. А именно, каждому сайту из каждого так найденного сигнала сопоставляется число, которое равно сумме качеств по всем полученным сигналам, которые включают этот сайт. Здесь под качеством понимается качество не всего сигнала, а качество слова из сигнала (который его содержит) по отношению ко всему этому сигналу, т.е. сумма значений $F(x,y)$, где x – упомянутое слово, а y пробегает все остальные слова этого сигнала. Таким образом, сайты, входящие в биологический сигнал, будут помечены в исходных последовательностях числами, которые заметно больше чисел, которые помечают другие сайты.

Еще один вариант состоит в том, чтобы рассматривать каждую систему как отдельный потенциальный сигнал или выбрать одну систему, в которой попарная схожесть слов наибольшая. В программе IRSA реализованы оба варианта и для каждой задачи выбирает-

ся более подходящий из них. Наша практика показывает, что удобнее работать с одной наилучшей системой (см. Введение).

В основном нами использовался вариант реализации алгоритма в виде программы на языке Object Pascal в среде программирования Delphi.

Глава 2. Тестирование программы

В этой главе приведены результаты тестирования нашей программы на различных искусственных и природных выборках. В § 2.1 приведены результаты тестирования на искусственных выборках, которые порождались двумя разными указанными ниже способами. В § 2.2 приведены результаты аналогичного тестирования на природных выборках.

В § 2.1 при *первом способе* исходная выборка сначала содержала искомый сигнал и затем он ослаблялся путем добавления к выборке новых последовательностей, уже не содержащих сигнала («мусорных последовательностей»), а кроме того – и путем «порчи» сайтов самого исходного сигнала. А именно, генерировались выборки из 10 бернуллиевских последовательностей каждая длиной 200 в четырехбуквенном алфавите {A, C, T, G} и в каждую последовательность сначала подставлялось одно и тоже слово длины 16. Затем в каждом из вхождений этого слова случайным образом «портилось» несколько букв (имитация ослабления сигнала), а также добавлялись новые бернуллиевские последовательности, не содержащие сигнала (мусорные последовательности – имитация загрязнения выборки). Такой искусственный сайт считался *найденным*, если полученный в результате работы нашей программы сайт перекрывался с ним не менее, чем на половину его длины. Результаты таковы: сайты длиной 16 устойчиво находились при внесении в исходный сигнал до 3 независимых ошибок (в каждый из его сайтов для выборки из этих 10 последовательностей), а также – когда число мусорных последовательностей не превышало число всех 10 исходных последовательностей (табл. 1). При ошибках в 4 позициях исходного сигнала результат зависел от *чистоты* выборки: приемлемые результаты получались, когда число мусорных последовательностей не превосходило чисел 3-4. А именно, в большинстве испытаний искомые сайты правильно определялись практически во всех исходных последовательностях. При дальнейшем загрязнении выборки некоторые сайты в сигнале могли не обнаружиться, а доля таких результатов, естественно, повышалась с увеличением числа мусорных последовательностей. При ошибках в 5 позициях сайтов исходного сигнала менее пяти из этих сайтов в каждом сигнале обнаруживалось (табл. 1).

Таблица 1. Результаты тестирования выборок из 10 исходных последовательностей. Номер строки указывает на число добавленных мусорных последовательностей (от 0 до 10). В заголовке столбца указано количество измененных букв во вхождениях исходного слова (от 0 до 5). На пересечении строки и столбца приведено число найденных исходных сайтов, где каждый знак соответствует отдельному независимому испытанию и знак X=10. В первом столбце по 1 испытанию для 4-х случаев, во втором – по 10 испытаний для одного случая, в третьем – по 4 испытания для одного случая). В скобках указано среднее число найденных исходных сайтов в % для соответствующей серии испытаний.

	0, 1, 2, 3	4	5
0	X (100%)	XXXXX9XXXX (99%)	2444 (35%)
1	X (100%)	XXXXX9XXXX (99%)	3005 (20%)
2	X (100%)	XXX997XX9X (94%)	4005 (23%)
3	X (100%)	X8X68XXXXX (92%)	3003 (15%)
4	X (100%)	885367X869 (70%)	3204 (23%)
5	X (100%)	3646065539 (47%)	2020 (10%)
6	X (100%)	0673090423 (34%)	0023 (13%)
7	X (100%)	0450054536 (32%)	0222 (15%)
8	X (100%)	2604403027 (28%)	0002 (5%)
9	X (100%)	0660436023 (30%)	0000 (0%)
10	X (100%)	7520273057 (38%)	2000 (5%)

В § 2.1 при *втором способе* ослабление сигнала достигалось за счет увеличения длин исходных последовательностей. Такое тестирование аналогично тому, которое выбрано в [11] для демонстрации качества представленных там алгоритмов. Там оно применялось к 8 исходным выборкам, каждая из которых содержит по 20 последовательностей длины n, а n меняется от 100 до 1000, с заранее имеющимися в них сигналами длиной 15 с 4 бернуллиевскими независимыми заменами в каждом сайте каждого сигнала. В [11] такие сигналы названы (15, 4)-сигналами. В работе [11] ее авторы на этих 8 выборках тестировали ряд типовых программ и ряд их собственных программ для поиска оптимального сигнала с целью их сравнения между собой. Это были программы CONSENSUS, Gibbs sampler, MEME, WINNOWER, SP-STAR. В табл. 2 для всех этих программ приведены результаты из [11] и к ним добавлен результат такого же тестирования и нашей программы. В табл. 2 на пересечении строки и столбца приводится *средний* (по всем сайтам и всем выборкам) коэффициент нахождения сайта, где последний определен в [11] следующим образом. Если для данной последовательности обозначить K множество позиций исходного сайта и обозначить P множество позиций сайта, предсказанного каким-то одним из перечисленных алгоритмов, то *коэффициент нахождения сайта* равен числу общих позиций у K и P, деленному на число позиций в объединении множеств K и P. Это тестирование позволяет сравнить эффективность нашей программы с другими наиболее употреби-

тельными программами. Из табл. 2 видно, что наша программа IRSA на выборках, предложенных в [11] для тестирования всех таких алгоритмов, находится на втором месте после алгоритмов WINNOWER и SP-STAR, предложенных самими авторами работы [11]. Отметим, что выборки, предложенные в [11], специально ориентированы на поиск именно (15, 4)-сигналов. Что касается второго места нашего алгоритма, то заметим, что для алгоритмов, занявших первое место на этих фиксированных выборках, известна только экспоненциальная верхняя оценка числа их шагов, а для нашего алгоритма нами получена полиномиальная верхняя оценка с низкими степенями вида $n^2 \cdot m^3 \cdot l^3$, где n – число последовательностей, m – длина последовательности, l – длина искомого сигнала.

Таблица 2. Результаты тестирования (средний коэффициент нахождения сайта) нашей программой IRSA в сравнении с другими известными программами на данных из [11].

Программы	**Длина последовательностей (m)**									
	100	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
CONSENSUS	0.92	0.94	0.53	0.31	0.29	0.07	0.15	0.09	0.01	0.04
GibbsDNA	0.93	0.96	0.51	0.46	0.29	0.12	0.09	0.34	0.00	0.12
MEME	0.91	0.78	0.59	0.37	0.17	0.10	0.02	0.03	0.00	0.00
WINNOWER (k=2)	0.98	0.98	0.97	0.95	0.97	0.92	0.58	0.02	0.02	0.02
WINNOWER (k=3)	0.98	0.98	0.97	0.94	0.97	0.92	0.90	0.93	0.90	0.88
SP-STAR	0.98	0.98	1	0.96	0.96	0.84	0.83	0.69	0.64	0.23
IRSA	*0.99*	*0.95*	*0.91*	*0.74*	*0.64*	*0.60*	*0.47*	*0.37*	*0.31*	*0.28*

В § 2.2 приводятся результаты тестирования нашей программы IRSA для поиска регуляторных сайтов на природных выборках регуляторных областей, которые постепенно портились. А именно, в качестве 3 исходных выборок были взяты регуляторные области перед генами (бактерии *Escherichia coli)*, которые регулируются соответственно тремя белками-регуляторами PurR (пуриновый регулон), ArgR (аргининовый регулон), CRP (регулон катаболитной репрессии). Для каждой из трех выборок сигнал постепенно портился путем удаления из выборки по одному наилучшему[2] из имеющихся в ней биологических сайтов. Таким образом, могли появляться мусорные последовательности и уменьшалось число сайтов в сигнале. Сайты удалялись таким образом до тех пор, пока их в общей сложности оставалось не менее 3 и пока среднее попарное сходство всех остающихся сайтов строго превышало число 1. Наш алгоритм IRSA искал сигнал с сайтами той же длины, что и у сайтов рассматриваемого биологического сигнала.

[2] Так называется сайт, на котором достигает максимума функция его суммарной похожести на все другие сайты данного сигнала. Если таких сайтов несколько, то выбирается один из них.

Результаты тестирования оценивался с помощью двух функций S_f и S_h. Первая из них определяется как доля *найденных* биологических сайтов (в %) к общему числу таких сайтов, где биологический сайт считается найденным, если алгоритмически полученный сайт пересекается с ним не менее, чем на половину их общей длины. Вторая функция определяется как доля всех найденных сайтов (в %) к числу всех выданных алгоритмом сайтов. Перейдем к описанию результатов.

Пуриновый регулон. Здесь на вход алгоритма подавалась выборка регуляторных областей генов, регулируемых пуриновым репрессором PurR. Она состояла из 19 последовательностей каждая длиной 200 нуклеотидов и содержала в общей сложности 21 сайт длиной по 16 нуклеотидов. Две последовательности содержали по два сайта, остальные – по одному. Результаты таковы (табл. 3): даже если выборка более чем наполовину состояла из мусорных последовательностей, то больше половины остающихся сайтов *опознавалось правильно* в том смысле, что найденный нашей программой сайт и биологический сайт (одинаковой длины) совпадали не менее, чем на половину их длины. Когда в одной последовательности содержится два сайта, то после удаления первого из них второй находился правильно. Первые ошибки появляются при удалении 8 последовательно наилучших из этих сайтов.

Аргининовый регулон. Здесь на вход алгоритма подавалась выборка регуляторных областей генов, регулируемых аргининовым репрессором ArgR. Она состояла из 9 последовательностей каждая длиной 200 нуклеотидов и содержала в общей сложности 19 сайтов длиной по 18 нуклеотидов. Одна последовательность содержала три сайта, остальные – по два. Аргининовый бокс – слабый сигнал, и специфичность регуляции осуществляется здесь за счет кооперативного узнавания мультимерными комплексами молекул репрессора пар сайтов, расположенных на фиксированном расстоянии в 2-4 пары нуклеотидов друг от друга. Результаты таковы (табл. 4), что, тем не менее, сайты связывания аргининового репрессора опознавались правильно даже после удаления 4-х последовательно наилучших сайтов. Первые потери обнаруживаются после удалении 5 сайтов. Как и в пуриновой выборке, при удалении первого уже найденного сайта второй сайт той же последовательности также опознавался правильно. Аналогично в случае трех сайтов в одной последовательности после удаления двух из них третий находился правильно.

Регулон катаболитной репрессии. Здесь на вход алгоритма подавалась выборка регуляторных областей генов, регулируемых белком CRP. Она состояла из 31 последовательности каждая длиной 200 нуклеотидов и содержала в общей сложности 48 сайтов длиной по 22 нуклеотида. В 16 последовательностях содержался один сайт, в остальных – от двух до четырех. Выборка сайтов связывания CRP содержит много слабых сайтов, мно-

гие из них не были найдены даже в исходной выборке. Результаты таковы (табл. 5): после удаления 6 последовательно наилучших сайтов правильно опознавались сайты в более, чем в половине всех последовательностей. Следует отметить, что взаимодействия CRP с регуляторными участками сложны и включают динамические переключения с одних сайтов на другие. Поэтому нельзя исключить, что некоторые из сайтов, найденных нашим алгоритмом, но не соответствующих известным, и вправду являются сайтами связывания CRP. Некоторые из таких предполагаемых сайтов приведены ниже в табл. 6.

Таблица 3. Результаты тестирования для выборки PurR

Количество сайтов в выборке	Количество последовательностей, не содержащих сайты	Количество найденных сайтов	Значение функции чувствительности S_f (в %)	Значение функции специфичности S_h (в %)
21	0	19	90	100
20	0	19	95	100
19	1	18	95	95
18	2	17	94	89
17	3	16	94	84
16	4	15	94	79
15	5	14	93	74
14	6	13	93	68
13	7	11	85	58
12	8	10	83	53
11	9	6	55	32
10	10	6	60	32
9	11	5	56	26
8	12	5	63	26
7	13	2	29	11
6	14	2	33	11
5	15	2	40	11
4	16	0	0	0
3	16	0	0	0

Таблица 4. Результаты тестирования для выборки ArgR

Количество сайтов в выборке	Количество последовательностей, не содержащих сайты	Количество найденных сайтов	Значение функции чувствительности S_f (в %)	Значение функции специфичности S_h (в %)
19	0	9	47	100
18	0	9	50	100
17	0	9	53	100
16	0	9	56	100
15	0	9	60	100
14	0	8	57	89
13	0	8	62	89
12	1	4	33	44
11	1	2	18	22
10	1	2	20	22

9	1	2	22	22
8	1	2	25	22
7	2	2	29	22
6	3	2	33	22
5	4	2	40	22
4	5	0	0	0
3	6	0	0	0

Таблица 5. Результаты тестирования для выборки CRP.

Количество сайтов в выборке	**Количество последовательностей, не содержащих сайты**	**Количество найденных сайтов**	**Значение функции чувствительности S_f (в %)**	**Значение функции специфичности S_h (в %)**
48	0	27	56	87
47	0	23	49	74
46	0	24	52	77
45	1	25	56	81
44	2	19	43	61
43	2	18	42	58
42	2	17	40	55
41	2	14	34	45
40	3	15	38	48
39	3	13	33	42
38	4	10	26	32
37	5	9	24	29
36	6	9	25	29
35	6	6	17	19
34	7	8	24	26
33	8	6	18	19
32	9	6	19	19
31	9	7	23	23
30	10	6	20	19
29	11	5	17	16
28	11	4	14	13
27	12	5	19	16
26	12	5	19	16
25	12	5	20	16
24	13	3	13	10
23	14	6	26	19
22	14	4	18	13
21	14	3	14	10
20	14	3	15	10
19	15	4	21	13
18	16	3	17	10
17	17	3	18	10
16	17	4	25	13
15	18	4	27	13
14	19	4	29	13
13	20	3	23	10
12	21	2	17	6
11	21	0	9	3

Таблица 6. Некоторые потенциальные сайты связывания белка CRP, найденные нашим алгоритмом, но не соответствующие известным сайтам.

Ген	Сайт
ansB	taaattgtttaacgtcaaattt
crp	ctatgctaaaacagtcaggatg
cya	tatgtagcgcatctttctttac
cytR	acggttacagaattttcatgaa
ompA	aaaagtcttgtataaggtatgt

Глава 3. Применение программы для поиска потенциальных сигналов связывания транскрипционных факторов в ортологичных рядах генов организмов из групп энтеробактерий и бациллы/клостридии

Для анализа регуляции применялся сравнительный подход, который основан на предположении, что родственные организмы имеют сходную регуляцию соответствующих метаболических путей. Таким образом, истинные регуляторные сайты располагаются перед ортологичными генами, а ложные («перепредсказанные») сайты разбросаны случайным образом. Считается, что пара генов (по одному из двух геномов) одинаково регулируется, если:

1. эти гены являются ортологами, т.е. гомологичными генами, дивергенция которых связана не с дупликацией, а с расхождением видов, и которые, скорее всего, выполняют в клетке одну и ту же функцию;
2. они имеют в их регуляторных областях потенциальные сайты рассматриваемого вида.

Ортологичные пары генов искались нами по признаку их наибольшей взаимной похожести в данной паре геномов. Затем пары ортологов объединялись в ряды, содержащие наибольшее возможное число генов. При этом транзитивность не требовалась и малые различия в уровне сходства игнорировались, одному гену могло соответствовать более одного ортолога в другом геноме, рис. 5.

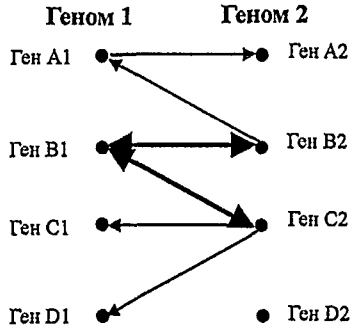

Рисунок 5. Ортологичные гены. Гены B1, B2 и C2 считаются ортологами. Толщина стрелок указывает на степень похожести соответствующих генов.

Регуляторная область определялась длиной в 200 нуклеотидов перед началом гена или как весь межгенный интервал, если он был короче 200 нуклеотидов.

Исходные выборки для поиска сигна-

лов регуляции совпадали с регуляторными областями так полученных рядов ортологичных генов из восьми геномов γ-протеобактерий: *Escherichia coli, Escherichia coli O157, Salmonella typhi, Salmonella typhimurium, Yersinia pestis, Vibrio cholerae, Haemophilus influenzae, Pasteurella multocida* и десяти геномов грам-положительных бактерий (группы бациллы/клостридии): *Bacillus subtilis, Bacillus halodurans, Staphylococcus aureus M (strain MU50), Staphylococcus aureus N (strain N315), Streptococcus pneumoniae, Streptococcus pyogenes, Lactococcus lactis, Listeria monocytogenes, Listeria innocua, Clostridium acetobutylicum.*

После этого нами выполнялась фильтрация так, чтобы в каждой из этих выборок регуляторных областей удалить слишком попарно похожие области (при этом, по возможности, оставив области из *E. coli*). Критерием похожести являлось совпадение 35 нуклеотидов из 40 подряд идущих; а отсев областей шел в порядке их нумерации. Цель фильтрации состояла в том, чтобы поиск консервативных регуляторных сигналов происходил без интерференции со стороны недостаточно дивергировавших областей из близких геномов (штаммов).

В тестировании участвовали выборки, состоявшие из трёх и более областей. Для каждой из двух выше указанных групп организмов (γ-протеобактерий и грам-положительных бактерий) было обработано около 2000 выборок регуляторных областей. Предсказанные для каждой выборки сайты мы сравнивали с известными сайтами из баз данных dpinteract [12] для *E. coli* и DBTBS [6] для *B. subtilis.* Оказалось, что из известных 311 сайтов в *E. coli* и 49 сайтов в *B. subtilis*, которые фактически присутствовали в исходных выборках, наш алгоритм нашел соответственно 99 и 28 сайтов (табл. 7, 8). Можно предположить, что остальные сайты не были найдены из-за слишком слабого сигнала или того, что ортологичные гены потеряли регуляцию. Среди прочего, были обработаны выборки, соответствующие ортологичным рядам, которые ранее, насколько нам известно, не изучались. Поэтому наши результаты могут содержать новые потенциальные сигналы, примеры которых приведены ниже в табл. 9 и были предложены для экспериментальной проверки.

Таблица 7. Результаты поиска регуляторного сигнала ортологичных генов бактерий родственных *E. coli*.

Регулятор	Кол-во известных сайтов	Кол-во известных сайтов, присутствующих в выборках			Найденных сайтов	Доля найденных сайтов от известных, присутствующих в выборках (%)
		в обоих направлениях	на прямой цепи	на обратной цепи		
ArcA	14	9	2	7	9	100
argR	17	20	14	6	3	15
cpxR	12	6	4	2	2	33
crp	49	41	26	15	3	7
cspA	4	6	3	3	1	16
cynR	2	4	2	2	1	25
cytR	5	4	4	0	0	0
deoR	3	1	1	0	0	0
dnaA	8	4	3	1	1	25
fadR	7	9	7	2	1	11
farR	4	8	2	6	4	50
fnr	14	10	6	4	1	10
fruR	12	6	3	3	4	66
fur	9	9	6	3	4	44
galR	7	4	3	1	2	50
gcvA	4	1	1	0	1	100
glpR	13	14	9	5	4	28
hns	15	11	6	5	4	36
hu	3	1	1	0	0	0
iclR	2	1	0	1	1	100
lacI	3	0	0	0	0	0
lexA	19	17	13	4	9	52
malT	10	17	9	8	5	29
melR	2	4	2	2	0	0
metJ	15	20	13	7	11	55
metR	8	10	7	3	1	10
narL	11	7	4	3	1	14
narP	8	10	6	4	0	0
ntrC	5	4	3	1	1	25
ompR	9	7	5	2	2	28
pdhR	2	2	2	0	1	50
purR	22	15	12	3	8	53
rpoN	6	4	3	1	3	75
torR	4	12	8	4	6	50
tyrR	17	13	13	0	5	38
Всего:	**345**	**311**	**203**	**108**	**99**	**31**

Таблица 8. Результаты поиска регуляторного сигнала ортологичных генов бактерий родственных *B. subtilis*.

Регулятор	Кол-во известных сайтов	Кол-во известных сайтов, присутствующих в выборках			Найденных сайтов	Доля найденных сайтов от известных, присутствующих в выборках (%)
		в обоих направлениях	на прямой цепи	на обратной цепи		
araR	5	3	3	0	0	0
treR	2	1	1	0	1	100
ahrC	5	3	2	1	1	33
mta	3	2	1	1	2	100
lacI(CcpA)	33	15	11	4	11	73
gntR	1	1	0	1	0	0
LysR	5	5	4	1	3	60
DeoR	12	2	2	0	1	50
ComA	8	2	2	0	0	0
LuxR/UhpA	22	3	1	2	0	0
Crp(fnr)	2	1	1	0	1	100
LexA(dinR)	8	6	6	0	5	83
merR	8	5	4	1	3	60
Всего:	**114**	**49**	**38**	**11**	**28**	**57**

Таблица 9. Несколько новых потенциальных сигналов, предсказанных нашим алгоритмом в ортологичных рядах.

Организм	Ортологичный ряд	Сайты
γ-протеобактерии		
E. coli	EC aspS	ataaagtggtaacga
Y. pestis	YP aspS	ataaagtgttaataa
P. multocida	VK aspS	ataaagtggcgtaat
V. cholerae	VC VC1166	agcaaggggtaagaa
E. coli	EC asnA	agattgtcgatcagat
Y. pestis	YP asnA	agattatcgatctgat
P. multocida	VK asnA	agattatcaatattgt
H. influenzae	HI HI0564	aaactatcaatgttgt
E. coli	EC yaeG	ttaggcatttgcacaa
S. typhimurium	SY cdaR	ttgtgcatttgcacaa
Y. pestis	YP YPO3978	ctgacctttacctcaa
H. influenzae	HI HI0093	ctgtaatagatctcat
V. cholerae	VC VCA0905	ttgtgcatagtcacaa
E. coli	EC accD	tgttttaatgtgcaacattc
Y. pestis	YP accD	tggtttaatgagtaacattt
P. multocida	VK accD	tggtgtaatacatcgaattt
H. influenzae	HI HI1260	tgttctaatacgcgcaattt
V. cholerae	VC VC1000	tgttttaatccacacgcatt

E. coli	EC_panB	tttatcaggatacgttatgaaa
E. coli O157	ECO_panB	gccatcaggatacgttatgaaa
S. typhimurium	SY_panB	accatcaggaaacgttatgaaa
S. typhi	TY_STY0200	cctataacgaaccgcaacgcaa
Y. pestis	YP_panB	aatttcaggagacagagtgatg
V. cholerae	VC_VC0592	cgagtaaggactaaacatgaaa
грам-положительных бактерии		
B. subtilis	BS_aroA	ctttatcacttaaaa
B. halodurans	HD_aroA	ctttagtacttaaaa
S. aureus N	SAN_SA1558	ttttattgctttaaa
S. pyogenes	ST_SPy1576	cgttatcccatagag
L. monocytogenes	LO_aroA	ctttaatgcttaaaa
C. acetobutylicum	CA_CAC0892	tgttaaggcaacaca
B. subtilis	BS_alaS	tggtaccgcgagacag
B. halodurans	HD_alaS	tggtaccgcgtgactt
S. aureus N	SAN_alaS	tggtaccgcgataacg
S. pneumoniae	PN_SP1383	ctgtgtcgcgattgac
S. pyogenes	ST_alaS	tggtgtgattacatta
L. lactis	LL_alaS	tggtaccgcggtataa
L. monocytogenes	LO_alaS	tggtaccgcgatttca
C. acetobutylicum	CA_CAC0906	tggagaaatgtcagca
C. acetobutylicum	CA_CAC1678	tggtaccgcggaatta
B. subtilis	BS_acpA	tgacggcggaatggtgatgtaa
B. halodurans	HD_acpA	aaatggcggaatggtcatgtaa
S. aureus N	SAN_hmrB	taaagacgcagtaatcaataaa
L. monocytogenes	LO_acpA	tgatggcggaatggtgatgtaa
C. acetobutylicum	CA_CAC1747	tgacgacagcaattatatgtaa
B. subtilis	BS_yrbF	ctttgagcgttacggctataac
B. halodurans	HD_BH1229	ttttgagcaatatggcttcaat
S. aureus N	SAN_SA1464	cttcgagcaatatggattaaat
L. monocytogenes	LO_lmo1529	ttttgagcaatatggattcaat

Глава 4. Применение программы для исследования регуляции метаболизма глицерол-3-фосфата

В этой главе описано применение нашего алгоритм IRSA для анализа GlpR-регулонов, отвечающих за метаболизм глицерола и глицерол-3-фосфата (Г3Ф) в геномах α-, β- и γ-протеобактерий.

Регулятор GlpR, принадлежащий к семейству регуляторов DeoR, контролирует экспрессию генов метаболизма глицерола и Г3Ф. GlpR-регулон хорошо изучен в *Escherichia coli* [16, 8, 17] и частично охарактеризован в *Pseudomonas aeruginosa* [13]. Глицерол поступает извне в цитоплазму путем облегченной диффузии (см. рис. 6), обеспечиваемой продуктом гена *glpF*, а Г3Ф активно транспортируется продуктом гена *glpT*. Внутрикле-

точный глицерол фосфорилируется глицеролкиназой (*glpK*), давая Г3Ф. Г3Ф затем может быть превращен в дигидроксиацетонфосфат под действием одной из двух имеющихся у *E. coli* Г3Ф дегидрогеназ: аэробной (*glpD*) или анаэробной (*glpA*). Кроме того, к GlpR регулону *E. coli* относится ген *glpQ*, кодирующий периплазматическую глицерофосфодиэстеразу, гидролизующую глицерофосфодиэфиры с высвобождением Г3Ф, гены *glpB* и *glpC*, кодирующие дополнительные структурные компоненты анаэробной Г3Ф дегидрогеназы, а также гены *glpE*, *glpG* и *glpX*, функции которых не ясны. Вышеназванные гены собраны в три локуса на хромосоме *E. coli*: *glpTQ/glpABC*, *glpEGR/glpD* и *glpFKX* (/ разделяет опероны, ориентированные в разные стороны).

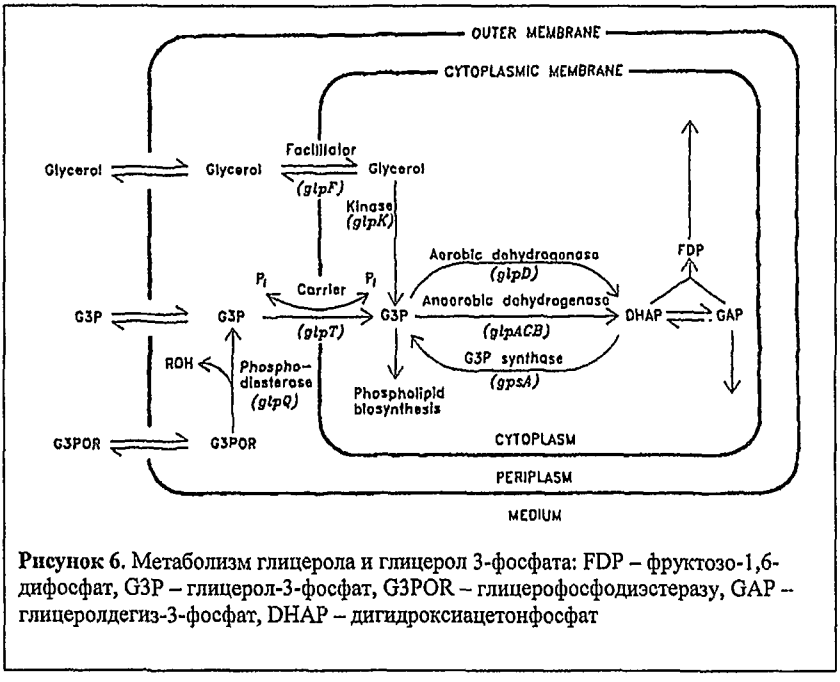

Рисунок 6. Метаболизм глицерола и глицерол 3-фосфата: FDP – фруктозо-1,6-дифосфат, G3P – глицерол-3-фосфат, G3POR – глицерофосфодиэстеразу, GAP – глицеролдегиз-3-фосфат, DHAP – дигидроксиацетонфосфат

Близкие гомологи GlpR были обнаружены во многих геномах α-, β- и γ-протеобактерий. Цель этой главы – поиск сайтов связывания белка GlpR. Для этого нами проведен дополнительный анализ гомологии GlpR-регулируемых генов и определена их оперонная структура в ряде геномов (рис. 7). Были рассмотрены следующие геномы.

γ-протеобактерии: *Escherichia coli, Salmonella typhi, S. typhimurium, Klebsiella pneumoniae, Erwinia carotovora, Yersinia pestis, Y. enterocolitica, Vibrio cholerae, V. vulnificus, V. fischeri, Pasteurella multocida, P. haemolytica, Haemophilus influenzae, H. ducrey, H. somnus, Pseudomonas aeruginosa, P. fluorescens, P. putida, P. syringae,*

H. somnus, Pseudomonas aeruginosa, P. fluorescens, P. putida, P. syringae, Actinobacillus actinomycetemcomitans, Xanthomonas axonopodis, X. campestris;

β-протеобактерии: *Burkholderia fungorum, B. pseudomallei, B. cepacia;*

α-протеобактерии: *Bordetella parapertussis, Ralstonia eutropha, R. solanacearum, Mesorhizobium loti, Sinorhizobium meliloti, Rhizobium leguminosarum, Agrobacterium tumefaciens, Rhodopseudomonas palustris, Brucella melitensis, Rhodobacter sphaeroides.*

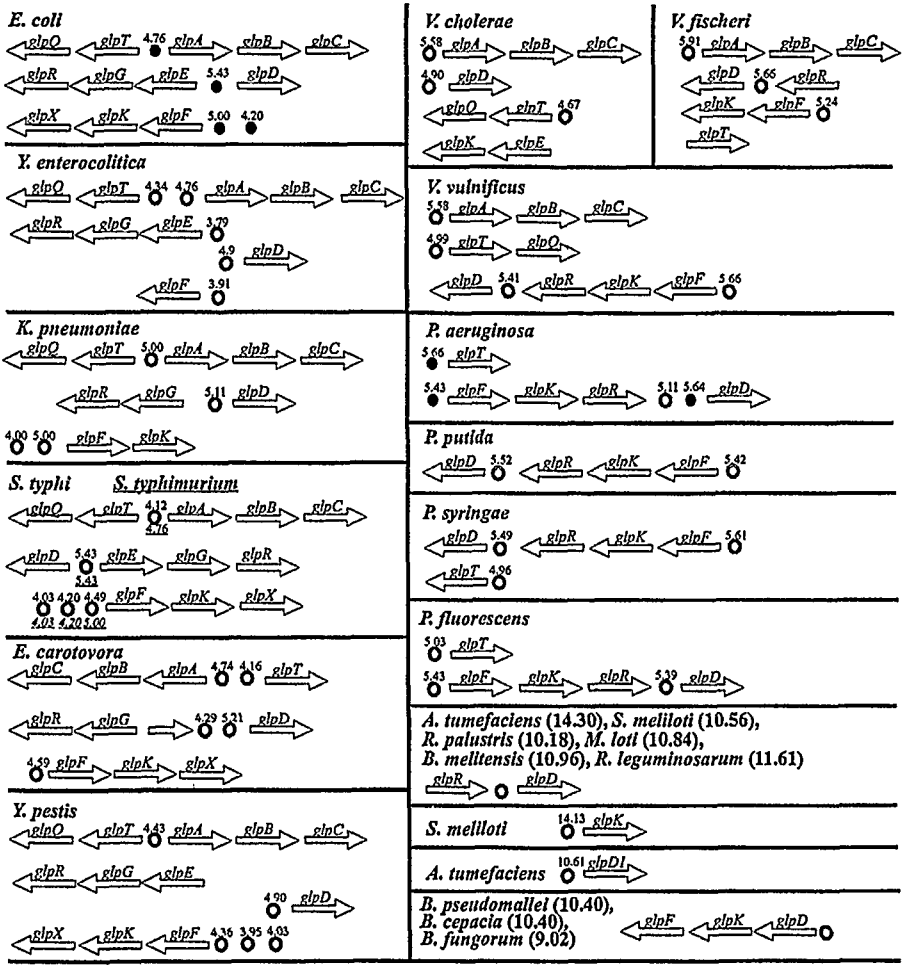

Рисунок 7. Оперонная структура GlpR-регулонов в α-, β- и γ-протеобактериях. Закрашенные кружки отмечают известные сайты, а незакрашенные – предсказанные нами с указанием их весов. Организм *S. typhimurium* имеет одинаковую оперонную структуру с *S. typhi* и подчеркнут для того, чтобы отметить соответствующий ему вес.

Для выравнивания последовательностей белков и построения филогенетического дерева нами использовались соответственно программы ClustalW [14] и Phylip [9]. Сначала наша программа IRSA применялась к части регуляторных последовательностей, которые в табл. 5 соответствуют жирному выделению букв. Эти части отбирались на основе компьютерного поиска частей, которые содержат достаточно сильный сигнал. Таким образом полученный сигнал назовем *базисным* (в табл. 5 он отмечен жирным шрифтом). Уже по нему строилась матрица позиционных весов (таким образом, он служил обучающей выборкой). Для ее построения использовалась программа SignalX [18], а для сканирования геномов – программа GenomeExplorer [18]. Таким образом были получены результаты, которые мы приведем ниже по группам организмов.

γ-Протеобактерии, семейство Enterobacteriaceae.

Сначала рассматривались регуляторные области из четырех геномов *E. coli*, *E. carotovora*, *Y. enterocolitica*, *K. pneumoniae*, и по ним был получен базисный сигнал, включающий уже известные сайты *E. coli* с консенсусом TGTTCGATAACGAACA. По базисному сигналу как по обучающей выборке была построена матрица позиционных весов для поиска палиндромных сайтов длины 16. С помощью этой матрицы были еще найдены сайты в дополнительных геномах *Y. pestis*, *S. typhimurium*, *S. typhi* (табл. 10а).

γ-Протеобактерии, семейство Vibrionaceae.

В регуляторных областях из трех геномов *V. cholerae*, *V. vulnificus*, *V. fischeri* был выделен базисный палиндромный сигнал длины 18 с консенсусом AATGCTCGATCGAGCATT. Базисный сигнал включает сайты в геноме *V. cholerae* перед ортологами генов *glpA, glpD, glpT,* в *V.vulnificus* – перед *glpA, glpD, glpT, glpF*, в *V. fischeri* – перед *glpA, glpD, glpF* (табл. 10б). При сканировании геномов с использованием матрицы позиционных весов, построенной по базисному сигналу, новых потенциальных сайтов не было обнаружено.

γ-Протеобактерии, семейство Pseudomonadaceae.

В регуляторных областях генов *glpD, glpF* из четырех геномов семейства *P. aeruginosa*, *P. fluorescens*, *P. putida*, *P. syringae* был найден палиндромный базисный сигнал wTTTTCGTATACGAAAAw длины 18, включающий сайты, ранее предсказанные в работе [13] у *P. aeruginosa*. По этому базисному сигналу была построена позиционная матрица, с помощью которой были найдены новые потенциальные сайты связывания GlpR перед генами *glpT* в *P. aeruginosa*, *P. syringae* и *P. fluorescens*, а также еще один сайт в регуляторной области гена *glpD* в *P. aeruginosa* (табл. 10в).

α-, β-Протеобактерии.

В регуляторных областях гена *glpD* в геномах α-протеобактерий *M. loti*, *S. meliloti*, *A. tumefaciens*, *B. melitensis*, *R. Palustris* и гена *glpK* в *S. meliloti* и еще одного ортолога гена *glpD* в *A. tumefaciens* были найдены 3-4 тандемных повтора слова TTTCGTT, идущих друг за другом через 3-4 нуклеотида (табл. 10г), которые составили базисный сигнал. При исследовании β-протеобактерий с помощью матрицы позиционных весов, построенной по этому базисному сигналу, аналогичные повторы были обнаружены перед генами *glpD* в геномах бактерий рода *Burkholderia: B. fungorum*, *B. pseudomallei*, *B. cepacia*.

Интересно, что для одного регулятора в одном классе протеобактерий выделился как палиндромный сигнал, так и сигнал на основе тандемного повтора. Поскольку данные о трехмерной структуре регуляторов семейства DeoR отсутствуют, нет оснований полагать, что эти регуляторы во всех случаях образуют димеры, связывающиеся только с палиндромными сайтами: возможны конформации белка, кооперативно связывающиеся с тандемными повторами.

Таблица 10. Сайты перед генами, входящими в ГЗФ регулон в ряде геномов. Вес сайта указывается относительно базисного сигнала (обучающей выборки). Базисный сигнал выделен жирным шрифтом. В скобках указана длина найденных сайтов. Большими буквами в сайтах указаны нуклеотиды, совпадающие с консенсусом.

Геном	Ген	Вес сайта	Сайт
а) семейство Enterobacteriaceae (16)			
E. coli	glpD	5,41	**TGTTCGATAaCGAACA**
E. coli	glpF	4,99	**TGcTCGtTAaCGAtaA**
E. coli	glpT	4,76	**TGTTtGATtTCGcgCA**
E. carotovora	glpD	5,20	**TGcTCGAaAaCGAACA**
E. carotovora	glpT	4,72	**TGTTtGATAaaGAgCA**
E. carotovora	glpF	4,59	**TtcTCGtTtTCGctCA**
K. pneumoniae	glpD	5,10	**TGagCGATATCGAgCA**
K. pneumoniae	glpT	5,00	**TGTTtGATtTCGAgCA**
K. pneumoniae	glpF	4,99	**TGcTCGtTAaCGAtaA**
Y. enterocolitica	glpD	4,89	**TGagCGAaAaCGAACA**
Y. enterocolitica	glpT	4,74	**cGcTCGtTATgGAACA**
E. coli	glpF	4.20	gGcgCGATAaCGctCA
E. carotovora	glpD	4.29	TGTTtGtTtTCGAttA
E. carotovora	glpA	4.16	TGTTCtATtaCGAACg
S. typhi	glpD	5.43	TGTTCGATAaCGAACA
S. typhi	glpF	4.49	TGcTCGtTAgCGAtaA
S. typhi	glpF	4.20	gGcgCGATAaCGctCA
S. typhi	glpT	4.12	TGTTtGATtTCGcgCg
S. typhimurium	glpD	5.43	TGTTCGATAaCGAACA
S. typhimurium	glpF	5.00	TGcTCGtTAaCGAtaA
S. typhimurium	glpT	4.76	TGTTtGATtTCGcgCA

S. typhimurium	glpF	4.20	gGcgCGATAaCGctCA
Y. enterocolitica	glpA	4.34	TGTTCcATAaCGAgCg
Y. pestis	glpD	4.90	TGTTCGtTtTCGctCA
Y. pestis	glpA	4.43	TGTTtctTATCaAtCA
Y. pestis	glpF	4.36	cGcTCGtTAaCGAtaA
б) семейство Vibrionaceae (18)			
V. cholerae	glpA	5.57	AATGCTCGtTCGcGCtTT
V. cholerae	glpD	4.92	AATatTCGAgCGctCATT
V. cholerae	glpT	4.56	AtTGCTCGtTCGccatTT
V. fischeri	glpA	5.91	AATGCgCGAaCGAGCATT
V. fischeri	glpD	5.66	AATGtTCGtTCGctCATT
V. fischeri	glpF	5.24	tgTGCTCGAaCGctCATT
V. vulnificus	glpF	5.69	tATGCTCGAaCGcGCATT
V. vulnificus	glpA	5.66	AATGtTCGAaCGctCATT
V. vulnificus	glpD	5.36	AATGCTCGtTCGAaCAaa
V. vulnificus	glpT	5.02	ttTGCTCGtTCGcaCAcT
в) семейство Pseudomonadaceae (18)			
P. aeruginosa	glpD	5.64	ATTTTCGaATtCGAAcAA
P. aeruginosa	glpF	5.43	TTTTTCGaAactGAAcAA
P. fluorescens	glpF	5.43	TTTTTCGaATctGAAtAA
P. fluorescens	glpD	5.39	ATTTTCGcAaAtGAAcAT
P. putida	glpD	5.52	ATTTTCGcAaACGAAcAT
P. putida	glpF	5.42	TTTTTCGTtTctGAAtAA
P. syringae	glpF	5.61	TTTTTCGTtTACGAAtAT
P. syringae	glpD	5.49	ATTTTCGgAaAtGAAcAT
P. aeruginosa	glpT	5.66	TTTTTCaTtTACGAAAAA
P. aeruginosa	glpD	5.11	ATgTTCGTtTcaGAAAAA
P. fluorescens	glpT	5.03	ATTTTCGgtaACGAAAcT
P. syringae	glpT	4.96	TTTTTCtgtaAtGAAAAT
г) α-, β-протеобактерии (3-4 повтора через 3-5 нуклеотидов)			
A. tumefaciens	glpD	14.30	gTTCGTTtatTTTCtTTgacaTTCGTTttgtTTTCGcT
A. tumefaciens	glpD1	10.61	TTTCGTTtgacaTTCGTTttgtCTTCGAA
B. melitensis	glpD	10.96	TTTCGTTtgatTTTCaTTtgcTTTCGTa
M. loti	glpD	10.84	TTTCGTTtgacaTTCGTTatgagTTCGaa
R. leguminosarum	glpD	11.61	aTTCGTTtgacaTTCGTattccTTTCGTT
R. palustris	glpD	10.18	TTTCGTTttggTTtGTgctttaTTCGTT
S. meliloti	glpK	14.13	TTTCGTTtgacaTTCGTTttcTaTCGTattgaagTCGTT
S. meliloti	glpD	10.56	aTTCGTTtgacaTTCGaaatatTTTCGcT
B. pseudomallei	glpD	10.40	TTTCGaTtatgTTCGTTaaaTTTCGaa
B. cepacia	glpD	10.40	TTTCGaTtccgTTCGTTaaaTTTCGaa
B. fungorum	glpD	9.02	TTTCGaatatgTTCaTTaaagTTCGaa

ВЫВОДЫ.

1. Создано и тестировано эффективное средство – алгоритм и компьютерная программа IRSA для поиска сайтов белок-дезоксирибонуклеиновой регуляции в бактериальных геномах.
2. Показано, что с помощью этой программы можно адекватно искать сигналы транскрипционных факторов белок-ДНКового взаимодействия.
3. На ее основе предсказаны новые сайты связывания репрессора GlpR в γ-протеобактериях (палиндромные сигналы) и в α-, β-протеобактериях (тандемные повторы).
4. На основании предсказанных нами сайтов предположены два типа связывания белка GlpR.
5. На ее основе найдены потенциальные регуляторные сигналы для ортологичных генов в группе энтеробактерий (γ-протеобактерии) и в группе бациллы/клостридии (Грам-положительные бактерии).

РАБОТЫ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

1. *Л.В. Данилова, К.Ю. Горбунов, М.С. Гельфанд, В.А. Любецкий.* Алгоритм выделения регуляторных сигналов в последовательностях ДНК (1) // *Мол. биол.*, 2001, том 35, № 6, с. 987-995.
2. *Л.В. Данилова, К.Ю. Горбунов, М.С. Гельфанд, В.А. Любецкий.* Алгоритм выделения регуляторных сигналов в последовательностях ДНК (2) // *Информационные процессы*, Том 1, № 1, 2001, с. 56-63.
3. *Л.В. Данилова, В.А. Любецкий.* Алгоритм выделения регуляторных сигналов: тестирование и биологические применения. // Труды 3-ей международной конференции «Проблемы управления и моделирования в сложных системах», Самара, РАН, 2001, с. 632-634.
4. *Л.В. Данилова, М.С. Гельфанд.* Поиск регуляторных сайтов в группах ортологичных генов гамма-протеобактерий. // *Информационные процессы*, Том 2, № 1, 2002, с. 59-61.
5. *L.V. Danilova, M.S. Gelfand.* Search for regulatory signals in groups of orthologous genes of gamma – proteobacteria. // *Proc. 3d Int. Conf. on Bioinformatics of Genome Regulation and Structure BGRS'2002*, vol. 2, 2002, p. 21-22.
6. *L.V. Danilova, V.A. Lyubetsky, M.S. Gelfand.* An algorithm for identification of regulatory signals in unaligned DNA sequences, its testing and parallel implementation. // *In*

Silico Biology, V. 3, N 1-2, 2003, p. 33-47. (Электронная версия: http://www.bioinfo.de/isb/2003/03/0004/.)

7. *Л.В. Данилова, М.С. Гельфанд, В.А. Любецкий, О.Н. Лайкова.* Компьютерный анализ регуляции метаболизма глицерол-3-фосфата в геномах протеобактерий // *Мол. биол.*, 2003. Т. 37, № 5, с. 843-849.
8. *L.V. Danilova, V.A. Lyubetsky, O.N. Laikova.* Computer detecting of glycerol-3-phosphate metabolism regulation in proteobacterial genomes // *Proc. Moscow Conference on Computational Molecular Biology (MCCMB'03)*, 2003, p. 52-54

СПИСОК ЛИТЕРАТУРЫ

1. *Bailey T.L., Elkan C.* Unsupervised learning of multiple motifs in biopolymers using expectation maximization // *Machine Learning J.* V. 21, 1995, p. 51-83.
2. *Eskin E., Pevzner P.A.* Finding composite regulatory patterns in DNA sequences. // *Bioinformatics*. 2002; 18, p. 354-363.
3. *Geman, S and Geman, D.* Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images // *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1984, 6, 621-641.
4. *Hertz G.Z., Stormo G.D.* Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. // *Bioinformatics*. 1999. V. 15. P. 563-577.
5. *Hu Y.-J., Sandmeyer S., McLaughlin C., Kibler D.* Combinatorial motif analysis and hypothesis generation on a genomic scale. // *Bioinformatics*. 2000. V. 16. P. 222-232.
6. *Ishii, T., Yoshida, K., Terai, G., Fujita, Y., and Nakai, K.* DBTBS: A database of *Bacillus subtilis* promoters and transcription factors // *Nucleic Acids Res.*, 2001, 29, 278-280.
7. *J. Buhler, M. Tompa* Finding motifs using random projections. // *J. Comp. Biol.*, V. 9, N 2, 2002, p. 225-242.
8. *Larson T.J., Cantwell J.S., van Loo-Bhattacharya A.T.* Interaction at a Distance between Multiple Operators Controls the Adjacent, Divergently Transcribed glpTQ-glpABC Operons of Escherichia coli K-12. // *J. Biol. Chem.* 1992. V. 267. N. 9. P. 6114-6121.
9. *Lim A, Zhang L.* WebPHYLIP: a web interface to PHYLIP. // *Bioinformatics*. 1999 Dec. V. 15(12), p. 1068-1069.
10. *Pesole G., Prunella N., Liuni S., Attimonelli M., Saccone C.* WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. // *Nucleic Acids Res.* 1992. V. 20. P. 287102875.

11. *Pevzner, P.A., Sze, S.-H.* Combinatorial approaches to finding subtle signals in DNA sequences. // *Proc. 8th Int. Conf. on Intelligent Systems for Molecular Biology ISMB'2000*, 2000, P. 269-278.

12. *Robison K., McGuire A.M., Church G.M.* A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. // *J. Mol. Biol.* 1998. V. 284. P. 241-254.

13. *Schweizer H.P., Po C.* Regulation of Glycerol Metabolism in Pseudomonas aeruginosa: Characterization of glpR Repressor Gene. // *J. Bacteriol.* 1996, 178. P. 5215-5221.

14. *Thompson J.D., Higgins D.G., Gibson T.J.* CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positionspectific gap penalties and weight matrix choice. // *Nucl. Acids Res.* 1994. V. 22. P. 4673–4680.

15. *Waterman M.S.* Multiple sequence alignment by consensus. // *Nucl. Acids Res.*, 14, 9095, 1986.

16. *Weissenborn D.L., Wittekindt N., Larson T.J.* Structure and Regulation of the glpFK Operon Encoding Glycerol Diffusion Facilitator and Glycerol Kinase of Escherichia coli K-12. // *J. Biol. Chem.* 1992. V. 267. N. 9. P. 6122-6131.

17. *Yang B., Larson T.J.* Action at a Distance for Negative Control of Transcription of the *glpD* Gene encoding *sn*-Glucerol 3-Phosphate Dehydrogenase of *Escherichia coli* K-12. // *J. Bacteriol.* 1996. V. 178. N. 24. P. 7090-7098.

18. *Миронов А.А., Винокурова Н.П., Гельфанд М.С.* Программное обеспечение анализа бактериальных геномов. // *Мол. биол.* 2000. Т. 34. № 2. С. 253-262.

Формат 60x90/16. Бумага офсетная №1, Печать офсетная.
Тираж 80 экз. Заказ № 0218-20111. Отпечатано в ООО «Эребус».