ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 44 2008 Вып. 4

УДК 621.391.1

© 2008 г. **В.А. Любецкий**[1], **Е.А. Жижина**[2], **Л.И. Рубанов**[1]

ГИББСОВСКИЙ ПОДХОД В ЗАДАЧЕ ЭВОЛЮЦИИ РЕГУЛЯТОРНОГО СИГНАЛА ЭКСПРЕССИИ ГЕНА

Предложен новый подход к моделированию эволюции нуклеотидных последовательностей с учетом вторичной структуры в них. Подход основан на оптимизации некоторого функционала, в котором помимо стандартной эволюционной динамики первичной структуры заложено требование консервативности вторичной структуры. Обсуждаются результаты моделирования на примере эволюции классической аттенюаторной регуляции.

§ 1. Введение и общая постановка задачи

Задача реконструкции эволюции множества видов или множества генов (белков) по современному состоянию этого множества хорошо известна и давно изучается (см., например, [1–3]). Эволюция описывается *филогенетическим деревом*, которое определяет родство состояний процесса эволюции и описывает ход эволюционных событий, приведших от предковой последовательности (в корне) к современным наблюдаемым последовательностям, заданным в листьях дерева. При этом задача реконструкции эволюции ставится в одном из двух вариантов: либо строится филогенетическое дерево вместе с предковыми состояниями во всех его внутренних вершинах, либо дерево предполагается уже известным и ищутся только предковые состояния, т.е. последовательности во внутренних вершинах. Предковые состояния, как и современные, задаются последовательностями в некотором алфавите. В статье будут рассмотрены последовательности в алфавите из четырех букв {A, C, T, G}, которые называются нуклеотидами.

По современным представлениям в геноме вида главную роль играют гены и в равной мере особые участки генома, обычно расположенные перед генами. Такой участок включает и поддерживает определенный достаточно высокий уровень функционирования своего гена ("экспрессирует" этот ген) или, напротив, выключает его работу, а точнее, понижает уровень его функционирования ("не экспрессирует" ген). Такой участок будем называть сайтом регуляции, или регуляторным участком. Экспрессирование или не экспрессирование гена представляют собой два альтернативных состояния сайта регуляции: одно из них называют антитерминацией (тогда соответствующий ген экспрессируется), а другое – терминацией (тогда ген не экспрессируется). Эти состояния реализуются посредством специальных сложных механизмов, которые могут применяться по отдельности или совместно в некоторой комбинации. В статье рассматривается один из таких механизмов, называемый классической аттенюаторной регуляцией. Этот механизм описан на биологическом

[1] Работа выполнена при частичной финансовой поддержке Международного научно-технического центра (номер проекта 3807).

[2] Работа выполнена при частичной финансовой поддержке Российского фонда фундаментальных исследований (номера проектов 07-01-92216-НЦНИЛ-а; 08-01-00105-а) и Американского фонда гражданских исследований и развития (CRDF) (Grant RUM1-2693-MO-05).

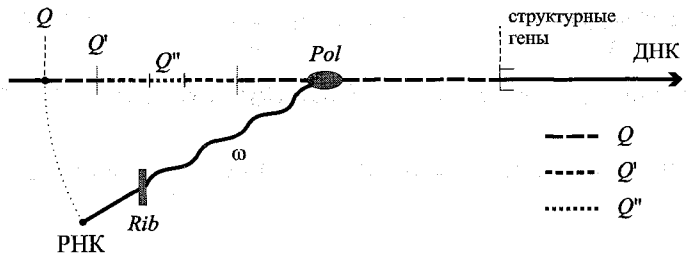

Рис. 1. Механизм классической аттенюаторной регуляции. РНК-полимераза *Pol* транскрибирует по возможности последовательность Q до начала структурных генов. Рибосома *Rib* транслирует ген лидерного пептида Q'. Движение *Rib* на регуляторных кодонах Q'' контролируется концентрацией регулируемой аминокислоты. Вторичная структура ω на матричной РНК между *Rib* и *Pol* тормозит *Pol* и иногда срывает ее с последовательности Q. Если *Pol* достигает структурных генов, то они экспрессируются, т.е. в свою очередь транскрибируются и затем транслируются. Мы используем одни и те же обозначения Q, Q' и Q'' для, соответственно, всей последовательности перед структурными генами, гена лидерного пептида и регуляторных кодонов на ДНК и РНК

уровне в [4] и на более строгом математическом уровне в [5]. Классическая аттенюаторная регуляция впервые была предсказана в [6], принципиально важный шаг в ее моделировании был сделан в [7].

Хотя в этой статье механизм функционирования аттенюаторной регуляции непосредственно не изучается, мы напомним его в более биологическом контексте на примере классической аттенюаторной регуляции, к которой как раз относятся приводимые ниже примеры 1–3. Согласно основной догме молекулярной биологии считывание информации с ДНК происходит в два этапа: сначала синтезируется матричная РНК (однонитевый аналог ДНК – цепочка нуклеотидов, переносящая информацию), а потом на ней синтезируется белок. Синтез РНК-копии называется транскрипцией и осуществляется молекулярной машиной, называемой РНК-полимеразой. Синтез белка по РНК называется трансляцией и осуществляется молекулярной машиной, называемой рибосомой. При синтезе белка рибосома, считывая по три нуклеотида, т.е. по одному кодону, присоединяет к растущей цепи белка аминокислоту, соответствующую прочитанному кодону (см. рис. 1). Кодон – это тройка нуклеотидов, которая по универсальному закону кодирует одну аминокислоту; часто несколько кодонов кодируют одну и ту же аминокислоту. Регуляция уровня экспрессии гена (количества синтезированного белка) зависит от внешних условий. Регуляция происходит на нескольких уровнях. В данной статье рассматривается один тип регуляции – аттенюация.

Аттенюаторная регуляция основана на возможности формирования альтернативных структур, одна из которых разрешает синтез белков, а другая запрещает. Формирование таких структур происходит следующим образом. Нуклеотиды матричной РНК, даже находящиеся в сайте на большом расстоянии друг от друга, могут спариться последовательно расположенными парами из некоторого фиксированного списка возможных комплементарных пар нуклеотидов, и в этом смысле спариваются четверками нуклеотидов. Обычно этот список включает пары G-C и T-A, а с некоторыми ограничениями и пару G-T. При этом каждая буква в каждый момент времени может участвовать только в одном спаривании. Спаривание представляет собой некоторый род гидрофобной связи, называемой стекингом; вместе с нею иногда учитывается водородная связь соответствующих нуклеотидов. Все эти понятия подробно формулируются, например, в [4,5]. В [5] изложен один из возможных способов вычисления энергии спирали, образованной в результате стекинг-взаимодействия.

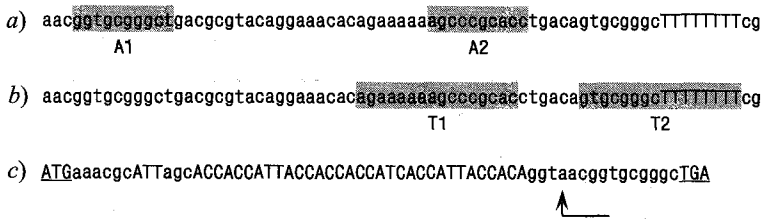

Рис. 2. Пример регуляторного сигнала – классической аттенюаторной регуляции: *a*) – последовательность, где показаны два плеча спирали антитерминатора – состояние А сайта регуляции; *b*) – та же последовательность, на которой показаны два плеча спирали терминатора – состояние Т сайта регуляции. Т-участок показан заглавными буквами; *c*) – последовательность, которая служит геном лидерного пептида: она содержит старт- и стоп-кодоны, показанные заглавными буквами с подчеркиванием и 12 регуляторных кодонов, показанных заглавными буквами без подчеркивания. Последовательности *a*) и *b*) являются продолжениями последовательности *c*), начиная с позиции, показанной стрелкой

Группа последовательно идущих пар формирует спираль. Участки сайта, которые формируют спираль, называют ее плечами. Участок между двумя плечами спирали называют петлей. Плечи спиралей – не обязательно сплошные участки, они могут содержать небольшое число неспаренных нуклеотидов, такие места называются выпячиваниями; различают односторонние и двусторонние выпячивания, последние еще называют внутренними петлями. Образование таких пар приводит к тому, что в РНК формируется набор спиралей – вторичная структура. Часто в одной молекуле РНК может сформироваться несколько альтернативных вторичных структур: в их числе антитерминаторная спираль А и терминаторная спираль Т (см. рис. 2). Спираль А формируется за счет спаривания одной пары участков сайта регуляции (на рис. 2,*a* они помечены А1 и А2), а спираль Т формируется за счет спаривания другой пары участков сайта (на рис. 2,*b* они помечены Т1 и Т2). Принципиально важно, что плечи А2 и Т1 существенно пересекаются: это означает, что нахождение одновременно в состояниях А и Т невозможно, так как любой нуклеотид может участвовать только в одном спаривании. Поэтому подается ровно один сигнал: А или Т. Сам сайт, или последовательность, иначе называется *первичной структурой*, а сайт вместе с одной или несколькими спиралями в нем – *вторичной структурой*. Спираль Т обладает дополнительным свойством – после нее в последовательности идет несколько нуклеотидов Т, образующих Т-участок (полиурацил).

Для механизма классической аттенюаторной регуляции обязательно наличие гена лидерного пептида, расположенного перед антитерминатором, как показано на рис. 2,*c*. Ген лидерного пептида, как и любой ген, состоит из последовательной цепочки кодонов, которые располагаются друг за другом без пересечения и разрывов. Ген лидерного пептида всегда начинается со старт-кодона, включает некоторое число регуляторных кодонов, кодирующих ту аминокислоту, концентрация которой в клетке регулируется, и заканчивается стоп-кодоном. Иногда вместо одной аминокислоты регулируется группа биохимически связанных между собой аминокислот. Ген лидерного пептида регулирует скорость считывания регуляторного сайта в зависимости от концентрации регулируемой аминокислоты в клетке. При наличии регулируемой аминокислоты рибосома транслирует ген лидерного пептида и по ходу своего движения расплетает спираль А. В результате может образоваться спираль Т, которая служит сигналом для прекращения синтеза РНК. Если же регулируемая аминокислота отсутствует, или ее мало, то рибосома останавливается на регуляторных кодонах и не может разрушить спираль А. В результате спираль Т не образуется, и синтез РНК продолжается. Поэтому в отсутствие регулируемой ами-

нокислоты синтезируются РНК-копии структурных генов, ответственных за биосинтез регулируемой аминокислоты. Тем самым, устанавливается обратная связь между имеющимися в клетке молекулами регулируемой аминокислоты и биосинтезом новых молекул той же аминокислоты. В Приложении указано, как в нашей модели учитывается ген лидерного пептида.

Другими примерами аттенюаторных регуляций служат Т-боксовая регуляция, регуляции, основанные на РНК-переключателях, LEU-элементе и др. Аттенюаторные регуляции в основном служат для экспрессирования структурных генов, которые кодируют ферменты метаболизма аминокислот, нуклеозидов, витаминов, и некоторые другие, например, аминоацил-тРНК-синтетазы. Важную роль в функционировании клетки играют и не аттенюаторные регуляции, например, белок-ДНК регуляция, которая применяется к уже более широкому кругу структурных генов. Заметим, что общее число типов регуляторных механизмов экспрессии генов, повидимому, сильно ограничено, и многие из них уже известны. Поэтому представляется обоснованным моделирование эволюции каждого из них сначала в отдельности. Перечисленные здесь общебиологические понятия, как и описание, в частности, самой классической аттенюаторной регуляции, можно найти, например, в [4, гл. 3].

В статье рассматривается задача реконструкции эволюции сайтов классической аттенюаторной регуляции на основе биологически мотивированных принципов стандартной эволюции первичной структуры и консервативности вторичной структуры. При этом структура филогенетического дерева предполагается известной (см. рис. 3, 5 в примерах 1, 2).

Свойства марковских процессов и гиббсовских полей на деревьях хорошо изучены в случае простого спинового пространства, например, для модели Изинга на деревьях [8–10]. Математическими моделями эволюции в основном также служат марковские процессы на деревьях с более сложным спином, представляющим собой последовательность конечной длины. В этих моделях эволюционирующая последовательность может быть нуклеотидной, т.е. в алфавите из четырех букв {A,C,T,G}, соответствующих четырем нуклеотидам, или аминокислотной, т.е. в алфавите из двадцати букв, соответствующих двадцати аминокислотам. Существует несколько очень простых моделей эволюции нуклеотидных последовательностей (Джукса – Кантора, Кимуры и т.д.), каждая из которых определяется матрицей размера 4×4, задающей скорости замены одного нуклеотида на другой, при этом предполагается, что все нуклеотиды в последовательности эволюционируют независимо (обзор таких моделей приведен, например, в [11]). Таким образом, в этих простейших вероятностных моделях предполагается, что нуклеотиды в процессе эволюции заменяют друг друга (мутируют) в соответствии с фиксированной переходной матрицей. В других моделях кроме замены нуклеотидов допускаются делеции (удаления одного нуклеотида или целого участка эволюционирующей последовательности), инсерции (вставки нуклеотида или участка последовательности), дупликации (удвоения) участка генома и другие виды рекомбинации (перестройки) генома. Перечисленные изменения последовательности – замены букв, вставки, удаления и т.д. – соответствуют реальным процессам в клетке, они описывают изменение ее первичной структуры и не учитывают взаимодействие между далекими участками последовательности, возникающее в результате образования вторичной структуры.

Все известные авторам модели, учитывающие вторичную структуру, строятся однотипно и представляют собой совокупность независимых марковских цепей, каждая из которых моделирует эволюцию одного нуклеотида или одной пары нуклеотидов с разными матрицами переходных вероятностей (см. например, [12–15]). В этих моделях позиции в последовательности разбиваются на два типа: связанные пары позиций и свободные одиночные позиции. В последних стоят нуклеотиды, которые эволюционируют по одной из моделей замены букв независимо от других нуклеотидов в последовательности. Нуклеотиды, находящиеся в связанных позициях, эво-

люционируют парами. Для них вводятся так называемые принудительные мутации, когда изменение одного нуклеотида из пары с большой вероятностью влечет изменение второго, так чтобы снова получилась допустимая пара с высокой энергией (от стекинг взаимодействия). При этом каждая пара эволюционирует независимо от других пар связанных позиций. Состояние в такой модели описывается последовательностью, в которой на каждой свободной позиции находится один из четырех нуклеотидов, и в каждой паре связанных позиций находится одна из допустимых пар нуклеотидов. В зависимости от модели рассматриваются либо шесть допустимых пар нуклеотидов, наиболее часто спариваемых в плечах, либо к этим парам добавляется еще одно состояние, соответствующее всем остальным редко встречающимся парам, и наконец, могут рассматриваться все 16 возможных пар (см. [12,13]). Таким образом, эти модели учитывают эволюцию вторичной структуры весьма ограниченным образом, так как представляется затруднительным и неестественным заранее разделить позиции сайта на свободные и связанные.

В статье предлагается новый подход к моделированию эволюции вдоль данного филогенетического дерева сайта регуляции экспрессии гена вместе с вторичной структурой в нем. При этом вторичная структура не привязывается к выделенным позициям сайта, а задается некоторым довольно сложным нелокальным потенциалом взаимодействия. Наш подход основан на гиббсовской форме апостериорного распределения, что приводит к поиску конфигураций, на которых достигается глобальный минимум некоторого функционала "энергии" H. Конфигурациями в этой модели служат наборы сайтов, приписанных всем вершинам данного дерева. Конфигурации, на которых функционал энергии H достигает глобального минимума, назовем *минимальными конфигурациями*, и их множество обозначим $E_{\min}$. Для поиска глобального минимума используется стохастическая процедура аннилинга, которая строит дискретную траекторию $\{\sigma(n) \mid n \in \mathbb{N}\}$, всегда сходящуюся к одной из минимальных конфигураций $\widehat{\sigma} \in E_{\min}$. Алгоритм представляет собой стохастическую итеративную схему, в которой на каждом шаге от n к $n+1$ необходимо принять некоторое стохастическое решение (описание алгоритма см. в § 3). В русскоязычной литературе аннилинг также называют "искусственным отжигом", или процедурой "отпуска".

Итак, наш подход сводится к построению всех минимальных конфигураций, которые за счет специального слагаемого в H согласуются с современными данными – известными регуляторными сайтами, приписанными листьям дерева. Заметим, что мы не предполагаем какой-либо вторичной структуры, заданной в современных сайтах в листьях, и тем более не предполагаем известным какое-либо множественное выравнивание этих сайтов. Наоборот, при тестировании мы задавали только первичную структуру в листьях, а полученную в результате наших вычислений вторичную структуру в листьях сравнивали с вторичной структурой, независимо предсказанной на основе биоинформатических методов и экспериментальных данных.

В качестве приложения нашего метода рассмотрена важная задача построения множественного выравнивания последовательностей с учетом предполагаемой общей вторичной структуры в них. Индуцированная эволюционным процессом вторичная структура из минимальной конфигурации позволяет выделить эволюционно консервативные спирали в современных сайтах и на этой основе получить их множественное выравнивание с учетом сохранения этих консервативных спиралей. Понятие множественного выравнивания приведено ниже (также см., например, [11,16]).

Наш подход к моделированию эволюции нуклеотидной последовательности с учетом ее вторичной структуры был представлен в [17].

§ 2. Модель эволюции последовательности с учетом вторичной структуры в ней

Дано конечное дерево G эволюции нуклеотидной последовательности с множеством узлов (вершин) V и заданными длинами ребер $\{t_j\}$, которые интерпретиру-

ются как время эволюции вдоль соответствующего ребра. Мы рассматриваем здесь случай бинарного дерева, хотя наша модель не зависит от этого обстоятельства. Множество листьев (концевых вершин) дерева G обозначим $V_1 \subset V$. Также дана функция θ, которая каждому листу приписывает последовательность в четырехбуквенном алфавите – современные данные. Конфигурация σ определяется как отображение множества V узлов в множество Q всех конечных последовательностей в четырехбуквенном алфавите, множество $\Sigma = Q^{|V|}$ назовем множеством конфигураций, $\sigma \in \Sigma$. В контексте гиббсовских полей последовательность в каждом узле дерева G можно назвать спином; позицию в последовательности (спине) в биологической литературе часто называют сайтом, используя это слово, конечно, в другом смысле, чем сайт регуляции.

Каждому спину σ_k, $k \in V$, сопоставим множество h_k всех потенциальных спиралей $h_k = \{h_{k,m}\}$ в σ_k. Обычно в это множество включают спирали с энергией ниже некоторого порога (более устойчивое состояние соответствует в нашей модели меньшему значению энергии) и удовлетворяющие ряду других ограничений, например, длины плеча и петли не менее трех. Энергия спирали, порожденная стекинг-взаимодействием, определяется, например, в [5]. Последовательности из какой-то конфигурации σ, которые сопоставлены листьям дерева G, назовем концевыми последовательностями этой конфигурации σ.

Ниже предлагается функционал $H(\sigma)$, точки глобального минимума которого – аргумент $\widehat{\sigma}$ – описывают возможные варианты искомой эволюции последовательности вдоль дерева с учетом вторичной структуры в ней. Функционал $H(\sigma)$ включает три условия (ограничения) на искомую конфигурацию $\widehat{\sigma}$: 1) для любого узла k в соответствующей ему последовательности σ_k в каждой ее позиции $i = 1, \ldots, n$ происходит независимая замена букв в соответствии с матрицей замен R и также происходят вставки и удаления (слагаемое H_1 – парное априорное взаимодействие между спинами, соседними по ребру); 2) значения концевых последовательностей конфигурации σ близки к соответствующим современным последовательностям θ (слагаемое H_2 – влияние данных); 3) последовательности σ_k из конфигурации σ по возможности сохраняют вторичную структуру от начала ребра к его концу и вдоль целого пути в дереве, т.е. в течение многих поколений; при этом функционал меньше, если такие пути длиннее и их больше (слагаемое H_3 – нелокальное парное априорное взаимодействие, отражающее требование консервативности вторичной структуры). Слагаемое H_1 описывает стандартную динамику первичной структуры, а слагаемое H_3 – динамику вторичной структуры.

В число эволюционных изменений первичной структуры входят делеции и инсерции, поэтому последовательности, приписанные концам любого ребра, имеют разные длины, и отсюда нарушается естественное соответствие позиций. Впрочем, и при одинаковых длинах этих последовательностей часто неверно, что буквы, расположенные в одной и той же позиции, эволюционируют согласованно. По этой причине для каждого ребра необходимо установить соответствие позиций в последовательностях s и s', приписанных его концам. Это делается с помощью процедуры, называемой парным выравниванием, которая состоит в добавлении в одну или в обе последовательности знаков пробела. Выравненные последовательности, т.е. последовательности, полученные в результате этой процедуры, являются словами в пятибуквенном алфавите и всегда имеют одинаковую длину. Пара выравненных последовательностей называется выравниванием. Процедура, в которой устанавливается соответствие позиций не двух, а n последовательностей, $n \geq 3$, называется множественным выравниванием. Если $n \geq 5$, то она гораздо сложнее с вычислительной точки зрения, чем парное выравнивание. Расстановка пробелов в парном выравнивании выполняется так, чтобы получить максимум функции $\varphi(s, s')$ сходства новых последовательностей $\overline{s}$ и $\overline{s}'$, получаемых в результате выравнивания. Эту функцию

можно определить следующим образом:

$$\varphi(s,s') = N_e a_e + N_t a_t + N_v a_v + \sum_k (a_d + a_g(l_k - 1)). \quad (1)$$

Здесь N_e – число позиций в выравнивании с одинаковыми буквами в $\overline{s}$ и $\overline{s}'$, N_t – число позиций, в которых имеет место "родственная" замена нуклеотидов, т.е. когда A заменяется на G (и наоборот) или C на T (и наоборот), N_v – число позиций с "перекрестными" заменами, т.е. когда имеют место все оставшиеся варианты перемены букв. Суммирование по k ведется по всем связным участкам длины $l_k \geq 1$, которые определяются так, что в каждой позиции участка одна из двух последовательностей содержит знак пробела. В приведенных ниже примерах параметры этой функции $\varphi(s,s')$ принимают следующие значения: $a_e = 1$, $a_t = -0{,}8$, $a_v = -1{,}2$, $a_d = -2$, $a_g = -1$. Для построения такого парного выравнивания (в отличие от множественного выравнивания) известны быстрые алгоритмы на основе динамического программирования, приведенные, например, в [16].

Итак, пусть

$$H(\sigma) = H(\sigma,\theta) = H_1(\sigma) + H_2(\sigma,\theta) + \lambda H_3(\sigma) \quad (2)$$

– функционал энергии. Напомним: первое слагаемое $H_1(\sigma)$ отражает энергию парного взаимодействия системы спинов и, в частности, априорную информацию о дереве G, второе слагаемое $H_2(\sigma,\theta)$ отражает влияние данных θ в листьях, а третье слагаемое $H_3(\sigma)$ отражает требование консервативности вторичной структуры на концах каждого ребра и вдоль путей в дереве.

Теперь уточним слагаемые в $H(\sigma)$. Обозначим через σ_j и σ'_j последовательности, приписанные в конфигурации σ началу (ближайшей к корню дерева вершине на ребре) и концу ребра j соответственно. Для вычисления слагаемого $H_1(\sigma)$ нужно для каждого ребра сначала выполнить парное выравнивание последовательностей σ_j и σ'_j, как описано выше. Результат выравнивания – две последовательности, отличающиеся от исходных только добавлением в них некоторого числа пробелов. Они имеют одинаковую длину n_j, зависящую от ребра j, и обозначаются, соответственно, $\overline{\sigma}_j$ и $\overline{\sigma'}_j$. Итак,

$$H_1(\sigma) = \sum_j H_1\left(\overline{\sigma}_j, \overline{\sigma'}_j\right) = -\sum_j \left(\ln \prod_{i=1}^{n_j}{}' \left(e^{\gamma_i t_j R}\right)\left(\overline{\sigma}_{ji}, \overline{\sigma'}_{ji}\right) - \kappa \sum_m (l_{j,m}+1)^q \right), \quad (3)$$

где внешняя сумма берется по всем ребрам j в дереве G, n_j – длина выравнивания на ребре j, произведение $\prod'$ берется только по тем позициям выравнивания, в которых у обеих последовательностей находятся нуклеотиды, R – матрица скоростей замещения букв в алфавите нуклеотидов, одна и та же в каждой позиции и в каждом узле, t_j – длина j-го ребра. Значение γ_i скорости эволюции последовательности по i-й позиции принято считать случайной величиной, распределенной согласно гамма-распределению с фиксированными двумя параметрами, и затем результаты в некотором смысле усреднять по этому распределению. В рассмотренных ниже примерах 1, 2 для простоты принимается $\gamma_i = 1$ для всех позиций i. Здесь e^{cR} – матричнозначная экспонента с аргументом cR, и $\left(e^{\gamma_i t_j R}\right)(\alpha,\beta)$ – тот элемент матрицы $e^{\gamma_i t_j R}$, который соответствует переходу от буквы α к букве β.

Вторая сумма в (3) берется по тем участкам выравненных последовательностей $\overline{\sigma}_j$ и $\overline{\sigma'}_j$, у которых в каждой позиции участка стоит пробел в одной из последовательностей. Здесь m пробегает число таких участков, и $l_{j,m}$ – длина m-го участка на j-м ребре. Параметры κ и q устанавливают значимость эволюционных событий делеции и инсерции по сравнению с заменой буквы на букву; в примерах 1, 2 используются значения $\kappa = 10$, $q = 1$.

Второе слагаемое определяется из следующего условия: оно должно иметь минимум на конфигурациях, концевые последовательности которых совпадают с современными последовательностями θ в листьях. Поэтому его можно задать, например, как

$$H_2(\sigma) = -\sum_{k \in V_1} \varrho\left(\varphi(\overline{\sigma}_k, \overline{\theta}_k)\right),$$

где $\varrho\left(\varphi(\overline{\sigma}_k, \overline{\theta}_k)\right)$ – некоторая функция, имеющая единственный максимум в точке $\varphi(\overline{\sigma}_k, \overline{\theta}_k) = n(\theta_k)$, где $n(\theta_k)$ – длина последовательности θ_k. В примерах 1, 2 мы рассматриваем предельный случай, когда

$$H_2(\sigma) = \begin{cases} 0, & \text{если} \quad \varphi(\overline{\sigma}_k, \overline{\theta}_k) = n(\theta_k), \quad \forall k \in V_1, \\ +\infty & \text{в противном случае,} \end{cases} \tag{4}$$

т.е. у любой конфигурации концевые последовательности совпадают с современными данными в листьях.

Третье слагаемое можно определить как

$$H_3(\sigma) = H_3(\sigma, h) = -\sum_{j \in V} \Phi(h_j, h'_j), \tag{5}$$

где $h = \langle h_j, h'_j \rangle$ и $h_j = \{h_{jm}\}$, $h'_j = \{h'_{jk}\}$ – два множества спиралей с достаточно низкой энергией, которые построены по двум данным последовательностям σ_j и σ'_j, приписанным концам ребра j. Потенциал Φ отражает сохранность вторичной структуры вдоль ребер дерева G. В Приложении приведены рассмотренные нами более сложные варианты слагаемого $H_3(\sigma)$.

Точный вид потенциала Φ зависит от вида искомой вторичной структуры. В случае классической аттенюаторной регуляции, когда вторичная структура содержит взаимно исключающие терминатор $T = (t_{m1}, t_{m2})$ с плечами t_{m1}, t_{m2} и антитерминатор $A = (a_{m1}, a_{m2})$ с плечами a_{m1}, a_{m2}, где m пробегает множество всех таких пар (A, T), потенциал естественно задать в виде

$$\Phi(h_j, h'_j) = \frac{1}{n_{mk}} \sum_{m,k} \left[\varphi(t_{m1}, t'_{k1}) + \varphi(t_{m2}, t'_{k2}) + \varphi(a_{m1}, a'_{k1}) + \varphi(a_{m2}, a'_{k2})\right]^{X+}, \tag{6}$$

где $[u]^{X+} = u$ при $u > X$ и $[u]^{X+} = 0$ при $u \le X$, и X – фиксированный порог, а φ – функция, определенная формулой (1), n_{mk} – число ненулевых слагаемых под знаком суммы. Иначе говоря, для вычисления $\Phi(h_j, h'_j)$ в двух множествах спиралей h_j и h'_j выбираются всевозможные пары антитерминатор – терминатор, после чего эти пары сопоставляются между собой путем независимого парного выравнивания соответственных плеч антитерминаторов и терминаторов, и для тех пар, у которых сходство выше порога X, вычисляется средняя величина этого сходства. В примерах 1, 2 использовалось значение порога $X = 0$.

Замечание. В качестве $\Phi(h_j, h'_j)$ можно рассмотреть потенциал, характеризующий схожесть вторичных структур *в целом* в последовательностях σ_j и σ'_j на концах ребра j. А именно, если в качестве h_j и h'_j взять множества *всех спиралей*, соответственно, в σ_j и σ'_j с достаточно низкой энергией, то $\Phi(h_j, h'_j)$ можно определить как

$$\Phi(h_j, h'_j) = \frac{1}{n_{mk}} \sum_{m,k} \left[\varphi(h_{m1}, h'_{k1}) + \varphi(h_{m2}, h'_{k2})\right]^{X+}.$$

В этом случае вторичная структура хуже сохраняется вдоль путей в дереве. Результаты вычислений для модели с таким потенциалом показали, что не все пути, составленные из регуляторных структур, доходят от листьев до корня.

Далее мы описываем алгоритм для поиска конфигураций $\widehat{\sigma}$

$$E_{\min} = \arg\min H(\sigma, \theta), \quad \widehat{\sigma} \in E_{\min}, \tag{7}$$

на которых достигается глобальный минимум функционала энергии $H(\sigma)$, определенного формулами (2)–(6), и обсуждаем результаты тестирования предложенной модели эволюции.

§ 3. Алгоритм на основе стохастической динамики

Чтобы не возникло затруднения в понимании, уточним терминологию: для процесса эволюции будет употребляться термин "эволюционная динамика", а динамику, на которой основывается вычислительный процесс, будем называть "стохастической динамикой". В первой динамике случайный процесс развивается во времени от корня дерева к листьям, во второй – происходит случайное изменение конфигурации в целом, т.е. на всем дереве; "время" стохастической динамики никак не связано с "эволюционным временем".

Функционал энергии (2) зависит от огромного числа взаимодействующих переменных и имеет много минимумов, в том числе локальных, лежащих в непосредственной близости друг от друга. Поэтому для поиска его глобального минимума естественно использовать один из стохастических алгоритмов. Мы предлагаем рассмотреть алгоритм аннилинга, построенный на основе стохастической динамики Метрополиса – Хастингса (см., например, [18, 19, 1]). В отличие от детерминированных алгоритмов, быстро приводящих к ближайшему локальному минимуму, стохастические алгоритмы позволяют перебрать все многочисленные минимумы функционала H, не останавливаясь в его локальных минимумах. Физически аннилинг имитирует медленное охлаждение системы до нулевой температуры, в результате чего распределение процесса сосредоточивается на конфигурациях из множества $E_{\min}$ независимо от начальных условий:

$$\lim_{n\to\infty} P(X(n) \in E_{\min}) = 1. \tag{8}$$

Алгоритм реализуется как неоднородная марковская цепь, переходные вероятности которой зависят от текущей конфигурации $\sigma(n)$ и параметра β_n, характеризующего температуру системы. А именно, с помощью алгоритма строится последовательность конфигураций $\{\sigma(n)\}$ ("траектория динамики"), которая начинается, вообще говоря, с произвольной конфигурации $\sigma(0)$ и сходится по вероятности к одной из минимальных конфигураций для любого начального условия.

Введем распределение вероятностей на Σ по гиббсовской формуле:

$$\pi_\beta(\sigma) = \frac{1}{Z_\beta}\, e^{-\beta H(\sigma)}, \tag{9}$$

где $Z_\beta = \sum\limits_{\sigma} e^{-\beta H(\sigma)}$ – нормировочная константа.

Возможные изменения конфигурации $\sigma(n)$ за один итерационный шаг при переходе к $\sigma(n+1)$ состоят в следующем: замена буквы в одной из позиций, вставка, удаление. А именно, в очередном k-м узле дерева (в смысле фиксированного линейного порядка на узлах: перебираются все внутренние узлы дерева в порядке убывания их расстояния до корня дерева) равновероятно выбирается одна из позиций в последовательности σ_k, приписанной этому узлу. Затем выбирается тип события, которое происходит в этой позиции этого узла, в соответствии со следующими вероятностями: $P_c = 0{,}992$ – вероятность замены буквы, $P_i = 0{,}004$ – вероятность вставки и $P_d = 0{,}004$ – вероятность удаления (вставка и удаление считаются равновероятными). Если в качестве типа события выбрана замена буквы, то она выполняется

по симметричной матрице для вероятностей замен, в которой на диагонали находятся нули, для родственных замен принимаются вероятности перехода $5/6$, для перекрестных – вероятности $1/12$, так что сумма по строкам и столбцам равна единице. Указанные значения вероятностей, описывающих изменение конфигурации, вычисляются по явным формулам, которые зависят от двух параметров модели: отношения частоты родственных замен букв (транзиций) к частоте перекрестных замен букв (трансверсий), которое в примерах 1, 2 принимается равным 5, и отношения частоты транзиций к частоте удалений (делеций) или вставок (инсерций), которое полагается равным 100. Наша модель устойчива относительно небольших изменений этих параметров.

Вставка в выбранной позиции производится так: выбирается длина $l = 1, \ldots, 32$ с вероятностью $\frac{2^{-l}}{c}$, $c = 1 - 2^{-32}$, и вставляемое слово состоит из l независимых равновероятных букв. Удаление производится аналогичным образом, причем выбранная позиция считается "серединой" участка, который должен быть удален. Таким образом, предлагается новая конфигурация $\widetilde{\sigma}$, которая отличается от σ не более чем в одной (k-й) вершине. Указанная процедура изменения конфигурации определяет некоторую симметричную переходную матрицу на пространстве спинов Q, которая задает так называемое распределение предложения.

После того как выбрана новая последовательность $\widetilde{\sigma}_k \in Q$ в узле $k \in V$, обозначим через $\widetilde{\sigma}$ новую конфигурацию, отличающуюся от конфигурации $\sigma(n) = \sigma$ значением только в этом одном узле k. Новая конфигурация $\widetilde{\sigma}$ принимается, т.е. $\sigma(n+1) = \widetilde{\sigma}$, с вероятностью

$$q(\sigma, \widetilde{\sigma}) = \exp\left\{-\beta\left[H(\widetilde{\sigma}) - H(\sigma)\right]^+\right\}, \tag{10}$$

где $[u]^+ = u$ при $u \geq 0$ и $[u]^+ = 0$ при $u < 0$. Соответственно, прежняя конфигурация σ сохраняется, т.е. $\sigma(n+1) = \sigma$, с вероятностью $1 - q(\sigma, \widetilde{\sigma})$.

Операции вставки или удаления приводят к изменению длины последовательности, приписанной одному узлу некоторого ребра j. Но даже и в случае замены букв выравнивание последовательностей на концах ребра j может измениться. Поэтому при вычислении энергии $H_1(\widetilde{\sigma})$ взаимодействия спинов в новой конфигурации после выравнивания пары последовательностей $\widetilde{\sigma}_j$ и σ'_j буквы, прежде располагавшиеся в одной позиции, занимают, вообще говоря, новые позиции в новом выравнивании, т.е. старые связи (по позициям) между буквами нарушаются и возникают новые связи. Эта ситуация, естественная для моделей эволюции, радикально отличается от ситуации в моделях статистической физики и приводит к "нелокальному" изменению энергии H_1. Иными словами, при вычислении (10) нужно подсчитывать оба слагаемых $H_1(\overline{\widetilde{\sigma}_j, \sigma'_j})$ в $H_1(\widetilde{\sigma})$ и $H_1(\overline{\sigma_j, \sigma'_j})$ в $H_1(\sigma)$, в то время как в локальном случае изменение энергии H_1 можно получить, вычисляя изменение H_1 только в одной позиции для двух последовательностей $\widetilde{\sigma}_j$ и σ'_j. Аналогично вычисляются изменения энергий H_2 и H_3.

Таким образом, возникает последовательность конфигураций

$$\sigma(0) \Rightarrow \sigma(1) \Rightarrow \ldots \Rightarrow \sigma(n) \Rightarrow \ldots$$

Заданный здесь процесс $\sigma(n)$ на Σ обратим относительно распределения π_β (9) при любом фиксированном β, и в частности, распределение (9) является стационарным распределением этого процесса. Отсюда следует:

Для любой начальной конфигурации $\sigma(0)$ и при любом заданном β

$$\lim_{n\to\infty} P\left(\sigma(n) = \eta \middle| \sigma(0)\right) = \pi_\beta(\eta).$$

Если параметр β (он задает так называемый режим охлаждения) выбирается не постоянным, а достаточно медленно возрастающим к бесконечности $\beta_n \to \infty$, то

построенный выше процесс не является обратимым. Встает вопрос, каким образом задать скорость изменения этого параметра (режим аннилинга), чтобы предельная мера сконцентрировалась на множестве $E_{\min}$. В [18,19] было доказано: если $\beta_n \to \infty$ так, что

$$\lim_{n\to\infty} \frac{\log n}{\beta_n} > C, \tag{11}$$

где константа C зависит от вида функции $H(\sigma)$, то выполняется соотношение (8). При этом от исходной системы взаимодействующих спинов на графе требуется выполнение двух условий: величины

$$H_{\min} = \min_{\sigma} H(\sigma), \quad H_{\max} = \max_{\sigma} H(\sigma), \quad \Delta = H_{\max} - H_{\min}$$

– конечные, а потенциал взаимодействия – локальный, т.е. каждый спин взаимодействует с конечным числом соседних спинов. В нашей модели эти условия выполняются: взаимодействие происходит не более чем с тремя соседними по дереву спинами. Следовательно:

Если параметр β_n меняется согласно (11), *то для любой начальной конфигурации $\sigma(0)$ и любого $\eta \in \Sigma$*

$$\lim_{n\to\infty} P\left(\sigma(n) = \eta |\ \sigma(0)\right) = \widehat{\pi}(\eta),$$

где $\widehat{\pi}$ – мера, такая что $\widehat{\pi}(E_{\min}) = 1$.

Предельные конфигурации $\widehat{\sigma}$, получающиеся в результате применения описанного выше итерационного стохастического алгоритма аннилинга с параметром $\beta_n \to \infty$, растущим согласно (11), образуют множество $E_{\min}$ всех минимальных конфигураций (7).

§ 4. Результаты тестирования модели

Ниже приводятся результаты тестирования модели для случая классической аттенюаторной регуляции. В них принималось $\lambda \in [0{,}2;\ 0{,}3]$, $\beta_n = C \ln^p(n+1)$, где n – номер итерации, а C и p – параметры модели: $C = 0{,}01$, $p = 1{,}5$. При указанных

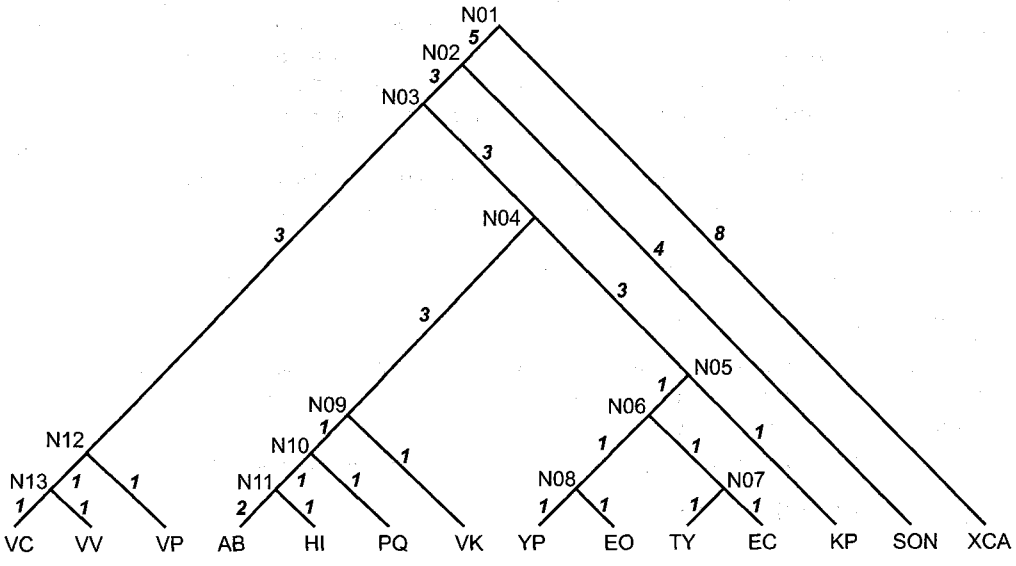

Рис. 3. Дерево видов из примера 1

```
gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc      NO1 -29.2
gGTTGGGGCGGGCTgctgtactcaaaaaatttAAAGAcGAGCCGGCATCCAACaaaGATGCGGGCTTtTTTTTt      NO2 -51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      NO3 -45.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa      N12 -61.3
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAATAaGATACAGGCCTTTTTTTa      N13 -47.5
tGTTGGGGCAGGCTgctgagcgcaaaatttcacAAAAAAGGCCTGTATCCAACcGATACAGGCCTTTTTTta       VC -234.3

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc      NO1 -29.2
gGTTGGGGCGGGCTgctgtactcaaaaaatttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      NO2 -51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      NO3 -45.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa      N12 -61.3
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAATAaGATACAGGCCTTTTTTTa      N13 -61.3
TGTTGGGGCAGGCTgctgagcgaaagaacaaatttcAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa     VV -248.1

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc      NO1 -29.2
gGTTGGGGCGGGCTgctgtactcaaaaaatttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      NO2 -51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      NO3 -45.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcaAAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTTa      N12 -57.1
TGTTGGGGCAGGCTgctgagcgaaagaaattcacAAAAAAGGCCTGTATCCAACAaGATACAGGCCTTTTTTta      VP -182.6

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc      NO1 -29.2
gGTTGGGGCGGGCTgctgtactcaaaaaatttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      NO2 -51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      NO3 -39.1
TGatGGTGCGGGCTgatgcgcacaagaaaaatcAGAAAAAAGCCCGCACCCAacaaaaTGCGGGCTTTTTTTTa      NO4 -24.6
aGAtgGTGCGGGTTagtgctgacaaaaaaaatgaacAAAAAACCCGCATCAactaaaaAGCGGGTTTTTTtata      NO9 -39.0
aaTGGTGCGGGTTagtactggcaaaaaaaatgaacAAAAAACCCGCAaCTCAactaaaAGCGGGTTTTTTtata      N10 -51.0
aaTGGTGCGGGTTagtacggcaaaaaaaagaaacAAAAAACCCGCAaCTCAactgaaAGCGGGTTTTTTtata       N11  -6.2
aaTGGGGCGGGctagtgcgttgaagaatagaattcatGAACCCGCaTTTCCCGAGaGCGGGTTTtttatg          AB -240.5

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc      NO1 -29.2
gGTTGGGGCGGGCTgctgtactcaaaaaatttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      NO2 -51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      NO3 -39.1
TGatGGTGCGGGCTgatgcgcacaagaaaaatcAGAAAAAAGCCCGCACCCAacaaaaTGCGGGCTTTTTTTTa      NO4 -24.6
aGAtgGTGCGGGTTagtgctgacaaaaaaaatgaacAAAAAACCCGCATCAactaaaAGCGGGTTTTTTtata      NO9 -39.0
aaTGGTGCGGGTTagtactggcaaaaaaaatgaacAAAAAACCCGCAaCTCAactaaaAGCGGGTTTTTTtata      N10 -51.0
aaTGGTGCGGGTTagtacggcaaaaaaaagaaacAAAAAACCCGCAaCTCAactgaaAGCGGGTTTTTTtata       N11 -35.0
aaTGGTGCGGGTTagtgcagcaaaaacaagatacAGAAAACCCGCGATTCAactGAATaGCGGGTTTTTTtata      HI -269.3

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc      NO1 -29.2
gGTTGGGGCGGGCTgctgtactcaaaaaatttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      NO2 -51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      NO3 -38.8
TGatGGTGCGGGCTgatgcgcacaagaaaaatcAGAAAAAAGCCCGCACCCAacaaaaTGCGGGCTTTTTTTTa      NO4 -27.8
aGAtgGTGCGGGTTagtgctgacaaaaaaaaATGAAcaaAAAACCCGCACTCaacaaaaaGCGGGTTTTTTTATa     NO9 -46.0
aaTGGTGCGGGTTagtactggcaaaaaaaATGAAcaaAAAACCCGCAacTCAactaaaaGCGGGTTTTTTTATa      N10  -8.6
catagtGCGGGTTTTaatTGGctgaaataatgaaagaATAAACCGAAAACCCGCTacaAGCGGGTTTTTTTGTa       PQ -201.9

gGTTGGGGCGGGCcgctgtcttcgaaaaattttaatgacGAGCCCGCATCCAATaaaGATGCGGGCattTCcctc      NO1 -29.2
gGTTGGGGCGGGCTgctgtactcaaaaaatttAAAGAcGAGCCCGCATCCAACaaaGATGCGGGCTTtTTTTTt      NO2 -51.3
TGTTGGGGCGGGCTgctgcgcacaagaaattccAAAAAAAAGCCCGCATCCAACAaGATGCGGGCTTTTTTTTa      NO3 -39.1
TGatGGTGCGGGCTgatgcgcacaagaaaaatcAGAAAAAAGCCCGCACCCAacaaaaTGCGGGCTTTTTTTTa      NO4 -24.6
aGAtgGTGCGGGTTagtgctgacaaaaaaaatgaacAAAAAACCCGCACTCaacaaaaAGCGGGTTTTTTtata      N09 -19.4
atagtGTGCGGGTTagtgcgtaacaaaaagatcgaattccacAAAACCCGTACTgaataaaaAGTGCGGGTTTTtttatg VK -163.6
```

Рис. 4. Результаты работы алгоритма на данных из примера 1

параметрах в зависимости от начальной точки для достижения одного из предположительно глобальных минимумов функционала H обычно достаточно $10^5 - 10^7$ итераций в работе алгоритма. В однопроцессорной реализации оптимальная продолжительность расчетов для ПК Pentium-4, 3 ГГц составляет от 10 часов до 3 – 5 суток. Также использовался 12-узловой кластер, предоставленный Институтом космических исследований РАН, что привело к ускорению вычислений примерно в 40 раз. Предполагается в дальнейшем реализовать этот алгоритм в параллельном варианте, что качественно ускорит вычисления.

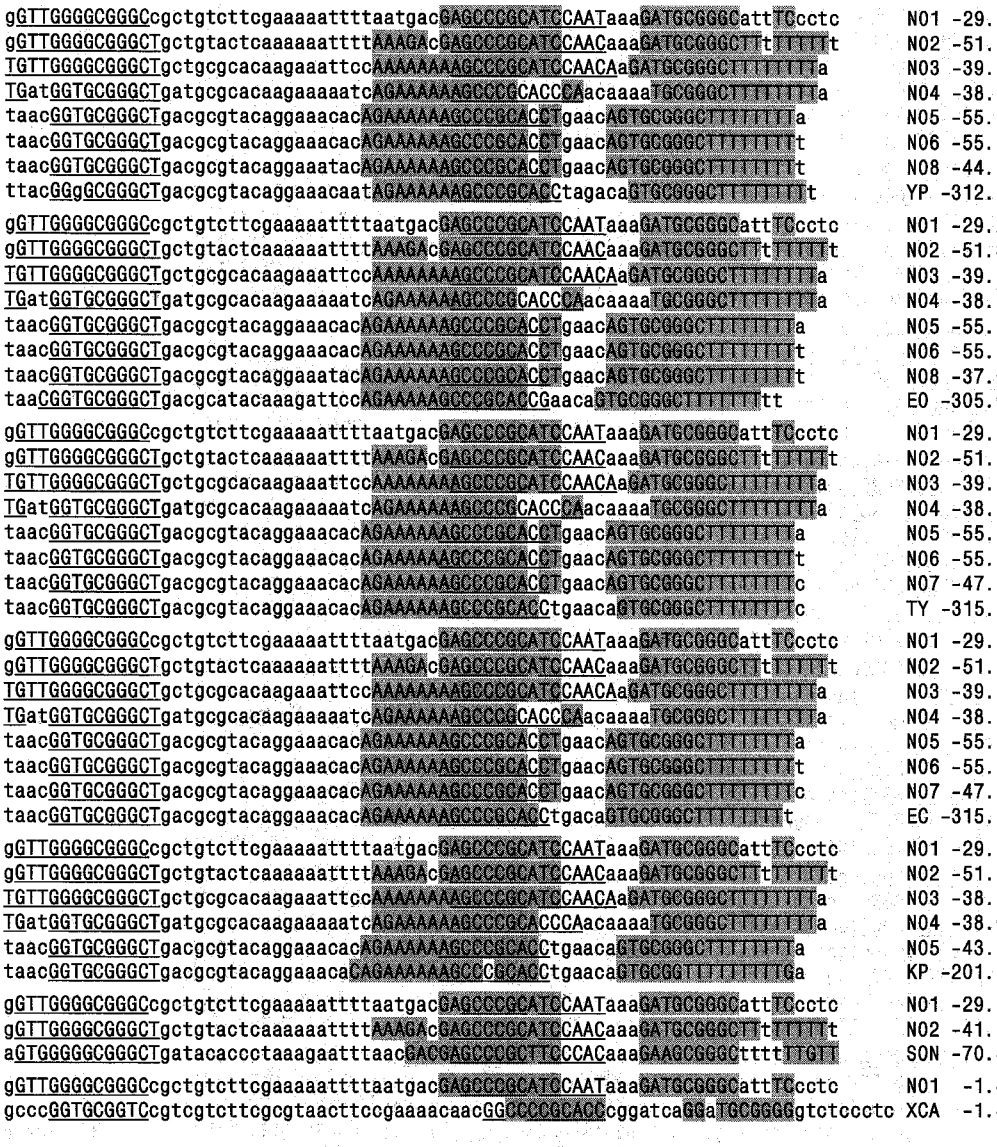

Рис. 4 (продолжение)

Пример 1 (классическая аттенюаторная регуляция биосинтеза треонина у гамма-протеобактерий). Исходные сайты в листьях без знаков пробела заимствованы из [20]. Рассматривается стандартное дерево видов (см. рис. 3) с 27 вершинами и 14 листьями, каждому ребру приписана его филогенетическая длина в условных единицах. Листья помечены именами видов в соответствии с сокращениями: EC – *Escherichia coli*, TY – *Salmonella typhi*, KP – *Klebsiella pneumoniae*, EO – *Erwinia carotovora*, YP – *Yersinia pestis*, HI – *Haemophylus influenzae*, VK – *Pasterella multocida*, AB – *Actinobacillus actinomycetemcomitans*, PQ – *Mannheimia haemolytica*, VC – *Vibrio cholerae*, VV – *Vibrio vulnificus*, VP – *Vibrio parahaemolyticus*, SON – *Shewanella oneidensis*, XCA – *Xanthomonas campestris*.

На рис. 4 показан один из результатов работы нашего алгоритма при $\lambda = 0{,}2$: минимальная конфигурация, для которой $H = 1154$, $H_1 = 1352$, $H_2 = 0$, $H_3 = -990$.

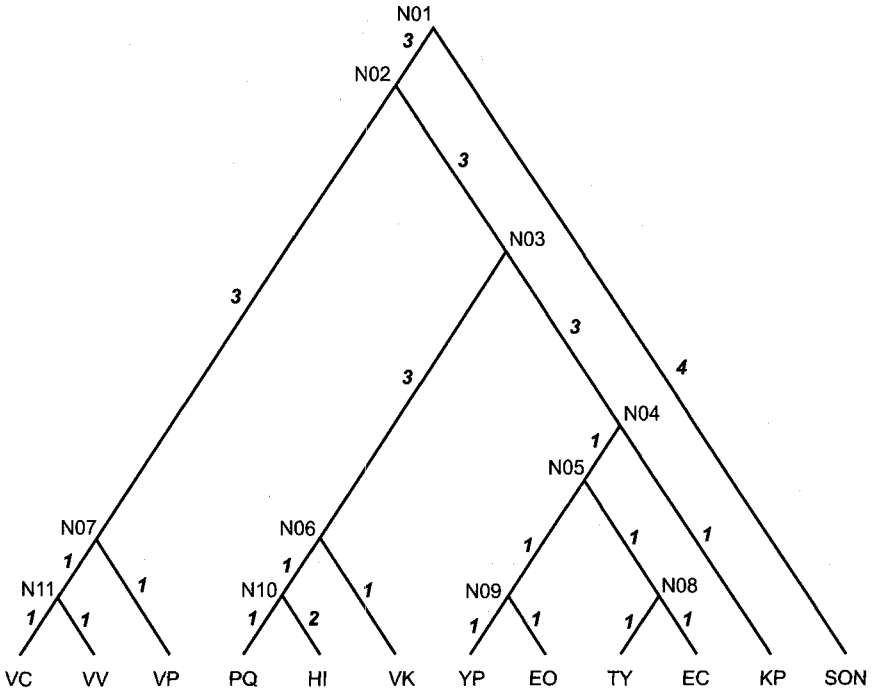

Рис. 5. Дерево видов из примера 2

Предковые последовательности, составляющие эту минимальную конфигурацию, организованы в блоки, каждый из которых определяет путь от одного листа до корня дерева, и на путях показана консервативная вторичная структура. Во всех предковых последовательностях отмечены найденные алгоритмом терминатор (серым фоном) и антитерминатор (подчеркиванием). Напомним, что терминатор и антитерминатор могут содержать небольшие одно- и двусторонние выпячивания, здесь это – нуклеотиды внутри плеч, которые, соответственно, не отмечены серым или не подчеркнуты. Таким образом, выделенные регуляторные структуры для каждого листа, помеченного именем вида, образуют *путь* от пары антитерминатор – терминатор в листе до соответствующей пары в последовательности с номером N01, приписанной корню дерева; например, путь от VC до корня N01. Отметим, что такие пути (возможно, не единственные) существуют для каждого листа. На рис. 4 показаны пути с наименьшей суммарной энергией H_3 по всем ребрам вдоль данного пути. В правом столбце при каждой последовательности указан номер ее вершины на дереве и, кроме того, для внутренних вершин – значение H_3 на соответствующем ребре, а для листьев – суммарное значение H_3 на всех ребрах этого пути.

Пример 2 (классическая аттенюаторная регуляция биосинтеза лейцина у гамма-протеобактерий). Рассматривается дерево видов (см. рис. 5) с 23 вершинами и 12 листьями, являющееся частью дерева, изображенного на рис. 3. На рис. 6 представлен один из результатов работы нашего алгоритма на данных из этого примера при $\lambda = 0{,}25$. Это минимальная конфигурация, для которой $H = 1718$, $H_1 = 1796$, $H_2 = 0$, $H_3 = -310$. На рис. 6 показаны блоки, состоящие из предковых последовательностей этой минимальной конфигурации, которые образуют пути, идущие от каждого листа до корня, с высокой консервативностью вторичной структуры вдоль каждого пути. Остальные обозначения такие же, как в примере 1.

Рис. 6. Результаты работы алгоритма на данных из примера 2

§ 5. Выводы

При моделировании эволюции классической аттенюаторной регуляции предложенный алгоритм строит в предковых вершинах разумную регуляторную вторичную структуру того же типа, который по современным биоинформатическим и экспериментальным данным имеется в первичных структурах, заданных в листьях. Вторичная структура, индуцируемая моделью в исходные первичные структуры в листьях, совпадает или близка к той, которая предсказывается в них по независимым от модели данным (см., например, [20]). Кроме того, первичные структуры, принадлежащие минимальной конфигурации, в листьях и во внутренних вершинах филогенетического дерева имеют хорошее множественное выравнивание, что указывает на хорошую согласованность первичных структур.

Анализ структуры минимальных конфигураций (основных состояний нашей модели) в зависимости от величины параметра λ показал, что в области "умеренных" значений $\lambda \in [0{,}2;\ 1]$ поддерживается сильная регуляторная структура одного типа вдоль всего дерева эволюции. При $\lambda = 0$, т.е. при учете в модели только первичной структуры эволюционирующей последовательности, во всех тестированиях отсутствуют пути, ведущие от современных регуляторных сайтов в листьях к регуляторному сайту в корне, на которых сохранялась бы вторичная структура. Аналогичная картина наблюдается при достаточно малых $\lambda \leq 0{,}1$. При умеренных значениях λ структура минимальных конфигураций меняется: практически от каждого листа до корня строится путь, на котором сохраняется вторичная структура. Наконец, при больших $\lambda > 2$, когда в нашем функционале связь по первичной структуре становится менее существенной, структура минимальных конфигураций снова меняется. Это проявляется в том, что число консервативных по вторичной структуре длинных путей, идущих от листьев к корню, резко сокращается: остаются лишь куски этих путей, наиболее энергетически выгодные для слагаемого H_3. Теперь, в отличие от случая малых λ, последовательности в прикорневых вершинах значительно различаются на уровне первичной структуры и в них можно выделить разнообразные вторичные структуры, которые не участвуют в формировании длинных путей от листьев до корня, а существенны только на отдельных ребрах. Таким образом, при больших значениях λ наблюдается тенденция к формированию изолированных коротких участков с консервативной вторичной структурой, которые оказываются наиболее энергетически выгодными для слагаемого H_3.

Итак, впервые введенное в статье слагаемое H_3, отвечающее за консервативность вторичной структуры в функционале энергии H, существенно меняет свойства минимальных конфигураций; оно оказывается важным для нахождения конфигураций, в которых поддерживается фиксированная регуляторная структура вдоль всего дерева эволюции.

Для сравнения результатов нашего и стандартных алгоритмов к последовательностям в листьях применялись известные компьютерные программы реконструкции предковых последовательностей только по первичной структуре, такие как PAML, PAUP и др. (см. `http://evolution.genetics.washington.edu/phylip/software.serv.html`). На входе предлагались последовательности с пробелами, *заранее выравненные с учетом известной вторичной структуры*, которая приведена в [20]. Даже в такой ситуации программа PAML вообще не построила консервативных вторичных структур нужного типа в предковых последовательностях. Программа PAUP строит их, по-видимому, основываясь на уже заданной вторичной структуре в листьях, но при этом значения H_3 примерно вдвое меньше по абсолютной величине соответствующих значений H_3 для наших минимальных конфигураций.

Модель тестировалась на добавление шума в искусственных и биологических примерах и показала устойчивость результата.

Чтобы убедиться в функциональности предковых сигналов, мы тестировали их с помощью соответствующей модели регуляции. В случае классической аттенюа-

торной регуляции для этого использовалась модель из [5] и основанный на ней сайт `http://lab6.iitp.ru/rnamodel`. Такое тестирование предполагает наличие гена лидерного пептида, поэтому модель применялась к более длинным исходным последовательностям в листьях, которые теперь уже включают ген лидерного пептида (отсутствующий в примерах 1, 2). Результат кратко обсуждается в конце Приложения.

Модель, предложенная в статье, естественно формулируется и для случая бесконечного дерева. Для исследования таких моделей естественно использовать методы и подходы теории гиббсовских случайных полей. В силу сложной структуры спинового пространства и сложного вида взаимодействия спинов полученная система должна иметь весьма нетривиальные свойства, отражающие эволюцию участков генома определенного типа – регуляторных сигналов.

Отметим несколько частных особенностей предложенной модели. Мощность множества $E_{\min}$ всех минимальных конфигураций большая: результаты счета по каждой отдельной траектории в зависимости от выбора начальной точки значительно различаются на уровне первичной структуры особенно в прикорневой области дерева. Но при этом во всех минимальных конфигурациях находятся одинаковые наборы путей с точностью до небольшого сдвига по позиции, составленные из вторичных структур (соответствующих регуляций), которые идут от всех листьев до корня с высокой консервативностью структуры вдоль каждого пути.

Иногда у полученных минимальных конфигураций в отдельных узлах или вдоль некоторых путей, ближе к корню, регуляция становится слабой. Можно предположить, что она там не функционирует, и тем самым, модель может указывать эволюционные периоды, в которых один тип регуляции сменял другой. В отдельных узлах модель предсказывает ансамбль из нескольких антитерминаторов и терминаторов. Например, в узле N04 примера 1 один терминатор, выделенный серым тоном на рис. 4, имеет петлю длины 6 и выпячивание в левом плече длины 4, другой терминатор имеет петлю длины 10 и без выпячиваний в левом плече. В минимальной конфигурации, показанной на рис. 4, пути через N04 до корня идут: от листьев AB, HI, VK, YP, EO, TY, EC – через терминатор с петлей 6, а от листьев PQ и KP – через терминатор с петлей 10. Для этого же примера в узле N05 имеются два терминатора и в узле N09 два терминатора и два антитерминатора, т.е. четыре варианта вторичной структуры. Это может говорить как о роли ансамблей регуляторных структур, так и о эволюционном предпочтении разных структур из однажды возникшего ансамбля при последующей эволюции сигнала вдоль дерева видов.

Кроме описанной выше, были развиты другие модели для построения эволюции регуляторного сигнала и его характеристик с учетом вторичной структуры. В первой из них [21] вторичная структура во внутренних вершинах реконструируется, начиная с листьев, на основе принципа минимальной эволюции, и одновременно от листьев к корню строится множественное выравнивание всех последовательностей. Во второй [22] был разработан алгоритм, который реконструирует матрицы частот нуклеотидов от листьев дерева видов к его корню. Полученные в этих моделях результаты согласуются между собой и с результатами представленной здесь модели.

ПРИЛОЖЕНИЕ

Другие варианты описания консервативности вторичной структуры. Вопрос о том, какой характер консервативности вторичной структуры выразить в слагаемом $H_3(\sigma)$, является наиболее трудным. Как говорилось, биологически мотивированным был бы учет в $H_3(\sigma)$ длин ребер и также консервативности вторичной структуры вдоль целых путей в G, т.е. в течение многих поколений. Соответствующий функционал приводит к модели, в которой взаимодействие между спинами на ребрах дерева G становится нелокальным, теперь еще за счет рассмотрения путей, что заметно усложняет анализ такой модели. Учет длины ребра зависит от описания

среды, в которой происходит эволюция, например, нами рассмотрена простейшая модификация вида $U(\Phi) = t^g\Phi$, где $g = 1{,}5$ – параметр модели.

Для учета путей вместо суммы по отдельным ребрам в слагаемом $H_3(\sigma)$ рассматривались следующие два варианта. Первый задается так:

$$H_3(\sigma) = -\sum_{k\in V_1}\max_{p_k}\sum_{m\in p_k}[\varphi(t_{m1},t_{m'1})+\varphi(t_{m2},t_{m'2})+\varphi(a_{m1},a_{m'1})+\varphi(a_{m2},a_{m'2})]^{X_+},$$

где p_k – какой-то путь от листа $k \in V_1$ до корня дерева, составленный из плеч a_{m1}, a_{m2} антитерминаторов и плеч t_{m1}, t_{m2} терминаторов, взятых в последовательностях σ_m, приписанных вершинам m вдоль этого пути. Моделирование с таким слагаемым $H_3(\sigma)$ приводит к результатам, аналогичным результатам, показанным на рис. 4, 6: вторичная структура вдоль путей от всех листьев до корня становится еще более консервативной, но первичная – менее консервативной.

Второй вариант для $H_3(\sigma)$, более содержательный с биологической точки зрения, задается выражением $H_3(\sigma) = -\sum\limits_{j\in G}\sum\limits_{p(j)\in G} U_{p(j)}(\Phi,\{t_j\})$, где $p(j)$ – пути по дереву G, идущие от ребра j вверх до корня, Φ определяется формулой (6), в которой принимается $U_{p(j)}(\Phi,\{t_j\}) = \prod\limits_{l\in p(j)}\Phi(h_l,h'_l)\dfrac{1}{(1+rt_l)}$ с некоторым параметром r.

Пример 3 (учет гена лидерного пептида). Для учета гена лидерного пептида мы рассматривали те же последовательности, что в примере 1, на этот раз взятые, начиная со старт-кодона гена лидерного пептида. При этом вместо последовательности PQ, у которой лидерный пептид по доступным нам данным известен не полностью, использовалась последовательность AS – *Actinobacillus succinogenes*. Для учета присутствия гена лидерного пептида в функционал энергии H было добавлено дополнительное слагаемое $H_4(\sigma)$, уменьшающее энергию, если в последовательности имеется лидерный пептид, находящийся в правильном месте, т.е. вблизи левого плеча антитерминатора вплоть до его петли, и линейно зависящее от числа регуляторных кодонов в нем до некоторого порогового значения m: $H_4(\sigma) = \begin{cases} -\mu r & \text{при } r \le m, \\ -\mu m & \text{при } r > m, \end{cases}$ где r – число регуляторных кодонов, μ и m – параметры модели. Во всех случаях построенная нашим алгоритмом минимальная конфигурация содержит ген лидерного пептида во всех предковых узлах – тогда как без слагаемого $H_4(\sigma)$ лидерный пептид можно было обнаружить в лучшем случае лишь в 3–4 узлах из 13 внутренних узлов. Полностью пример построенной конфигурации для $\mu = 5$, $m = 12$ приводится на сайте `http://lab6.iitp.ru/docs/anneal/ex4_a.htm`.

Для независимой оценки качества реконструированных предковых последовательностей мы воспользовались моделью из [5] без какого-либо специального подбора ее параметров. С помощью этой модели для всех последовательностей минимальной конфигурации строилась зависимость $p(c)$ частоты преждевременной терминации от концентрации c регулирующих аминокислот (треонина и изолейцина) в интервале от 0 до 1 с шагом 0,05. Для каждого значения концентрации частота исходов (т.е. терминации или антитерминации) оценивалась по результатам 1000 различных траекторий метода Монте-Карло.

Для листьев дерева обнаружено, что в диапазоне концентраций $c \in [0{,}15; 1]$ у всех современных последовательностей наблюдается монотонный с незначительным отклонением рост частоты преждевременной терминации со средней величиной отношения Q максимальной частоты преждевременной терминации к минимальной свыше 3,5 (табл. 1). Как поясняется, например, в [5], монотонный рост зависимости частоты терминации $p(c)$ на достаточно большом интервале значений c и с большим значением размаха Q говорит в пользу функциональности предсказанной регуляторной структуры.

Таблица 1

Доля событий преждевременной терминации (в %) в зависимости от концентрации аминокислоты в листьях дерева видов

c	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	1,0	Q
AB	9	13	20	25	27	35	37	42	44	48	48	50	51	52	54	53	53	55	6,1
EC	16	14	17	17	21	26	28	34	40	48	48	52	54	57	61	63	65	66	4,7
EO	14	12	11	13	19	25	28	35	40	44	50	52	57	62	60	66	68	67	6,2
HI	16	18	20	20	21	23	23	27	25	24	26	26	28	32	29	30	31	32	2,0
KP	21	22	20	25	28	33	36	39	41	47	51	53	58	58	64	61	65	68	3,4
AS	22	22	28	32	35	40	45	50	54	55	58	62	64	64	65	67	65	67	3,0
SON	18	21	23	32	36	41	46	53	56	58	60	63	66	69	69	69	70	74	4,1
TY	21	17	19	23	24	30	33	41	44	48	48	52	58	59	62	60	66	66	3,9
VC	10	14	16	24	34	39	48	51	57	63	64	69	69	70	71	72	75	75	7,5
VK	27	29	32	38	45	50	53	59	61	63	67	70	69	72	73	70	72	72	2,7
VP	48	49	52	51	59	61	64	65	68	68	71	74	72	73	76	74	75	78	1,6
VV	47	46	51	53	56	57	62	65	66	67	69	73	72	74	73	74	75	75	1,6
XCA	26	27	27	28	33	35	37	39	41	39	41	46	43	44	44	46	48	47	1,8
YP	48	51	53	53	59	61	62	65	67	68	69	72	70	72	77	73	74	76	1,6

Таблица 2

Доля событий преждевременной терминации (в %) в зависимости от концентрации аминокислоты во внутренних вершинах дерева видов

c	0,30	0,35	0,40	0,45	0,50	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	Q
N01	15	17	17	16	18	24	21	24	24	23	25	26	24	27	2,0
N02	18	21	26	31	31	37	43	46	50	51	55	54	58	58	3,9
N03	27	29	31	35	40	44	46	47	50	52	52	57	57	59	2,1
N04	30	33	36	40	42	46	50	53	55	58	58	57	61	60	2,1
N05	28	34	39	42	48	56	58	62	62	66	69	71	68	74	3,6
N06	49	50	53	55	59	65	67	67	69	73	73	75	71	76	3,6
N07	24	27	36	36	43	46	53	58	60	63	66	67	69	73	3,4
N08	54	54	58	59	64	65	67	70	68	74	73	76	77	78	2,6
N09	37	41	45	51	55	60	63	66	68	68	72	71	70	71	2,8
N10	54	57	55	57	61	59	63	63	67	66	68	71	71	71	1,4
N11	18	24	24	27	32	35	38	41	44	46	45	51	52	50	2,4
N12	53	53	55	57	59	63	65	68	69	70	70	72	73	73	1,4
N13	46	50	56	56	58	60	64	67	64	70	70	69	72	75	1,5

Во всех внутренних узлах имеет место аналогичная картина (табл. 2), однако диапазон монотонного роста зависимости $p(c)$ оказывается у́же, $c \in [0{,}3;\ 0{,}95]$, и отношение Q составляет в среднем 2,5.

Таким образом, функционал энергии H вместе со слагаемым H_4 позволяет успешно моделировать эволюцию всего регуляторного сайта, включая ген лидерного пептида. Во всех реконструированных последовательностях восстанавливается этот ген и, более того, модель из [5] указывает на наличие аттенюаторного регуляторного сигнала обсуждаемого типа, качество которого, хотя и слабо, улучшается по мере приближения к современному состоянию сигнала.

СПИСОК ЛИТЕРАТУРЫ

1. *Ewens W., Grant G.* Statistical Methods in Bioinformatics: An Introduction. NY: Springer, 2001.
2. Mathematics of Evolution and Phylogeny. NY: Oxford University Press, 2005.
3. *Lyubetsky V., Gorbunov K., Rusin L., V'yugin V.* Algorithms to Reconstruct Evolutionary Events at Molecular Level and Infer Species Phylogeny // Bioinformatics of Genome Regulation and Structure II. Springer, 2005. P. 189–204.
4. *Сингер М., Берг П.* Гены и геномы. Т. 2. М.: Мир, 1998.

5. *Lyubetsky V.A., Pirogov S.A., Rubanov L.I., Seliverstov A.V.* Modeling Classic Attenuation Regulation of Gene Expression in Bacteria // J. Bioinformatics and Computational Biology. 2007. V. 5. № 1. P. 155–180.
6. *Lee F., Yanofsky C.* Transcription Termination at the trp Operon Attenuators of *Escherichia coli* and *Salmonella typhimurium*: RNA Secondary Structure and Regulation of Termination // Proc. Natl. Acad. Sci. USA. 1977. V. 74. № 10. P. 4365–4369.
7. *Миронов А.А., Кистер А.Е.* Теоретический анализ кинетики формирования вторичной структуры РНК в процессах транскрипции и трансляции // Молекулярная биология. 1985. Т. 19. № 5. С. 1350–1357.
8. *Bleher P.M., Ruiz J., Zagrebnov V.A.* On the Purity of Limiting Gibbs State for the Ising Model on the Bethe Lattice // J. Stat. Phys. 1995. V. 79. № 1–2. P. 473–482.
9. *Evans W., Kenyon C., Peres Y., Schulman L.J.* Broadcasting on Trees and the Ising Model // Ann. Applied Prob. 2000. V. 10. № 2. P. 410–433.
10. *Martinelli F., Sinclair A., Weitz D.* Glauber Dynamics on Trees. Boundary Conditions and Mixing Time // Comm. Math. Phys. 2004. V. 250. № 2. P. 301–334.
11. *Дурбин Р., Эдди Ш., Крог А., Митчисон Г.* Анализ биологических последовательностей. М. – Ижевск: Регулярная и хаотическая динамика, 2006.
12. *Muse S.V.* Evolutionary Analyses of DNA Sequences subject to Constraints on Secondary Structure // Genetics. 1995. V. 139. P. 1429–1439.
13. *Savill N.J., Hoyle D.C., Higgs P.G.* RNA Sequence Evolution with Secondary Structure Constraints: Comparison of Substitution Rate Models Using Maximum-Likelihood Methods // Genetics. 2001. V. 157. № 1. P. 399–411.
14. *Telford M.J., Wise M.J., Gowri-Shankar V.* Consideration of RNA Secondary Structure Significantly Improves Likelihood-Based Estimates of Phylogeny: Examples from the Bilateria // Mol. Biol. Evol. 2005. V. 22. № 4. P. 1129–1136.
15. *Kosakovsky Pond S.L., Mannino F.V., Gravenor M.B., Muse S.V., Frost S.D.* Evolutionary Model Selection with a Genetic Algorithm: A Case Study Using Stem RNA // Mol. Biol. Evol. 2007. V. 24. № 1. P. 159–170.
16. Математические методы для анализа последовательностей ДНК. М.: Мир, 1999.
17. *Любецкий В.А., Жижина Е.А., Горбунов К.Ю., Селиверстов А.В.* Модель эволюции нуклеотидной последовательности // Математические методы распознавания образов (ММРО-13): Сборник докладов 13-й Всероссийской конференции. М.: МАКС Пресс, 2007. С. 605–609.
18. *Geman S., Geman D.* Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images // IEEE Trans. Pattern Anal. Machine Intelligence. 1984. V. 6. P. 721–741.
19. *Kirkpatrick S., Gelatt C.D., Vecchi M.P.* Optimization by Simulated Annealing // Science. 1983. V. 220. № 4598. P. 671–680.
20. *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. 2004. V. 234. № 2. P. 357–370.
21. *Горбунов К.Ю., Любецкий В.А.* Модель эволюции нуклеотидной последовательности с учетом ее вторичной структуры // Междунар. научн. конф. "Вычислительная филогеномика и геносистематика". М., 16–19 ноября, 2007. М.: Изд-во МГУ, 2007. С. 68–71.
22. *Gorbunov K., Lyubetsky V.* Reconstruction of Ancestral Regulatory Signals along a Transcription Factor Tree // Molecular Biology. 2007. V. 41. № 5. P. 836–842.

Любецкий Василий Александрович
Жижина Елена Анатольевна
Рубанов Лев Израилевич
Институт проблем передачи информации
им. А.А. Харкевича РАН
lybetsk@iitp.ru
ejj@iitp.ru
rubanov@iitp.ru

Поступила в редакцию
12.02.2008
После переработки
22.04.2008