

# СРАВНЕНИЕ МОЛЕКУЛ РНК: ЭНЕРГИЯ СВЯЗЫВАНИЯ И СТАТИСТИЧЕСКИЕ СВОЙСТВА СЛУЧАЙНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

*О. В. Вальба*<sup>a\*</sup>, *С. К. Нечаев*<sup>b,c\*\*</sup>, *М. В. Тамм*<sup>d\*\*\*</sup>

<sup>a</sup> *Московский физико-технический институт  
141700, Долгопрудный, Московская обл., Россия*

<sup>b</sup> *LPTMS, Université Paris Sud  
91405, Cedex, France*

<sup>c</sup> *Физический институт им. П. Н. Лебедева Российской академии наук  
119991, Москва, Россия*

<sup>d</sup> *Московский государственный университет им. М. В. Ломоносова  
119992, Москва, Россия*

Поступила в редакцию 29 апреля 2011 г.

Предлагается новый статистический подход для решения задачи сравнения («выравнивания») двух последовательностей РНК. Данная проблема рассматривается с точки зрения связывания двух взаимодействующих полимеров, имеющих иерархическую кактусообразную структуру, характерную для молекул РНК. Выравнивание двух последовательностей характеризуется числом совпадающих и несовпадающих букв, а также числом пропусков. Для каждого выравнивания определяется весовая функция, имеющая смысл свободной энергии, которая, помимо непосредственной энергии контактов, учитывает комбинаторный вклад образования различных кактусообразных вторичных структур. Определяется свободная энергия связывания двух молекул РНК и обсуждаются такие статистические свойства, как флуктуации средней энергии связывания двух молекул РНК и распределение длин петель в образованной структуре. Анализ зависимости удельной свободной энергии комплекса двух длинных случайных молекул РНК от числа типов нуклеотидов  $c$  позволил выдвинуть гипотезу о выделенной роли используемого природой алфавита с  $c = 4$ .

## 1. ВВЕДЕНИЕ

Роль таких биополимеров, как ДНК и РНК, в механизмах клеточной регуляции общеизвестна. Их взаимодействие является одним из необходимых этапов клеточного цикла, связанного с хранением и передачей генетической информации. Помимо механизмов трансляции и транскрипции информации, в которых участвуют комплексы ДНК–РНК, исключительно важную роль играют взаимодействия РНК–РНК. Эти взаимодействия имеют ключевое значение для регуляции экспрессии генов [1–3]. Схематично образование комплексов РНК–РНК проис-

ходит следующим образом: молекулы РНК посредством образования комплементарных пар связываются с матричной РНК (мРНК) или ее участком и тем самым останавливают трансляцию генов с данной мРНК [3]. Молекулы РНК, участвующие в процессах данного типа, называются некодирующими РНК (нкРНК). Это название обусловлено тем, что они сами не транслируются в белки [2] и, следовательно, исключены из непосредственного процесса транскрипции.

Важная биологическая роль взаимодействий РНК–РНК обуславливает необходимость построения эффективного алгоритма, который бы позволил по первичным последовательностям теоретически вычислять энергию связывания двух РНК, а также предсказывать вторичную структуру молекул

\*E-mail: valbaolga@gmail.com

\*\*E-mail: sergei.nechaev@gmail.com

\*\*\*E-mail: thumm.m@gmail.com

нкРНК (т.е. оптимальную с термодинамической точки зрения архитектуру связей между звеньями биополимера). Как будет показано ниже, эта задача тесно связана с проблемой «выравнивания» (alignment) двух произвольных линейных последовательностей типа ДНК. Существенным отличием задачи о выравнивании молекул РНК от аналогичной задачи для ДНК является наличие нетривиальной вторичной структуры у молекул РНК.

Молекулы РНК относятся к классу ассоциирующих полимеров. По определению в ассоциирующих полимерах помимо сильных ковалентных связей между мономерами первичной цепи, возможно образование относительно слабых связей между удаленными по цепи мономерами. Так, в биополимерах мономеры могут связываться друг с другом короткодействующими водородными связями, расположение которых и задает вторичную структуру биополимера. Биополимеры несут на себе электрические заряды, однако в реальных условиях, соответствующих протеканию биологических процессов, система биополимер–растворитель является, как правило, электронейтральной, поскольку заряды на биополимере полностью компенсированы контрионами растворителя. Кроме того, поскольку ионная сила биологических сред весьма велика, электростатические взаимодействия между различными звеньями цепи сильно экранированы (радиус Дебая много меньше характерных размеров макромолекулы), благодаря чему электростатическое отталкивание звеньев можно рассматривать как эффективное короткодействующее взаимодействие. Таким образом, наличие электростатических взаимодействий лишь перенормирует микроскопические характеристики биополимера, например, персистентную длину (см. [4]), и влияет на вторичную структуру биополимеров только через перенормированные характеристики.

Наличие и конкретный вид вторичной структуры молекулы ассоциирующего полимера определяется его свободной энергией, вклад в которую дают несколько факторов, явно зависящих от гетерогенности (неоднородности первичной структуры): энергия водородных связей; комбинаторный вклад от выбора того, какие именно мономеры образуют связь; потери конформационной энтропии полимерной цепи из-за наличия связей. Особенностью молекул РНК в отличие от других биополимеров (белки, ДНК) [5] является то, что они образуют вторичные структуры иерархического типа, называемые в литературе «клеверными листьями» или «кактусами».

В данной работе предлагается новый статисти-

ческий подход к определению энергии связывания двух молекул РНК, учитывающий вторичную структуру каждой из молекул, и исследуются статистические свойства возникающих комплексов для случая случайной последовательности звеньев РНК. Структура работы такова. В разд. 2 мы приводим основные сведения об известных рекурсивных алгоритмах поиска наибольшей общей подпоследовательности двух произвольных линейных последовательностей. В разд. 3 показываем, что рекурсивный алгоритм выравнивания последовательностей может рассматриваться формально как предел нулевой температуры в статистической модели связывания двух полимерных цепочек с произвольной первичной структурой. Данная идея была кратко высказана в работе [6] и развернуто сформулирована для РНК-подобных молекул в разд. 4. Подробный анализ статистических свойств основного состояния кактусообразных (РНК-подобных) структур, приведенный в разд. 5.1–5.3 является оригинальным. В разд. 5.4 мы высказываем гипотезу о возможных причинах выделенности генетического алфавита из четырех типов нуклеотидов с точки зрения статистики выравнивания последовательностей РНК.

## 2. ВЫРАВНИВАНИЕ ЛИНЕЙНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Рассмотрим задачу о выравнивании двух линейных последовательностей и опишем стандартный алгоритм динамического программирования, используемый для определения оптимального выравнивания. Задача о выравнивании двух последовательностей — это задача нахождения эффективного алгоритма поиска максимальной общей подпоследовательности (МОП) двух произвольных линейных последовательностей. Данная проблема является одной из ключевых задач вычислительной эволюционной биологии. В частности, она позволяет судить о том, насколько далеко (в эволюционном смысле) разошлись друг от друга два рассматриваемых гена и какие гены могут являться их общими предками.

Задача о поиске МОП двух последовательностей формулируется следующим образом. Рассмотрим две произвольные последовательности (в качестве примера рассматриваются последовательности РНК, составленные из четырехбуквенного алфавита А, С, G, U):  $S_1 = \{A, C, G, C, U, A, C\}$  длиной  $m = 7$  и  $S_2 = \{C, U, G, A, C\}$  длиной  $n = 5$ . Общая подпоследовательность — это подпоследовательность, содержащая буквы (нуклеоти-

ды) как первой последовательности, так и второй, причем подпоследовательность необязательно содержит буквы, идущие непосредственно друг за другом. Так, для двух последовательностей  $S_1$  и  $S_2$  можно выделить несколько различных общих подпоследовательностей, например,  $\{C, U, A, C\}$  или  $\{G, A, C\}$ . Число возможных общих подпоследовательностей с ростом длин  $m$  и  $n$  полимеров растет экспоненциально. Алгоритм для определения оптимального выравнивания двух последовательностей впервые был сформулирован в работе [7]. В наиболее общем смысле каждое выравнивание двух последовательностей характеризуется числом совпадающих и несовпадающих букв, а также числом пропусков («делений») в выравненных последовательностях. Для каждого выравнивания можно ввести «весовую функцию» (cost function), имеющую значение энергии [5–7]:

$$F = N_{match} + \mu N_{mis} + \delta N_{gap}. \quad (1)$$

В формуле (1)  $N_{match}$ ,  $N_{mis}$ ,  $N_{gap}$  — соответственно число пар совпадающих букв, число пар несовпадающих букв и число делений в рассматриваемом выравнивании. Величины  $\mu$  и  $\delta$  — вклады в весовую функцию от пары несовпадающих букв и деления; энергию комплементарной связи, без потери общности, можно считать равной единице. В таком представлении функция  $F$  удовлетворяет очевидному закону сохранения:

$$n + m = 2N_{match} + 2N_{mis} + N_{gap}. \quad (2)$$

Используя (2), формулу (1) можно переписать в виде

$$\tilde{F} = N_{match} + \gamma N_{mis}, \quad (3)$$

где

$$\gamma = \frac{\mu - 2\delta}{1 - 2\delta}. \quad (4)$$

Здесь интерес представляет область  $0 \leq \gamma \leq 1$ , так как случай  $\gamma < 0$  неотличим от случая  $\gamma = 0$ , а случай  $\gamma > 1$  соответствует тому, что «несовпадения» более выгодны, чем «совпадения» и может быть учтен простым переопределением этих понятий. Заметим, что, хотя предлагаемая нами теория приложима ко всему доступному интервалу значений  $\gamma$ , все численные результаты настоящей работы получены для случая  $\gamma = 0$ , который представляется наиболее физически осмысленным. Задача поиска МОП заключается в определении выравнивания с

максимальным значением весовой функции  $F$ . Оказывается, что для нахождения весовой функции  $F$  удобнее всего использовать рекурсивный алгоритм, известный как метод динамического программирования:

$$\tilde{F}_{i,j}^{max} = \max \left[ \tilde{F}_{i-1,j}^{max}, \tilde{F}_{i,j-1}^{max}, \tilde{F}_{i-1,j-1}^{max} + \zeta_{i,j} \right], \quad (5)$$

где

$$\zeta_{i,j} = \begin{cases} 1 & \text{для } S_1(i) = S_2(j), \\ \gamma & \text{для } S_1(i) \neq S_2(j). \end{cases} \quad (6)$$

Выражения (5), (6) имеют следующий смысл. Начиная с левых концов последовательностей, на каждом шаге выбирается такое положение букв в выравнивании, которое вносит наибольший вклад в функцию  $F$ . Члены в правой части (5) соответствуют трем возможным ситуациям: пропуску буквы в первой последовательности, пропуску во второй последовательности и случаю, когда  $i$ -я буква первой последовательности выравнена с  $j$ -й буквой второй последовательности.

### 3. ВЗАИМОДЕЙСТВИЕ ДВУХ ЛИНЕЙНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Как уже отмечалось во Введении, основной целью данной работы является разработка статистического алгоритма вычисления весовой функции, которая бы характеризовала «похожесть» двух заданных последовательностей со сложной вторичной структурой, типа молекул РНК. Эта функция должна включать как энергетический вклад от непосредственного взаимодействия мономеров друг с другом, так и энтропийный вклад, обусловленный наличием ансамбля пространственных конформаций макромолекул. При этом мы постараемся по возможности остаться в рамках статистической физики и избежать неконтролируемых эвристических соображений, апеллирующих к опыту, полученному в результате анализа экспериментальных данных.

Прежде всего покажем, что рекуррентное соотношение (5) имеет прозрачный физический смысл в терминах статистической физики и формулы (5), (6) можно рассматривать как свободную энергию статистической модели, описывающей комплексообразование двух взаимодействующих линейных полимеров в пределе нулевой температуры. Затем, учитывая возможность того, что каждый из полимеров может, помимо собственно комплексообразования, образовывать сложную иерархическую структуру,

обобщим выражение для статистической суммы (соответствующей ненулевой температуре) на комплексы с внутренней иерархической структурой. Переходя в конечном выражении снова к пределу  $T \rightarrow 0$ , найдем искомую весовую функцию.

Стоит отметить, что при «физиологических» значениях температуры ( $t \approx 36.6^\circ\text{C}$ ) энергия водородных связей между комплементарными нуклеотидами РНК превышает тепловую энергию  $T$  в несколько десятков раз (здесь и далее в настоящей статье абсолютная температура  $T$  измеряется в энергетических единицах, т. е. принята система единиц, в которой константа Больцмана  $k_B = 1$ ). Предел нулевой температуры, как представляется, должен хорошо описывать комплексообразование молекул РНК в таких условиях. Более того, для выполнения своей физиологической роли всякая молекула или комплекс двух молекул РНК должен «самособирается» в строго определенную вторичную структуру (или, максимум, в одну из нескольких строго определенных структур), а это возможно только при условии того, что различные структуры отделены друг от друга энергетическими щелями, существенно превосходящими  $T$ .

Рассмотрим вспомогательную статистическую модель, описывающую взаимодействие двух линейных полимеров с произвольными первичными последовательностями. Пусть длины этих последовательностей, измеренные в единицах мономерных звеньев, равны соответственно  $m$  и  $n$ . Каждый мономер может быть выбран из  $c$  различных мономеров  $A, B, C, D, \dots$  (Для последовательностей РНК  $c = 4$ ). Мономеры первой последовательности могут образовывать связи с мономерами второй последовательности. В молекулах РНК такие связи образуются согласно комплементарности азотистых оснований. Пары нуклеотидов в молекуле РНК, аденин (A)–урацил (U) и гуанин (G)–цитозин (C), являются комплементарными. Для простоты положим энергию связи между комплементарными нуклеотидами равной  $u$ , а энергию между некомплементарными —  $v$ . Предположим также, что некоторые части полимеров могут образовывать петли. На рис. 1 схематически представлена возможная конфигурация структуры комплекса двух взаимодействующих РНК. Очевидно, что петли соответствуют делециям в задаче о выравнивании двух последовательностей.

Наша задача заключается в вычислении свободной энергии описанной модели при достаточно низких температурах, при которых энтропийным вкладом можно пренебречь по сравнению с энергетиче-

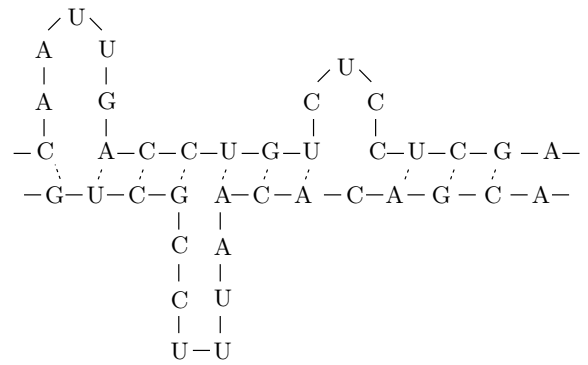


Рис. 1. Структура комплекса двух линейных последовательностей РНК

ским. Пусть  $G_{m,n}$  — статистическая сумма рассматриваемого комплекса. По смыслу  $G_{m,n}$  — это сумма по всем возможным конфигурациям связей. При низких температурах  $G_{m,n}$  можно представить как

$$G_{m,n} = 1 + \sum_{i,j=1}^{m,n} \beta_{i,j} G_{i-1,j-1}; \tag{7}$$

$$G_{m,0} = 1, \quad G_{0,n} = 1, \quad G_{0,0} = 1.$$

Смысл данной формулы очевиден: начиная с левого конца последовательностей (рис. 1), находим первый существующий контакт между  $i$ -м мономером первой цепи и  $j$ -м мономером второй, а далее суммируем по всем возможным расположениям этого контакта. Статистические веса связей  $\beta_{i,j}$  можно представить в виде матрицы контактов:

$$\beta_{i,j} = \begin{cases} \beta^+ \equiv e^{u/T}, & \text{если } S_1(i) \text{ и } S_2(j) \\ & \text{комплементарны;} \\ \beta^- \equiv e^{v/T}, & \text{если } S_1(i) \text{ и } S_2(j) \\ & \text{не комплементарны.} \end{cases} \tag{8}$$

Прямое вычисление показывает, что статистическая сумма вида (7) удовлетворяет рекуррентному соотношению:

$$G_{m,n} = G_{m-1,n} + G_{m,n-1} + (\beta_{m,n} - 1) G_{m-1,n-1}. \tag{9}$$

В свою очередь, статистическая сумма связана со свободной энергией  $F_{m,n}$  комплекса и температурой  $T$  известным соотношением  $G_{m,n} = \exp\{F_{m,n}/T\}$ . Переходя в уравнении (9) к пределу  $T \rightarrow 0$ , получим

$$F_{m,n} = \lim_{T \rightarrow 0} T \ln \left( e^{F_{m-1,n}/T} + e^{F_{m,n-1}/T} + (\beta_{m,n} - 1) e^{F_{m-1,n-1}/T} \right). \tag{10}$$

Формулу (10) можно переписать в виде

$$F_{m,n} = \max [F_{m-1,n}, F_{m,n-1}, F_{m-1,n-1} + \eta_{m,n}], \quad (11)$$

где

$$\eta_{m,n} = T \ln(\beta_{m,n} - 1) = \begin{cases} \eta^+ = T \ln(e^{u/T} - 1), & \text{если} \\ S_1(i) \text{ и } S_2(j) \text{ комплементарны;} \\ \eta^- = T \ln(e^{v/T} - 1), & \text{если} \\ S_1(i) \text{ и } S_2(j) \text{ не комплементарны.} \end{cases} \quad (12)$$

Принимая  $\eta^+$  за единицу энергии, перепишем формулу (10) в виде

$$\tilde{F}_{m,n} = \max [\tilde{F}_{m-1,n}, \tilde{F}_{m,n-1}, \tilde{F}_{m-1,n-1} + \tilde{\eta}_{m,n}], \quad (13)$$

где

$$\tilde{\eta}_{m,n} = \begin{cases} 1, & \text{если } S_1(i) \text{ и } S_2(j) \\ & \text{комплементарны;} \\ a = \frac{\eta^-}{\eta^+} = \frac{\ln(e^{v/T} - 1)}{\ln(e^{u/T} - 1)} \Big|_{T \rightarrow 0} = \frac{v}{u}, & (14) \\ & \text{если } S_1(i) \text{ и } S_2(j) \\ & \text{не комплементарны.} \end{cases}$$

Функция  $\tilde{F}_{m,n}$  удовлетворяет начальным условиям  $\tilde{F}_{0,n} = \tilde{F}_{n,0} = \tilde{F}_{0,0} = 0$ . Видно, что выражение для свободной энергии двух взаимодействующих полимеров с линейной структурой имеет вид, совпадающий с (7).

Таким образом, мы обнаружили, что рекурсия, используемая в методе динамического программирования, является ничем иным, как рекуррентным соотношением для свободной энергии комплекса, состоящего из двух взаимодействующих гетерополимеров в пределе нулевой температуры. В природе существует множество примеров образования подобных гетерополимерных комплексов, например, образование двойной спирали ДНК.

Отметим, что предложенная выше модель является лишь грубым первым приближением к описанию образования реальных линейных комплексов. Известно (см., например, [4]), что для точного количественного описания образования, например, двойной спирали ДНК необходимо учесть еще ряд факторов. Во-первых, мы не учли так называемые «петлевые факторы»: при образовании петли возможные

конформации полимера ограничены условием, что ее концы обязаны сойтись в одной точке пространства, поэтому образование каждой петли приводит к снижению энтропии комплекса. Во-вторых, в реальной ДНК имеется выраженная кооперативность образования связей: вероятность образования связи выше, если соседние мономеры также образуют связь. В-третьих, мы не учли, что гибкость полимера конечна и существует ограничение на минимальную длину петли. И наконец, мы не приняли во внимание то обстоятельство, что комплементарные пары А–U и С–G имеют различные значения энергии связи.

Обобщение выражений (7)–(11) с учетом кооперативности образования связи, минимальной длины петли и различных значений энергии комплементарных связей — задача вычислительно сложная, но не требующая качественного изменения предложенного формализма, так как эти факторы влияют только на локальные свойства полимерных цепей. С другой стороны, петлевой фактор — характеристика нелокальная, зависящая от расстояния между мономерами, образующими связи, и, учитывая его, мы не можем описать состояние комплекса уравнениями динамического программирования вида (11). Но поскольку петлевые факторы имеют энтропийную природу, в пределе низких температур ( $T \rightarrow 0$ ) их вклад в свободную энергию гетерополимерного комплекса становится пренебрежимо мал. Ситуация, однако, усложняется, если сами петли могут образовывать вторичную структуру (т. е. если внутри петли имеет место взаимодействие между мономерами), а именно такая ситуация типична для последовательностей РНК. В этом случае энергетический вклад от вторичной структуры петли сохраняется и в пределе нулевой температуры, и его учет становится необходимым.

#### 4. ВЗАИМОДЕЙСТВИЕ ДВУХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ РНК

В этом разделе мы обобщим модель комплекса двух взаимодействующих полимеров на случай, когда полимеры могут образовывать сложную кактусообразную вторичную структуру типа «клеверного листа» или «кактуса», т. е. помимо связей между двумя рассматриваемыми последовательностями возможно образование связей внутри петель каждого из полимеров. Будем рассматривать структуры петель типа «кактуса» (рис. 2а, в), структуры типа «псевдоузлов» (см. рис. 2б, г) в данной работе не рас-

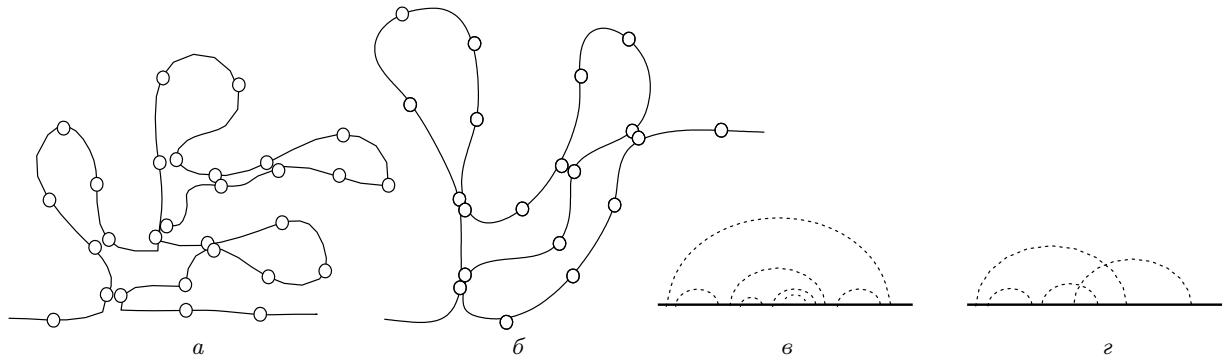


Рис. 2. Разрешенная кактусообразная структура (а) и запрещенная структура — псевдоузел (б); *v, z* — диаграммы структур соответственно *a, б*

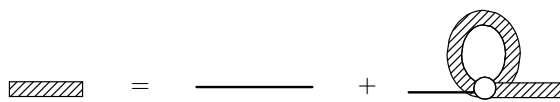


Рис. 3. Диаграмма для вычисления статистического веса *g* последовательности

считаются. Как и в предыдущем разделе, для простоты не будем учитывать кооперативность образования связей и различие в энергиях комплементарных пар. Однако, как уже указывалось, наша модель может быть обобщена с учетом этих факторов. Согласно [6] можно переписать выражение (7) для статистической суммы  $G_{m,n}$  комплекса двух взаимодействующих РНК в виде

$$G_{m,n} = g_{1,m}^{(1)} g_{1,n}^{(2)} + \sum_{i,j=1}^{m,n} \beta_{i,j} G_{i-1,j-1} g_{i+1,m}^{(1)} g_{j+1,n}^{(2)}; \quad (15)$$

$$G_{m,0} = g_{1,m}^{(1)}, \quad G_{0,n} = g_{1,n}^{(2)}, \quad G_{0,0} = 1,$$

где  $g_{i,j}^{(1)}$  и  $g_{i,j}^{(2)}$  — статистические веса участков (с *i*-го нуклеотида до *j*-го) соответственно первой и второй последовательности, удовлетворяющие уравнениям [10, 11]

$$g_{k,n}^{(a)} = 1 + \sum_{i=k}^{n-1} \sum_{j=i+1+\ell}^n \beta'_{i,j} g_{i+1,j-1}^{(a)} g_{j+1,n}^{(a)}; \quad (16)$$

$$g_0^{(a)} = 1, \quad a = 1, 2.$$

Эти уравнения отвечают за топологию кактусообразной структуры, свойственной молекулам РНК (диаграмма, описывающая такие структуры, представлена на рис. 3). Коэффициенты  $\beta'_{i,j}$  — констан-

ты, описывающие взаимодействие внутри последовательности, они аналогичны  $\beta_{m,n}$ . Суммирование по *j* ведется от  $i + 1 + \ell$  до *n*, для того чтобы исключить петли длиной меньше  $\ell$  мономеров. В последующих вычислениях, как правило, предполагается, что  $\ell=0$ , также обсуждается случай  $\ell = 3$ . Напомним еще раз, что, поскольку мы интересуемся низкими температурами, можно пренебречь вкладом, связанным с потерей энтропии при образовании петель.

Сложную систему уравнений для статистических весов петель  $g_{i,j}^{(a)}$ ,  $a = 1, 2$  (16) можно решить следующим образом. Для каждой из последовательности РНК можно построить матрицу *g*, (*i, j*)-й элемент которой определяет статистический вес петли, начинающейся с *i*-го нуклеотида и заканчивающейся *j*-м. Таким образом, статистические веса всех возможных петель описываются матрицами размера  $m \times m$  для первой последовательности и  $n \times n$  — для второй. Из граничных условий (16) мы можем однозначно определить элементы  $g_{i,i+1}^{(a)}$ . Из формулы (16) следует, что элементы последующих субдиагоналей  $g_{i,i+k}^{(a)}$  зависят только от элементов предыдущих субдиагоналей  $g_{i,i+k-\ell}^{(a)}$  матрицы:

$$g_{i,i+k}^{(a)} = g_{i+1,i+k}^{(a)} + \sum_{s=i+1}^{i+k} \beta'_{i,s} g_{i+1,s-1}^{(a)} g_{s+1,i+k}^{(a)}. \quad (17)$$

Определенные таким образом матрицы статистических весов  $g^{(a)}$  всех возможных петель позволяют вычислить статистическую сумму комплекса двух взаимодействующих РНК (15).

Как и в случае комплекса двух линейных последовательностей, мы можем выполнить переход к пределу нулевой температуры, см. выражения (7)–(13). Элементы матрицы свободной энергии при этом можно представить в виде

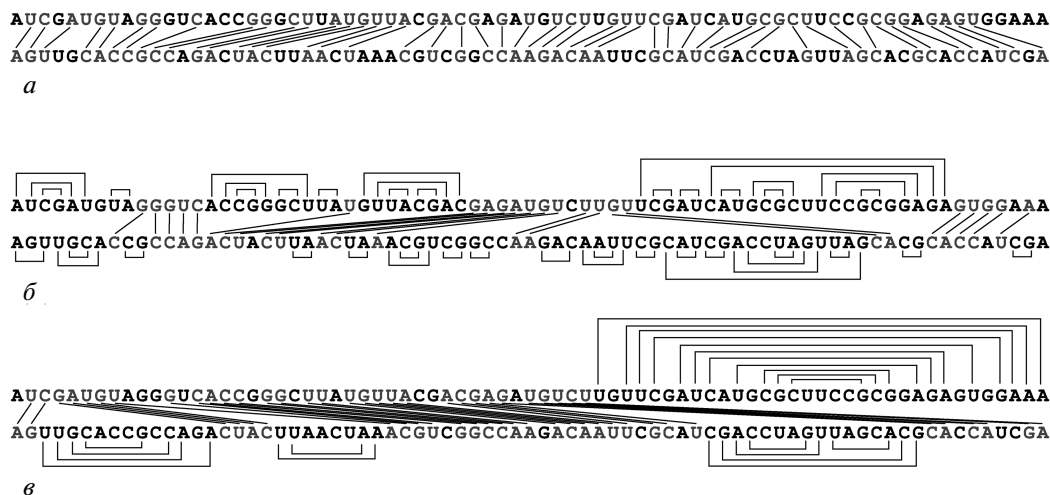


Рис. 4. Структура комплекса двух взаимодействующих биополимеров в случае линейных последовательностей (а) и последовательностей РНК с кактусообразной структурой при  $\ell = 0$  (б), 3 (в)

$$F_{m,n} = \max_{\substack{i=1,\dots,m \\ j=1,\dots,n}} [f_{1,m}^{(1)} + f_{1,n}^{(2)} Q_{i,j}^{m,n}], \quad (18)$$

где

$$f_{i,j}^{(a)} = \lim_{T \rightarrow 0} [T \ln g_{i,j}^{(a)}], \quad a = 1, 2,$$

— свободные энергии отдельных последовательностей с  $i$ -го нуклеотида по  $j$ -й,  $Q_{i,j}^{m,n}$  —  $(i, j)$ -й элемент суммы (15), который в пределе нулевой температуры равен

$$Q_{i,j}^{m,n} = F_{i-1,j-1} + f_{i+1,m}^{(1)} + f_{j+1,n}^{(2)} + \tilde{\eta}_{i,j}. \quad (19)$$

Элемент  $Q_{i,j}$  описывает энергию комплекса двух РНК, которые не взаимодействуют после контакта в позиции  $(i, j)$ . Для свободной энергии участков последовательностей из (17) можно записать:

$$f_{i,i+k}^{(a)} = \max \left[ f_{i+1,i+k}^{(a)}, \max_s \left( f_{i+1,s-1}^{(a)} + f_{s+1,i+k}^{(a)} + \tilde{\eta}'_{i,s} \right) \right]. \quad (20)$$

Здесь  $\tilde{\eta}_{i,j}$  — величина, определяемая формулой (14),  $\tilde{\eta}'_{i,s}$  — аналогичная величина, описывающая взаимодействие внутри петель. Из формулы (15) следуют граничные условия для свободной энергии:

$$\begin{aligned} F_{0,0} &= 0; \\ F_{i,0} &= f_{1,i}^{(1)}, \quad 1 \leq i \leq m; \\ F_{0,j} &= f_{1,j}^{(2)}, \quad 1 \leq j \leq n. \end{aligned} \quad (21)$$

Таким образом, для того чтобы вычислить энергию основного состояния комплекса двух взаимодействующих РНК, необходимо построить матрицы

$f^{(1)}$  и  $f^{(2)}$  для индивидуальных последовательностей, далее, применяя (18), следует определить элементы матрицы свободной энергии  $F$ .

Стоит отметить, что алгоритмы для определения свободной энергии комплекса, описанные выше, можно использовать для решения обратной задачи — восстановления вторичной структуры основного состояния гетерополимерного комплекса. Подробно этот вопрос изложен в работе [12]. На рис. 4 приведены восстановленные в соответствии с этим алгоритмом структуры комплекса двух взаимодействующих РНК [13] длиной  $m = n = 75$ . Вычисления были проведены для трех различных случаев: комплекс двух линейных РНК (а) и комплекс двух кактусообразных РНК с  $\ell = 0$  и  $\ell = 3$ .

### 5. СТАТИСТИЧЕСКИЕ СВОЙСТВА ДВУХ ВЗАИМОДЕЙСТВУЮЩИХ РНК

Данный раздел посвящен обсуждению ряда статистических свойств линейных и РНК-подобных комплексов полимеров со случайной последовательностью мономеров. Вначале обсуждается средняя энергия и флуктуации энергии комплексов случайных РНК в зависимости от длин рассматриваемых последовательностей. Далее будет описана модель, которая позволила оценить вероятность связывания случайных полимеров в РНК-подобных комплексах. И наконец, будут приведены результаты для распределения длин петель в структурах комплексов и предложены модели для их описания.

Величины свободной энергии приведены в единицах энергии комплементарной связи  $u$ . Для простоты разница в энергиях пар А–U и С–G не учитывалась, энергия некомплементарной связи была принята равной нулю.

### 5.1. Энергия и флуктуация энергии комплекса двух взаимодействующих РНК

Задача поиска основного состояния линейного комплекса неоднократно рассматривалась в литературе (см., например, [14, 15]) в рамках так называемой модели «бернуллиевского сравнения», т. е. в предположении о том, что матричные элементы  $\eta_{m,n}$  (14) являются независимыми случайными величинами, принимающими значение 1 с вероятностью  $p = c^{-1}$  и значение 0 с вероятностью  $q = 1 - p$ . В работе [15] было показано, что при  $n, m \gg 1$  распределение энергии основного состояния имеет вид

$$\langle F \rangle = \frac{2\sqrt{pmn} - p(m+n)}{q} + \frac{(pmn)^{1/6}}{q} \times \left[ (1+p) - \sqrt{\frac{p}{mn}}(m+n) \right]^{2/3} \chi, \quad (22)$$

где  $\chi$  — случайная величина с распределением Трейси–Видома,  $\langle \chi \rangle = -1.7711\dots$  и  $\langle \chi^2 \rangle - \langle \chi \rangle^2 = 0.8132\dots$  (более подробное описание этого распределения можно найти, например, в недавнем обзоре [16]). При  $m = n$  энергию комплекса можно представить как

$$\langle F_{m,n} \rangle \approx \frac{2}{1 + \sqrt{c}} n + f(c) \langle \chi \rangle n^{1/3}, \quad (23)$$

где

$$f(c) = \frac{c^{1/6} (\sqrt{c} - 1)^{1/3}}{\sqrt{c} + 1}.$$

Флуктуации свободной энергии имеют вид [15]

$$\sigma \equiv \sqrt{\langle F^2 \rangle - \langle F \rangle^2} \approx \sqrt{\langle \chi^2 \rangle - \langle \chi \rangle^2} f(c) n^{1/3}. \quad (24)$$

Показатель  $1/3$  является типичным для стохастической динамики сильно коррелированных систем и относится к классу универсальности KPZ (Kardar–Parisi–Zhang) (см., например, обзор [17]).

Численные результаты для двух линейных полимеров представлены на рис. 5. Угловой коэффициент прямой  $k_l \approx 0.65$  (рис. 5), что хорошо согласуется с величиной

$$k_l = \lim_{n \rightarrow \infty} \frac{\langle F \rangle}{n} \rightarrow \frac{2}{3},$$

вычисленной по формуле (23). Для флуктуации энергии наклон равен 0.34 (рис. 5), что близко к значению  $1/3$ . Таким образом, уравнение (24), полученное в приближении бернуллиевского сравнения, удовлетворительно описывает экспериментально наблюдаемую зависимость энергии основного состояния линейного комплекса двух случайных цепей от их длины.

Аналогичный анализ был проведен и для двух последовательностей, образующих кактусообразную (т. е. РНК-подобную) структуру с  $\ell = 0$ . Графики зависимости энергии и флуктуации энергии представлены на рис. 5. Как и для комплекса линейных полимеров,  $\langle F(n) \rangle = k_c n$  при  $n \gg 1$ , но угловой коэффициент прямой  $k_c \approx 0.92$  гораздо выше, что обусловлено взаимодействием нуклеотидов внутри петель. Зависимость флуктуации энергии основного состояния остается такой же, см. рис. 5.

В двух следующих разделах мы оценим аналитически величину коэффициента  $k_c$  в зависимости свободной энергии от длины цепи для кактусообразных структур и проведем анализ распределения петель по длинам в случае комплекса линейных полимеров и последовательностей РНК.

### 5.2. Оценка свободной энергии основного состояния комплекса двух взаимодействующих РНК

Будем рассматривать комплекс, который образуют две случайные последовательности РНК, как структуру, состоящую из петель различных иерархических уровней с индексом « $i$ » (см. рис. 6). Каждую петлю  $i$ -го иерархического уровня можно рассматривать как комплекс двух взаимодействующих подпоследовательностей, из которых она состоит. Из выражения (22) следует, что наибольший вклад в свободную энергию наблюдается для комплекса, состоящего из двух последовательностей равной длины,  $m = n$  (23). Это позволяет сделать оценку сверху свободной энергии петли как свободной энергии двух взаимодействующих половинок этой петли. Представление комплекса двух молекул РНК в виде иерархической структуры позволяет использовать идеи ренормализационной группы [18]. А именно, комплексы  $i$ -го иерархического уровня содержат петли, которые будем считать комплексами  $(i+1)$ -го уровня (рис. 6),  $i = 1, 2, \dots$

Формализуя эту идею, будем полагать, что комплекс двух молекул РНК иерархического уровня  $i$  является комплексом линейных последовательностей, в которых энергия взаимодействующих мо-



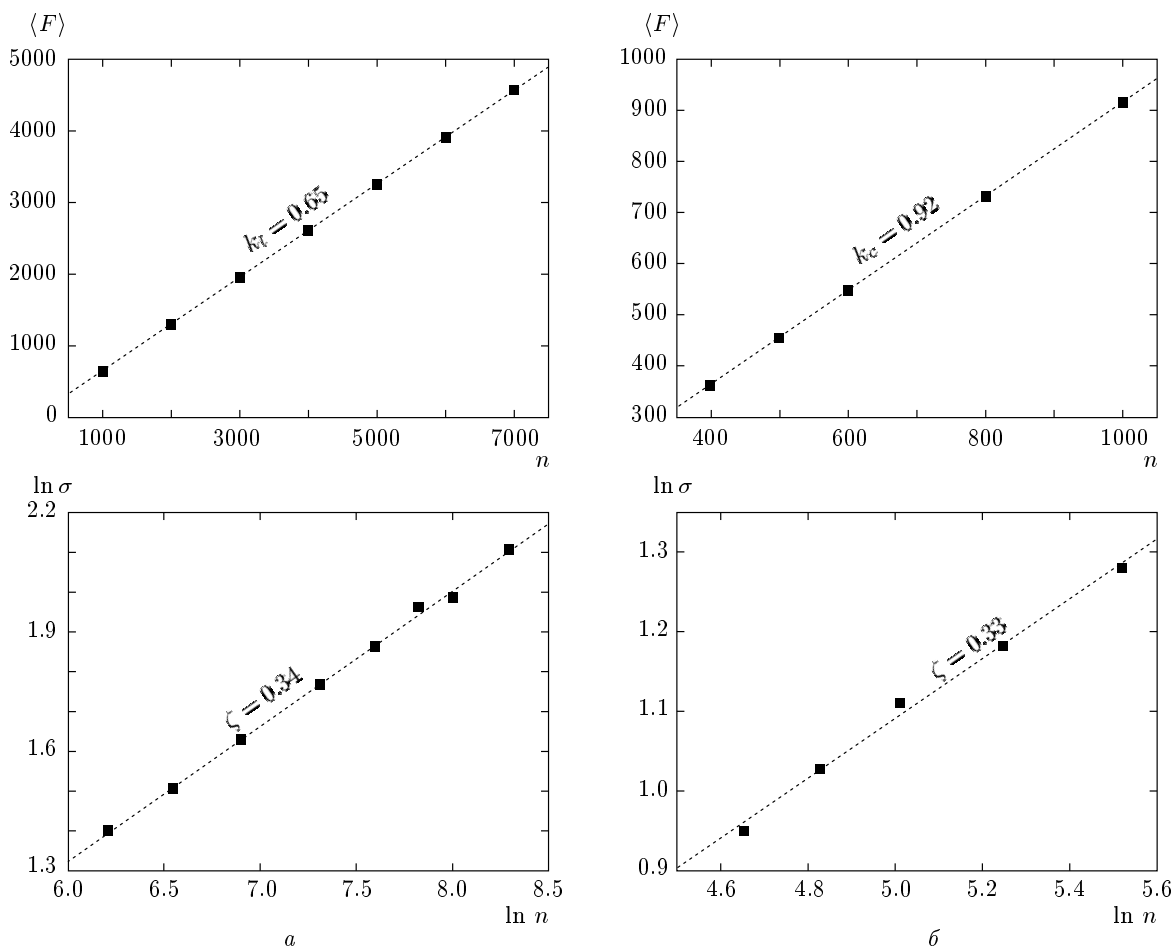


Рис. 5. Зависимости энергии основного состояния  $\langle F(n, T = 0) \rangle$  и флуктуации энергии  $\sigma(n)$ : для комплекса линейных полимеров (a) и для комплекса двух РНК (б). При каждом значении  $n$  было выполнено 1000 накоплений

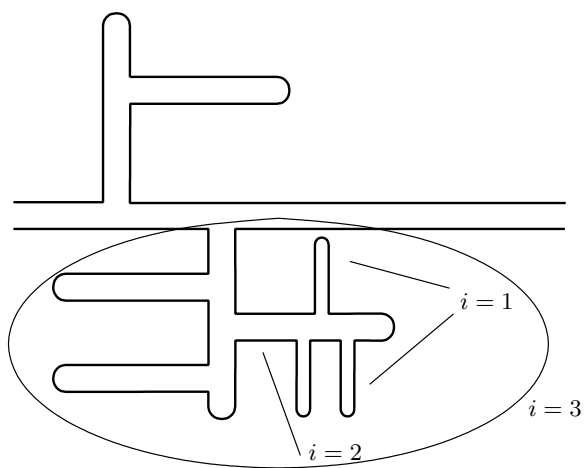


Рис. 6. Иерархическая модель для определения энергии связывания в РНК. Петли первого ( $i = 1$ ), второго ( $i = 2$ ) и третьего ( $i = 3$ ) иерархических уровней

номеров перенормирована энергией петель иерархического уровня  $i + 1$ . Пользуясь тем, что энергия петель в первом приближении пропорциональна длине (23), представим ее в виде  $F_L^{(i)} \approx k_r^{(i)} L$ . Подставляя в формулу (15) статистические веса петель  $g_{i,i+L} = e^{k_r L/T}$ , получим выражение для определения свободной энергии комплекса двух случайных РНК-последовательностей:

$$F_{m,n}^{(i+1)} = \max \left[ F_{m-1,n} + k_r^{(i)}, F_{m,n-1} + k_r^{(i)}, (F_{m-1,n-1} + u) \mathcal{P}(m, n) \right]. \quad (25)$$

Выражение (25) нужно понимать следующим образом. Прежде всего, определим свободную энергию комплекса  $F_{m,n}^{(2)}$ , в котором могут образовываться петли только первого иерархического уровня. Далее определим энергию связывания на один мономер в петлях второго уровня как

$$k_r^{(2)} = \frac{F_{m,n}^{(2)}}{m+n}. \quad (26)$$

Подставляя полученный коэффициент связывания снова в формулу (25), получим значения энергии для петель третьего иерархического уровня,  $k_r^{(3)}$ , и т. д. Величина  $\mathcal{P}(m, n)$  учитывает ограничение на минимальное количество мономеров, которые могут образовать петлю  $i$ -го иерархического уровня:

$$\mathcal{P}(m, n) = \begin{cases} 1, & \text{мономеры } m \text{ и } n \\ & \text{могут образовать связь;} \\ 0 & \text{в противном случае.} \end{cases} \quad (27)$$

Будем считать, что мономеры  $m$  и  $n$  могут образовывать связь, если

а) участок  $[m - \ell, m - 1]$  последовательности  $S_1$  не имеет связей с участком  $[n - \ell, n - 1]$  подпоследовательности  $S_2$ , где  $\ell$  — минимальное количество нуклеотидов, необходимых для формирования петли определенного уровня (если  $m < \ell$  и/или  $n < \ell$ , то рассматриваются соответственно участки последовательностей  $[1, m]$  и/или  $[1, n]$ );

б)  $m - 1$  мономеров первой последовательности взаимодействует с  $n - 1$  мономерами второй последовательности и при замене  $(m - 1) \rightarrow m$ ,  $(n - 1) \rightarrow n$  выполняется пункт а) (или б)).

В таблице приведены значения для коэффициента связывания и минимальное количество нуклеотидов в петлях  $i$ -го уровня; вычисления проводились для случайных последовательностей равной длины  $m = n = 10^4$ . Длины последовательностей слабо влияют на средний коэффициент связывания, однако рассмотрение больших длин позволяет провести оценку для большего количества иерархических уровней. Отметим, что коэффициент связывания, определяемый по данной иерархической процедуре, медленно (логарифмически) стремится к единице с ростом количества иерархических уровней (т. е. при  $n \rightarrow \infty$ ). Логарифмическая зависимость обусловлена экспоненциальным ростом минимального числа мономеров, которые могут образовать петлю,  $l_i = 3l_{i-1} + 6$  ( $i > 2$ ) с увеличением номера иерархического уровня  $i$ , см. таблицу.

Таким образом, численно наблюдаемый коэффициент связывания  $k_c$  (рис. 5б) в действительности зависит (логарифмически) от длин рассматриваемых последовательностей и полученное нами значение  $k_c \approx 0.92$  лишь указывает на то, что последовательности длиной 400–1000 мономеров образуют структуру всего с двумя–тремя иерархическими уровнями.

### 5.3. Распределение петель по длинам в линейных и РНК-подобных комплексах

Мы проанализировали распределение длин петель в структуре комплекса для двух линейных полимеров и для последовательностей типа РНК. На рис. 7а представлена зависимость  $W(s)$  числа петель различной длины  $s$  для комплекса двух взаимодействующих линейных полимеров. Видно, что зависимость с хорошей точностью является экспоненциальной. Такое распределение характерно для системы, в которой связывание различных мономеров в цепи происходит независимо (т. е. вероятность того, что следующий по цепи мономер образует связь, никак не зависит от того, образует ли связь предыдущий мономер). Действительно, величину  $k_l = \langle F \rangle / n$  при  $n \gg 1$  можно рассматривать как вероятность связывания мономера в комплексе. Считая, что связывание мономера не зависит от того, участвует ли предыдущий мономер во взаимодействии, число петель длиной  $s$  в структуре комплекса из двух линейных цепей длины  $n$  можно оценить как

$$W(s) = nk_l^2(1 - k_l)^s. \quad (28)$$

Такое распределение длин петель при  $n \gg 1$  удовлетворяет очевидному соотношению:

$$\sum_{s=1}^n sW(s) = (1 - k_l)n.$$

На рис. 7а видно, что численные результаты хорошо аппроксимируются в логарифмическом масштабе прямой  $y(s) = a - bs$ , где с хорошей точностью  $a \approx \ln(nk_l^2)$  и  $b \approx \ln(1 - k_l)$  (см. (28)). Таким образом, в линейных комплексах статистика петель выглядит в точности так, как при независимом связывании мономеров.

Существенно иное поведение имеет статистика петель в комплексах кактусообразной структуры, характерной для РНК. На рис. 7б представлена зависимость числа петель с длиной  $s$  по набору из 200 пар случайных последовательностей. Отметим особенности наблюдаемого распределения. Во-первых, для зависимости характерно степенное поведение. Показатель степенной зависимости для РНК разной длины меняется в интервале  $[1.38, 1.5]$ . Во-вторых, распределение для РНК с различной длиной  $n$  совпадают, что позволяет проводить вычисления для набора коротких последовательностей. В-третьих, при малых  $s$  ( $s \leq 5$ ) характерно небольшое число петель с нечетной длиной и большое число петель с четной длиной (см. рис. 7б). Последнее обстоятельство

Таблица. Вероятность связывания мономеров в зависимости от числа уровней

Число учтенных иерархических уровней, $i$	2	3	4	5	6	7
Минимальное число мономеров на $i$ -м уровне, $l_i$	2	6	24	78	240	726
Коэффициент связывания при учете $i$ уровней, $k_r^{(i)}$	0.851	0.912	0.931	0.937	0.94	0.941

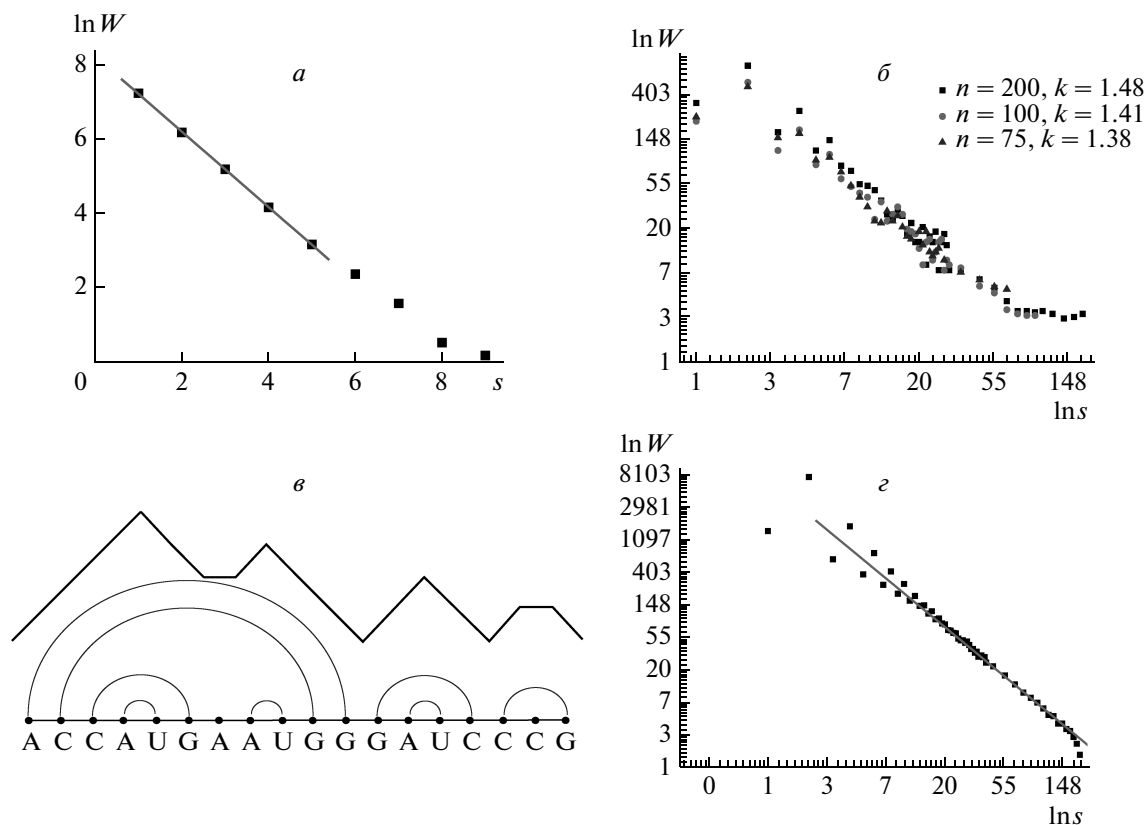


Рис. 7. Распределения длин петель в структуре линейного комплекса двух последовательностей (а) (вычисления были выполнены для случайных полимеров длины  $n = 10^4$ , результаты усреднялись по набору из 1000 последовательностей) и в кактусообразной структуре (б) (вычисления проводились для последовательностей с длинами  $n = m = 75, 100, 200$ , для каждого  $s$  было выполнено 400 накоплений, для  $s \geq 30$  функция распределения сглаживалась по 10 соседним значениям); в — структура петли и соответствующий ей путь Моцкина; z — распределение путей Моцкина по длинам (длина пути случайного блуждания 200 шагов, количество накоплений  $10^3$ , для  $s \geq 30$  функция распределения сглаживалась по 10 соседним значениям)

связано с тем, что для структуры комплекса двух взаимодействующих РНК с  $\ell = 0$  характерно высокое значение средней энергии на один нуклеотид ( $k_c \approx 0.92$ ), которое обусловлено внутривитчелковым взаимодействием, а образование петли малой длины с нечетным числом нуклеотидов приводит к потере, по крайней мере, одной возможной связи внутри петли. Таким образом, образование петель с нечетным числом мономеров энергетически невыгодно. И наконец, для распределения характерно наличие плато при больших  $s$ , что обусловлено эффектом конеч-

ного размера (см., например, [19], где построена теория аналогичного эффекта в несколько иной системе).

Обсудим полученные численные распределения. Каждой вторичной структуре полимера поставим в соответствие одномерное случайное блуждание на  $1 + 1$ -мерной решетке, построенное следующим образом (см. рис. 7в). Каждому мономерному звену соответствует один шаг блуждания. Этот шаг направлен направо вверх, если мономер является «началом петли» (т. е. связан с мономером, расположен-

ным после него по цепи), направо вниз, если он является «концом петли» (т. е. связан с мономером, расположенным до него по цепи) или горизонтально, если мономер не образует связи. Легко видеть, что такое построение задает взаимно однозначное соответствие между РНК-подобными вторичными структурами полимера и так называемыми путями Моцкина [20] — состоящими из горизонтальных и диагональных участков, дискретными случайными блужданиями в верхней полуплоскости, концы которых закреплены на оси абсцисс. Возвращение на ось абсцисс соответствует образованию одной петли в структуре комплекса двух взаимодействующих РНК. Как известно [21], количество различных путей Моцкина длины  $n$  с заданным количеством горизонтальных шагов  $t$  определяется числами Каталана:

$$P(n, t) = \binom{n}{t} C_{(n-t)/2} = \binom{n}{t} \frac{1}{\frac{n-t}{2} + 1} \binom{n-t}{\frac{n-t}{2}}, \quad (29)$$

где  $\binom{n}{t}$  — биномиальные коэффициенты,  $C_{(n-t)/2}$  — числа Каталана. При  $n \gg 1$  (29) имеет асимптотическую зависимость  $P(n, t) \sim n^{-3/2}$  от длины пути. Мы построили распределение длин петель в случайных путях Моцкина, приняв вероятность диагонального шага вверх или вниз равной  $p_m \approx k_c/2 = 0.46$ , где  $k_c = 0.92$  — экспериментально полученное значение вероятности образования связи в случайной РНК, а вероятность горизонтального шага  $1 - 2p_m$ . Результат приведен на рис. 7г. Видно, что зависимость обладает всеми характерными свойствами, наблюдаемыми для распределения длин петель в структуре комплекса двух РНК.

#### 5.4. Выделенность алфавита из четырех букв

Представление структур РНК в виде путей Моцкина, статистика которых известна, позволяет сделать следующее наблюдение. Зададимся вопросом о том, к какому пределу стремится доля мономеров, образующих связи, в основном состоянии длинной ( $n \rightarrow \infty$ ) цепи РНК. Другими словами, мы интересуемся удельной (в расчете на одно звено) энергией основного состояния длинной РНК. Оказывается, что это предельное значение энергии существенно

зависит от числа используемых различных нуклеотидов (букв в алфавите РНК)  $c$ . Существует критическое значение числа букв  $c = c_{cr}$ , такое что при  $c < c_{cr}$  доля связанных звеньев стремится к единице, тогда как при  $c > c_{cr}$  предельная доля связанных звеньев меньше единицы. Убедиться в этом можно следующим образом. Для того чтобы доля связанных звеньев в «кактусе», образуемом случайной последовательностью, была равна единице, каждой последовательности из  $c^n$  возможных должен соответствовать так называемый путь Дика (т. е. путь Моцкина, в котором нет горизонтальных шагов). Количество путей Дика  $G(n)$  длины  $n$  определяется формулой (29), при  $k = 0$  и при  $n \gg 1$  имеет асимптотическое выражение:

$$G(n) \sim \frac{4^{n/2}}{n^{3/2}}. \quad (30)$$

Заметим, что один и тот же путь Дика может описывать основное состояние разных полимеров. Действительно, каждая пара «подъем/спуск» в пути Дика может быть независимо от остальных реализована с разными способами (в случае РНК возможные варианты — это А-Т, Т-А, С-Г, Г-С). Таким образом, число различных первичных структур, для которых существуют совершенные (т. е. полностью связанные) вторичные структуры, не превышает

$$W(n) = G(n)c^{n/2} \sim \frac{(4c)^{n/2}}{n^{3/2}}. \quad (31)$$

Это оценка сверху, так как, вообще говоря, одной и той же последовательности может соответствовать несколько различных путей Дика. Тем не менее, естественно ожидать, что при  $c \rightarrow c_{cr}$  число таких последовательностей с двумя и более совершенными вторичными структурами становится малым. В таком случае, сравнивая (31) с полным числом возможных первичных структур  $W_0(n) = c^n$ , получаем  $c_{cr} = 4$ : при  $c < c_{cr}$  в пределе  $n \rightarrow \infty$  практически любой последовательности соответствует совершенная вторичная структура и энергия последовательности на одну пару нуклеотидов стремится к единице, в то время как для случайных полимеров с  $c > c_{cr}$  доля последовательностей, допускающих совершенную вторичную структуру, экспоненциально мала. Добавление горизонтальных шагов в пути Моцкина приводит к увеличению возможных кактусообразных конфигураций (29), что позволяет сопоставить каждой случайной последовательности с  $c \geq c_{cr}$  путь в случайном блуждании, соответствующий ее оптимальной вторичной структуре. Однако в этом случае доля связанных звеньев в оптимальной вторичной структуре остается в пределе  $n \rightarrow \infty$

меньше единицы. При  $c_{cr} = 4$  (что, напомним, соответствует реальным РНК) мы ожидаем критическое поведение: при  $n \rightarrow \infty$  коэффициент связывания стремится к единице логарифмически (см. таблицу).

Таким образом,  $c_{cr} = 4$  оказывается выделенным значением для числа букв: это наибольшее число, при котором длинная случайная РНК в термодинамическом пределе может образовывать почти совершенную вторичную структуру.

По современным представлениям [22] возникновение жизни было связано со спонтанной саморепликацией случайных молекул РНК. В связи с этим представляется, что приведенное наблюдение может иметь отношение к объяснению того факта, что природные ДНК и РНК имеют именно четыре различных нуклеотида. Развернутое обсуждение данного вопроса приведено в работе [23].

## 6. ЗАКЛЮЧЕНИЕ

В настоящей работе предложен новый подход к определению энергии связывания двух биополимеров, способных образовывать сложную иерархическую вторичную структуру, свойственную молекулам РНК. Напомним еще раз основные положения модели. Во-первых, задачу о выравнивании двух буквенных последовательностей мы представили как задачу о вычислении свободной энергии основного состояния (т.е. при  $T \rightarrow 0$ ) статистической модели, описывающей комплексообразование двух линейных полимеров. Во-вторых, учли способность каждого из полимеров образовывать сложную иерархическую РНК-подобную структуру и получили выражение для статистической суммы такого комплекса при температуре, отличной от нуля. В-третьих, переходя в конечном выражении снова к пределу  $T \rightarrow 0$ , получили нелокальные рекуррентные соотношения для определения энергии основного состояния комплекса двух молекул РНК.

Следует отметить, что структура образующегося комплекса двух полимеров сильно зависит от деталей модели. Так, структуры, представленные на рис. 4б,в отличаются только одним параметром в модели — минимальным размером петли  $\ell$ . Сильная чувствительность глобальной топологии оптимальной структуры к микроскопическим деталям модели ясно показывает, что, для того чтобы получать экспериментально достоверные результаты, необходимо иметь подробную информацию о точных значениях петлевого фактора, энергий связей и параметра кооперативности. Как уже указывалось, при

необходимости все эти параметры можно учесть, не выходя за рамки предложенной модели.

В данной работе мы проанализировали такие важные статистические свойства комплекса двух случайных последовательностей РНК, как флуктуации средней энергии связывания двух макромолекул, а также распределение длин петель (gaps) в образуемой структуре. Результаты показали, что средние энергии и линейных и РНК-подобных структур удовлетворяют статистике, принадлежащей классу универсальности KPZ. Распределение длин петель в линейных комплексах совпадает со статистикой, характерной для независимого связывания мономеров в цепи. Распределение петель в «кактусах» хорошо описывается моделью случайного блуждания. Кроме того, проанализировав поведение удельной свободной энергии комплекса двух случайных молекул РНК в зависимости от числа различных мономеров  $c$  («букв в алфавите»), мы высказали гипотезу о наличии критического поведения при  $c = 4$ , что делает используемый природой алфавит из четырех букв выделенным.

Авторы благодарны А. А. Миронову и В. А. Аветисову за обсуждение биофизических и статистических аспектов задачи. Работа выполнена при частичной поддержке гранта ERASysBio+ № 66; М. В. Тамм и О. В. Вальба благодарны за теплый прием лаборатории LPTMS, где была выполнена основная часть данной работы.

## ЛИТЕРАТУРА

1. K.-M. Schmitz, C. Mayer, A. Postepska, and I. Grummt, *Genes Dev.* **24**, 2264 (2010).
2. S. R. Eddy, *Cell* **109**, 137 (2002).
3. V. Ambros, *Cell* **107**, 862 (2001).
4. A. Grosberg and A. Khokhlov, *Statistical Physics of Macromolecules*, AIP, New York (1994).
5. V. Pande, A. Grosberg, and T. Tanaka, *Rev. Mod. Phys.* **72**, 259 (2000).
6. M. V. Tamm and S. K. Nechaev, *Phys. Rev. E* **75**, 031904 (2007).
7. M. S. Waterman, *Bull. Math. Biol.* **46**, 473 (1984).
8. R. Bundschuh and T. Hwa, *Discrete Appl. Math.* **104**, 113 (2000).
9. D. Drasdo, T. Hwa, and M. Lassig, *J. Comp. Biol.* **7**, 115 (2000).

10. P. de Gennes, *Biopolymers* **6**, 715 (1968).
11. И. Я. Ерухимович, *Высокомолек. соед.* **20B**, 10 (1978).
12. S. K. Nechaev, M. V. Tamm, and O. V. Valba, *J. Phys. A: Math. Theor.* **44**, 195001 (2011).
13. S. R. Morgan and P. G. Higgs, *J. Phys. A: Math. Gen.* **31**, 31533170 (1998).
14. J. B. de Monvel, *European Phys. J. B* **7**, 293 (1999); *Phys. Rev. E* **62**, 204 (2000).
15. S. N. Majumdar and S. Nechaev, *Phys. Rev. E* **69**, 011103 (2004).
16. T. Kriecherbauer and J. Krug, *J. Phys. A: Math. Theor.* **43**, 403001 (2010).
17. T. Halperin-Healy and Y. C. Zhang, *Phys. Rep.* **254**, 215 (1995).
18. Ш. Ма, *Современная теория критических явлений*, Мир, Москва (1980).
19. М. В. Тамм, Н. Г. Лисаченко, И. Я. Ерухимович, В. А. Иванов, *Высокомолек. соед.* **47**, 348 (2005).
20. С. К. Ландо, *Лекции о производящих функциях*, МЦНМО, Москва (2004).
21. W. Feller, *An Introduction to Probability Theory and Its Applications*, J. Wiley & Sons (1968).
22. L. E. Orgel, *Trends Biochem. Sci.* **23**, 491 (1998).
23. O. V. Valba, S. K. Nechaev, and M. V. Tamm, arXiv:1109.5410v1[q-bio.GN].