На правах рукописи

ЛЮБЕЦКАЯ Елена Васильевна

МАССОВЫЙ ПОИСК АТТЕНЮАТОРНОЙ РЕГУЛЯЦИИ В ГЕНОМАХ ПРОТЕОБАКТЕРИЙ

05.13.17 – Теоретические основы информатики,

03.00.28 – Биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата физико-математических наук

Москва – 2004

2007-4
14939

Работа выполнена в Институте проблем передачи информации
Российской Академии наук

Научный руководитель: член-корреспондент РАН, доктор биологических наук,
профессор Л.М. ЧАЙЛАХЯН

Официальные оппоненты: доктор физико-математических наук,
профессор В.Г. ТУМАНЯН,
доктор физико-математических наук,
профессор П.В. ГОЛУБЦОВ.

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, ГСП-4, Б. Каретный переулок, 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

2497532

ВВЕДЕНИЕ. ОБЩАЯ ХАРАКТЕРИСТИКА ПРОБЛЕМЫ

Актуальность темы. Текущие концентрации молекул в клетке в значительной мере зависят от протекающих биохимических реакций в ней (в работе рассматривается случай прокариот). Как правило, реакция протекает по мере поступления в цитоплазму клетки соответствующего набора ферментов, что зависит от экспрессии соответствующих групп генов (регулонов и их полицистронных случаев – оперонов). Таким образом, текущая жизнь клетки в значительной степени состоит в регуляции групп генов в зависимости от ее внутренних и внешних условий жизнедеятельности. Известны разные типы регуляции: основанная на белок-ДНКовом взаимодействии (позитивная или негативная, когда активатор или репрессор связывается с соответствующим сайтом в лидерной области оперона, обычно расположенным внутри промотера или вблизи него); или регуляция, основанная на образовании специфических вторичных структур мРНК (например, альтернативных и, в частности, аттенюаторных – в последнем случае механизм регуляции зависит от взаимного продвижения РНК-полимеразы и рибосомы в процессах транскрипции и трансляции); аллостерическая (когда конечный продукт катализируемой ферментом реакции ингибирует работу самого фермента) и другие. В процессах РНКовой регуляции большую роль играют: лидерный пептид с регуляторными кодонами, стабилизирующие белки, тРНК, молекулы-эффекторы и т.п. Часто в регуляции одного оперона участвуют несколько разных типов регуляции.

Регуляция с помощью белков-репрессоров или активаторов, а также аллостерическая регуляции изучаются сравнительно давно. Фундаментальная важность альтернативной регуляции обнаружена недавно, когда были найдены новые ее примеры.

В настоящее время расшифровано и доступно более 100 полных геномов, несколько сотен полных геномов секвенируются и будут доступны в ближайшее время, не говоря уже о секвенировании частей геномов. Такой огромный объем информации делает невозможным лабораторный биохимический анализ подавляющего большинства геномов, поэтому необходимы **алгоритмы** компьютерного анализа геномов и, в частности, *поиска потенциальных аттенюаторных структур* в достаточно полно секвенированных геномах, которые были бы применимы для **массового** анализа сразу всех организмов из данной таксономической группы.

В основном для поиска *регуляторных сигналов* до сих пор применялись два **подхода**: составлялось распознающее правило (по выборке лидерных областей, содержащих достаточно сходные регуляторные сайты); или такой сигнал искался непосредственно в каждой из последовательностей, входящих в выборку, на основе существенной консервативности сигнала. Оба эти подхода плохо применимы в случае массового поиска аттенюаторных структур. Ситуация особенно усложняется, когда речь идет о поиске регуляторного сигнала для генов с неизвестной функцией или для геномов, у которых еще не выяснена структура интересующего нас оперона.

Цель работы. Создание алгоритмов и программы для массового поиска аттенюаторной регуляции экспрессии генов. Тестирование эффективности алгоритмов и программы на различных

РОССИЙСКАЯ НАЦИОНАЛЬНАЯ БИБЛИОТЕКА С.Петербург

искусственных и биологических данных. Применение этих алгоритмов для решения биологической задачи поиска аттенюаторных сигналов регуляции у бактерий.

Методика исследования. Построение алгоритмов и компьютерных программ для поиска аттенюаторной регуляции в ***одном*** исходном нуклеотидном фрагменте, и затем применение их для массового поиска аттенюаторной регуляции у бактерий. Сначала методы сравнительной геномики применяются для построения предполагаемой оперонной структуры генов (в нашем случае биосинтеза некоторых аминокислот), отсюда в основном вручную выделяются потенциальные регуляторные области. К ним по отдельности применяются разработанные нами программы. Для подтверждения полученных потенциальных аттенюаторных структур для данного оперона у ряда родственных бактерий проводилось выравнивание участков регуляторных областей, содержащих найденные структуры (с некоторыми полями влево и вправо) так, чтобы оказались выровненными терминаторы, антитерминаторы и паузные шпильки, лидерные пептиды и другие аттенюаторные элементы. После такого подтверждения найденных сигналов (или после указания алгоритма об их отсутствии) уточнялись оперонные структуры в геномах и проводился окончательный анализ соответствующего метаболического пути биосинтеза аминокислот.

Таким образом, был проведен массовый поиск аттенюаторной регуляции у протеобактерий и в некоторых группах грамположительных бактерий.

Научная новизна. Предложенные алгоритмы – одни из первых для поиска регуляторных аттенюаторных сигналов по одной исходной нуклеотидной последовательности. Алгоритмы были реализованы в виде программного приложения, разнообразно тестированы и применены в задаче поиска сигналов аттенюаторной регуляции в геномах протеобактерий и грамположительных бактерий. Для последних эта задача до сих пор не рассматривалась.

Основные результаты. В диссертации получены следующие основные результаты.

- Предложены и реализованы в виде компьютерной программы, *называемой далее* **LLLM**, алгоритмы построения потенциальных структур аттенюаторной регуляции в геномах бактерий.
- Показана практическая эффективность и надежность созданной программы LLLM на основе ее детального тестирования на искусственных и биологических нуклеотидных последовательностях.
- Проведен массовый поиск и во многих случаях найдены потенциальные сигналы аттенюаторной регуляции (а в иных случаях алгоритм указал на предположительную причину их отсутствия) у гамма-, альфа- и бета-протеобактерий (биосинтез разветвленных и ароматических аминокислот, гистидина и треонина, фенилаланил-тРНК-синтетазы), а также и у грамположительных бактерий: из групп Bacillales, Lactobacillales, Clostridiales, Bacteroidetes/Chlorobi и Thermotogales (биосинтез гистидина); описана эволюционная динамика аттенюаторной регуляции транскрипции.

- Установлены потенциальные оперонные структуры для генов биосинтеза некоторых аминокислот (триптофана, фенилаланина, треонина, гистидина и разветвленные аминокислот) в различных геномах.

- Предсказано новое семейство гистидиновых транспортеров – ортологов *yuiF* у *B. subtilis* (например, HI0325 у *H. Influenzae*) и два гистидиновых транспортера BC0629 у *B. cereus* (ортолог *yvsH* у *B. subtilis*) из белкового семейства APA и ген у *L. lactis* (ортолог *lysQ* у *E. coli*) из семейства APC.

- Получена предположительная функциональная аннотация ряда генов, кодирующих ферменты и находящихся под аттенюаторной регуляцией, а именно:

  - гену *ygeA* у *Pasteurella multocida* приписана функция рацемазы разветвленных аминокислот;

  - не ортологичные гены *vatB*, *actX2* и *actX3*, соответственно, у *Pasteurella multocida*, *Mannheimia haemolytica*, *Polaribacter filamentus* кодируют ацетилтрансферазы, участвующие в метаболизме гистидина.

- Показано, что биосинтез изолейцина у Xanthomonadales использует треонин дегидратазу TdcB, в отличие от IlvA у *E. coli*.

- Показано, что у Pasteurellales бифункциональный ген *thrA* аспартат киназы/гомосерин дегидрогеназы регулируется не только треонинином и изолейцином (как это имеет место у *E. coli*), но и метионином.

- Предсказано, что у альфа-протеобактерий ацетолактат синтаза IlvIH регулируется аттенюацией с регуляторными кодонами лейцина, изолейцина и валина.

- Предсказано, что оперон *his* биосинтеза гистидина регулируется гистидин-зависимыми аттенюаторами у *Bacillus cereus* и *Clostridium difficile*, но и в тоже время регулируется гистидиновыми T-боксами у *Lactococcus lactis* и *Streptococcus mutans*.

Хорошее выравнивание этих биологических предсказаний рассматривается нами как еще одно подтверждение правильности работы программы LLLM.

Теоретическая и практическая ценность. До недавнего времени алгоритмы для поиска потенциальных альтернативных структур *в одной регуляторной области* не предлагались (наши алгоритмы излагаются в Главе 1). Публикации автора по таким алгоритмам были одними из первых работ, сравнение наших алгоритмов с немногочисленными другими, см.[1], приводится в тексте диссертации. Исследование аттенюаторной регуляции в классе протеобактерий начато сравни-

[1] Gorodkin J., Stricklin S.L., Stormo G.D. (2001) Discovering common stem-loop motifes in unligneted RNA sequences. Nucleic Acids Reseach, Vol. 29, No. 10, p. 2135-2144. Lathe W.C., Suyama M., Bork P. Identification of attenuation and antitermination regulation in prokaryotes. Genome Biology 2002, 3: prepr. 3. p. 1-60. Eddy S.R. (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. BMC Bioinformatics. 3:18, p.1-16.

тельно недавно, см., например, работы[2,3]; проведенный нами массовый поиск в классе грамположительных бактерий является первым. Нами предсказаны (Глава 3) новые регуляторные сигналы этого типа (включая лидерные пептиды и регуляторные кодоны) у гамма-, альфа- и бета-протеобактерий, у фирмикутов из групп Bacillales, Lactobacillales и Clostridiales, у бактерий из групп Bacteroidetes/Chlorobi и Thermotogales. Соответствующие выравнивания нуклеотидных участков исходных последовательностей показали хорошую согласованность между собой предсказанных нами регуляторных сигналов. В Главе 2 приводятся результаты систематического тестирования наших алгоритмов для аттенюаторных и Т-бокс структур. В частности, были независимо найдены все ранее известные случаи аттенюаторной регуляции в классе гамма-протеобактерий[2,3].

Апробация. Результаты диссертации докладывались на:

3-ей международной конференции «Проблемы управления и моделирования в сложных системах», Самара, РАН, 4-9 сентября 2001.

3rd 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 6 печатных работ.

Структура и объем работы. Диссертация состоит из 3 глав и 3 приложений; последние содержат основные результаты работы предложенных нами алгоритмов. Объем работы 110 страниц машинописного текста, в том числе, 13 таблиц и 26 рисунков.

Глава 1 содержит биологическую и математическую постановки задачи, описание полученных автором алгоритмов поиска аттенюаторной регуляции, краткое описание реализующей их компьютерной программы, сравнение двух полученных алгоритмов между собой.

---

[2] Panina, E.M., Vitreschak, A.G., Mironov, A.A. and Gelfand, M.S. (2001) Regulation of aromatic amino acid biosynthesis in gamma-proteobacteria. J. Mol. Microbiol. Biotechnol. 3, 529-543.

[3] Landick, R., Turnbough, C.L. and Yanovsky, C. (1994) Transcriptional attenuation. In: Escherichia coli and Salmonella. Cellular and molecular biology (Neidhardt, F.C., Ed.), p. 1263-1286. American Society for Microbiology, Washington, DC.

Глава 2 содержит результаты компьютерного тестирования наших двух алгоритмов на биологических последовательностях, для которых аттенюаторы были известны, и на искусственных последовательностях. Тестирование этих алгоритмов продолжается в Главе 3 на задаче массового поиска аттенюаторной регуляции в обширных классах бактерий.

В главе 3 последовательно рассматриваются результаты массового поиска потенциальных сигналов аттенюаторной регуляции у протеобактерий и фирмикутов, у бактерий из групп Bacteroidetes/Chlorobi и Thermotogales для случаев биосинтеза разветвленных и ароматических аминокислот, гистидина и треонина. Для каждого из этих случаев приводится и кратко комментируется соответствующая потенциальная оперонная структура. Затем приводятся выравнивания найденных РНК-структур.

В конце работы приведены выводы и список из 6 публикаций диссертанта.

СОДЕРЖАНИЕ РАБОТЫ

ГЛАВА 1. Два алгоритма и компьютерная программа поиска потенциальных аттенюаторных регуляторных структур мРНК

Рассматривается задача поиска аттенюаторной регуляции в лидерной области гена. Условимся *обозначать* плечи паузной шпильки 1 и 2, плечи терминатора 3 и 4, а плечи антитерминатора 2'и 3'.

Первый алгоритм применяется для исследования вопроса о наличии аттенюаторной регуляции в **одной** данной лидерной области: в ней ищется сама аттенюаторная структура или обоснование ее отсутствия (т.е. ищется алгоритмическое указание на отсутствующий элемент такой структуры). Нами обычно рассматривались случаи, когда лидерная область имеет длину в 450-800 нуклеотидов. Алгоритм ищет *структуру*, состоящую из следующих элементов в исходной нуклеотидной последовательности: рамки считывания, кодирующей лидерный пептид с набором регуляторных кодонов в нем; тройки слов s_1, s_2, s_3, которые служат «основой» для построения плеч: s_1 плеч 2 и 2', s_2 плеч 3 и 3', s_3 плеча 4; и трех шпилек – терминатора, антитерминатора и паузной (в соответствии с этими плечами); участок остатков урацила.

В алгоритме используются следующие основные параметры: минимальная длина лидерного пептида (обычно, 30 нуклеотидов); длина участка, содержащего остатки урацила (обычно 7); минимальное количество самих остатков урацила (обычно 5) в этом участке; длина каждого из слов s_1, s_2, s_3 (обычно 8); максимальное расстояние (разница между началом первой буквы следующего слова и последней буквой предыдущего слова) между словами s_1 и s_2, s_2 и s_3 (обычно 100); максимальное число различных (в паре s_1 и s_3) и некомплементарных (в парах s_1 и s_2, s_2 и s_3) нуклеотидов (обычно 2); максимальная расстояние между началом слова s_3 и концом слова s_2 (обычно 30); максимальное число отбираемых алгоритмом троек слов s_1, s_2, s_3 (обычно 15); максимальное рас-

стояние между началом участка остатков урацила и концом слова s_3 (обычно 9); минимальное отношение числа пар GC к числу пар AT в словах s_2 и s_3 (обычно 1.3); интервальное значение длины фрагмента до начала петли антитерминатора – паузная шпилька ищется в этом фрагменте (обычно от 100 до 50).

Первый алгоритм последовательно находит:

1) кандидатов в лидерные пептиды. Для этого перебираются все открытые рамки считывания, в каждой из которых ищется скопление регуляторных кодонов соответствующей аминокислоты (список этих кодонов может расширяться в процессе счета – он также является параметром алгоритма).

2) Для каждого найденного кандидата в лидерные пептиды ищутся все подходящие участки остатков урацила.

3) Для каждой пары таких объектов ищутся тройки слов s_1, s_2, s_3; тройки сортируются в порядке увеличения расстояния между s_2 и s_3, из них отбирается заданное число первых.

4) Затем последовательно по s_2 и s_3 строится терминатор из условия минимума энергии, и аналогично – антитерминатор и паузная шпилька (при заданных параметрах).

5) Среди так полученных структур отбираются по одному представителю из каждого класса «подобных» структур (определяемого нами некоторым отношением эквивалентности). Для наших случаев, как правило, фактически оказывалось по одному-двум, но не более трех таких представителей.

Итак, терминатор образуется спариванием слов s_2 и s_3 и продолжением спаривания в обе стороны. Альтернативность терминатора и антитерминатора поддерживается за счет слова s_2: антитерминатор – это шпилька, включающая спаривание s_1 и s_2 с изменением свободной энергии сворачивания, меньшим некоторого параметра (например, –10 ккал/моль, что обеспечивает возможность сворачивания антитерминатора).

Правильность работы первого алгоритма подтверждается сравнением результатов счета с известными биологическими данными, а также результатами совместных выравниваний найденных нами новых и уже известных сигналов аттенюаторной регуляции. Выравнивания говорят, в частности, о консервативности указанных троек слов. Действительно,

1. При нашем выравнивании для многих оперонов и независимо от алгоритма определяются слова s_1, s_2, s_3 и они оказываются консервативными. Так, для *trp*-оперона эти слова выровнялись даже у гамма-, альфа- и бета-протеобактерий. Для каждого из 8 изученных оперонов у организмов из одной группы эти слова оказывались одинаковыми с точностью до нескольких букв (см. рисунки 3-6 в Приложении 2).

2. Для многих оперонов (например, *thrABC* у гамма-протеобактерий) антитерминаторы, полученные алгоритмом для разных организмов, имели сходную структуру: близкие значения числа

отрезков и их длин, близкие значения длин выпячиваний и их типов (односторонние или двусторонние), и т.д.

Алгоритм обеспечивает следующие условия:

1. Наличие лидерного пептида с полем регуляторных кодонов соответствующих аминокислот.
2. Пересечение терминатора и антитерминатора.
3. Отсутствие пересечения паузной шпильки и терминатора.
4. Расстояние от конца поля регуляторных кодонов до начала левого плеча антитерминатора более 5 нуклеотидов.

Кроме того, алгоритм проверяет полученные им ответы на выполнение условия, которое наблюдалось во многих известных случаях аттенюаторной регуляции: расстояние от конца лидерного пептида до начала левого плеча антитерминатора от –3 до +3 нуклеотидов.

Упомянутое хорошее выравнивание по разным причинам не является обязательным условием наличия аттенюаторной регуляции. В некоторых организмах (например, *Y. pestis*) фрагменты последовательности, участвующие в такой регуляции, не являются консервативными в других гамма-протеобактериях.

Второй алгоритм основан на идее, предложенной в работе[4], и применяется, как и первый, для поиска аттенюаторного сигнала в **одной** нуклеотидной последовательности. Его применение целесообразно в случае слабо выраженного сигнала: он выдает больше вариантов ответа, но может находить альтернативный сигнал, мало похожий на «классическую» аттенюацию.

Этот алгоритм (примерно квадратичный от размера исходных данных) по любой нуклеотидной последовательности выдает список потенциальных аттенюаторных структур в ней. Каждая такая структура состоит теперь из тройки шпилек <терминатор, антитерминатор, пауза>. Ниже в качестве примера указываются численные значения, которые, конечно, являются параметрами этого алгоритма и могут варьироваться.

Алгоритм состоит из двух этапов, которые в автореферате описаны схематично, а в тексте диссертации подробно.

Этап 1: *порождение множества локально оптимальных шпилек.* Количество элементов в этом множестве приблизительно равно длине исходной последовательности. Этап 1 состоит из следующих двух шагов.

1. Последовательность делится на фрагменты фиксированной длины и внутри каждого из них индуктивно порождаются локально оптимальные шпильки (т.е. максимальные элементы в смысле некоторого фиксированного отношения частичного порядка).

---

[4] Верещагин Н.К., Любецкий В.А. Алгоритм определения вторичной структуры РНК. Труды научно-исследовательского семинара логического центра ИФ РАН, выпуск 14, Москва, Издательство РАН, 2000, с. 99-109.

2. На индуктивном шаге переопределяются лишь параметры шпилек. Только малое количество специально отобранных алгоритмом шпилек порождается полностью, т.е. как совокупность пар комплементарных нуклеотидов.

Этап 2: *построение самой аттенюаторной структуры*, который состоит из следующих двух шагов.

1. В множестве всех локально оптимальных шпилек отбираются те внешние петли (этих шпилек), начала и концы (B,C) которых повторяются более чем $p = 9$ раз. Такие пары (B,C) будем называть *частыми*.

2. В так полученном списке внешних петель отбираются петли, у которых частые пары (B,C) расположены рядом с участком остатков урацила U. Они считаются петлями будущих терминаторов. Затем для каждой такой петли строится сам терминатор, а для него последовательно порождаются еще две оставшиеся шпильки – сначала антитерминатор и затем паузная шпилька.

Построение терминатора для данных (B,C) и участка остатков урацила U происходит «вытягиванием» отрезков от (B,C), т.е. последовательным спариванием как можно большего числа нуклеотидов в 1-2 отрезка.

При построении антитерминатора и паузы используется понятие ядра. *Ядром* для данного множества шпилек с одной и той же внешней петлей (B,C) с координатами начала B и конца C петли называется консенсусная шпилька, т.е. шпилька наилучшим образом согласованная со всеми шпильками из этого множества. Например, в качестве ядра нами бралась шпилька, состоящая из пар нуклеотидов, которые являются спаренными в более чем половине от всех шпилек из этого множества. Для наших случаев, как правило, получалось ядро, состоящее из двух отрезков (иногда оно содержало от 1 до 5 отрезков). Заметим, что в большинстве случаев так вычисляемые ядра несколько короче биологических шпилек, участвующих в аттенюации.

Итак, для поиска антитерминатора при уже полученном терминаторе берется частая пара (B_1,C_1) слева и ближайшая к петле (B,C) терминатора. Рассмотрим все локально оптимальные шпильки с данной парой (B_1,C_1) , включая и их подшпильки; и среди них в качестве антитерминатора отберем все те, у которых плечо C_1D_1 имеет не менее 5 общих нуклеотидов с плечом AB, и также $A > B_1$, $C > D_1$. Если таким образом нашлось пустое множество, то в качестве антитерминатора возьмем все ядра, построенные для множества всех локально оптимальных шпилек с данным (B_1,C_1).

Паузную шпильку построим по каждому уже полученному антитерминатору аналогично тому, как антитерминатор строился по полученному терминатору. Для каждого терминатора выбираем одну соответствующую ему пару антитерминатор-пауза как пару с наибольшей суммарной мощностью. Так полученные структуры ранжируются по мощности терминатора, а при одинаковой его мощности по возрастанию координаты B его петли.

Алгоритмы реализованы на языке Object Pascal в среде[5] Delphi 5; вспомогательные алгоритмы – на языке Perl.

Ряд дополнительных описаний, касающихся работы первого и второго алгоритмов, приведены в тексте диссертации.

Сравнение результатов работы первого и второго алгоритмов. Отметим общие результаты сравнения их работы.

1. Терминаторы с выраженными признаками (т.е. наличие участка остатков урацила U, отсутствие выпячиваний, GC-насыщенность, большое число комплементарных пар нуклеотидов) одинаково хорошо находятся с помощью обоих алгоритмов.

2. Первый алгоритм, в отличие от второго, с хорошей точностью находит антитерминаторы и паузные шпильки.

3. Если в исходной последовательности отсутствует лидерный пептид или антитерминатор неканонический, то первый алгоритм не находит структуру, но она может успешно быть найденной вторым алгоритмом. Например, в случае регуляции гена *pheA* в организме *Y. pestis* перед аттенюаторной структурой встраивается мобильный элемент, и лидерный пептид хотя, конечно, существует, но находится на большом расстоянии от вторичной структуры и фактически не попадает в исходную последовательность разумной длины. В этом примере, антитерминатор также не содержит в последнем от его петли отрезке комплементарных слов s_1 и s_2, что рассматривается нами как случай неканонического антитерминатора, и первый алгоритм его не находит.

Продолжая сравнение алгоритмов, можно добавить следующее: первый алгоритм оперирует с бóльшим числом параметров и объектов, поэтому «классическая» аттенюаторная регуляция находится им более точно. Второй алгоритм из биологических особенностей использует только энергетические соображения, наличие сгущения нуклеотидов U и т.п. Поэтому, если искомая структура содержит не все стандартные элементы классической аттенюаторной регуляции, то второй алгоритм, в отличие от первого, может ее найти. Для успешной работы второго алгоритма желательно присутствие мощного терминатора в исходной последовательности, а для успешной работы первого алгоритма важно наличие там лидерного пептида.

Аттенюаторные структуры, **найденные вторым алгоритмом и не найденные первым** среди 18 тестовых примеров, перечисленных в таблице 2.5 Приложения 1, приведены в таблице 1.

[5] Алгоритмы также реализованы на языке ANSI C для параллельной вычислительной архитектуры с протоколом MPI – этот результат не включается в диссертационную работу. Часть вычислений велась на суперкомпьютере МВС-1000М (в МСЦ Миннауки, РАН, МГУ и РФФИ).

В качестве этих тестовых брались лидерные области, в которых аттенюация известна[6]; в случаях *E. coli_ilvBN* (и *E. coli_G*) она получена экспериментально[7].

Таблица 1.

Название организма и гена	**Наличие лидерного пептида**	**Комплементарная тройка слов – существенный параметр первого алгоритма**
Y. pestis_trpE	да	нет
Y. pestis_pheA1	нет	да
Y. pestis_pheA2	нет	нет
E. coli_ilvBN	да	нет

ГЛАВА 2. Тестирование алгоритмов

Тестирование двух приведенных выше алгоритмов проводилось для следующих семи типов условий. Для каждого типа тестировался один из алгоритмов.

1) Второй алгоритм: тестирование на случайных последовательностях. На вход алгоритма подавалась случайная бернуллиевская последовательность длины 450 (порождаемая с помощью генератора случайных чисел стандартной библиотеки языка Perl). На выходе появлялось сообщение вида <Терминатор, Антитерминатор или -, Пауза или -, Участок U или ->, в котором знак «-» говорит об отсутствии соответствующего элемента. (или сообщение об ее отсутствии). Участком U считались фрагменты последовательности длины 7, в которых нуклеотид U встречался не менее 5 раз и допускались разрывы: два из одной буквы или один из двух букв. Найденными считались структуры, содержащие терминатор из не менее, чем трех пар комплементарных нуклеотидов, в то время как антитерминатор и/или паузная шпилька, участок U могли отсутствовать.

Было проведено 67 запусков алгоритма с указанными выше параметрами. Результаты приведены в таблице 2.1.1 Приложения 1. Из них видно, что в случайной последовательности, по крайней мере, терминатор находится почти в 40% случаев.

Таблица 2.1.1. Второй алгоритм: поиск аттенюаторных структур в бернуллиевских случайных последовательностях.

Аттенюаторная структура в случайной последовательности:	Число соответствующих случаев:
Найдена (с участком U)	29

[6] Panina, E.M., Vitreschak, A.G., Mironov, A.A. and Gelfand, M.S. (2001) Regulation of aromatic amino acid biosynthesis in gamma-proteobacteria. J. Mol. Microbiol. Biotechnol. 3, 529-543.

[7] Landick, R., Turnbough, C.L. and Yanovsky, C. (1994) Transcriptional attenuation. In: Escherichia coli and Salmonella. Cellular and molecular biology (Neidhardt, F.C., Ed.), pp. 1263-1286. American Society for Microbiology, Washington, DC.

Не найдена (но участок U имеется)	21
Не найден даже участок U	17

Рассмотрим число пар (B,C), повторяющихся более чем *p* раз в множестве локально оптимальных шпилек, которое образовано вторым алгоритмом по исходной нуклеотидной последовательности. Такую пару (B,C) мы назвали *частой*. Число частых пар при значении порога $p = 9$ для тех же случайных и 20 тестовых биологических последовательностей приводятся в таблице 2.1.2 Приложения 1, третий столбец. Из нее и других наших вычислений видно, что число различных пар (B,C) в множествах локально оптимальных шпилек примерно на порядок меньше длины исходной последовательности, и число частых пар примерно в 4 раза меньше числа всех пар (B,C).

2) *Второй алгоритм: тестирование на случайных последовательностях с участком* U. На вход алгоритма подавались 20 бернуллиевских случайных последовательностей, использованных в пункте 1 тестирования, в которые дополнительно в случайных местах вставлялись 1-2 фрагмента из 5 нуклеотидов U, которые находились от начала последовательности на расстоянии не менее 100 нуклеотидов (с тем, чтобы алгоритм имел возможность построить структуру перед участком U). На выходе появлялось такое же как выше сообщение <Терминатор мощности не менее 3, Антитерминатор или -, Пауза или -, Участок U или -> или сообщение об отсутствии такого терминатора.

В этом случае структуры находились чаще, чем в пункте 1, что, конечно, нежелательно. Поэтому было предложено указанное ниже дополнительное ко второму алгоритму условие, которое позволяет отличить последовательность, случайную даже с вставленным в нее (как выше) участком U, от биологической последовательности, содержащей аттенюацию. Таким образом, второй алгоритм, дополненный этим условием, может применяться и для решения задачи различения случайной последовательности от биологической, содержащей аттенюацию. Численные значения указанных ниже параметров соответствуют рассмотренному нами случаю последовательностей длины 450. Это *условие* состоит в следующем. Биологическая последовательность, содержащая аттенюаторную регуляцию, отличается от случайной последовательности с вставленными в нее 1-2 участками U, если для структур, найденных в ней вторым алгоритмом, выполняется одно из трех следующих свойств 1-3.

1. Ответ содержит ровно одну структуру и она «хорошая» в смысле:

а) участок U около терминатора найденной структуры содержит не менее 6 нуклеотидов подряд;

б) найденная структура имеет пересекающиеся (не менее чем на 5 нуклеотидов) терминатор и антитерминатор, а также паузную шпильку;

в) терминатор имеет не менее 5 пар комплементарных нуклеотидов.

2. Среди двух или трех найденных структур хотя бы одна удовлетворяет условиям 1б и 1в.

3. Найдено не менее 4 структур с различными парами (B,C) координат петли терминатора.

Из таблицы 2.2 Приложения 1 видно, что случайная последовательность с участком U устойчиво отличается от биологической последовательности, содержащей аттенюацию. Причем, как правило, эта аттенюация выдается нашим алгоритмом в качестве *первого* ответа, когда все находимые ответы *ранжируются* по убыванию мощности терминатора, а при одинаковой мощности по возрастанию координаты B из пары (B,C) терминатора. Напомним, что *мощность* шпильки это – число комплементарных пар в ней.

3) Второй алгоритм: тестирование на случайных последовательностях, содержащих биологически значимые терминаторы. На вход алгоритму подавались случайные бернуллиевские последовательности, использованные в пункте 2 тестирования, в которые перед участком U на расстоянии в 3 нуклеотида помещался фрагмент, содержащий биологически значимый терминатор из регуляторной области гена *trpE* организма *E. coli*, а именно, фрагмент `cagcccgcctaatgagcgggc`. На выходе алгоритм выдавал такое же как выше сообщение.

На таких последовательностях второй алгоритм не дает удовлетворительных результатов: он не может систематически отличать такие последовательности от биологически значимых. Это видно из таблицы 2.3 Приложения 1.

4) Второй алгоритм: тестирование на биологических последовательностях, содержащих аттенюаторную структуру. На вход алгоритму подавались регуляторные области длины 450, содержащие уже известную аттенюаторную регуляцию. На выходе алгоритма выдавалось такое же сообщение. Результаты тестирования приводятся в таблице 2.4 Приложения 1, из которой видно, что алгоритм находит аттенюацию с высокой точностью. Естественно, что он находит терминатор с более высокой точностью, чем антитерминатор, и последний с более высокой точностью, чем паузу.

5) Первый алгоритм: тестирование на биологических последовательностях, содержащих аттенюаторную структуру. На вход алгоритму подавались регуляторные области длины 450 из предыдущего пункта, содержащие аттенуаторную регуляцию. На выходе алгоритм выдавал четверку <Лидерный пептид, Терминатор, Антитерминатор, Пауза> или сообщение об ее отсутствии. В подавляющем большинстве случаев алгоритм нашел ответ с высокой точностью, что видно из таблицы 2.5 Приложения 1.

6) Второй алгоритм: тестирование на биологических последовательностях, содержащих альтернативные структуры типа Т-бокс. На вход алгоритма подавались регуляторные области длины 350-550, содержащие альтернативную регуляцию типа Т-бокс. На выходе алгоритм выдавал пары шпилек вида <антитерминатор, терминатор> или сообщение об их отсутствии. В этом случае терминаторы нашлись с точностью до 1-2 нуклеотидов; точность нахождения антитерминатора указана в таблице 2.6 Приложения 1 и обычно она была весьма высокой.

7) Второй алгоритм: тестирование на биологических последовательностях, не содержащих аттенюаторов по результатам работы первого алгоритма. На вход второму алгоритму подавались лидерные области генов, в которых первый алгоритм не нашел аттенюаторной регуляции (в частности, не нашел лидерного пептида). Из таблицы 2.7 Приложения 1 видно, что примерно в 80% случаев второй алгоритм также не нашел в них аттенюаторной регуляции.

ГЛАВА 3. Массовый поиск аттенюаторной регуляции

В этой главе разработанная нами компьютерная программа LLLM, реализующая два выше описанных и тестированных алгоритма, применялась для массового поиска аттенюаторной регуляции у протеобактерий, у фирмикутов и у бактерий из групп Bacteroidetes/Chlorobi и Thermotogales.

Процедура поиска заключалась в следующем: полные и частично секвенированные последовательности бактериальных геномов выгружались из базы данных Genbank (института NCBI). Список рассмотренных нами геномов приводятся в таблице 1 Приложения 2. Похожесть белков определялась с помощью алгоритма Смита-Уотермана, реализованного программой GenomeExplorer[8]. Ортологичные белки находились как наилучшие двунаправленные хиты[9] сразу для нескольких геномов. В некоторых случаях приходилось дополнительно строить филогенетические деревья белков. Например, у *P. multocida* был найден ген, высоко гомологичный двум генам *ilvG* и *ilvB* из *E. coli*, и только филогенетическое дерево построенное для данного семейства белков позволило аннотировать его как *ilvG*. Филогенетические деревья строились методом максимального правдоподобия, реализованного в пакете[10] Phylip, множественные выравнивания выполнялись программой[11] CLUSTAL W. Для аннотации генов также применялась программа поиска трансмембранных сегментов (TMpred, http://www.ch.embnet.org/software/TMPRED_form.html) и использовались базы, содержащие функциональную и структурную аннотацию белков, в частности[9,12] COG и InterPro. Затем, с учетом найденных сигналов регуляции, были получены оперонные

---

[8] Mironov, A.A., Vinokurova, N.P. and Gelfand, M.S. (2000) GenomeExplorer: software for analysis of complete bacterial genomes. Mol. Biol. 34, 222-231.

[9] Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. and Koonin, E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29, 22-28.

[10] Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368-376.

[11] Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876-4882.

[12] Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R., Courcelle, E., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Griffith-Jones, S., Haft, D., Hermjakob, H., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Orchard, S., Pagni, M., Peyruc, D., Ponting, C.P., Servant, F. and Sigrist, C.J. (2002) InterPro: an integrated documentation resource for protein families, domains and functional sites. Brief Bioinform. 3, 225-235.

структуры для рассмотренных нами организмов и генов биосинтеза аминокислот (приведены в таблицах 2-5 Приложения 2).

Выявление потенциальных оперонных структур позволило найти предполагаемые лидерные области оперонов для поиска в них различных типов регуляции. Поиск аттенюаторных регуляторных структур выполнялся с помощью нашей программы LLLM, а также (в случае белок-ДНКовой регуляции) с помощью программы[13] IRSA или иными средствами поиска других упомянутых в таблицах 2-5 типов регуляции. На защиту выносятся только случаи аттенюаторной регуляции, поэтому другие регуляции подробно не обсуждаются.

Средством подтверждения обоснованности наших результатов об аттенюаторной регуляции служило выравнивание лидерных участков, содержащих соответствующие структуры (с небольшими полями) от начала лидерного пептида до окончания терминатора и участка остатков урацила (при его наличии). Эти выравнивания приведены на рисунках 3-6 Приложения 2. Соответствующие метаболические карты приведены на рисунке 2 Приложения 2.

Биосинтез изолейцина, лейцина и валина (сокращенно ILV). Новые возможные транскрипционные аттенюаторы найдены у гамма-, альфа-, бета-протеобактерий. Возможные аттенюаторы оперонов *ilvGMEDA* найдены у энтеробактерий, у Pasteurellales (только в *P. multocida*), у Vibrionales, Alteromonadales (в *S. oneidensis*), у Xanthomonadales. У Pseudomonadales найдена возможная аттенюация отдельно лежащего гена *ilvA*. У энтеробактерий найдена аттенюаторная регуляция оперона *ilvBN*, содержащего гены, кодирующие один из изозимов ацетолактат-синтазы. Гены лидерных пептидов оперона *ilvG* содержат регуляторные кодоны трех аминокислот – изолейцин, лейцин и валин (как и в экспериментально изученном случае *E. coli*), а оперона *ilvBN* регуляторные кодоны только двух аминокислот – лейцина и валина.

Структура потенциальных оперонов биосинтеза ILV весьма разнообразна. Например, у энтеробактерий и Vibrionales она имеет классический вид *ilvGMEDA*, а у Xanthomonadales – вид *ilvCGM-tdcB-leuA*. Здесь ген *tdcB* предположительно регулируется вместе с этим опероном и кодирует треонин дегидратазу, которая участвует в биосинтезе изолейцина.

У *P. multocida* отмечен ген с неизвестной функцией (гомологичный гену *ygeA* у *E. coli*), который локализуется в *ilv*-опероне *ilvGM-ygeA-ilvDA*. Ген *ygeA* слабо похож на аспартат рацемазу *racX* из *B. subtilus*; весьма вероятно, что он кодирует новый тип рацемазы разветвленных аминокислот.

Среди гамма-протеобактерий оперон *leu* регулируется аттенюацией у энтеробактерий, Pasteurellales, Vibrionales, Alteromonadales, но не у Pseudomonadales и других видов. Регуляторными во всех этих случаях являются лейциновые кодоны.

---

[13] Данилова Л.В., Горбунов К.Ю., Гельфанд М.С., Любецкий В.А. (2001) Алгоритм выделения регуляторных сигналов в последовательностях ДНК. Мол. биол., том 35, № 6, с. 987-995.

В группе альфа-протеобактерий впервые обнаружена потенциальная аттенюаторная регуляция единственной ацетолактат синтетазы *ilvIH* у Rhizobiales (*Sinorhizobium meliloti, Agrobacterium tumefaciens, Mesorhizobium loti, Bradyrhizobium japonicum, Rhodopseudomonas palustris, Brucella melitensis*), у *Rhodobacter* spp., *Magnetospirillum magnetotacticum* и у *Caulobacter crescentus*. Регуляторными кодонами этой аттенюации являются кодоны изолейцина, лейцина и валина. Заметим, что у гамма-протеобактерий опероны, участвующие в биосинтезе двух ацетолактат синтетаз IlvGM и IlvBN (у энтеробактерий), но не IlvIH, регулируются аттенюацией.

Группа ортологов гена *leuA* изопропилмалат синтазы из *E. coli* обнаружена у гамма-протеобактерий (исключая Pseudomonadales) и у некоторых альфа-протеобактерий. Группа генов, названная нами *leuA2*, которые похожи на гены, кодирующие изопропилмалат синтазу у *Corynebacterium glutamicum*, была найдена у альфа-протеобактерий, у некоторых бета-протеобактерий и у Pseudomonadales. В альфа-протеобактериях оба типа 2-изопропилмалат синтазы LeuA и LeuA2 имеют потенциальные аттенюаторы. Эти предполагаемые аттенюаторы имеют лидерные пептиды с лейциновыми регуляторными кодонами, но их терминаторы слабые и не имеют участков остатков урацила. Эта ситуация кажется подобной регуляции оперонов *trpE* и *trpGDC* у Pseudomonadales, где имеется аттенюаторная регуляция в отсутствие ро-независимой терминаторной структуры.

Биосинтез гистидина. Потенциальные аттенюаторы были найдены у многих гамма-протеобактерий, у фирмикутов и у бактерий из групп Bacteroidetes/Chlorobi, и Thermotogales. В большей части гамма-протеобактерий (энтеробактерии, Pasteurellales, Vibrionales, Alteromonadales) имеется единый *his*-оперон, предположительно регулируемый аттенюаторами с гистидиновыми регуляторными кодонами, и имеется выраженная терминатор/антитерминаторная структура. Аттенюация не была найдена для *his*-генов у Pseudomonadales, Xanthomonadales и у некоторых других гамма-протео-бактерий.

У Bacillus/Clostridium, Bacteroidetes/Chlorobi, Thermotogales гистидиновые регулоны включают большую часть обычных *his*-генов и также ген *hisZ* или *hisS* гистидил-тРНК-синтетазы; они регулируются аналогично.

Отметим разнообразие механизмов регуляции *his*-генов. Например, у *Lactococcus lactis* и *Streptococcus mutans his*-оперон регулируется с помощью антитерминаторного механизма с образованием структуры Т-бокса[14], а у *Bacillus cereus* и *Clostridium difficile* он предположительно регулируется классической аминокислотной аттенюацией; в тоже время, другие *Streptococcus* spp., а также *Entrerococcus* spp. лишены *his*-генов.

Что касается гистидил-тРНК-синтетазы, то у *Bacillus cereus* имеются три кодирующие ее гена *hisZ*, *hisZ2* и *hisS*. Оба гена *hisZ* и *hisZ2* регулируются аттенюацией: один в составе *his-*

[14] Delorme, C., Ehrlich, S.D. and Renault, P. (1999) Regulation of expression of the Lactococcus lactis histidine operon. J. Bacteriol. 181, 2026-2037. Витрешак А.Г. Частное сообщение.

оперона, другой как отдельный ген. Третий ген *hisS*, как и его ортологи у *Bacillus* spp., *Listeria* spp., *Enterococcus* spp. и *Lactococcus lactis*, (предположительно) регулируется с помощью антитерминаторного механизма с образованием структуры Т-бокса[15].

Показана регуляция гистидином следующих генов. У *H. influenzae* ген *HI0325* кодирует возможный транспортер (с 10 трансмембранными сегментами) и имеет потенциальный аттенюатор. В ряде геномов (в частности, у *Fusobacterium nucleatum* и *Bacillus halodurans*) этот ген кластеризуется с генами утилизации гистидина (*hut* локус). Возможно, этот ген и его ортологи (например, *yuiF* у *B. subtilis*) образуют новое семейство гистидиновых транспортеров.

У *B. cereus* ген *BC0629*, возможно, регулируется аттенюатором с гистидиновыми регуляторными кодонами. Этот ген ортологичен *yvsH* у *B. subtilis*, гомологичен аргинин:орнитин антипортеру *arcD* у *Pseudomonas aeruginosa* и лизиновому транспортеру *lysI* у *Corinobacterium glutamicum*. Эти белки относятся к семейству APA антипортеров аминокислот и полиаминов.

У *B. cereus* имеются два паралога *yvsH* (BC0629 и BC0865), из которых первый имеет аттенюаторную регуляцию с гистидиновыми регуляторными кодонами, а второй (предполагаемый лизиновый транспортер) регулируется лизином с помощью лизин-специфичного регуляторного элемента[16]. Подобная ситуация наблюдается у *L. lactis*, где видны два паралогических транспортера LysP и LysQ. Оба белка подобны (более, чем на 50%) экспериментально подтвержденной лизиновой пермеазе LysP из *E. coli*. У *L. lactis* ген *lysP* регулируется *LYS*-элементом (и, по видимому, участвует в транспорте лизина), а ген *lysQ* предположительно регулируется аттенюатором с гистидиновыми регуляторными кодонами. Таким образом, эти два транспортера могут иметь различное сродство к лизину и гистидину, и потому по разному регулироваться.

Все гены гистидинового регулона были найдены во всех анализируемых бактериях, за исключением гистидинол фосфатазы HisB у *Pseudomonas* spp.

Найдены три негомологичных гена с неизвестной функцией (*actX2*, *vatB*, *actX3* соответственно у *Mannheimia haemolytica, Pasteurella multocida, Polaribacter filamentus*), возможно кодирующих ацетилтрансферазы и регулируемые совместно с *his*-генами. Эти белки могли бы катализировать превращение гистамина в 4-бета-ацетиламиноэтил-имидазол (EC 2.3.1.-).

Биосинтез треонина. В опероне биосинтеза треонина у энтеробактерий, Pasteurellales, Vibrionales, Alteromonadales и Xanthomonadales наблюдается обычный порядок генов *thrABC*. У Pseudomonadales и еще у некоторых бактерий гены треонинового биосинтеза разбросаны по геному. Более того, у энтеробактерий, Pasteurellales, Vibrionales, Alteromonadales и Xanthomonadaels ген *thrA* кодирует бифункциональный белок – аспартат киназу/гомосерин дегидрогеназу, а у

[15] Chopin, A., Biaudet, V. and Ehrlich, S.D. (1998) Analysis of the Bacillus subtilis genome sequence reveals nine new T-box leaders. Mol. Microbiol. 29, 662-664. Витрешак А.Г. Частное сообщение.

[16] Rodionov, D.A., Vitreschak, A.G., Mironov, A.A. and Gelfand, M.S. (2003) Regulation of lysine biosynthesis and transport genes in bacteria: yet another RNA riboswitch? Nucleic Acids Res. 31, 6748-6757.

Pseudomonadales и у некоторых других гамма-протеобактерий *thrA2* (аспартат киназа) и *hom* (гомосерин дегидрогеназа) расположены в разных локусах. У Pseudomonadales наблюдались два гена гомосерин киназы *thrB2* и *thrH*, которые не гомологичны *thrB* из *E. coli*.

Треониновые опероны регулируются аттенюацией у энтеробактерий, Pasteurellales, Vibrionales, Alteromonadales и *Xanthomonas campestris*. Все предполагаемые аттенюаторы имеют треониновые и изолейциновые регуляторные кодоны, а также выраженные терминаторы и антитерминаторы.

Нами установлено, что у Pasteurellales (*Haemophylus influenzae, Pasteurella multocida, Actinobacillus actinomycetemcomitans* и *Mannheimia haemolytica*) регуляторными кодонами служат не только треонин и изолейцин, но и метионин. В самом деле, предполагаемая регуляция *thr*-оперона у Pasteurellales концентрациями треонина, изолейцина и метионина может быть обоснована наличием у Pasteurellales только одной аспартат киназы/гомозерин дегидрогеназы, вместо двух изозимов ThrA и MetL у других гамма-протеобактерий, которая входит в метаболические пути синтеза этих трех аминокислот.

Третья монофункциональная аспартат киназа LysC имеется у трех из пяти Pasteurellales, а именно, у *P. multocida, Haemophylus ducrei* и *M. Haemolytica*; и экспрессия гена *lysC* предположительно регулируется лизином посредством *LYS*-элементов (как у *E. coli*).

Биосинтез ароматических аминокислот (триптофана, фенилаланина) и фенилаланил-тРНК-синтетазы. Были найдены предполагаемые *trp*-, *pheA*- и *pheST*-опероны у альфа-, бета- и у многих новых гамма-протеобактерий.

Аттенюаторная регуляция *trp* была показана для энтеробактерий, Vibrionales, Alteromonadales с регуляторными кодонами триптофана и при наличии сильных терминаторов и антитерминаторов. Ген *trp(E/G)*, возникший в результате слияния и кодирующий две компоненты антранилат синтазы (первый шаг триптофанового биосинтеза), возможно, регулируется аттенюацией во всех проанализированных нами альфа-протеобактериях из группы Rhizobiales кроме *Brucella melitensis*.

Оперон *pheA*, возможно, регулируется аттенюаторами (с регуляторными кодонами фенилаланина) у энтеробактерий, Vibrionales, Alteromonadales, а оперон *pheST* регулируется таким же образом только у энтеробактерий.

Опероны *trpE* и *trpGDC* у Pseudomonadales имеют некоторые особенности: несмотря на экспериментальные данные об аттенюаторной регуляции этих оперонов[17], нами найдены только потенциальные лидерные пептиды с парой близких триптофановых кодонов, которые хорошо выравниваются для пяти Pseudomonadales, но наша программа не нашла для них ро-независимых

[17] Olekhnovich, I. and Gussin, G.N. (2001) Effects of mutations in the Pseudomonas putida miaA gene: regulation of the trpE and trpGDC operons in P. putida by attenuation. J. Bacteriol. 183, 3256-3260.

терминаторов. Возможно, это объясняется тем, что в этих случаях имеются менее выраженные и стабильные терминаторы и антитерминаторы.

ПРИЛОЖЕНИЯ

Текст диссертация содержит **Приложение 1** с подробными таблицами результатов тестирований, о которых говорилось в Главе 2.

Затем содержит **Приложение 2**, примыкающее к материалу Главы 3, в которое включены таблицы оперонных структур и выравнивания сигналов регуляции биосинтеза аминокислот. А именно, таблица 1 содержит список геномов, в которых алгоритмом LLLM нами искалась аттенюаторная регуляция, с указанием таксономических групп. Таблица 2 содержит предсказанные нами оперонные структуры и регуляции для ILV генов (биосинтез разветвленных аминокислот). Таблица 3 содержит предсказанные оперонные структуры и регуляции для HIS генов (биосинтез гистидина). Таблица 4 содержит предсказанные оперонные структуры и регуляции для THR генов (биосинтез треонина). Таблица 5 содержит предсказанные оперонные структуры и регуляции для *trp*, *pheA* и *pheST* генов (биосинтез триптофана и фенилаланина).

Кроме того, в Приложении 2 приводится рисунок 1, показывающий типичную («классическую») аттенюацию. И рисунок 2, содержащий пути биосинтеза аминокислот у гамма- и альфа-протеобактерий для четырех случаев: (a) ILV (изолейцина, лейцина, валина), (b) HIS (гистидина), (c) THR (треонина), (d) ароматических аминокислот (триптофана, тирозина, фенилаланина).

В Приложении 2 также приводятся важные для нас рисунки, примыкающие к материалу Главы 3: выравнивание предсказанных структур транскрипционной аттенюации биосинтеза разветвленных аминокислот у гамма- и альфа-протеобактерий (рис. 3). Затем выравнивание предсказанных структур транскрипционной аттенюации биосинтеза гистидина у гамма-, альфа-, фирмикут и других бактерий (рис. 4). Выравнивание предсказанных структур транскрипционной аттенюации биосинтеза треонина у гамма-протеобактерий (рис. 5). Выравнивание предсказанных структур транскрипционной аттенюации *trp*, *pheA* and *pheST* оперонов у гамма- и альфа-протеобактерий (рис. 6).

Приложение 3 примыкает к Главе 3 и содержит рисунки всех основных аттенюаторных структур (с указанием в нуклеотидах терминатора, антитерминатора, паузной шпильки, лидерного пептида с регуляторными кодонами и участка остатков урацила, включая указание спарившихся нуклеотидов), которые найдены нами для упомянутых оперонов в классе протеобактерий.

ВЫВОДЫ

1. Предложены новые алгоритмы и соответствующая компьютерная программа для поиска потенциальных структур аттенюаторной регуляции в геномах бактерий. Показана эффективность и надежность этой программы на основе ее детального тестирования на искусственных и биологических последовательностях.

2. Предсказано 108 предполагаемых аттенюаторных структур для оперонов биосинтеза разветвленных и ароматических аминокислот, гистидина и треонина у протеобактерий, у фирмикут и у бактерий из групп Bacteroidetes/Chlorobi и Thermotogales. В некоторых из этих групп аттенюация обнаружена впервые.

3. Получена предположительная функциональная аннотация ряда генов, кодирующих ферменты и находящихся под аттенюаторной регуляцией, а именно:
- гену *ygeA* у *Pasteurella multocida* приписана функция рацемазы разветвленных аминокислот;
- не ортологичные гены *vatB*, *actX2* и *actX3*, соответственно, у *Pasteurella multocida*, *Mannheimia haemolytica*, *Polaribacter filamentus* кодируют ацетилтрансферазы, участвующие в метаболизме гистидина.

4. Показано, что биосинтез изолейцина у Xanthomonadales использует треонин дегидратазу TdcB, в отличие от IlvA у *E. coli*.

5. Предсказано новое семейство гистидиновых транспортеров – ортологов *yuiF* у *B. subtilis* (например, HI0325 у *H. Influenzae*) и два гистидиновых транспортера BC0629 у *B. cereus* (ортолог *yvsH* у *B. subtilis*) из белкового семейства APA и ген у *L. lactis* (ортолог *lysQ* у *E. coli*) из семейства APC.

6. Предсказано, что оперон *his* биосинтеза гистидина регулируется гистидин-зависимыми аттенюаторами у *Bacillus cereus* и *Clostridium difficile*, и в тоже время регулируется гистидиновыми Т-боксами у *Lactococcus lactis* и *Streptococcus mutans*.

7. Показаны следующие особенности аттенюаторной регуляции:
ген *thrA* аспартат киназы/гомосерин дегидрогеназы у Pasteurellales регулируется не только треонинином и изолейцином (как у *E. coli*), но и метионином;
ацетолактат синтаза IlvIH у альфа-протеобактерий имеет регуляторные кодоны лейцина, изолейцина и валина.

ПУБЛИКАЦИИ автора по теме диссертации

1. Горбунов К.Ю., *Любецкая Е.В.*, Любецкий В.А. О двух алгоритмах поиска альтернативной вторичной структуры РНК. Информационные процессы, РАН, том 1, №2, 2001, стр. 178-187.

2. Леонтьев Л.А., *Любецкая Е.В.*, Любецкий В.А. Модифицированный алгоритм поиска альтернативных вторичных структур РНК и результаты счета. Информационные процессы, РАН, 2002, том 2, №1, с. 100-105.

3. *Lyubetsky E.V.*, Lyubetsky V.A. Algorithm for searching alternative secondary RNA structures. Proceedings of the third international conference of bioinformatics of genome regulation and structure, BGRS'2002, Novosibirsk, Russia, July 14-20, 2002, v. 3, p. 15-17.

4. *Любецкая Е.В.*, Леонтьев Л.А., Гельфанд М.С., Любецкий В.А. Поиск альтернативных вторичных структур РНК, регулирующих экспрессию бактериальных генов. Молекулярная биология, том 37, № 5, 2003, с. 834-842.

5. *Любецкая Е.В.*, Леонтьев Л.А., Любецкий В.А. Поиск альтернативных вторичных структур в классе гамма-протеобактерий. Информационные процессы, РАН, том 3, №1, 2003, с. 23-38.

6. *Lyubetskaya E.V.*, Leontiev L.A., Lyubetsky V.A. Algorithm for detecting alternative secondary RNA structures and mass analysis attenuator regulation in proteobacteria, MCCMB'03, 2003, p. 144-145.

РНБ Русский фонд
2007-4
14939

15 МАР 2004