ТОПОЛОГИЧЕСКИЕ СВОЙСТВА РНК-ПОДОБНЫХ МОЛЕКУЛ СО СЛУЧАЙНОЙ ПЕРВИЧНОЙ СТРУКТУРОЙ

О. Вальба[1, 2], М. Тамм[1, 3], С. Нечаев[4, 5]

[1]*Национальный исследовательский университет Высшая школа экономики, Москва*

[2]*Институт химической фихики им. Н.Н. Семенова РАН, Москва*

[3]*Московский государственный университет им. М.В. Ломоносова, Москва*

[4]*Universitet Paris-Sud/CNRS,Orsay, France*

[5]*Физический институт им. П.Н. Лебедева РАН, Москва*

ovalba@hse.ru

Поступила 20.04.2015

Работа посвящена применению методов статистической физики и теории случайных процессов для исследования топологических свойств РНК-подобных гетерополимеров со случайной первичной структурой. В частности, описывается критического изменение топологии РНК-подобных структур в зависимости от алфавита, используемого в случайной последовательности, приводится аналитической оценка точки перехода в рамках комбинаторного и матричного описания.

УДК 538.9

I. ВВЕДЕНИЕ

Структура важнейших биологических макромолекул, таких как дезоксирибонуклеиновые кислоты (ДНК), рибонуклеиновые кислоты (РНК) и белки, играет ключевую роль в их правильном функционировании в клетке. Различают три уровня структурной

упорядоченности биомакромолекул. Одна из основных их особенностей состоит в гетерополимерности. Последовательность звеньев в ДНК, РНК и белках индивидуального организма, она называется первичной структурой, строго зафиксирована. Далее, биополимерные цепи могут формировать спиралеобразные и складчатые участки небольшого масштаба, как в белках, или комплементарно спаренные и петлевые участки, как в РНК. Такие фрагменты называются элементами вторичной структуры. Различают также третичную и четвертичную пространственные структуры биополимеров.

Данная работа посвящена исследованию топологических свойств вторичной структуры молекул РНК-типа. Известно, что биомакромолекулы являются «слабо отредактированными случайными гетерополимерами» [1, 2]. Более того, для ряда свойств распределение мономерных звеньев в первичной структуре, например, функциональных РНК можно считать случайным [3, 4]. В этом случае, модель случайной первичной структуры является базовой моделью, описывающей основной (нулевой) вклад в наблюдаемые физические явления. Основное внимание при этом сфокусировано на нетривиальной вторичной структуре РНК-подобных полимеров, для описания которой привлекаются разнообразные техники, в том числе, техники квантовой теории поля и моделей Изинга[5].

Структура работа такова. В разделе II приводятся алгоритмы описания РНК-подобной структуры и вычисления свободной энергии ее основного состояния. Формулируется вспомогательная статистическая модель, описывающая взаимодействия мономерных звеньев в РНК-подобной структуре с петлевыми участками. Далее, в предложенной модели учитывается вклад внутрипетлевого взаимодействия мономеров и приводится соответствующий алгоритм динамического программирования для вычисления энергии такой иерархической структуры.

Раздел III посвящен определению свойств распределения свободной энергии ансамбля случайных последовательностей РНК. Обсуждаются такие характеристики, как среднее значение свободной энергии в ансамбле, флуктуация средней энергии, распределение по длинам петель в пространственных структурах.

Анализ топологических свойств в модели случайной первичной структуры РНК-подобной молекулы в зависимости от используемой в последовательности алфавита описан в отдельный раздел IV. Показывается, что в зависимости от алфавита РНК-подобная структура характеризуется либо максимально связанной вторичной структурой без пропусков (неспаренных мономеров), либо структурой с конечной долей несвязанных мономеров. Для определения точки такого топологического перехода фор-

мулируется модель Бернулли. В рамках предложенной модели приводятся численные и аналитические оценки критической точки перехода.

II. АЛГОРИТМЫ ВЫЧИСЛЕНИЯ СВОБОДНОЙ ЭНЕРГИИ РНК-ПОДОБНЫХ СТРУКТУР

A. Выравнивание последовательностей

Задача о выравнивании двух последовательностей – это задача нахождения эффективного алгоритма поиска наибольшей общей подпоследовательности (НОП) 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 об НОП широко исследовалась в биологии, компьютерных науках, теории вероятности и позже в статистической физике.

Задача о поиске НОП двух последовательностей формулируется следующим образом. Рассмотрим две произвольные последовательности (в качестве примера рассматриваются последовательности РНК, составленные из 4-х буквенного алфавита $\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}$): $S_1 = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{C}, \mathsf{U}, \mathsf{A}, \mathsf{C}\}$ длины $m = 7$ и $S_2 = \{\mathsf{C}, \mathsf{U}, \mathsf{G}, \mathsf{A}, \mathsf{C}\}$ длины $n = 5$. Далее, везде под алфавитом подрузамевается количество различных мономерных хвеньев в первичной структуре. Общая подпоследовательность – это подпоследовательность, содержащая буквы (нуклеотиды) как первой, так и второй последовательности, причем подпоследовательность необязательно содержит буквы, идущие непосредственно друг за другом. Так, например, для двух последовательностей S_1 и S_2 можно выделить несколько различных общих подпоследовательностей, например, $\{\mathsf{C}, \mathsf{U}, \mathsf{A}, \mathsf{C}\}$ или $\{\mathsf{G}, \mathsf{A}, \mathsf{C}\}$ – обе эти подпоследоавтельности содержатся в S_1 и S_2, и являются для них общими. Число возможных общих подпоследовательностей с ростом длин m и n полимеров растет экспоненциально. Алгоритм для определения оптимального выравнивания двух последовательностей впервые был сформулирован в [6]. В наиболее общем смысле каждое выравнивание двух последовательностей характеризуется числом совпадающих и несовпадающих букв и числом пропусков (делеций) в выравненных последовательностях. Для каждого выравнивания можно ввести весовую функцию (cost function), имеющую значение энергии [6]:

$$F = N_{\text{match}} + \mu\, N_{\text{mis}} + \delta\, N_{\text{gap}}. \tag{1}$$

В формуле (1) N_{match}, N_{mis} и N_{gap} – число пар совпадающих букв, число пар несовпадающих букв и число делеций в рассматриваемом выравнивании, соответственно. Величины μ и δ – это вклады в весовую функцию от пары несовпадающих букв и делеции; вклад от пары совпадающих нуклеотидов, без потери общности, можно считать равным 1. В таком представлении функция F удовлетворяет очевидному закону сохранения:

$$n + m = 2N_{\text{match}} + 2N_{\text{mis}} + N_{\text{gap}}. \tag{2}$$

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

$$\tilde{F} = N_{\text{match}} + \gamma N_{\text{mis}}, \tag{3}$$

где

$$\gamma = \frac{\mu - 2\delta}{1 - 2\delta}. \tag{4}$$

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

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

где

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

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 (5) соответствуют трем возможным ситуациям: пропуску буквы в первой последовательности, пропуску во второй

последовательности и случаю, когда i-ая буква первой последовательности выравнена с j-ой буквой второй последовательности.

B. Комплементарное связывание биополимеров

Цель работы заключается в разработке статистического алгоритма вычисления весовой функции, которая бы характеризовала «похожесть» двух заданных последовательностей со сложной вторичной структурой типа 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 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 компексообразования, 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 $T \to 0$, найдем искомую весовую функцию.

Рассмотрим вспомогательную статистическую модель, описывающую взаимодействие двух линейных полимеров с произвольными первичными последовательностями. Пусть длины этих последовательностей, измеренные в единицах мономерных звеньев, равны m и n, соответственно. Каждый мономер может быть выбран из c различных мономеров $A, B, C, D, ...$ (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. Будем считать энергию связи между комплементарными нуклеотидами равной $-u$, а энергию между некомплементарными равной $-v$, где u и v — некоторые положительные величины ($|v| > |u|$). Предположим также, что некоторые части полимеров могут образовывать петли. На Рис. 1 схематически представлено взаимодействие

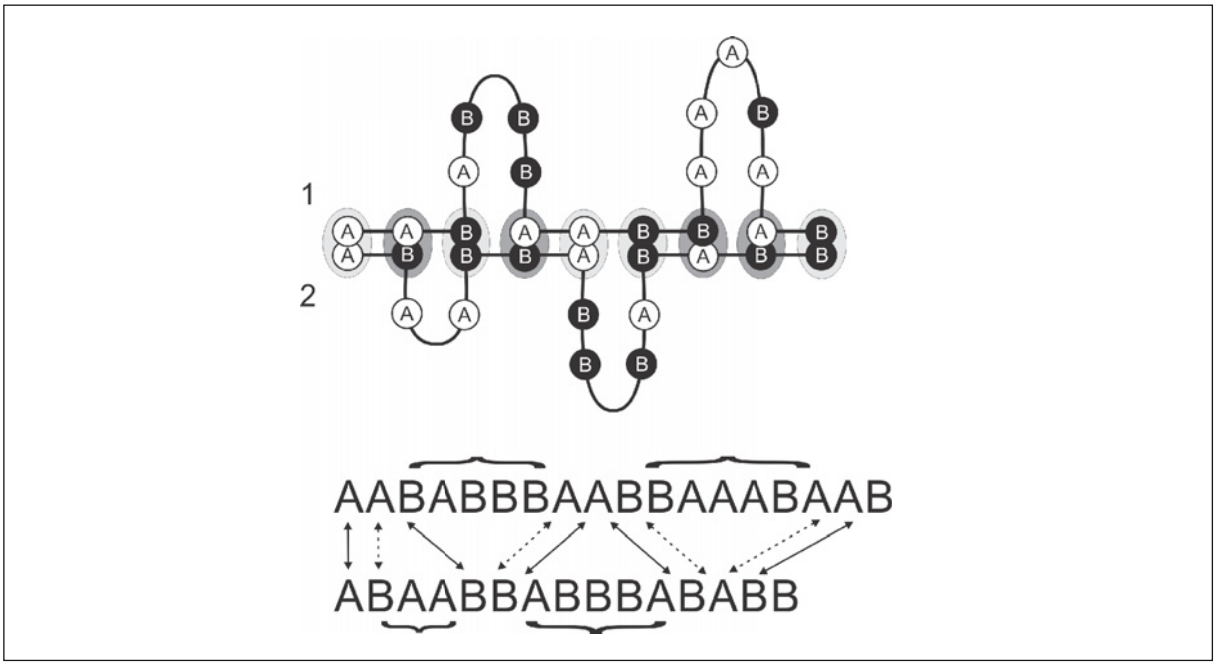


Рис. 1. Взаимодействие мономерных звеньев в РНК-подобной структуре с петлевыми участками как выравнивание соответствующих последовательностей (первичных структур).

двухбуквенных полимеров. Очевидно, что петли соответствуют делециям в задаче о выравнивании двух последовательностей.

Задача заключается в вычислении свободной энергии описанной модели при достаточно низких температурах, при которых энтропийным вкладом можно пренебречь по сравнению с энергетическим. 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:

$$\begin{cases} G_{m,n} = 1 + \sum\limits_{i,j=1}^{m,n} \beta_{i,j}\, G_{i-1,j-1} \\ G_{m,0} = 1;\ G_{0,n} = 1;\ G_{0,0} = 1. \end{cases} \tag{7}$$

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 $\beta_{i,j}$ определяются энергией контакта между i-ым и j-ым мономерами:

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

Здесь и далее, $T \equiv k_B T$. Легко проверить, 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 комплекса $F_{m,n}$ и температурой T известным соотношением $G_{m,n} = \exp\{-F_{m,n}/T\}$. Будем интересоваться значением свободной энергии с точностью до знака, тогда для величины $\tilde{F}_{m,n} = -F_{m,n}$, переходя в уравнении (9) к пределу $T \to 0$, получим:

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

Формулу (10) можно переписать в виде:

$$\tilde{F}_{m,n} = \max\left[\tilde{F}_{m-1,n},\, \tilde{F}_{m,n-1},\, \tilde{F}_{m-1,n-1} + \eta_{m,n}\right], \tag{11}$$

где введено обозначение:

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

Принимая η^+ за единицу энергии, перепишем формулу (11) в виде:

$$\tilde{F}_{m,n} = \max\left[\tilde{F}_{m-1,n},\, \tilde{F}_{m,n-1},\, \tilde{F}_{m-1,n-1} + \tilde{\eta}_{m,n}\right], \tag{13}$$

с

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

И функция $\tilde{F}_{m,n}$ удовлетворяет начальным условиям: $\tilde{F}_{0,n} = \tilde{F}_{n,0} = \tilde{F}_{0,0} = 0$. Видно, что выражение свободной энергии связывания двух полимеров без петлевых взаимодействий имеет вид, совпадающий с (5). Далее, все результаты численного моделирования представлены для $\frac{v}{u} = 0$.

Таким образом, рекурсия, используемая в методе динамического программирования является ничем иным, как рекуррентным соотношением на свободную энергию взаимодействия 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 первым приближением к описанию комплексообразования биополимеров. Известно (см., например, [2]), что

для точного количественного описания такого связывания, например, двойной спирали ДНК необходимо учесть еще ряд факторов. Во-первых, не учтены 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, не учтено, что гибкость полимера конечна и, тем самым, существует ограничение на минимальную длину петли. И наконец, не было принято во внимание то обстоятельство, что комплементарные пары $A-T$ и $C-G$ имеют различную энергию связи и, что помимо комплементарных пар, возможно образование неканонических пар.

Обобщение выражений (7)–(14) с учетом кооперативности образования связи, минимальной длины петли и различной энергией комплементарных связей — задача вычислительно сложная, но не требующая качественного изменения предложенного формализма, т.к. эти факторы влияют только на локальные свойства полимерных цепей. С другой стороны, петлевой фактор — характеристика нелокальная, зависящая от расстояния между мономерами, образующими связи и в этом случае нельзя описать состояние комплекса уравнениями динамического программирования, вида (7). Однако, поскольку петлевые факторы имеют энтропийную природу, в пределе низких температур ($T \to 0$) их вклад в свободную энергию гетерополимерного комплекса становится пренебрежимо мал. Ситуация 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 необходим.

C. Связывание РНК с внутрипетлевым взаимодействием

В этом разделе обобщается модель взаимодействия двух сополимеров на случай, когда возможно комплементарное связывание внутри петель комплекса. Будем рассматривать иерархические структуры петель типа клеверного листа (Рис. 2(а)), структуры типа псевдоузлов (Рис. 2(б)) в данной работе не рассматриваются.

Как и в предыдущем параграфе для простоты не будем учитывать кооперативность образования связей и различие в энергиях комплементарных пар. Однако, как уже

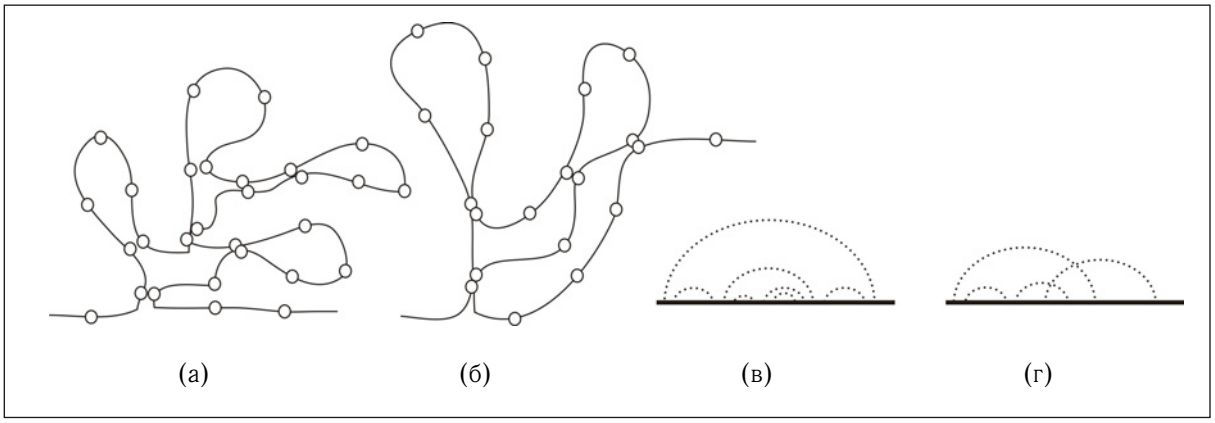


Рис. 2. Клеверная структура РНК (а) и псевдоузел (б); (в) и (г) — арочное представление (а) и (б), соответственно

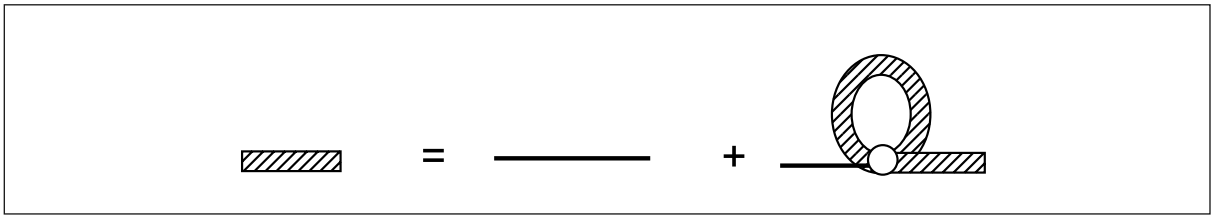


Рис. 3. Диаграмма для вычисления статистического веса g последовательности

указывалось, модель может быть обобщена с учетом этих факторов. Выражение (7) для статистической суммы $G_{m,n}$ двух взаимодействующих сополимеров можнопереписать в виде:

$$\begin{cases} G_{m,n} = g^{(1)}_{1,m}\, g^{(2)}_{1,n} + \sum\limits_{i,j=1}^{m,n} \beta_{i,j}\, G_{i-1,j-1}\, g^{(1)}_{i+1,m}\, g^{(2)}_{j+1,n} \\ G_{m,0} = g^{(1)}_{1,m}; \;\; G_{0,n} = g^{(2)}_{1,n}; \;\; G_{0,0} = 1, \end{cases} \tag{15}$$

где $g^{(1)}_{i,j}$ и $g^{(2)}_{i,j}$ обозначены статистические веса участков (с i-го нуклеотида до j-го) первой и второй последовательности, соответственно, удовлетворяющие уравнениям:

$$\begin{cases} g^{(a)}_{i,j} = 1 + \sum\limits_{k=i}^{j-1} \sum\limits_{l=i+1+\ell}^{j} \beta'_{k,l} g^{(a)}_{k+1,l-1}\, g^{(a)}_{l+1,j}; \\ g^{(a)}_{i,i} = 1, \quad a = 1, 2. \end{cases} \tag{16}$$

Эти уравнения отвечают за топологию кактусообразной структуры, свойственной молекулам РНК, диаграмма, описывающая такие структуры представлена на Рис.3. Коэффициенты $\beta'_{i,j}$ – это константы, описывающие взаимодействие внутри последовательности, аналогичные $\beta_{m,n}$. Суммирование по j ведется от $i+1+\ell$ до n для того чтобы исключить петли длиной меньше ℓ мономеров. В последующих вычислениях, как правило, предполагается, что ℓ=0, также обсуждается случай $\ell = 3$. Напомним еще раз, что так

как интерес представляет низкие температуры, можно пренебречь вкладом, связанным с потерей энтропии при образовании петель.

Сложную систему уравнений на статистические веса петлевых участках $g^{(a)}_{i,j}$, $a = 1, 2$ (16) можно решить следующим образом. Для каждой из последовательности РНК можно построить матрицу g, (i, j)-й элемент которой определяет статистический вес участка, 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) можно однозначно определить элементы $g^{(a)}_{i,i+1}$. Из (16) следует, что элементы последующих субдиагоналей $g^{(a)}_{i,i+k}$ зависят только от элементов предыдущих субдиагоналей $g^{(a)}_{i,i+k-l}$ матрицы:

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

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

Как и в случае связывания последовательностей без петлевых участков, можно выполнить переход к пределу нулевой температуры – см. выражения (7)-(14). Элементы матрицы свободной энергии при этом можно представить в виде:

$$F_{m,n} = \max_{\substack{i=1,\dots,m \\ j=1,\dots,n}} \left[f^{(1)}_{1,m} + f^{(2)}_{1,n}, Q^{m,n}_{i,j} \right] \tag{18}$$

где $f^{(a)}_{i,j} = \lim\limits_{T \to 0} \left[T \ln g^{(a)}_{i,j} \right]$ $(a = 1, 2)$ имеют смысл с точностью до знака свободных энергий петлевых участков последовательностей с i-го нуклеотида по j-й, $Q^{m,n}_{i,j}$ – (i, j)-ый элемент суммы (15), который в пределе нулевой температуры есть:

$$Q^{m,n}_{i,j} = F_{i-1,j-1} + f^{(1)}_{i+1,m} + f^{(2)}_{j+1,n} + \tilde{\eta}_{i,j}. \tag{19}$$

Элемент $Q_{i,j}$ описывает энергию комплекса взаимодействующих РНК, не имеющих контакта правее пары (i, j). Из (17) следует, что функции $f^{(a)}_{i,j}$ удовлетворяют:

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

здесь величина $\tilde{\eta}_{i,j}$ – величина, как в (14), $\tilde{\eta}'^{(a)}_{i,s}$ – аналогичная величина, описывающая взаимодействие внутри петель. На свободную энергию накладываются граничные условия, как это следует из (15):

$$\begin{cases} F_{0,0} = 0; \\ F_{i,0} = f^{(1)}_{1,i};\ 1 \le i \le m \\ F_{0,j} = f^{(2)}_{1,j};\ 1 \le j \le n. \end{cases} \tag{21}$$

Таким образом, для того, чтобы вычислить энергию основного состояния комплекса двух взаимодействующих РНК, необходимо построить матрицы $f^{(1)}$ и $f^{(2)}$ и, далее, применяя (18)-(19), определить элементы матрицы F.

Отметим, что выражения (17), (20) можно использовать для непосредственного вычисления свободной энергии основного состояния одноцепочечной РНК.

III. СВОЙСТВА РНК СТРУКТУР СО СЛУЧАЙНОЙ ПОСЛЕДОВАТЕЛЬНОСТЬЮ ЗВЕНЬЕВ

В данном разделе обсуждаются свойства распределения свободной энергии основного состояния в ансамбле РНК-подобных молекул со случайной первичной структурой. Также, приводятся результаты для распределения длин петель в РНК-подобных структурах и обсуждаются аналитические модели их описания.

A. Свободная энергия основного состояния

1. Связывание двух РНК с петлевыми участками

Задача поиска оптимальной конфигурации линейного выравнивания случайных последовательностей неоднократно рассматривалась в литературе (см., например, [7]) 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 значения 1 с вероятностью $p = c^{-1}$ и 0 с вероятностью $q = 1 - p$, где c — алфавит, используемый в случайной первичной структуре полимера. В работе [7] было показано что для длин последовательностей $n, m \gg 1$ распределение энергии основного состояния имеет вид:

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

где χ – случайная величина с распределением Трейси–Видома ($\langle\chi\rangle = -1.7711...$ и $\langle\chi^2\rangle - \langle\chi\rangle^2 = 0.8132...$)(более подробное описание этого распределения можно найти, например, в обзоре [8]). При $m = n$, оптимальная конфигурация характеризуется:

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

где

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

Флуктуации свободной энергии подчиняются [7]:

$$\sigma \equiv \sqrt{\left\langle F_{n,n}^2\right\rangle - \langle F_{n,n}\rangle^2} \approx \sqrt{\langle\chi^2\rangle - \langle\chi\rangle^2} f(c) n^{1/3}. \tag{24}$$

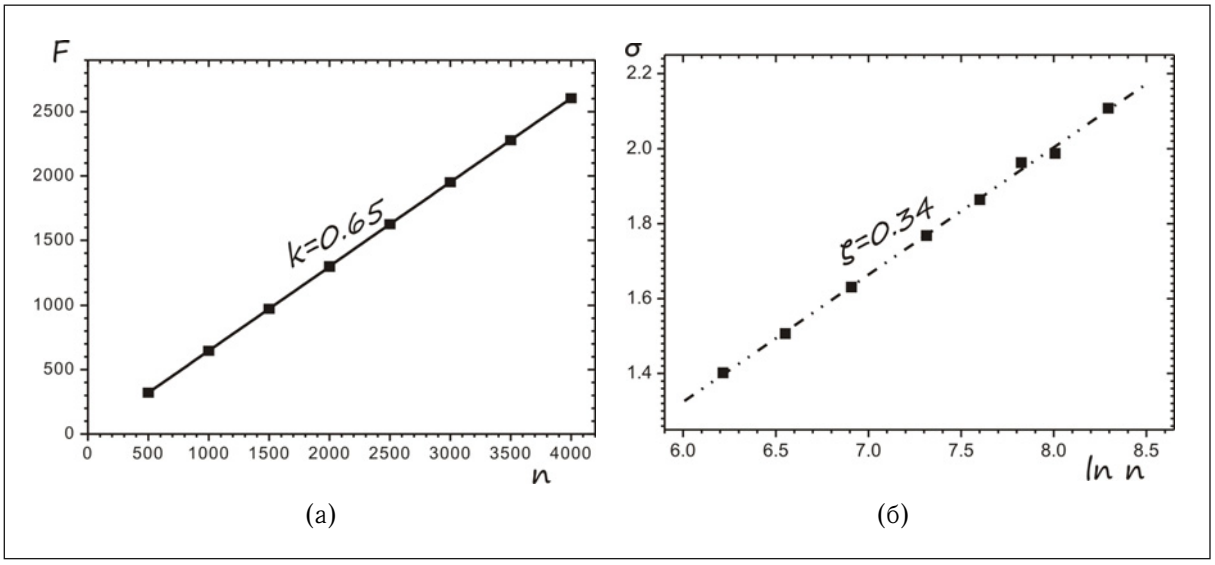
Показатель $1/3$ является типичным для стохастической динамики сильно коррелированных систем и относится к классу универсальности Кардара-Паризи-Занга (Kardar–Parisi–Zhang (KPZ)) [9].

Результаты численного моделирования распределения свободной энергии основного состояния для ансамбля случайных первичных структур РНК представлены на Рис. 4. Угловой коэффициент прямой $k \approx 0.65$ (Рис. 4(а)), что хорошо согласуется с величиной $k = \lim\limits_{n\to\infty} \frac{\langle F_{n,n}\rangle}{n} \to \frac{2}{3}$, вычисленной по формуле (23). Для флуктуации энергии полученный наклон 0.34 (Рис.4(б)) также близок к значению $\frac{1}{3}$. Таким образом, уравнение (23), полученное в приближении бернуллиевского сравнения, удовлетворительно описывает численно наблюдаемую зависимость энергии основного состояния при связывании сополимеров с петлевыми участками от длины случайных цепей.

2. Связывание двух РНК с внутрипетлевым взаимодействием

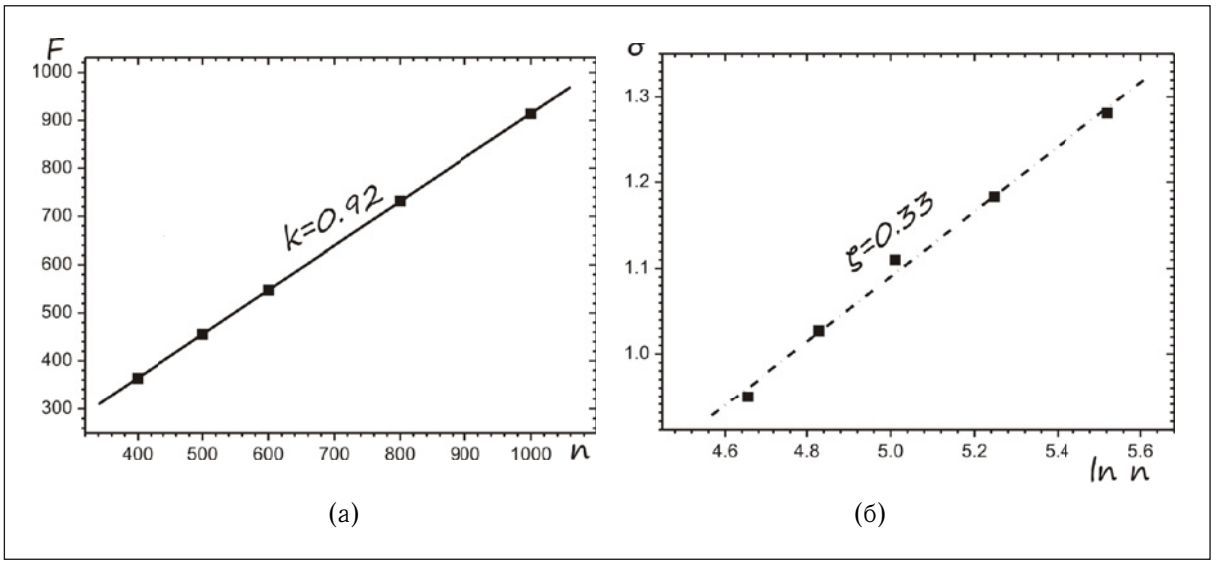
Аналогичный анализ был проведен и для двух последовательностей, образующих структуру с внутрипетлевым взаимодействием и минимальной длиной петли $\ell = 0$. Соответствующие графики зависимости свободной энергии и флуктуации энергии представлены на Рис. 5. Как и для взаимодействия с петлевыми участками, $\langle F_{n,n}\rangle(n) = kn$ при $n \gg 1$ (Рис. 5), но угловой коэффициент прямой $k \approx 0.92$ гораздо выше, что обусловлено взаимодействием нуклеотидов внутри петель. Зависимость флуктуации энергии основного состояния остается такой же (см. Рис. 5(б)).

Оценим аналитически величину коэффициента k в зависимости свободной энергии от длины цепи для внутрипетлевого взаимодействия (Рис. 5). Будем рассматривать комплекс, который образуют две случайные последовательности РНК, как структуру, состоящую из петель различных иерархических уровней, занумерованных индексом i (см. Рис. 6).



(а) (б)

Рис. 4. Взаимодействие РНК с петлевыми участками: зависимость среднего значения свободной энергии основного состояния $F_{n;n}$ (а) и флуктуации энергии σ (б) от длины случайной последовательности n. Усреднение проводилось по ансамблю из 10^5 случайных пар последовательностей для каждого значения длины.



(а) (б)

Рис. 5. Связывание РНК с внутрипетлевым взаимодействием: зависимость энергии основного состояния $F_{n;n}$ (а) и флуктуации свободной энергии σ (б) от длины случайной последовательности n. Усреднение проводилось по ансамблю из 10^5 случайных пар последовательностей для каждого значения длины

Каждую петлю i-ого 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, $m = n$. Это позволяет оценить сверху свободную энергию петли как свободную энергию двух взаимодействующих половинок этой петли. Представление комплекса двух молекул РНК в виде иерархиче-

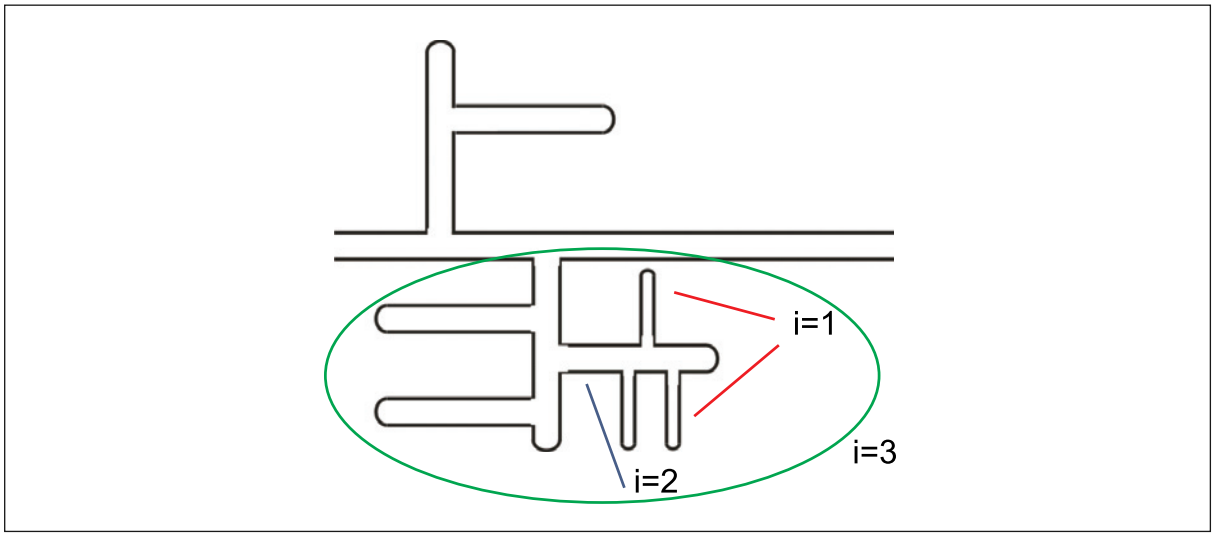


Рис. 6. Иерархическая модель связывания двух полимеров с внутрипетлевым взаимодействием. Петли первого ($i = 1$), второго ($i = 2$) и третьего ($i = 3$) иерархических уровней

ской структуры позволяет использовать идеи ренормализационной группы. А именно, комплексы i-ого иерархического уровня содержат петли, которые будем считать комплексами $(i+1)$-ого уровня (Рис.6) $(i = 1, 2, ...)$.

Формализуя эту идею, будем полагать, что комплекс двух молекул РНК иерархического уровня i – это комплекс двух последовательностей с петлевыми участками, в которых энергия взаимодействующих мономеров перенормирована энергией петель иерархического уровня $(i+1)$. Пользуясь тем, что энергия петель в первом приближении пропорциональна длине (23), представим ее в виде: $F_s^{(i)} \approx k_r^{(i)} s$, где s — длина петли, а k_r — соответствующий i-ому уровню коэффициент связывания. Подставляя в формулу (15) статистические веса петель $g_{i,i+s} = e^{-k_r s/T}$, получим выражение для определения свободной энергии комплекса двух случайных РНК–последовательностей [1]:

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

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

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

[1] Здесь, как и ранее, F имеет смысл свободной энергии с обратным знаком

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-ого иерархического уровня:

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

Будем считать, что m-й и n-й мономеры могут образовать связь, если:
а) участок $[m - l^i_{min}, m-1]$ последовательности S_1 *не имеет* связей с участком $[n - l^i_{min}, n-1]$ подпоследовательности S_2, где l^i_{min} – минимальное количество нуклеотидов, необходимых для формирования петли определенного уровня (если $m < l^i_{min}$ и/или $n < l^i_{min}$, то рассматриваются соответственно участки последовательностей $[1,m]$ и/или по $[1,n]$);
б) $m-1$-й мономер первой последовательности взаимодействует с $n-1$-м мономером второй последовательности, и при замене $(m-1) \to m$, $(n-1) \to n$ выполняется а) (или б)).

В таблице I приведены значения для коэффициента связывания и минимальное количество нуклеотидов в петлях i-ого 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 1 с ростом количества иерархических уровней (т.е. при $n \to \infty$). Логарифмическая зависимость обусловлена экспоненциальным ростом минимального числа мономеров, которые могут образовать петлю, $l^i_{min} = 3l^{i-1}_{min} + 6$ $(i > 2)$ с увеличением номера иерархического уровня i(см. Табл. I).

Таблица 1

Вероятность связывания мономеров в зависимости от числа уровней в иерархической модели взаимодействия двух полимеров

Уровень, i	2	3	4	5	6	7
Минимальная длина петли	2	6	24	78	240	726
Коэффициент связывания	0.851	0.912	0.931	0.937	0.94	0.941

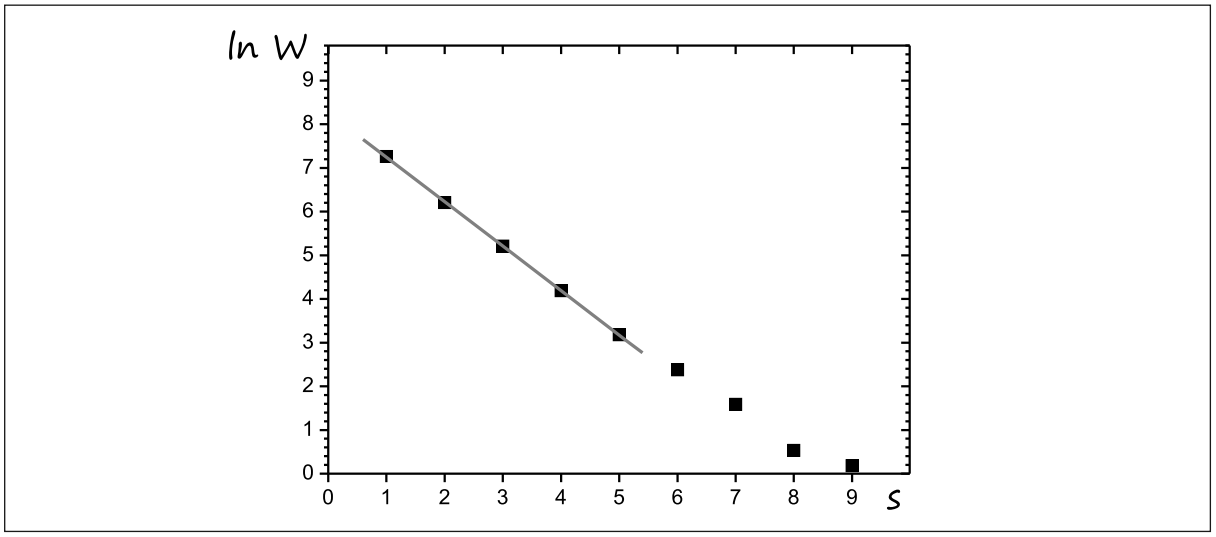


Рис. 7. Распределение длин петель в структуре комплекса с петлевыми участками. Вычисления были выполнены для случайных последовательностей длины $N = 10^4$, результаты усреднялись по набору из 10^5 сополимеров.

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

B. Распределение длин петель в РНК-подобных структурах

1. Связывание двух РНК с петлевыми участками

Было проанализировано распределение длин петель в структуре комплекса с петлевыми участками и внутрипетлевым взаимодействием. На Рис. 7 представлена зависимость $W(s)$ числа петель различной длины s для структуры с петлевыми участками. 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 $k = \frac{\langle F_{n,n} \rangle}{n}$ при $n \gg 1$ можно рассматривать, как вероятность связывания мономера в структуре. Считая, что взаимодействие мономеров независимым, число петель длиной s в структуре двух взаимодействующих сополимеров длиной n можно оценить, как:

$$W(s) = nk^2(1-k)^s. \tag{28}$$

Такое распределение длин петель при $n \gg 1$ удовлетворяет очевидному соотношению $\sum_{s=1}^{n} sW(s) = (1-k)n$. Из Рис. 7 видно, что численные результаты хорошо аппроксимируются в логарифмическом масштабе прямой $y(s) = a - bs$, где с хорошей точностью $a \approx \ln(nk^2)$ и $b \approx \ln(1-k)$ (см. (28)). Таким образом, в связывании сополимеров с петлевыми участками статистика петель выглядит в точности так, как происходит при *независимом* связывании мономеров. Однако стоит отметить, что модель независимого связывания дает хорошие результаты для последовательностей, в которых количество различных сортов мономеров $c \geq 4$. Для двухбуквенных и трехбуквенных алфавитов, взаимодействие сополимеров оказывается коррелированным, и формула (22) плохо описывает энергию оптимальной конфигурации.

2. Связывание двух РНК с внутрипетлевым взаимодействием

Существенно иное поведение имеет статистика петель в комплексах с внутрипетлевым взаимодействием. На Рис.8(а) представлена зависимость числа петель с длиной s по набору из 10^3 пар случайных последовательностей. Отметим особенности наблюдаемого распределения. Во-первых, для данной зависимости характерно степенное поведение. Показатель степенной зависимости для РНК разной длины меняется в интервале $[1.38, 1.5]$. Во-вторых, распределения для РНК с различной длиной n совпадают, что позволяет проводить вычисления для набора коротких последовательностей. В-третьих, при малых n ($n \leq 5$) характерно небольшое число петель с нечетной длиной и большое число петель с четной длиной. Последнее обстоятельство связано с тем, что для структуры комплекса с внутрипетлевым взаимодействием и $\ell = 0$ характерно высокое значение средней энергии на один нуклеотид ($k \approx 0.92$), обусловленное связыванием внутри петель, 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. Наконец, для распределения характерно наличие плато при больших s, что обусловлено эффектом конечного размера (см., например, [10], где построена теория аналогичного эффекта).

Полученные численные распределения можно интерпретировать следующим образом. Поставим каждой вторичной структуре полимера в соответствие одномерное случайное блуждание на (1+1)-мерной решетке, построенное следующим образом (см. Рис. 9). Каждому мономерному звену соответствует один шаг блуждания. Этот шаг

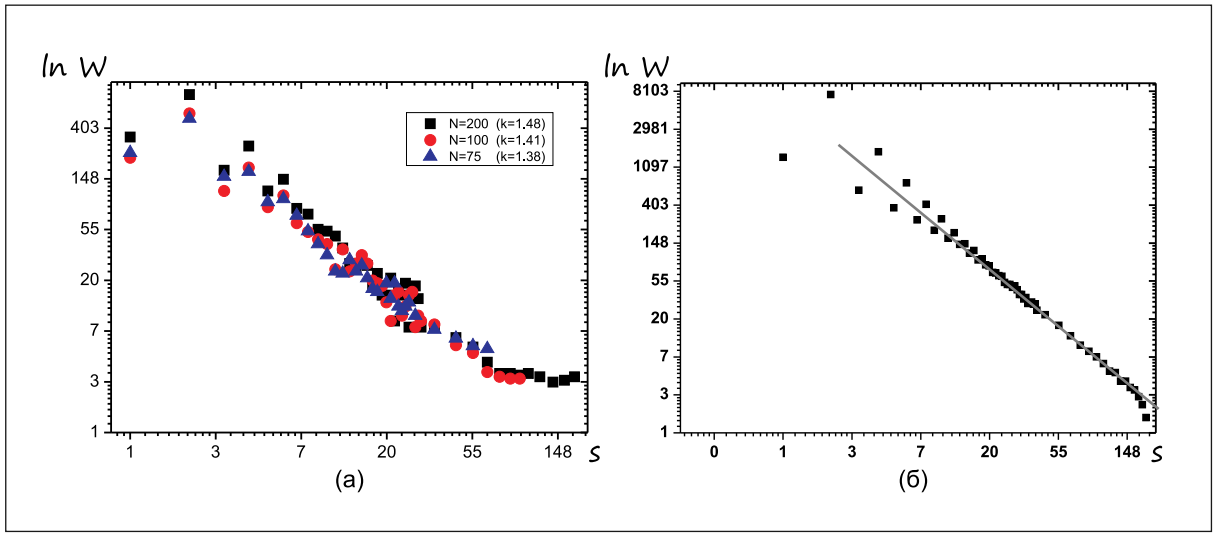


Рис. 8. (а) Распределение длин петель в структуре комплекса с внутрипетлевым взаимодействием. Вычисления проводились для последовательностей с длинами $n = m = 75$; 100 и 200, для каждого n было выполнено 10^3 накоплений, для $n \geq 30$ функция распределения сглаживалась по 10 соседним значениям); (б) Распределение путей Моцкина по длинам (длина пути случайного блуждания 200 шагов, количество накоплений -10^4, для $n \geq 30$ функция распределения сглаживалась по 10 соседним значениям).

направлен направо вверх, если мономер является «началом петли» (т.е. связан с мономером, расположенным *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 соответствие между РНК-подобными вторичными структурами и так называемыми путями Моцкина [11] — 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. Как известно, [12], количество различных путей Моцкина $W_M(s,t)$ длины s с заданным количеством горизонтальных шагов t определяется числами Каталана:

$$W_M(s,t) = \begin{pmatrix} s \\ t \end{pmatrix} C_{(s-t)/2} = \begin{pmatrix} s \\ t \end{pmatrix} \frac{1}{\frac{s-t}{2}+1} \begin{pmatrix} s-t \\ \frac{s-t}{2} \end{pmatrix}, \qquad (29)$$

где $\begin{pmatrix} s \\ t \end{pmatrix}$ – биномиальные коэффициенты, $C_{(s-t)/2}$ – числа Каталана. При $s \gg 1$ (29) имеет асимптотическую зависимость $W_M(s,t) \sim s^{-3/2}$ от длины пути. Было построено распределение длин петель для случайных путей Моцкина с вероятностью диагонального шага вверх или вниз равной $p_M \approx \frac{k}{2} = 0.46$, где $k = 0.92$ — наблюдаемое в численном моделировании значение вероятности образования связи, а вероятность

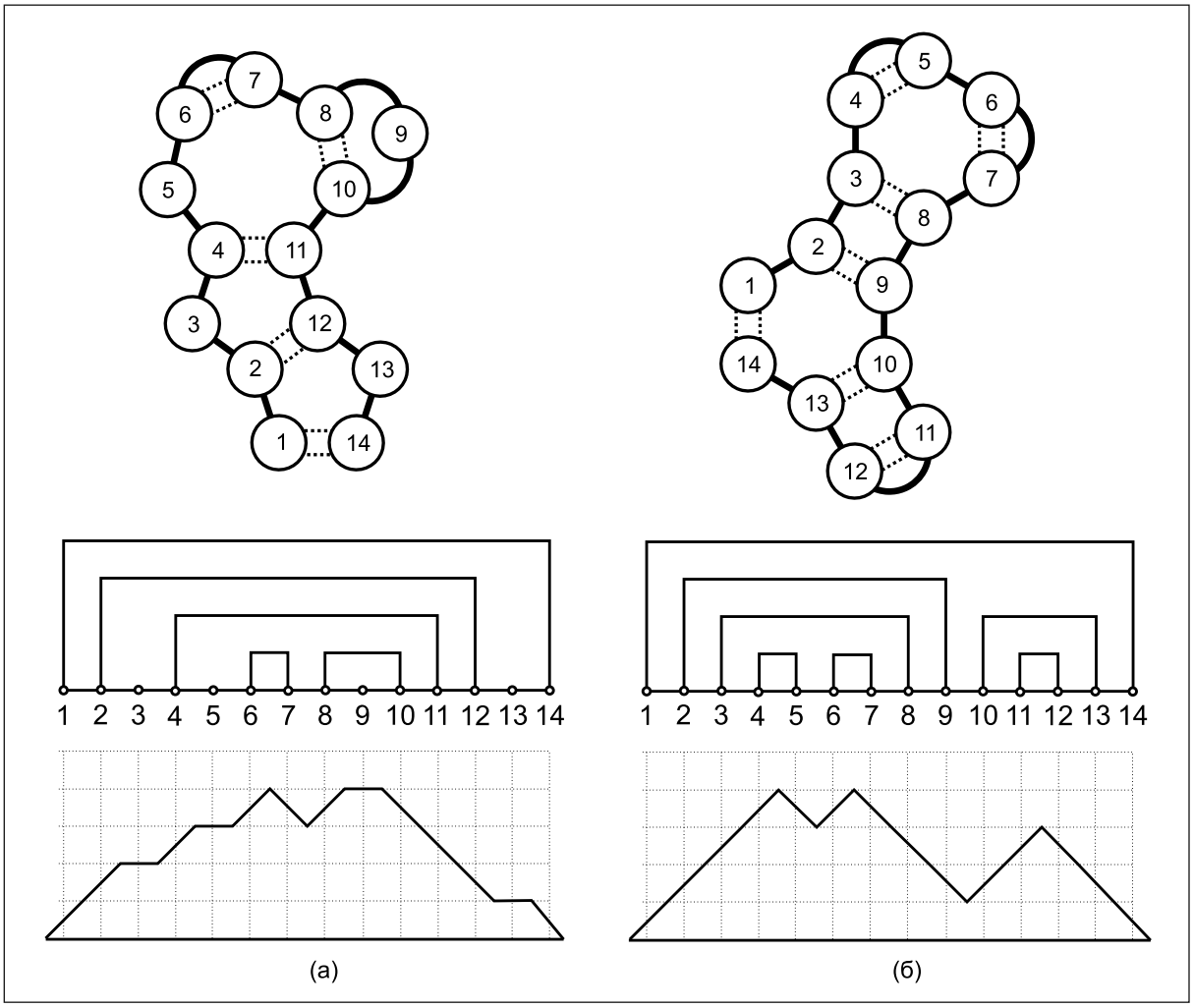


Рис. 8. (а) Распределение длин петель в структуре комплекса с внутрипетлевым взаимодействием. Вычисления проводились для последовательностей с длинами $n = m = 75$; 100 и 200, для каждого n было выполнено 10^3 накоплений, для $n \geq 30$ функция распределения сглаживалась по 10 соседним значениям); (б) Распределение путей Моцкина по длинам (длина пути случайного блуждания 200 шагов, количество накоплений -10^4, для $n \geq 30$ функция распределения сглаживалась по 10 соседним значениям).

горизонтального шага $1 - 2p_M$. Результат приведен на рисунке 8(б). Видно, что зависимость обладает всеми характерными свойствами, наблюдаемыми для распределения длин петель в структуре с внутрипетлевыми взаимодействиями.

Представление структур РНК в виде путей Моцкина, статистика которых известна, позволяет сделать интересное наблюдение. А именно, показать, что для РНК-подобных структур характерно критическая зависимость структуры основного состояния в зависимости от числа различных сортов мономеров, используемых в последовательности.

IV. ТОПОЛОГИЯ РНК-ПОДОБНЫХ МОЛЕКУЛ В ЗАВИСИМОСТИ ОТ АЛФАВИТА СЛУЧАЙНОЙ ПЕРВИЧНОЙ СТРУКТУРЫ

Раздел посвящен исследованию топологии пространственной структуры РНК-подобной молекулы в основном состоянии и ее изменений в зависимости от алфавита, используемого в случайной первичной структуре. А именно, показывается, что существует некоторая критическая точка (критический алфавит) в которой происходит изменение топологии основного состояния РНК-подобной молекулы. В разделе приводятся аналитические и численные оценки критической точки топологического перехода и обсуждается связь данного топологического перехода с температурным фазовым переходом в замороженное состояние.

A. Зависимость свободной энергии РНК-подобных структур от алфавита

Рассмотрим случайную последовательность длиной L и алфавита c, образующую вторичную структуру типа РНК (Рис. 2(а)). Зададимся вопросом о том, к какому пределу стремится доля комплементарных пар в основном состоянии длинной ($L \to \infty$) цепи РНК. Другими словами, интерес представляет удельная (в расчете на одно звено) энергия основного состояния длинной РНК. Вначале, приведем доводы, подтверждающие наличие критического изменения удельной энергии в зависимости от алфавита. Предположим, что существует критическое значение алфавита $c = c_c$ такое, что при $c < c_c$ доля связанных мономерных звеньев стремится к 1, тогда как при $c > c_c$ предельная доля связанных звеньев меньше 1. Убедиться в этом можно следующим образом. Для того чтобы доля связанных звеньев в РНК-подобной структуре, образуемом случайной последовательностью, была равна 1, каждой последовательности из c^L возможных должен соответствовать так называемый путь Дика (т.е. путь Моцкина, в котором нет горизонтальных шагов) (Рис. 9(б)). Количество путей Дика $G(L)$ длины L определяется формулой (29) ($G(L) = W_M(L, 0)$). При $t = 0$ и при $L >> 1$ $G(L)$ имеет асимптотическое выражение

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

Заметим, что один и тот же путь Дика может описывать несколько РНК-подобных структур. Действительно, каждая пара подъем/спуск в пути Дика может быть, независимо от остальных, реализована c разными способами (в случае РНК возможные

варианты — это A-U, U-A, C-G и G-C). Таким образом, число различных первичных структур, для которых существуют полностью связанные вторичные структуры, не превышает

$$W(c,L) = G(L)c^{L/2} \sim \frac{(4c)^{L/2}}{L^{3/2}}. \tag{31}$$

Это оценка сверху, т.к., вообще говоря, одной и той же последовательности может соответствовать несколько различных РНК-подобных структур и, таким образом, несколько путей Дика. Тем не менее, естественно ожидать, что при $c \to c_c$ число таких последовательностей с двумя и более полностью связанными вторичными структурами становится малым. В таком случае, сравнивая (31) с полным числом возможных первичных структур $W_0(c,L) = c^L$, можно записать ($L \gg 1$):

$$\begin{cases} \lim\limits_{L\to\infty} \dfrac{1}{L} \ln W(c,L) > \lim\limits_{L\to\infty} \dfrac{1}{L} \ln W_0(c,L), \text{ для } 2 \le c < c_c \\ \lim\limits_{L\to\infty} \dfrac{1}{L} \ln W(c,L) < \lim\limits_{n\to\infty} \dfrac{1}{L} \ln W_0(c,L), \text{ для } c > c_c. \end{cases} \tag{32}$$

Откуда, $c_c = 4$. Подчеркнем, что несмотря на то, что эта оценка является грубой оценкой сверху, она демонстрирует характерное изменение свойств РНК-структур со случайной первичной структурой.

Таким образом, при $c < c_c$ в пределе $L \to \infty$ практически любой последовательности соответствует полностью связанная вторичная структура, и энергия оптимальной конфигурации на одну пару нуклеотидов стремится к 1, в то время как для случайных цепочек с $c > c_c$ доля последовательностей, допускающих совершенную вторичную структуру, экспоненциально мала. Добавление горизонтальных шагов в пути случайных блужданий приводит к увеличению возможных РНК-подобных конфигураций (29), что позволяет сопоставить каждой случайной последовательности из ансамбля c^L путь в случайном блуждании, соответствующий ее оптимальной вторичной структуре. Однако в этом случае доля связанных звеньев в оптимальной вторичной структуре остается в пределе $L \to \infty$ меньше единицы. Путь Моцкина длиной L, включающий t горизонтальных шагов определяется (согласно (29)) как

$$W_M(L,t) = \frac{L!}{t!(L-t)!} C_{(L-t)}. \tag{33}$$

Для нечетных $(L-t)$ функция $W_M(L,t)$ равна 0. Для четных $(L-t)$ воспользуемся (30) и формулой Стирлинга для оценки асимптотического поведения:

$$\begin{aligned} &\frac{1}{L} \ln W_M(L,a) = -a \ln a - (1-a) \ln \frac{1-a}{4} + o\left(\frac{\ln L}{L}\right) \\ &\frac{\partial}{\partial a} \frac{1}{L} \ln W_M(L,a) \to +\infty, \text{ при } a \to +0, \end{aligned} \tag{34}$$

где введено обозначение $a = \frac{t}{L}$ ($a > 0$). Последнее выражение (34) показывает рост W_M для малых, но конечных a.

Как много различных структур могут иметь один и тот же путь Моцкина (Рис. 9(a))? Как и в случае полностью связанных структур, каждая связанная пара имеет вырожденность c, тогда как каждый несвязанный мономер также может быть выбран c разными способами. Суммарная вырожденность Z имеет вид

$$Z(c, L, a) = c^{(L-aL)/2} c^{aL} = c^{L(1+a)/2}, \tag{35}$$

и является возрастающей функцией a.

А теперь оценим минимальное количество несвязанных мономеров (горизонтальных шагов в пути Моцкина), $a(c) = 1 - f(c)$, в основном состоянии при $c > 4$. Наибольшее количество структур, имеющих в основном состоянии долю несвязанных мономеров меньше или равной a определяется выражением:

$$W(c, L, a) = \sum_{j=0}^{aL} Z(c, L, j/L) W_M(j, L). \tag{36}$$

Для $c > 4$ и $a = 0$ эта сумма меньше $W_0(c, L) = c^L$, и растет с увеличением a так, что при некотором $\bar{a}$ величины $W(c, L, a)$ и $W_0(c, n)$ сравниваются. Для $L \gg 1$ сумму (36) можно оценить методом перевала. Введем обозначение

$$\Delta w(a, c) = \lim_{L\to\infty} \frac{1}{L} \ln \frac{W(c, L, a)}{W_0(c, L)}.$$

Тогда

$$\Delta w(a, c) = \begin{cases} -(1-a) \ln \dfrac{\sqrt{c}(1-a)}{2} - a \ln a; & a < a_{\mathrm{m}} \\ \ln\left(1 + \dfrac{\sqrt{c}}{2}\right) > 0; & a > a_{\mathrm{m}}, \end{cases} \tag{37}$$

где $a_{\mathrm{m}} = \frac{\sqrt{c}}{2+\sqrt{c}}$. Для $a < a_{\mathrm{m}}$ сумма в (37) определяется вкладом от верхней границы, тогда как для $a > a_{\mathrm{m}}$ максимум достигается в точке a_{m} и, таким образом, не зависит от верхнего предела суммирования. Величина $\bar{a}(c)$ определяется из уравнения $\Delta w(a, c) = 0$. На Рис. 10 представлена функция $f(c) = 1 - a(c)$. Напомним, что данная оценка является верхней границей, так как не учитывает корреляции между оптимальными конфигурациями.

Оценка (37) сделана в предположении так называемого среднего поля: связывание на каждой паре подъем/спуск Рис. 9 происходит независимо с вероятностью $1/c$ и все пути случайных блужданий считаются статистически независимыми. В разделе IV C

приводится более точная оценка критического алфавита, учитывающая корреляции между конфигурациями.

Результаты численного моделирования для РНК-подобных структур со случайной последовательностью звеньев различного алфавита представлены на Рис. 10. Для простоты предполагается, что комплементарные связи образуются согласно правилу A–A, т.е., только одинаковые мономеры могут комплементарно связываться. Напомним, что в реальных молекулах РНК действуют перекрестные правила комплементарности. Однако, анализ случайных последовательностей показал, что правила комплементарности незначительно влияют на свойства РНК-подобных структур. Тогда как, комплементарное связывания типа A–A позволяет исследовать цепочки не только с четным алфавитом как в случае перекрестного связывания, но и с нечетным. Соответственно, для каждого значения $c = 3, 4, ..7$ были построены зависимости удельной энергии $f =< F > /L$ от длины случайной первичной структуры. Как видно (Рис. 10(а)), удельная энергия при $L \to \infty$ действительно стремится к некоторому усредненному значению f_∞, которое является только функцией от c (Рис. 10(б)). Результаты численного моделирования существенно расходятся с аналитической оценкой (Рис. 10(б)). Зависимость, полученная в численном моделировании имеет критическую точку топологического перехода $c_c = 2$, которая является очевидной оценкой снизу. Действительно, рассмотрим произвольную двухбуквенную последовательность, например, $ABAABBBAAABBABAAAB$ и будем последовательно находить комплементарые пары (в предположении A–A связывания) следующим образом. Ближайшие соседи по цепи одного сорта образуют комплементарную пару, и далее, вычеркиваются из последовательности. Легко видеть, что такая процедура приводит к формированию РНК-подобной структуры. Рассматриваемая цепочка после первой итерации будет выглядеть: $ABBAABAB$, последующее вычеркивание приведет к $ABAB$. Понятно, что данная процедура для любой случайной двухбуквенной последовательности приведет к тому, что, либо в остатке будет $ABAB$, либо последовательность будет полностью вычеркнута. Второй вариант означает, что все мономеры участвуют в формировании вторичной структуры, удельная энергия которой $f = 1$. В случае остатка $ABAB$, данный участок цепочки образует конфигурацию с двумя пропусками, но, в термодинамическом пределе, $f_\infty = 1$. Если структура образуется согласно перекрестным правилам комплементарности, то остатком будет $|L_1 - L_2|$ букв одного сорта, где L_1 и L_2— количество мономеров A и B соответственно. В случайной последовательности $|L_1 - L_2| \sim 1/\sqrt{L}$. Таким образом, для случайных последовательностей с алфавитом $c = 2$ можно записать:

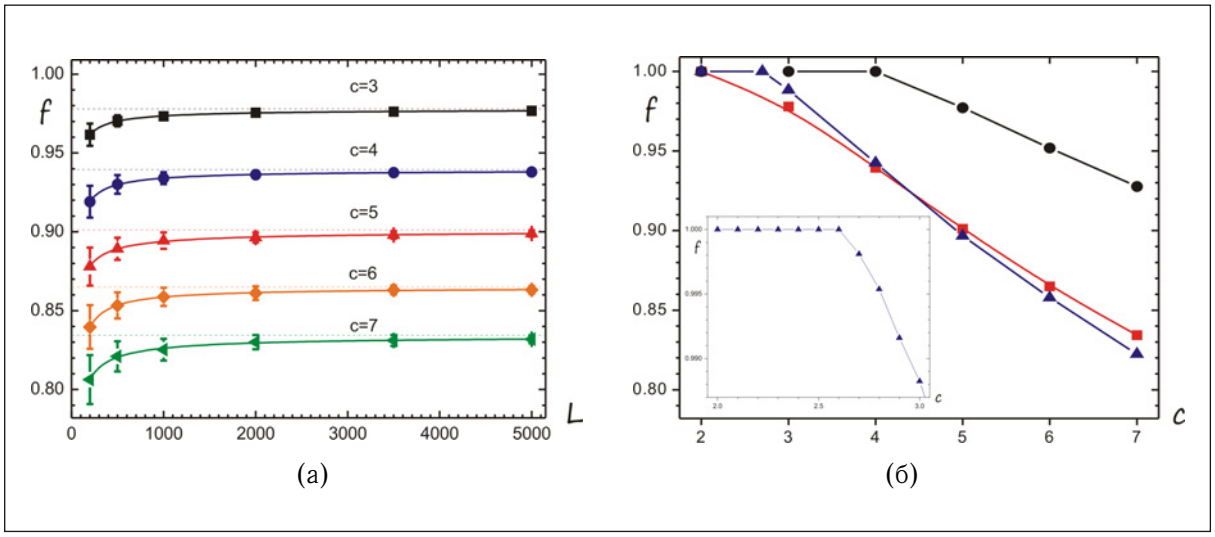


Рис. 10. (а) Зависимость удельной энергии f от длины случайной последовательности с заданным алфавитом c; (б) зависимость предельного значения энергии f_∞ от алфавита для последовательностей с дискретным алфавитом (красным), в модели Бернулли (синим), и верхняя оценка энергии (черным) в модели независимого связывания. Дополнительный график: зависимость предельного значения энергии от алфавита в модели Бернулли демонстрирует, что критический алфавит является нецелым и принадлежит интервалу $2 < c_c < 3$.

$$f(L) = \begin{cases} 1 - \dfrac{const}{L}; & \text{для A-A связывания} \\ 1 - \dfrac{const}{\sqrt{L}}; & \text{для A-B связывания.} \end{cases} \tag{38}$$

Таким образом, аналитическое рассмотрение буквенных последовательностей позволило говорить о том, что критическое значение алфавита $2 \leq c_c \leq 4$. Более строгое рассмотрение [13] показало, что критическое значение алфавита лежит в интервале: $2 < c_c < 3$.

Подводя итог, подчеркнем еще раз, что при изменении алфавита, используемого в первичной структуре случайной последовательности, существует переход от полностью связанной РНК-подобной структуры до структуры с конечной долей несвязанных мономеров. Такой переход в работе называется топологическим. Критическая точка топологического перехода принадлежит интервалу $(2 < c_c < 3)$, т.е. эффективно является *нецелым*.

Как можно трактовать нецелый алфавит в случайных последовательностях типа РНК? Далее, попробуем ответить на этот вопрос.

В. Топологический переход в модели Бернулли

Модель случайной последовательности с эффективно нецелым алфавитом может быть построена следующим образом. Будем считать, что матрица контактов η' в урав-

нении (20) является случайной: вероятность того, что $\eta'_{i,j} = 1$, равна p, а вероятность $\eta'_{i,j} = 0$ равна соответственно $1-p$. То есть теперь случайная последовательность характеризуется не первичной структурой — последовательностью мономеров из c различных типов, как это было раньше, а некой матрицей контактов, (i, j)-элемент которой разрешает или запрещает образование комплементарной пары между i и j мономером цепи. Мономеры цепи в данной модели не различаются по сортам и, в целом, любой мономер может образовать связь с любым другим в цепи, однако, в среднем, вероятность такого события равна p. Каждой последовательности в рассматриваемой модели можно сопоставить граф Эрдёша–Реньи, изображающего все возможные контакты между L мономерами. Основное отличие данной модели от дискретных буквенных последовательностей — нарушение свойства транзитивности. Если 1-й мономер может образовать связь со 2-м, а 2-й с 3-м, отсюда, вообще говоря, не следует (как это было для последовательностей с дискретным алфавитом), что 1-й мономер может связаться с 3-м. Однако, как, например, уже упоминалось, подобная модель бернуллиевского сравнения в задачах выравнивания случайных последовательностей является хорошей аппроксимацией. Вероятности p случайной матрицы контактов соответствует алфавит, равный:

$$c_{eff} = \frac{1}{p}. \tag{39}$$

Таким образом, оказывается возможным генерировать случайную последовательность с *любым* нецелым значением алфавита c. На Рис. 10(б) приведена зависимость удельной энергии f_∞ в термодинамическом пределе от алфавита c (39), полученная в численном моделировании. Во-первых, отметим, что значения f_∞ для бернуллиевского алфавита не более, чем на 1% отличается от соответствующих величин для случайных последовательностей с дискретным алфавитом, что оправдывает применимость данной модели. Случайный бернуллиевский полимер характеризуется критической вероятностью p_c. Для $p > p_c$, в термодинамическом пределе, $f_\infty = 1$ (так называемая «полочка» на зависимости удельной энергии (см. дополнительный график на Рис. 10(б)), что соответствует полностью связанной вторичной структуре, тогда как для $p < p_c$, даже в пределе бесконечной длины, основное состояние характеризуется $O(L)$ количеством несвязанных мономеров. Критическое значение вероятности согласно (39) соответствует критическому значению алфавита $c_c = 2.6$. Таким образом, модель Бернулли позволяет численно получить точку перехода.

Для более точной оценки критической точки топологического перехода были проведены следующие численные эксперименты. Рассмотрим ансамбль, состоящий из

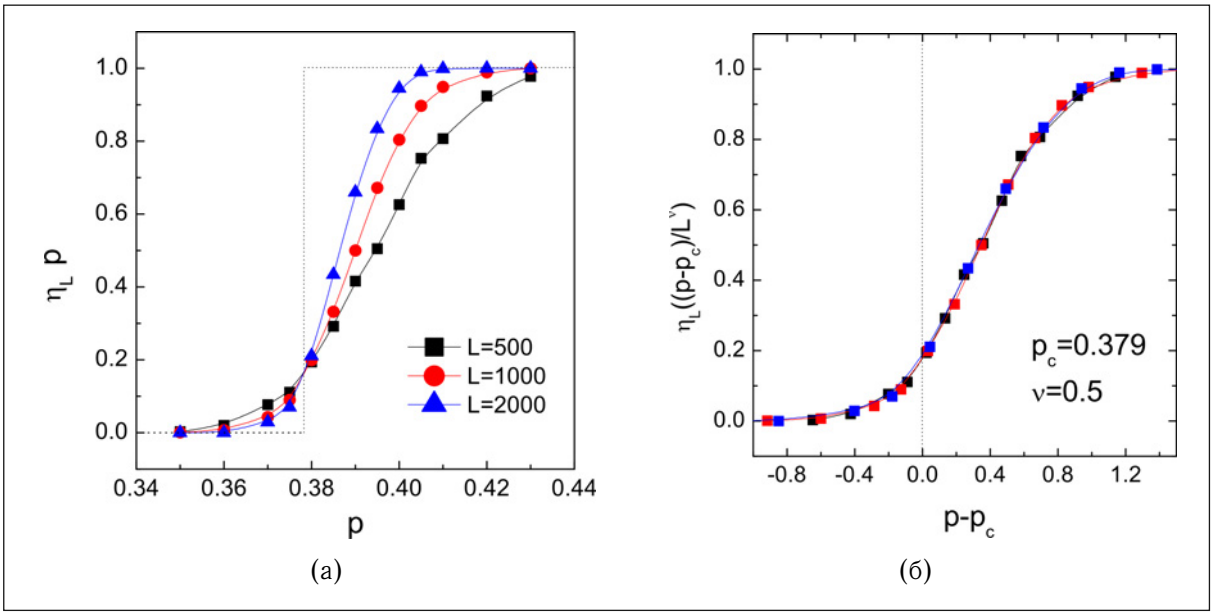


Рис. 11. Зависимость доли полностью связанных РНК-подобных структур в ансамбле случайных первичных структур различной длины (а) от параметра p модели Бернулли; скейлинг-анализ полученных зависимостей (б). Для каждого значения p и L было выполнено 10^5 накоплений.

N ($N = 10^5$) случайных бернуллиевских полимеров длиной L и подсчитаем количество последовательностей с полностью связанной вторичной структурой N_c. Доля полностью связанных структур в таком ансамбле $\eta_L = N_c/N$ есть функция p (см. Рис. 11). Естественно ожидать, что в пределе $L \to \infty$ (Рис. 11(а)), функция $\eta_L(p)$ вырождается в ступенчатую функцию. Скейлинг-анализ полученных зависимостей $\eta_L(p)$ обеспечивает критическое значение $p_c = 0.37$, что соответствует алфавиту:

$$c_c \approx 2.67.$$

Можно провести аналогию между данным топологическим переходом и переходом, наблюдаемым в теории перколяции [14]. В перколяционной теории задача формулируется следующим образом (одна из возможных формулировок). Рассмотрим протекание жидкости через пористую среду, причем пористую среду будем моделировать дискретной решеткой (сетью) — набором сайтов, между которыми есть связи — каналы. Жидкость протекает по этим каналам, которые могут быть открыты или закрыты с вероятностью p и $1-p$ соответственно. Существует пороговое значение вероятности p_{th} выше которой, протекание через данную среду возможно, т.е. существует связанный кластер на решетке, а ниже которой, построить связанный кластер невозможно. Переход между этими двумя состояниями в теории перколяции называют геометрическим фазовым переходом и относят к переходам второго рода [14].

Таким образом, можно предполагать, что топологический переход между полностью связанной РНК-подобной структурой и структурой с пропусками является фазовым переходом второго рода. В пользу этого предположения также свидетельствует непрерывное изменение вырожденности основного состояния (числа полностью связанных РНК-подобных структур) от вероятности p.

Был также выполнен анализ областей алфавита, лежащего выше и ниже критической точки топологического перехода в модели Бернулли. Во-первых, области характеризуются различной зависимостью от длины случайной последовательности: L:

$$\begin{cases} f(L) \sim 1 - C_1 e^{-L/\ell} & \text{для } p > p_c \\ f(L) \sim f_\infty - C_2 L^{-\alpha} & \text{для } p < p_c, \end{cases} \tag{40}$$

где C_1 и C_2 — некоторые константы. Для допереходной фазы ($p > p_c$) характерно экспоненциальное приближение к предельному значению удельной энергии ($f_\infty = 1$), тогда как в области больших алфавитов ($p < p_c$) энергия приближается к своему предельному значению степенным образом (Рис. 12). Показатель степени α в (40) находится в пределах $[0.75, 1]$ (сравните с (Рис. 10(а))). В допереходной области случайная последовательность из алфавита p может быть охарактеризована некоторой релаксационной длиной ℓ, указывающей на характерный масштаб длин, на котором энергия основного состояния сходится к своему предельному значению $f_\infty = 1$. Ясно, что зависимость релаксационной длины ℓ от вероятности p имеет вертикальную асимптоту в точке $p = p_c$. Естественно ожидать, что асимптотическое поведение $f(L)$ зависит от выбранной модели случайного полимера, в частности от правил комплементарности — см. (38).

Области отличаются также зависимостями флуктуаций свободной энергии от длины случайной последовательности L. Допереходная область характеризуется быстрым (экспоненциальным) падением флуктуаций с ростом L. Тогда как для $p < p_c$, характерен степенной рост флуктуаций с увеличением длины последовательности L (см. Рис. 5(б)).

C. Аналитическая оценка критической точки топологического перехода в модели Бернулли

1. Метод среднего поля

Для простоты переформулируем задачу в терминах планарных диаграмм. Рассмотрим граф, вершины которого (мономеры вдоль цепочки) перенумерованы, а матрица

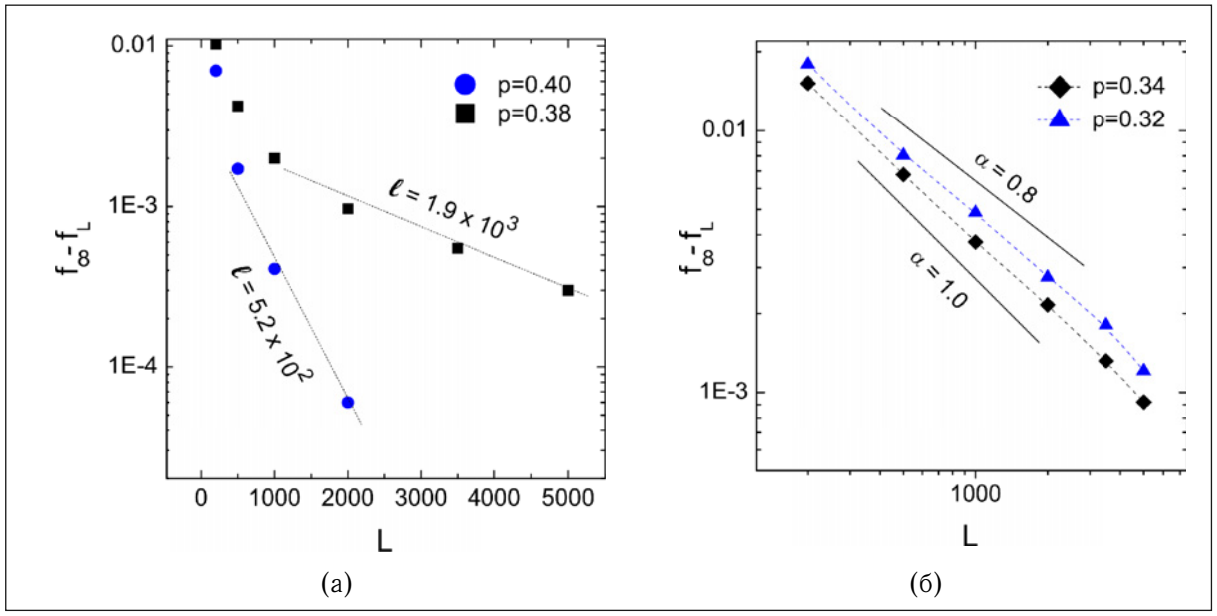


Рис. 12. Асимптотическое поведение удельной энергии $f(L)$ до (а) и после (б) топологического перехода. Зависимость $(f_\infty - f(L))$ в логарифмическом масштабе (а) и двойном логарифмическом масштабе (б) (см. (40)).

контактов V — матрица инцидентности графа. Задача о полностью связанной РНК-подобной структуре на данном графе сводится к вопросу о том, как выбрать среди разрешенных контактов $L/2$ связей, которые обеспечивают планарную структуру на заданном случайном графе, т.е. все вершины входят в конфигурацию ровно один раз и любые пары связей $(i_1, j_1$и (i_2, j_2) удовлетворяют соотношению [15]:

$$(j_1 - i_1)(j_2 - i_1)(j_1 - i_2)(j_2 - i_2) > 0. \tag{41}$$

Другими словами, как разместить $L/2$ непересекающихся арок, принимая во внимание ограничения, накладываемые матрицей V. В модели Бернулли каждый элемент V_{ij} равен 1 либо 0 с соответствующими вероятностями p и $1-p$, кроме того, матрица контактов — симметричная с нулевыми диагональными элементами:

$$P(V_{ij}) = \left([p\delta(V_{ij}-1) + (1-p)\delta(V_{ij})]\,\theta(i-j) + \delta(V_{ji}-V_{ij})\theta(j-i)\right)(\delta_{ij}-1). \tag{42}$$

Здесь $\delta(x)$ и $\theta(x)$ — дельта-функция Дирака и функция Хевисайда, соответственно. Для $p = 1$ (когда все элементы V_{ij} равны 1), количество всех возможных арочных структур, удовлетворяющих (41) определяется числами Каталана (см. (33))

$$\# = C_{L/2} = \frac{L!}{(\frac{L}{2})!(\frac{L}{2}-1)!}. \tag{43}$$

Когда $p \neq 1$, некоторые из конфигураций $\#$ запрещены матрицей контактов V. Введем обозначение p_1 — вероятность, того, что одна выбранная из $\#$ конфигурация разрешена.

Очевидно, что

$$p_1 = p^{L/2}. \tag{44}$$

Аналогично, определим p_k как вероятность, что k диаграмм из $\#$ разрешены, для $k = 2$, например

$$p_2 = p^{L/2} p^{L/2} p^{-n_{1\cap 2}} = p^L p^{-\kappa_2 L}, \tag{45}$$

где $n_{1\cap 2} \equiv \kappa_2 L$ равно количеству общих арок для двух случайно выбранных планарных диаграмм, усредненному по ансамблю $\#$. Для p_3 можно записать:

$$p_3 = (p^{L/2})^3 p^{-n_{1\cap 2\cap 3}} = p^{3L/2} p^{-C_3^2 \kappa_2 L} p^{\kappa_3 L}. \tag{46}$$

Величины κ_k могут быть вычислены с любой точностью. К примеру, κ_2 лежит строго в интервале $[1/15, 1/14.8]$. Вероятность иметь по крайней мере одну планарную конфигурацию для данной заполненности p матрицы V (42) определяется как:

$$\mathcal{P} = \# p_1 - \frac{\#(\# - 1)}{2} p_2 + C_{\#}^3 p_3 + \ldots \tag{47}$$

Предполагая, что все диаграммы в ансамбле $\#$ независимы, т.е., $p_k = p_1^k$, для $\mathcal{P}$ из (47) можно записать:

$$\mathcal{P} = 1 - (1 - p_1)^{\#} = 1 - \exp(-p_1 \#). \tag{48}$$

В пределе больших L, величина $\mathcal{P}$ равна либо нулю, либо единице, в зависимости от соотношения между $\#$ и p_1. Используя (44), для критического значения вероятности можно записать уравнение:

$$\lim_{L\to\infty} p_c \, [\#]^{2/L} = 1. \tag{49}$$

Условие (49) можно интерпретировать как то, что переход наблюдается в точке, при которой плотность единиц в матрице контактов V такая, что *в среднем* разрешена только одна планарная конфигурация. Вспоминая, асимптотику чисел Каталана (45), для критического значения вероятности получим $p_c = 1/4$, что совпадает с верхней оценкой $c_c = 4$ из (IV B).

2. Комбинаторная оценка

Предположение о независимости планарных конфигураций соответствует так называемому приближению среднего поля. Естественным следующим шагом является введение ненулевых корреляций между конфигурациями: $\kappa_k \neq 0$. Чтобы учесть корреляции между различными планарными диаграммами, поступим следующим образом. Перепишем (49) как:

$$\lim_{L\to\infty} \xi(p_c)\,[\#]^{2/L} = 1, \tag{50}$$

где $\xi(p)$ — некоторая функция, учитывающая корреляции между планарными диаграммами. Основная идея дальнейшего рассмотрения следующая: арки разной длины встречаются в оптимальной планарной конфигурации с различной вероятностью. Рассмотрим полностью связанную планарную конфигурацию, состоящую из $\mathcal{N} = \frac{L}{2}$ арок, соединяющих L точек. Возвращаясь к представлению планарных диаграмм через пути Дика (см.Рис. 9), можно увидеть, что арка между i-ой и j-ой точками возможна, только если i-й и j-й шаг имеют одну и ту же пространственную координату y. Тогда можно определить вероятность арки между i-ой и j-ой точками как:

$$P(i,j) = \frac{1 \times C_{(j-i-1)/2} \times 1}{2^{j-i+1}}. \tag{51}$$

В знаменателе правой части (51) стоит суммарное число возможных шагов вверх/вниз на длине $(j-i+1)$, в числителе — "1" соответствуют выбору шага вверх и вниз на позициях i и j соответственно; число Каталана $C_{(j-i-1)/2}$ описывает все возможные конфигурации петли между парой (i,j) (так как i-й и j-й шаги находятся на одной высоте, петля между ними должна быть тоже путем Дика). Вероятности $P(i,j)$ зависят только от длины арки $(j-i)$ и не равны нулю только для арок нечетной длины, т.е., $P(i,i+1) = \frac{1}{4}$, $P(i,i+3) = \frac{1}{16}$, $P(i,i+5) = \frac{2}{32}$, т.д.. Если просуммировать $P(i,j)$ по всем возможным арочным длинам, то результатом будет $\sum_{k=1}^{L-1} P(i,i+k) = \frac{1}{2}$ — вероятность того, что в i-ой позиции находится левая граница арки (шаг вверх).

Отметим, что доля коротких арок чрезвычайно высока. Действительно, вероятность, в типичной арочной конфигурации иметь арку длиной $\ell = 1$ равна $\frac{1}{4}$, арку длиной $\ell = 3$, уже $\frac{1}{16}$, и т.д.. С другой стороны, количество всех возможных кратчайших арок — $(L-1)$. Поэтому, в типичной конфигурации $\frac{1}{4}$ среди них должны быть «разрешены». Естественно, что веса таких коротких арок в бернуллиевской модели (элементы $V_{i,i+1}$ матрицы контактов) выше, чем длинных арок.

Принимая во внимание эту выделенность коротких арок, оценим функцию $\xi(p)$ в (50). Вместо независимого выбора набора арок, теперь предположим, что построение типичной арочной конфигурации происходит следующим образом:

1. выбор $\frac{L}{4}$ непересекающихся коротких арок ($\ell = 1$) из $(L-1)$ возможных
2. выбор остальных $\mathcal{N} - \frac{L}{4} = \frac{L}{4}$ из длинных ($\ell > 2$) арок

Так как общее число длинных арок порядка $L^2 \gg \frac{L}{4}$, будем считать, что длинные арки

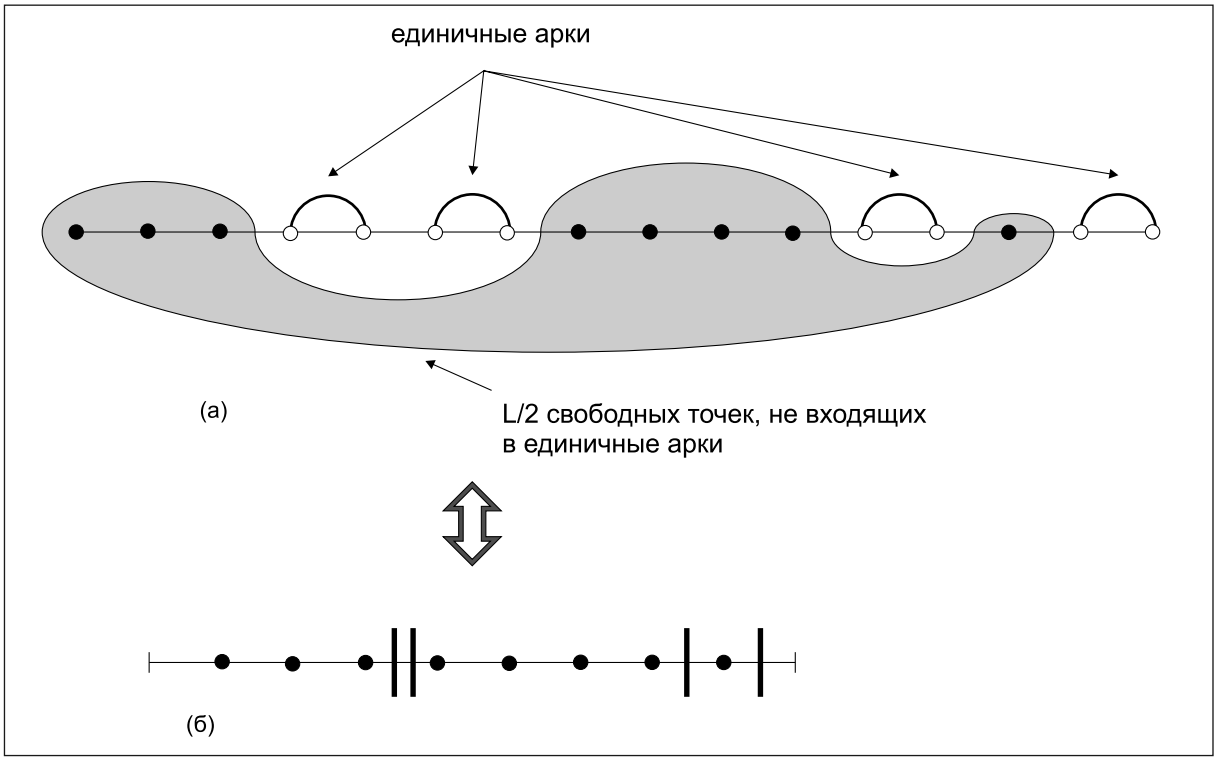


Рис. 13 Пояснение к вычислению $\mathcal{Z}(p)$: (а) Выбор $\frac{L}{4}$ единичных арок на L вершинах случайного графа ($\frac{L}{2}$ свободных вершин) аналогичен комбинаторной задаче о расположении $\frac{L}{2}$ точек по $\frac{L}{4}-1$ ящикам (б).

выбираются независимо друг от друга с вероятностью p. И, таким образом, вклад от длинных арок в функцию $\xi(p)$ равен $p^{L/4}$.

Иная ситуация при выборе кратчайших арок длиной “1”. Для бернуллиевского полимера с матрицей контактов V только pL единичных арок разрешены. Таким образом, выбор коротких арок для оптимальной конфигурации без пропусков оказывается сильно ограниченным. Вероятность выбрать $L/4$ непересекающихся арок из pL разрешенных можно оценить следующим образом. Определим сначала число способов $\mathcal{Z}$ выбора $\frac{L}{4}$ непересекающихся единичных арок из всех $(L-1)$ возможных (Рис. 13). Единичные арки можно рассматривать как стенки ящиков, тогда задачу можно переформулировать следующим образом. Будем интересоваться количеством способов, которыми можно заполнить $(\frac{L}{4}-1)$ ящика $L/2$ свободными точками (шарами). Результат известен из комбинаторики и $\mathcal{Z}=C_{L/4}^{3L/4-1}$, где C_m^n — число сочетаний m по n.

Можно считать, что среди них $p\left(3\frac{L}{4}-1\right)$ арок разрешены первичной структурой полимера (матрицей контактов V) и величина $C_{L/4}^{p(3L/4-1)}$ описывает вес коротких арок в полностью связанной РНК-подобной структуре случайного полимера. Учет корреляций

между планарными конфигурациями на уровне единичных дуг приводит к следующими выражению для $\xi(p)$ (50):

$$\xi(p)^{L/2} = p^{L/4}\, C_{L/4}^{p(3L/4-1)} \left[C_{L/4}^{3L/4-1}\right]^{-1}. \tag{52}$$

В пределе $L \to \infty$, после упрощений, получим:

$$\ln \xi(p) = \frac{1}{2}\ln p + \frac{3p}{2}\ln\frac{3p}{2} - \frac{3p-1}{2}\ln\frac{3p-1}{2} - \frac{3}{2}\ln\frac{3}{2}. \tag{53}$$

Подставляя этот результат в (50):

$$\begin{aligned} &\ln \xi(p_c) = -\ln 4; \\ &p_c \approx 0.35 \ (c_c = 2.87). \end{aligned} \tag{54}$$

3. Матричный подход

Еще один подход оценки критического алфавита основан на матричном описании вторичной структуры РНК. Статистическую сумму $Z_L(N,V)$ случайного полимера можно представить через случайные эрмитовы матрицы ϕ как [16]:

$$Z_L(N,V) = \frac{\int d\phi_1 ... d\phi_L e^{-S_0} \dfrac{1}{N}\mathrm{tr}\,(\phi_1 ... \phi_L)}{\int d\phi_1 ... d\phi_L e^{-S_0}} \equiv \langle \phi_1 ... \phi_L \rangle_{S_0}, \tag{55}$$

где

$$S_0 \equiv S_0\{V, \phi_1, \ldots, \phi_L\} = \frac{N}{2}\sum_{i,j}(V^{-1})_{ij}\mathrm{tr}(\phi_i\phi_j). \tag{56}$$

В отсутствии замороженного беспорядка, т.е., если $V_{ij} \equiv 1$, задача (55) может быть решена точно. В частности, множитель $a_{L/2,0}$ перед $\upsilon^{L/2}$, описывающий планарные конфигурации с $L/2$ арками, т.е. полностью связанные структуры, вносит наибольший вклад в общую статистическую сумму полимера и определяется числами Каталана:

$$\lim_{N\to\infty} Z_L(N;V) = C_{L/2} \sim \frac{4^{L/2}}{(L/2)^{3/2}\sqrt{\pi}}. \tag{57}$$

Как и ранее будем вычислять функцию $\xi(p)$ в (50) усредняя статистическую сумму $Z_L(N,V)$ по распределению (42). Для этого выполним стандартное преобразование Хаббарда-Стратоновича и будем интегрировать по V с весом (42):

$$\begin{aligned} &\int dV \ P(V)\, Z_L(N,V) = \\ &const \int \prod_{k=1}^{L} d\phi_k \frac{1}{N}\mathrm{tr}\,(\phi_1 \ldots \phi_L) \int \prod_{m=1}^{L} dh_m e^{iN\sum_i \mathrm{tr}(h_i\phi_i)} e^{\mathcal{S}}, \end{aligned} \tag{58}$$

где $\mathcal{S} = \mathcal{S}_0 + U$, и

$$\mathcal{S}_0 = -\frac{pN}{2}\sum_{ij}\mathrm{tr}(h_i h_j), \tag{59}$$

$$\begin{aligned} U = \; & \frac{p(1-p)N^2}{8}\sum_{ij}[\mathrm{tr}(h_i h_j)]^2 \\ & -\frac{p(1-p)(1-2p)N^3}{48}\sum_{ij}[\mathrm{tr}(h_i h_j)]^3 + \ldots \end{aligned} \tag{60}$$

Величина $\mathcal{S}_0$ соответствует единичной матрице контактов с дополнительным фактором p. Учет только этого слагаемого, после обратного преобразования Хаббарда-Стратоновича, приводит к $\xi(p) = p$, и оценке $p_c = \frac{1}{4}$, совпадающей с оценкой в предположении среднего поля. Действие U (60) сдвигает значение p_c в сторону меньших значений. Но, так как U содержит бесконечное число слагаемых (60), теория возмущений в данном случае неприменима. В этой связи было предложено следующее приближение: все поля $\{h_i\}_{i=1,\ldots,L}$ в (60) эквивалентны, поэтому можно считать, что в среднем, $\langle N\mathrm{tr}(h_i h_j)\rangle_{\mathcal{S}_0} \equiv T$ не зависит от (i,j). В рамках данного средне-полевого приближения можно сделать замену $e^{\mathcal{S}} = e^{\mathcal{S}_0}e^{\langle U\rangle}$, где:

$$\begin{aligned} \langle U\rangle = \; & \frac{p(1-p)N}{8}T\sum_{ij}\mathrm{tr}(h_i h_j) \\ & -\frac{p(1-p)(1-2p)N}{48}T^2\sum_{ij}\mathrm{tr}(h_i h_j) + \ldots \end{aligned} \tag{61}$$

Упрощение выражения (61) приводит к следующему уравнению на пропагатор T:

$$\frac{1}{T} = \frac{-2}{T}\log\left[1 - p + p\exp\left(-\frac{T}{2}\right)\right]. \tag{62}$$

Выражение (62) дает $T = -2\log\left[1 - \frac{1-1/\sqrt{e}}{p}\right]$, и окончательно можно написать:

$$\mathcal{S} = -\frac{\xi(p)N}{2}\sum_{ij}\mathrm{tr}(h_i h_j), \tag{63}$$

где

$$\xi(p) = \frac{1}{T} = \left(-2\log\left[1 - \frac{1-1/\sqrt{e}}{p}\right]\right)^{-1}. \tag{64}$$

Подстановка (64) в (50) приводит к оценке критического алфавита $p_c^* = 0.4551$. Большая расходимость полученного результата с численным $p_c = 0.37$ означает, что предложенного приближения недостаточно для описания топологического перехода.

D. Переход случайной РНК в замороженное состояние, ограниченный топологическим переходом

Рассмотрим как данный топологический переход ограничивает фазовый переход в замороженное состояние. С пионерских работ Бундшу и Хва [17, 18] принято считать, что в этой системе имеет место фазовый переход в «замороженное» состояние при низких температурах. Основываясь на репличном анализе, Лассиг и Визе, [19] и Давид и Визе [20] сформулировали задачу о переходе в терминах теории поля. Ниже приводятся доводы Бундшу и Хва, доказывающие существование фазового перехода и обсуждаются характерные свойства разных фаз.

В зависимости от температуры, случайная РНК находится в одной из фаз: i) «расплавленная» высокотемпературная фаза (molten phase) или ii) «замороженная» низкотемпературная фаза (glass phase). В высокотемпературной фазе большую роль играет энтропия цепочки, нежели порядок мономеров в первичной структуре. Данная фаза хорошо описывается в модели гомополимера, комплементарное связывание не играет роли, и эффективно можно заменить все мономеры мономерами одного типа A. Низкотемпературная фаза, наоборот, определяется, в первую очередь, первичной структурой цепочки, то есть, основной вклад в свободную энергию обусловлен комплементарным связыванием мономеров. Такую фазу принято характеризовать замороженным беспорядком [17, 18]. Температура, при которой РНК переходит из одной фазы в другую, называется температурой фазового перехода и в литературе обозначается T_g.

Был предложен следующий подход к определению температуры фазового перехода. Рассмотрим пару мономеров, чье взаимодействие приводит к образованию петли наибольшего размера, т.е. нуклеотидов с номером 1 и $L/2$ по цепи для последовательности длиной L. Определим энергию выигрыша данного контакта, которая определяется как $\Delta F(L) = k_B T \ln P_{1,L/2}$, где $P_{1,L/2}$ — вероятность связывания 1 и $L/2$ мономера цепи. Данную энергию называют энергией пинча, и из выражения для статистической суммы цепочки следует, что:

$$\Delta F(L) = F_{1,L} - (F_{1,L/2} + F_{L/2+1,L}). \tag{65}$$

Отметим, что аналогичный вопрос исследуется и в теории перколяции, где тоже предполагается взаимосвязь перколяционного перехода и температурного фазового перехода, наблюдаемого, например, в модели Изинга [21].

Были проанализированы температурные зависимости свободной энергии пинча (2) случайной последовательности в модели Бернулли разной вероятности p. Как уже об-

суждалось, температура перехода в замороженное состояние T_g непосредственно связано со средним числом пропусков в структуре основного состояния. В [18] было показано, что температура перехода не превосходит T^*

$$T^* = \lambda^{-1}\sigma, \tag{66}$$

где σ — среднее число пропусков на пару мономеров, а λ определяется из зависимости наибольшего общего непрерывного сегмента ℓ двух половинок последовательности РНК: $\ell = \lambda^{-1}\ln L$ (см. Рис.1.4). Известно, что для цепочек РНК $\lambda = \ln 2$. Для случайного бернуллиевского процесса λ определяется как $\lambda = \ln(1/p)$ [6]. Таким образом, выражение (66) можно переписать в виде

$$T^* = \frac{\sigma}{\ln(1/p)}. \tag{67}$$

Доля несвязанных мономеров σ растет с ростом алфавита $1/p$ сильнее, чем логарифм (см. Рис. 10(б)) и из (67) непосредственно следует, что в допереходной области ($p > p_c$) фазовый переход в замороженное состояние наблюдаться не будет. Температура перехода T_g эффективно равна нулю, т.е., случайный полимер во всем температурном диапазоне находится в расплавленной фазе. Данное предположение дополнительно подтверждается наблюдением того, что для случайных последовательностей с алфавитом $c = 2$ переход имеет место только при накладывании ограничений на структуру, а именно, введением минимального размера петли [22].

Результаты численного моделирования представлены на Рис. 14. Был проанализирован температурный коэффициент $a(T)$ (4) для последовательностей с разной вероятностью p. Температура перехода определяется точкой, в которой нарушается линейная зависимость $a(T) = \frac{3}{2}T$, характерная для расплавленной фазы. Из полученных данных видно, что температура перехода уменьшается с ростом вероятности p и в допереходной области становится равной нулю ($p = 0.5$ на Рис. 14). Вблизи критического значения p_c численный эксперимент усложняется тем, что корректный анализ требует рассмотрения достаточно длинных случайных цепочек (с длиной, превышающей соответствующую релаксационную длину $\ell(p)$, см. (40)), что приводит к существенному увеличению времени численного моделирования. Также стоит отметить, что в связи с наблюдаемой степенной зависимостью свободной энергии основного состояния от длины последовательности ((40)), аппроксимация уравнением (2) вблизи точки $T = 0$, вообще говоря, неверна.

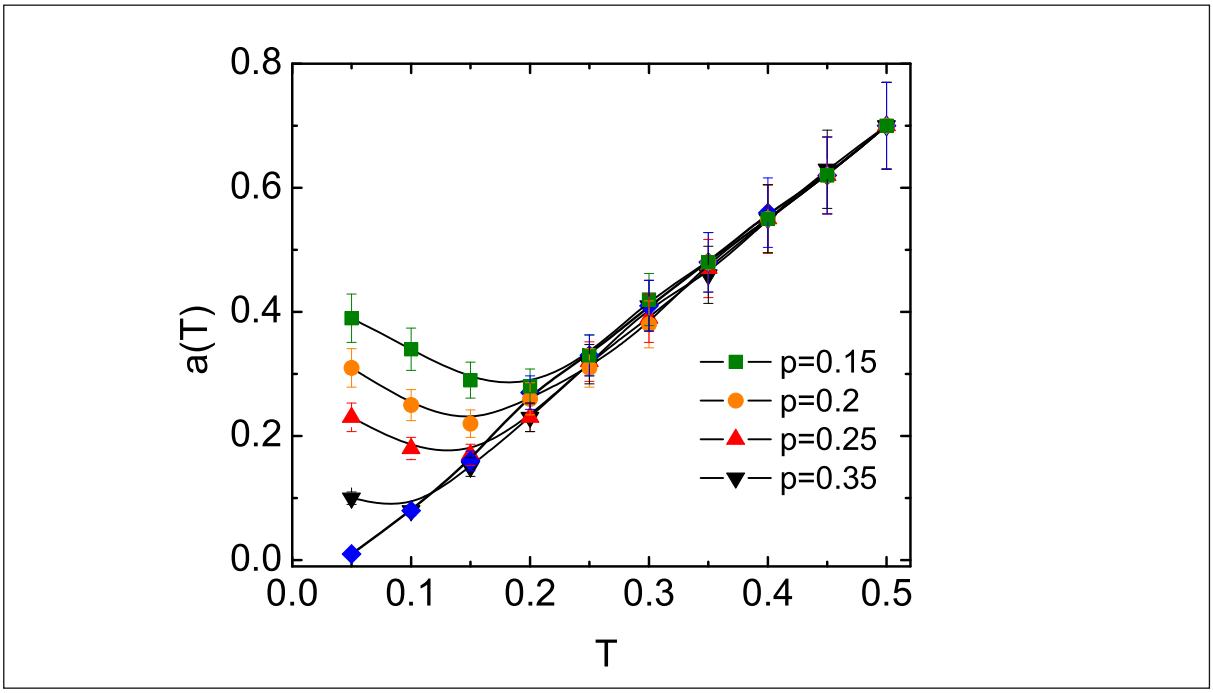


Рис. 14. Зависимость коэффициента $a(T)$ для случайной последовательности разной вероятности p в модели Бернулли.

Предполагается, что критическая точка топологического перехода между полностью связанной РНК-подобной структурой и структурой с пропусками является пороговым значением для термодинамического перехода. В области последовательностей $p > p_c$ возможна только расплавленная фаза вне зависимости от температуры. Рис. 15 показывает фазовую диаграмму на (T, p) плоскости. Это предположение подтверждается исследованием энергии пинча от длины случайной последовательности в точке $T = 0$. Точка пересечения зависимостей для разных длин (см. дополнительный график на Рис. 15) разделяет два топологических режима и близка к наблюдаемому критическому алфавиту.

E. Другие модели нецелого алфавита

Основной недостаток бернуллиевской модели полимера заключается в отсутствии ясного соответствия матрицы контактов V для произвольного p и первичной структуры полимера. Как уже указывалось, в модели Бернулли нет разделения на сорта мономеров, все мономеры, рассматриваются однотипными. В этом разделе, речь пойдет о некоторых подходах генерации полимера с нецелым алфавитом и разными сортами мономеров.

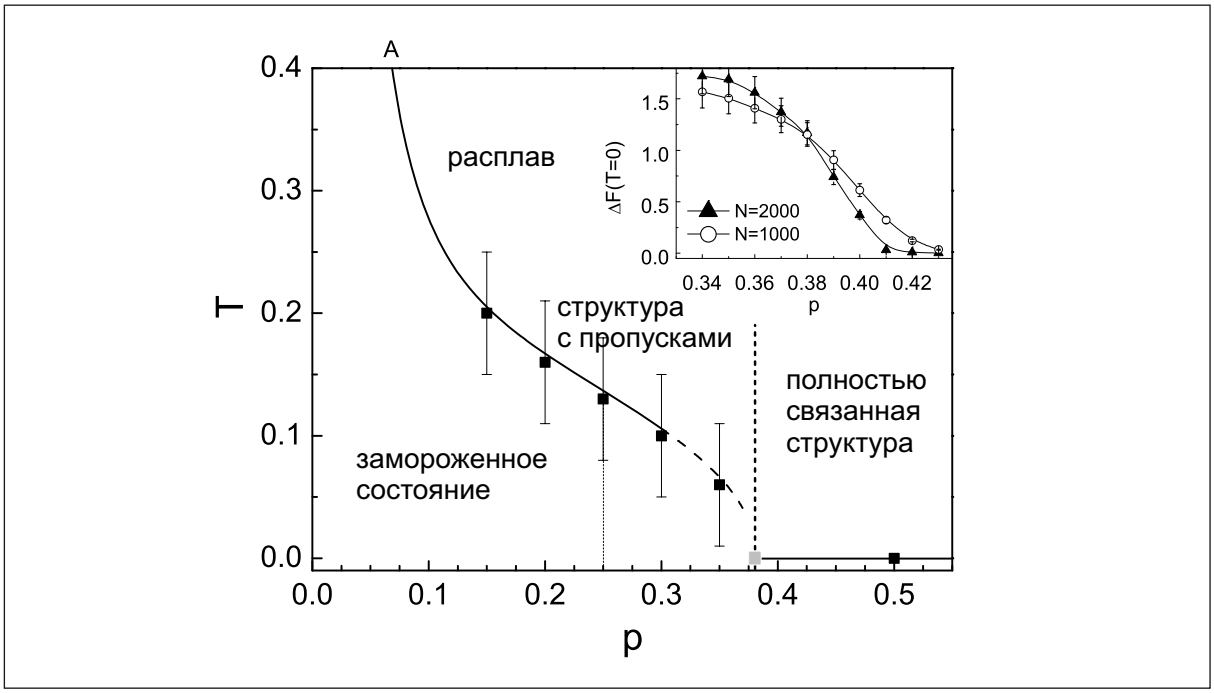


Рис. 15. Фазовый переход в замороженное состояние, ограниченный топологическим переходом в модели Бернулли. Дополнительный график: зависимость энергии пинча в пределе $T \to 0$ от вероятности p.

1. Метод концентраций

Одна из самых простых моделей нецелого алфавита, сохраняющего свойство транзитивности — модель концентраций. В такой модели предполагается, что случайный полимер состоит из трех типов мономеров, A, B и C, но мономеры распределены в цепочке не случайно, а коррелировано. В модели так называемых «локальных концентраций» предполагается, что концентрация, количество мономеров третьего типа $[C]$ не равна концентрации мономеров $[A] = [B]$, а зависит от алфавита p. В частности, концентрацию $[C]$ можно определить по заполненности матрицы контактов V. Изменение концентрации $[C]$ от 0 до $\frac{1}{3}$ описывает последовательности с нецелым алфавитом от $c = 2$ до $c = 3$. Однако, для алфавитов, немного превышающих $c = 2$ ($c = 2 + \epsilon$), данная модель приводит к случайной двухбуквенной последовательности, слабо разбавленной третьим типом мономеров C. Из-за малого количества $[C]$, эти мономеры появляются редко в цепочке, и из-за специфического комплементарного взаимодействия C–C, это приводит к сильным ограничениям на конфигурации основного состояния. Как уже упоминалось, основное состояние характеризуется большим количеством коротких арок, т.е., взаимодействием ближайших соседей в цепочке. Таким образом, важнее оказывается не количество различных типов мономеров в первичной структуре, а их

распределение. Модель концентраций может быть улучшена, если распределять $[C]$ мономеров третьего типа не случайно, а согласно некому распределению, характерному для случайных трехбуквенных (двухбуквенных) последовательностей. Грубо говоря, это приведет к тому, что мономеры третьего типа C будут появляться в первичной структуре блоками. Но даже такая модель обладает существенным недостатком — выделенной ролью мономеров типа C, по сравнению с мономерами типа A и B.

F. Коррелированная случайная последовательность

Модель, которая устраняет этот недостаток — модель так называемой коррелированной последовательности, в которой распределение трех типов мономеров является не случайным, а сильно коррелированным. Различный алфавит p определяется в таких последовательностях не количеством (концентрацией) мономеров, а тем, насколько сильно скоррелировано появление мономеров различного типа в цепочке. Распределение мономеров в первичной структуре определяется согласно Марковскому процессу [12]:

	A	B	C
A	$1-2\epsilon$	ϵ	ϵ
B	ϵ	$1-2\epsilon$	ϵ
C	ϵ	ϵ	$1-2\epsilon$,

т.е., вероятность встретить, например, мономер типа A за мономером A не равна вероятности появления A после B (или C). Изменение ϵ в диапазоне $[0, \frac{1}{3}]$ обеспечивает диапазон алфавитов $[1, 3]$. Взаимосвязь между параметром модели ϵ и алфавитом c можно установить, используя определение информационной энтропии по Шеннону [23]:

$$c = \left(\frac{1}{\epsilon} - 2\right)^{2\epsilon} \frac{1}{1-2\epsilon}. \tag{68}$$

Результатом данной модели является полимер с блочной первичной структурой, причем размер блока зависит от параметра модели ϵ. На Рис. 16 представлены результаты численного моделирования в модели такой коррелированной последовательности. Скачкообразного изменения удельной энергии от алфавита c в численном моделировании не наблюдается. Отметим, что даже для алфавитов, эффективное значение которых меньше 2, идеальные полностью связанные структуры не образуются. Объяснить такую зависимость можно следующим образом. После процедуры вычеркивания коррелированный полимер с трехбуквенным алфавитом сводится к последовательности со

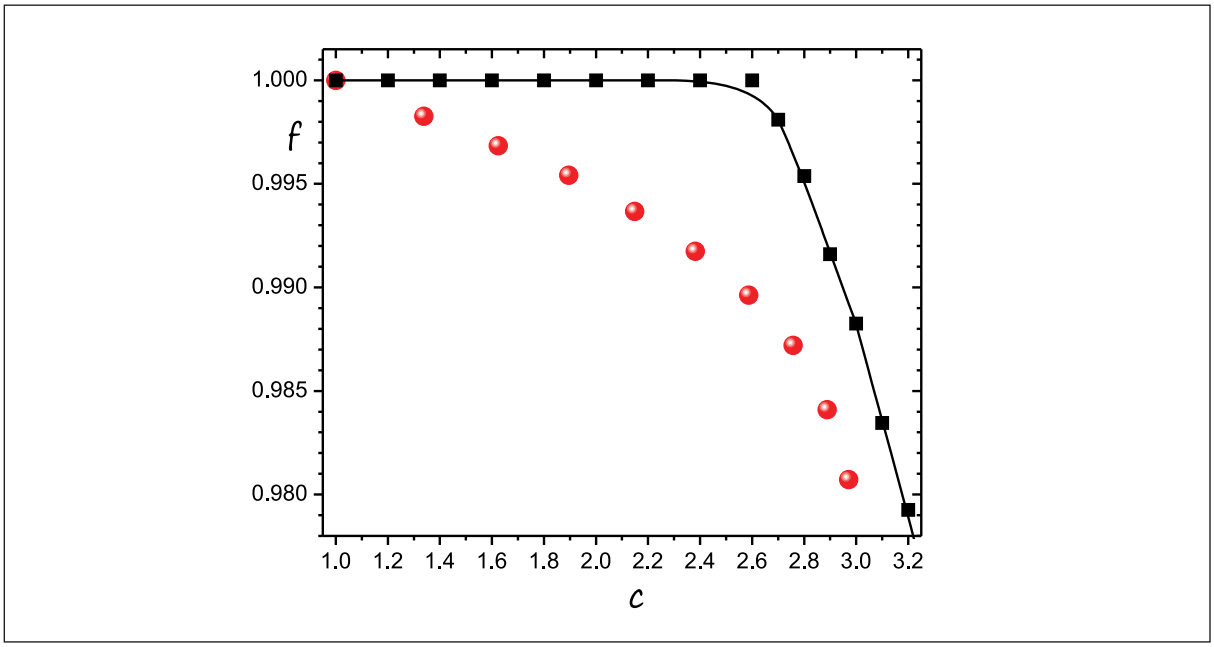


Рис. 16. Модель коррелированной последовательности: зависимость удельной энергии f_∞ от алфавита (красным); для сравнения приведена зависимость предельной энергии в модели Бернулли (черным).

случайно распределенными мономерами трех сортов, для которой, как было показано (Рис. 10) свойственно образование $O(L)$ пропусков. Длина остаточного полимера (после вычеркивания) зависит от параметра ϵ, и для достаточно длинных последовательностей пропорциональна L. Таким образом, любая модель с тремя сортами мономеров (коррелированная последовательность или модель концентраций) всегда сводится к модели случайной последовательности с алфавитом $c = 3$, который лежит в послепереходной области. Тем не менее, стоит отметить, что для зависимости удельной свободной энергии (см. Рис. 16) в модели коррелированной последовательности характерен резкий спад удельной энергии основного состояния при $c > c_c$.

1. Рациональный алфавит

Другая модель, частично сохраняющая свойство транзитивности, — модель рационального алфавита — заключается в следующем. Последовательность с алфавитом $\frac{P}{Q}$ можно представить, как полимер, состоящий из P сортов мономеров, правила комплементарности для которых разрешают Q связей для каждого мономера. Например, алфавит $c = \frac{5}{2}$ означает пятибуквенный алфавит в первичной структуре и с правилами комплементарности, организованными, к примеру, по пятиугольнику (Рис. 17). Такие правила можно построить для рационального алфавита любой величины. Численные

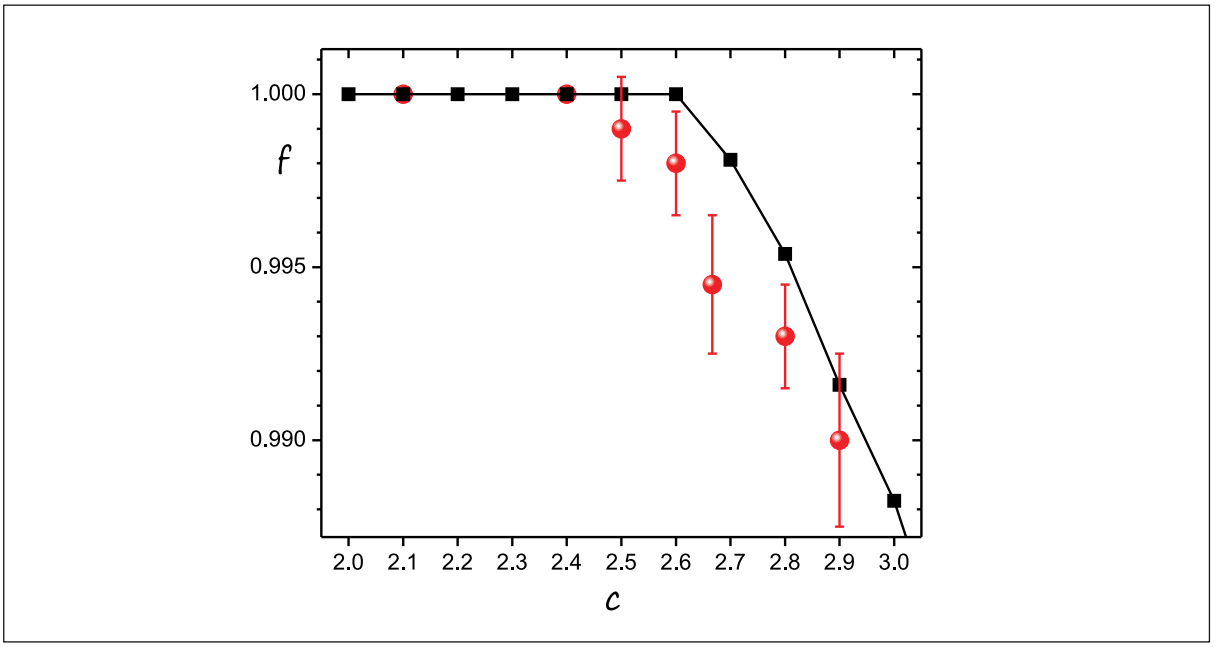


Рис. 17. Модель рационального алфавита: зависимость удельной энергии основного состояния f_∞ от алфавита (красным). Для сравнения приведена зависимость, полученная в модели Бернулли (черным).

результаты для данной модели приведены на Рис. 17 и показывают критическое изменение топологии вторичной структуры РНК.

Отметим, что модель чувствительна к выбору P и Q, так например, один и тот же алфавит $c = 2.8$, представленный как $\frac{14}{5}$ и $\frac{28}{10}$ дает разные результаты для удельной энергии основного состояния. В пределе $P \to L$ рассматриваемая модель сводится к модели Бернулли. Модель рационального алфавита, интуитивно, кажется ближе к алфавиту, используемого природой в молекулах РНК. Как указывалось, в молекулах РНК, помимо комплементарных пар, образуются неканонические пары, т.е. правила образования связей, во-первых, не транзитивны, а во-вторых, система правил похожа на систему связей в модели рационального алфавита (Рис. 17).

Каков алфавит в реальных молекулах РНК? Понятно, что учет неканонических пар эффективно приводит к уменьшению алфавита. С другой стороны, учет, к примеру, минимальной длины петли увеличивает алфавит в последовательностях РНК. Образование псевдоузлов и стэкинг взаимодействия приводит к сдвигу алфавита к меньшим значениям. Таким образом, фактический алфавит в молекулах РНК определяется многими факторами. Однако, логично предполагать, что алфавит в РНК находится вблизи критического. Почему выгодно реальным молекулам РНК иметь алфавит вблизи критического? Для того чтобы РНК выполняла свою биологическую функцию, она должна удовлетворять следующим критериям: i) ее фолдинг должен быть достаточно уникален

и ii) структура должна быть устойчива к тепловому шуму. Короткие алфавиты $c < c_c$ не обеспечивают первый критерий, так как для допереходной фазы характерна сильная вырожденность основного состояния. С другой стороны, длинным алфавитам $c >> c_c$ свойственны структуры с длинными петлями — несвязанными мономерами, которые неустойчивы к тепловым флуктуациям. Таким образом, биологический алфавит, по-видимому, находится вблизи критического.

V. ЗАКЛЮЧЕНИЕ

В работе представлен анализ топологических свойст РНК-подобных молекул со случайной первичной структурой методами статистической физики и теории случайных процессов. Получено выражение для статистической суммы, описывающие взаимодействие двух сополимеров, учитывающий способность каждого из сополимеров образовывать РНК-подобную структуру с иерархией петлевых участков. Разработан соответствующий алгоритм динамического программирования вычисления свободной энергии основного состояния таких РНК-подобных молекул. Численно и аналитически показано критическое поведение РНК-подобной структуры в зависимости от используемого в первичной структуре алфавита. Существует две области: для алфавитов $c < c_c$ свойственна максимально связанная вторичная структура без пропусков, тогда как для $c > c_c$ вторичная структура содержит конечную долю несвязанных мономеров. Аналитическая оценка точки топологического перехода $c_c = 2.87$ близка к наблюдаемой в численном моделировании $c_c = 2.67$.

Литература

[1] *Птицын Б.О., Финкельштейн А.* Физика белка: Курс лекций // Москва: Университет, 2002, 376 С.

[2] *Гросберг Ю.А., Хохлов Р.А.* Статистическая физика макромолекул // Москва: Наука, 1989, 344 С.

[3] *Workman C., Krogh A.* No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution // Nucleic Acids Research, 1999,**27** (24), P. 4816-4822.

[4] *Clote P., Ferre F., Kranakis E., Krizanc D.* Structural RNA has lower folding energy than randomRNA of the same dinucleotide frequency // RNA, 2005, **11**(5), P. 578-591.

[5] *Brezin E.E., Itzykson C., Parisi G., Zuber J.B.* Planar diagrams // Communications in Mathematical Physics,1978, **59**, P. 5-51.

[6] *Waterman M.S., Vingron M.* Sequence comparison significance and poisson approximation // Statistical Science, 1994, **9**, P. 367-381.

[7] *Majumdar S.T., Nechaev S.K.* Exact asymptotic results for the bernoulli matching model of sequence alignment // Physical Review E, 2005, **72** (2), P. 020901.

[8] *Kriecherbauer T., Krug J.* A pedestrian's view on interacting particle systems: KPZ universality and random matrices // Journal of Physics A: Mathematical and Theoretical, 2010, **43**(40), P. 403001.

[9] *Kardar M., Parisi G., Zhang Y.C.* Dynamic scaling of growing interfaces // Physical Review Letters, 1986, **56** (9), P. 889-892.

[10] *Тамм М.В., Лисаченко Н.Г., Ерухимович И.Я., Иванов В.А.* Эффекты конечного объема в системе равновесных циклических полимеров: теория и компьютерное моделирование // Высокомолекулярные соединения, 2005, **47** (7), С. 348-352.

[11] *Ландо К.* Лекции о производящих функциях // Москва: Московский центр непрерывного математического образования, 2007, 144 С.

[12] *Feller W.* An introduction to probability theory and its applications // New York: Wiley, 1968, 509 p.

[13] *Владимиров А.А.* Паросочетания без пересечений // Проблемы передачи информации, 2013, **49**(1), С. 61-65.

[14] *Grimmett G.* What is Percolation? // New York: Springer, 1999, 444 P.

[15] *Tamm M., Nechaev S.* Necklace-cloverleaf transition in associating RNA-like diblock copolymers // Physical Review E, 2007, **75** (3), P. 031904.

[16] *Zee A.* Random matrix theory and RNA folding // Acta Physica Polonica B, 2005, **36** (9), P. 2829-36.

[17] *Bundschuh R., Hwa T.* RNA secondary structure formation: A solvable model of heteropolymer folding // Physical Review Letters, 1999, **83** (7), P. 1479-1482.

[18] *Bundschuh R., Hwa T.* Statistical mechanics of secondary structures formed by random RNA sequences // Physical Review E, 2002, **65** (3), P. 031903.

[19] *Lassig M., Wiese K.J.* Freezing of random RNAs // Physical Review Letters, 2006, **96** (22), P. 228101.

[20] *David F., Wiese K.J.* Systematic field theory of the RNA glass transition // Physical Review Letters, 2007. **98** (12), P. 128102.

[21] *Toninelli C., Biroli G., Fisher D.* Jamming percolation and glass transitions in lattice models // Physical Review Letters, 2006, **96** (3), P. 035702.

[22] *Pagnani A., Parisi G., Ricci-Tersenghi F.* Glassy transition in a disordered model for the RNA secondary structure // Physical Review Letters, 2000, **84** (9), P. 2026-2030.

[23] *Shannon C.E., Weaver W.* A mathematical theory of communication // The Bell System Technical Journal, 1948, **27**, P. 379-423.

TOPOLOGICAL PROPERTIES OF RNA-LIKE MOLECULES WITH A RANDOM PRIMARY STRUCTURE

O. Valba[1, 2], M. Tamm[1, 3], S. Nechaev[4, 5]

[1]*National Research University HSE*
[2]*Semenov Institute of Chemical Physics*
[3]*Moscow State University*
[4]*Universitet Paris-Sud / CNRS*
[5]*Lebedev Physical Institute*

ovalba@hse.ru

Received 20.04.2015

The paper focuses on the application of methods of statistical physics and stochastic processes for the study of the topological properties of RNA-like heteropolymers with a random primary structure. In particular, it describes the critical behavior of RNA-like secondury structure topology in dependence on the alphabet used in a random sequence. The analytical evaluation of the critical transition point based on the combinatorial and matrix description is given.