

Поиск альтернативных вторичных структур в классе гамма-протеобактерий

Е.В.Любецкая, Л.А.Леонтьев, В.А.Любецкий

Институт проблем передачи информации РАН, 101447, Москва, Большой Каретный переулок, 19, Россия,
e-mail: lyubetsk@iitp.ru

Поступила в редколлегию 28.12.2002

Аннотация—Приводится алгоритм и результаты счета для одной из фундаментальных задач алгоритмической (или, как иногда говорят, вычислительной) генетики — задачи определения альтернативной вторичной структуры РНК.

1. ВВЕДЕНИЕ

1.1. Биологическая мотивировка задачи. Регуляция многих бактериальных генов осуществляется на уровне трансляции или взаимодействия процессов трансляции и транскрипции. При этом основным регуляторным сигналом является образование вторичной структуры РНК. Часто (хотя и не всегда) в рассматриваемой области могут образоваться две альтернативные структуры РНК, что и служит регуляторным переключателем. Примерами такой регуляции являются аттенуаторы оперонов синтеза аминокислот [6] и регуляторные структуры некоторых оперонов рибосомальных белков кишечной палочки [7].

Стандартный подход при поиске таких регуляторных сигналов состоит в построении вторичных структур для новых геномов на основе уже известных структур (беря их в качестве шаблонов) у хорошо изученных геномов. В частности, этот подход был применен в упомянутых работах. Однако он имеет естественное ограничение: с его помощью можно переносить известные данные о регуляции на новые геномы, однако невозможно изучать новые регуляторные системы. В то же время существуют таксономические группы бактерий, не содержащие хорошо изученных организмов; но и даже для “изученных” геномов нет гарантии, что у них известны все регуляторные системы данного вида (например, даже у кишечной палочки *Escherichia coli* или сенной палочки *Bacillus subtilis*).

Поэтому возникает задача поиска регуляторных структур РНК без использования всяких шаблонов, “с самого начала”. В общем виде она, по-видимому, не решается, так как количество потенциальных структур в РНК даже для относительно коротких фрагментов РНК велико, и неясно, как выделить среди них регуляторные структуры (никаких общих признаков регуляторной структуры в РНК не известно). Однако, можно воспользоваться тем, что многие регуляторные взаимодействия построены на механизме образования альтернативных структур. Есть надежда, что количество фрагментов, содержащих потенциальные альтернативные структуры, как и количество самих этих структур относительно невелико (поскольку требование альтернативности структуры накладывает существенные ограничения на соответствующий участок фрагмента). Отметим, что это — лишь предположение, никаких вероятностных оценок здесь не известно. Поэтому при наличии компьютерной программы, осуществляющей поиск альтернативных структур в РНК, стратегия исследования представляется следующим образом:

- (1) Сканировать межгенные участки и искать потенциальные альтернативные вторичные структуры;
- (2) Если структура настолько “хороша” (включает протяженные сильно и “правильно” перекрывающиеся спирали, шпильки), что образование ее по случайным причинам кажется невозможным, то она

подлежит непосредственной экспериментальной проверке;

(3) Если предсказание альтернативной структуры не так надежно (т.е. неясен точный вид альтернативных спиралей или структура “слаба” в том смысле, что имеет короткие или “слабо регулярные” спирали), то следует анализировать возможность образования аналогичной (в некотором неформальном смысле) структуры перед тем же геном в другом родственном геноме. При этом предполагается, что истинные структуры консервативны даже при изменении последовательности (это предположение выполняется во многих случаях, см. цитированные работы). Этот анализ производится вручную, поскольку нет четко сформулированного критерия консервативности. Для облегчения ручного этапа анализа разрабатываемая компьютерная программа должна выводить набор субоптимальных структур, поскольку используемые критерии оптимальности по необходимости приближительны и “наилучшая” альтернативная структура вовсе не обязана быть биологически правильной.

1.2. Исторические замечания, относящиеся к задаче. О механизме альтернативной регуляции. Предсказание регуляции экспрессии генов является важным разделом компьютерной генетики. Методы сравнительной геномики позволяют предсказывать оперонную структуру и механизмы регуляции экспрессии генов в этих геномах. Результаты последних лет показывают, что в регуляции экспрессии генов большую роль играет вторичная структура РНК, и поэтому важно предсказание такого рода регуляции в разных функциональных системах для секвенированных геномов. Хотя большинство известных механизмов регуляции экспрессии генов основаны на ДНК-белковом взаимодействии, также распространены механизмы регуляции, в которых существенное участие принимает вторичная структура мРНК. Уже известны ряд систем, в которых вторичная структура мРНК принимает участие в регуляции экспрессии генов. Экспериментально изучена регуляция следующих генов: (1) генов биосинтеза аминокислот и генов, кодирующих аминокил-тРНК синтетазы [7]; (2) генов биосинтеза нуклеотидов [8], [9]; (3) катаболических оперонов (утилизация сахаров и аминокислот) [10]; (4) генов биосинтеза витаминов [11], [12]; (5) генов устойчивости к антибиотикам [13]; (6) генов, кодирующих рибосомальные белки [14]; и прочих генов, принадлежащих различным системам. В связи с быстрыми темпами секвенирования возникает задача предсказания регуляции в расшифрованных геномах на основе сравнения с известными регуляторными системами. Методы сравнительной геномики применялись для анализа различных регуляторных систем, в которых вторичная структура РНК играет регуляторную роль [2], [3], [4]. В этой работе рассматривается регуляция экспрессии генов биосинтеза аминокислот в гамма-протеобактериях. В настоящее время нами ведется аналогичная работа для альфа-бактерий, от которой мы представляем некоторые из предварительных результатов.

Механизм регуляции биосинтеза некоторых аминокислот и аминокил-тРНК синтетаз в протеобактериях основан на соотношении скоростей транскрипции и трансляции. Он экспериментально показан для нескольких аминокислотных оперонов (*trp*, *his*, *ilvGMEDA*, *ilvBN*, *phe*, *thr*, *leu*) в *Escherichia coli* и *Salmonella typhimurium* [7]. Роль посредника в регуляции выполняет рибосома. Регуляция осуществляется следующим образом. На рисунке 1 схематически изображены две альтернативные конформации вторичной структуры лидерной (некодирующей) части мРНК, расположенной на 5' конце перед аминокислотными генами.

Цифрами обозначены различные участки мРНК, вовлеченные в процесс формирования вторичной структуры. Первая конформация (обозначена как 2:3) называется антитерминирующей, вторая (1:2 и 3:4) — терминирующей. При укладке вторичной структуры мРНК в антитерминирующую конформацию РНК-полимераза синтезирует весь транскрипт, при укладке в терминирующую конформацию транскрипция преждевременно обрывается, не доходя до структурных генов оперона. Как видно из рисунка, шпилька, называемая антитерминатором, складывается из частей паузной шпильки и терминатора. Показанные на рисунке структуры (антитерминатор и терминатор + паузная шпилька) являются альтернативными. Ранее [6] были предсказаны аттенуаторы транскрипции ароматических аминокислотных оперонов (*trp*, *pheA*, *pheST*) в ряде гамма-протеобактерий. В нашей

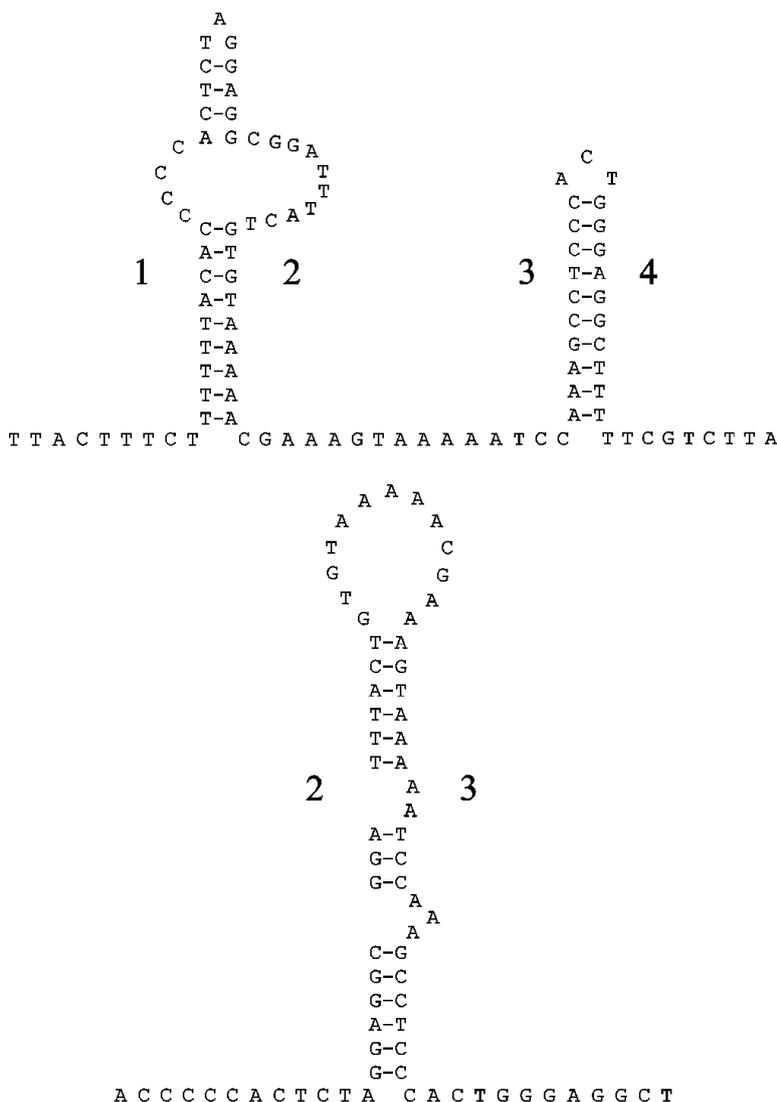


Рис. 1. Вторичная структура в лидерной (не-кодирующей) области *Shewanella putrefaciens* мРНК

работе на основании оригинального алгоритма представлены результаты предсказания регуляции экспрессии аминокислотных оперонов (*his*, *ilv*, *leu* и *thr*) в гамма-протеобактериях.

2. БАЗОВОЕ ОПИСАНИЕ АЛГОРИТМА

Предлагается логически довольно сложный алгоритм, который однако отличается быстротой (примерно квадратичный от размера исходных данных) и эффективностью. Он основан на идеях алгоритма, предложенного в [1], и состоит из нескольких частей.

В **части 1** дана *одна* последовательность в 4-х буквенном алфавите (длины 450 — 600 или больше). Даны параметры алгоритма: максимально возможная длина шпильки, измеряемая числом нуклеотидов от А до Д (обозначаемая максдлина), максимально возможный размер выпячивания или внутренней петли (обозначаемый максвыпячивание), минимально возможная длина одного отрезка (т.е. число подряд склеенных пар нуклеотидов, обозначаемое минподряд), минимально и максимально возможный размер петли спирали (обозначаемые минпетля и макспетля).

В этой части 1 ищется набор **Sp** спиралей мощностей, близких к максимальной, а при включении соответствующего параметра алгоритма — и с некоторыми ограничениями на спирали (например,

нуклеотиды А и Т спариваются с весом 1, а С и G с весом 2, G и T с весом 0, причем G и T не могут быть в началах и концах отрезков; возможны и другие ограничения на спирали такого типа). В этом наборе спирали располагаются по убыванию В, см. часть 1с. Программа может запускаться и сразу на большом файле из многих исходных последовательностей.

В **части 2** набор спиралей Sp статистически обрабатывается нашим алгоритмом следующим образом: составлялся новый набор **Sp1** из всех спиралей набора Sp , у которых значения параметров В и С повторяются строго более p раз (часто выбиралось $p = 9$). В файл **Sp2** пишем по строкам самые длинные спаривания с одной парой выпячиваний длины не более 2, которые возможны в исходной последовательности для параметров В и С из файла **Sp1** (т.е. образуются спирали с 1-2 отрезками).

В версии алгоритма для трехшпильчатых и Т-букс структур из $Sp2$ отбираются в новый файл **TSp** все спирали, удовлетворяющие условию наличия Т-поля (от 5 нуклеотидов Т с одним произвольным пропуском длины 1). Все спирали из файла **TSp**, упорядоченные по убыванию мощности, рассматриваются алгоритмом как кандидаты в терминаторы и выдаются на печать.

Часть 3 алгоритма организована по разному для разных типов структур. Опишем подход, который мы используем для случая трехшпильчатых структур. Он основан на выборе лучшего пути в следующем *дереве спиралей*. Для каждого кандидата t в терминаторы с какими-то параметрами В и С из файла **TSp** берется первая слева пара B_1 и C_1 из файла $Sp1$ и среди всех обрезаний спиралей из $Sp1$ с данным B_1 и C_1 выбираются все те t_l , которые зацепляются с t некоторым “биологически правильным” образом. Эти спирали t_l рассматриваются как кандидаты на антитерминаторы. Для каждой t_l аналогичным образом строится набор спиралей t_2 — кандидатов в паузные спирали. В так полученном дереве спиралей (глубины 3) каждому пути приписывается определенное качество и затем в нем выбирается путь наилучшего качества. Несколько таких путей образуют кандидатов в искомые трехшпильчатые вторичные структуры. Они выдаются алгоритмом на печать.

Опишем подробнее часть 1 и начнем с **части 1а**, в которой вычисляются *качества спиралей*.

Переменная X пробегает по всей последовательности; для каждого X рассматривается отрезок от $H = \max(1, X - (\text{максдлина}))$ до $K = \text{Конец} = \min(\text{длина последов.}, X + \frac{1}{2} (\text{максдлина}))$. Вместо $1/2$ можно взять и другое значение, не большее 1.

Пара $\langle m, C \rangle$, составленная по какой-то спирали, называется (ее) *качеством м-С*; эти пары рассматриваются с поординатным порядком. Пара $\langle \langle m, C \rangle, D \rangle$, составленная по какой-то спирали, называется (ее) *качеством м-С-Д*. Пары $\langle \langle m, C \rangle, D \rangle$ будут рассматриваться с лексикографическим порядком: поординатным по первой координате и обратным по второй координате.

Будут рассматриваться столбцы из попарно не сравнимых пар качеств пар одного или другого типа (по возрастанию m), а также — (индексированные) наборы таких столбцов.

Файл **шпилька** $[X, A, k]$ состоит из качеств m, C, D , расположенных от A до D , где $D \leq k$ и $X \leq V$ (последнее условие может отсутствовать или быть замененным на подобное). Этот файл при $k = \text{Конец}$ обозначается $M[X, A]$. Элементы обоих файлов индексируются i , A восстанавливается по i как $H + i$, где A меняется от H до k .

Файл **продолж** $[X, A, k]$ состоит из столбцов, индексируемых j , каждый из которых состоит из качеств m, C ; где D восстанавливается по k как $D = k - \text{максвыпяч} + j$. A именно, каждый столбец состоит из качеств m, C *шпилек* от A до D , где D пробегает отрезок $[k - \text{максвыпяч}, k]$. Его длина $\text{максвыпяч} + 1$. Этот файл по k рассматривается далее как очередь, т.е. при новом k образующийся столбец кладут в конец, выбрасывая 1-й столбец. Здесь j от 0 до максвыпяч . Этот файл состоит из спиралей, которые могут быть продолжены склейкой A с $k + 1$.

Файл **предшпилька** $[X, A, k]$ — это последовательность длины $\text{минподряд} - 1$. Для наглядности будем считать, что $\text{минподряд} = 3$ и тогда в этом файле всегда 2 элемента, каждый из которых — (в нашем случае) содержит качества m, C соответственно: 1-й элемент (=столбец) состоит из качеств $\langle m, C \rangle$ *предшпильек* от A до k кратности 1, а 2-й столбец — из аналогичных качеств предшпильек

кратности 2. *Предшпилька* определяется как объект, отличающийся от шпильки только в начале первого ее отрезка (левого плеча, и соответственно — в конце последнего отрезка ее правого плеча); а именно, в этих отрезках может быть по одному нуклеотиду (тогда говорят: кратности 1) или по 2 нуклеотида (тогда говорят: кратности 2).

1. (Начальный шаг по k , т.е. $k=X$.)

Пусть $k=X$ (заполняется для всех A от H до k). $M[X,A,k]=$ пусто, $\text{продолж}[X,A,k]=$ пусто, $\text{предшп}[X,A,k]=$ пусто (так как нуклеотид сам с собой не может склеиваться).

Основной цикл по k от $X+1$ до Конец, т.е. в целом k меняется от X до Конец.

(Индуктивный шаг от (k) к $(k+1)$.)

Цикл по A от H до $k+1$.

Случай: A не комплементарно $k+1$.

$\text{продолж}[X,A,k+1]=\{$ в $\text{продолж}[X,A,k]$ 1-й элемент выбрасываем, остальные сдвигаем к началу, последний вектор пустой $\}$;

$\text{шпилька}[X,A,k+1]=M[X,A,k]$;

$\text{предшп}[X,A,k+1]=$ пусто.

Основной случай: A комплементарно с $k+1$.

1) $\text{продолж}[X,A,k+1]=$

$\{$ продолж $[X,A,k]_0$ удаляем, остальные подвигаем к началу; и добавляем:

$\text{продолж}[X,A,k+1]_{\text{максвып}} = \text{Слияние}\{\text{предшп}[A+1,k]_2, \text{продолж}[A+1,k]_{\text{максвып}}\}$ и ко всем m прибавляем $+1$;

2) $\text{шпилька}[X,A,k+1]=$

$\text{Слияние}\{M[X,A,k], \{\text{предшп}[A+1,k]_2, \text{продолж}[A+1,k]_{\text{последний}}\}$ с приписанным $k+1$ в качестве $D\}$;

3) $\text{предшп}[X,A,k+1]_2 = \text{предшп}[X,A+1,k]_1$ и k мощности m прибавляем $+1$; затем вычисляем:

$\text{предшил}[X,A,k+1]_1 = \{$ Если разность $((k+1)-A)$ между параметрами минпетля и макспетля, и, кроме того, $A \geq X$ (т.е. A выступает в роли B или, если без последнего условия, то в роли конца “или какой-то еще точки”), то: первой парой столбца пишем $(1,k+1)$ ($k+1$ в роли C). Иначе — ничего не пишем, затем в этот же столбец (в любом случае) включаем:

$\text{Слияние}\{\text{продолж}[X,A+1,k], \dots, \text{продолж}[X,A+1+\text{максвып}ч,k]\}$.

На выходе этой части 1 для каждого X получается массив $M[X,A,k]$, где $k=\text{Конец}$ (это некоторый набор столбцов). В нем возьмем последний (нумерация по A) столбец $M[X,X,\text{Конец}]$, где $X=A$, а в нем качество с наибольшим значением m (это последняя строка этого столбца по возрастанию m). Эти качества образуют массив четверок M,A,C,D , которые обрабатываются в следующей части 1б.

В части 1б по данным параметрам M,A,C,D образуется параметр B с помощью индуктивного заполнения двух массивов **полн** и **неполн**.

Члены массива $M[X,A]$ (A всегда имеет смысл начала спирали) удобнее представлять себе как четверки $\langle M,A,C,D \rangle$. Идет цикл по элементам этого массива. Далее рассматривается один шаг этого цикла для *фиксированной пятерки* $\langle X,M,A,C,D \rangle$. Цель этой части — для каждой пятерки найти соответствующее B (так, чтобы шестерка $\langle X,M,A,B,C,D \rangle$ описывала некоторую одну спираль с этими параметрами).

Далее участвуют массивы: **полн** $[A,C,D;B,c]$ и **неполн** $[A,C,D;B,c]$ (где B , *возрастая, меняется от A до $(C-\text{минпетля})$, а c , убывая, меняется от D до C).*

Значения m в **полн** $[B,c]$ — это максимальная мощность *недошпильки с параметрами A, B, c, D* или символ “—бесконечн”, если нет ни одной такой шпильки.

Недошпилька — это обычный объект без верхнего ограничения на длину внешней петли.

Файл $\text{неполн}[B,c]$ — это последовательность длины (минподряд–1) (в нашем случае равной 2), в которой 1-й член — максимальная мощность *преднедошпилки* с параметрами A, B, c, D *кратности* 1 или аналогично (–бесконечн), а 2-й член — аналогичная величина *кратности* 2.

Преднедошпилка — это обычный объект без верхнего ограничения на длину внешней петли, а также длина последнего отрезка (в левом плече и соответственно первого отрезка в правом плече) может быть равна 1 или 2.

Массив Полн *заполняется до такого номера В строки*, что $C - B$ меньше макспетля и $\text{полн}[B,C]=M$ (здесь $c=C$). Это и есть искомое B (его нахождение — первая цель Части 1б).

Цикл по B. Начальный шаг, когда $B=A$.

$\text{Полн}[A,c]$ (здесь мощности *недошпилек*) равно символу (–бескон) для всех c (за счет условия, что минподряд строго больше 1).

$\text{неполн}[A,c]$ (если c не равно D) (здесь мощности *преднедошпилек*) = (–бескон, –бескон) (по всем таким c). В ином случае $\text{неполн}[A,D]=(1, -\text{бескон})$.

2. Индуктивный шаг по B (к $B+1$).

Если $B+1$ не комплементарно c, то $\text{Полн}[B+1,c]=\text{неполн}[B+1,c]_{1,2} = -\text{бескон}$.

Если $B+1$ комплементарно c, то

$\text{полн}[B+1,c] = 1 + (\text{макс} \{ \text{неполн}[B,c+1]_2, \text{полн}[B,c+1] \})$;

$\text{неполн}[B+1,c]_2 = (\text{неполн}[B,c+1]_1) + 1$;

$\text{неполн}[B+1,c]_1 = 1 + (\text{макс} \{ \text{полн}[i,j] \})$, где i от (макс $\{A, B - \text{максвып}\}$) до B и j от $c+1$ до $\text{минс} + 1 + \text{максвып}, D$. (Конец индуктивного шага.)

После этого искомое B определяется как номер последней заполненной строки массивов Полн и Неполн.

В **части 1с** по каждой такой пятерке данным $\langle X, M, A, B, C, D \rangle$ и соответствующим ей файлам *полн* и *неполн* вычисляется одна соответствующая этим данным спираль. Эти спирали и образуют матрицу Sp (располагаясь по ее строкам), строки индексируются переменной $X=A$, затем матрица переупорядочивается по убыванию B. Строки этой матрицы вычисляются независимо друг от друга. Приведем индукцию для вычисления какой-то одной строки S как последовательности пар нуклеотидов. Файл S заполняется индукцией по номеру места в нем.

Начальный шаг: в S на 0-ое место заносится (B,C). Полагаем $b=B, c=C$.

Полагаем $i = 1$.

(Начало цикла по i ; индуктивный шаг.)

Если выполнено $M > i$ (иначе выходим из цикла по i), то

$b=B-i, c=C+i$.

Если $\text{полн}[b,c] = M-i$ или $\text{неполн}[b,c]_2 = M-i$, то:

полагаем $S = \{(B,C), \dots, (b,c)\}$, т.е. заносим в S еще одну пару (b,c) и полагаем $i = i + 1$;

Если $\text{полн}[b,c] = M-i$ или $\text{неполн}[b,c]_2 = M-i$, или $\text{неполн}[b,c]_1 = M-i$, то полагаем $S = \{(B,C), \dots, (b,c)\}$, т.е. заносим в S еще одну пару (b,c) и также $i = i + 1$.

Иначе: ищем начало нового отрезка, т.е. последовательно перебираем пары $b_1=b-1, b-2, \dots$ и $c_1=c+1, c+2, \dots$ до тех пор, пока не выполнится условие $\text{полн}[b_1,c_1]=M-i$. Тогда полагаем $b=b_1, c=c_1$ и $S = \{(B,C), \dots, (b,c)\}$.

3. ДАННЫЕ И РЕЗУЛЬТАТЫ

Рассматривался класс гамма-протобактерий (10 организмов, которые перечислены в таблице 1) и 99 оперонов (*trpEDCBA, pheA, pheST, hisGDCBHAFI, ilvBN, ilvGMEDA, leuABCD, thrABC*), которые

были построены как ортологичные к соответствующим восьми оперонам в *E. coli* (названия оперонов, естественно, те же, как у *E. coli*) с помощью программы Gene explorer. В одном случае оперон не был найден, см. таблицу 1. Регуляторные области, содержащие поле T, были найдены у 72 оперонов — на регуляторных областях этих оперонов запускался алгоритм.

В таблице 1 наших результатов, помещенной ниже, используются следующие обозначения.

** — биологически правдоподобные структуры с лидерными пептидами, найденные нашим алгоритмом и известные из публикации [6] и книги [7]; таких случаев нашлось 17;

* — кандидаты в регуляторные структуры с лидерными пептидами, предсказанные нашим алгоритмом; таких случаев нашлось 28;

— алгоритм нашел “хорошие” альтернативные структуры (т.е. с мощным G-C богатым терминатором, “хорошим” антитерминатором, “правильно” зацепляющимся с терминатором, и аналогично для паузной шпильки и антитерминатора), которые, однако, не удовлетворяют естественным биологическим требованиям: нет “хорошего лидерного пептида” (т.е. нет поля соответствующих кодонов, нет старт кодона atg в разумной окрестности слева от антитерминатора, длина нуклеотидного слова не соответствует биологическому представлению); или алгоритм не обнаружил хорошей структуры (в этих случаях иногда нами находился хороший лидерный пептид); или, наконец, алгоритм не запускался, так как соответствующая регуляторная область не содержала T-поля или вообще отсутствовала.

Таким образом, алгоритм работал успешно в 45 случаях из 72 (причем в 28 случаях, когда алгоритм не нашел трехшпильчатой регуляторной структуры в соответствующей регуляторной области, и которые мы классифицируем как неуспешную работу нашего алгоритма, может оказаться, что такая структура действительно отсутствует в какой-то части из них).

E.coli	**trpEDCBA	**pheST	#pheA	**hisGDCBHAFI
	**thrABC	**leuABCD	**ilvGMEDA	**ilvBN
S.typhi	**trpED #trpCBA	**pheST	**pheA	#hisGDCBHAFI
	*thrABC	*leuABCDxx	*ilvGMEDA	*ilvBN
Y.pestis	**trpEGDCBA	#pheST	**pheA1 **pheA2	*hisGDCBAHAFI
	*thrABC	*leuABCD	*ilvGMEDA	*ilvBN
V.cholerae	**trpEGDCBA	#pheST	**pheA	*hisGDCBHAFI
	*thrABC	*leuABCDx	*ilvGMEDA	#ilvBN
H.influenzae	#trpEGDC #trpBA	#pheST	**pheA	*hisGDCBHAFI
	*thrABC	*leuABCD	#ilvGDA	*ilvBN
S.putrefaciens	*trpEGDCBA	#pheST	*pheA	*hisGDCBHAFI
	*thrABC	*leuABCD	#ilvDA	#ilvBN
A.actinomyces temcomitans	#trpEG #trpD	#pheST	**pheA	#hisC
	#trpFC #trpxBA			
	#thrAB #thrCx	No gene	#ilvGE	#ilvBN
P.multocida	#trpEG #trpDC	#pheSTxx	#pheA	*hisG #hisDCBxHAFxA
	#trpBA			
	*thrABCxxxx	*leuABCD	*ilvGMxDA	#ilvBN
K.pneumoniae	#trpEDC		#pheA	#hisD #hisBHA #hisFI
	#trpB			
	#thrABC	#leuA #leuC	*ilvGxEDA	*ilvB
P.aeruginosa	#trpE	#pheST		#hisGDC #hisBH #hisAF
	#thrA #thrC	#leuA #leuB #leuC	#ilvE #ilvD	#ilvBN
		#leuD	#ilvA	

Таблица 1.

Ниже приводятся найденные авторами выравнивания терминаторов (точнее, их левых и правых плеч) для трехшпильчатых (альтернативных) вторичных структур из таблицы 1 — каждый терминатор приводится на пересечении строки-названия организма и столбца-названия оперона.

	ilvGMEDA
Escherichia coli	Aaacccecgacc – ggtccgggggttt
Salmonella typhi	agacccecgacc – gtccgggggttt
Yersinia pestis	gaacccecgacc – ggtccgggggttt
Vibrio cholerae	aaacccecgacc – agtccgggggttt
Shewanella putrefaciens	–
Pasterella multocida	aaccctcgtctg – cagatgagggt
Haemophilus influenzae	
Actinobacillus actinomycetemcomitans	–
Pseudomonas aeruginosa	–
Klebsiella Pneumoniae	aagacccecgacc – ggtccgggggttt
	ilvBN
Escherichia coli	aaacccecgccg – ccggcggggttt
Salmonella typhi	aaacccecgccg – ccggcggggttt
Yersinia pestis	aaacccecgccg – gccgagcggggttt
Vibrio cholerae	–
Shewanella putrefaciens	–
Pasterella multocida	–
Haemophilus influenzae	aatgcccga – tgtcatttt
Actinobacillus actinomycetemcomitans	–
Pseudomonas aeruginosa	–
Klebsiella Pneumoniae	aaacccecgccg – ccgagcggggttt
	ThrABC
Escherichia coli	agaaaaagcccgcac – tgcgggttttttt
Salmonella typhi	agaaaaagcccgcac – gtgcgggttttttt
Yersinia pestis	agaaaaagcccgcac – gtgcgggttttttt
Vibrio cholerae	aaaaagcctgtatc – gatacaggccttttt
Shewanella putrefaciens	gcccgcttc – gaagcgggt
Pasterella multocida	
Haemophilus influenzae	Aaaaccgcgattc – gaatagcgggttt
Actinobacillus actinomycetemcomitans	aaccgcattt – agagcgggt
Pseudomonas aeruginosa	–
Pausterella multocida	Aaaaccgtact – agtgcgggttt
	hisGDCBHAFI
Escherichia coli	Tgagaaagcccccgaag – ctccggggcctttttat
Yersinia pestis	agagagccctcgaaga – ctccgaggcctttt
Vibrio cholerae	accctcggga – ctccgagggt
Shewanella putrefaciens	aaccct_ggaa – ttccgggggttt
Pausterella multocida	Aaccctcggaa – ttcgagagggt
Haemophilus influenzae	Accctcggaa – ttcgagggt
Actinobacillus actinomycetemcomitans	No such gene
Pseudomonas aeruginosa	–
Klebsiella Pneumoniae	–

	leuABCD
Escherichia coli	aaaaaaccegcgc – gcgcgggttttt
Salmonella typhi	Aaacaacaccegcgc – tgtgcgggttttt
Yersinia pestis	cctcegcgc – gccggagg
Vibrio cholerae	aaaaaaccegcgc – tgcgggttttt
Shewanella putrefaciens	aaacccgcgc – ttgcgggtttt
Pausterella multocida	aattagacccgc – gcgggtcttttt
Haemophilus influenzae	aaacccgcactt – aagttgcgggttt
Actinobacillus actinomycetemcomitans	No such gene
Pseudomonas aeruginosa	–
Klebsiella Pneumoniae	–

Таблица 2.

Ниже, в таблице 3 без подробностей приводятся результаты поиска (альтернативной) вторичной структуры Т-боксов у некоторых грам-положительных бактерий. Указаны только терминаторы и качества найденных антитерминаторов, так как имеется полное побуквенное совпадение с известными биологически ответами для терминаторов и высокая степень совпадения для антитерминаторов.

Имена организма и генов	Найденный терминатор	Качество соответствующего антитерминатора
BS SERS	agagctctcgtccc – gggatgagggtctct	+
BE SERS	gctcccgtccc – gggacgggggc	+
HD SERS	gctctcgtccct – agggatgagagc	–
BQ SERS	aactcccgtccct – agggacgggagtt	–
SA SERS	gaccacgtccct – agggatgtggct	–
EF SERS	agtcatttcgtcccaagaat – tattttgggggtgagatggct	+
LO SERS	cttctctg – cggaag	+
LLX SERS	gcgcctaattc – gaattgggcgc	+

Таблица 3.

Ниже приводятся наши предварительные результаты, относящиеся к классу альфа-протеобактерий. Для организмов *S. meliloti*, *M. loti* и *A. tumefaciens* нами выделены опероны *trpEGDCFBA* и для них найдены биологически правдоподобные регуляторные структуры. В таблице 4 приводятся терминаторы и лидерные пептиды этих структур, а также для подтверждения ортологичности — аналогичные данные для оперона *trpEGDCFBA* у организма *E. coli*.

Организм	Терминатор	Лидерные пептиды
<i>S. meliloti</i>	agccgccgga – ttcgaggcggc	<i>atggcaaacacg-cgaacattc gatctggtggtgggc tgcctga</i>
<i>M. loti</i>	gtggccgcaacgga-tccggcgggcca	<i>atgcgttcgaccaagac-catttgggtgggacctgctga</i>
<i>A. tumefaciens</i>	aagccgcc – ggcggctt	<i>atgaatatcgtgtctaa-gaacattgctaactggtggtggag-cagcttttgcggacctga</i>
<i>E. coli</i>	agcccgc – gcgggct	<i>atgaaag-caatttcgtactgaaaggttgggtg gcgcacttctga</i>

Таблица 4.

В регуляторных аттенуаторных структурах гена *trpE* (первого гена оперонов *trpEDCBA*, *trpED* и гомологичных им) организмов, перечисленных ниже в таблице 5, были найдены указанные в этой же таблице достаточно консервативные слова. Слова находятся в левом плече терминатора и повторяются в конце правого плеча антитерминатора (в большинстве случаев — в самом конце антитерминатора). Интересно, что во всех этих словах из перечисленных гамма-протеобактерий присутствует слово *gccccg*, а во всех перечисленных организмах (уже гамма- и альфа-протеобактериях) присутствует слово *gcc*. В приложении 1 после таблицы 6 приводятся таблицы 7-8 гомологичных оперонов для ряда бета-протеобактерий.

Название Организма	Общий отрезок в нуклеотидах
Гамма-протеобактерии	
<i>E.coli</i>	<i>gccccga</i>
<i>Y.pestis</i>	<i>gccccga</i>
<i>Sh. Putrefaciens</i>	<i>acgccccgaaaagacaact</i>
<i>V.cholerae</i>	<i>cgccccga</i>
<i>S.typhi</i>	<i>gccccgc</i>
Альфа-протеобактерии	
<i>S. meliloti</i>	<i>gccgc</i>
<i>A. tumefaciens</i>	<i>gccgc</i>
<i>M. loti</i>	<i>gccgc</i>

Таблица 5.

В приложении 1 приводятся таблицы 6-8 обсуждавшихся выше, в таблицах 1-5 организмов и оперонов.

Сходные результаты были получены для ряда других оперонов и организмов гамма-, альфа- и бета-протеобактерий. Они будут приведены в следующей публикации авторов.

Авторы благодарят М.С. Гельфанда и А.Г. Витрешака за помощь и многочисленные объяснения биологической задачи, обсуждения результатов счета. Визуализация результатов счета была выполнена М.А. Ширшиным, авторы также благодарят его за помощь в программировании и счете.

Приложение 1.

В таблице 6 собрана информация об оперонах и регуляторных структурах генов *trpE*, *pheS* и *pheA* в различных геномах. В ней используются следующие обозначения:

GRS — биологически правдоподобная структура;

GRS? — биологически правдоподобная структура, не содержащая подходящего лидерного пептида;

GRS^{new} — биологически правдоподобная структура, не известная на момент расчета;

GRS*¹ — биологически правдоподобная структура, известная из статьи [6] и найденная алгоритмом;

NT — регуляторная область не содержит полей T;

NU — регуляторная область отсутствует (длина лидерной области меньше 100 нуклеотидов)

BRS — биологически неправдоподобная структура;

NRS — регуляторные структуры не найдены;

— — продолжение оперона;

TrpE(<< 0) — ген TrpE, направленный влево; расстояние до следующего гена оперона — 0.

TrpEG[trpE(>> 15)-trpG] — фузия генов *trpE* и *trpG*;

{trpEG} — конец оперона *trpEG*;

(482) — расстояние между генами; направление не указано, т.к. гены в одном опероне сонаправлены и направление указывается только для первого гена оперона;

gX — неопределенный ген (его функция и название не определяются из соображений ортологичности).

Т	Организм		1		2		3		4		5
	<i>E. coli</i> (1 оперон, 0 фузий)	GRS*1	TrpE(<<0)	-	TrpD(1)	-	TrpC(12)	-	TrpB(0)	-	TrpA {trpEDC-BA}
	<i>Sal. Typhi</i> (2 оперона, 0 фузий)	GRS*1	TrpE(<<0)	-	TrpD {trpED} (412)	BRS	TrpC (<<10)	-	TrpB(0)	-	TrpA {trpC-BA}
	<i>Yers. Pestis</i> (1 оперон, 0 фузий)	GRS*1	trpE(<<0)	-	TrpGD (trpG (14)-trpD)(4)	-	trpC(157)	-	trpB(0)	-	trpA {trpEGD-CBA}
	<i>Vibrio ch.</i> (1 оперон, 1 фузия)	GRS*1	trpE(<<0)	-	trpGD[trpG (24)-trpD](9)	-	trpC(44)	-	trpB(3)	-	trpA {trpEGD-CBA}
	<i>Haemophilus influenzae</i> (2 оперона, 1 фузия)	BRS	trpE(>>13)	-	trpD[trpG (488)-trpD](37)	-	TrpC {trpEGDC}	BRS	TrpB (>>0)	-	TrpA {trpBA}
	<i>Shewanella p.</i> (1 оперон, 0 фузий)	FBRS	PabB(trpE)	-		-		-		-	
	<i>Shewanella p.</i> (1 оперон, 1 фузия)	GRS ^{new}	trpE(<<11)	-	TrpGD [trpG (0)-trpD](6)	-	trpC(120)	-	trpB(6)	-	trpA {trpEGD-CBA}
	AA(4 оперона, 3 фузии)	FBRS	TrpEG [trpE (>>15)-trpG](482) {trpEG}	FBRS	TrpD(>>674){trpD}	NT	TrpFC [trpF (>>19)-trpC] {trpFC}	FBRS	TrpB [gX (<<10)-trpB]	-	TrpA {trpBA}
	<i>Pm</i> (3 оперона, 0 фузий)	FBRS	TrpEG [trpE (<<11)-trpG_1] (1069) {trpEG}	RBRS	TrpD (<<37)	-	TrpC {trpDC} (996)	FBRS	TrpB (<<-13)	-	TrpA {trpBA}
	<i>Kleb. Pn.</i> (2 оперона, 0 фузий)	NT	TrpE(<<979)	-	TrpD(27)	-	TrpC {trpEDC}	NT	TrpB (<<) {trpB}		
	<i>Pseudomonas aeruginosa</i> (1 оперон, 0 фузий)	NT	TrpE {trpE}								

	Actinobacillus actinomyces-temcomitans	GRS?	pheS(<<20)	-	PheT {pheST}
	Pasteurella multocida	GRS?	pheS(<<11)-Pm0630(23)	-	pheT(4)-himA(89)-nlpC {pheST-G-G}
P	Pseudomonas aeruginosa	NT	pheS(29)	-	PheT {pheST}
h	Vibrio cholerae	FNRS	pheS(>>47)	-	PheT {pheST}
e	Shewanella putrefaciens	RBRs	pheS(>>19)	-	PheT {pheST}
S	pseudomonas putida	NT	PheS(>>36)	-	PheT {pheST}
T	Escherichia coli	GRS	pheS(<<15)	-	pheT {pheST}
	Salmonella typhi	GRS	pheS(<<16)	-	pheT {pheST}
	Yersinia pestis	BRS	pheS(<<14)	-	pheT {pheST}
	Vibrio cholerae	BRS	pheS(>>47)	-	pheT {pheST}
	Haemophilus influenzae	BRS	pheS(>>34)	-	pheT {pheST}
D	Escherichia coli	BRS	pheA(>>)		
	Salmonella typhi	GRS* ¹	pheA(>>)		
	Yersinia pestis	GRS* ¹	pheA1(>>) pheA2(>>)		
	Vibrio cholerae	GRS* ¹	pheA(>>)		
	Haemophilus influenzae	GRS* ¹	pheA(>>)		
	Shewanella putrefaciens	GRS ^{new}	PheA(<<)		
	Actinobacillus actinomyces-temcomitans	GRS* ¹	PheA(>>)		
	Pasteurella multocida	GRS?	PheA(>>)		
	Klebsiella pneumoniae	FBRS	PheA(<<)		

Escherichia coli	GRS ^{new}	HisG(>>6)	-	HisD(-3)	-	HisC(-3)	-	HisB(0)	-	HisH(-3)
Escherichia coli	-	HisA(-18)	-	HisF(-6)	-	HisI {hisGD-CBHAFI}				
Salmonella typhi	BRS	HisG (>>107)	-	hisD(1)	-	HisC(1)	-	HisB(4)	-	HisH(4)
Salmonella typhi	-	HisA(-14)	-	HisF(-2)	-	HisI {hisGD-CBHAFI}				
Yersinia pestis	GRS ^{new}	HisG(<<4)	-	HisD(3)	-	HisC(-3)	-	HisB (-1)	-	HisH(-144)
Yersinia pestis	-	HisA(-18)	-	HisF(48)	-	HisI {hisGD-CBHAFI}				
Vibrio cholerae	GRS ^{new}	HisG (>>-3)	-	HisD(5)	-	HisC(5)	-	HisB (-8)	-	HisH(5)
Vibrio cholerae	-	HisA(-98)	-	HisF(-8)	-	HisI {hisGD-CBHAFI}				
Shewanella_ putrefaciens	GRS ^{new}	HisG(<<4)	-	HisD(63)	-	HisC(83)	-	HisB (145)	-	HisH(105)
B Shewanella_ putrefaciens	-	HisA(-18)	-	HisF(83)	-	HisI {hisGD-CBHAFI}				
Pausterella multocida	GRS ^{new}	HisG (>>921) {hisG}	NU	HisD(>>4)	-	HisC(49)	-	HisB (-3)	-	GX(3)
Pausterella multocida	-	HisH(46)	-	HisA(-18)	-	HisF(34)	-	GX(13)	-	HisI {hisD-CBxHAFxI}
Haemophilus influenzae	GRS	HisG (>>93)	-	HisD(69)	-	HisC(112)	-	HisB(66)	-	HisH(36)
Haemophilus influenzae	-	HisA(-18)	-	HisF(0)	-	HisI {hisGD-CBHAFI}				
Actinobacillus actinomycescomitans	NU	HisC								
Pseudomonas aeruginosa	NT	HisG(>> 118)	-	HisD(32)	-	hisC {his-GDC}	NT	HisB (>>-473)	-	HisH(268) {hisBH}
Pseudomonas aeruginosa	BRS	HisA (>>-150)	-	HisF {hisAF}						
Klebsiella Pneumoniae	NT	HisD(<<){hisD}	BRS	HisB(<<-19)	-	550_0005 [hisH-hisA] {hisBHA}	NT	hisF (<<121)	-	HisI {hisFI}

T	Escherichia coli	GRS ^{new}	thrA(>>2)	-	thrB(1)	-	ThrC {thrABC}		
h	Salmonella typhi	GRS ^{new}	thrA(>>6)	-	ThrB(8)	-	ThrC {thrABC}		
r	Yersinia pestis	GRS ^{new}	ThrA(>> -14)	-	ThrB(8)	-	ThrC {thrABC}		
A	Vibrio cholerae	GRS ^{new}	ThrA(<<9)	-	ThrB(0)	-	ThrC {thrABC}		
B	Shewanella putrefaciens	GRS ^{new}	ThrA(<<-3)	-	ThrB(73)	-	ThrC {thrABC}		
C	Haemophilus influenzae	GRS ^{new}	ThrA(<<13)	-	ThrB(43)	-	ThrC {thrABC}		
	Actinobacillus actinomycescomitans	BRS	ThrA(>>17)	-	ThrB {thrAB}	NRS	ThrC(>>63)	-	gX {thrCX}
	Pseudomonas aeruginosa	NT	ThrA(<<) {thrA}	NT	thrC(>>) {thrC}				
	Pausterella multocida	GRS ^{new}	thrA(>>24)	-	thrB(20)	-	ThrC	-	GX - GX
	Pausterella multocida	-	GX	-	GX {thrABCXXXX}				
	Klebsiella Pneumoniae	NRS	thrA[gX(39)-gX(133)-gX](>>384)	-	ThrB(8)	-	ThrC {thrABC}		
	Escherichia coli	GRS ^{new}	leuA(<<-3)	-	leuB(3)	-	leuC(11)	-	LeuD {leuAB-CD}
	Salmonella typhi	GRS ^{new}	LeuA(<<0)	-	LeuB(3)	-	LeuC(11)	-	LeuD (30) - gX(-52)
	Salmonella typhi	-	GX {leuABCDXX}						
	Yersinia pestis	GRS ^{new}	LeuA(<<3)	-	LeuB(-67)	-	LeuC(15)	-	LeuD {leuAB-CD}
B	Vibrio cholerae	GRS ^{new}	LeuA(>>107)	-	LeuB(38)	-	LeuC(5)	-	LeuD (93) - GX {leuABCDX}
	Shewanella putrefaciens	GRS ^{new}	LeuA(>>56)	-	LeuB(4)	-	LeuC(29)	-	LeuD {leuAB-CD}
	Haemophilus influenzae	GRS ^{new}	LeuA(>>86)	-	LeuB(176)	-	LeuC(24)	-	LeuD {leuAB-CD}
	Pseudomonas aeruginosa	NU	LeuA(>>)	NU	LeuB(>>)	NU	LeuC(>>)	NU	LeuD (>>)
	Pausterella multocida	GRS ^{new}	leuA(<<110)	-	leuB(341)	-	leuC(203)	-	LeuD {leuAB-CD}
	Klebsiella Pneumoniae	NRS	LeuA(<<)	BRS	LeuC(<<)				

Таблица 6.

Таблица 7 содержит опероны в *альфа-протеобактериях ортологичные* с соответствующими оперонами у *E.coli* (в опероне *pheA* все гены х ортологичны гену *kdsB*, т.е. он выглядит как *kdsBpheA*).

	S. meliloti	R. palustris	M. loti	A. tumefaciens	B. melitensis
thrABC	thrA thrC	thrA, #thrC	xxthrAxx #thrC	#thrC	thrA,thrB, thrC
trpABCDEFG	GtrpE	BtrpEG, #trpDC, #trpFBA	GtrpEGx, #trpDC, #trpFBA	#trpEG, #trpDC, #trpFBA	#trpDCx, #trpFBA
ilvBN xxilvG	ilvIH <i>or</i> xxilvB1 <i>or</i> xilvB ilvN not found	#ilvB <i>or</i> #ilvB1 <i>or</i> <i>or</i> ilvB- ilvN not found	ilvBH <i>or</i> ilvBgipx ilvN not hisDx. . . ilvN not found	ilvBH <i>or</i> x caib iivG found found	ilvBH ilvN not
pheA pheST	xpheA pheST	#xpheA pheS, pheT	xpheA pheSx- pheTx	xpheA xpheST	<i>not found</i> pheSTx
leuABCD	xleuA1, xleuA2, leuBx, leuCx leuDx	leuA, leuB, leuC, #leuD	leuA, #xxleuB, #xleuC, #leuD	leuA, leuB, leuC, leuD	xxleuA, leuC
hisGDCBHAFI	xxxhisDxx, xhisC2, hisBxHAFE coaA	xhisG, hisD, hisBH(A/F)	hisG, hisD (3), hisC, hisBxHxA xFIcoaA	hisSxhisShisG	not found

Таблица 7.

Таблица 8 содержит опероны в альфа-протеобактериях ортологичные с S. meliloti. Перед названием оперона знак # показывает, что для него апстрим не был найден.

	S. meliloti	R. palustris	M. loti	M. magnetotacticum	A. tumefaciens	B. melitensis
thrABC	xxthrArecJ thrB, thrC	#xthrA, #thrB, #thrC	xxthrAxx #thrC		xthrA, #thrB, #thrC	xxthrA, #thrB, thrC
trpABCDEFG						
IlvCADE- MEIHNB						
pheA						
leuDCBA						
hisGDCBHAFI						

Таблица 8.

СПИСОК ЛИТЕРАТУРЫ

1. Vereschagin N.K., Lyubetsky V.A. Algorithm for determination of alternate secondary RNA structures. *Transaction of research seminar of the logical centre of the philosophy institute of RAS*, 2000, Issue 14, Moscow, Nauka, pp. 99-109.

2. Pesole G., Liuni S., D'Souza M. PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics*, 2000 May, vol. 16, no. 5, pp. 439-450.
3. Vitreschak A.G., Rodionov D.A., Mironov A.A., Gelfand M.S. Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. *Nucleic Acids Res.*, 2002 Jul 15, vol. 30, no. 14, pp. 3141-3151.
4. Rodionov D.A., Vitreschak A.G., Mironov A.A., Gelfand M.S. Comparative genomics of thiamin biosynthesis in prokaryotes: new genes and regulatory mechanisms. *J. Biol. Chem.*, 2002 Oct 9.
5. Vitreshchak A., Bansal A.K., Titov I.I., Gelfand M.S. Computer analysis of regulatory signals in complete bacterial genomes. Translation initiation of ribosomal protein operons. *Biofizika*, 1999 Jul-Aug, vol. 44, no. 4, pp. 601-610 (Russian).
6. Vitreshchak A.G., Gelfand M.S. Computer analysis of signals of regulation in bacterial genomes. Attenuators of operons of aromatic amino acids metabolism. *Mol. Biol. (Mosk)*, 2000 Jul-Aug, vol. 34, no. 4, pp. 545-552 (Russian).
7. Landick R., Yanofsky C., et al. *Escherichia coli* and *Salmonella*. *Cellular and Molecular Biology*, Neidhardt F.C., Ed., Washington DC: ASM Press, 1996, vol. 1, Ch. 81, pp. 1263-1286.
8. Switzer R.L., Turner R.J., Lu Y. Regulation of the *Bacillus subtilis* pyrimidine biosynthetic operon by transcriptional attenuation: control of gene expression by an mRNA-binding protein. *Prog Nucleic Acid Res Mol. Biol.*, 1999, vol. 62, pp. 329-367.
9. Liu J., Turnbough C.L.Jr. Effects of transcriptional start site sequence and position on nucleotide-sensitive selection of alternative start sites at the *pyrC* promoter in *Escherichia coli*. *J. Bacteriol.*, 1994 May, vol. 176, no. 10, pp. 2938-2945.
10. Rutberg B. Antitermination of transcription of catabolic operons. *Mol. Microbiol.*, 1997 Feb, vol. 23, no. 3, pp. 413-421. (Review)
11. Perkins J.B. and Pero J.G. Vitamin biosynthesis. In A.L. Sonenshein, J.A. Hoch, and R. Losick (ed.), *Bacillus subtilis* and its relatives: from genes to cells. *American Society for Microbiology*, 2001, pp. 279-293, Washington DC.
12. Miranda-Rios J., Navarro M., Soberon M. A conserved RNA structure (thibox) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc. Natl. Acad. Sci. USA*, 2001 Aug 14, vol. 98, no. 17, pp. 9736-9741.
13. Stasinopoulos S.J., Farr G.A., Bechhofer D.H. *Bacillus subtilis* tetA(L) gene expression: evidence for regulation by translational reinitiation. *Mol. Microbiol.*, 1998 Dec, vol. 30, no. 5, pp. 923-932.
14. Keener J. and Nomura M. *Escherichia coli* and *Salmonella*. *Cellular and Molecular Biology*, Neidhardt F.C., Ed., Washington DC: ASM Press, 1996. vol. 1, Ch. 90, pp. 1417-1431.
15. Decatur A., McMurry M.T., Kunkel B.N., Losick R. Translation of the mRNA for the sporulation gene *spoIIID* of *Bacillus subtilis* is dependent upon translation of a small upstream open reading frame. *J. Bacteriol.*, 1997 Feb, vol. 179, no. 4, pp. 1324-1328.