Российская Академия наук

Институт проблем передачи информации

На правах рукописи

УДК 519.178; 577.053

Селиверстов Александр Владиславович

АЛГОРИТМ ПОИСКА КЛИКИ В ГРАФЕ, ПРЕДСКАЗАНИЕ РЕГУЛЯТОРНЫХ СТРУКТУР РНК И МОДЕЛИРОВАНИЕ РЕГУЛЯЦИИ БИОСИНТЕЗА ТРИПТОФАНА

05.13.17 – Теоретические основы информатики,

03.00.28 – Биоинформатика

ДИССЕРТАЦИЯ

на соискание учёной степени кандидата физико-математических наук

Научный руководитель:

доктор физ.-мат. наук, профессор В.А. Любецкий

Москва – 2006

ОГЛАВЛЕНИЕ

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В информатике исключительно велико значение направления, которое состоит в поиске быстрых и эффективных алгоритмов, в частности алгоритмов решения комбинаторных задач, включая задачу поиска клики в графе. Столь же велико значение анализа эффективности и, в частности, времени работы (*вычислительной сложности*) предлагаемого алгоритма. Известно, что алгоритмы, имеющие хорошую асимптотическую сложность, часто оказываются не эффективными на входных данных малой длины. Однако в биоинформатике реально возникают входные данные весьма большого размера – графы со многими тысячами вершин. Поэтому оценка асимптотической сложности предлагаемого алгоритма и доказательство её полиномиальности представляет реальный интерес.

Многопроцессорные вычислительные комплексы в принципе позволяют эффективно реализовывать и недетерминированные алгоритмы – это делает обоснованным изучение в связи с биоинформатическими проблемами класса задач, разрешимых за полиномиальное время недетерминированными алгоритмами. Такие задачи *называются NP-задачами*, и класс всех таких задач – *классом NP*.

К настоящему времени доступно более 300 полностью секвенированных прокариотических геномов и десятки эукариотических геномов, и также более 500 не полностью секвенированных геномов. Столь огромный объём информации делает невозможным лабораторные чисто биохимические исследования подавляющего большинства геномов, по крайней мере, со скоростью сопоставимой со скоростью пополнения базы данных геномной информации. Это приводит к необходимости разрабатывать эффективные и быстрые алгоритмы для компьютерного анализа таких баз данных и, в частности, для поиска потенциальных *регуляторных структур РНК, что в рассматриваемом случае сводится к*

задаче поиска клики в графе. Эти регуляторные структуры обеспечивают *регуляцию экспрессии генов*.

Ранее многими авторами, в том числе М.С. Гельфандом, отмечалась возможность сведения к поиску клики в многодольном графе задачи нахождения *консервативных участков* в *наборе* невыравненных лидерных областей перед гомологичными генами родственных видов или перед генами, кодирующими ферменты одного метаболического пути. Эти консервативные участки составляют упомянутые выше *регуляторные структуры* (*сигналы*) – статику *регуляции экспрессии* соответствующих генов. Однако практическое применение этого подхода затруднялось из-за отсутствия эффективных методов поиска клики. Другие методы поиска сигнала рассмотрены в работах А.А. Миронова, П.А. Певзнера, М.С. Уотермена и др.

Альтернативный путь изучения регуляторных структур по одной последовательности РНК, впервые рассмотренный А.А. Мироновым, состоит в моделировании кинетики вторичной структуры РНК. Однако, многие регуляторные системы, включая классическую аттенюаторную регуляцию экспрессии генов, не исследовались подобным образом. Более того, невозможность прямого измерения некоторых параметров ставит нетривиальную обратную задачу: выбор параметров модели, соответствующих наблюдаемым зависимостям. И после уточнения модели – решение вопроса о наличии регуляции в *одной* лидерной области, без множественного выравнивания и поиска сигналов, что представляет собой весьма трудную задачу.

Цели работы. Разработать алгоритмы для поиска клики в многодольном графе, для получения нижней оценки числа клик в графе, исследовать эффективность и вычислительную сложность таких алгоритмов; на основе этих алгоритмов провести массовый поиск регуляторных структур (сигналов) и предложить механизмы регуляции в лидерных областях генов у актинобактерий и хлоропластов; построить

математическую модель классической аттенюаторной регуляции биосинтеза триптофана у бактерий.

Методы исследования. В работе использовались методы комбинаторного анализа, теории графов, теории групп, линейного программирования, вычислительной математики, статистической физики, проверки математических моделей проведением компьютерного эксперимента. Построение модели регуляции опирается на сведения из молекулярной биологии и биологической химии.

Научная новизна. Разработаны алгоритмы для поиска клики в многодольном графе, исследована их вычислительная сложность и предсказаны как новые потенциальные *типы регуляции* экспрессии генов на уровнях трансляции и транскрипции, так и много новых потенциальных *регуляторных структур* перед отдельными генами.

Положения, выносимые на защиту.

Доказано, что существование n-клики в n-дольном графе с двумя вершинами в каждой доле эквивалентно непустоте многогранника, стороны которого вычисляются за полиномиальное от n время. Таким образом, предложен алгоритм поиска клики в указанном многодольном графе с помощью алгоритма линейного программирования. (Этот многогранник называется *многогранником квазиклик* – только часть его вершин соответствует кликам; он отличается от *многогранника клик*, у которого все вершины соответствуют кликам.)

Разработан алгоритм полиномиального времени для решения *неявно* заданной системы однородных линейных уравнений над конечным бимодулем и математически доказана его корректность. Алгоритм позволяет, в частности, оценивать снизу число клик в многодольном графе.

Разработан эвристический алгоритм поиска клики в многодольном графе в общем случае, и на его основе получен алгоритм для поиска сигнала в наборе невыравненных последовательностей – лидерных областей генов.

С помощью этого эвристического алгоритма найдены новые потенциальные сайты связывания белков с мРНК у хлоропластов в 5'-нетранслируемых областях генов *atpF*, *clpP*, *petB* и генов *psaA*, *psbA*, *psbB*, кодирующих белки фотосистем. Предложена гипотеза, объясняющая задержку начала трансляции до завершения сплайсинга у ряда этих генов за счёт специального белок-РНКового связывания.

Потенциальные структуры классической аттенюаторной регуляции предсказаны для: оперонов, кодирующих ферменты биосинтеза триптофана, у *Corynebacterium* и *Streptomyces*; гена *trpS*, кодирующего триптофанил-тРНК синтетазу, у *Streptomyces avermitilis*; генов *leuS*, кодирующих лейцил-тРНК синтетазу, у *Streptomyces*; оперонов *ilv* у многих актинобактерий. Предсказаны у многих актинобактерий: новый потенциальный тип регуляции трансляции гена *leuA*, кодирующего 2-изопропилмалат синтазу, Т-боксовая регуляция *трансляции* гена *ileS*, кодирующего изолейцил-тРНК синтетазу, потенциальная Rho-зависимая аттенюаторная регуляция биосинтеза цистеина.

Разработана модель классической аттенюаторной регуляции, которая позволяет вычислять вид зависимости уровня транскрипции оперонов от концентрации триптофана.

Практическая значимость работы. Работа носит теоретический характер. В то же время, данное исследование представляет интерес, поскольку сравнительный анализ геномов позволяет лучше понять механизмы возникновения устойчивости бактерий к антибиотикам и найти пути создания более эффективных промышленных штаммов. Компьютерный анализ проведён в части регуляции экспрессии перечисленных выше генов. К актинобактериям принадлежат индустриальные продуценты аминокислот (*Corynebacterium glutamicum*, *Corynebacterium efficiens*) и антибиотиков (*Streptomyces* spp.), симбионты человека (*Bifidobacterium longum*, *Propionibacterium acnes*), возбудители опасных инфекционных болезней (*Corynebacterium diphtheriae,*

Mycobacterium spp.). В то же время актинобактерии составляют отдельную филогенетическую группу, и они исследованы гораздо меньше, чем кишечная палочка (представитель протеобактерий) или сенная палочка (представитель фирмикутов).

В случае хлоропластов предложена гипотеза, объясняющая задержку начала трансляции до завершения сплайсинга для многих генов за счёт белок-РНКового связывания, что может представлять интерес для изучения сплайсинга у других организмов и также углубленного изучения процессов фотосинтеза у водорослей и растений на геномном уровне.

Предложенные алгоритмы и программа поиска клики и сигнала могут быть применены для исследования широкого класса задач. С их помощью найдено большое число потенциальных регуляторных сигналов, перечисленных выше.

Аппробация работы. Результаты диссертации неоднократно излагались на семинаре Учебно-Научного центра «Биоинформатика» Института проблем передачи информации РАН, на семинаре «Алгоритмы в геномике» кафедры математической логики и теории алгоритмов механико-математического факультета МГУ им. Ломоносова, на Научном семинаре по биоинформатике Института проблем передачи информации РАН и на следующих четырёх конференциях: шестая международная конференция РАН «Проблемы управления и моделирования в сложных системах» (14-17 июня 2004, Самара); четвёртая международная конференция по биоинформатике, геномной регуляции и структуре генома (25–30 июля 2004, Новосибирск); седьмая международная конференция РАН «Проблемы управления и моделирования в сложных системах» (27 июня – 1 июля 2005, Самара); вторая международная московская конференция по вычислительной молекулярной биологии (18–21 июля 2005, Москва).

Публикации. По теме диссертации опубликовано 18 работ. Все результаты из этих работ, включенные в диссертацию, получены автором.

Структура и объём работы. Работа состоит из введения, трёх глав, заключения и списка литературы. Список литературы содержит 60 наименований. Объём работы составляет 102 страницы, включая 24 таблицы и 8 рисунков.

ВВЕДЕНИЕ

0.1 Обзор алгоритмических результатов, относящихся к диссертационному исследованию

Граф G называется *n-дольным*, если кроме самого графа указано разбиение множества всех его вершин на n множеств («долей») такое, что концы любого ребра в этом графе принадлежат разным долям. Многодольный граф G называется *полным многодольным*, если каждые две его вершины из разных долей соединены ребром. Полный многодольный граф является полным графом, если и только если каждая доля состоит из одной вершины. Полный подграф, содержащий m вершин, называется *m-кликой*.

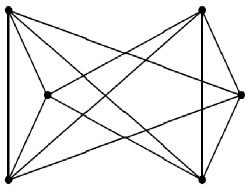

Рис. 1. Полный 3-дольный граф

Рациональный *многогранник* – это ограниченное замкнутое множество точек, выделяемое системой линейных неравенств с рациональными коэффициентами. Многогранник совпадает с выпуклой оболочкой всех его вершин. *Стороной* d-мерного многогранника называется грань размерности d-1. Напомним две теоремы о многогранниках, доказательство которых содержится, например, в книге [Схрейвер, 1991].

Теорема Фаркаша. *Если система линейных неравенств от d переменных неразрешима, то в ней имеется неразрешимая подсистема из не более чем d+1 неравенства.*

Длина двоичной записи положительного целого числа n обозначается Size(n) и равна округлению до большего целого числа величины $\log_2(n+1)$. Длина двоичной записи рационального числа, равного несократимой дроби

n/m, определяется как Size(n/m)=1+Size(|n|)+Size(|m|). Длина двоичной записи рациональной матрицы размера $n \times m$ определяется как Size(M)=nm + Σ Size(M_{ij}).

Теорема Хачияна. *Существует алгоритм для проверки совместности системы рациональных линейных неравенств за полиномиальное от размера записи время. Более того, этот алгоритм выдаёт решение системы, если оно существует.*

Литерал – это пропозициональная переменная или её отрицание. 2-КНФ есть конъюнкция дизъюнкций пар литералов, 3-КНФ – конъюнкция дизъюнкций троек литералов. КНФ позитивная, если каждый её литерал является пропозициональной переменной. Моделью для КНФ называется такая оценка пропозициональных переменных со значениями истина или ложь, при которой КНФ истинна.

Напомним, что класс NP состоит из множеств, распознаваемых недетерминированными алгоритмами за время, ограниченное полиномом от длины входа. Эквивалентно, множество A принадлежит классу NP, если существует такое множество пар B, разрешимое за полиномиальное время, что x принадлежит множеству A тогда и только тогда, когда некоторая пара (x, y), в которой длина записи y ограничена полиномом от длины записи x, принадлежит множеству B.

Множество A называется NP-*трудным*, если для любого множества B из класса NP существует такая функция f, вычислимая за полиномиальное время, что x принадлежит B тогда и только тогда, когда $f(x)$ принадлежит множеству A. Множество A называется NP-*полным*, если оно принадлежит классу NP и является NP-трудным. Известными NP-полными множествами являются множество выполнимых 3-КНФ и множество n-дольных графов, имеющих n-клику, [Сэвидж, 1998].

Теорема Шефера. *Множество позитивных* 3-*КНФ, имеющих такую модель, в которой каждая дизъюнкция содержит ровно один истинный литерал, является* NP-*полным.*

Доказательство. [Schaefer, 1978] Пусть пропозициональная формула $\varphi(\lambda,\mu,\nu)$ истинна тогда и только тогда, когда *ровно* одна из переменных λ, μ или ν истинна, а две другие ложны.

Формула $\lambda \vee \mu \vee \nu$ равновыполнима формуле

$$\varphi(\lambda,\xi_1,\xi_4) \wedge \varphi(\mu,\xi_2,\xi_4) \wedge \varphi(\xi_1,\xi_2,\xi_5) \wedge \varphi(\xi_3,\xi_4,\xi_6) \wedge \varphi(\nu,\xi_3,\xi_7) \wedge \varphi(\xi_7,\xi_7,\xi_8)$$

Формула $\mu \equiv \neg\nu$ равновыполнима формуле $\varphi(\mu,\nu,\xi_1) \wedge \varphi(\xi_1,\xi_1,\xi_2)$.

Для любой 3-КНФ легко построить равновыполнимую коньюнкцию формул вида $\varphi(\lambda,\mu,\nu)$, где λ, μ и ν – пропозициональные переменные. Заменяя в ней подформулы вида $\varphi(\lambda,\mu,\nu)$ на дизъюнкции $\lambda \vee \mu \vee \nu$, получим позитивную 3-КНФ, которая имеет модель, в которой каждая дизъюнкция содержит ровно один истинный литерал, тогда и только тогда, когда исходная 3-КНФ выполнима. Так известная NP-полная проблема выполнимости 3-КНФ сводится за полиномиальное время к задаче распознавания рассматриваемого множества. *Теорема доказана.*

Отметим, что поиск *n*-клики в *n*-дольном графе с двумя вершинами в каждой доле сводится за полиномиальное время к поиску модели пропозициональной конъюнктивной нормальной формы с двумя литералами в каждой дизъюнкции (2-КНФ), которая в свою очередь может быть найдена за полиномиальное время, [Even, Itai, Shamir, 1976].

Задача поиска клики, в свою очередь, тесно связана с проблемой описания сторон многогранника, вершины которого соответствуют кликам полного многодольного графа. Существование алгоритма полиномиального времени для распознавания сторон такого многогранника влечёт самодвойственность класса NP, состоящего из множеств, разрешимых недетерминированными машинами Тьюринга за время, ограниченное полиномом от длины входа. Напомним, что класс coNP состоит из дополнений множеств, входящих в класс NP; «самодвойственность» класса NP означает совпадение классов NP и coNP. Поэтому нахождение

упомянутого даже эвристического алгоритма представляет фундаментальную трудность.

Если в n-дольном графе существует n-клика, то остаётся открытым вопрос о поиске других клик. Нижнюю оценку на число n-клик можно получить, вычислив группу автоморфизмов графа, сохраняющих разбиение множества вершин на доли. Поскольку для каждой перестановки вершин легко проверить, является ли она автоморфизмом графа, мы приходим к задаче о поиске скрытой подгруппы в группе перестановок, то есть такой подгруппы, для проблемы принадлежности к которой имеется распознающий алгоритм полиномиального времени, а требуется найти порождающие и порядок этой подгруппы. Эту задачу можно решить за полиномиальное время алгоритмом Симса [Симс, 1976]. Однако время его работы довольно велико. [Hoffmann, 1982].

В случае, когда граф имеет по две вершины в каждой доле, группа автоморфизмов, сохраняющих доли, изоморфна группе решений системы однородных линейных уравнений с коэффициентами в поле из двух элементов.

Заметим, что поиск подгруппы решений *явно* заданной системы однородных линейных уравнений легко найти методом, являющимся обобщением алгоритма Гаусса, [Боревич, Шафаревич, 1985].

Выделение сигнала в наборе невыравненных последовательностей. Поиск клики в многодольном графе позволяет искать консервативные участки в наборе регуляторных областей. В этой задаче по n данным последовательностям в алфавите {A, C, G, U} строится n-дольный граф G, вершинами которого служат слова из этих последовательностей фиксированной длины. Две вершины соединены ребром в графе G, если они являются словами из разных последовательностей и похожи друг на друга больше некоторого фиксированного порога. Например, они отличаются друг от друга в не более чем фиксированном числе позиций.

Системе попарно похожих слов (по одному слову в каждой из q последовательностей) соответствует q-клика в графе G.

Другим возможным приложением служит поиск кластеров ортологичных генов. В этом случае доля графа соответствует геному, вершины – генам, смежными являются вершины, соответствующие гомологичным генам.

Важным методом анализа регуляции экспрессии генов служит поиск коротких консервативных в большинстве геномов у представителей филогенетческой группы 5'-нетранслируемых участков мРНК. Типичные примеры – поиск сайта связывания белка с РНК и поиск спирали РНК с консервативными плечами.

Существует несколько подходов к поиску по набору нуклеотидных последовательностей сигнала – набора попарно похожих слов одинаковой длины по одному из каждой последовательности, выбранных из данной доли входных последовательностей. В типичной ситуации нас интересует сигнал, включающий сайты из не менее чем 80% входных последовательностей, но заранее не известно какие последовательности не содержат сайтов.

Оптимизационные алгоритмы строят последовательность сигналов, качество которых (то есть значение некоторого функционала) монотонно возрастает. Примером является алгоритм SeSiMCMC [Favorov, Gelfand, Gerasimova, Ravcheev, Mironov, Makeev, 2005] и алгоритм IRSA [Данилова, Горбунов, Гельфанд, Любецкий, 2001]. Комбинаторные алгоритмы, например MITRA [Eskin, Pevzner, 2002], основаны на поиске консенсуса, то есть слова или в общем случае весовой матрицы, который близок к некоторым словам из большинства входных последовательностей.

Задача поиска сигнала тесно связана с задачей множественного выравнивания. В действительности, набор сигналов, состоящих из сайтов маленькой длины можно объединить в общее множественное выравнивание. И наоборот, зная множественное выравнивание легко

выделить наиболее консервативные участки. Однако популярные программы множественного выравнивания, например, CLUSTAL [Thompson, Higgs, Gibson, 1994] нестабильно работают при добавлении невыравниваемых последовательностей, а также при поиске коротких сигналов. Поэтому множественное выравнивание выполнялось для тех участков, на которых обнаружены консервативные слова с помощью алгоритма на основе поиска клики.

Обычно, рассматриваемые участки не являются абсолютно консервативными, но некоторые нуклеотиды встречаются чаще остальных. Эти нуклеотиды часто описывают, используя обозначения, указанные в табл. 1.

Для поиска вторичной структуры РНК, согласованной с множественным выравниванием, и для выравнивания белков автором использовалась программа множественного выравнивания MultAlign, реализованная А.А. Мироновым. А также применялась программа, реализующая алгоритм из работы [Горбунов, Миронов, Любецкий, 2003].

Таблица 1. Алфавиты нуклеотидных и аминокислотных последовательностей.

Нуклеотиды			
G	Гуанин	S	G или C
A	Аденин	W	A или U
C	Цитозин	H	A или C или U
T	Тимин	B	G или U или C
U	Урацил	V	G или C или A
R	Пурин (A или G)	D	G или U или A
Y	Пиримидин (C или U)	N	G или A или U или C
M	A или C	*	Делеция
K	G или U		

Таблица 1. (Продолжение)

Аминокислоты			
F	Фенилаланин Phe	H	Гистидин His
M	Метионин Met	K	Лизин Lys
P	Пролин Pro	C	Цистеин Cys
Y	Тирозин Tyr	G	Глицин Gly
N	Аспарагин Asn	I	Изолейцин Ile
E	Глютаминовая кислота Glu	S	Серин Ser
R	Аргинин Arg	A	Аланин Ala
L	Лейцин Leu	Q	Глютамин Gln
V	Валин Val	D	Аспарагиновая кислота Asp
T	Треонин Thr	W	Триптофан Trp

Качество найденного сигнала оценивается экспертом по совокупности признаков: консервативности участков РНК, согласованности консервативных участков со структурой РНК, наличию характерных особенностей, например, расположению по отношению к открытым рамкам считывания.

Качество выравнивания E двух белков оценивается обычным способом и вычисляется по формуле $E = Kmn\exp\left(-\frac{S}{\lambda}\right)$, где числа m и n равны длинам выравниваемых аминокислотных последовательностей, K и λ – некоторые константы, S – величина, зависящая от относительных частот аминокислотных замен и от частот делеций на выравнивании. При этом вес замены аминокислот вычисляется в соответствии с матрицей BLOSUM62.

0.2 Обзор результатов по регуляции экспрессии генов у хлоропластов и бактерий

Транскрипция происходит от 5’ к 3’ концу возникающей РНК, трансляция от N к C концу соответствующего белка. РНК часто образует сложную вторичную структуру, состоящую из спиралей; каждая спираль состоит из пары участков РНК, которые называются плечами спирали. Длины плеч спирали равны, и *k*-й нуклеотид от 5’-конца левого плеча комплементарен *k*-му нуклеотиду от 3’-конца правого плеча. Линейно вложенные друг в друга спирали образуют *шпильку*.

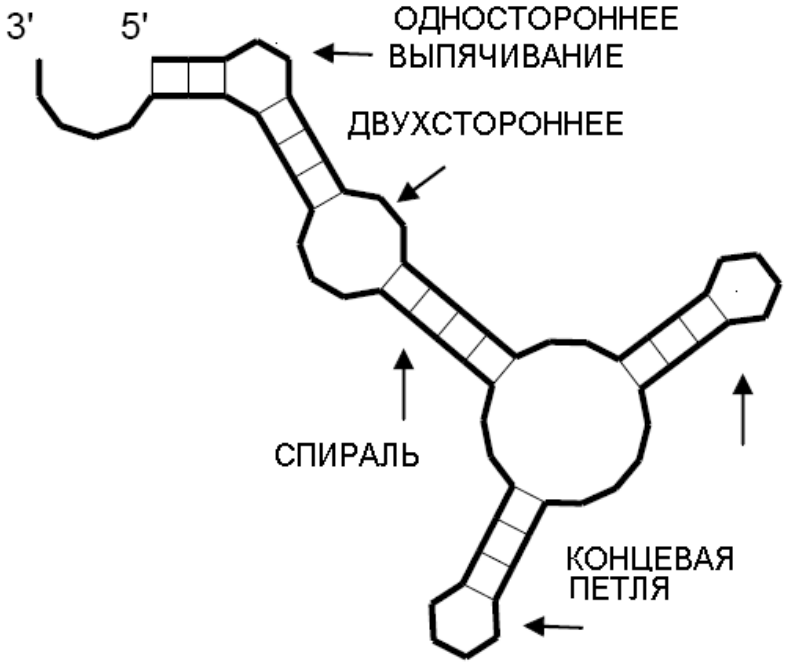

Рис. 2. Вторичная структура РНК без псевдоузла.

Две спирали, у которых плечо одной лежит между плечами другой, образуют *псевдоузел*.

Образование вторичной структуры РНК может прерывать транскрипцию (в этом случае обычно образуется длинная спираль, к которой примыкает U-богатый участок) или препятствовать инициации трансляции. В этом случае спирали перекрывают сайт связывания

рибосомы с РНК вблизи инициирующего кодона трансляции. Этот сайт часто называют областью Шайна-Дальгарно. Консенсусом для неё служит слово GGAGGA. Сайт связывания рибосомы мало консервативен и не всегда может быть точно предсказан по одной последовательности, но обычно выделяется на множественном выравнивании ортологичных генов.

На образование вторичной структуры может влиять взаимодействие РНК с белками, трансляция лидерных пептидов (классическая аттенюация), взаимодействие матричной РНК с транспортной РНК (Т-бокс) и другие факторы. Ниже предсказаны как новые случаи ранее известных механизмов регуляции, так и совершенно новые механизмы, предсказанные теоретически путём поиска высоко консервативных структур. Перечислим основные регуляторные структуры, рассмотренные во второй главе.

Белок-мРНК взаимодействие в хлоропластах. Экспрессия многих генов хлоропластов водорослей и растений регулируется белками, кодируемыми ядерной ДНК, которые связывают мРНК хлоропластов, [Nickelsen, 2003]. Эти белки влияют на редактирование (editing) и инициацию трансляции мРНК. Детальные экспериментальные исследования по поиску соответствующих сайтов известны для одной водоросли *Chlamydomonas reinhardtii* [Hauser, Gillham, Boynton, 1996] и небольшого числа растений [Nickelsen, 2003] и [Zerges, 2000].

Классическая аттенюаторная регуляция. Для этого типа регуляции транскрипции характерными признаками являются короткий лидерный пептид, содержащий кодоны тех аминокислот, для которых концентрация соответствующих загруженых тРНК влияет на уровень транскрипции, терминатор транскрипции. В зависимости от скорости трансляции лидерного пептида, происходит либо терминация транскрипции после транскрипции короткого фрагмента, либо транскрипция всей мРНК. При классической аттенюации терминатор представляет собой шпильку с U-богатым участком (см. главу 3), конкурирующую со шпилькой антитерминатора. Этот механизм детально описан в [Сингер, Берг, 1998].

Ранее экспериментально показана классическая регуляция генов синтеза триптофана у двух актинобактерий: у *Corynebacterium glutamicum* [Heery, Dunican, 1993] и у *Stryptomyces venezuelae* [Lin, Pradkar, Vining, 1998]. Предсказано несколько новых аттенюаторов.

При другом механизме рибосома непосредственно перекрывает сайт связывания белка Rho. В этом случае терминация не связана с образованием шпильки. Такой механизм регуляции транскрипции для гена катаболизма триптофана подробно описан в статьях [Konan, Yanofsky, 2000] и [Gong, Yanofsky, 2003]. Для оперонов синтеза цистеина аналогичный механизм предсказан впервые.

Регуляторная структура с участием Т-бокса. Хорошо известен механизм регуляции транскрипции с участием тРНК. [Grundy, Henkin, 2003]. Незагруженная тРНК стабилизирует структуру мРНК, которая включает терминатор транскрипции. Таким образом, уровень экспрессии зависит от концентрации загруженных тРНК. Ниже впервые предсказаны Т-боксы, связанные с регуляцией трансляции.

Рибопереключатели. Регуляция, как транскрипции, так и трансляции часто связана с образованием характерной структуры РНК, которая стабилизируется небольшой молекулой-лигандом. [Rodionov, Vitreschak, Mironov, Gelfand, 2003], [Mandal, Breaker, 2004], [Vitreschak, Rodionov, Mironov, Gelfand, 2004].

0.3 Обзор результатов по моделированию кинетики образования вторичной структуры РНК

Моделирование классической аттенюации позволяет предсказывать эффективность регуляции транскрипции по одной нуклеотидной последовательности. Важно, что при этом можно предсказать не только наличие регуляции, но и получить количественные оценки. Такой механизм регуляции у кишечной палочки хорошо известен [Сингер, Берг, 1998]. История исследований классической аттенюаторной регуляции и результаты массового поиска такой регуляции у протеобактерий изложена в [Vitreschak, Lyubetskaya, Shirshin, Gelfand, Lyubetsky, 2004].

В целом биоинформатические работы в этой области можно разделить на *систематические исследования* по поиску регуляции и на *немногочисленные попытки* моделирования этого процесса или составляющих его частей.

В работах [Миронов, Кистер, 1985], [Миронов, Кистер, 1989], [Mironov, Lebedev, 1993] и [Danilova, Pervouchine, Favorov, Mironov, 2006] рассматривается моделирование методом Монте-Карло кинетики сворачивания вторичной структуры РНК на уровне микросостояний и поставлена задача моделирования этого процесса на уровне макросостояний. В работе [Xayaphoummine, Bucher, Isambert, 2005] метод вероятностного моделирования Монте-Карло применяется для изучения процесса формирования псевдоузлов у вторичной структуры РНК. В них предлагается оригинальный прием для ускорения процедуры Монте-Карло, который позволяет исключить повторение пройденных состояний марковской цепи. В модели используется другая, но также исключающая повторения и быстрая организация процедуры Монте-Карло. В работе [Elf, Ehrenberg, 2005] вероятность антитерминации вычисляется по явной формуле: как сумма двух слагаемых: первое из них – вероятность того, что рибосома находится на одном из регуляторных кодонов и происходит

формирование антитерминатора в то время, как полимераза доходит до U-богатого участка, а второе из них – умноженная на 0.5 вероятность того, что рибосома покинет стоп-кодон, когда антитерминатор ещё не сформировался. Коэффициент 0.5 мотивируется тем, что в такой ситуации с вероятностью 0.5 формируется что-то одно – терминатор или антитерминатор.

ГЛАВА 1. Алгоритм поиска клики в многодольном графе

1.1 Алгоритм поиска клики в случае двух вершин в каждой доле и доказательство корректности его работы за полиномиальное время

Доказано, что существование n-клики в n-дольном графе с двумя вершинами в каждой доле эквивалентно непустоте многогранника, стороны которого можно вычислить за полиномиальное от n время. Более того, размерность этого многогранника позволяет получить оценку числа n-клик.

Многогранник, включающий клики данного графа. Ниже индексы p, q, r равны 1 или 2, а индексы i, j, k пробегают значения от 1 до n. Для целого числа n определим многогранник *квазиклик* P_n в $4n^2$-мерном пространстве, выделяемый следующей системой равенств и неравенств:

(1) для всех i, j, p, q $X_{ijpq} = X_{jiqp}$,

(2) для всех i $X_{ii11} + X_{ii22} = 1$,

(3) для всех i $X_{ii12} = 0$,

(4) для всех i, j, p $X_{ijp1} + X_{ijp2} = X_{iipp}$,

(5) для всех i, j, p, q X_{ijpq} неотрицательно и не превосходит единицу,

(6) для всех i, j, k, p, q, r сумма $X_{iipp} + X_{jkqr}$ больше либо равна сумме $X_{ijpq} + X_{ikpr}$.

В этом пространстве координаты соответствуют весам, приписанным вершинам и рёбрам графа, см. рис. 3. Координата вида X_{iipp} соответствует p-й вершине i-й доли графа, координата вида X_{ijpq} соответствует ребру, соединяющему p-ю вершину i-й доли и q-ю вершину j-й доли графа. Вес равен нулю, если соответствующее ребро или вершина отсутствует. Первое равенство соответствует неориентированности графа. Третье равенство соответствует несмежности вершин из одной доли. Второе равенство и пятое неравенство означают, что n-клика n-дольного графа содержит по одной вершине из каждой доли.

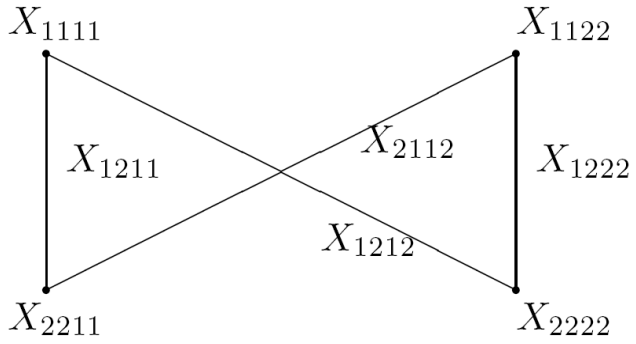

Рис. 3. Веса вершин и рёбер графа

Лемма 1. *Для любого отображения f сегмента* {1, 2, ..., *n*} *во множество* {1, 2} *точка X с координатами*

X_{ijpq}=1, *если* $p=f(i)$ *и* $q=f(j)$,

X_{ijpq}=0, *иначе,*

является вершиной многогранника P_n. *Заметим, что такие точки соответствуют n-кликам полного n-дольного графа с двумя вершинами в каждой доле*.

Доказательство. Координаты точки X равны нулю или единице. При этом

$X_{ijpq} = 1$ тогда и только тогда, когда $X_{iipp} = 1$ и $X_{jjqq} = 1$.

Если $i > j$ и $X_{ijpq} = 0$, то $X_{iipp} = 0$ или $X_{jjqq} = 0$.

Равенства (1) и (3) очевидны. Равенства (2) следуют из того, что в каждой сумме одно слагаемое равно единице, а другое равно нулю.

Равенства (4) следуют из того, что в каждой сумме не более одного слагаемого равно единице, а другие равны нулю.

Проверим неравенство (6). Если $X_{ijpq}=X_{ikpr}=1$, то обе части неравенства (6) равны двум. Если среди чисел X_{ijpq}, X_{ikpr} одно равно единице, то левая и правая части равны единице. Если координаты $X_{ijpq}=X_{ikpr}=0$, то правая часть неравенства (6) равна нулю и не превосходит левую.

Итак, точка X принадлежит многограннику P_n. В точке X одно из каждой пары неравенств (5) обращается в равенство. Поэтому X является единственной точкой пересечения некоторых сторон многогранника P_n. Следовательно, точка X является вершиной. *Лемма доказана.*

n-Дольному графу G, имеющему по две вершины в каждой доле, сопоставим аффинное подпространство H(G), выделяемое всеми уравнениями $X_{ijpq}=0$, где индекс i не равен индексу j и p-я вершина i-й доли не соединена ребром с q-й вершиной j-й доли.

Точке X многогранника P_n сопоставим n-дольный граф $g(X)$, имеющий две вершины в каждой доле, у которого p-я вершина i-й доли соединена ребром с q-й вершиной j-й доли, где $j>i$, если $X_{ijpq}>0$. Очевидно, если точка X принадлежит пересечению многогранника P_n с подпространством $H(G)$, то граф $g(X)$ является подграфом графа G.

Удобно считать, что положительная координата точки X многогранника P_n равна весу вершины или ребра графа $g(X)$.

Лемма 2. *Если точка X принадлежит многограннику P_n, то граф $g(X)$ содержит n-клику. Более того, эта n-клика может быть явно описана за время, ограниченное полиномом от числа n.*

Доказательство. Пусть точка X принадлежит многограннику P_n. Будем шаг за шагом выделять в каждой из долей графа $g(X)$ некоторую вершину так, что в конечном итоге они образуют n-клику. При этом удобно рассматривать такую нумерацию вершин внутри доли, чтобы в клике оказались первые вершины каждой доли. Пусть для любого индекса i координата X_{ii11} не меньше координаты X_{ii22}.

Согласно (2), множество индексов долей $\{1,\ldots, n\}$ равно объединению двух множеств

$S_0 = \{i \mid X_{ii11}>1/2\}$

$S_1 = \{i \mid X_{ii11}=X_{ii22}=1/2\}$

Из равенств (4) следует, что для любых индекса i из множества S_0 и любого индекса j не равного i выполнено

$X_{ij11}=X_{ii11}-X_{ij12}>0$.

Более того, если индекс j принадлежит множеству S_1, то

$X_{ij12}=X_{ii11}-X_{ij11}>0$.

Поэтому первая вершина в каждой доле из S_0 связана ребром с первой вершиной в каждой другой доле из S_0 и с каждой вершиной в любой доле из S_1. Если множество S_1 содержит не более одного элемента, то искомый полный подграф включает первую вершину в каждой доле.

Если для индексов i и p число $X_{iipp}>0$, то p-я вершина i-й доли в графе $g(X)$ соединена рёбрами с некоторой вершиной каждой другой доли в силу (4). Покажем, что в графе $g(X)$ есть такая клика, что каждая доля из S_1 содержит вершину, принадлежащую этой клике. Если множество $S_1=\{i, j\}$ содержит два элемента, то найдутся такие индексы p и q, что $X_{ijpq}>0$.

Пусть множество S_1 содержит больше двух элементов. Из равенств (4) следует, что для любых индексов i, j из S_1 для любого индекса p существует такой индекс q, что X_{ijpq} не меньше 1/4.

Фиксируем элемент a из множества S_1. Можно считать, что для всех индексов j из $S_1\backslash\{a\}$ X_{aj11} не меньше X_{aj12}. Множество $S_1\backslash\{a\}$ является объединением двух непересекающихся множеств

$S_{10}=\{j \text{ из } S_1\backslash\{a\} \mid X_{aj11}>1/4\}$

$S_{11}=\{j \text{ из } S_1\backslash\{a\} \mid X_{aj11}=X_{aj12}=1/4\}$

Из неравенств (6) следует, что для всех индексов i, j, k из S_1 и для всех индексов p, q, r сумма $X_{ijpq}+X_{ikpr}$ не превосходит сумму $1/2+X_{jkqr}$.

Для любых разных индексов i и j из S_{10} первые вершины в a-й, i-й и j-й долях соединены рёбрами. Более того, для любых индексов i из S_{10} и j из $S_1\backslash\{a, i\}$ первые вершины в a-й, i-й долях соединены рёбрами с каждой вершиной j-й доли. Если множество S_{11} пусто или содержит единственный элемент, то искомая клика включает первые вершины каждой доли. Поэтому остаётся построить клику, вершины которой лежат в долях с номерами из множества S_{11}, когда оно имеет не меньше двух элементов. Для этого фиксируем элемент b из S_{11} и разбиваем множество $S_{11}\backslash\{b\}$ в

объединение двух непересекающихся подмножеств S_{110} и S_{111}. Повторяя этот процесс, мы либо построим n-клику в графе $g(X)$, либо за не более чем n шагов придём к задаче построения клики, вершины которой лежат в долях с номерами из такого множества S, что для любых индексов i, j из S и любых индексов p, q координата $X_{ijpq}=1/4$. Очевидно, что можно выбрать любую вершину в каждой доле из множества S. *Лемма доказана.*

Теорема 1. *Пусть n-дольный граф G имеет по две вершины в каждой доле.*

Если пересечение I многогранника P_n и пространства $H(G)$ не пустое, то граф G имеет хотя бы одну n-клику.

Если размерность пересечения I многогранника P_n и пространства $H(G)$ равна единице, то граф G имеет либо одну, либо две n-клики.

Доказательство. Из лемм 1 и 2 следует, что граф G содержит n-клику тогда и только тогда, когда пересечение многогранника P_n и пространства $H(G)$ непустое. Более того, если найдена точка X в пересечении многогранника P_n с пространством $H(G)$, то n-клика определяется эффективно по лемме 2. Многогранник P_n задан системой из $O(n^3)$ равенств и неравенств. Поиск n-клики сводится к задаче линейного программирования и требует лишь полиномиального времени.

С другой стороны, n-клики соответствуют вершинам многогранника P_n. Если граф G включает три n-клики, то пересечение I содержит три вершины многогранника. Поскольку никакая из вершин не является выпуклой комбинацией двух других, то размерность пересечения I не меньше двух. Противоречие доказывает, что существует не больше двух n-клик. *Теорема доказана.*

Теорема 2. *Размерность многогранника P_n равна $n(n+1)/2$.*

Доказательство. Поскольку координаты вида X_{ij11} однозначно определяют точку многогранника P_n, то размерность не превосходит $n(n+1)/2$.

С другой стороны, среди вершин, соответствующих n-кликам полного n-дольного графа, существует $n(n+1)/2+1$ аффинно независимая точка. Это точка с координатами $X_{ii11}=0$, n точек, у которых среди координат X_{ii11} одна единица, а остальные равны нулю, и $n(n-1)/2$ точек, у которых среди координат X_{ii11} две единицы, а остальные равны нулю. Поэтому размерность многогранника P_n равна $n(n+1)/2$. *Теорема доказана.*

Обозначим Q_n выпуклую оболочку точек, соответствующих n-кликам полного n-дольного графа. Этот многогранник ***называется многогранником клик***.

Теорема 3. *Выпуклая оболочка любых двух вершин многогранника Q_n является его ребром.*

Доказательство. Рассмотрим две вершины X и Y многогранника Q_n. Они соответствуют паре n-клик, также обозначенных через X и Y. Отметим в каждой из двух клик X и Y по одному ребру, которое не принадлежит другой клике. Рассмотрим линейную форму

$$F(X)=\sum_{i>j} a_{ijpq}X_{ijpq}$$

где коэффициенты a_{ijpq} определены следующим образом. Рассмотрим ребро графа между p-й вершиной i-й доли и q-й вершиной j-й доли. Тогда

$a_{ijpq}=2$, если рассматриваемое ребро не принадлежит ни клике X ни клике Y,

$a_{ijpq}=1$, если рассматриваемое ребро отмечено,

$a_{ijpq}=0$ во всех остальных случаях.

Равенства $F(X)=1$ и $F(Y)=1$ выполнены на вершинах X и Y многогранника, а для любой другой вершины Z значение формы $F(Z)$ не меньше двух. Следовательно, вершины X и Y являются концами ребра многогранника Q_n. *Теорема доказана.*

Многогранник Q_2 имеет четыре вершины и размерность три. Поэтому он является симплексом. Многогранник Q_3 имеет восемь вершин и размерность шесть. Его стороны определяются шестёрками вершин, т.е.

всеми кроме двух. Исключаемые вершины отвечают паре треугольников в полном трёхдольном графе с двумя вершинами в каждой доле. Возможны три случая взаимного расположения этих треугольников.

Треугольники не имеют общей вершины. Без ограничения общности можно считать, что первый треугольник состоит из первых вершин, а второй – из вторых вершин каждой доли, рис. 4. Остальные шесть вершин многогранника Q_3 принадлежат стороне, заданной уравнением $X_{1211}+X_{1311}+X_{2311}+X_{1222}+X_{1322}+X_{2322}=1$.

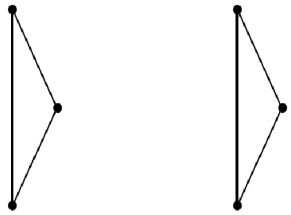

Рис. 4. Два треугольника без общей вершины

Треугольники имеют общее ребро. Без ограничения общности полагаем, что первый треугольник состоит из первых вершин каждой доли, и общее ребро соединяет первую и вторую доли графа, рис. 5. Выпуклая оболочка остальных шести вершин многогранника является стороной, заданной уравнением $X_{1211}=0$.

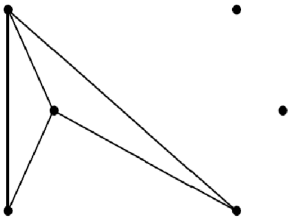

Рис. 5. Два треугольника с общим ребром

Треугольники имеют одну общую вершину. Без ограничения общности полагаем, что первый треугольник состоит из первых вершин каждой доли, и общая вершина принадлежит первой доле графа, рис. 6. В этом случае выпуклая оболочка остальных шести вершин многогранника не является стороной. Форма $X_{1222}+X_{1312}+X_{1212}+X_{1221}$ принимает значения нуль

и два на исключённых вершинах многогранника и значение единица на остальных шести вершинах многогранника.

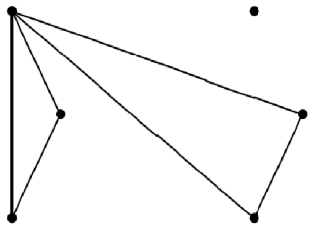

Рис. 6. Два треугольника с одной общей вершиной

Легко показать, что в общем случае число сторон многогранника Q_n не меньше чем 2^n.

Каждой 2-КНФ от m переменных, имеющей l дизъюнкций пар литералов, сопоставим $(l+m)$-дольный граф $G[\varphi]$, имеющий по две вершины в каждой доле. Доли и вершины графа пронумерованы, каждой вершине сопоставлен литерал. Для индексов i от 1 до l вершины i-й доли графа $G[\varphi]$ соответствуют литералам из i-го конъюнктивного члена. Для индексов j от 1 до m первая вершина $(l+j)$-й доли соответствует j-й переменной, вторая вершина – отрицанию j-й переменной. Две вершины из разных долей графа $G[\varphi]$ соединены ребром, если соответствующие литералы не являются отрицанием один другого.

Лемма 3. *Для 2-КНФ* φ *от* m *переменных с* l *дизъюнкциями граф* $G[\varphi]$ *содержит* $(l+m)$*-клику тогда и только тогда, когда 2-КНФ* φ *выполнима; если граф* $G[\varphi]$ *содержит* $(l+m)$*-клику, то литералы, связанные с вершинами* $(l+m)$*-клики, истинны на некоторой модели 2-КНФ* φ. *Следовательно, разным моделям соответствуют разные* $(l+m)$*-клики.*

Доказательство. Переменным, от которых зависят литералы в вершинах клики, можно независимо придать такие значения, что все эти литералы будут истинны. Обратно, если 2-КНФ истинна при некоторой оценке переменных, то каждый конъюнктивный член φ содержит литерал, истинный при этой оценке. *Лемма доказана.*

Заметим, что разные (l+m)-клики графа $G[\varphi]$ могут соответствовать одной и той же модели.

Лемма 4. *Множество таких пар (позитивная 2-КНФ φ, двоичная запись числа k), что 2-КНФ φ имеет модель, на которой оценка k литералов – ложь, и в каждой модели не больше чем k ложных литералов, NP-трудное.*

Доказательство. Для позитивной 3-КНФ φ, имеющей l дизъюнкций, определим индукцией по длине формулы позитивную 2-КНФ φ^*

$$(\lambda \vee \mu \vee \nu)^* = (\lambda \vee \mu) \wedge (\lambda \vee \nu) \vee (\mu \vee \nu)$$

$$(\vartheta \wedge \psi)^* = \vartheta^* \wedge \psi^*$$

На любой модели 2-КНФ φ^* имеет не более $2l$ ложных литералов, не более одного в каждой из подформул вида φ, поскольку для каждой подформулы ложность одной переменной влечёт истинность двух других. Следовательно, инвертируя оценки переменных в модели для 2-КНФ φ^*, где l ложных литералов, получим модель для 3-КНФ φ, на которой каждая дизъюнкция имеет один истинный литерал.

С другой стороны, инвертируя оценки всех переменных в модели для 3-КНФ φ, на которой в каждой дизъюнкции ровно один истинный литерал, получим модель для 2-КНФ φ^*, на которой ровно $2l$ ложных литералов. Поиск такой модели сводится к поиску модели для 2-КНФ φ^*, на которой не менее $2l$ ложных литералов. Так известная NP-полная проблема сводится к рассматриваемой за время, ограниченное полиномом от длины входа. *Лемма доказана.*

Лемма 5. *Модель для позитивной 2-КНФ φ от m переменных с l дизъюнкциями, на которой наибольшее число литералов ложны, соответствует вершине многогранника $Q_{l+m} \cap H(\mathrm{G}[\varphi])$, в которой линейный функционал $f(X) = \sum_i N_i X_{(l+i)(l+i)22}$, где N_i равно числу вхождений i-й переменной, достигает максимального значения.*

Доказательство. Если 2-КНФ φ выполнима, то граф $G[\varphi]$ имеет n-клику. Поэтому пересечение $Q_{l+m} \cap H(G[\varphi])$ непустое. Это пересечение является гранью многогранника Q_{l+m}. Поэтому вершины многогранника $Q_{l+m} \cap H(G[\varphi])$ соответствуют моделям для 2-КНФ φ. Линейный функционал f достигает на многограннике максимального значения на некоторой вершине X. Если индекс i не превосходит m, первая вершина $(l+i)$-й доли графа $G[\varphi]$ соответствует i-й переменной, а вторая – её отрицанию. Поэтому вершина X соответствует модели, на которой больше всего ложных литералов. *Лемма доказана*.

Теорема 4. *Если существует недетерминированный алгоритм для распознавания сторон многогранника Q_n за время, ограниченное полиномом от n, то классы coNP = NP.*

Доказательство. За полиномиальное время принадлежность входа множеству из класса coNP сводится к проверке несовместности системы линейных неравенств, причём для любого неравенства за полиномиальное время проверяется его принадлежность системе. По теореме Фаркаша надо проверить несовместность недетерминировано выбранной подсистемы полиномиального размера. Проверить несовместность можно за полиномиальное время алгоритмом Хачияна. Итак, классы coNP = NP. *Теорема доказана*.

1.2 Алгоритм решения неявно заданной системы однородных линейных уравнений над конечным бимодулем и нижняя оценка числа клик

Пусть A и B – некоторые кольца, а R – конечный A-B-бимодуль из r элементов. В частности, любое ассоциативное (но, вообще говоря, некоммутативное и без единицы) кольцо R является R-R-бимодулем.

Рассматривается произвольная система однородных линейных уравнений вида $\sum_{k,j} a_{kj} x_k b_{kj} = 0$ над бимодулем R. Множество решений

системы из m таких уравнений, каждое от n переменных образует подгруппу в R^n, которую будем обозначать B_m. Отметим два простых факта.

Лемма 6. *Пусть G – конечная абелева группа, H и K – её подгруппы. Индекс $(K: K\cap H)$ не превышает индекса $(G:H)$.*

Лемма 7. *Множество решений в R^n любого однородного линейного уравнения – подгруппа индекса не больше r, где r – порядок группы R.*

Доказательство. Рассмотрим линейное уравнение $f=0$ в R^n. Число его решений равно числу решений системы из двух уравнений $f=z$ и $z=0$ в R^{n+1}, где переменная z новая, не встречается в f. Уравнение $f=z$ выделяет в R^{n+1} подгруппу H индекса r, поскольку любое значения переменных $x_1,\ldots,x_n$ определяют ровно одно значение переменной $z=f(x_1,\ldots,x_n)$; индекс подгруппы H равен $r^{n+1}/r^n=r$.

Уравнение $z=0$ выделяет в R^{n+1} подгруппу $K=R^n$. По предыдущей лемме индекс $(K: K\cap H)$ не превосходит r, т.е. исходное уравнение $f=0$ выделяет в R^n подгруппу индекса не больше r. *Лемма доказана.*

Теорема 5. *Существует и ниже описан алгоритм, который для произвольной однородной линейной системы уравнений над R образует систему порождающих в группе B_m всех её решений и указывает число всех решений, выполняя $O\left((m+n)^2\cdot m\cdot L\cdot r^4\right)$ операций в R. Здесь m – число уравнений, n – число переменных, L – число операций, необходимых для проверки выполнимости любого уравнения системы при любых значениях переменных, r – число элементов в R.*

Алгоритм применим и к системе неявно заданных уравнений, когда коэффициенты уравнений не известны, но дана эффективная процедура для проверки выполнимости любого уравнения системы при любых значениях переменных. Процедура говорит «да» или «нет».

Доказательство. Образуем цепь коммутативных подгрупп

$$B_m \subseteq B_{m-1}\subseteq\ldots\subseteq B_2 \subseteq B_1\subseteq B_0=R^n,$$

где B_i – множество решений первых i уравнений системы.

Обозначим через A_i множество решений i-го уравнения этой системы. Очевидно, подгруппа$B_i=B_{i-1} \cap A_i$. По лемме индекс B_i в B_{i-1} не больше индекса A_i в R^n, который не превосходит r. Продолжим предыдущую цепь групп

$$\{0\}=B_{m+n}\subseteq B_{m+n-1}\subseteq\ldots\subseteq B_{m+1}\subseteq B_m\subseteq B_{m-1}\subseteq\ldots\subseteq B_2\subseteq B_1\subseteq B_0=R^n,$$

полагая для всех $i<n$ подгруппу $B_{m+i} = B_{m+i-1}\cap C_i$, где C_i – подгруппа в R^n, определяемая уравнением $x_i=0$. Индекс $(R^n: C_i)=r$. По лемме индекс B_{m+i} в B_{m+i-1} не больше r. Итак, для всей цепи групп индекс $(B_{i-1}: B_i)$ не больше r.

Представим группу B_i как дизъюнктное объединение $B_i = \sum_d B_{i,d} + B_{i+1}$, где $B_{i,d}$ – все смежные классы группы B_i по подгруппе B_{i+1} за исключением класса B_{i+1} нуля. (Интуитивно говоря, это – представление B_i в виде суммы с точностью до B_{i+1}.) Тогда $B_j = \sum_{i=j}^{m+n-1} \sum_d B_{i,d}$. Ясно, что B_j порождается любым набором представителей классов $B_{i,d}$. Напомним, что нас интересует случай $j=m$, когда набор представителей порождает (в виде сумм) как раз группу B_m всех решений исходной системы.

Фиксируем в $B_0=R^n$ тривиальную систему порождающих, состоящую из $(rn-n)$ элементов вида (0,..., 0, x, 0,..., 0), где x не равен нулю. Ниже описан алгоритм, который строит множества W_i представителей (не обязательно всех) смежных классов в группе B_i по подгруппе B_{i+1}. При этом из каждого смежного класса выбирается не более одного представителя, а представителем класса B_{i+1} нуля всегда берётся нуль.

Сначала опишем вспомогательный алгоритм. Этот алгоритм получает на вход ненулевой элемент g в R^n и набор множеств $\{W_i\}$. Если g не удаётся разложить в сумму элементов из множеств W_i, то алгоритм добавляет к множествам W_i новые элементы так, чтобы указанное разложение стало возможным.

Описание вспомогательного алгоритма.

1. Ищем наибольший индекс k, для которого g принадлежит множеству B_k.

2. Перебирая элементы множества W_k, ищем такой элемент x из множества W_k, что разность (g-x) принадлежит к B_{k+1}. Если такого элемента x нет, то присоединяем сам элемент g к множеству W_k. В ином случае, когда такой элемент x найден, рекурсивно применяем этот вспомогательный алгоритм к элементу (g-x).

Проверка принадлежности элемента g к группе B_k сводится к проверке k равенств, что требует $O(k \cdot L)$ операций. Если $k>m$, то эта проверка требует $O(m \cdot L)$ операций в исходном бимодуле R. Итак, вспомогательный алгоритм требует $O((m+n) \cdot m \cdot L \cdot r)$ операций.

Описание алгоритма для решения неявно заданной системы линейных уравнений.

1. Полагаем $W_i = \{0\}$ для всех i.

2. Последовательно применяем вспомогательный алгоритм ко всем элементам указанной выше тривиальной системы порождающих группы R^n.

3. Для каждого i, начиная от нуля, и до $m+n$-1 применяем вспомогательный алгоритм ко всем ненулевым элементам вида $x+y$, где x и y принадлежат множеству W_i, до тех пор, пока множество W_i не перестанет пополняться новыми элементами. Точнее, пока для W_i не перестанут появляться новые пары (x, y).

Выполнение пункта 2 требует $O((m+n) \cdot m \cdot L \cdot r^2 \cdot n)$ операций. Поскольку множество W_i содержит не более r элементов, то при каждом i просматривается не более r^2 пар элементов. Заметим, что применение вспомогательной процедуры к суммам элементов из W_i не добавляет новых элементов в W_j при $j<i$. Итак, выполнение пункта 3 требует $O((m+n)^2 \cdot m \cdot L \cdot r^3 \cdot n)$ операций.

Элемент g из R^n, входящий в B_k и не входящий в B_{k+1}, назовём *элементом ранга k*.

Перейдем к **обсуждению алгоритма**. Он строит множество W_i представителей (не обязательно всех) смежных классов в группе B_i по подгруппе B_{i+1}, для всех i.

Класс B_{i+1} всегда представляется элементом 0.

Элементы множества W_i (кроме 0) имеют ранг i.

Объединение множеств W_i для всех индексов i от нуля до $m+n$-1 порождает группу R^n.

Более того, каждый элемент g из R^n ранга k представим в виде суммы элементов из множества $\bigcup_{i=k}^{m+n-1} W_i$, $g=x_1+...+x_j$, где x_1 ранга k, и ранги всех элементов x_1, ..., x_j строго возрастают.

Доказательство очевидно из двойной индукции по рангу и числу элементов одного ранга. Действительно, если $g = y_1+...+y_l$ – некоторое представление элемента g ранга k (пусть ранги элементов y_1,..., y_l не убывают), то y_1 имеет ранг k, и для каждого множества элементов одинакового ранга, начиная с наименьшего ранга, заменим (пусть y_1 и y_2 одинакового ранга) сумму y_1+y_2 на равную ей сумму элементов (из объединения множеств $\bigcup_{i=k}^{m+n-1} W_i$) попарно различных рангов. Тогда число элементов ранга k в правой части исходного равенства уменьшится на единицу.

Отсюда сразу следуют такие свойства построенной системы множеств W_i.

1. Множество W_{m+n-1} равно B_{m+n-1}.

2. Элемент g ранга k (т.е. g принадлежит разности множеств $B_k \backslash B_{k+1}$) представим как сумма элементов из множества $\bigcup_{i=k}^{m+n-1} W_i$. В частности, это множество порождает группу B_k.

3. Множество W_i содержит по одному представителю каждого смежного класса в группе B_i по подгруппе B_{i+1}. Действительно, для любого смежного класса $[g]$ в B_i имеем $g=x_1+...+x_j$, и представителем класса $[g]$ является x_1.

4. Множество $\bigcup_{i=k}^{m+n-1} W_i$ порождает группу B_m, которая является искомым множеством решений системы линейных уравнений.

5. Построение системы множеств W_i требует выполнения $O\left((m+n)^2 \cdot m \cdot L \cdot r^4\right)$ операций в исходном бимодуле R.

6. Число элементов в множестве B_m равно произведению мощностей множеств $W_m,..., W_{m+n-1}$. *Доказательство завершено.*

Заметим, что без всяких изменений этот алгоритм и соответствующее доказательство переносятся на произвольную конечную абелеву группу с конечными множествами произвольных (левых и правых) операторов, действующих на ней.

Нижняя оценка числа *n*-клик в *n*-дольном графе, содержащем хотя бы одну *n*-клику.

Пусть абелева группа G действует на множествах вершин каждой доли графа. Тогда определено действие прямой суммы n экземпляров группы G на множестве вершин графа, сохраняющего доли. Подгруппа H в рассматриваемой прямой сумме экземпляров группы G, действие которой сохраняет также и рёбра графа, то есть отношение смежности на вершинах, выделяется системой однородных линейных уравнений над группой G, рассматриваемой как модуль над кольцом целых чисел. Очевидно, если в графе есть хотя бы одна n-клика, то общее число n-клик не меньше чем порядок группы H. Аналогично можно оценить снизу число клик произвольного размера, если рассматривать подграф, состоящий из соответствующих долей.

В частном случае, когда граф имеет ровно две вершины в каждой доле, группа его автоморфизмов, сохраняющих доли, является подгруппой в *n*-кратной прямой сумме группы из двух элементов. В этом случае вычисление группы автоморфизмов сводится к решению системы линейных уравнений над полем вычетов по модулю два. Если отношение смежности на вершинах графа задано неявно, то и соответствующая система линейных уравнений задана неявно.

1.3 Алгоритм поиска клики в многодольном графе в общем случае и поиск консервативных участков в наборе последовательностей на основе этого алгоритма, учёт дерева видов

Алгоритм поиска сигнала, который использован во всех биологических приложениях, основан на поиске клики в соответствующем многодольном графе. А поиск клики достаточно большого размера в свою очередь основан на подсчёте числа клик малого размера в этом графе. Этот *эвристический алгоритм* принципиально отличается от алгоритма, описанного в пункте 1.1 и основанного на задаче линейного программирования. В пункте 1.1 было приведено нетривиальное доказательство применимости для поиска клики в многодольном графе с не более чем двумя вершинами в каждой доле, такого алгоритма линейного программирования. Применимость для таких специальных графов алгоритмов проверки выполнимости 2-КНФ очевидна. Хотя последний алгоритм линейный и в этом смысле наилучший из возможных, алгоритм линейного программирования в среднем также является весьма быстрым и практически удобным, хотя в худшем случае он только полиномиальный. Фактически при поиске сигнала применялся указанный ниже эвристический алгоритм.

Алгоритмы подсчёта клик малого размера. Напомним, что для любых двух матриц M и N одинакового размера *произведением Адамара* $M*N$ называется матрица, элементы которой – произведения

соответствующих элементов матриц M и N; эта операция коммутативна. Тривиальный алгоритм (обычного) умножения двух квадратных матриц порядка V требует $O(V^3)$ операций (в кольце целых чисел). Алгоритм Штрассена для такого умножения требует $O(V^w)$ операций, где $w = \log_2 7 \approx 2.81$. Известны алгоритмы умножения матриц, которые требуют $O(V^w)$ операций, где $w<2.376$, [Сэвидж, 1998]. (Все оценки сверху.)

Рассмотрим неориентированный граф G без петель и обозначим V число его вершин, E число его рёбер, который задаётся матрицей смежности: симметричной 0-1-значной матрицей A порядка V, на главной диагонали которой стоят только нули.

Рассмотрим матрицу $A^*(A^2)$, вычисление которой требует $O(V^w)$ операций. Её элемент на месте ij, не лежащий на главной диагонали, равен числу *треугольников*, т.е. числу 3-клик, содержащих ребро ij. Заметим в связи с 6-кликами, которые будут рассмотрены ниже, что на главной диагонали матрицы A^3 находятся удвоенные числа треугольников, содержащих соответствующую вершину. А также заметим, что другой возможный способ подсчёта числа треугольников с наперёд заданным ребром ij состоит в перечислении всех треугольников. Это можно сделать тривиально за $O(VE)$ шагов. При небольшом числе E рёбер этот способ может быть быстрее, чем вычисление матрицы $A^*(A^2)$, так как вычисление матрицы за $O(V^w)$ шагов только асимптотически быстрее перебора. В алгоритме использовано именно вычисление матрицы.

4-клика определяется парой рёбер *без общей вершины*. Число 4-клик, содержащих какое-то фиксированное ребро ij, вычисляется прямым перебором всех рёбер, что происходит за $O(E)$ шагов.

Построим ещё один вспомогательный граф G^* следующим образом. Вершины графа G^* соответствуют рёбрам графа G. Две вершины графа G^* смежны, если они соответствуют двум рёбрам без общей вершины, которые определяют 4-клику в G; тогда каждой 6-клике в G соответствует ровно 15 треугольников в G^*. Поэтому, подсчитав число треугольников в графе G^* с

помощью третьей степени матрицы смежности для графа G*, как об этом говорилось выше, и поделив на число 15, получим число 6-клик в исходном графе G. Итак, подсчёт числа 6-клик в графе G требует $O(E^w)$ операций, где E – число рёбер в графе G.

Описание эвристического алгоритма поиска сигнала и клики. По исходному набору из n невыравненных последовательностей ищется *сигнал* – набор наиболее попарно похожих слов одной и той же фиксированной длины, при этом из каждой последовательности берётся не более чем по одному слову. Определим граф G, *вершинам* которого взаимно однозначно приписаны все слова фиксированной длины, взятые из всех исходных последовательностей. А *долей* объявляются все вершины, которым приписаны слова из какой-то одной исходной последовательности. Таким образом, в графе G ровно n долей. В графе G две вершины соединяются *ребром*, если величина сходства между словами, приписанными этим вершинам, превышает некоторый порог, который является параметром алгоритма. Алгоритм ищет в таком многодольном графе список клик данного размера (т.е. с числом вершин равным) q, где q – параметр. Такие клики будем называть *q-кликами*. В процессе поиска сигнала значение q постепенно уменьшается. Рассмотрим текущее фиксированное значение q.

Следующий алгоритм порождает список q-клик, состоящий не обязательно из всех q-клик или даже пустой список. В начале *текущим графом* G' объявляется исходный граф G и *текущим списком* CL q-клик объявляется пустой список.

В текущем графе G' сначала *исключаются вершины* (и все инцидентные им рёбра), которые соединены хотя бы одним ребром с долями таким образом, что суммарное число этих долей строго меньше q-1. Затем *исключаются рёбра*, которые принадлежат строго меньшему, чем q-2 *числу* 3-клик, *или* строго меньшему, чем $(q\text{-}2)(q\text{-}3)/2$ *числу* 4-клик. Такое исключение последовательно повторяется сначала для всех вершин и затем для всех рёбер текущего графа G' до тех пор, пока это возможно.

Когда это невозможно и при этом *удалены все рёбра*, то алгоритм *завершает работу* и выдает текущий список CL q-клик, сформированный к этому моменту. Если при этом *остались рёбра* в текущем графе, то алгоритм проверяет наличие какой-то вершины R степени ровно q-1 в текущем графе. Если такая вершина *найдена*, то алгоритм тривиально проверяет, образует ли эта вершина вместе со всеми смежными ей вершинами q-клику, и затем *удаляет* вершину R из текущего графа (вместе со всеми инцидентными ей ребрами). Если вершина R вместе со смежными ей вершинами образует q-клику, то алгоритм *включает эту q-клику* в текущий список CL q-клик.

Если вершина R степени ровно q-1 *не найдена*, то одно ребро текущего графа, входящее в наименьшее число треугольников в нём, удаляется.

К так полученному текущему графу G' снова с самого начала применяется описанный алгоритм до тех пор, пока это возможно.

Каждая клика из списка q-клик определяет свой сигнал, состоящий из слов, которые приписаны вершинам клики. Каждому сигналу приписывается его *качество*, измеряемое числом, которое тем больше, чем больше число высоко консервативных позиций в сигнале.

Как уже говорилось, этот алгоритм не гарантирует нахождение всех клик размера q. В частности, если соответствующий граф полный n-дольный, то алгоритм найдёт ровно одну n-клику, проделав при этом очень большое число операций, за счет того, что из исходного графа нужно удалить почти все рёбра. Однако, если исходный граф состоит из некоторого числа непересекающихся q-клик и небольшого числа других рёбер, как это часто бывает в биологических данных, то алгоритм корректно и очень быстро определяет все клики.

Алгоритм может быть улучшен за счёт включения в описанные выше условия на удаление рёбер числа 6-клик, для чего необходимо предварительно вычислить *число* всех 6-клик. И аналогично для 7-клик и

т.д., что, однако, приводит (начиная с 7) к быстрому росту сложности алгоритма. Вообще проблема вычисления числа малых клик в многодольном графе важна для биоинформатики.

Программа, реализующая указанный алгоритм, имеет графический интерфейс и написана на языке Object Pascal в среде программирования Delphi для операционной системы Windows.

Поиск сигнала при известном дереве видов. Если известно дерево тех видов, из которых брались исходные нуклеотидные последовательности, то рассмотренный алгоритм может быть усовершенствован следующим образом и это приводит к более надёжным результатам.

Рассмотрим набор 5'-нетранслируемых областей перед ортологичными генами бактерий и соответствующее дерево видов бактерий. Выделим филогенетические поддеревья, соответствующие небольшим таксономическим группам. Для каждой такой группы ищем один или несколько сигналов из слов переменной длины m с помощью алгоритма поиска клики в многодольном графе, как описано выше. По каждому найденному для одной группы сигналу формируем $4\times m$ *весовая матрица*, где k-й столбец матрицы, $1\le k\le m$, содержит частоты вхождения букв в k-й позиции сигнала. Теперь образуем множество всех весовых матриц всех сигналов во всех филогенетических группах и кластеризуем его. Кластеризация матриц происходит обратно пропорционально расстояниям между матрицами и между предками соответствующих филогенетических групп в дереве видов. Затем матрицы в составе каждого кластера заменяются снова исходными сигналами, образуя “составной” сигнал, у которого сайты в каждой последовательности могут состоять из нескольких слов.

Заметим, что кластеризацию множества матриц можно выполнить также с помощью алгоритма поиска клики. Составной сигнал можно

построить и перебором, если в каждой таксономической группе было выделено небольшое число первичных сигналов.

1.4 Тестирование алгоритма

Подтверждением корректности работы алгоритма служит успешный поиск консервативных участков в 5'-нетранслируемых областях мРНК хлоропластов и актинобактерий, подробно описанный во второй главе. При этом подтверждение биологической значимости найденных результатов основано на независимом от поиска клик анализе вторичной структуры РНК и, в некоторых случаях, на экспериментальных данных о регуляции экспрессии генов. С другой стороны, для контроля *отсутствия недопредсказаний* применялся поиск структур РНК по образцу, выполняемый независимо от поиска клики с помощью вспомогательных программ, описанных ниже. Вычисления подтверждают отсутствие недопредсказаний в рассмотренных случаях.

Следующий пример показывает теоретическую возможность применения алгоритма для поиска сайтов связывания белков с ДНК. Хотя такие исследования не проводились автором систематически, вычисления были проведены на хорошо изученном примере из биологии и служат дополнительным подтверждением корректности работы алгоритма.

Тестирование на сайтах связывания белка PurR с ДНК в кишечной палочке. В этом случае на вход были поданы 19 последовательностей – регуляторных областей генов одного регулона; каждая длиной примерно 240 знаков. Сайты длиной 16 найдены во всех последовательностях (в данном случае q=19), по одному слову в каждой последовательности, кроме двух, в каждой из которых получено по два слова, что соответствует двум кликам такого размера. Был построен граф с 3560 вершинами. В зависимости от критического расстояния между словами DistMax числа рёбер менялись следующим образом: (0, 0), (1, 2), (2, 64), (3, 344), (4, 1480), (5, 6696), (6, 30018), (7, 116350), (8, 381004), (9,

1046796). При DistMax<8 не найдено ни одного сигнала. При DistMax=8 было найдено три сигнала, у двух по пять абсолютно консервативных позиций, у третьего – только три. Найденные сигналы соответствуют известным биологическим данным. В данном случае результат не зависит от флага Thetrahedron.

Ниже приведены исследованные последовательности, найденные сайты выделены прописными буквами.

purR: gctccgtgtcgtttttccggcgtaccgcaacacttttgttgtgcgtaaggtgtgtaaAGGCAA ACGTTTACCTtgcgattttgcaggagctgaagttagggtctggagtgaaatggaatggcaacaataaaaga tgtagcgaaacGAGCAAACGTTTCCACtacaactgtgtcacacgtgatcaacaaaacacgtttcgtcg ctgaagaaacgcgcaacgccgtgtgggcagcgattaaa

purEK: AGGAAAACGGTTGCGTggctgtgaaatcagcaaagttgcgggtttttttaaaac ggaaaatgaatcagctcaacgtcatccgccgtgactttcaccattgaaccttccgtatgccaggcacccagtaccac gcgaaaagcaggttgctgattggcgataagttcatgcaccgccgggcgatgggtatgcccgtggatcagc

cvpApurF: AAGAAAACGTTTGCGTagggatttccttcccgcgcatcaataaaaatggc gctgaaaaaatattcaacgccatcgacttttttatgcctttgcggcatcgggcaatgcgtgtcggatgcggcgtaaacg ccttatccgacctacggttctacccctgcgtaggcctgataagacgcgccagcgtcgcatcaggcaagaccg

purC: ACGCACACGTTTGCGTatcatatcagaaaaaagggccggatgattccagccctg tatttttacttgctaaacgcagcctggaagacagctaccagcgcgtcgttctgactctgagtcagagtatgacctttcg gatcgatgaactgtaggctgctgcggttatctaaatcgccaacctgcagtttatagtcaccggatgc

purMN: tccgaaaactaacctttaccctggcacaagtcttctttcgccgcgcgcctggggaaaagacgt gcaaaaaggttgtgtaaagcagtcTCGCAAACGTTTGCTTtccctgttagaattgcgccgaattttatttt tctaccgcaagtaacgcgtggggacccaagcagtgaccgataaaacctctcttagctacaaagatgccggt

purL: ACGAAACCGTTTGCGTggaaataaaatcaccatcgtgaattagcaacgcgtgcc gccaatggctgtaataagttgccatctggcgcaggtttacgcaaaatgccgctcatttatgagtaaacctttcactattat tacgtttttttcaagctgggacgcgcacgacacagagaattaactaattgaaaaaattaaagatta

purB: AGGTAACCGATTGCGTcggcagcaaaagcaggaacttccttgatcgctgcgtgc gcggccagtttttgcagccaacgtacttcaacttgtacacggaatttcagcaaaccatattcgctgaaaatcccgcgc agcgcgctgactttatcgccgtagcgtccatcgacaggggaaacggcggtcagtgaggataattccat

guaBA: GCGTAACCGATTGCATctacccctttttgcaaaaaatgcttgctatccccgaag ggcgggttactatcgactgaataacctgctgatttagaatttgatctcgctcacatgttaccttctcaatcccctgcaattt ttaccgttagtcgcctgaatcaaacggttcgtctgctgcttgagcatgagatgggacaggtttg

purHD: ACGAAAACGTTTGCGCaacgctcgcgaatttttctctttcaatggtgatcacaat tttgactgtggttaccgtgggcaaaatacagaaattacattgatgattgtggataactctgtgcgtaaaaaggtataaag cgggcttttgctggggaatgcagcagtcagtcatttttctgcaatttttctattgcggcctgcgg

glyA: ACGCAAACGATTACCTtcaggctacgcaaggctttggagaataaagagcttgca accggaaacggatttcttttcaggtttgtgatgcaaatttttcacttcatcacattctttctgaaaaacaccaaagaaccat ttacattgcagggctatttttttataagatgcatttgagatacatcaattaagatgcaaaaaaag

pyrD: tcgcaccgtgtcatgtcgcctggcgggcgataaatgattacggcgggttgagtgcaaagaagg agcaaaatctgccctgaaacaggttCGGAAAACGTTTGCGTttttttgccgcaggtcaattccctttttg gtccgaactcgcacataatacgcccccggtttgcacaccgggaatccaggagagttcatgtactacc

prsA: GCGAAAACGTTTTCTTgcgataacctcgcaatcattgctgaattcacattttagcaa cgtttctattgacctttttggcgtattgcgcgggagttcaggatctttaaaaatcaccggagtcggaatacttacaccag ggtgcgccaccagataccacttctctggcggatccaccggcgttagtatttcaccaacgcctt

glnB: TTGAAATCGTTTGCATccagctcgtgtcgggaaagcagtttataaaattctgtccg gttgcgccccgccattctcgccgcgtgggtgacgttgcctttggtgatttgcagcagcttACGCAAATAGT TGAGTtcaaactgattacgtgcctcaacaaaggttggcagcgccgtattttcaccctccagcgcctgctccac

purA: cggtggacttgtggttgcgggcgttgtggtctactacatgttgAGGAAAACGATTG GCTgaacaaaaaacagactgatcgaggtcatttttgagtgcaaaaagtgctgtaactctgaaaaagcgatggtag aatccatttttaagcaaacggtgattttgaaaaatgggtaacaacgtcgtcgtactgggcacccaatgg

codBA: ttttacaccgataatttttcccccacctttttgcactcattcatataaaaaatatatttccccACG AAAACGATTGCTTtttatcttcagatgaatagaatgcggcggatttttgggtttcaaacagcaaaaaggg ggaatttcgtgtcgcaagataacaactttagccaggggccagtcccgcagtcggcgcggaaag

pyrC: GCGGAAACGTTTTCCTttgcacgaaaaataaagggcgcgaatgcgccctcgtga ttaatcagtaaatggaatgacaatttcgcctggcttcacttcaatgcctttcgccagttttttcgccattgcttcgccctgg ctgccatcttcgcgcaggacgtaagcaggttgctggttaaagtaattgcgtaatgcctggttca

purT: gtatttaacttcaaccgccatttgcagcctctcataataactgtgattttatacagtatatttcttttcgg ttgagaaatcaacatcagcaataaagacacACGCAAACGTTTTCGTttatactgcgcgcggaattaat caggggatattcgttatgacgttattaggcactgcgctgcgtccggcagcaactcgcgtga

gcvTHP: ACGCAATCGTTCTCTTttgcctgaacttaccaccgaaacagactgttaaccat aaggtaaaattgatcatcacattagcttatggttaaaaaatgcaaaaatcgcgacagaataaaaaaccaaaaaaataca ccagtttctatacaaagatgatgtgatgagaaagtcaatttgaataagacaatattaagagctaaaaaa

speAB: GCGCAACCGGTTTCTTtttcataacattattaagcacataaccgaacgtaagtgt gaaagttcggcgaaaccacgagaaaactcttgtttttacaagagcgcccttgttcagtcctcagtaactgtaaccagct cttgaatcctgagaagcgccgagatgggtataacatcggcaggtatgcaaagcagagatgcagagt

1.5 Вспомогательные программы: выделение лидерных областей генов и поиск спиралей и слов специального вида по их параметрам в аннотированных геномах

Для исследования геномов использовалась разработанная автором программа для выделения лидерных областей генов, содержащих спираль с ограничениями на её длину и на расстояние от инициирующего кодона гена, и для поиска участков данной длины, близких к образцу.

Программа имеет графический интерфейс и реализована для операционной системы Windows. На вход программа получает файл *.gbk с аннотированным геномом из ГенБанка. В результате программа выдаёт список в формате FASTA, состоящий из координат и нуклеотидных последовательностей, примыкающих к инициирующим кодонам генов. Выбор файла происходит после выбора опции Open главного меню. Опция меню Save позволяет сохранить выходные данные в дисковый файл. Опция Font позволяет изменить шрифт. Пользователю предоставлен выбор следующих параметров, которым должны удовлетворять отобранные участки:

- длина рассматриваемого участка;
- такое образцовое слово в алфавите {A, C, G, U, R, Y, N}, что участок включает слово, близкое к образцу и максимальное число ошибок в нём;
- длина плеча спирали, 5’-плечо которой состоит из A или G, и максимальное расстояние от начала 3’-плеча до инициирующего кодона;
- длина плеча произвольной спирали на рассматриваемом участке.

Все параметры, кроме первого, учитываются, только если выставлены соответствующие флаги. Если выставлен флаг RNA, то выходные данные представлены в алфавите {A, C, G, U}, иначе в алфавите {a, c, g, t}.

Другая программа, также разработанная автором, предназначена для поиска пары слов, близких к соответствующим образцам. Программа имеет графический интерфейс и реализована для операционной системы Windows. На вход программа получает файл с нуклеотидной последовательностью. В результате программа выдаёт список в формате FASTA, состоящий из координат и найденных участков нуклеотидной последовательности от начала первого найденного слова до конца второго слова. Выбор файла происходит после выбора опции Open главного меню. Опция меню Save позволяет сохранить выходные данные в дисковый файл. Опция Font позволяет изменить шрифт. Пользователю предоставлен выбор следующих параметров, которым должны удовлетворять отобранные участки:

- первое образцовое слово в алфавите {a, c, g, t, r, y, x} так, что начало участка близко к образцу, и число ошибок;
- второе образцовое слово в алфавите {a, c, g, t, r, y, x} так, что конец участка близок к образцу, и число ошибок;
- верхняя граница на расстояние между искомыми словами.

Возможность поиска не только в 5'-нетранслируемых областях, но по всему геному, позволяет исправить возможные ошибки в аннотации. Кроме того, так был найден LEU-элемент в составе предполагаемой транспосазы из *Bifidobacterium longum*.

ГЛАВА 2. Предсказание структур РНК, регулирующих экспрессию генов, у хлоропластов и бактерий на основе алгоритма поиска клики

2.1 Регуляция трансляции посредством взаимодействия белков с РНК для различных генов у хлоропластов

Найдены консервативные структуры РНК в 5'-нетранслируемых областях генов фотосистем у хлоропластов многих водорослей и растений и перед генами *atpF*, *petB* и *clpP* из хлоропластов растений.

Материалы. Геномы хлоропластов получены из базы данных ГенБанка (NCBI). В качестве набора последовательностей были взяты 5'-нетранслируемые области перед генами хлоропластов у *Euglena gracilis, Cyanidioschyzon merolae, Cyanidium caldarium, Gracilaria tenuistipitata, Guillardia theta, Nephroselmis olivacea, Odontella sinensis, Porphyra purpurea, Chlamydomonas reinhardtii, Chaetosphaeridium globosum, Mesostigma viride, Anthoceros formosae, Adiantum capillus-veneris, Huperzia lucidula, Marchantia polymorpha, Psilotum nudum, Pinus thunbergii, Amborella trichopoda, Arabidopsis thaliana, Atropa belladonna, Calycanthus floridus, Cucumis sativus, Epifagus virginiana, Lotus corniculatus, Nicotiana tabacum, Nymphaea alba, Panax ginseng, Spinacia oleracea, Oryza nivara, Oryza sativa, Triticum aestivum, Zea mays*.

Заметим, что *Epifagus virginiana* не является фотосинтезирующим. Гены фотосистем в его хлоропластах отсутствуют.

В аннотации к хлоропласту из *Psilotum nudum* ортолог гена *psbB* назван *psbT*.

Результаты. Применение алгоритма поиска консервативных участков к 5'-нетранслируемым областям генов позволило выделить протяжённые консервативные участки, включающие шпильки РНК. Кратко результаты представлены в табл. 2.

Таблица 2. Распределение сайтов связывания перед генами у хлоропластов. Обозначения: "+" – предполагаемый сайт связывания белка с РНК найден, "-" – сайт не найден, "s" – гены содержит интроны, "n" – отсутствие гена, указанного в столбце, у вида, указанного в строке.

Отдел	Вид	*atpF*	*clpP*	*petB*	*psaA*	*psbA*	*psbB*
Euglenozoa	*Euglena gracilis*	-s	-	-s	-s	-s	-s
Bacillariophyta	*Odontella sinensis*	-	-	-	+	+	-
Cryptophyta	*Guillardia theta*	-	-	-	+	+	-
Rhodophyta	*Cyanidioschyzon merolae*	-	-	-	-	+	-
	Cyanidium caldarium	-	-	-	-	-	-
	Porphyra purpurea	-	-	-	+	+	+
	Gracilaria tenuistipitata	-	-	-	-	+	-
Chlorophyta	*Chlamydomonas reinhardtii*	-	-	-	-s	+s	-
	Nephroselmis olivacea	-	-	-	+	+	+
Streptophyta	*Chaetosphaeridium globosum*	-	+s	-s	+	+	+
	Mesostigma viride	-	-	-	+	-	-
Anthocerophyta	*Anthoceros formosae*	+s	+s	+s	+	+	+
Hepatophyta	*Marchantia polymorpha*	+s	+s	+s	+	+	+
Lycopodiophyta	*Huperzia lucidula*	+s	+s	+s	+	+	+
Pteridophyta	*Adiantum capillus-veneris*	+s	+s	-s	+	+	+
Psilophyta	*Psilotum nudum*	+s	+s	+s	+	+	+
Pinophyta	*Pinus thunbergii*	+s	+	+s	+	+s	+

Таблица 2 (продолжение)

Отдел	Вид	*atpF*	*clpP*	*petB*	*psaA*	*psbA*	*psbB*
Magnoliophyta (двудольные)	*Amborella trichopoda*	+s	+s	+s	+	-	+
	Arabidopsis thaliana	+s	+s	+s	+	+	+
	Atropa belladonna	+s	+s	+s	+	+	+
	Calycanthus floridus	+s	+s	+s	+	+	+
	Cucumis sativus	+s	+s	+s	+	+	+
	Epifagus virginiana	n	+s	n	n	n	n
	Lotus corniculatus	+s	+s	+s	+	+	+
	Nicotiana tabacum	+s	+s	+s	+	+	+
	Nymphaea alba	+s	+s	+s	+	+	+
	Panax ginseng	+s	+s	+s	+	+	+
	Spinacia oleracea	+s	+s	+s	+	+	+
Magnoliophyta (однодольные)	*Oryza nivara*	+s	+s	+s	+	+	+
	Oryza sativa	+s	+s	+s	+	+	+
	Triticum aestivum	+s	+s	+s	+	+	+
	Zea mays	+s	+s	+s	+	+	+

АТФ синтаза. Ген *atpF*, кодирующий одну из субъединиц АТФ синтазы, имеет интроны в хлоропластах растений (кроме *Epifagus virginiana*). У *Adiantum capillus-veneris* инициирующий кодон AUG получается из кодона ACG при редактировании мРНК. В остальных случаях инициирующим кодоном является RUG. В лидерных областях гена *atpF* у растений найдены по два консервативных слова, близких соответственно к RAUNAAAAAA и WAUCUAUAAGAGGAGANNA, причём второе расположено вблизи старта трансляции и перекрывает предполагаемый сайт связывания рибосомы. Здесь и далее обозначения нуклеотидов берутся из табл. 1. Между участками расположен AU богатый участок от 19 до 61 нуклеотидов, неконсервативный даже у цветковых растений. См. табл. 3.

Цитохром b6. В 5’-лидерной области гена *petB* (cytochrome b6) хлоропластов у всех растений кроме *Adiantum capillus-veneris* и *Epifagus virginiana* найден консервативный участок, примыкающий к инициирующему кодону AUG и близкий к консенсусу GGUAGUUCGA YCGYGGAAUU*YUUU***GUUUNNGUAUUUYYGGAAU. Здесь же расположена консервативная спираль.

Этот участок перекрывает предполагаемый сайт связывания рибосомы, но имеет очень большую длину. Перед ним расположен неконсервативный участок.

У *Epifagus virginiana* ортолог гена *petB* отсутствует. У *Adiantum capillus-veneris* соответствующая 5’-лидерная область не выравнивается. У всех растений ген *petB* содержит интроны. См. табл. 4.

АТФ-зависимая протеаза. В 5’-нетранслируемой области гена *clpP* (протеолитическая субъединица АТФ-зависимой протеазы Clp) из хлоропластов у растений найден консервативный участок, близкий к консенсусу UUACGYUUYCAYAUYARAGNRNARU. Здесь же расположена консервативная спираль РНК. См. табл. 5.

Фотосистема I. В 5’-нетранслируемых областях гена *psaA* (фотосистема I, P700 апопротеин A1) у многих хлоропластов найдены длинные консервативные AG-богатые участки, примыкающие к инициирующему кодону AUG, с консенсусом GUURGYRRGUYUYUUY *UAUN*****NUYGUCYGRARAGAGGAGRA*CUCR, найдена консервативная спираль около сайта связывания рибосомы, см. табл. 6.

Таблица 3. Множественное выравнивание 5’-нетранслируемых областей пред геном *atpF* хлоропластов.

Anthoceros formosae	**aaugaauaau**agcuuaagcuuuccuauuuuaaguaggauuggauaaggaagagacauacuaagacuuaaag**aaccuaugaugggagagag**agu
Marchantia polymorpha	**aaugaaaaaa**acuguugaguaaacaaaaaacucca**********uaauuuucaauaauauaauaa**cgaaaaaaagaggacagc****
Adiantum capillus-veneris	**aauaauuaau**gugucaauucuugcuaggcuggauca**********gaaauugcccaaaacguaaaa**ccuucgaggagggaaaaga**au*
Huperzia lucidula	**aaggaaaaaa**cuauguaaacuu**********ggauaau**aaccuguaaugggagaaaa**gu*
Psilotum nudum	**aauaaaagaa**aaacuuugucaa**********aauuaga**uagucauuaugggagaggu**auu
Pinus thunbergii	**aauaagaaaa**uaacaaaauucugua**********gaacauauccu**uaucuaugaggggagagcg**u**
Amborella trichopoda	**aauaaaaaua**agauaugggugaagugauacaaaaaga**********acuccguuugguuuuguuagucc**uaucuagaagaggagagua**u**
Arabidopsis thaliana	**aauaaaaaaa**aaaggacagaguuccu**********uuuuuauaguu**uagcuagaagaggagauua**u**
Atropa belladonna	**aauaaauaaa**aagaggggcgaagugcuacaaaaaga**********acucuguucgauuuuuuaguc**uaucuauaagaggagauca**u**
Calycanthus floridus	**aaucaaaaaa**aaggggggggcgaagugauacaaaaagaacu******cugcgcaauuuuguuagcccuaucu**auucuauaagaggaaagca**u**
Cucumis sativus	**aauaaaaaaa**auagaaagaaauaga**********uaauuaguuu**uaucuauaaaaggagauca**u**
Lotus corniculatus	**aaaaaaaaaa**uaaagaaaaaaucuacaaaaauagga**********aucauauaaagagaauucguu**uauccauaagaggagauca**u**
Nicotiana tabacum	**aauaaauaaa**aagaggggcgaagugcuacaaaaaga**********acucuguucgauuuuuuaguc**uaucuauaagaggagauca**u**
Nymphaea alba	**gauacaaaaa**aggaaaucuuuuuuuguauuuug**********uuaguccuauuccuaucc**auaagauaagaggagagca**u**
Panax ginseng	**uauaaaaaaa**gaacaggggcaagguuauacaaaaagaac********ucuguucuuuuuuuuuuuuuaguc**uaucuauaagaggagauca**u**
Spinacia oleracea	**aauaaaaaaa**auaaaauucauuuagaaguagcaaacaauu******gaaauaauacaacgauuuuuuuguu**uaucuauaagaggagauca**u**
Oryza nivara	**auuaaaaaaa**ccgaucaaaaagggcgagcgaaguaagugau****cgaaaaacuuuguucuuuguucgucc**uaucuauaagaggagagca**u**
Oryza sativa	**auuaaaaaaa**ccgaucaaaaagggcgagcgaaguaaaguga***ucgaaaaacuuuguucuuuguucgucc**uaucuauaagaggagagca**u**
Triticum aestivum	**gaucaaaaaa**gggcgagcgaaguaagugaucgaaa**********aacuuuguucuuuguucgucc**uaucuauaagaggagagc**au**
Zea mays	**gaucaaaaaa**ggacgagcgaaguaaguaaucgaaa**********aacuuucuucuuuguucgucc**uaucuauaagaggagagc**au**
Консенсус	raunaaaaaannnnnnnnnnn**********nnnnnwaucuauaagaggagann****

Таблица 4. Выравнивание 5’-нетранслируемых областей пред геном *petB* хлоропластов.

Спирали выделены прописными буквами.

Anthoceros formosae	***uuuucccagug*gugGUAGUUuaaucgugcAACUACugaaaaaaaaaaggauuuuugaaau
Marchantia polymorpha	cauuuuuuuaauuuu*aggUAGUUuaauuguguAAUUA*uuaa**auucaaggauuu*uugaau
Huperzia lucidula	uugauccuucccuuu*ugguAGUUuaaucguguAAUU*cuga***aucaaaggaucuuuagaau
Psilotum nudum	ucauaaaaaaagac*gaggcagUUGaaucacgCAAauuauua***auuuaaugauguuuguaau
Pinus thunbergii	agcuuaucuuguuc**cacUAGUUugaucguguAAUUAcuuuu**cucuaaggauuuuuggaau
Amborella trichopoda	uggguuucuagguu*a*gguaGUUCGaccguGCAAUUccuuu***guuucgguauuuccggaau
Arabidopsis thaliana	**ccuauucuccuu**ugguAGUUCGaccgCGAAAUuuuuuucugcauuguauauuuccggaau
Atropa belladonna	cauucuauuuucuuu*ugguAGUUCgaucgugGAAUUucuuu***guuucuguauuuccggagu
Calycanthus floridus	uuuucuaggccauuc*ugguAGUUCgaccgugGAAUUccguu***guuucgguauuuccggaau
Cucumis sativus	uuagccuacucuuuuuugguAGUUCgaucgugGAAUUuauuu****uuucuguauuuccggaau
Lotus corniculatus	cauuccuuuuuuauu*ugguAGUUCgaucgugGAACUucuuu***guuucuguauuuccggaau
Nicotiana tabacum	cauucuauuuucuuu*ugguAGUUCgaucgugGAAUUucuuu***guuucuguauuuccggagu
Nymphaea alba	caucucauucuguu**ugguAGUUCgaccgcgGAAUUcuuuu***guuucgguauuuccggaau
Panax ginseng	cagcccauucuauuu*ugguAGUUCgaccgcgGAAUUucuuu***guuucuguauuucuggaau
Spinacia oleracea	uauuucuaucccuuu*ugguAGUUCgaucgcgGAAUUucuuu***cuuucuguauuuccggaau
Oryza nivara	cauuucuagacauuc*ugguAGUUCgaccgugGAAUU*uuuug**guuucgguaucucuggaau
Oryza sativa	cauuucuagacauuc*ugguAGUUCgaccgugGAAUU*uuuug**guuucgguaucucuggaau
Triticum aestivum	cauuucuagauauu*augguAGUUCgaccgcgGAAUUuuuuu***guuucgguaucucuggaau
Zea mays	cauuucuagacauuc*ugguAGUUCgaccguGGAUUu*uuuu***guuuugguaucucuggaau
Консенсус	gguAGUUCgaycgygGAAUU*yuuu***guuunnguauuuyyggaau

Таблица 5. Выравнивание 5’-нетранслируемых областей пред геном *clpP* хлоропластов.

Спирали выделены прописными буквами.

Chaetosphaeridium globosum	uuauaUUCUuacauuuuAGAAacua*uauauauacu**********
Anthoceros formosae	ucuuacgUUUUCauaucaGAGAGaaauuu*ccaguugauaacaagaaagaa*********
Marchantia polymorpha	UUUUAcguuuuuuuauuaUAGAAgaguau*uuuguuuguggaagaaaaaaaaa*******
Huperzia lucidula	aguuacGCUUUCguauuaGAGGGCaauau*aauauuuagccucagggaaagaagaa****
Adiantum capillus-veneris	acUUACgcccacaauucagagGUAAcucc*aagacguuuugauaauuucuau********
Psilotum nudum	auuuauGCUUUcgcaucaGAGGUuuauuc*ucuuuuuauuuucuau**********
Pinus thunbergii	CUUUAcguuuccauauUAGAGuauagugc*uucacuucuuuccauuaaaacaa*******
Amborella trichopoda	ucuuaCGUUUccgcaucaAAGUGaaauuauagugcuuaacuccuuuuuccuuc*******
Arabidopsis thaliana	uuuuaCGUUUccacaucaAAGUGaaauag*agaacuucauucucuuuuuuuuucauuuca
Atropa belladonna	uguuaCGUUUccaccucaAAGUGaaauau*aguauuuaguucuuucuuucauuuaa****
Calycanthus floridus	uauuaCGUUUccacaucaAAGUGaaguag*aguacuuaacucccuuuugaauuuca****
Cucumis sativus	uguuACGcuuccacauuaaCGUuagcuau*acuacuuacuucguuuugcuuugcuuua**
Epifagus virginiana	uguuaCGUUUccauaucaAAGUGaaauuu*aguauuuaguucuuuuuuuuuauua****
Lotus corniculatus	uaUUACgUUUucacaccAAAGUAAgauau*augacuucauuuuuucuucauuga******
Nicotiana tabacum	uguuaCGUUUccaccucaAAGUGaaauau*aguauuuaguucuuucuuucauuua*****
Nymphaea alba	uguuaCGUUUccacaucaAAGUGaaguau*aguacuuaaccccguuuucuuuaaugua**
Panax ginseng	uguuaCGUUUccacaucaAAGUGaaauau*aguacuuaauucuuuuuucuuuuuuuc***
Spinacia oleracea	uauuaCGUUUccacaucaAAGUGaaauag*aguacuuaauuuuuuucuuucauuua****
Oryza nivara	ucUUACGUUUccauauuaAAGUGUAGuuuucuuacuuaaauuuaauaauauuaaucuaau
Oryza sativa	ucUUACGUUUccauauuaAAGUGUAGuuuucuuacuuaaauuuaauaauauuaaucuaau
Triticum aestivum	ucUUACGUUUccauauuaAAGUGUGGuuuucuuacuuaaauuuaauaauauuaaucuaau
Zea mays	ucUUACGUUUccauauuaAAGUGUAGuuuuuuuacuuaaauuuaauaauauuaaucuaau
Консенсус	UUACGYUUycayauyaRAGNRNARu

Таблица 6. Выравнивание 5’-нетранслируемых областей пред геном *psaA* хлоропластов.

Спирали выделены прописными буквами.

Odontella sinensis	cuuaugagaguuucau*aaauu*****UUCgucUCCcaaaaGGAGAAaguca
Guillardia theta	auaaaguaagaguuuuuagauu*****gcugUCUCaaaagagGAGAaccuca
Porphyra purpurea	uagaaauaagcguuuu**gauu****ccuugUCUCaagagagGAGAaucuca
Nephroselmis olivacea	agccaggaagacuauuu*cauu****CCUCgugugaagaGAGGagaaucucg
Chaetosphaeridium globosum	uguuguuaaguauuuucuuagc*****CUCgUCUgaaaAGAgGAGaauuucg
Mesostigma viride	uagaggugaguuuuuuu*ugug****cCUCaUCUaaaaAGAgGAGaaucucc
Anthoceros formosae	uuguuggcggucuuuuc*caug****CCUCgucugaaagGAGGauaauaucg
Marchantia polymorpha	uguugguagguuuuucc*uaug****CCUCgucugaagaGAGGagaaccucg
Huperzia lucidula	ucuuggcgggguuuuuuc*uaug****CCUCgucuggaaaGAGGagaaccucg
Adiantum capillus-veneris	uguugguagguuguugc*uauc****cCCUgCUCgaaGAGAGGagaguccca
Psilotum nudum	ugcuggcagguuguugc*uauu****CCUCgucucgagaGAGGagaaucuca
Pinus thunbergii	uauuggcagguuucuauuauuuaagucccgUCCgaaaagaGGAgaa*uuca
Amborella trichopoda	ucuuggcgggucucuucguaug****uguugUCCggaaagaGGAgga*cuca
Arabidopsis thaliana	uguuggcggguuucuuuguaug****uguugUCCggaaagaGGAgga*cuca
Atropa belladonna	uguuggcgggucucuuuguaug****uguugUCCggaaagaGGAgga*cuca
Calycanthus floridus	uguuggcggguuucuuuguaug****uguugUCCggaaauaGGAgga*cuca
Cucumis sativus	uauuggcgggucucuuuguaug****uguugUCCggaaagaGGAgga*cuca
Lotus corniculatus	uauuggcaggucucuuuguaug****uguugUCCggaaagaGGAgga*cuca
Nicotiana tabacum	uguuggcgggucucuuuguaug****uguugUCCggaaagaGGAgga*cuca
Nymphaea alba	uguuggcgggucucuucguaug****uguugUCCggaaagaGGAgga*cuca
Panax ginseng	uguuggcgggucucuuuguaug****uguugUCCggaaagaGGAgga*cuca
Spinacia oleracea	uguuggcaggucucuuuguaug****ucuugUCCggaaagaGGAgga*cuca
Oryza nivara	aguuggcgggucucuuuguaug****ucuugUCCggaaagaGGAgga*cuua
Oryza sativa	aguuggcgggucucuuuguaug****ucuugUCCggaaagaGGAgga*cuua
Triticum aestivum	aguuggcgggucucuuuguaug****ucuugUCCggaaagaGGAgga*cuua
Zea mays	aguuggcgggucucuuuguaug****ucuugUCCggaaagaGGAgga*cuua
Консенсус	wguurgyrrguyuyuuy*uaun*****nuygUCYgraragaGGAgra*cucr

Таблица 7. Выравнивание 5’-нетранслируемых областей пред геном *psbA* хлоропластов.

Спирали выделены прописными буквами.

Odontella sinensis	aGUAAAaacuugggaagaau**********uaaaUUUAUaaaaauguuauu*
Guillardia theta	auagAUUAcuugggaaguauauuacuauaaaauu****uaaacaUAAUauaaaaca****
Cyanidioschyzon merolae	aguacuuacUUGggaagagaaaaa******aucuuucuuaaaCGAcu**********
Porphyra purpurea	agauauuguUUGGgaagcaucuua**guuaaauccuaaaaaaUUGAuaaaaaaauu*****
Gracilaria tenuistipitata	aaAUAUUUcuuGGGAaguauccuc******aauuaUCCUaaacuAAAUAUauaauauu**
Chlamydomonas reinhardtii	auauuuuacGGAGaaa**********uuaaaaCUUUaaaaaaauuaacau
Nephroselmis olivacea	caaaaaagcUUGGgagaucacuu**********acaaaaUCAAuaaauucaccuacu
Chaetosphaeridium globosum	uuauaguUUGGguaaauuauuauaa********cuuuauaaaCCAAuauauacc******
Anthoceros formosae	uuauauGCUUGGgaacuuauuauua********uuucacaaaCCAAGUuuuacc******
Marchantia polymorpha	uuaaauAUUUGGgaaacucuuaauua*******uuuuaaaaaCCAAGUuuuacu******
Huperzia lucidula	uuauauGCUUGGggaacuuauuau****cacuuuauaaaaaaCUAAGU**uuuacc****
Adiantum capillus-veneris	uuacuagccUGGggaaucacuaau*****uccuuuucgaaaaCCAa**aaauuccc****
Psilotum nudum	uuuauugacUUGGgaacuuuagaa******uugcuuuagaaaCCAA**auauuauc****
Pinus thunbergii	gccuuauGCUUgggagccucugaugauu*******uuauaaacgAAGU*ucugacc****
Arabidopsis thaliana	uuugugcgCUUGGgagucccugauuauu*******aaauaaaCCAAGgauuuuacc****
Atropa belladonna	cucgugugCUUGGgagucccugaugauu*******aaauaaaCCAAG*auuuuacc****
Calycanthus floridus	ucugugugCUUGGgagucccugaugauu*******caauaaaCCAAG*aucuuacc****
Cucumis sativus	uucaugugCUUGGgagucccugaugauuaauuauuaaauaaaCCAAG*auuuuacc****
Lotus corniculatus	uucgugugCUUGGgagucccugaugauu*******aaauaaaCCAAG*auuuuacc****
Nicotiana tabacum	cuagugugCUUGGgagucccugaugauu*******aaauaaaCCAAG*auuuuacc****
Nymphaea alba	ucugugugCUUGGgagcuccugaugauu*******gaauaaaCCAAG*aucuuacc****
Panax ginseng	uucgugugCUUGGgagucccugaugauu*******aaauaaaCCAAG*auuuuacc****
Spinacia oleracea	uuugugugCUUGGgagucccugaugauuaa**auuaaauaaaCCAAG*auuuuacc****
Oryza nivara	uacgugugCUUGGgaguccuugcaauuu*******gaauaaaCCAAG*aucuuacc****
Oryza sativa	uacgugugCUUGGgaguccuugcaauuu*******gaauaaaCCAAG*aucuuacc****
Triticum aestivum	uacguguGCUUGGggguccuugcaauuu*******gaauaaaCCAAGU*ucuuacc****
Zea mays	uacgugugCUUGGgaguccuugcaauuu*******gaauaaaCCAAG*aucuuacc****
Консенсус	YUUGGgaryyyy**********aaaCYAAG

Таблица 8. Выравнивание 5’-нетранслируемых областей пред геном *psbB* хлоропластов.

Porphyra purpurea	uuuguaagaaagucaacaaaguauguucuuauu******cauaggaggcauguaguca*
Nephroselmis olivacea	auagcaagaaagucauaccguuauaaaggucucgaau*agaccuagaaggaguauag**
Chaetosphaeridium globosum	acugcaagaaagucacaaaua*guuuguuuuuuuc*****uuaacaaagagguauuuac
Anthoceros formosae	aaugcaagaaauuuacguagu*gucuauucuucu*******ggauaaagggguaucuuc
Marchantia polymorpha	aaugcaaaaaaguuacauagc*gucuaauucucuu******ugagaaagggguauuuuu
Huperzia lucidula	aacguaagaaagucauaugau*gucuaccuaucuuugguaaggggaaaggggggacucaa
Adiantum capillus-veneris	auugcaagaaaguuacgcagugaucaguugucuccaauauucaagaaaggggguuuuuc*
Psilotum nudum	aacgcaagaaaguuacguagu*auugacuaaa**********aaaaagagguauuuaa
Pinus thunbergii	aaugugagaaaguuacauagu*gucuacuuuuuc*******cgauaaagggguguuugc
Amborella trichopoda	aaugcgauaaaguuacauagu*gucuauuuug**********augaagggguauuucc
Arabidopsis thaliana	aaugcaauaaaguuacauagu*gucuauuuuucgu******ugauaaagggguauuucc
Atropa belladonna	aaugcaauaaaguuacguaga*uuuaucuuuga**********uauaagggguauuucc
Calycanthus floridus	aaugcgauaaaguuacauagu*gucuauuuuuucu*****uugauaaagggguauuuuc
Cucumis sativus	agugcaauaaaguuacauagu*gucuauuuuuccu******ugauaaagggguauuucc
Lotus corniculatus	agugcaauaaagucacauagu*gucuauuuucugu*****ugauagaagagguauucuc
Nicotiana tabacum	aaugcaauaaaguuacguagu*gucuauuuaucuu*****ugauauaagggguauuucc
Nymphaea alba	aaugugacaaaguuacauagu*gucuauuuuucuu******ugaugaagggguauuucc
Panax ginseng	aaugcaauaaaguuacguagu*gucuuuuuuucuuug*auauaaauaagggguauuucc
Spinacia oleracea	aaugcaauaaaguuacauagu*gucauuuuucuu*******ugauaaagggguauuucc
Oryza nivara	gauaaaauaaagugacaucgu*gucuauuuuucuu******ugcuaaagggguauuucc
Oryza sativa	gauaaaauaaagugacaucgu*gucuauuuuucuu******ugcuaaagggguauuucc
Triticum aestivum	gauaaaauaaagcgacaucgu*gucuauuuuucuu******ugcuaaagggguauuucc
Zea mays	gauaaaauaaagcgacaucgu*gucuauuuuucuu******ugcuaaagggguauuucc
Консенсус	aaagunacruagu*gucuayuunn**********nnnaaggggguruuu

Фотосистема II. В 5'-нетранслируемых областях генов *psbA* (белок D1) и *psbB* (P680 хлорофилл A) у многих хлоропластов найдены консервативные участки, примыкающие к инициирующему кодону AUG гена, с консенсусами

YUUGGGARYYYY**********NAAACYAAG

(см. табл. 7) и

AAAGUNACRUAGU*GUCUAYUUNN**********NNNAAGGGGURUUU

(см. табл. 8). Исключение составляет ген *psbB* из *Adiantum capillus-veneris*, инициирующий кодон которого ACG подвергается редактированию. Перед геном *psbA* расположены консервативные шпильки. В *Mesostigma viride* 5'-нетранслируемые области обоих генов *psbA* и *psbB* не содержат рассматриваемого сигнала.

Обсуждение. У многих хлоропластов гены белков имеют интроны, или мРНК образуется транс-сплайсингом из разных транскриптов. Поэтому трансляция не может происходить сразу после транскрипции. Однако аппарат трансляции у хлоропластов очень близок к таковому у бактерий, у которых рибосома движется вдоль мРНК сразу после РНК-полимеразы. Более того, дойдя до края экзона, рибосома перекрыла бы порядка 10 нуклеотидов последующего интрона.

Вероятно, найденные консервативные участки перед генами *atpF* и *petB* связаны с регуляцией трансляции, предотвращая начало трансляции до завершения сплайсинга. Сейчас нет достаточных оснований, чтобы решить, основана ли эта регуляция на связывании протеина с мРНК. Перед геном *petB* консервативный участок не содержит очевидного сайта связывания рибосомы, но имеется консервативная шпилька, что позволяет предположить: после транскрипции происходит модификация лидерной области гена или сюда присоединяется белок, активирующий трансляцию.

Вероятно, найденные перед генами *clpP*, *psbA* и *psbB* консервативные участки мРНК связаны с регуляцией трансляции. Роль консервативного

участка перед геном *psaA* менее понятна. Возможно, AG-богатая 5'-область связана со стабилизацией мРНК.

Консервативные участки у 5'-нетранслируемых областей генов *clpP* и *psbA* найдены перед всеми ортологами этих генов, содержащих интроны, и перед некоторыми ортологами этих генов, не содержащих интроны. Подобная регуляция трансляции гена *psbA* белка D1 фотосистемы II экспериментально показана, например, у *Chlamydomonas reinhardtii*, где транскрипция происходит непрерывно, а трансляция активируется на свету некоторым белком с массой 47 кДа, который связывает мРНК в комплексе с другими белками, напрямую не связывающими мРНК, [Hauser, Gillham, Boynton, 1996]. Этот комплекс инактивируется в темноте.

Консервативность рассматриваемой структуры у растений и водорослей позволяет предположить, что зависимая от света регуляция трансляции гена *psbA* сформировалась на ранних стадиях эволюции, до появления интронов в генах белков и до расхождения зелёных и пурпурных водорослей.

Отметим, что обычно консервативная область РНК содержит спираль с консервативными плечами, действующая совместно с белком-медиатором, что характерно для большого числа регуляторных элементов.

Во всех случаях у *Adiantum capillus-veneris* соответствующая 5'-нетранслируемая область значительно дивергировала. И именно для этого хлоропласта характерно частое редактирование РНК.

Не удалось обнаружить общих консервативных участков в нетранслируемых областях РНК у рассматриваемых хлоропластов и у хлоропластов из *Euglena gracilis*.

Хотя ген *ycf3*, кодирующий связанный с первой фотосистемой белок Ycf3, имеет интроны и протяжённую 5'-нетранслируемую область, которая не перекрывается другими генами в хлоропластах растений, алгоритм не нашёл здесь длинного консервативного участка.

Небольшой консервативный участок с консенсусом ARGGAGGGACYU непосредственно перед геном *rbcL* у высших растений включает сайт связывания рибосомы и нет основания предполагать здесь сайт связывания регуляторного белка. Этот ген *rbcL* имеет интроны в хлоропластах *Euglena gracilis* и *Chlamydomonas reinhardtii*, где его трансляция регулируется белками, взаимодействующими с мРНК. Однако, в последнем случае состав 5'-нетранслируемой области этого гена совершенно другой, чем у высших растений, что может быть связано с отсутствием интронов в гене *rbcL* у рассматриваемых растений.

2.2 Различные системы регуляции экспрессии генов биосинтеза аминокислот и аминоацил-тРНК синтетаз у актинобактерий

Материалы. Геномы актинобактерий *Actinomyces naeslundii* (An), *Bifidobacterium longum* (Bl), *Corynebacterium diphtheriae* (Cd), *Corynebacterium efficiens* (Ce), *Corynebacterium glutamicum* (Cg), *Kineococcus radiotolerans* (Kr), *Leifsonia xyli* (Lx), *Mycobacterium avium* (Ma), *Mycobacterium bovis* (Mb), *Mycobacterium leprae* (Ml), *Mycobacterium marinum* (Mm), *Mycobacterium smegmatis* (Ms), *Mycobacterium tuberculosis* (Rv and Mt), *Nocardia farcinica* (Nf), *Propionibacterium acnes* (Pa), *Rubrobacter xylanophilus* (Rx), *Streptomyces avermitilis* (Sa), *Streptomyces coelicolor* (Sc), *Thermobifida fusca* (Tf), *Tropheryma whipplei* (Tw) получены из ГенБанка. Кроме того, использована последовательность нуклеотидов из *Streptomyces venezuelae* (Sv) [Lin, Pradkar, Vining, 1998].

Результаты. Применение алгоритма поиска консервативных участков к 5'-нетранслируемым областям генов позволило выделить протяжённые консервативные участки, имеющие сложную консервативную вторичную альтернативную структуру РНК. Кратко результаты перечислены в табл. 9.

Таблица 9. Распределение предсказанных регуляторных структур РНК у актинобактерий.

"A" означает классическую аттенюаторную регуляцию, "R" – Rho-зависимую аттенюаторную регуляцию, "LEU" – регуляцию на уровне трансляции при участии LEU-элемента, "T" – регуляцию на уровне трансляции при участии T-бокса.

Род	триптофан	цистеин		лейцин		изолейцин
	trp	*cys*	*cbs*	*leuA*	*leuS*	*ileS*
Actinomyces				LEU		T
Bifidobacterium			R			T
Corynebacterium	A			LEU		T
Kineococcus				LEU		T
Leifsonia				LEU		
Mycobacterium		R		LEU		T
Nocardia				LEU		T
Propionibacterium		R				T
Rubrobacter						T
Streptomyces	A			LEU	A	T
Thermobifida				LEU		
Tropheryma						

Таблица 10. Аттенюаторная регуляция *trp* оперонов у актинобактерий

a) Координаты *trp* генов

Вид	Локус	Ген	Координаты гена	Белок
C. diphtheriae	NC_002935	*trpB1*	2456701..2458032	NP_940652
		trpB2	2465139..2466365	NP_940660
C. efficiens	NC_004369	*trpE*	3052837..3054504	NP_739478
C. glutamicum	NC_003450	*trpE*	3233404..3234960	NP_602223
S. avermitilis	NC_003155	*trpS2*	complement(5757496..5758491)	NP_825902
		trpE1	complement(7320283..7322268)	NP_827260
S. coelicolor	NC_003888	*trpE*	2276703..2278607	NP_626374

b) Оперонная структура и лидерные пептиды

Вид	Оперон	Лидерный пептид
C. diphtheriae	*trpB1EGDC1*	2456514 ******MNAHN**WWW**RA***** 2456543
C. diphtheriae	*trpB2A*	2464983 *****MNAAFKF**WW**RA***** 2465015
C. efficiens	*trpEGDCBA*	3052621 VNNFCQSQGTQ**WWW**RAR**** 3052671
C. glutamicum	*trpEGDCBA*	3233152 VNNSCLSQSTQ**WWW**RAN**** 3233199
S. avermitilis	*trpS2*	5758647 ***MTTRTCTQQ**WW**AA***** 5758609
S. avermitilis	*trpE1*	7322414 ***MFAHSIQN**WWW**TAHPAAH 7322361
S. coelicolor	*trpE*	2276540 ***MFAHSTRN**WWW**TAHPAAH 2276593

c) Выравнивание аттенюаторов. Антитерминатор подчёркнут, терминатор выделен прописными буквами. Полужирным – UGG и стоп-кодон.

Вид	Оперон	РНК
C. diphtheriae	*trpB1EGDC1*	**ugguggugg**cgcgcu**uaa**cc*gcgggcc*guuuu***cacgcauucauuuc**********aac**AGGCUCGCCUUGUcca****AC*AAGcaGCGGGCCUuuuuguuagc
C. diphtheriae	*trpB2A*	uuc**uggugg**cgcgcc**uag**caggcgggccccuuuugugugagcauucaccacacaacuuuuggaaacacAAGCCCGCGuau**********C*GCGGGCUUuuucguauau
C. efficiens	*trpEGDCBA*	**ugguggugg**cgcgcuaga**uaa**gcgggcccacggaucaccaaguuguuuucacacugaagauuu***cAAGGCUCGuguaCUUCGUucgACGAAGcagCGGGCCUUuu*gugguu
C. glutamicum	*trpEGDCBA*	**ugguggugg**cgcgcuaac**uaa**gcgagccugacaccucaaguuguuuucacuu**ugaugaauuuuuuuAAGGCUCGu**aCUUCGUucgACGAAGaagCGGGCCUUuu*gugguu
S. avermitilis	*trpS2*	cag**uggugg**gccgcc**uga***cggcg*gccguacacacguauguacuc**********AACGGC*CGCCGccu**********CGGCGGCCGUUcucguuucu
S. avermitilis	*trpE1*	**ugguggugg**accgcucauccggcg*gcccac**uga**cugcgcgu*********acgcaagacuuCGCGAAGGC*CGCCC**********gagGGGCGGCCUUuCGUGuuucc
S. coelicolor	*trpE*	**ugguggugg**accgcucacccggcg*gcccac**uga**cugcgcgcg*********acucaagacucgCGAAGGC*CGCCC**********gagGGGCGGCCUUCGguguuuuc
S. venezuelae	*trpE*	**ugguggugg**accgcucacccggcg*gcccac**uga**ucgcgcgu********acacggaucacacgcacAGGC*CGCCC**********gagGGGCGGCCUuucucg

Таблица 11. Аттенюаторная регуляция *cys* и *cbs* оперонов у актинобактерий

a) Координаты *cys* и *cbs* генов

Вид	Локус	Ген	Координаты гена	Белок
M. avium	NC_002944	*MAP2122*	2351330..2352622	NP_961056
M. bovis	NC_002945	*cysK1*	2586392..2587324	NP_856011
M. tub CDC1551	NC_002755	*cysK*	2604640..2605572	NP_336875
M. tub H37Rv	NC_000962	*cysK*	2608794..2609726	NP_216850
M. leprae	NC_002677	*ML0840*	complement(997285..998589)	NP_301634
M. marinum	Sanger_216594	*cysK*	complement(136548..137477)	
P. acnes	NC_006085	*cysK*	1047389..1048324	YP_055674
B. longum	NC_004307	*cbs*	1006495..1007721	NP_696325

b) Оперонная структура и лидерные пептиды

Вид	Оперон	Лидерный пептид
M. avium	*XcysKE*	2351124 MQHRLQPRFAPSRCLVVACCCCCCR 2351177
M. bovis	*cysK1E*	2586122 MQQAIQLRFILPRRLAVGCCCC*** 2586187
M. tub CDC1551	*cysKE*	2604371 MQQAIQLRFILPRRLAVGCCCC*** 2604436
M. tub H37Rv	*cysKE*	2608526 MQQAIQLRFILPRRLAVGCCCC*** 2608591
M. leprae	*XcysKE*	0998791 MHQSTQPRFVFTRRFTVDCYCRCC* 0998742
M. marinum	*cysKE*	0138059 MQQAAQLSFVLTRCPAVDCCCC*** 0137994
P. acnes	*cysK*	1047061 MTSAMMVCICRCCC* 1047102
B. longum	*cbs*	1007876 MQIISCCCR* 1007850

Таблица 12. Аттенюаторная регуляция *cys* оперонов у микобактерий.
Полужирным выделены инициирующий кодон, цистеиновые кодоны и стоп-кодон. Прописными буквами выделены повторяющиеся UC-богатые участки, характерные для области связывания белка Rho.
Нуклеотидные последовательности для *M. bovis* и *M. tuberculosis* совпадают.

	RBS Старт
M. avium	uauaguggugac**aug**caacaccgccuacagccgcgcuuu
M. bovis, tub	uauaguggg**aug**caacaggccauacagcugcgcuuu
M. leprae	uauaguggaccu**aug**caucaguccacacagccacgcuuu
M. marinum	uauaguagagcc**aug**caacaggccgcacagcugagcuuu
	Cys кодоны
M. avium	gccccgucgcgcugccuugucguggcc**uguuguugcuguuguuguc**gu
M. bovis, tub	auccucccgcgccgccucgccgugggc**uguuguuguuguu***********
M. leprae	gucuuuacgcgccgcuuuaccguggac**ugu**uau**ugu**cgc**uguugc***
M. marinum	guccucacgcgcugccccgccguggac**uguuguuguuugcugu********
	Стоп и область связывания белка Rho
M. avium	**ugA**UUUCCgcaaGCCCUCugacgcuguagaaAUCCCCgcgcucGCCCCUgcccg
M. bovis, tub	**ugA**UUCCUg*gcguccacagcaAUUCCUcgcGCUCUUgcccg
M. leprae	**ugA**UUCCUgac*ACCUUUuaacGCUCUCagcaaaucauucacGUUCUCgccua
M. marinum	**ugA**UUCCUgac*gcguucugaccguccaguaaucgucGCCUCUgucgccucaugg

Биосинтез триптофана. Найдена классическая аттенюация для оперонов биосинтеза триптофана во всех *Corynebacterium* spp. и *Streptomyces* spp. У *C. diphtheriae* классическая аттенюация предсказана для двух оперонов $trpB_1EDGC$ и $trpB_2A$. У *S. avermitilis* такая аттенюация предсказана для триптофанил-тРНК синтетазы $trpS_2$. Аттенюаторы приведены в табл. 10.

Лидерные пептиды перед *trp* оперонами имеют двойной или тройной повтор регуляторного кодона UGG. Все терминаторы содержат консервативную GC-богатую шпильку, за которой следует участок остатков урацила. Шпильки антитерминатора и терминатора во всех случаях содержат комплементарную тройку gGCC-rGCy-GGCC, где абсолютно консервативные нуклеотиды показаны прописными буквами. Эти аттенюаторы аналогичны таковым у протеобактерий.

Биосинтез цистеина. 5'-области оперонов *cys* у всех *Mycobacterium* spp. кроме *M. smegmatis*, у *P. acnes*, и оперона *cbs* у *B. longum* содержат открытую рамку считывания с последовательностью цистеиновых кодонов непосредственно перед стоп кодоном, см. табл. 11. Возможно, регуляция транскрипции основана здесь на Rho-зависимой терминации. Эта ситуация аналогична той, которая известна для триптофаназы *tna* у *E. coli* [Konan, Yanofsky, 2000], [Gong, Yanofsky, 2003]. Отметим, что рассматриваемые геномы содержат гены *rho, nusG, nusA, nusB*, кодирующие белки, участвующие в Rho-зависимой терминации. 3'-нетранслируемая область предполагаемого лидерного пептида содержит UC-богатый мотив, характерный для связывания белка Rho с РНК, см. табл. 12. Дополнительным подтверждением служит то, что стоп-кодоном предполагаемого лидерного пептида служит именно тот из трёх возможных кодонов, который часто служит стоп-кодоном какого-либо гена из генома. Наиболее ярко это проявилось у *P. acnes*, см. табл. 13.

Итак, предполагаемая схема регуляции такова:

- При *недостатке* цистеина участок мРНК вокруг стоп кодона лидерного пептида закрыт рибосомой длительное время, за которое РНК полимераза успевает уйти далеко и транскрипция не прерывается.
- При *избытке* цистеина рибосома быстро завершает трансляцию лидерного пептида, после чего открывается следующий за ним UC-богатый участок РНК, характерный для Rho-зависимого терминатора. Транскрипция прерывается.

Табл. 13. Частоты стоп-кодонов у рассматриваемых актинобактерий.
Частота стоп-кодон лидерного пептида выделена полужирным шрифтом.

Вид	UGA	UAG	UAA
Bifidobacterium longum	56	20	**24**
Mycobacterium avium	**63**	26	11
Mycobacterium bovis	**54**	30	16
Mycobacterium leprae	**46**	30	24
Mycobacterium tuberculosis	**55**	29	16
Propionibacterium acnes	**81**	7	12

Первыми генами оперонов биосинтеза цистеина в *M. avium* и *M. leprae* являются гипотетические гены *X*. Соответствующие им протеины (MAP2122 в первом случае и ML0840 во втором) ортологичны друг другу. Они идентичны между собой на 62% процента аминокислот и не имеют других близких гомологов.

Биосинтез лейцина. Перед генами *leuA* 2-изопропилмалат синтазы у большинства актинобактерий (*A. naeslundii, Corynebacterium spp., K. radiotolerans, L. xyli, Mycobacterium spp., N. farcinica, Streptomyces spp., T. fusca*, см. табл. 14) в 5'-нетранслируемой области содержится характерная консервативная структура, названная LEU-элементом. А именно, LEU элемент – это: короткая открытая рамка считывания (лидерный пептид, см. табл. 15) с участком лейциновых кодонов, вторичная структура РНК с

псевдоузлом, плечи спиралей которой высоко консервативны по нуклеотидному составу, и альтернативная к ней структура, в которой псевдоузел не образуется. Эти две альтернативные структуры лежат в петле спирали, 5'-плечо которой перекрывает область лейциновых кодонов лидерного пептида, а 3'-плечо перекрывает область Шайна-Дальгарно гена *leuA*.

Таблица 14. Координаты генов *leuA* перед которыми найден LEU-элемент.

Вид	Локус	Координаты гена	Белок
C. diphtheriae	NC_002935	complement(228555..230372)	NP_938656
C. efficiens	NC_004369	complement(233589..235439) (добавлено 35 кодонов)	NP_736826
C. glutamicum	NC_003450	complement(266151..268001)	NP_599502
K. radiotolerans	AAEF02000060	complement(3238..4965)	EAM73829
M. avium	NC_002944	333789..335633	NP_959246
M. bovis	NC_002945	4091088..4093193	NP_857375
M. tub CDC1551	NC_002755	4145949..4147928	NP_338367
M. tub H37Rv	NC_000962	4153737..4155671	NP_218227
M. leprae	NC_002677	2754640..2756463	NP_302512
M. marinum	Sanger_216594	192528..194345	
M. smegmatis	TIGR_246196	6334690..6336495	
S. avermitilis	NC_003155	6774328..6776049	NP_826778
S. coelicolor	NC_003888	complement(2725480..2727201)	NP_733575
T. fusca	NZ_AAAQ 02000002	349237..350943 (добавлено 9 кодонов)	ZP_ 00293601
L. xyli	NC_006087	complement(1501628..1503400)	YP_062368
N. farcinica	NC_006361	complement(322994..324787)	YP_116514
A. naeslundii	TIGR_240017	594374..596211	

Таблица 15. Лидерные пептиды перед генами *leuA* у актинобактерий.

	Вид	Лидерный пептид		
Cd	*C. diphtheriae*	230506	MNRANLLLLRRGGSQA*	230459
Ce	*C. efficiens*	235612	MFSSHERSALLLRRGGSQRS	235553
Cg	*C. glutamicum*	268124	MTSRANLLLLRRGGSQRS	268095
Kr	*K. radiotolerans*	5097	VARLENLLLRRRGGAS*	5050
Ma	*M. avium*	333705	VADVQRVLLLGRRDGV**	333752
Mb	*M. bovis*	4090959	VLHVQRVLLLGRRDGV**	4091006
Mt	*M. tub CDC1551*	4145866	VLHVQRVLLLGRRDGV**	4145913
	M. tub H37Rv	4153611	VLHVQRVLLLGRRDGV**	4153658
Ml	*M. leprae*	2754521	VQQVLLLERRDGV**	2754559
Mm	*M. marinum*	192399	VLCVQRVLLLGRRDG***	192443
Ms	*M. smegmatis*	6334564	VLGVQRVLLLGRRGGV**	6334611
Sa	*S. avermitilis*	6774199	MRFGLLLLSCRGEGL*	6774243
Sc	*S. coelicolor*	2727361	MRFGLLLLSCRGEGL*	2727317
Tf	*T. fusca*	349104	MLRELLLLSGRGGGR*	349148
Lx	*L. xyli*	1503533	MRVTLGLVYGLILLSCRDES**	1503474
Nf	*N. farcinica*	324906	MQRALLLGRRDGV**	324868
An	*A. naeslundii*	594266	VSLLLSRRGGA**	594298

Множественные выравнивания LEU-элементов, построенные на основе поиска консервативных участков РНК, приведены в табл. 16 и 17.

Таблица 16. Структура LEU-элемента с псевдоузлом РНК.

Курсивом выделена область Шайна-Дальгарно RBS. Плечи спиралей, образующих псевдоузел выделены подчёркиванием и двойным подчёркиванием. Черенок LEU-элемента, перекрывающий область Шайна-Дальгарно, выделен прописными буквами.

```
Cd cuucUCCUUCuu******cgccgcggcgggucacaggcuuaacgucccuua
Ce GCUCuUCUUcuu******cgccgcggcgggucccagaggucauaa******
Cg cuacuUCUUCUU******cgccgcggcgggucccagaggucuuaa******
Kr aacCUCCUCCUUC***gucgccgcggcggggccag***************
Ma cgggUGCUCCuccucggacgccgcgacggggucugauu************
Mb cgggUGCUCCuccucggacgccgcgacggggucugau*************
Ml cagguaCUCCuccucgaacgccgcgacggggucugau*************
Mm cgggUGCUCCuccucggacgccgcgacggggccugau*************
Ms cgggguGCUCCuuucucggacgccgcggcggg*******gucuga*******
Sa gggcuGCUCCUCcuuagcugccgcggcgaggggcuguaag***********
Sc GGGCUgCUUCuccuuagcugccgcggcgaggggcuguag************
Tf gagcuGCUCCUGcuuagcggccgcggcgggggccgauaa************
Lx ggcCUGAUUCUCCuuagcugccgcgacgaauccuaag**************
Nf cgggcucUUCUUCucggcccgccgacggggucugau**************
An gugagcCUCCUGcuuagucgccgcggcggggccuga**************
Bl ggcguggaUCUGGagggcggccgcgacgug**cugggc*************

Cd cacacagccggcuc*cccgucgcggaguucu******aguguagccggcug********
Ce ***gcgaccggcac*cccgucgcggaguuu********guguugccggucgugaacccg
Cg **cacgaccggcau*cccgucgcggaguuu*******gguguugccggucgug******
Kr **cuaggccggucuccccgucgcgggaccucgucgu**gcg*cgccggcc*********
Ma **ccagaccggcuu*cccgucgcgggu*guucgc***gaug*cgccggucug*******
Mb **ccagaccggcuu*cccgucgcgggacguucgc***gaug*cgccggucug*******
Ml *cccagaccggcug*cccguuguggaa*guucacu***aug*cgccggucug*******
Mm **ccagaccggcuu*cccgucgcggg*uguucg***cgaug*cgccggucugaag****
Ms **ucagaccggcuu*cccgucgcggg*uguuu**cgcgaug*cgccggucga*******
Sa *cagaggccgaccccucccgcgg**agucuggcguugcgccgucggccg********
Sc ****aggccgacucccucccgcgg**agcuuggu**ggugccgucggccguccuuccg
Tf ****gggccggcucccucgccgcggagguucgaccugucugcugucggccg********
Lx uuccgggcc***uccuucgucgcgg*aguucguc*********guuggcucuccc****
Nf ***cggaccggc**ucccgucgcgg*gguu******aagccgugccggucgaccc****
An ***caggccggcaccccgaccgcggcugacucgu*****ccugcucggccacguucgcg
Bl aucugggc**gucg*ccgccgcggaggg*cgcacgcuauuggcugucggugcucac**

Cd ****************************caacaagaacccacguGAAGGAaacuacca
Ce caacagcgcuagaguuugauuccagaaaacaagcgcacacuccacgaAAGAuGAGCacccauc
Cg ***************************gacccacccaaaacuuuuuAAGAAGGuugaacaca
Kr ********************gccgcaccagccgcugaagaccgcGAACGAGGAGaacgaa
Ma ******************************agguuccuucugauaucccccGGAGCAaucacc
Mb ******************************agguuccuucucaccauccccGGAGCAacuacc
Ml ******************************agguuccuucucacauc*ccGGAGcaauuauu
Mm ******************************uuccuucucgccacccccGGAGCAacuacc
Ms ******************************guccc**guccaacucccGGAGCcaagaacuu
Sa ******************************uccuuccggacaccacGAGGAGCccacgcauc
Sc *****gacacgcggacgacgcggacaccgccgagauccgcggacaucacGAGGAGCCCacgccauc
Tf ************************cacgaccgcaagaaaaagucucaCGGGAGCguauucac
Lx ***************************gaccagaccgcgaaGAGAuAUCGGacc
Nf ***************************auuacugggauuccaccaacccuGGAGAAuugc
An ************************************gccgcguuccuCAGGAGucag
Bl *********************************cgagcugaagaaCCGGGgc
```

Таблица 17. Альтернативная структура LEU-элемента без псевдоузла РНК.

Курсивом выделена область Шайна-Дальгарно RBS. Плечи спиралей, образующих псевдоузел выделены подчёркиванием и двойным подчёркиванием. Черенок LEU-элемента, перекрывающий область Шайна-Дальгарно, выделен прописными буквами.

```
Cd cuucUCCUUCuu******cgccgcggcgggucacaggcuuaacgucccuua
Ce GCUCuUCUUcuu******cgccgcggcgggucccagaggucauaa******
Cg cuacuUCUUCUU******cgccgcggcgggucccagaggucuuaa******
Kr aacCUCCUCCUUC***gucgccgcggcggggccag***************
Ma cgggUGCUCCuccucggacgccgcgacggggucugauu************
Mb cgggUGCUCCuccucggacgccgcgacggggucugau*************
Ml cagguaCUCCuccucgaacgccgcgacggggucugau*************
Mm cgggUGCUCCuccucggacgccgcgacggggccugau*************
Ms cggguGCUCCuuucucggacgccgcggcggg*******gucuga*******
Sa gggcuGCUCCUCcuuagcugccgcggcgagggccuguaag**********
Sc GGGCUgCUUCuccuuagcugccgcggcgagggccuguag***********
Tf gagcuGCUCCUGcuuagcggccgcggcgggggccgauaa***********
Lx ggcCUGAUUCUCCuuagcugccgcgacgaauccuaag**************
Nf cgggcucUUCUUCucggccgccgcgacggggucugau**************
An gugagcCUCCUGcuuagucgccgcggcgggggccuga**************
Bl ggcguggaUCUGGagggcggccgcgacgug**cugggc*************

Cd cacacagccggcuc*cccgucgcggaguucu******aguguagccggcug********
Ce ***gcgaccggcac*cccgucgcggaguuu********guguugccggucgugaacccg
Cg **cacgaccggcau*cccgucgcggaguuu*******gguguugccggucgug******
Kr **cuaggccggucucccgucgcgggaccucgucgu**gcg*cgccggcc*********
Ma **ccagaccggcuu*cccgucgcgggu*guucgc***gaug*cgccggucug*******
Mb **ccagaccggcuu*cccgucgcgggacguucgc***gaug*cgccggucug*******
Ml *ccagaccggcug*ccgguuguggaa*guucacu***aug*cgccggucug*******
Mm **ccagaccggcuu*cccgucgcggg*uguucg***cgaug*cgccggucugaag****
Ms **ucagaccggcuu*cccgucgcggg*uguuu**cgcgaug*cgccggucga*******
Sa *cagaggccgacccccucccccgcggag**ucuggcguugcgccgucggccg********
Sc ****aggccgacuccccucccgcggag**cuuggu**ggugccgucggccguccuuccg
Tf ****gggccggcuccccucgccgcggagguucgaccugucugcugucggccg********
Lx uuccgggcc***uccuucgucgcggag*uucguc*********guuggcucuccc*****
Nf ***cggaccggc**ucccgucgcgggg*uu******aagccgugccggucgaccc*****
An ***caggccggcacccccgaccgcggcugacucgu*****ccugcucggccacguucgcg
Bl aucugggc**gucg*ccccgccgcggaggg*cgcacgcuauuggcugucggugcucac**

Cd ****************************caacaagaacccacguGAAGGAaacuacca
Ce caacagcgcuagaguuugauuccagaaaacaagcgcacacuccacgaAAGAuGAGCacccauc
Cg ***************************gacccacccaaaacuuuuuAAGAAGGuugaacaca
Kr ********************gccgcaccagccgcugaagaccgcGAACGAGGAGaacgaa
Ma ******************************agguuccuucugauaucccccGGAGCAaucacc
Mb ******************************agguuccuucucaccaucccGGAGCAacuacc
Ml ******************************agguuccuucucacauc*ccGGAGcaauuauu
Mm ******************************uuccuucucgccacccccGGAGCAacuacc
Ms ******************************guccc**guccaacucccGGAGCcaagaacuu
Sa ******************************uccuuccggacaccacGAGGAGCccacgcauc
Sc *****gacacgcggacgacgcggacaccgccgagauccgcggacaucacGAGGAGCCCacgccauc
Tf ************************cacgaccgcaagaaaaagucucaCGGGAGCguauucac
Lx ***************************gaccagaccgcgaaGAGAuAUCGGacc
Nf ***************************auuacugggauuccaccaacccuGGAGAAuugc
An ************************************gccgcguuccuCAGGAGucag
Bl *********************************cgagcugaagaaCCGGGgc
```

Для *M. bovis* эти альтернативные структуры LEU-элемента показаны на рис. 7.

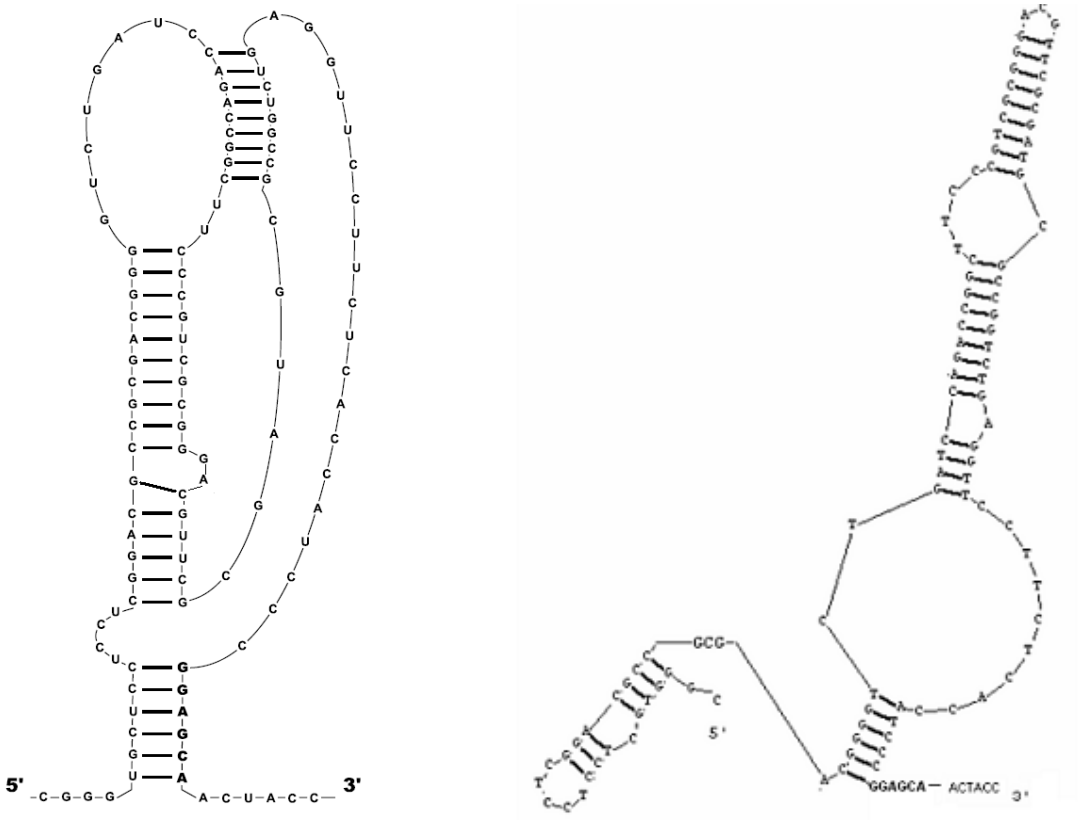

Рис. 7. Альтернативная структура LEU-элемента с псевдоузлом (слева) и без него (справа) у *M. bovis*.

Высокая консервативность участков, которые перекрывают область Шайна-Дальгарно, позволила уточнить положение инициирующего кодона гена *leuA* у *C. efficiens* и *T. fusca*. Это предположение хорошо согласуется со множественным выравниванием соответствующих белков. Таким образом, изучение регуляторных механизмов дополняет другие методы определения начала гена [Baytalulk, Gelfand, Mironov, 2002]

Рассмотрим гипотезу, относящуюся к LEU-элементу. В случае псевдоузла черенок стабилен, а в случае его альтернативы, которая отделена от черенка большими выпячиваниями, черенок нестабилен и не происходит

перекрытия области Шайна-Дальгарно. LEU-элемент обнаружен также внутри открытой рамки считывания гипотетической транспосазы из *B. longum.* LEU-элемент содержит сравнительно мало консервативных нуклеотидов, хотя стабильность такой структуры должна избирательно зависеть от концентрации лейцина и не зависеть от концентраций изолейцина и валина. Возможно, в регуляции принимает участие протеин, который, образуя комплекс с лейцином, формирует псевдоузел LEU-элемента. Филогенетический профиль, наиболее близкий к филогенетическому профилю LEU-элемента, имеют гомологи гипотетического белка ML1624 (596 аа) из *M. leprae*. Гомологи этого белка с качеством выравнивания *E* меньшем, чем 10^{-170}, найдены у всех актинобактерий, которые имеют LEU-элемент перед геном *leuA*. И в тоже время достаточно близкие гомологи этого протеина отсутствуют во всех остальных бактериях. Хотя гомолог с качеством выравнивания *E* равным примерно 10^{-108} имеется у *P. acnes*, не имеющей LEU-элемента, но это единственное исключение и в этом случае гомология заметно меньше. У протеина ML1624c помощью базы PFAM найден N-концевой домен (аминокислоты с 34 по 193), характерный для DEAD/DEAH бокса хеликаз. Этот домен характерен для многих протеинов, вовлечённых в метаболизм РНК, включая транскрипцию, трансляцию, распад РНК и образование рибосомальных РНК. Можно предположить, что этот протеин участвует в образовании псевдоузла в LEU-элементе оперона *leuA*, и в результате секвестирование области Шайна-Дальгарно регулируется концентрацией лейцина. Во всех случаях, кроме *B. longum*, 5'-конец LEU-элемента включает открытую рамку считывания с последовательностью лейциновых кодонов. Её роль в регуляции не вполне ясна. Мутации этих лейциновых кодонов в *S. coelicolor* [Craster, Potter, Baumberg, 1999], как будто, не оказывают влияние на уровень экспрессии гена *leuA*. Но именно в этом случае черенок LEU-элемента смещён и захватывает соседние нуклеотиды, не входящие в состав лейциновых кодонов. Другая особенность LEU-элемента состоит в его присутствии внутри транспосазы из *B. longum.*

Возможно, это объясняется многочисленными горизонтальными переносами LEU-элемента в составе транспозона у актинобактерий: у *B. longum* эта транспосаза сохранилась, а у других актинобактерий она эволюционно трансформировалась в регуляторный элемент перед геном *leuA*.

Биосинтез разветвлённых аминокислот. 5'-нетранслируемые области генов *ilvB*, кодирующих большую субъединицу ацетолактат синтазы (часто в составе оперонов *ilvBNC* или *ilvBHC*, где *ilvN* и *ilvH* кодируют малую субъединицу ацетолактат синтазы, *ilvC* кодирует кетол-ацид редуктоизомеразу) у видов из родов *Corynebactecterium, Mycobacterium, Streptomyces* содержат открытую рамку считывания с повтором кодонов изолейцина, лейцина и валина, за которой следует консервативный терминатор транскрипции, они указанны в табл. 18.

Таблица 18. Лидерные пептиды и терминаторы транскрипции перед *ilv* опероном у актинобактерий

а) Координаты гена *ilvB*.

Вид	Локус	Координаты гена *ilvB*	Белок
C. diphtheriae	NC_002935	1082013..1083971	NP_939459
C. efficiens	NC_004369	1432330..1434327	NP_737975
C. glutamicum	NC_003450	1338131..1340011	NP_600493
M. avium	NC_002944	complement(3379032..3380900)	NP_961972
M. tub H37Rv	NC_000962	complement(3361127..3362983)	NP_217519
M. tub CDC1551	NC_002755	complement(3355506..3357362)	NP_337598
M. bovis	NC_002945	complement(3317745..3319601)	NP_856673
M. leprae	NC_002677	complement(2044335..2046212)	NP_302166
M. marinum	Sanger_216594	complement(164709..166565)	
S. avermitilis	NC_003155	complement(3354433..3356283)	NP_823909
S. coelicolor	NC_003888	6003117..6004958	NP_629647

b) Лидерные пептиды

Вид	Лидерный пептид		
C. diphtheriae	1081747	MNIIRLVVITTRRLP	1081791
C. efficiens	1432212	MTSIRPVVIVAARRLP*	1432259
C. glutamicum	1337840	MTIIRLVVVTARRLP	1337884
M. avium	3381051	MLVVI*RRVGA	3381022
M. tuberculosis H37Rv	3363152	MDKAGKPGMLVVIGRRVGA	3363096
M. tuberculosis CDC1551	3357528	MDKAGKPGMLVVIGRRVGA	3357472
M. bovis	3319767	MDKAGKPGMLVVIGRRVGA	3319711
M. leprae	2046378	MLVVICQRVGG	2046346
M. marinum	166742	MDTAGTPGKLVVLGRRVVA	166686
S. avermitilis	3356481	MRTRILVLGKRVG	3356443
S. coelicolor	6002909	MRTRILVLGKRVG	6002947

c) Терминаторы транскрипции перед *ilv* опероном. Шпильки выделены прописными буквами. Нуклеотидные последовательности у *M. tuberculosis* H37Rv, *M. tuberculosis* CDC1551 и *M. bovis* совпадают.

Вид	Терминатор
C. diphtheriae	aaaagcG***CCCUCGaCAG****CAccacacaUGCUGAGCGGGGGCuuuccuuau
C. efficiens	caa*gcG***CCCUCGACAGUACccaccacaGUGCUGuuUCGAGGGCuuuguugu*
C.glutamicum	caa*gcG***CCCUCGaCAACACUcaccacAGUGUUGgaaCGAGGGCuuucuuguu
M. avium	caacgcgcAACCCUCGugCAGCaca*********aGCUGuCG*GGGGUUuuuuguu
M.tuberculosis	caacgcg**ACCCUCGugCAGCagc*******ugaGCUGgCGA*GGGUuuuuucuu
M. bovis	caacgcg**ACCCUCGugCAGCagc*******ugaGCUGgCGA*GGGUuuuuucuu
M. leprae	caacgcgcAACCCUCGugCAGCUag*******ucAGCUGuCGA*GGGUUuuuuguu
M. marinum	caacgcgcAACCCUCGUgCAGCagc*******ugaGCUGACG*GGGGUUuuuuguu
S. avermitilis	cggcgcgcuCCCCUCGcuUGCC**********ucacGGCACGAGGGGuuuuuuguu
S. coelicolor	cgacgcgcuCCCCUCGcuUGCC**********uuacGGCACGAGGGGuuuuuuguu

Найденные консервативные элементы характерны для классической аттенюации, но в данном случае характер регуляции не вполне понятен. Согласно экспериментальным данным [Craster, Potter, Baumberg, 1999] замена кодонов в составе открытой рамки считывания перед опероном *ilvBNC* из *S. coelicolor* не влияет на экспрессию.

Биосинтез аминоацил-тРНК синтетаз. Для генов *ileS* изолейцил-тРНК синтетаз большинства актинобактерий (*A. naeslundii, B. longum, Corynebacterium spp., K. radiotolerans, Mycobacterium spp., N. farcinica, P. acnes, R. xylanophilus, Streptomyces spp.*) предсказана регуляция трансляции

с участием Т-бокса. См. табл. 19. Специфицирующая шпилька, содержащая изолейциновый анти-анти-кодон, у *S. coelicolor* показана на рис. 8.

В отличие от обычной регуляции с участием Т-бокса на уровне транскрипции, здесь нет терминатора. Незагруженная тРНК стабилизирует структуру РНК, при которой область Шайна-Дальгарно открыта для инициации трансляции. В ином случае, сайт связывания рибосомы перекрывается длинной спиралью альтернативной структуры, предотвращая трансляцию гена *ileS*. Выравнивание последовательностей РНК показано в табл. 20.

Высокая консервативность участков, которые перекрывают область Шайна-Дальгарно, позволила уточнить положение инициирующего кодона гена *ileS* у *C. efficiens*. Это предположение хорошо согласуется со множественным выравниванием соответствующих белков.

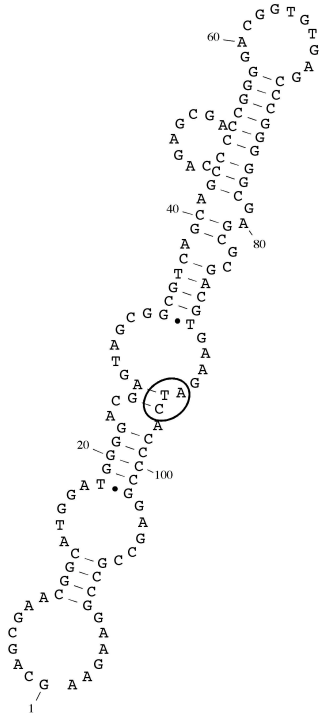

Рис. 8. Специфицирующая шпилька Т-бокса перед геном *ileS* у *S. coelicolor* . Изолейциновый кодон выделен.

Таблица 19. Гены *ileS* у актинобактерий, перед которыми предсказан Т-бокс.

Вид	Локус	Координаты гена *ileS*	Белок
A. naeslundii	TIGR_240017	complement(1311947..1315252)	
C. diphtheriae	NC_002935	complement(1617227..1620385)	NP_939931
C. efficiens	NC_004369	complement(2160737..2164195) удалено 49 кодонов	NP_738653
C. glutamicum	NC_003450	complement(2270986..2274150)	NP_601350
M. avium	NC_002944	1324371..1327532	NP_960180
M. bovis	NC_002945	1720532..1723657	NP_855215
M. tub H37Rv	NC_000962	1736519..1739644	NP_216052
M. tub CDC1551	NC_002755	1736672..1739797	NP_336040
M. marinum	Sanger_216594	complement(184205..187372)	
M. leprae	NC_002677	1410785..1413964	NP_301871
N. farcinica	NC_006361	1932119..1935247	YP_117986
P. acnes	NC_006085	268050..271394	YP_054935
R. xylanophilus	NZ_AAEB01000029	complement(26358..29492)	ZP_00187197
S. avermitilis	NC_003155	complement(7371348..7374491)	NP_827306
S. coelicolor	NC_003888	2227237..2230380	NP_626335
T. fusca	NZ_AAAQ02000011	complement(75752..78934)	ZP_00291779

Таблица 20. Т-боксы перед геном *ileS* у актинобактерий. Т-бокс и область Шайна-Дальгарно RBS выделены курсивом, спирали выделены подчёркиванием и прописными буквами.

An 1315386	ccgucccggaugggcgcgcaguacggcaa*gcgaggugguaccgcgg*ugcggcaccagccgggcaccagccccggu
Cd 1620486	uacaucagaugccucugguggaaugcucaa*gcggggugguaccgcgc*gga**********
Ce 2164019	*gguggccuguuggugggccgcagguucaa*gcagggugguaccgcg*uccggauca**********
Cg 2274270	aacgaaguggagcuaguuaauuuagcucaa*gcugggugguaccgcg*uccguuu**********
Ma 1324265	*****gaguggccacgcgaaagcgcggcaa*gcggggugguaccgcggc*gcucgcgcag**********
Mb 1720398	****cgagcggccgcgcaucggcguggcaa*gcggggugguaccgcggc*guucgcgca**********
Mt 1736385	****cgagcggccgcgcaucggcguggcaa*gcggggugguaccgcggc*guucgcgca**********
Mt 1736538	****cgagcggccgcgcaucggcguggcaa*gcggggugguaccgcggc*guucgcgca**********
Ml 1410679	****aguggccgugcguucgcgugcggcaa*gcggggugguaccgcggc*gcucgcgcac**********
Mm 187479	aaauugagcggccgcacucaggugcggcaa*gcggggugguaccgcggc*gcucgcgca**********
Nf 1931988	**gagguccggugcguccgacgccggacaa*acggggugguaccgcgg*uuucggcgcac**********
Pa 267949	********cgacgucguugacgucgugcaa*ggagggugguaccgcgg*guacccggaga**********
Rx 29622	agcgguccggggccgcgaggccucgggcaa*gcagggugguaccgcga*gagccgcuucuuuggagaaaga*******
Sa 7374620	***ggugcacacagggcgccggggagccaa*ggagggugguaccgcggg*agcgcgccgcacacggcguacggaaaga
Sc 2227135	***gagcacacgacgcaccggccgggccaa*ggagggugguaccgcggg*agca**********
Tf 79034	*****ggcaggacgacggccgcgcggccaa*ggagggugguaccgcggg*gcguc**********
	Т-бокс
An	cgggagccgac*gGUCGuCCUCGu*caggccc**********cgggcacccgccC*GAGGCGG*Caggaacga****
Cd	*********aa*cgcgUCCCCgC*ACUUUaaggc**********agaaugcuugcg*AAAGU*GaaGGAG*Aaaa****
Ce	********ag*gggcGUCCCcgc*aaguaca*******ugaccaucauuggcacuugcgaaggauua*aGGGA*ccgacucac**
Cg	******uuu*agggcGCCCCc*gc*agguagaacgaua*******auuauuguuacuugcgugaaggau*GGGACc*gaacacac**
Ma	*ccagcgcgu*CGUCGUCCCCg*guuugca**********ccguggcaca*GGAGA*CAACGcgcauc*
Mb	*ccggcgu*ggCGUCGUCCCCg*aGCCUggauugcaGGCACgcaGUGCCgaacggugcugGGGCcu*GGGGAg*ACGacgcgcaaa
Mt	*ccggcgu*ggCGUCGUCCCCg*aGCCUggauugcaGGCACgcaGUGCCgaacggugcugGGGCcu*GGGGAg*ACGacgcgcaaa
Mt	*ccggcgu*ggCGUCGUCCCCg*aGCCUggauugcaGGCACgcaGUGCCgaacggugcugGGGCcu*GGGGAg*ACGacgcgcaaa
Ml	*cuagcgcgu*CGUCgUCCccgu*gucuacuugu**********guuaaguggccca*GGAGA*CGu*********
Mm	cugagcgcgu*CGUCgUCCccgu*gccg**********ugugauuucuggcaca*GGAGA*Ccg*********
Nf	cgggcgcc*gaGGUCGUCCCCGU*GCCcacacagacacgcg******cccugcggcgcggu*GGCAC*GAGGAGA*Cgcauccgcg*
Pa	*auccggu*gugcucGUCCCUCG*gugacc**********cgagaC*GAAGGA*Ccacccgcugcg**
Rx	******GG*GCUCCCGUCCCuGC*GGCCggaga**********GGUCGCc*GGGGCGGGAG*CCUggcuuuucaacgggag
Sa	****cucg*gcucucgUCCCUCCGGaCGGAagg**********agaaagUCCGCC*GGAGGAag*cucgccg*
Sc	******cg*gcucucgUCCCUCCG*aCGGAagg**********cagcacgUCCGcC*GGAGGAag*cucgcug*
Tf	********u*gccucgUCCCUCCG*UCAGGugaccag**********caccCCUGAU*GGAAAGG*uacgccac*****
	RBS

Потенциальная **классическая аттенюация** обнаружена перед геном *leuS* лейцил-тРНК синтетазы у *S. avermitilis* и *S. coelicolor* с лидерным пептидом, терминатором и антитерминатором, а также перед геном $trpS_2$ триптофанил-тРНК синтетазы у *S. avermitilis*. Аттенюаторы приведены в табл. 21.

Таблица 21. Аттенюаторная регуляция транскрипции гена *leuS* у *Strepmyces*.

a) Координаты гена *leuS*.

Вид	Локус	Ген	Координаты гена	Белок
S. avermitilis	NC_003155	*leuS*	6661895..6664783	NP_826665
S. coelicolor	NC_003888	*leuS*	complement(2775536..2778436)	NP_626809

b) Аттенюаторы. Антитерминаторы подчёркнуты, терминаторы выделены прописными буквами. В верхней строке указаны аминокислоты лидерного пептида.

```
   M  R  A  V  R  L  L  L  S  E  P  R
Sa augcgugccguacgccuucugcuuagcgagccgcgcugaucagcccagaccacugacga
Sc augcgugccguacgccuucugcuuagcgagccgcgcugaucagucccgaccccggucgu

Sa **uuc*guggucggaaucggcgcggcguCCCCUCcugugcGAGGGGuuuuuucauu
Sc aguccgguggccggaaucggcgcggcgUCCCCUCcugugcGAGGGGAuuuuucauu
```

2.3 Регуляция трансляции гена *ykoE* АВС транспортёра посредством тиаминового рибопереключателя у актинобактерий

Геномы бактерий получены из базы данных GenBank (NCBI). В качестве набора последовательностей были взяты 5'-нетранслируемые области перед гомологами гена *ykoE* гипотетического АВС транспортёра у актинобактерий *Brevibacterium linens*, *Kineococcus radiotolerans*, *Leifsonia xyli*, *Propionibacterium acnes*, *Thermobifida fusca*, *Corynebacterium diphtheriae*, *Corynebacterium glutamicum*, перед которыми был найден рассматриваемый рибопереключатель (см. табл. 22), а также у

Bifidobacterium longum, *Corynebacterium jeikeium*, *Corynebacterium efficiens* и *Tropheryma whipplei*, содержащих гомологи гена *ykoE*, но у которых сигнал не был обнаружен.

Таблица 22. Краткие обозначения геномов и номера белков, гомологичных YkoE в сенной палочке.

Краткое и полное название вида		Белок, гомологичный YkoE
Br	*Brevibacterium linens*	ZP_00378910.1
Kr	*Kineococcus radiotolerans*	ZP_00619644.1
Lx	*Leifsonia xyli*	YP_062345.1
Pa	*Propionibacterium acnes*	YP_054871.1
Tf	*Thermobifida fusca*	YP_288648.1
Cd	*Corynebacterium diphtheriae*	NP_939314.1
Cg	*Corynebacterium glutamicum*	YP_225369.1

Всего найдено шесть консервативных участков с длинами десять, двенадцать, десять, тринадцать, шесть и шесть нуклеотидов. При этом выявились консервативные спирали, характерные для THI-рибосвитча. Консервативная структура РНК показана на множественном выравнивании в табл. 23.

Мы предполагаем, что найденная консервативная структура связана с регуляцией трансляции. При этом у *B. linens*, *K. radiotolerans*, *L. xyli*, *P. acnes* и *T. fusca* область связывания рибосомы перекрывается короткой дополнительной спиралью, а у *C. diphtheriae* и *C. glutamicum* рибопереключатель примыкает непосредственно к области связывания рибосомы.

Рассмотренные рибопереключатели у трёх актинобактерий *T. fusca*, *C. diphtheriae* и *C. glutamicum* описаны ранее в статье [Rodionov, Vitreschak, Mironov, Gelfand, 2002]. Там же регуляция экспрессии на основе тиаминового рибопереключателя предсказана и у некоторых других генов

актинобактерий, но соответствующие участки мРНК менее консервативны и не определяются предлагаемым в диссертации методом при сопоставлении с таковыми перед геном *ykoE*.

Таблица 23. Множественное выравнивание, построенное на основе поиска консервативных слов в 5'-нетранслируемых областях гена *ykoE*. Прописными буквами выделены нуклеотиды, входящие в состав шести спиралей консервативной вторичной структуры РНК. Эти консервативные спирали характерны для THI-рибопереключателя.

Br	acAGGGgAGCGCCga**********uaggGGCGCUgagagUGCAGa*********
Kr	acAGGGgAGCGCCg**********uggGGCGCUgagagUGCGG**********
Lx	acACGGgAGUCCGGu**********gagCCGGGCUgagagGAAGCUU********
Pa	acAGGGgAGCAUCg**********ucgGAUGCUgagagUGGGC**********
Tf	acAGGGgAGCGCcu**********cuaGCGCUgagagUGCGGC*********
Cd	ucACGGguGCUggacGGCAuacguuUGCCacaaAGCugagaCAGGGcgagaagacgu
Cg	acACGGguGCUCGGuga*********aaauCCGGGCugagaucUGGC**********
	->1>*->->-2->->->****<-<-<-2<-<-<****->->3->*********
Br	*ugaagCUGCAgaCCCUc*gaaCCUgauGCGGcuagcaCCGCcga*AGGaag
Kr	**guuuCCGCAgaCCCUc*gaaCCUGauCCGGuucagaCCGGcg*UAGGgag
Lx	auccAAGCUUCgaCCGUc*gaaCCUGauCUGGgucaugCCAGcg*CAGGgag
Pa	***accGCCCAgaCCCUc*gaaCCUGaaCCGGuuaggaCCGGcg*UAGGgag
Tf	**acaGCCGCAgaCCCUuacuaCCUgauCUGGguaaugCCAGcga*AGGaag
Cd	gcacguCCCUGaaCCGUu*gaaCCUGauCCGGguaauaCCGGcgaUAGGaag
Cg	***auaGCCAcgaCCGUc*gaaCCUGauCCGGauaaugCCGGcgaUAGGgag
	****<-3<-<-**<1<-*****>4->**->5>******<5<-***<6<-***

2.4 Регуляция трансляции гена *alr3806* с участием Т-бокса трансляции у цианобактерии *Nostoc* PCC7120

Добавляя к набору последовательностей, содержащих известный хороший сигнал, новую последовательность, можно предсказывать наличие сигнала в ней. Известна регуляция транскрипции у многих фирмикутов с участием Т-бокса [Henkin, Glass, Grundy, 1992]. Сопоставление регуляторных областей позволило предсказать регуляцию трансляции с участием Т-бокса в цианобактерии *Nostoc* PCC7120 гена *alr3806*, который не имеет ортологов ни в каком другом полном геноме из базы NCBI.

Открытая рамка считывания *alr3806* длиной 450 аминокислот из *Nostoc* PCC7120 предсказана теоретически и не соответствует экспериментально подтверждённому белку. Перед этой открытой рамкой на расстоянии 147 нуклеотидов на комплементарной цепи расположена другая открытая рамка считывания *alr3805*, экспрессия которой также не подтверждена экспериментом. Предполагаемый Т-бокс почти целиком расположен в промежутке между этими рамками, включая инициирующий кодон гена *alr3805*. Предлагаемая структура имеет слово с правильным консенсусом (собственно Т-бокс) и шпильки, характерные для Т-бокса. Важно, что тРНК стабилизирует такую структуру РНК, которая не препятствует трансляции гена *alr3806*. В противном случае возникает спираль, перекрывающая область связывания рибосомы гена *alr3806*, препятствуя трансляции.

Специфицирующим кодоном, вероятно, является аргининовый кодон AGG в выпячивании на 3’-плече соответствующей шпильки. Такое расположение кодона соответствует ранее известным примерам Т-боксов. Таким образом, формирование структуры РНК зависит от концентрации аргинина и, следовательно, связанного азота.

Ортологов для гена *alr3806* не найдено, но его N-концевой домен обладает АТФазной активностью. Ближайшие гомологи гена *alr3806* как в той же цианобактерии *Nostoc* PCC7120 (ген *all4835*), так и других цианобактериях порядка Nostocales (*Anabaena variabilis* ATCC29413 и *Nostoc punctiforme* PCC73102) ортологичны друг другу и принадлежат COG1066: АТФ-зависимые сериновые протеазы.

Сказанное выше о предполагаемой регуляции и гомологах гена *alr3806* позволяет предположить, что соответствующий белок участвует в деградации цианофицина и освобождении аргинина. Однако цианофициновые гранулы характерны для других цианобактерий, не имеющих ортолога для гена *alr3806*.

ГЛАВА 3. Моделирование классической аттенюаторной регуляции биосинтеза триптофана у бактерий

3.1 Математическая модель классической аттенюаторной регуляции

Определения микросостояний. Константы скоростей переходов. Предполагается, что дана и далее везде фиксирована последовательность в четырёхбуквенном алфавите {A, C, U, G} – регуляторная область в геноме бактерии или случайная последовательность. Например, область от сайта связывания рибосомы перед лидерным пептидом и до конца терминатора транскрипции, включая участок остатков урацила.

В исходной последовательности выделяются отрезки длиной не менее трёх нуклеотидов – плечи будущих спиралей. При спаривании каких-то отрезков одинаковой длины получается спираль γ_i – везде предполагается, что спираль γ_i *непродолжаемая,* и промежуток между отрезками (концевая петля спирали) имеет длину не менее трёх нуклеотидов. Модель допускает, вообще говоря, любой список *исходных спиралей*; выше определён лишь один из возможных вариантов, в котором в качестве исходных берутся все непродолжаемые спирали с указанными ограничениями на плечо и петлю.

Все эти представления, как и описание самой классической аттенюаторной РНКовой регуляции экспрессии генов в зависимости от концентрации аминокислоты (или загруженной тРНК, последняя в свою очередь определяется концентрациями аминокислоты и аминоацил-тРНК синтетазы), изложены, например, в [Сингер, Берг, 1998].

Гипоспиралью спирали γ_i называется любая непустая часть $\overline{\gamma}_i$ спирали γ_i, состоящая из двух связных плеч длины *не менее* трёх нуклеотидов. Здесь и далее *плечами* называются спариваемые отрезки гипоспирали или спирали, концы которых будут стандартно *обозначаться*

(считая от 5'-начала исходной последовательности) буквами A, B, C, D. *Концевой петлей* называется участок цепи РНК между двумя плечами гипоспирали.

Микросостоянием называется совместный набор гипоспиралей, не продолжаемых *в этом наборе* и без псевдоузлов, для которого никакие две гипоспирали не соприкасаются (т.е. A и D одной из них не являются оба соседними нуклеотидами к B и C другой из них); кроме того, отдельным «начальным» микросостоянием является пустое множество $\varnothing$. *Псевдоузлом* называется пара гипоспиралей, у которой ровно одно плечо одной из них пересекается с петлей другой (и, следовательно, находится в этой петле). Объединение всех гипоспиралей от одной спирали, вошедших в данное микросостояние, называется *подспиралью* этой спирали в данном микросостоянии.

Для любого микросостояния каждая из его гипоспиралей и подспиралей получает тот же номер, что и непродолжаемая спираль, из которой она взята; при этом все спирали исходной последовательности нумеруются в каком-то заранее фиксированном порядке.

Любому набору, состоящему из спиралей $\gamma_1,\ldots,\gamma_k$, соответствует множество *реализующих его микросостояний*: это любой нерасширяемый (в себе) набор *подспиралей* $\overline{\gamma}_1 \subseteq \gamma_1,\ldots,\overline{\gamma}_k \subseteq \gamma_k$ (от каждой спирали γ_i берется ровно один непустой и не обязательно связный участок $\overline{\gamma}_i$) без псевдоузлов. Как и выше, *соседние* гипоспирали (т.е. у которых пара A и D нуклеотидов расположена непосредственно вслед за парой B и C) объединяются.

Каждой гипоспирали $\overline{\gamma}_i$ из данного микросостояния ω приписывается число l_i нуклеотидов в её концевой петле, не вошедших в петли и плечи других гипоспиралей из этого микросостояния. Это число, зависящее от микросостояния, называется *длиной* концевой петли гипоспирали $\overline{\gamma}_i$ и обозначается l_i.

Микросостоянию ω по определению приписываются: *свободная энергия связи* гипоспиралей и *свободная энергия петель* гипоспиралей из ω. Далее везде рассматриваются относительные энергии, т.е. деленные на $R \cdot T$, где температура T равна $310K$, а R – универсальная газовая постоянная. Поэтому все формулы для вычисления энергий дают безразмерные величины. Энергия связи микросостояния ω принимается равной:

$$G_{hel}(\omega) = \frac{1}{RT} \sum_j E_{\overline{\gamma}_j},$$

где j пробегает все гипоспирали из микросостояния ω. Вычисление *энергии связи* $E_{\overline{\gamma}_j}$ гипоспирали происходит по схеме и с численными значениями, взятыми из [Mathews, Sabina, Zuker, Turner, 1999] и [Mathews, Disney, Childs, Schroeder, Zuker, Turner, 2004].

Энергия петель микросостояния ω принимается равной сумме:

$$G_{loop}(\omega) = \sum_i \left(1,77 \cdot \ln(l_i + 1) + B + \frac{C}{l_i} \right),$$

где i пробегает все гипоспирали из микросостояния ω.

Эта формула хорошо согласуется с обширными таблицами из [Mathews, Disney, Childs, Schroeder, Zuker, Turner, 2004] для энергий всех петель при всех $l_i > 2$, если положить B=6.5 для концевых петель, B=0 для двусторонних выпячиваний, B=4 для односторонних выпячиваний. Случаи $l_i \le 2$ рассматриваются отдельно в соответствии с таблицами из [Mathews, Disney, Childs, Schroeder, Zuker, Turner, 2004]. А именно, принимаются значения энергий петель: для двустороннего выпячивания 0.8 (при l=2), для одностороннего выпячивания 6.2 (при l=1) и 4.5 (при l=2). Концевые петли с такими длинами в модели исключаются. Константа C=5.

Переходы между микросостояниями делятся на быстрые и медленные. В модели предполагается, что на множестве микросостояний, между любыми двумя из которых возможен быстрый переход,

устанавливается стационарное распределение вероятностей Больцмана-Гиббса:

$$p(\omega)=\frac{\exp\left(-(G_{loop}(\omega)+G_{hel}(\omega))\right)}{z(\Omega)},\ \text{где}\ z(\Omega)=\sum_{\omega\in\Omega}\exp\left(-G_{loop}(\omega)-G_{hel}(\omega)\right).$$

Такое множество микросостояний объединяется в одно макросостояние. Для медленных переходов между микросостояниями, константы скоростей переходов между микросостояниями в модели вычисляются по формуле:

$$K(\omega\to\omega')=\kappa\cdot\exp\left[\frac{1}{2}\left((G_{loop}(\omega)+G_{hel}(\omega))-(G_{loop}(\omega')+G_{hel}(\omega'))\right)\right].$$

Аналогичное правило вычисления скоростей переходов рассмотрено, например, в [Flamm C., Fontana W., Hofacker I.L., Schuster P. RNA folding at elementary step resolution // RNA, V.6, 2000. P.325–338], для численного моделирования процесса формирования вторичной структуры РНК. Подобная формула предложена Кавасаки [Kawasaki K. Diffusion constants near the critical point for timedependent Ising models // Phys. Rev. V.145, 1966. P.224–230.] в связи с теорией кинетической модели Изинга. Легко видеть, что выполнен принцип детального равновесия:

$$\frac{K(\omega\to\omega')}{K(\omega'\to\omega)}=\exp[G(\omega)-G(\omega')],$$

где $G(\omega)$ – энергия, приписываемая микросостоянию ω.

Отметим, что многие авторы, например, [Danilova, Pervouchine, Favorov, Mironov, 2006], предлагают различать случаи распада и образования гипоспирали. Константы скоростей переходов между микросостояниями вычисляются по следующим несимметричным формулам. В случае медленного перехода – распада гипоспирали $K(\omega\to\omega')=\kappa\cdot\exp\left(G_{hel}(\omega)-G_{hel}(\omega')\right)$. В случае присоединения гипоспирали константа скорости $K(\omega'\to\omega)$ зависит только от длин петель спиралей, составляющих микросостояния ω и ω'.

В приведенных ниже расчетах принималось значение константы $\kappa = 1000\ c^{-1}$.

После очевидного усреднения по всем парам микросостояний $\omega \in \Omega, \omega' \in \Omega'$ получим следующую формулу для *скорости перехода из одного макросостояния* Ω в *другое* макросостояние Ω':

$$K(\Omega \to \Omega') = \sum_{\omega \in \Omega} \sum_{\omega' \in \Omega'} p(\omega) \cdot K(\omega \to \omega').$$

Величина замедления полимеразы вторичной структурой, образующейся на участке мРНК между рибосомой и РНК-полимеразой. Согласно экспериментальным данным, включая результаты из [Wilson, von Hippel, 1995], [Uptain, Chamberlin, 1997] и [Yin, Artsimovitch, Landick, Gelles, 1999], вероятность терминации в зависимости от длины шпильки терминатора имеет вид кривой, которая в физической литературе называется «резонансной». Не претендуя на обсуждение физического процесса взаимодействия шпильки с полимеразой, сделана попытка использовать такую кривую для описания зависимости константы скорости перескока полимеразы от вторичной структуры РНК. Итак, в модели принимается, что «сила» F замедления *шпилькой* ω полимеразы, имеющая смысл величины эффективного уменьшения константы скорости движения полимеразы по цепи ДНК и измеряемая в c^{-1}, определяется по формуле:

$$F(\omega) = \frac{\delta}{L_1^2 \cdot (p - p_0)^2 + 1} \cdot \exp\left(-\frac{r}{r_0}\right),$$

где r – *расстояние* от конца D шпильки ω до начала полимеразы. Параметры L_1, p_0, r_0, δ, зависят только от свойств полимеразы, число p зависит только от шпильки.

Движение РНК-полимеразы по цепи ДНК. Рассмотрим ситуацию перескока полимеразы с нуклеотида, *принадлежащего U-богатому участку*. Если 3'-конец полимеразы, обозначаемый везде далее z, находится на n-м нуклеотиде, то возможен её перескок на $(n+1)$-й нуклеотид или срыв с нуклеотидной последовательности. *U-богатый*

участок определяется следующим образом. Нуклеотид z назовём U-богатым, если существует хотя бы одно слово, содержащее z на любом его месте, которое по длине больше порога (по умолчанию, 6) и по плотности U больше порога (по умолчанию, 0.8). Это слово может содержать исключения, т.е. не букву U, в любом месте, включая концы; само z также может не быть буквой U. Во множестве всех U-богатых нуклеотидов образуем все интервалы максимальной длины. Они называются U-богатыми участками, и не пересекаются.

Полимераза из положения $z = n$ («основное состояние») может перейти с константой скорости $\overline{\lambda}_{pol}$ в положение $z = n+1$ или с некоторой константой скорости в «возбужденное» состояние n^*, из которого может с константой λ_{ur} соскочить или с некоторой константой скорости вернуться назад в основное состояние $z = n$, [Yin, Artsimovitch, Landick, Gelles, 1999]. Если переходы между n и n^* быстрые, то эту схему движения полимеразы можно заменить на её усреднение: пусть переход из n в $n+1$ происходит с константой $\beta \cdot \overline{\lambda}_{pol}$ и срыв из n с константой $(1-\beta) \cdot \lambda_{ur}$, где β – вероятность найти полимеразу в основном состоянии, а $(1-\beta)$ – вероятность найти её в возбужденном состоянии. Приравнивая $\beta \cdot \overline{\lambda}_{pol} = \nu(\Omega) = \overline{\lambda}_{pol} - F(\Omega)$, получим $(1-\beta) \cdot \overline{\lambda}_{pol} = F$, т.е. $(1-\beta) = \dfrac{F}{\overline{\lambda}_{pol}}$, и окончательно имеем для срыва полимеразы константу скорости $\mu = \mu_{out} = \dfrac{\lambda_{ur} \cdot F}{\overline{\lambda}_{pol}}$. Отношение $\dfrac{\overline{\lambda}_{pol}}{\lambda_{ur}}$ является естественным параметром модели и по данным из [Yin, Artsimovitch, Landick, Gelles, 1999] равно четырём.

Движение рибосомы по цепи мРНК. На *нерегуляторных* кодонах константа скорости λ_{rib} сдвига рибосомы на один кодон принимается равной $\lambda_{rib} = \overline{\lambda}_{rib} = 15$ c^{-1}. На *регуляторных* кодонах она зависит от концентрации c по формуле Микаэлиса-Ментен:

$$\lambda_{rib}(c) = \frac{\overline{\lambda}_{rib} \cdot c}{c_0 + c},$$

где c – концентрация аминоацил-тРНК и c_0 – концентрация аминоацил-тРНК, при которой рибосома движется по регуляторным кодонам со скоростью равной половине от максимальной скорости такого движения $\overline{\lambda}_{rib} = 15\,\text{с}^{-1}$, и $\overline{\lambda}_{rib}$ – значение этой функции, при столь большой концентрации c, что прохождение рибосомой регуляторных кодонов происходит с той же скоростью, что и нерегуляторных.

Вообще говоря, концентрация триптофанил-тРНК зависит от концентраций триптофана, тРНК и триптофанил-тРНК синтетазы. При моделировании регуляции экспрессии оперонов биосинтеза триптофана естественно полагать, что последние две концентрации достаточно велики. В этом случае величина c имеет смысл концентрации триптофана. Напротив, при моделировании регуляции экспрессии гена триптофанил-тРНК синтетазы, естественно полагать концентрацию триптофана высокой. Тогда величина c отражает концентрацию триптофанил-тРНК синтетазы.

3.2 Проверка модели методом Монте-Карло

Цель моделирования состояла в численном определении зависимости $p(c)$ вероятности терминации от концентрации c аминоацил-тРНК или от концентрации c аминокислоты в клетке. Для построения зависимости $p(c)$ при каждом значении c из сетки с некоторым шагом узлов указанный в модели процесс проигрывался определенное число раз (например, 10^3-10^4 раз, что дает примерно одинаковый результат) и вычислялось значение $p(c)$ как доля случаев, в которых происходила терминация. Параметр c_0 полагаем равным единице, т.е. единицей измерения по оси c является c_0, а r_0 подбирался в интервале от двух до восьми.

Для исходной фиксированной последовательности РНК *текущее состояние* модели характеризуется:

(1) Окном между положениями 3’-края x рибосомы и начала y полимеразы. «Размер» рибосомы от её Р-участка до её 3’-края обозначим s_0 (от десяти до двенадцати нуклеотидов, по умолчанию – двенадцать), а «размер» полимеразы от y – места выхода цепи РНК до точки транскрипции через s_1 (от двух до семи нуклеотидов, по умолчанию пять). Точку транскрипции обозначим z, всегда выполняется $z=y+s_1$. В окне происходит перестройка вторичной структуры от одного макросостояния к другому – при этом макросостояния могут включать только спирали, пересекающиеся с окном обоими плечами хотя бы по трем нуклеотидам, т.е. речь идёт о макросостояниях в окне (для текущего окна).

(2) Списком T (потенциальных) спиралей, пересекающихся с окном обоими плечами (хотя бы по минимальной длине гипоспирали, т.е. по трём нуклеотидам); это тривиальная компонента состояния в том смысле, что каждый раз её можно вычислять заново по исходному списку спиралей.

(3) Макросостоянием Ω в окне.

До посадки полимеразы окна нет (*пустое макросостояние*), а после посадки полимеразы и до посадки рибосомы окно начинается в первом нуклеотиде исходной последовательности – точке 0 и заканчивается в текущем положении начала полимеразы. В окне может впервые появиться непустое макросостояние Ω, состоящее из одной спирали. Затем к этой спирали может добавиться вторая спираль или, наоборот, макросостояние может вернуться к исходному – пустому, и так далее.

Отслеживается один из *двух возможных исходов* моделирования: срыв полимеразы на одном из нуклеотидов участка остатков урацила исходной последовательности, или прохождение полимеразой всего участка остатков урацила.

Инициация процесса РНКовой регуляции: Как только полимераза транскрибировала инициирующий кодон лидерного пептида и ещё s_0+s_1 нуклеотидов, на область Шайна-Дальгарно пытается сесть рибосома с константой скорости, которая отражает зависимость от качества этой

области и от вторичной структуры, закрывающей ее. Как только это произошло, рибосома занимает положение на инициирующем кодоне лидерного пептида. В этот момент фиксируются: левый конец x окна в точке «начало лидерного пептида» + s_0, и правый конец y окна в том положении, которое на тот момент занимает начало полимеразы.

Переходы в процессе РНКовой регуляции после формирования окна $[x, y]$.

(1) *Сдвиг* полимеразы на один нуклеотид вправо, при этом окно увеличивается на один нуклеотид, а список спиралей T может расшириться. Или *срыв* полимеразы на U-богатом участке.

(2) *Сдвиг* рибосомы на один кодон вправо; окно уменьшается на три нуклеотида и, вообще говоря, список спиралей T сокращается, а макросостояние Ω меняется. Из Ω исключается гипоспираль, 5'-плечо которой перекрывается рибосомой. Так полученное макросостояние – новое Ω, может быть, пустое – оно фиксируется в текущем окне.

(3) *Перестройка* вторичной структуры, т.е. смена макросостояния в окне; при этом само окно и список спиралей T не меняются.

(4) *Окончание моделирования*. При наступлении события срыва полимеразы на участке урацилов моделирование прекращается; в ином случае, полимераза проходит весь участок урацилов и моделирование также прекращается. Если на каком-то переходе рибосома не сдвигается, то можно фиксировать и время до наступления перехода, эти времена суммируются вплоть до наступления события первого сдвига рибосомы. Распределение этих времен несёт полезную информацию.

Организация переходов при моделировании. Моделирование выполняется с использованием метода Монте-Карло стандартным образом. *Состояние* описывается набором (x, y, z, T, Ω). *Окрестностью* данного состояния Ω (с *центром* в Ω) называется набор всех состояний, в которые можно (с ненулевой вероятностью) перейти из Ω. Если окрестность состоит из n состояний и соответствующие константы скоростей переходов

равны соответственно $k_1, \ldots, k_n$ (пусть $k = \sum k_i$), то состояние, в которое переходим (которое считается *следующим* на данной траектории), определяется как реализация случайной величины $i \to \frac{k_i}{k}$. Заметим, что порядки величин $\lambda_{sd}, \lambda_{rib}, \lambda_{pol}$ и $K(\Omega \to \Omega')$ часто значительно отличаются.

3.3 Тестирование модели и обсуждение результатов

Исходная последовательность бралась от области Шайна-Дальгарно лидерного пептида до конца участка урацилов, т.е. до полиурацилового тракта терминатора транскрипции.

В табл. 24 приведены результаты счёта вероятности терминации транскрипции при регуляции оперонов, включающих ген антранилат синтазы, у различных бактерий. А также при регуляции гена *trpS*, кодирующего триптофанил-тРНК синтетазу, у *Streptomyces avermitilis*. Неизвестные заранее параметры модели подобраны так, что для *E. coli* и *C. diphtheriae*, у которых аттенюаторная регуляция подтверждена экспериментально, при малой концентрации наблюдается рост частоты терминации с увеличением концентрации триптофана, а при больших концентрациях – насыщение и выход на плато. Было исследовано поведение функции $p(c)$ при различных значениях параметров и выбраны те, при которых у этих оперонов наиболее выражен эффект регуляции транскрипции.

Для параметров F приняты значения $\delta = 25$, L_1=14.5, p_0=0.167, r_0=2.

Влияние «размеров» рибосомы и полимеразы имеет место, но заметно слабее и одинаковое для всех рассмотренных организмов и генов. Это позволило выбрать для них общие значения s_0=12, s_1=5.

С другой стороны, для гена *trpE* у *S. avermitili*, *S. coelicolor* и *S. venezuelae* не удалось подобрать такие значения параметров, при которых аттенюация проявляется столь же заметно. В этих случаях рассчётная вероятность терминации оказывается высокой даже при малых значениях

концентраций триптофана, что объясняется высокой устойчивостью шпильки терминатора. И сравнительно медленно растёт с увеличением концентрации. В случае гена *trpS* у *S. avermitilis* вычисленное изменение вероятности терминации также мало при различных значениях параметров, но монотонно возрастает от 0.06 до 0.30.

При значениях параметров L_1=14.5, p_0=0.167, r_0=2, получены разумные зависимости для вероятности $p(c)$ для триптофанового оперона у актинобактерии *C. glutamicum*, у гамма-протеобактерии *Vibrio cholerae* и у альфа-протеобактерий из статьи [Vitreschak, Lyubetskaya, Shirshin, Gelfand, Lyubetsky, 2004]: *Agrobacterium tumefaciens*, *Bradyrhizobium japonicum*, *Rhodopseudomonas palustris*, *Rhizobium leguminosarum*, *Sinorhizobium meliloti*.

Для *C. glutamicum* вычисленная вероятность терминации возрастает при росте концентрации триптофана в два раза, но остаётся очень низкой. В то же время для некоторых альфа-протеобактерий предсказанная вероятность терминации изменяется весьма значительно: для *R. palustris* в 48 раз, для *S. meliloti* в 7.6 раза. Также значительный рост вероятности терминации предсказан для *Vibrio cholerae*: в 16.6 раза.

Значение константы замыкания, равное $\kappa = 10^3$ с$^{-1}$, является близким к оптимальному. При меньшем её значении перестройка вторичной структуры почти не происходит, большинство событий в процессе моделирования – это сдвиг рибосомы или полимеразы. Напротив, при увеличении κ число смен вторичной структуры в текущем окне быстро растёт, а зависимость вероятности терминации от концентрации триптофана становится менее выраженной. При значении $\kappa = 10^6$ с$^{-1}$ вероятность терминации для случая *E. coli* близка к 0.7 не зависимо от концентрации c.

Табл. 24 содержит также данные моделирования при $\kappa = 10^3$ с$^{-1}$ для лидерных областей перед геном *trpE* у кишечной палочки, где один нуклеотид G заменён нуклеотидом A в соответственно 75 и 132 позициях от начала

транскрипции. В первом случае (trpL75) мутация нарушает антитерминирующую структуру и наблюдается высокая вероятность терминации. Во втором случае (trpL132) мутация нарушает шпильку терминатора, снижая частоту терминации, но не нарушает регуляцию полностью. Влияние этих мутаций было экспериментально подтверждено [Das, Crawford, Yanofsky, 1982].

Из таблицы видно, что у мутанта trpL75 вычисленная вероятность терминации мало зависит от концентрации. Это служит дополнительным подтверждение корректности модели, поскольку невозможность формирования антитерминатора соответствует отсутствию регуляции.

Таблица 24. Указаны вероятности терминации $p(c)$ в зависимости от концентрации c триптофана (или триптофанил-тРНК) для различных видов бактерий. А также для последовательностей trpL75 и trpL132, полученных замещением 75 и 132 нуклеотида от начала транскрипции РНК у кишечной палочки. Константа $\kappa = 10^3$ с$^{-1}$. Концентрация измеряется в условных единицах c_0=1.

Вид	Концентрация c										
	0.00	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
C. diphtheriae	0.34	0.34	0.39	0.46	0.50	0.54	0.53	0.53	0.53	0.52	0.54
C. glutamicum	0.05	0.06	0.08	0.10	0.10	0.09	0.09	0.09	0.10	0.10	0.10
S. avermitilis, trpS	0.06	0.13	0.21	0.26	0.28	0.29	0.30	0.30	0.32	0.32	0.30
A. tumefaciens	0.49	0.50	0.62	0.70	0.74	0.78	0.77	0.78	0.82	0.80	0.79
B. japonicum	0.19	0.20	0.24	0.26	0.28	0.26	0.26	0.27	0.26	0.26	0.26
R. leguminosarum	0.23	0.30	0.42	0.55	0.60	0.65	0.67	0.70	0.71	0.71	0.71
R. palustris	0.01	0.22	0.40	0.48	0.56	0.59	0.60	0.60	0.63	0.61	0.62
S. meliloti	0.07	0.11	0.23	0.37	0.43	0.49	0.48	0.51	0.50	0.53	0.51
E. coli	0.34	0.46	0.54	0.68	0.70	0.70	0.71	0.73	0.75	0.75	0.74
E. coli (trpL75)	0.74	0.74	0.74	0.77	0.74	0.74	0.76	0.76	0.75	0.74	0.75
E. coli (trpL132)	0.03	0.11	0.22	0.23	0.25	0.28	0.28	0.31	0.31	0.32	0.31
V. cholerae	0.05	0.16	0.39	0.57	0.70	0.74	0.77	0.77	0.80	0.79	0.81

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

Доказано существование алгоритма полиномиальной сложности, который сводит решение задачи поиска *n*-клики в *n*-дольном графе, с двумя вершинами в каждой доле, к вопросу о непустоте многогранника, у которого стороны имеют полиномиальную сложность описания (алгоритм и многогранник описаны явно). В результате алгоритм сводит задачу поиска такой клики к задаче линейного программирования.

Разработан алгоритм полиномиальной сложности для поиска клики в многодольном графе в общем случае. На его основе получен алгоритм для предсказания сигналов в наборе невыравненных лидерных областей генов.

Разработан алгоритм полиномиальной сложности для решения неявно заданной однородной системы линейных уравнений, который, в частности, позволяет оценивать снизу число клик в графе.

С помощью этих алгоритмов получены следующие новые потенциальные регуляторные структуры РНК. Найдены консервативные структуры РНК, которые обеспечивают потенциальную регуляцию трансляции шести генов у хлоропластов посредством взаимодействия белков с РНК. Предложена гипотеза о том, что зависимая от света регуляция трансляции гена *psbA* сформировалась на ранних стадиях эволюции, а именно: до появления интронов в генах белков и до расхождения зелёных и пурпурных водорослей. Найдена классическая аттенюаторная регуляция перед генами биосинтеза триптофана, триптофанил- и лейцил-тРНК синтетазами некоторых актинобактерий. Перед геном *leuA* у многих актинобактерий найден регуляторный элемент нового типа, названный LEU-элементом. Найдены консервативные структуры РНК, включающие Т-бокс, которые обеспечивают потенциальную регуляцию трансляции гена *ileS* у многих актинобактерий и гена *alr3806* у *Nostoc*. Найдены ген лидерного пептида и консервативный участок РНК, которые обеспечивают потенциальную Rho-зависимую аттенюаторную регуляцию на уровне транскрипции генов биосинтеза

цистеина, и определена структура оперонов, содержащих эти гены у актинобактерий. Найдены новые тиаминовые рибопереключатели, вовлечённые в регуляцию трансляции некоторых атинобактерий.

Предложена математическая модель классической аттенюаторной регуляции экспрессии генов, кодирующих ферменты биосинтеза триптофана. Модельный счёт на регуляторных областях перед триптофановыми оперонами у *E. coli*, *C. diphtheriae*, *V. cholerae* и у нескольких альфа-протеобактерий приводит к результатам, качественно согласными с экспериментальными данными.

СПИСОК ЛИТЕРАТУРЫ

1. Боревич З.И., Шафаревич И.Р. Теория чисел. М.: Наука, 1985.

2. Горбунов К.Ю., Миронов А.А., Любецкий В.А. Поиск консервативных вторичных структур РНК // Молекулярная биология. Т. 37, 2003. С. 850–860.

3. Данилова Л.В., Горбунов К.Ю., Гельфанд М.С., Любецкий В.А. Алгоритм выделения регуляторных сигналов в последовательностях ДНК // Информационные процессы. Т. 1, 2001. С. 56–63.

4. Леонтьев Л.А., Селиверстов А.В., Любецкий В.А. Алгоритм массового поиска у бактерий вторичных структур, включающих Т-бокс // Молекулярная биология. Т. 39, 2005. С. 1076–1078.

5. Любецкий В.А., Горбунов К.Ю., Пирогов С.А., Рубанов Л.И., Селиверстов А.В. Алгоритм и результаты счета для модели регуляции экспрессии генов у бактерий на основе формирования вторичных структур РНК // Информационные процессы. Т. 5, 2005. С. 337–366.

6. Любецкий В.А., Рубанов Л.И., Селиверстов А.В., Пирогов С.А. Модель регуляции экспрессии генов у бактерий на основе формирования вторичных структур РНК // Молекулярная биология. Т. 40, № 3, 2006. С. 497–511.

7. Любецкий В.А., Селиверстов А.В. Вычисление эффективности регуляции биосинтеза триптофана у бактерий на основе модели классической аттенюации // Информационные процессы.Т. 6, 2006. С. 55–57.

8. Любецкий В.А., Селиверстов А.В. Геометрический метод поиска клики в графе и его применение для выделения сигнала // Труды VI международной конференции Проблемы управления и моделирования в сложных системах, 14–17 июня 2004. Самара: изд. Самарского научного центра РАН, 2004. С. 154–157.

9. Любецкий В.А., Селиверстов А.В. Многодольные графы с двумя вершинами в каждой доле // Информационные процессы. Т. 4, 2004. С. 127–132.

10. Любецкий В.А., Селиверстов А.В. Некоторые алгоритмы, связанные с конечными группами // Информационные процессы. Т. 3, 2003. С. 39–46.

11. Любецкий В.А., Селиверстов А.В. Регуляция трансляции у актинобактерий и цианобактерий с участием вторичных структур мРНК // Труды VII Международной конференции РАН Проблемы управления и моделирования в сложных системах, 27 июня – 1 июля 2005. Самара: изд. Самарского научного центра РАН, 2005. С. 216–221.

12. Любецкий В.А., Селиверстов А.В. Регуляция экспрессии генов биосинтеза аминокислот и аминоацил-тРНК синтетаз у Актинобактерий // Молекулярная биология. Т. 39, 2005. С. 1072–1075.

13. Миронов А.А., Кистер А.Э. Теоретический анализ кинетики образования вторичной структуры РНК в процессе транскрипции и трансляции. Учет дефектных спралей // Молекулярная биология. Т. 19, 1985. С. 1350–1357.

14. Миронов А.А., Кистер А.Э. Теоретический анализ структурных перестроек в процессе образования вторичных структур РНК // Молекулярная биология. Т. 23, 1989. С. 61–71.

15. Селиверстов А.В., Любецкий В.А. Алгоритм поиска консервативных участков нуклеотидных последовательностей // Информационные процессы. Т. 6, 2006. С. 33–36.

16. Селиверстов А.В., Любецкий В.А. Особенности синтеза цистеина у *Corynebacterium*, *Mycobacterium* и *Propionibacterium* // Информационные процессы. Т. 4, 2004. С. 247–250.

17. Селиверстов А.В., Любецкий В.А. Поиск консервативных участков в лидерных областях генов в случае известного дерева видов // Информационные процессы. Т. 5, 2005. С. 265–270.

18. Селиверстов А.В., Любецкий В.А. Регуляция трансляции в хлоропластах // Информационные процессы. Т. 5, 2005. С. 400–404.

19. Симс Ч.К. Вычислительные методы в изучении групп перестановок // Вычисления в алгебре и теории чисел. М.: Мир, 1976. С. 129–147.

20. Сингер М., Берг П. Гены и геномы. М.: Мир, 1998.

21. Схрейвер А. Теория линейного и целочисленного программирования. М.: Мир, 1991.

22. Сэвидж Дж. Сложность вычислений. М.: Факториал, 1998.

23. Baytalulk M.V., Gelfand M.S., Mironov A.A. Exact mapping of prokaryotic gene starts // Briefings in Bioinformatics. V. 3, 2002. P. 181–194.

24. Craster H.L., Potter C.A., Baumberg S. End-product control of branched-chain amino acid biosynthesis genes in *Streptomyces coelicolor* A3 (2): paradoxical relationships between DNA sequence and regulatory phenotype // Microbiology. V. 145, 1999. P. 2375–2384.

25. Danilova L.V., Pervouchine D.D., Favorov A.V., Mironov A.A. RNAKINETICS: A web server that models secondary structure kinetics of an elongating RNA // Journal of Bioinformatics and Computational Biology. V. 4, 2006. P. 589-596.

26. Das A., Crawford I.P., Yanofsky C. Regulation of tryptophan operon expression by attenuation in cell-free extracts of *Escherichia coli*. // The Journal of Biological Chemistry. V. 257, No. 15, 1982. P. 8795–8798.

27. Elf J., Ehrenberg M. What Makes Ribosome-Mediated Transcriptional Attenuation Sensitive to Amino Acid Limitation? // PLoS Computational Biology. V. 1, N. 1, 2005. P. 14–23 (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1183508)

28. Eskin E., Pevzner P.A. Finding composite regulatory patterns in DNA sequences // Bioinformatics. V. 18, 2002. P. 354–363.

29. Even S., Itai A., Shamir A. On the complexity of timetable and multicommodity flow problems // SIAM Journal on Computing. V. 5, 1976. P. 691–703.

30. Favorov A.V., Gelfand M.S., Gerasimova A.V., Ravcheev D.A., Mironov A.A., Makeev V.J. A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length // Bioinformatics. V. 21, 2005. P. 2240–2245.

31. Gong F., Yanofsky C. Rho's role in transcription attenuation in the *tna* operon of *E. coli* // Methods Enzymol. V. 371, 2003. P. 383–391.

32. Grundy F.J., Henkin T.M. The T box and S box transcription termination control systems // Front Biosci. V. 8, 2003. P. d20–31.

33. Hauser C.R., Gillham N.W., Boynton J.E. Translation regulation of chloroplast genes // The Journal of Biological Chemistry. V. 271, 1996. P. 1486–1497.

34. Heery D.M., Dunican L.K. Cloning of the *trp* gene cluster from a tryptophan-hyperproducing strain of *Corynebacterium glutamicum*: Identification of a mutation in the *trp* leader sequence // Applied and Environmental Microbiology. V. 59, 1993. P. 791–799.

35. Henkin T.M., Glass B.L., Grundy F.J. Analysis of the *Bacillus subtilis tyrS* gene: conservation of a regulatory sequence in multiple tRNA synthetase genes // J Bacteriol. V. 174, 1992. P. 1299–1306.

36. Hoffmann C.M. Group-Theoretic Algorithms and Graph Isomorphism // Lecture Notes in Computer Science. V. 136. Berlin / Heidelberg: Springer, 1982.

37. Konan K.V., Yanofsky C. Rho-dependent transcription termination in the tna operon of *Escherichia coli*: Roles of the boxA sequence and the rut site // Journal of Bacteriology. V. 182, 2000. P. 3981–3988.

38. Lin C., Pradkar A.S., Vining L.C. Regulation of an antranilate synthase gene in *Streptomyces venezuelae* by *trp* attenuator // Microbiology. V. 144, 1998. P. 1971–1980.

39. Lyubetsky V.A., Seliverstov A.V. Amino acid biosynthesis attenuation in bacteria // Proceedings of the fourth international conference on bioinformatics of genome regulation and structure, July 25–30, 2004. Новосибирск: ред.-изд. отдел ИЦиГ СО РАН, 2004. Т. 1. С. 307–310.

40. Lyubetsky V.A., Seliverstov A.V. Note on cliques and alignments // Информационные процессы. Т. 4, 2004, С. 241–246.

41. Lyubetsky V.A., Seliverstov A.V. Modeling classic attenuation regulation of gene expression in bacteria // Proceedings of the fifth international conference on bioinformatics of genome regulation and structure, July 16–22, 2006. Новосибирск: ред.-изд. отдел ИЦиГ СО РАН, 2006. Т. 1. С. 102–105.

42. Mandal M., Breaker R.R. Gene regulation by riboswitches // Nat Rev Mol Cell Biol. V. 5, 2004. P. 451–463.

43. Mathews D.H., Disney M.D., Childs J.L., Schroeder S.J., Zuker M., Turner D.H. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure // PNAS. V. 101, 2004. P. 7287–7292.

44. Mathews D.H., Sabina J., Zuker M., Turner D.H. Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure // J. Mol. Biol. V. 288, 1999. P. 911–940.

45. Mironov A.A., Lebedev V.F. A kinetic model of RNA folding // BioSystems. V. 30, 1993. P. 49–56.

46. Nickelsen J. Chloroplast RNA binding proteins // Current Genet. V. 43, 2003. P. 392–399.

47. Rodionov D.A., Vitreschak A.A., Mironov A.A., Gelfand M.S. Computational analysis of thiamin regulation in bacteria: Possible mechanisms and new THI-element-regulated genes // J. Biol. Chem. V. 277, 2003. P. 48949–48959.

48. Schaefer T.J. The Complexity of Satisfability Problems // Proceedings of the 10th Annual ACM Symposium on Theory of Computing, 1978, NY: ACM Press, 216–226.

49. Seliverstov A.V., Lyubetsky V.A. RNA regulatory structures in Actinobacteria and Cyanobacteria // Proceedings of the International Moscow Conference on Computational Molecular Biology. July 18–21, 2005. M., 2005. C. 351–353.

50. Seliverstov A.V., Lyubetsky V.A. Translation regulation in chloroplasts // Proceedings of the fifth international conference on bioinformatics of genome regulation and structure, July 16–22, 2006. Новосибирск: ред.-изд. отдел ИЦиГ СО РАН, 2006. Т. 1. С. 146-149.

51. Seliverstov A.V., Putzer H., Gelfand M.S., Lyubetsky V.A. Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria // BMC Microbiology. V. 5 N. 54, 2005. 14 p.

52. Thompson J.D., Higgs D.G., Gibson T.J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice // Nucl. Acids Res. V. 22, 1994. P. 4673–4680.

53. Uptain S.M., Chamberlin M.J. *Escherichia coli* RNA polymerase terminates transcription efficiently at rho-independent terminators on single-stranded DNA templates // Proc. Natl. Acad. Sci. USA V. 94, 1997. P. 13548–13553.

54. Vitreschak A.G., Lyubetskaya E.V., Shirshin M.A., Gelfand M.S., Lyubetsky V.A. Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis // FEMS Microbiology Letters. V. 234, 2004. P. 357–370.

55. Vitreschak A.G., Rodionov D.A., Mironov A.A., Gelfand M.S. Riboswitches: the oldest mechanism for the regulation of gene expression? // Trends in Genetics. V. 20, 2004. P. 44–50.

56. Wilson K., von Hippel P. Transcription termination at intrinsic terminators: the role of the RNA hairpin // Proc. Natl. Acad. Sci. USA. V. 92 1995. P. 8793–8797.

57. Xayaphoummine A., Bucher T., Isambert H. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots // Nucleic Acids Res. V. 33, 2005. P. W605–610.

58. Yin H., Artsimovitch I., Landick R., Gelles J. Nonequilibrium mechanism of translation termination from observations of single RNA polymerase molecules // PNAS. V. 96, 1999. P. 13124–13129.

59. Zerges W. Translation in chloroplasts // Biochimie. V. 82, 2000. P. 583–601.

60. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction // Nucleic Acids Res. V. 31, 2003. P.3406–3415.