МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2006, том 40, № 3, с. 497–511

МАТЕМАТИЧЕСКАЯ И СИСТЕМНАЯ БИОЛОГИЯ

УДК 577.053

МОДЕЛЬ РЕГУЛЯЦИИ ЭКСПРЕССИИ ГЕНОВ У БАКТЕРИЙ НА ОСНОВЕ ФОРМИРОВАНИЯ ВТОРИЧНЫХ СТРУКТУР РНК

© 2006 г. В. А. Любецкий, Л. И. Рубанов, А. В. Селиверстов*, С. А. Пирогов

Институт проблем передачи информации Российской академии наук, Москва, 127994

Поступила в редакцию 20.06.2005 г.
После последней доработки 07.11.2005 г.

Предлагается модель, в первую очередь, классической аттенюаторной РНКовой регуляции экспрессии генов с помощью терминации транскрипции. Модель опирается на представление о макросостоянии вторичной структуры в регуляторной области РНК между рибосомой и РНК-полимеразой, на формулы резонансного типа, определяющие величину замедления РНК-полимеразы набором шпилек в той же области, на представление процессов инициации и элонгации транскрипции и трансляции. Специальное внимание уделяется подбору параметров модели. Для проверки модели проведено компьютерное моделирование и получены, в частности, зависимости вероятности терминации транскрипции от величины концентрации нагруженных тРНК, от концентрации аминокислоты для многих регуляторных областей в геномах бактерий (данные приводятся для *trpE* генов из *Streptomyces* spp., *Bradyrhizobium japonicum* и *Escherichia coli*) при различных значениях трех выделенных нами и описанных в статье параметров, которые авторы рассматривают как основные. Полученные зависимости согласуются с доступными экспериментальными данными; в том числе, по форме графиков, относящихся к активности фермента в зависимости от концентрации аминокислоты в культуре (например, активность антранилатсинтазы от концентрации триптофана у *S. venezuela*). Одно из возможных применений модели таково. Сейчас аттенюаторная регуляция предсказывается обычно на основе множественного выравнивания и для этого требуется несколько последовательностей; получение с помощью модели по индивидуальной последовательности характерной для аттенюации (или ее отсутствия) кривой зависимости активности фермента от концентрации (при подходящих параметрах) могло бы рассматриваться как аргумент в пользу наличия или отсутствия аттенюации.

Ключевые слова: аттенюация, модель транскрипционной регуляции, математическая модель в генетике.

MODEL OF GENES EXPRESSION REGULATION IN BACTERIA BY MEANS OF FORMATION OF SECONDARY RNA STRUCTURES, by *V. A. Lyubetsky, L. I. Rubanov, A. V. Seliverstov*, S. A. Pirogov* (Institute of information transmission problems, Russian academy of Sciences, Moscow, 127994 Russia, *e-mail: slvstv@iitp.ru). In this article a model, first, classical attenuation RNA regulation of gene expression by means of transcription termination is offered. The model bases on representation about a macrostate of secondary structure in RNA regulatory region between a ribosome and a RNA polymerase, on the formulas of a resonant type defining the value of deceleration of a RNA polymerase by a set of hairpins in the same region. The special attention is given to selection of parameters of model. To check of model the computer simulation is carried out and the dependences of transcription termination probability from the value of concentration charged tRNA are obtained, in particular, and from concentration of amino acid for many regulatory regions in genomes of bacteria (here data are presented for *trpE* genes in *Streptomyces* spp., *Bradyrhizobium japonicum* and *Escherichia coli*) and at various values of three parameters, which authors consider as the main. The obtained dependences are compounded with the accessible experimental data; including, under the form of the graphs concerning to activity of an enzyme depending on concentration of amino acid (for example, anthranilate synthase from tryptophan in *S. venezuela*). One possible usage: now attenuation is predicted usually by means of multiple alignment, it needs some sequences; the obtaining with the help of model on an individual sequence characteristic for attenuation or its absence of a curve at approaching parameters could be considered as argument for the benefit of presence or absence of attenuation.

Key words: attenuation, model of regulation on transcription level.

*Эл. почта: slvstv@iitp.ru

В настоящее время признается, что регуляция экспрессии генов с помощью механизмов, связанных с формированием вторичных структур РНК, имеет важное значение. Эти механизмы влияют на элонгацию транскрипции или задерживают трансляцию и используют тот или иной медиатор (например, рибосому в случае классической аттенюаторной регуляции) или регуляторный белок, тРНК, кофактор в (случаях других аттенюаторных регуляций) [1–6]. Эти механизмы изучали, главным образом, на примере гамма-протеобактерий и бацилл [7–12]. Среди регуляторов нового типа можно упомянуть Т-боксы [15, 9], недавно открытые "рибосвитчи" [4, 5, 13, 14] и совсем новые гипотетические регуляторные элементы, названные LEU-элементами [16]. Предприняты массовые поиски некоторых новых регуляторных структур [17, 18]. Все это позволяет предполагать, какую функцию имеют гипотетические гены, и заполнить пробелы в схемах метаболизма бактерий [11, 14, 19, 20]. История исследований собственно классической аттенюаторной регуляции кратко изложена во введениях, например, к работам [6, 8, 21].

В целом биоинформатические работы в этой области можно разделить на *систематические исследования* по поиску самих регуляций как известного, так и нового типа, на основе сравнительной геномики и на работы, к сожалению, *немногочисленные* по моделированию этого процесса или составляющих его частей [22–27].

В некоторых из упомянутых работ [22–24] рассматривается моделирование по методу Монте-Карло кинетики сворачивания вторичной структуры РНК на уровне микросостояний и поставлена задача моделирования этого процесса на уровне макросостояний. В других работах [26, 27] метод вероятностного моделирования Монте-Карло применяют для изучения процесса формирования псевдоузлов у вторичной структуры РНК. В них предлагается модель кинетики сворачивания вторичной структуры, основанная на идеях, высказанных ранее [22–24], а также оригинальный прием для ускорения процедуры Монте-Карло, который позволяет исключить повторение пройденных состояний марковской цепи. В нашей модели используется другая, быстрая организация процедуры Монте-Карло, также исключающая повторения. В работе [25] вероятность антитерминации вычисляется по явной формуле, как сумма двух слагаемых. Первое из них – вероятность того, что рибосома находится на одном из регуляторных кодонов и что происходит формирование антитерминатора в тот момент, когда полимераза доходит до Т-богатого участка. Второе слагаемое – умноженная на 0.5 вероятность того, что рибосома покинет стоп-кодон, когда антитерминатор еще не сформировался. Коэффициент 0.5 мотивируется тем, что в такой ситуации с такой вероятностью может формироваться что-либо одно – терминатор или антитерминатор. Далее работа основана на счете по этой формуле, которая не представляется нам вполне ясной.

В нашей работе предлагается модель классической аттенюаторной РНКовой регуляции экспрессии генов с помощью терминации транскрипции, как она описана ранее [21, с. 172–189]. Модель опирается на представления о макросостоянии вторичной структуры в регуляторной области РНК между рибосомой и РНК-полимеразой и на формулы резонансного типа, определяющие величину замедления РНК-полимеразы набором шпилек в той же области. Кроме определения макросостояния мы ставили задачу получения формулы для оценки замедления РНК-полимеразы набором шпилек и, в конечном счете, текущим макросостоянием в изучаемый момент. Была поставлена также задача: представить процессы инициации и элонгации трансляции регуляторных и нерегуляторных кодонов.

Такая модель, как и модели, представленные ранее [22–29], зависит от многих, достаточно произвольных, решений о способах разложения всего процесса на составные части, о выборе математического аппарата для описания частей, о выборе списка и численных значений параметров, о способе сопоставления результатов моделирования с пока малочисленными экспериментальными данными и т.д. Авторы видят путь преодоления этих трудностей в обсуждении и сравнении предлагаемых моделей между собой и с экспериментальными данными. Можно надеяться, что это могло бы стимулировать и соответствующие эксперименты.

ОПИСАНИЕ МОДЕЛИ

Определения микро- и макросостояний. Константы скоростей переходов.

Предполагается, что дана и далее везде фиксирована последовательность в четырехбуквенном алфавите $\{A, C, T, G\}$, будь это регуляторная область в геноме бактерии или случайная последовательность. Например, область, начиная от промотора (в тех редких случаях, когда он известен) или от сайта связывания рибосомы перед лидерным пептидом и до конца участка остатков урацила.

В исходной последовательности выделяют отрезки длиной не менее 3 нуклеотидов – плечи будущих спиралей: $\ldots a_i, \ldots, b_j, \ldots$ При спаривании каких-либо отрезков a_i и b_j одинаковой длины (подразумевается образование водородных связей и стекинга между соответствующими нуклеотидами вдоль всей длины отрезков a_i и b_j) получается спираль γ_i, причем везде предполагается, что спираль γ_i *непродолжаемая*, т.е. к концам этих отрезков a_i и b_j (плечам спирали γ_i) примыкают (вне

плеч) некомплементарные пары нуклеотидов, а промежуток между отрезками (концевая петля спирали) имеет длину не менее 3 нуклеотидов. Модель допускает, вообще говоря, использование любого списка *исходных спиралей*. Выше определен лишь один из возможных вариантов, в котором в качестве исходных берутся все непродолжаемые спирали с указанными ограничениями на размеры плеча и петли.

Все эти представления, как и описание самой классической аттенюаторной РНК-регуляции экспрессии генов в зависимости от концентрации аминокислоты (или нагруженной тРНК, причем концентрация последней, в свою очередь, определяется концентрациями аминокислоты и аминоацил-тРНК–синтетазы) изложены, например, в монографии Сингера и Берга [21, с. 172–189].

Гипоспиралью спирали γ_i называется любая непустая часть $\bar{\gamma}_i$ спирали γ_i, состоящая из двух связанных плеч с длиной *не менее* 3-х нуклеотидов. Здесь и далее *плечами* называются спариваемые отрезки гипоспирали или спирали, концы которых будут стандартно *обозначаться* (считая от 5'-начала исходной последовательности) буквами *A, B, C, D*. *Концевой петлей* называется участок цепи РНК между двумя плечами гипоспирали.

Микросостоянием называется (непустой совместный) набор гипоспиралей, непродолжаемых *в этом наборе* и без псевдоузлов. В этом состоянии никакие две гипоспирали не соприкасаются (т.е. *A* и *D* одной из них не являются оба соседними нуклеотидами к *B* и *C* другой из них). Кроме того, отдельным "начальным" микросостоянием является пустое множество $\varnothing$. "Непродолжаемость в наборе" означает, что плечи входящих в нее гипоспиралей не могут быть удлинены. *Псевдоузлом* называется пара гипоспиралей, у которой ровно одно плечо одной из них пересекается с петлей другой (и, следовательно, находится в этой петле). Объединение всех гипоспиралей от одной спирали, вошедших в данное микросостояние, называется *подспиралью* этой спирали в данном микросостоянии.

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

Диаграммой микросостояния называется обычная скобочная структура, отражающая взаиморасположение всех гипоспиралей в данном микросостоянии, а каждой паре скобок (по другой терминологии: *хорде*) приписывают номер той спирали, из которой взята гипоспираль, соответствующая паре скобок (хорде). Скобочная структура отражает взаиморасположение гипоспиралей в соответствии с обычным правилом: нескольким расположенным друг за другом подряд гипоспиралям соответствует такое же число последовательных скобок ( )( )…( ), а расположению гипоспирали 1 в петле гипоспирали 2 соответствует вложение скобок $((\,)_1)_2$, где внутренняя пара скобок соответствует гипоспирали 1 и внешняя пара скобок – гипоспирали 2. В диаграмме номера спиралей могут неоднократно повторяться, так как из одной спирали можно вывести много гипоспиралей. По микросостоянию, т.е. фактически по списку всех спаренных нуклеотидов, легко выписывают его диаграмму. Но по диаграмме микросостояния нельзя восстановить само микросостояние: диаграмма сохраняет только "геометрию" взаиморасположения гипоспиралей и указание для каждой пары скобок, из какой спирали разрешено брать гипоспираль для этой пары скобок.

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

Макросостоянием называют любую *непустую* диаграмму; "непустая" в том смысле, что она имеет хотя бы одно реализующее ее микросостояние. Для любого микросостояния ω из макросостояния Ω диаграммы для ω и Ω совпадают.

Энергию связи $E_{\bar{\gamma}_j}$ гипоспирали $\bar{\gamma}_i$ получают при суммировании энергий связи всех последовательных пар ее спаренных нуклеотидов на основе стекинга – энергии связи соседних пар. При этом особым образом учитывают стекинг первой и последней пар гипоспирали $\bar{\gamma}_i$ и коаксиальный стекинг $\bar{\gamma}_i$, зависящий еще от микросостояния ω, из которого берется $\bar{\gamma}_i$. Вычисление этой энергии происходит по схеме и с численными значениями, взятыми из опубликованных работ [30–33].

Каждой гипоспирали $\bar{\gamma}_i$ из данного микросостояния ω приписывают число l_i нуклеотидов в ее концевой петле, которые не вошли в петли и плечи других гипоспиралей из этого микросостояния. Это число, зависящее от микросостояния, называется *длиной* концевой петли гипоспирали $\bar{\gamma}_i$ и обозначается l_i. В то же время, гипоспирали $\bar{\gamma}_i$ приписывают *полную длину* ее концевой петли, т.е. длину этой петли без учета спариваний в ней, относящихся к другим гипоспиралям данного микросостояния. Полная длина зависит только от гипоспирали $\bar{\gamma}_i$ и обозначается l'_j.

Микросостоянию ω по определению приписывают две свободные энергии: *энергию связи* гипоспиралей и *энергию петель* гипоспиралей из ω. Далее везде рассматривают нормированные свободные энергии, т.е. энергии, деленные на величину $R \cdot T$, где, например, T равно 310 K. Поэтому все наши формулы для вычисления энергий дают безразмерные величины. Укажем один из вариантов таких формул.

Энергию связи микросостояния ω принимают равной, как и в упомянутых работах [30–33]:

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

где j пробегает все гипоспирали из ω.

Однако для некоторых лидерных областей, в которых множественное выравнивание предсказывает аттенюаторную регуляцию (что экспериментально подтверждается у некоторых из выравненных организмов), расчет вероятности смены макросостояния в нашей модели по приведенной далее формуле (5), которая использует формулу (1), приводит к явно неправильному результату: вероятность терминации с ростом концентрации не возрастает (см. раздел "Результаты модельного счета и обсуждение"). Поэтому вместо формулы (1) нами предложена следующая более общая формула:

$$G_{hel}(\omega) = \frac{1}{RT}\sum_j \left(E_{\tilde{\gamma}_j} - \alpha \frac{l'_j}{1 + \dfrac{l'_j}{l_{\max}}} \right). \qquad (2)$$

Для упомянутых выше лидерных областей наша модель на основе уже формул (5) и (2) при $\alpha > 0$ дала разумные кривые вероятности терминации. Дополнительное слагаемое ("поправка")

$$E(l') = -\alpha \frac{l'}{1 + \dfrac{l'_j}{l_{\max}}}$$

включает параметры α и $l_{\max}$, где $l_{\max}$ равно длине петли l', при которой значение $E(l')$ равно половине его асимптотического значения. В нашем счете типичными были значения $l_{\max} = 10$, $\alpha = 0$ (во многих случаях) и $\alpha = 5$–10 (в некоторых случаях). Вопрос о физической природе взаимодействия, отвечающего за эту поправку, неясен: можно думать о дополнительной энергии связи участка РНК, на котором реализовано микросостояние ω, со стабилизирующими молекулами, об энергии третичной структуры этого участка, например, псевдоузлов и узлов на нем. Параметр α представляется биологически значимым, в частности, об этом свидетельствует анализ полученных в наших расчетах длин цикла смены макросостояний между двумя соседними переходами рибосомы или полимеразы. При значении $\alpha = 0$ в каждом таком цикле макросостояние обычно меняется от десятков до тысяч раз, иногда достигая 50000 и более. С увеличением значения α длина цикла заметно уменьшается, и при $\alpha > 10$ смена макросостояния происходит уже не в каждом цикле хотя бы один раз, т.е. наблюдается стабилизация вторичной структуры между соседними сдвигами рибосомы или полимеразы. Нельзя исключить, что дальнейшее уточнение значений энергий стекинга и петель снимет необходимость в такой поправке.

Энергия петель микросостояния ω принимается равной:

$$G_{loop}(\omega) = \sum_i (1.77 \ln(l_i + 1) + B), \qquad (3)$$

где i пробегает все гипоспирали из ω.

Эта формула хорошо согласуется с обширными таблицами [31, 33] для энергий всех петель при всех $l_i > 2$, если положить $B = 6.5$ для концевых петель и $B = 0$ для двусторонних выпячиваний, $B = 4$ для односторонних выпячиваний. Коэффициент 1.77 (параметр Флори) обоснован в теории несамопересекающихся случайных блужданий [34]. Случаи $l_i \le 2$ рассматриваются нами отдельно в соответствии с таблицами, приведенными в работах [31, 33]. А именно, принимаются следующие значения энергий петель: для двустороннего выпячивания – 0.8 (при $l = 2$), для одностороннего выпячивания – 6.2 (при $l = 1$) и 4.5 (при $l = 2$). Концевые петли с такими длинами в нашей модели исключаются.

Формула (3) является частью ряда типа разложения Эджуорта, но, по-видимому, имеющиеся экспериментальные данные еще не позволяют оценить значения его старших коэффициентов.

Переходы между микросостояниями делятся на "быстрые" и "медленные". *Быстрый* переход – это, по определению, переход без изменения соответствующего макросостояния. *Медленный переход* – это, по определению, переход, при котором макросостояние меняется ровно на одну пару скобок, т.е. меняется на ±1 хорду. При любом переходе может произойти изменение, вообще говоря, любого числа гипоспиралей.

Для нашей модели не существенны точные значения *вероятностей* (далее везде вместо этого будем говорить – *скоростей*) быстрого перехода из микросостояния ω в микросостояние ω' в *том же макросостоянии*. Используется лишь предположение, что в результате быстрых переходов на множестве всех микросостояний ω из данного макросостояния Ω устанавливается ста-

Численные характеристики потенциально возможных спиралей и состояний

Ширина окна (нуклеотидов)	Среднее число спиралей с длиной плеча:					Среднее число макросостояний	Среднее число микросостояний
	3	4	5	6	>6		
10	0.09	–	–	–	–	0.09	0.09
20	1.93	0.55	0.17	0.05	–	2.91	2.92
30	5.48	1.73	0.60	0.38	0.27	17.35	18.31
40	11.02	3.73	1.54	0.68	0.69	105.2	116.4
50	17.89	6.16	2.92	1.00	1.27	501	578
60	24.91	8.42	4.35	1.27	1.80	1981	2325
70	32.15	10.82	5.69	1.59	2.16	8265	9887
80	39.56	13.00	7.04	1.92	2.44	33713	40801
90	53.55	17.55	9.18	3.09	3.73	219097	284627

ционарное распределение вероятностей Больцмана–Гиббса:

$$p(\omega) = \frac{\exp(-(G_{loop}(\omega) + G_{hel}(\omega)))}{z(\Omega)}, \quad (4)$$

где $z(\Omega) = \sum_{w \in \Omega} \exp(-G_{loop}(\omega) - G_{hel}(\omega))$.

Для медленных переходов между микросостояниями, когда *макросостояние обязательно меняется ровно на одну хорду*, в модели приняты указанные ниже формулы (5, 6) для вероятностей переходов [28, 29], которые также нам представляются физически обоснованными в некотором приближении: скорость распада гипоспирали определяется ее энергией связи, а скорость присоединения гипоспирали – трудностью сближения ее будущих плеч.

Как другой вариант, для этих медленных переходов между микросостояниями принималась указанная ниже формула (7). Нельзя исключить, что эти формулы вероятностей медленных переходов должны выбираться с учетом филогенетической группы (см. раздел "Результаты модельного счета и обсуждение").

Все формулы и табличные значения вызываются в нашей программе по их именам, поэтому они могут быть легко изменены в программе, в том числе, и формулы (5–7). Авторы в значительной мере видели цель работы в создании достаточно *универсальной компьютерной программы*, которая позволяла бы реализовать модель с излагаемой здесь логикой, но с любыми формулами для заложенных в ней зависимостей. По исходной последовательности и списку всех формул, имеющих в статье номер, и по списку значений явно указанных здесь числовых параметров в ответ на запрос на адрес lin@iitp.ru авторы готовы проводить вычисление вероятности $p(c)$ терминации при любом значении концентрации c. Формулы и параметры, которые не будут указаны, подставляются по умолчанию, как они приводятся сейчас в статье.

Итак, в случае медленного перехода – **распада гипоспирали**, т.е. когда происходит уменьшение макросостояния на одну хорду, при котором происходит переход из некоторого микросостояния $\omega = \{\bar{\gamma}_{1i}, \ldots, \bar{\gamma}_{ki}\}$ (где указаны все его гипоспирали) в некоторое микросостояние $\omega' = \{\bar{\gamma}'_{1i}, \ldots, \bar{\gamma}'_{ki}\}$ (где также указаны все его гипоспирали, и $\bar{\gamma}'_{li} = \varnothing$ для каких-то одних l, i – фактически гипоспираль $\bar{\gamma}'_{li}$ отсутствует в ω'), гипоспирали $\bar{\gamma}_{li}$, $\bar{\gamma}'_{li}$ взяты от одной и той же спирали γ_l и соответствуют одной хорде, скорость медленного перехода задается формулой

$$K(\omega \longrightarrow \omega') = \kappa \exp(G_{hel}(\omega) - G_{hel}(\omega')). \quad (5)$$

В случае медленного перехода – **присоединения гипоспирали**, т.е. когда происходит увеличение макросостояния на одну хорду, в тех же обозначениях скорость обратного перехода задается формулой

$$K(\omega' \longrightarrow \omega) = \kappa \exp(G_{loop}(\omega') - G_{loop}(\omega)). \quad (6)$$

Авторы также рассмотрели иной, чем формулы (5, 6), вариант. В нем скорость любого из этих двух медленных переходов вычисляется по формуле:

$$K(\omega \longrightarrow \omega') = \kappa \exp\left[\frac{1}{2}(G_{loop}(\omega) + G_{hel}(\omega)) - (G_{loop}(\omega') + G_{hel}(\omega'))\right]. \quad (7)$$

Эта формула лучше, чем (5, 6), "обслуживает" ряд оперонов грамотрицательных организмов, включая триптофановый оперон у *E. coli*, и также, например, у некоторых стрептомицетов (см.

"Результаты модельного счета и обсуждение"). Авторы проводили счет каждой лидерной области в двух вариантах: на основе формул (5, 6) и формулы (7). Вопрос о выборе одной или нескольких формул для скоростей медленных переходов требует дальнейшего систематического изучения на основе моделирования и сравнения с экспериментальными данными.

Следуя рекомендациям авторов статей [22–24], в приведенных ниже расчетах принимали $\kappa = 10^6$ с$^{-1}$. Однако при значениях $\kappa = 10^4$ или 10^5 с$^{-1}$ можно также получить интересные результаты. По-видимому, правильное значение κ можно и было бы важно определить из эксперимента, значение κ связано с представлением о "вязкости" цитоплазмы.

Заметим, что соотношения (5–7) выбирали так, чтобы выполнялся принцип детального равновесия:

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

где $E(\omega)$ – энергия, приписываемая состоянию ω в том или ином варианте. Отсюда, в частности, возникает коэффициент $\frac{1}{2}$ в формуле (7).

Вычисление для *E. coli* по формулам (5–6) привело к заведомо неправильной зависимости частоты терминации от концентрации аминокислоты. В то время как формула (7) позволяет получить хороший результат (рис. 1). В других случаях результаты моделирования с использованием разных вариантов дают сопоставимые результаты, хотя для *S. venezuela* формула (7) дает лучший результат, по сравнению с формулами (5–6).

Если теперь описать динамику макросостояний на основе динамики реализующих их микросостояний, то возможны только два перехода: добавление к текущему макросостоянию Ω новой гипоспирали (хорды) γ и исчезновение из состояния Ω одной из бывших в нем гипоспиралей (хорд) γ. После очевидного усреднения по всем парам микросостояний $\omega \in \Omega$, $\omega' \in \Omega'$ получим следующую формулу для *скорости перехода из одного макросостояния* Ω в *другое макросостояние* Ω', которая относится к обоим случаям увеличения и уменьшения макросостояния на одну гипоспираль:

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

Авторами предложены эффективные способы реализации различных частей предлагаемой компьютерной модели; в частности, алгоритм вычисления этих сумм, не предполагающий перебора всех пар микросостояний; эти способы будут опубликованы отдельно.

Наши методы определения "быстрого" и "медленного" переходов находят комбинаторное обоснование в утверждении, доказательство которого приведено нами отдельно [28, 29]. Конечно, было бы желательно получить различение быстрых и медленных переходов в терминах значений констант скоростей, но пока этого не сделано.

Предложение 1. Пусть даны два микросостояния, реализующих одно макросостояние (что эквивалентно *изоморфизму деревьев микросостояний*: ребрам деревьев приписаны гипоспирали, которые считаются эквивалентными, если берутся из одной спирали, а порядок непосредственных потомков каждой вершины фиксирован и сохраняется при изоморфизме). Тогда от одного микросостояния к другому можно перейти, оставаясь внутри макросостояния, цепочкой "шагов" так, что каждый шаг включает не более двух разрывов и двух рождений пар спаренных нуклеотидов. И наоборот: если два микросостояния взяты из разных макросостояний, то между ними такая цепочка невозможна.

Величина замедления движения полимеразы вторичной структурой, образующейся на участке мРНК между рибосомой и РНК-полимеразой. *Шпилькой* называется цепочка пар спаренных отрезков, которые линейно расположены в петлях друг друга (т.е. соответствующее дерево линейное) с *небольшими* выпячиваниями между соседними парами отрезков и *произвольной* петлей на конце этой цепочки пар; первая пара отрезков называется стеблем шпильки. В шпильке каждая пара спаренных отрезков, т.е. какая-то гипоспираль, имеет свою петлю, включающую все последующие пары таких отрезков, выпячивания и петли. Заметим, что шпилька может не являться микросостоянием.

Согласно экспериментальным данным [34–36], вероятность терминации в зависимости от длины шпильки терминатора имеет вид кривой, которая в физической литературе называется "резонансной". Не претендуя на обсуждение физического процесса взаимодействия шпильки с полимеразой, мы попытались использовать такую кривую для описания зависимости константы скорости "перескока" полимеразы от вторичной структуры РНК. Можно добавить следующее наивное пояснение к нашей формуле (8): на полимеразе имеется "положительно заряженная область с отрицательно заряженным окружением", которая может находиться в кулоновском взаимодействии с отрицательно заряженной шпилькой. Тогда возникает картина, описываемая формулой (8) и ее следствием (17) при фиксированном расстоянии r: с ростом длины h стебля шпильки "сила торможения" F шпилькой полимеразы сначала возрастает до некоторого максимума, а затем снова убывает.

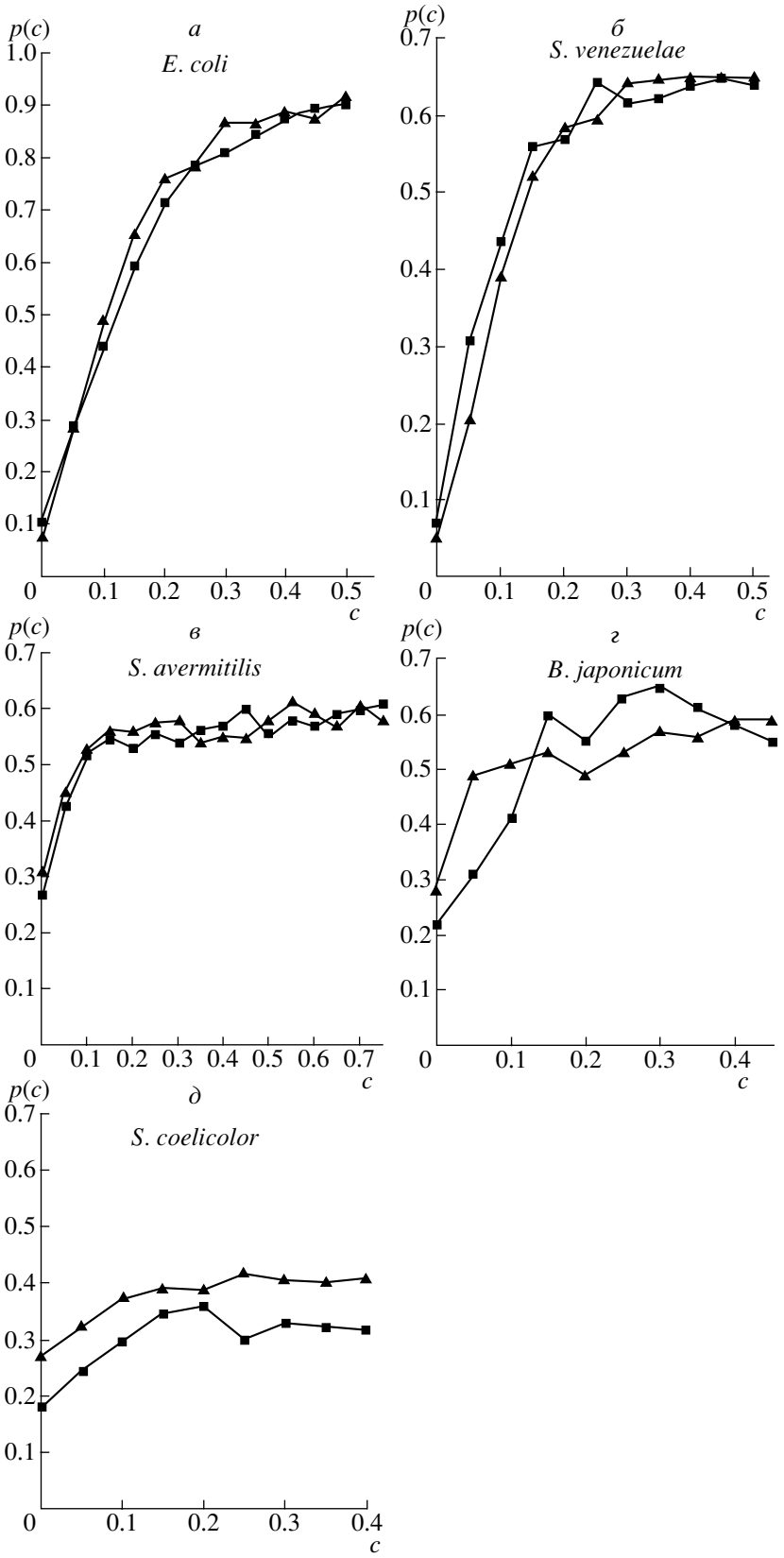

Рис. 1. Экспрессия гена *trpE* в *E. coli* в зависимости от концентрации триптофана у следующих организмов: *а* – *E. coli*; *б* – *S. venezuelae*; *в* – *S. avermitilis*; *г* – *B. japonicum*; *д* – *S. coelicolor*. Величины, обозначенные квадратиками, получены по формуле (9); треугольниками – по формуле (10). Параметр – $\alpha = 0$ в рис. 1*а–г*; $\alpha = 10$ в рис. 1*д*. По оси абсцисс – концентрация триптофана *c*; по оси ординат – вероятность терминации $p(c)$.

Итак, в модели принимается, что "сила" F замедления *шпилькой* ω полимеразы, имеющая смысл величины эффективного уменьшения константы скорости движения полимеразы по цепи ДНК и измеряемая в c^{-1}, определяется по формуле:

$$F(\omega) = \frac{\delta}{L_1^2(p - p_0)^2 + 1} \exp\left(-\frac{r}{r_0}\right), \qquad (8)$$

где r – *расстояние* от конца D шпильки ω до начала полимеразы. Биологический смысл параметров L_1, p_0, r_0, δ, зависящих только от свойств полимеразы, обсуждается ниже, волновое число p зависит только от шпильки. Если на участке РНК между рибосомой и полимеразой сформировалось несколько последовательно расположенных шпилек, составляющих набор $\{\omega_i'\}$ шпилек, то силу воздействия $F(\omega)$ этого набора на полимеразу можно вычислить как *сумму* сил от каждой из шпилек ω_i'. В свою очередь, этот набор $\{\omega_i'\}$ шпилек образуется по микросостоянию ω, которое устанавливается на участке РНК между рибосомой и полимеразой, по правилу, которое будет указано ниже. Итак,

$$F(\omega) = \sum_i F(\omega_i'), \qquad (9)$$

где $F(\omega_i')$ и соответствующее r_i вычисляют, как указано выше, по формуле (8). В формулы (8, 9) входит быстро убывающая с расстоянием экспонента, поэтому они имеют некоторое обоснование в обычном в физике приеме суммирования с экспоненциально затухающим весом.

В то же время требует рассмотрения основательный аргумент в пользу того, что с РНК-полимеразой взаимодействует только одна шпилька из набора $\{\omega_i'\}$. При этом возникает принципиальная трудность – как определить, какая именно. Чтобы рассмотреть этот аргумент, мы ввели в модель и программу вместо формулы (9) следующую формулу, в которой сумма заменена на максимум:

$$F(\omega) = \max_i\{F(\omega_i')\}. \qquad (10)$$

При этом слагаемое, соответствующее i-й шпильке, определяют по формуле (8), в которой расстояние r_i берется "по прямой", т.е. без учета всех шпилек между i-й шпилькой и полимеразой; точнее, при подсчете r_i от каждой из j-й шпильки между i-й и полимеразой сохраняются ровно два нуклеотида A_j и D_j – ее концы. Авторы проводили счет для каждой лидерной области в двух вариантах: согласно формулам (9) и (10). Сравнение этих результатов счета показывает, что итоговая кривая вероятности терминации в зависимости от концентрации аминокислоты не сильно зависит от выбора одного из вариантов – (9) или (10), но в отдельных случаях такая зависимость имеется. Совпадение результатов счета по формулам (9) и (10) объясняется тем, что для всех просчитанных нами биологических последовательностей одно из слагаемых в сумме из формулы (9) резко больше всех других слагаемых. Это слагаемое соответствует "самой массивной" из достаточно близких к полимеразе шпилек ω_i'. Однако не удается формально определить такое слагаемое, и отсюда возникает вариант с формулой (10). В разделе "Результаты модельного счета и обсуждение" приводятся кривые для обоих вариантов.

Теперь определим величину воздействия $F(\omega)$ от произвольного *микросостояния* ω на РНК-полимеразу, сведя его к определению некоторого специального набора шпилек $\{\omega_i'\}$ – *корня микросостояния* ω. Диаграмма микросостояния ω однозначно разлагается в цепочку *неразложимых диаграмм*, каждую из которых определяют по наличию стебля – самой внешней пары скобок с приписанным к ней номером спирали. В этой цепочке i-й неразложимой диаграмме ("*неразложимому микросостоянию*") соответствует гипоспираль γ_i с внешними концами A_i и D_i, приписанная самой внешней паре скобок.

Шпилька ω_i' по определению начинается с гипоспирали γ_i (т.е. с пары нуклеотидов $\langle A_i, D_i\rangle$) и продолжается по участку РНК между A_i и D_i исходной последовательности в соответствии со спариваниями в шпильке ω, оставляя без изменений мелкие выпячивания в ω вплоть до появления в ω большого выпячивания (по некоторому порогу, по умолчанию строго большего 2) или разветвления. Участки до этого момента считаются *плечами* шпильки ω_i', а участок, остающийся внутри, объявляется *петлей* шпильки ω_i'. Затем к шпильке ω_i' применяется формула (8).

Теперь определим $F(\Omega)$ как математическое ожидание по всем микросостояниям ω, реализующим данное макросостояние Ω:

$$F(\Omega) = \sum_{\omega \in \Omega} p(\omega)F(\omega). \qquad (11)$$

Константа скорости перехода полимеразы с нуклеотида на следующий нуклеотид определяют по формуле:

$$\nu(\Omega) = \bar{\lambda}_{pol} - F(\Omega). \qquad (12)$$

Предполагается $\delta < \bar{\lambda}_{pol}$. В случае (10) это означает, что $\nu(\Omega) > 0$. В случае (9) это также выполняется при всех расчетах.

Вывод принципиальной формулы для *вычисления величины p* подробно изложен ранее [28, 29].

Здесь приведем результат вывода. Для случая шпильки, состоящей только из черенка и петли, p находят из уравнения

$$\mathrm{tg}(ph) = \frac{2}{pl}, \quad 0 < ph < \frac{\pi}{2}, \tag{13}$$

где h – *длина стебля*, т.е. число спаренных нуклеотидов в нем. Для шпильки, состоящей из нескольких спаренных отрезков с небольшими выпячиваниями между ними и произвольной петлей на конце, допустимо следующее рассуждение. Пусть шпилька содержит s отрезков с длинами h_1, …, h_s и $s - 1$ выпячиваний между ними с длинами l_1, …, l_{s-1} и, наконец, петлю с длиной l. Тогда

$$p = \bar{p}\left(1 - \frac{1}{2h + l\sin^2(\bar{p}h(i))}\sum_{i=1}^{s-1} l_i \sin^2(\bar{p}h(i))\right), \tag{14}$$

где, *по определению*, $h(i) = h_1 + \ldots + h_i$, $h = h(n) = h_1 + \ldots + h_n$ и $\bar{p}$ можно вывести из аналогичного (13) уравнения

$$\mathrm{tg}(\bar{p}h) = \frac{2}{\bar{p}l}, \quad 0 < \bar{p}h < \frac{\pi}{2}. \tag{15}$$

Поскольку $0 < \bar{p}h < \frac{\pi}{2}$, то в формуле (14) все множители $\sin^2(\bar{p}h(i))$ монотонно возрастают по переменным $h(i)$.

Движение РНК-полимеразы по цепи ДНК

Рассмотрим ситуацию перескока полимеразы с нуклеотида, *принадлежащего Т-богатому участку*. Если 3'-конец полимеразы, обозначаемый везде далее как z, находится на n-м нуклеотиде, то возможен ее перескок на $(n + 1)$-й нуклеотид или срыв с нуклеотидной последовательности. *Т-богатый участок* определяется следующим образом. Нуклеотид z, *назовем* Т-богатым (внутри участка), если существует хотя бы одно слово, содержащее z на любом его месте, которое по длине больше порога (по умолчанию, 6) и по плотности "Т" больше порога (по умолчанию, 0.8). Это слово может содержать исключения, т.е. не букву "Т", на любом месте, включая концы; само z также может не быть буквой "Т". Во множестве всех Т-богатых нуклеотидов образуем все интервалы максимальной длины. Они называются Т-богатыми участками и не пересекаются.

Полимераза из положения $z = n$ ("основное состояние") может перейти с константой скорости $\bar{\lambda}_{pol}$ в положение $z = n + 1$ или с некоторой константой скорости в "возбужденное" состояние n^*, из которого может с константой λ_{ur} соскочить или с некоторой константой скорости вернуться назад в основное состояние $z = n$ [35]. Если переходы между n и n^* быстрые, то эту схему движения полимеразы можно заменить на ее усреднение: пусть переход из n в $n + 1$ происходит с константой $\beta\bar{\lambda}_{pol}$ и срыв из n с константой $(1 - \beta)\lambda_{ur}$, где β – вероятность найти полимеразу в основном состоянии, а $(1 - \beta)$ – вероятность найти ее в возбужденном состоянии. Приравнивая $\beta\bar{\lambda}_{pol} = \nu(\Omega) = \bar{\lambda}_{pol} - F(\Omega)$, получим $(1 - \beta)\bar{\lambda}_{pol} = F$, т.е. $(1 - \beta) = \frac{F}{\bar{\lambda}_{pol}}$, и окончательно имеем для варианта срыва полимеразы константу скорости

$$\mu = \mu_{out} = \frac{\lambda_{ur}F}{\bar{\lambda}_{pol}}. \tag{16}$$

Если упомянутые переходы не предполагать быстрыми, то можно выписать более сложную формулу, но по сути аналогичную формуле (16). Отношение $\frac{\bar{\lambda}_{pol}}{\lambda_{ur}}$ является естественным параметром модели и, по данным других авторов [35], равно 4.

Отвлечемся, чтобы применить эту схему *для численной оценки параметров* L_1, p_0, δ. Если шпилька состоит из одного стебля с пренебрежимо малой петлей, то $p = \frac{\pi}{2h}$ и

$$F(h) = \frac{\delta}{L_2^2\left(\frac{1}{h} - \frac{1}{h_0}\right)^2 + 1}\exp\left(-\frac{r}{r_0}\right), \tag{17}$$

где $L_2 = \frac{\pi}{2}L_1$, $h_0 = \frac{\pi}{2p_0}$. Вероятность того, что полимераза перейдет с n-го нуклеотида на $(n + 1)$-й и не сорвется, очевидно, равна $\frac{\nu}{\nu + \mu}$; таким образом, она отлична от 1 только на Т-богатом участке и только при наличии шпилек (т.е. когда $\mu > 0$ или, что то же самое, $F > 0$). Вероятность совершить N переходов по Т-богатому участку и не сорваться равна, очевидно,

$$\left(\frac{\nu}{\nu + \mu}\right)^N. \tag{18}$$

Эти формулы можно применить к следующим известным из эксперимента данным, полученным на *E. coli*, о зависимости частоты терминации от длины стебля (т.е. от числа пар h нуклеотидов в нем при участке урацилов длиной 8, $N = 7$) [35–37]: ⟨3; 0.2⟩, ⟨7; 0.8⟩, ⟨14; 0.2⟩. В этих парах первое число указывает на длину стебля, а второе – на частоту терминации. А именно, функция F в приближении, когда она записывается как функция от h, и при $r = 0$, зависит от трех параметров – h_0, L_1, δ;

получается система трех нелинейных уравнений с тремя неизвестными. Решая ее, получим:

$$h_0 = 7, \quad L_1 = 14.5, \quad \delta = 25,$$

откуда $p_0 = \frac{\pi}{14}$.

Эти численные значения являются ориентировочными и требуют уточнения, в том числе, с учетом филогенетической группы организма.

Значение r_0 выбирается близким к размеру РНК-полимеразы от точки выхода цепи РНК до точки транскрипции, так как для статистической оценки r_0 требуются дополнительные данные.

Движение рибосомы по цепи мРНК

На *нерегуляторных* кодонах константа скорости λ_{rib} сдвига рибосомы на 1 кодон принимается равной $\lambda_{rib} = \bar{\lambda}_{rib} = 15\ с^{-1}$. На *регуляторных* кодонах она зависит от концентрации c соответствующей аминокислоты по формуле Михаэлиса–Ментен:

$$\lambda_{rib}(c) = \frac{\bar{\lambda}_{rib} c}{c_0 + c}, \qquad (19)$$

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

Поскольку неясно, как экспериментально измерять такое c_0 (а модель должна ориентироваться на сравнение с результатами экспериментов), был использован следующий подход. Зависимость концентрации заряженных тРНК от концентрации аминокислоты определяется также по формуле Михаэлиса–Ментен. Результат ее подстановки в формулу (19) снова приводит к формуле такого же вида, в которой теперь c – концентрация аминокислоты в клетке. Но в опытах концентрация фиксируется не в клетке, а вне нее, в культуре. Поэтому, подставляя еще раз, получим ту же формулу (19), где *c – концентрация аминокислоты в культуре*. Соответствующее c_0 – параметр Михаэлиса–Ментен, отражающий эти два процесса: влияние концентрации аминокислоты в культуре на ее же концентрацию в клетке, влияние концентрации в клетке на концентрацию заряженных тРНК, а эта концентрация – на вероятность движения рибосомы на регуляторных кодонах. Таким образом, эта константа c_0 не имеет прямой биологической интерпретации. Здесь предполагается, что аминоацил-тРНК синтетаз и самих тРНК имеется в достаточном количестве. Значение $\bar{\lambda}_{rib}$ в этих трех случаях не меняется и совпадает со скоростью трансляции нерегуляторных кодонов.

Посадка рибосомы на область Шайна–Дальгарно

Как только область Шайна–Дальгарно (ШД) и старт-кодон (atg или gtg) лидерного пептида и еще $s_0 + s_1$ нуклеотидов (расстояние между Р-участком рибосомы и точкой транскрипции) транскрибированы, появляется возможность рибосоме связаться с мРНК. Представим себе (речь идет о рассуждении, которое лишь поможет нам вывести формулу), что комплекс "рибосома и нагруженная тРНК" являются двумя "плечами", которые должны связаться соответственно с областью ШД и старт-кодоном, при этом на участке мРНК, включающем ШД и старт-кодон, имеется некоторое макросостояние Ω. Возникает возможность переходов между состояниями $\langle \Omega, freerib \rangle$ и $\langle \Omega, boundrib \rangle$. Константы скоростей переходов между ними обозначим соответственно слева направо K_{in} и в обратную сторону K_{out}. По общему правилу

$$K_{in} = \sum_{\omega \in \Omega} \sum_{\omega' \in \langle \Omega, boundrib \rangle} p(\omega) K(\omega \longrightarrow \omega'), \qquad (20)$$

где $K(\omega \longrightarrow \omega') = \kappa_{SD} \exp(G_{loop}(\omega) - G_{loop}(\omega'))$ и $\kappa_{SD} = 10\ с^{-1}$. По-видимому, можно считать, что $G_{loop}(\omega) = G_{loop}(\omega')$, и получим $K(\omega \longrightarrow \omega') = \kappa_{SD}$, если в микросостоянии ω имеется не менее трех подряд открытых нуклеотидов в области ШД и старт-кодон открыт полностью. Иначе $K(\omega \longrightarrow \omega') = 0$.

Обратный переход имеет константу скорости $K_{out} = \sum_{\omega' \in \langle \Omega, boundrib \rangle} \sum_{\omega \in \Omega} p(\omega') K(\omega' \longrightarrow \omega)$, где $K(\omega' \longrightarrow \omega) = \kappa \exp(G_{hel}(\omega') - G_{hel}(\omega))$ и $\kappa = 10^6\ с^{-1}$ – стандартная константа скорости замыкания. Из состояния $\langle \Omega, boundrib \rangle$ возможны переход в состояние $\langle \Omega, freerib \rangle$ или сдвиг рибосомы на первый (после старта) кодон, в результате чего рибосома уже не может стать свободной (до достижения стоп-кодона лидерного пептида, когда рибосома мгновенно "срывается"), инициация заканчивается и рассматриваются переходы, характерные для установившегося движения, когда рибосома и полимераза, уже обе, находятся на цепи мРНК.

Если мы не принимаем во внимание процесс многократного присоединения и отсоединения рибосомы от цепи мРНК, то достаточно вычислить математическое ожидание времени между транскрипцией старт-кодона и переходом рибосомы на первый после старта кодон (связанное состояние). Это время равно

$$T = \frac{K_{in} + K_{out} + \bar{\lambda}_{rib}}{K_{in} \bar{\lambda}_{rib}}.$$

Поэтому можно считать, что после транскрипции старт-кодона (и еще $3 + s_0 + s_1$ нуклеотидов), становится возможным этап "связывание Р-участка рибосомы с первым кодоном после старт-кодона", причем константа скорости этого процесса равна

$$K_{in}^{eff} = \frac{K_{in}\bar{\lambda}_{rib}}{K_{in} + K_{out} + \bar{\lambda}_{rib}}.$$

После этого движение рибосомы по цепи мРНК продолжается до стоп-кодона. Моделирование посадки рибосомы возможно в случаях, когда указан старт транскрипции, поскольку только в этом случае определяется вторичная структура на участке, включающем область ШД.

ОПИСАНИЕ СХЕМЫ МОДЕЛИРОВАНИЯ

В случае классической аттенюаторной регуляции *цель моделирования* состояла в численном определении зависимости $p = p(c)$ вероятности терминации от концентрации c заряженных тРНК (или от концентрации c аминокислоты в клетке) для различных биологических регуляторных областей, а также ряда связанных с $p(c)$ зависимостей. Для построения зависимости $p = p(c)$ при каждом значении c из сетки с некоторым шагом узлов указанный в нашей модели процесс проигрывали определенное число раз (например, 10^3–10^4 раз, что дает примерно одинаковый результат) и вычисляли $p = p(c)$ как долю случаев, в которых происходит терминация. Параметр c_0 в формуле (19) полагаем равным 1, т.е. единицей измерения по оси c является c_0, а r_0 подбирался в интервале 2–8.

Из опытов известны немногочисленные характеристики, с которыми можно сравнивать результаты моделирования: например, отношение вероятностей $p = p(c)$ при достаточно большой и достаточно малой концентрациях. Известны графики зависимости активности фермента (например, антранилатсинтазы) от концентрации аминокислоты (например, триптофана в культуре [37]). Переход от условных единиц, которые сейчас приняты на осях графика функции $p(c)$, к физическим единицам измерения активности и концентрации требует отдельного изучения. Пока эти графики сравнивали на качественном уровне.

Для исходной фиксированной последовательности РНК *текущее состояние* модели характеризуется следующими моментами.

1) Имеется *окно* между положениями 3'-края x рибосомы и начала y полимеразы. "Размер" рибосомы от ее Р-участка до ее 3'-края обозначим s_0 (порядка 10–12 нуклеотидов, по умолчанию 12), а "размер" полимеразы от y – места выхода цепи РНК до точки транскрипции через s_1 (порядка 2–7 нуклеотидов, по умолчанию 5). Точку транскрипции обозначим z, всегда выполняется $z = y + s_1$. В *окне* происходит перестройка вторичной структуры от одного макросостояния Ω к другому Ω', при этом макросостояния могут включать только спирали, пересекающиеся с окном обоими плечами хотя бы по трем нуклеотидам, т.е. речь идет о *макросостояниях в окне* (для текущего окна).

2) Имеется *список T* (потенциальных) спиралей, пересекающихся с окном обоими плечами (*хотя бы* по минимальной длине гипоспирали, т.е. по 3 нуклеотидам); это тривиальная компонента состояния в том смысле, что каждый раз ее можно вычислять заново по исходному списку спиралей.

3) Имеется *макросостояние* Ω, оно же – непустая диаграмма, вторичная структура в окне.

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

Отслеживается один из *двух возможных исходов* моделирования: 1) срыв полимеразы на одном из нуклеотидов участка остатков урацила исходной последовательности или 2) прохождение полимеразой всего участка остатков урацила.

Инициация процесса РНК-регуляции (от посадки полимеразы до посадки рибосомы).

1) Полимераза садится на промотор и через некоторое число шагов приходит в точку старта лидерного пептида по общему правилу.

2) Как только полимераза транскрибирует старт лидерного пептида и еще $s_0 + s_1$ нуклеотидов, на область ШД пытается сесть рибосома с константой скорости, которая отражает зависимость от качества этой области и от вторичной структуры, закрывающей ее. Как только это произошло, рибосома занимает положение на старте ЛП. В этот момент *фиксируются*: левый конец x окна в точке "старт лидерного пептида" + s_0, и правый конец y окна в том положении, которое на тот момент занимает начало полимеразы.

Переходы в процессе РНК-регуляции после формирования окна [x, y]

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

2) *Сдвиг* рибосомы на 1 кодон вправо; окно уменьшается на 3 нуклеотида и, вообще говоря,

список спиралей T сокращается, а макросостояние Ω меняется. Из диаграммы Ω исключается *самая левая* скобка, если приписанная ей спираль не входит в новый список T (и, конечно, соответствующая правая скобка). Так полученное макросостояние – новое Ω, может быть, пустое – оно фиксируется в текущем окне.

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

Окончание моделирования

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

Организация переходов при моделировании

Моделирование выполняется с использованием метода Монте-Карло стандартным образом. *Состояние* описывается набором $\langle x, y, z, T, \Omega\rangle$. В период инициации в описание ситуации еще входит признак ς, определяющий, села рибосома или нет; затем его можно опускать. *Окрестностью* данного состояния Ω (*с центром* в Ω) называется набор всех состояний, в которые можно (с ненулевой вероятностью) перейти из Ω. Если окрестность состоит из n состояний и соответствующие константы скоростей переходов равны соответственно $k_1, \ldots, k_n$ (пусть $k = \sum k_i$), то состояние, в которое переходим (которое считается *следующим* на данной траектории), определяется как реализация случайной величины $i \longrightarrow \frac{k_i}{k}$. При этом в некоторые моменты моделирования дополнительно определяют время до наступления перехода как времени реализации случайной величины $t \longrightarrow ke^{-kt}$. Заметим, что порядки величин λ_{sd}, λ_{rib}, λ_{pol} и $K(\Omega \longrightarrow \Omega')$ часто значительно отличаются.

Представление данных в программе, реализующей эту модель, было нетривиальной задачей, поскольку для эффективной реализации вышеописанной схемы переходов, помимо упомянутого набора $\langle x, y, z, T, \Omega\rangle$, необходимо хранить все множество возможных макросостояний для текущего окна (или хотя бы для окрестности макросостояния Ω) вместе с множествами их всевозможных микросостояний. Вычисления показали, что 80% и более от всех спиралей любой последовательности имеют длину плеча 3–4, и поэтому подавляющее большинство потенциальных макросостояний содержат по одному микросостоянию, случаи нескольких микросостояний в одном макросостоянии редки. Представление о количественных характеристиках типичных спиралей и потенциально возможных вторичных структур дает таблица, содержащая данные обработки лидерной области перед триптофановым опероном у *S. avermitilis* MA-4680 скользящим окном фиксированной ширины с последующим усреднением указанных в таблице значений по всем различным позициям окна.

Видно, что мощность множеств макро- и микросостояний быстро растет с увеличением ширины окна и уже при ширине 80–100 приближается к миллиону, при том что 70% и более макросостояний имеют лишь по одному микросостоянию. Поэтому в реализованной структуре данных основой описания является набор всевозможных микросостояний текущего окна, а они группируются затем в макросостояния по признаку совпадения диаграммы. В итоге мы остановились на четырехуровневой смешанной линейно-списковой структуре данных, изображенной на рис. 2.

В первоначальных вариантах программы эту структуру строили заново при каждом изменении окна, что оказалось вполне допустимым для окон шириной до 40–50 нуклеотидов, но при дальнейшем увеличении окна времени были слишком велики. Поэтому в окончательной программе, начиная с некоторой пороговой ширины окна, после сдвига его правой или левой границ текущие наборы макро- и микросостояний не строили заново, а наоборот, перестраивали из уже имеющихся, что, хотя и сложнее алгоритмически, но дает заметное увеличение производительности программы. Счет показал, что в методе Монте-Карло 1000 повторений при каждом значении c концентрации приводит к кривой, которая не изменяется при дальнейшем увеличении этого числа.

РЕЗУЛЬТАТЫ МОДЕЛЬНОГО СЧЕТА И ОБСУЖДЕНИЕ

Исходную последовательность брали, начиная от области ШД лидерного пептида до конца участка урацилов, т.е. до полиурацилового тракта терминатора транскрипции.

В качестве примера приведем результаты счета для генов антранилатсинтазы трех стрептомицетов (*S. venezuelae* ISP5230, *S. avermitilis* MA-4680 и *S. coelicolor* A3(2)), выравнивание 5'-нетранслируемых областей которых, приведенное в одной из наших работ [16], дано на рис. 3. Приведем также результаты счета для гена *trpE* антранилатсинтазы альфа-протеобактерии *Bradyrhizobium japonicum* и гамма-протеобактерии *E. coli*. Массо-

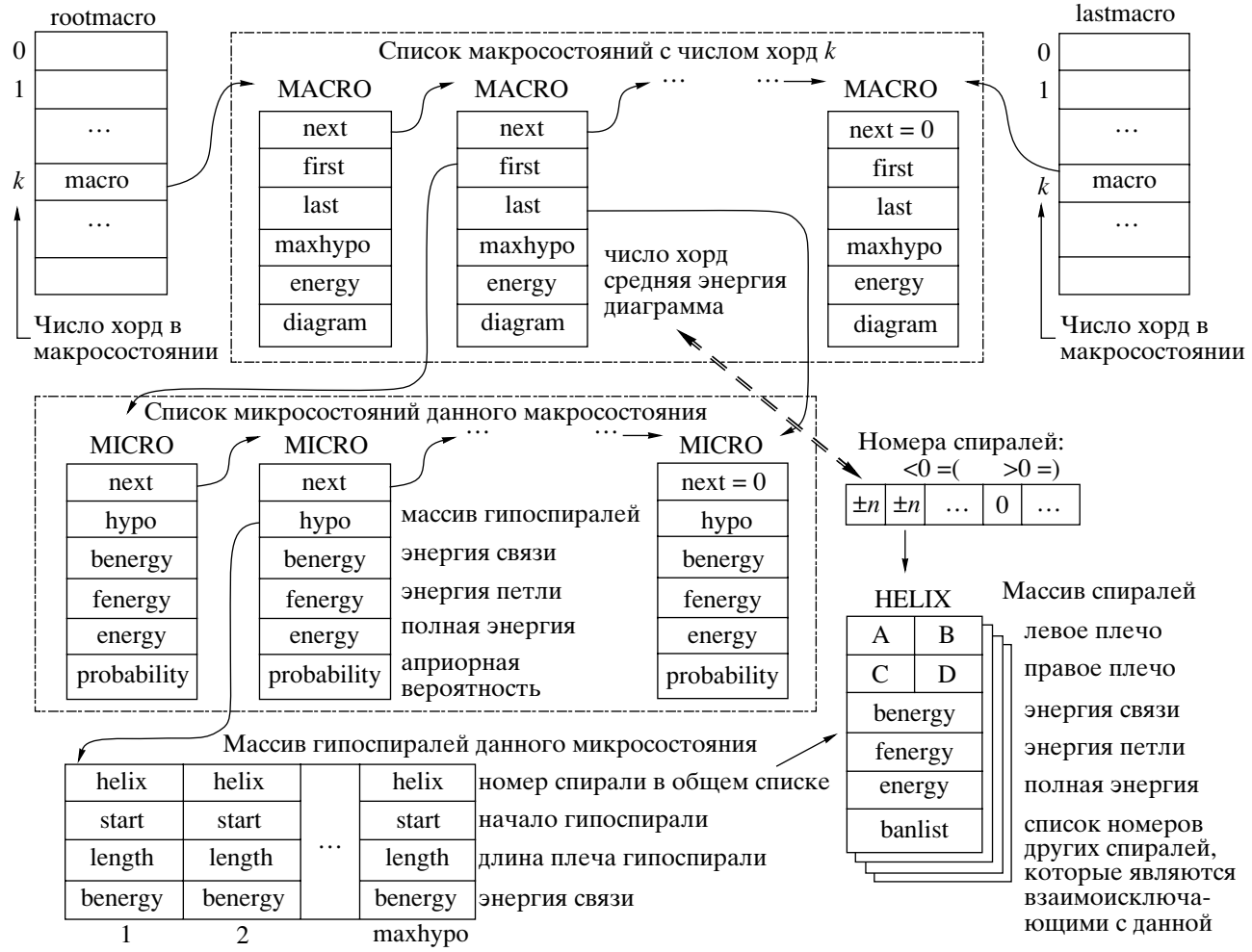

Рис. 2. Структура данных, представляющих одно состояние в модели.

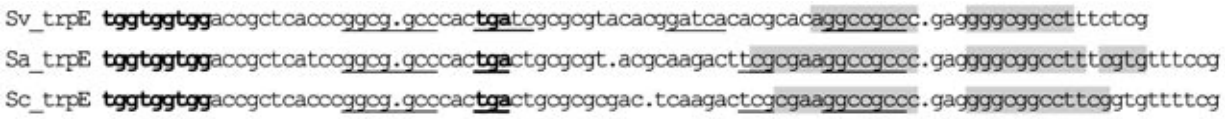

Рис. 3. Выравнивание 5'-нетранслируемых регионов генов *trpE* из трех стрептомицетов *S. venezuelae* ISP5230, *S. avermitilis* MA-4680 и *S. coelicolor* A3(2). Полужирным шрифтом выделены регуляторные и стоп-кодоны, подчеркиванием – антитерминатор, серым фоном – терминатор.

вый счет будет представлен в следующей публикации.

Анализ модели и численный счет показали, что *наиболее критическими параметрами* модели являются L_1, r_0, α. Счет проводили, варьируя эти параметры в указанных выше пределах. Для параметров F принимали значения $\delta = 25$. Для случаев, указанных на рис. 1*а* и 1*б*, принимали $L_1 = 14.5$, $p_0 = 0.167$, $r_0 = 2$ и вариант формулы (7), значение $\alpha = 0$. Для случаев, указанных на рис. 1*в–д*, принимали $L_1 = 10$, $p_0 = 0.12$, $r_0 = 5$ и вариант формул (5–6), значения α указаны в подписях к рисункам. На рисунках указаны оба варианта вычисления силы F по формуле (9) и по формуле (10).

Влияние “размеров” рибосомы и полимеразы имеет место, но заметно слабее и одинаково для всех рассмотренных организмов и генов. Это позволило выбрать для них общие значения $s_0 = 12$, $s_1 = 5$. Что касается параметра p_0, то, в принципе, его значение должно быть существенным, но, как показал счет, оно не слишком сильно влияет на характер зависимости, а приводит, в основном, лишь к сдвигу области значений вероятности терминации.

Результаты счета по формулам (9) и (10) для каждого из генов при вышеуказанных значениях параметров приведены в форме графиков на рис. 1, где по оси ординат откладывается величина $p(c)$. В случаях *S. venezuelae*, *S. avermitilis*, *B. japonicum* и *E. coli* устанавливается параметр $\alpha = 0$. Случай *S. coelicolor* является примером, для которого счет при $\alpha = 0$ дает строго убывающий график $p(c)$ вероятности терминации, что привело нас к необходимости использовать поправку в формуле (2) со значением параметра $\alpha = 10$. Трудно предположить, что в этом случае модель не работает, а в двух других работает, так как множественное выравнивание, приведенное на рис. 3, указывает, что в равной мере во всех трех случаях имеет место аттенюаторная регуляция, а в случае *S. venezuelae* такая регуляция даже подтверждена экспериментально.

Во всех пяти случаях при малой концентрации триптофана наблюдается рост частоты терминации с увеличением концентрации триптофана, а при больших концентрациях – насыщение и выход на плато. При этом в случаях *E. coli*, *S. venezuelae* и *S. avermitilis* (рис. 1, 4 и 5) различия между результатами, полученными при использовании формул (9) и (10), незначительны. В случае *B. japonicum* формула (9) кажется предпочтительнее, чем (10). В случае *S. coelicolor* формула (9) предпочтительнее при малых концентрациях, но при больших частота терминации транскрипции начинает уменьшаться, в то время как при использовании формулы (10) наблюдается почти монотонный рост $p(c)$.

Итак, предложена модель и ее компьютерная реализация для процесса аттенюаторной регуляции. Модель основана на явных, поддающихся анализу и строго сформулированных положениях (в части описания зависимостей величин и выбора значений параметров). По этой модели счет на биологических примерах приводит к результатам, качественно не расходящимся с экспериментальными данными. Предложена методика определения параметров по исходным данным. В компьютерной программе предусмотрена возможность широкого варьирования использованных в модели зависимостей и параметров. На основе счета получены заключения о чувствительности модели к одним параметрам и ее относительной устойчивости к другим; получены численные оценки параметров как биологически содержательных, так и артефактных и относящихся собственно к методу Монте-Карло. Получены хотя и внутренние, но существенные для любых моделей в этой области численные характеристики: типичные длины плеч, соотношения числа микро- и макросостояний, длины циклов между двумя соседними переходами рибосомы и полимеразы и т.п.

Секвестирование, т.е. перекрытие шпилькой РНК, области посадки рибосомы находит отражение в предлагаемой модели, но подробное исследование этого вопроса, как и учет белок-ДНКовой регуляции транскрипции, т.е. связывание с ДНК вблизи промотора белка-репрессора или активатора, будут представлены в другой статье. В дальнейшем на основе нашей модели предполагается получить предсказания о влиянии точечных мутаций в регуляторных областях на результат аттенюаторной регуляции, включая предсказания об эволюционной устойчивости организмов. А затем – включить модель в более широкую модель регуляции экспрессии генов и метаболизма у бактерий. Другое возможное использование модели таково: сейчас аттенюаторная регуляция предсказывается обычно на основе множественного выравнивания, и для этого требуется несколько последовательностей; получение с помощью модели по индивидуальной последовательности характерной для аттенюации (или ее отсутствия) кривой зависимости активности фермента от концентрации (при подходящих параметрах) могло бы рассматриваться как аргумент в пользу наличия или отсутствия аттенюации.

Авторы глубоко благодарны А.А. Миронову за многочисленные разъяснения по теме работы и за постановку этой задачи, а также М.С. Гельфанду, на семинаре которого доложена эта работа, за проявленный интерес и ценные критические замечания. Авторы благодарят К.Ю. Горбунова за советы и большую помощь в проведении расчетов. Авторы благодарны рецензенту, тщательная и продуманная критика которого позволила значительно улучшить текст статьи.

Работа получила финансовую поддержку фонда ISTC (2766).

СПИСОК ЛИТЕРАТУРЫ

1. Henkin T.M., Yanofsky C. 2002. Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions. *Bioessays*. **24**, 700–707.
2. Grundy F.J., Henkin T.M. 2003. The T box and S box transcription termination control systems. *Front. Biosci*. **8**, d20–31.
3. Grundy F.J., Henkin T.M. 2004. Regulation of gene expression by effectors that bind to RNA. *Curr. Opin. Microbiol*. **7**, (2), 126–131.
4. Mandal M., Breaker R.R. 2004. Gene regulation by riboswitches. *Nature Rev. Mol. Cell. Biol*. **5**, 451–463.
5. Vitreschak A.G., Rodionov D.A., Mironov A.A., Gelfand M.S. 2004. Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet*. **20**, 44–50.
6. Yanofsky C. 2004. The different roles of tryptophan transfer RNA in regulating trp operon expression in *E. coli* versus *B. subtilis*. *Trends Genetics*. **20**, (8) 367–374.

7. Panina E.M., Vitreschak A.G., Mironov A.A., Gelfand M.S. 2001. Regulation of aromatic amino acid biosynthesis in gamma-proteobacteria. *J. Mol. Microbiol. Biotechnology*. **3**, 529–543.
8. Vitreschak A.G., Lyubetskaya E.V., Shirshin M.A., Gelfand M.S., Lyubetsky V.A. 2004. Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis. *FEMS Microbiol. Letters*. **234**, 357–370.
9. Grundy F.J., Henkin T.M. 1994. Conservation of a transcription antitermination mechanism in aminoacyl-tRNA synthetase and amino acid biosynthesis genes in gram-positive bacteria. *J. Mol. Biol*. **235**, 798–804.
10. Grundy F.J., Henkin T.M. 1998. The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria. *Mol. Microbiol*. **30**, 737–749.
11. Murphy B.A., Grundy F.J., Henkin T.M. 2002. Prediction of gene function in methylthioadenosine recycling from regulatory signals. *J. Bacteriol*. **184**, 2314–2318.
12. Panina E.M., Vitreschak A.G., Mironov A.A., Gelfand M.S. 2003. Regulation of biosynthesis and transport of aromatic amino acid in low-GC Gram-positive bacteria. *FEMS Microbiol. Letts*. **222**, 211–220.
13. Sudarsan N., Barrick J.E., Breaker R.R. 2003. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA*. **9**, 644–647.
14. Rodionov D.A., Vitreschak A.A., Mironov A.A., Gelfand M.S. 2003. Computational analysis of thiamin regulation in bacteria: Possible mechanisms and new THI-element-regulated genes. *J. Biol. Chem*. **277**, 48949–48959.
15. Henkin T.M., Glass B.L., Grundy F.J. 1992. Analysis of the *Bacillus subtilis tyrS* gene: conservation of a regulatory sequence in multiple tRNA synthetase genes. *J. Bacteriol*. **174**, 1299–1306.
16. Seliverstov A.V., Putzer H., Gelfand M.S., Lyubetsky V.A. 2005. Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. *BMC Microbiology*. **5**, 54.
17. Barrick J.E., Corbino K.A., Winkler W.C., Nahvi A., Mandal M., Collins J., Lee M., Roth A., Sudarsan N., Jona I., Wickiser J.K., Breaker R.R. 2004. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl. Acad. Sci. USA*. **101**, 6421–6426.
18. Abreu-Goodger C., Ontiveros-Palacios N., Ciria R., Merino E. 2004. Conserved regulatory motifs in bacteria: riboswitches and beyond. *Trends Genet*. **20**, (10) 475–479.
19. Vitreschak A.A., Rodionov D.A., Mironov A.A., Gelfand M.S. 2002. Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. *Nucleic Acids Res*. **30**, 3141–3151.
20. Vitreschak A.G., Rodionov D.A., Mironov A.A., Gelfand M.S. 2003. Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. *RNA*. **9**, 1084–1097.
21. Сингер М., Берг П. 1998. *Гены и геномы*. М.: Мир.
22. Миронов А.А., Кистер А.Э. 1985. Теоретический анализ кинетики образования вторичной структуры РНК в процессе транскрипции и трансляции. Учет дефектных спралей. *Молекуляр. биология*. **19**, 1350–1357.
23. Миронов А.А., Кистер А.Э. 1989. Теоретический анализ структурных перестроек в процессе образования вторичных структур РНК. *Молекуляр. биология*. **23**, 61–71.
24. Mironov A.A., Lebedev V.F. 1993. A kinetic model of RNA folding. *BioSystems*. **30**, 49–56.
25. Elf J., Ehrenberg M. 2005. What Makes Ribosome-Mediated Transcriptional Attenuation Sensitive to Amino Acid Limitation? *PLoS Comput. Biology*. **1**, 1, e2.
26. Xayaphoummine A., Bucher T., Thalmann F., Isambert H. 2003. Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proc. Natl. Acad. Sci. USA*. **100**, 15310–15315.
27. Xayaphoummine A., Bucher T., Isambert H. 2005. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res*. **33** (Web Server issue), W605–10.
28. Пирогов С.А., Горбунов К.Ю., Любецкий В.А. 2005. Макро- и микросостояния в модели аттенюаторной регуляции экспрессии генов у бактерий. *Труды 7 Межд. конф. "Проблемы управления и моделирования в сложных системах"*, 210–216. Самара. РАН.
29. Любецкий В.А., Пирогов С.А. Модель аттенюаторной регуляции у бактерий. 2005. *Труды 7 Межд. конф. "Проблемы управления и моделирования в сложных системах"*, 205–210. Самара. РАН.
30. Mathews D.H., Sabina J., Zuker M., Turner D.H. 1999. Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. *J. Mol. Biol*. **288**, 911–940.
31. Mathews D.H., Disney M.D., Childs J.L., Schroeder S.J., Zuker M., Turner D.H. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA*. **101**, 7287–7292.
32. Dima I., Hyeon C., Thirumalai D. 2005. Extracting Stacking Interaction Parameters for RNA from the Data Set of Native Structures. *J. Mol. Biol*. **347**, 53–69.
33. RNA Structure, Turner Lab, http://rna.chem.rochester.edu.
34. Lawler G.F., Coyle L.N. 1999. Lectures on Contemporary Probability, AMS.
35. Yin H., Artsimovitch I., Landick R., Gelles J. 1999. Nonequilibrium mechanism of translation termination from observations of single RNA polymerase molecules. *Proc. Natl. Acad. Sci. USA*. **96**, 13124–13129.
36. Wilson K., von Hippel P. 1995. Transcription termination at intrinsic terminators: the role of the RNA hairpin. *Proc. Natl. Acad. Sci. USA*. **92**, 8793–8797.
37. Lynn S., Kasper L., Gardner J. 1988. Contributions of RNA secondary structure and length of the thymidine tract to transcription termination at the thr operon attenuator. *J. Biol. Chem*. **263**, 472–479.
38. Lin Cong, Paradkar A.S., Vining L.C. 1998. Regulation of an anthranilate synthase gene in *Streptomyces venezuelae* by a trp attenuator. *Microbiology*. **144**, 1971–1980.