На правах рукописи

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

АВТОРЕФЕРАТ

диссертации на соискание учёной степени
кандидата физико-математических наук

Москва – 2006

Работа выполнена в

Институте проблем передачи информации РАН

Научный руководитель:	д.ф.-м.н. проф. В.А. Любецкий
Официальные оппоненты:	д.б.н. проф. А.А. Миронов,
	д.ф.-м.н. проф. М.Р. Пентус
Ведущая организация:	Институт молекулярной биологии им. В.А. Энгельгарта РАН

Защита диссертации состоится _____2006 года в _____на заседании Диссертационного совета Д.002.077.01 при Институте проблем передачи информации РАН по адресу: 127994, Москва, Большой Каретный пер., 19.

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

Автореферат разослан _____2006 года.

Учёный секретарь диссертационного совета:

д.ф.-м.н. И.И. Цитович

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 нахо-

ждения *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 помо-

щью 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.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении**, раздел 0.1, даны основные определения и обзор алгоритмических результатов. Приведены теоремы Фаркаша, Хачияна и теорема Шефера о 3КНФ. Далее обсуждаются различные алгоритмы поиска *сигнала*, т.е. набора наиболее попарно похожих слов, и поиска множественного выравнивания последовательностей. Указана связь задач о поиске сигнала и о поиске клики в многодольном графе. В разделе 0.2 дан обзор результатов по регуляции экспрессии генов у хлоропластов и бактерий. Отмечены примеры белок-РНКового взамодействия у хлоропластов. Кратко описаны ме-

ханизм аттенюаторной регуляции транскрипции и механизм регуляции, основанной на Т-боксах и рибопереключателях, у бактерий. В разделе 0.3 приводится обзор методов для моделирования кинетики формирования вторичной структуры РНК.

В **первой главе** рассмотрены алгоритмы поиска клики в многодольном графе. В разделе 1.1 доказано, что существование n-клики в n-дольном графе с двумя вершинами в каждой доле эквивалентно непустоте многогранника, названного *многогранником квазиклик*, стороны которого можно вычислить за полиномиальное от n время, часть вершин которого взаимно однозначно соответствуют всем n-кликам в таком n-дольном графе. Размерность этого многогранника позволяет оценивать сверху число n-клик в таком графе. С другой стороны, определен многогранник, названный *многогранником клик*, уже все вершины которого взаимно однозначно соответствуют всем n-кликам в таком n-дольном графе. В первой главе доказано, что сложность описания сторон этого многогранника представляет собой алгоритмически трудную задачу в смысле, указанном в теореме 4.

Ниже индексы p, q, r равны 1 или 2, а индексы i, j, k пробегают значения от 1 до n. Для любого целого числа n *определим многогранник квазиклик*, *обозначаемый* P_n, в $4n^2$-мерном пространстве, как выделяемый следующей системой равенств и неравенств: для всех i, j, p, q пусть $X_{ijpq}= X_{jiqp}$, для всех i пусть $X_{ii11}+X_{ii22}=1$, для всех i пусть $X_{ii12}=0$, для всех i, j, p пусть $X_{ijp1}+X_{ijp2}=X_{iipp}$, для всех i, j, p, q пусть X_{ijpq} неотрицательно и меньше или равно единице, для всех i, j, k, p, q, r пусть сумма $X_{iipp}+X_{jkqr}$ больше или равна сумме $X_{ijpq}+X_{ikpr}$.

Лемма 1. *Для любого отображения f множества* {1, 2, ..., n} *во множество* {1, 2} *точка X с координатами $X_{ijpq}=1$, если $p=f(i)$ и $q=f(j)$,* и $X_{ijpq}=0$, *если такое условие не выполнено, является вершиной многогранника квазиклик. Эти точки взаимно однозначно соответствуют всем n-кликам полного n-дольного графа с двумя вершинами в каждой доле.*

По n-дольному графу G, имеющему по две вершины в каждой доле, *определим* аффинное подпространство H(G), выделяемое всеми уравнения-

ми вида $X_{ijpq}=0$, если индекс i не равен индексу j и p-я вершина i-й доли не соединена ребром с q-й вершиной j-й доли.

Теорема 1. *Пусть n-дольный граф G имеет по две вершины в каждой доле. Если пересечение многогранника квазиклик и пространства H(G) непустое, то граф G имеет хотя бы одну n-клику. Если размерность пересечения многогранника квазиклик и пространства H(G) равна единице, то граф G имеет одну или две n-клики.*

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

Обозначим Q_n выпуклую оболочку точек (из многогранника квазиклик), соответствующих всем n-кликам полного n-дольного графа G. Этот многогранник назовём *многогранником клик.*

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

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

В разделе 1.2 предложен алгоритм решения неявно заданной системы однородных линейных уравнений над конечным бимодулем. В частности, он позволяет оценить снизу число n-клик в n-дольном графе, если существует хотя бы одна n-клика. Рассмотренный алгоритм имеет меньшую сложность, чем ранее описанный алгоритм Симса.

Пусть A и B – некоторые кольца, а R – конечный A-B-бимодуль из r элементов. 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 пространстве R^n, которую обозначим B_m.

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 для этого* $O\left((m+n)^2 \cdot m \cdot L \cdot r^4\right)$ *операций в R. Здесь m – чис-*

ло 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, у которой коэффициенты уравнений не известны, но известна эффективная процедура для проверки выполнимости любого уравнения системы при любых значениях переменных. Процедура говорит «да» или «нет».

В разделе 1.3 описан *эвристический* алгоритм поиска клики в многодольном графе G в общем случае и на его основе предложен алгоритм поиска сигнала в наборе невыравненных последовательностей, например, лидерных областей в алфавите из четырех нуклеотидов. А также – модификация алгоритма поиска сигнала, в которой для получения сигнала сложной структуры учитывается дерево видов, из геномов которых взяты эти последовательности – лидерные области соответствующих генов. В этом эвристическом алгоритме существенно используется алгоритм для вычисления числа клик малого размера в данном 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-2)(q-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.

В разделе 1.4 приведены результаты тестирования алгоритма для поиска сигнала на примере известных сайтов связывания белка PurR с ДНК у кишечной палочки. Алгоритм нашёл три сигнала, содержащих 21 сайт в де-

вятнадцати последовательностях, две из которых включают по два сайта. Этот результат совпадает с опубликованными данными.

Алгоритм также нашёл консервативные участки в 5'-нетранслируемых областях мРНК хлоропластов и актинобактерий, которые представляются биологически адекватными; последние подробно представлены и обсуждаются во второй главе. Биологическая значимость этих последних сигналов также подтверждается независимым от поиска клик анализом вторичной структуры РНК и в некоторых случаях экспериментальными данными о соответствующей регуляции экспрессии генов. С другой стороны, для уменьшения *недопредсказаний* применялся поиск структур РНК по образцу, который строился по набору заранее найденных алгоритмом сигналов, с помощью вспомогательных программ.

В разделе 1.5 описаны вспомогательные программы для выделения лидерных областей генов, для поиска спиралей и слов специального вида по их параметрам в аннотированных геномах, для поиска по образцу.

Вторая глава содержит результаты массового поиска потенциальных структур РНК, регулирующих экспрессию генов, у хлоропластов и бактерий. Эти структуры предсказаны с помощью алгоритма поиска клики из первой главы.

В разделе 2.1 рассмотрена регуляция трансляции посредством взаимодействия белков с РНК для различных генов у хлоропластов. Алгоритм из первой главы, примененный для поиска консервативных участков в 5'-нетранслируемых областях генов, выделил протяжённые консервативные участки, включающие шпильки РНК. Эти результаты кратко представлены в табл. 1. Найденные консервативные участки перед генами, содержащими интроны, вероятно, связаны как с регуляцией трансляции, так и с задержкой начала трансляции до завершения сплайсинга.

Таблица 1. Распределение потенциальных сайтов связывания белка перед указанными в таблице генами у хлоропластов.
Обозначения: "+" – сайт связывания белка с РНК найден, "-" – сайт не найден, "s" – соответствующий ген содержит интроны, "n" – соответствующий ген отсутствует 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	+
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, Oryza sativa*	+s	+s	+s	+	+	+
	Triticum aestivum	+s	+s	+s	+	+	+
	Zea mays	+s	+s	+s	+	+	+

В разделе 2.2 рассмотрены системы регуляции экспрессии генов, кодирующих ферменты для биосинтеза аминокислот и аминоацил-тРНК син-

тетазы у актинобактерий. Применение того же алгоритма для поиска консервативных участков в 5'-нетранслируемых областях генов привело к выделению протяжённых участков с консервативной вторичной альтернативной структурой РНК. Кратко эти результаты приведены в табл. 2.

Таблица 2. Распределение предсказанных регуляторных структур РНК у актинобактерий.
Обозначения. "A" указывает на присутствие классической аттенюаторной регуляции, "R" – Rho-зависимой аттенюаторной регуляции, "LEU" – нового типа регуляции при участии LEU-элемента на уровне трансляции, "T" – T-боксовую регуляции на уровне трансляции.

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		

Биосинтез триптофана. Найдена классическая аттенюаторная регуляция оперонов биосинтеза триптофана у всех *Corynebacterium* spp. и у *Streptomyces* spp. У *C. diphtheriae* эта регуляция предсказана для двух оперонов $trpB_1EDGC$ и $trpB_2A$. У *S. avermitilis* она предсказана для триптофанил-тРНК синтетазы $trpS_2$.

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, в которой абсолютно консервативные нуклеотиды показаны прописными буквами. Найденные здесь

структуры классической аттенюаторной регуляции весьма похожи на таковые у протеобактерий.

Биосинтез цистеина. 5'-области оперонов *cys* у всех *Mycobacterium* spp., кроме *M. smegmatis*, у *P. acnes*, и также оперона *cbs* у *B. longum* содержат открытую рамку считывания с последовательностью цистеиновых кодонов непосредственно перед стоп кодоном. Возможно, регуляция транскрипции основана здесь на Rho-зависимой терминации. Эта ситуация аналогична той, которая известна для триптофаназы *tna* у *E. coli*. 3'-нетранслируемая область предполагаемого лидерного пептида содержит UC-богатый мотив, характерный для связывания белка Rho с РНК. Итак, предполагаемая схема регуляции такова: при *недостатке* цистеина участок мРНК вокруг стоп кодоном лидерного пептида закрыт рибосомой столь долго, что РНК полимераза успевает уйти далеко и транскрипция не прерывается. При *избытке* цистеина рибосома быстро завершает трансляцию лидерного пептида, и в результате открывается следующий за ним UC-богатый участок РНК, характерный для Rho-зависимого терминатора. Транскрипция прерывается.

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*) в 5'-нетранслируемой области присутствует характерная консервативная структура, названная LEU-элементом. А именно, LEU-элемент имеет две конформации, включающие короткую открытую рамку считывания с участком лейциновых кодонов. Первая из них включает вторичную структуру РНК с псевдоузлом, плечи спиралей которой высоко консервативны по нуклеотидному составу. Пвсевдоузел лежит в петле спирали (называемой черенком), 5'-плечо которой перекрывает область лейциновых кодонов лидерного пептида, а 3'-плечо перекрывает область Шайна-Дальгарно гена *leuA*. Вторая конформация альтернативна, не имеет псевдоузла и не перекрывает устойчиво область Шайна-Дальгарно этого гена.

Высокая консервативность участков, которые перекрывают область Шайна-Дальгарно, позволила уточнить положение инициирующего кодона гена *leuA* у *C. efficiens* и *T. fusca*. Это уточнение хорошо согласуется с множественным выравниванием соответствующих белков.

Предполагается следующий механизм аттенюации, связанной с LEU-элементом. В первой конформации, случай псевдоузла, черенок стабилен, а во второй конформации черенок нестабилен и область Шайна-Дальгарно не перекрывается. 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 белок, который, образуя комплекс с лейцином, формирует псевдоузел LEU-элемента. Филогенетический профиль, близкий к филогенетическому профилю LEU-элемента, имеют гомологи гипотетического белка ML1624 (596 остатков аминокислот) из *M. leprae*. Гомологи этого белка с качеством выравнивания меньшим, чем $E=10^{-170}$, найдены у всех актинобактерий, которые имеют LEU-элемент перед геном *leuA*. У этого белка с помощью базы PFAM найден N-концевой домен (аминокислоты с 34 по 193), обычный для DEAD/DEAH бокса хеликаз. Этот домен характерен для многих белков, вовлечённых в метаболизм РНК, включая транскрипцию, трансляцию, распад РНК и образование рибосомальных РНК. Можно предположить, что этот белок участвует в образовании псевдоузла в LEU-элементе оперона *leuA*, и в результате перекрытие области Шайна-Дальгарно регулируется концентрацией лейцина.

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. Найденные консервативные элементы характерны для классической аттенюаторной регуляции, но в данном случае характер регуляции не вполне ясен.

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 Т-боксовая регуляция трансляции. В отличие от обычной Т-боксовой регуляции на уровне транскрипции здесь отсутствует терминатор. 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*. Высокая консервативность участков, которые перекрывают область Шайна-Дальгарно, позволила уточнить положение инициирующего кодона гена *ileS* у *C. efficiens*. Это уточнение хорошо согласуется с множественным выравниванием соответствующих белков.

Потенциальная классическая аттенюаторная регуляция 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*.

В разделе 2.3 описана потенциальная регуляция трансляции гена *ykoE* ABC транспортёра посредством тиаминового рибопереключателя у актинобактерий. У актинобактерий *Brevibacterium linens*, *Kineococcus radiotolerans*, *Leifsonia xyli*, *Propionibacterium acnes*, *Thermobifida fusca*, *Corynebacterium diphtheriae*, *Corynebacterium glutamicum* найдена консервативная вторичная структура РНК, характерная для тиаминового рибопереключателя. Вероятно, 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 дополни-

тельной спиралью, а у *C. diphtheriae* и *C. glutamicum* рибопереключатель примыкает непосредственно к области связывания рибосомы.

В разделе 2.4 рассмотрена потенциальная Т-боксовая регуляция трансляции гена *alr3806* у цианобактерии *Nostoc* PCC7120. 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 перед геном *alr3806*, препятствуя трансляции.

Глава 3 посвящена моделированию классической аттенюаторной регуляции биосинтеза триптофана у бактерий. В разделе 3.1 описана математическая модель классической аттенюаторной регуляции. Даны определения микро- и макро- состояний РНК, формулы для констант скоростей переходов, формулы для вычисления замедления РНК полимеразы вторичной структурой, образующейся на участке мРНК между рибосомой и РНК-полимеразой. Выясняется зависимость скорости трансляции от концентрации триптофанил-тРНК и, следовательно, триптофана.

В разделе 3.2 описана 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 процесс, указанный в модели, проигрывался определенное число раз (обычно 10^3-10^4 раз, что дает примерно одинаковый результат) и вычислялось $p(c)$ как доля случаев, в которых происходила терминация.

В разделе 3.3 приведены результаты тестирования модели. В табл. 3 даны результаты вычисления вероятности терминации транскрипции в процессе атенюаторной регуляции оперонов, включающих ген антранилат синтазы, у актинобактерий *Corynebacterium diphtheriae* и *Corynebacterium glutamicum*, у альфа-протеобактерий *Agrobacterium tumefaciens*, *Bradyrhizobium japonicum*, *Rhodopseudomonas palustris*, *Rhizobium*

leguminosarum, Sinorhizobium meliloti и у гамма-протеобактерий *Escherichia coli* и *Vibrio cholerae*. И также для гена *trpS*, кодирующего триптофанил-тРНК синтетазу, у *Streptomyces avermitilis*. Параметры модели подбирались так, чтобы у *E. coli* и *C. glutamicum*, для которых аттенюаторные регуляции экспериментально подтверждены, при малой концентрации наблюдался рост частоты терминации с увеличением концентрации триптофанил-тРНК, а при больших концентрациях – насыщение и выход на константу.

Таблица 3. Вероятность терминации $p(c)$ в зависимости от концентрации c триптофанил-тРНК для различных видов бактерий.

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
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 альфапротеобактерий приводит к результатам, качественно согласными с экспериментальными данными.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Любецкий В.А., Селиверстов А.В. Регуляция экспрессии генов биосинтеза аминокислот и аминоацил-тРНК синтетаз у Актинобактерий // Молекулярная биология. Т. 39, № 6, 2005. С. 1072–1075.

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

3. 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.

4. 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.

5. 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. (http://www.bionet.nsc.ru/meeting/bgrs2004)

6. 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.

7. Seliverstov A.V., Lyubetsky V.A. Translation regulation in chloroplasts // Там же. Т. 1. С. 146–149. (http://www.bionet.nsc.ru/meeting/bgrs2006)

8. 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. С. 351–353.

9. 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. (http://www.biomedcentral.com/1471-2180/5/54).

10. Любецкий В.А., Селиверстов А.В. Некоторые алгоритмы, связанные с конечными группами // Информационные процессы. Т. 3, № 1, 2003. С. 39–46. (http://www.jip.ru/Contents.htm).

11. Любецкий В.А., Селиверстов А.В. Многодольные графы с двумя вершинами в каждой доле // Там же. Т. 4, № 2, 2004. С. 127–132.

12. Lyubetsky V.A., Seliverstov A.V. Note on Cliques and Alignments // Там же. Т. 4, № 3, 2004. С. 241–246.

13. Селиверстов А.В., Любецкий В.А. Особенности синтеза цистеина у *Corynebacterium*, *Mycobacterium* и *Propionibacterium* // Там же. Т. 4, № 3, 2004. С. 247–250.

14. Селиверстов А.В., Любецкий В.А. Поиск консервативных участков в лидерных областях генов в случае известного дерева видов // Там же. Т. 5, № 4, 2005. С. 265–270.

15. 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 // Там же. Т. 5, № 5, 2005. С. 337–366.

16. Селиверстов А.В., Любецкий В.А. Регуляция трансляции в хлоропластах // Там же. Т. 5, № 5, 2005. С. 400–404.

17. Селиверстов А.В., Любецкий В.А. Алгоритм поиска консервативных участков нуклеотидных последовательностей // Там же. Т. 6, № 1, 2006. С. 33–36.

18. Любецкий В.А., Селиверстов А.В. Вычисление эффективности регуляции биосинтеза триптофана у бактерий на основе модели классической атенюации // Там же. Т. 6, № 1, 2006. С. 55–57.