ЖЭТФ, 1999, том 116, вып. 2(8), стр. 620–635

О КРИТЕРИЯХ РЕГУЛЯРНОСТИ ДЛЯ ЛИНЕЙНЫХ ЦЕПЕЙ

В. В. Лобзин[a]*, *В. Р. Чечеткин*[b]

[a] *Институт земного магнетизма, ионосферы и распространения радиоволн Российской академии наук*
142092, Троицк, Московская обл., Россия
[b] *Троицкий институт инновационных и термоядерных исследований*
142092, Троицк, Московская обл., Россия

Поступила в редакцию 26 марта 1999 г.

В рамках спектрального подхода обсуждается проблема структурного анализа линейных цепей произвольной фиксированной формы. Форма цепи характеризуется набором скалярных и псевдоскалярных инвариантов, не изменяющихся при сдвигах и поворотах системы как целого. Статистические свойства набора инвариантов сравниваются с аналогичными характеристиками для свободно-сочлененной цепи. Предлагаемые критерии обладают свойством самоусредняемости для сравнительно коротких цепей из $\sim$ 100–300 звеньев и позволяют выявить возможные скрытые периодичности и симметрии в системе. В качестве приложения теории рассматривается структурный анализ для цепей, генерируемых случайными блужданиями на кубической решетке, и C_α-остовов белков.

PACS: 87.10; 02.50

1. ВВЕДЕНИЕ

Конформации длинных полимерных молекул в первом приближении характеризуются с помощью положений фиксированных атомов в повторяющихся молекулярных группах [1–3]. Выделенному таким образом остову молекулы соответствует некоторая пространственная цепь. Для целого ряда полимерных молекул, и прежде всего для широкого класса белков, складывание в трехмерную структуру при заданных физических условиях в растворе происходит строго определенным образом (с точностью до относительно незначительных тепловых колебаний). Именно специфичность складывания белков обеспечивает специфичность молекулярного узнавания и отбор химических реакций, в которых участвует конкретный белок. В современных базах данных собрана рентгеноструктурная информация о примерно 7500 белковых структурах. Предполагается, что через два года их число превысит 12000 [4].

Физический анализ фиксированных остовов молекул наталкивается на значительные трудности, поскольку изучаемые структуры содержат в себе одновременно черты как упорядоченных, так и случайных структур. Поэтому, например, остовы белков (или C_α-остовы по положению выделенного атома углерода) классифицируются обычно во внутренних терминах на основе сравнения структур между собой [5–7]. Такие интегральные характеристики, как радиус гирации или отношение главных осей эллипсоида инерции [1, 8] оказываются слишком грубыми и не отражают внутреннюю хиральность молекул. С другой стороны, длины типичных белков сравнительно невелики

*E-mail: lobzin@top.izmiran.troitsk.ru

(~ 100–400 звеньев) и не обладают ярко выраженным самоподобием. Поэтому скейлинговые зависимости различных характеристик для белковых остовов также довольно плохо отражают детальные особенности складывания (ср. [9, 10]).

Введение количественной меры регулярности для пространственных конфигураций линейных цепей имеет важный информационный аспект, так как со специфичностью складывания можно связать ту или иную меру сложности [11–14], которая может отражать специфичность молекулярных взаимодействий или эволюционные особенности. Прямое распространение классической теории информации [15] на этот случай приводит к «комбинаторному взрыву». Пусть, например, звенья линейной цепи описываются полярными углами $0 \le \varphi \le 2\pi$, $0 \le \theta \le \pi$ [1]. Тогда, если разбить пространство (φ, θ) на M сегментов, то фрагменты из n звеньев могут принимать M^n различных конформаций. Поэтому даже для грубого разбиения на квадранты ($M = 6$) возможный анализ ограничен уже трехзвенными фрагментами, поскольку число различных комбинаций M^n не должно превышать полную длину L (~ 100–400 для белков). С другой стороны, вводимые критерии регулярности должны отражать не только локальные, но и глобальные характеристики структуры цепи. Наконец, они должны обладать свойством самоусредняемости для возможно более коротких цепей, чтобы позволять анализировать отдельные структуры с $L \sim$ 100–400.

В данной работе мы покажем, что достаточно простые и удобные в практических приложениях критерии можно получить в рамках спектрального подхода. Ранее аналогичный подход использовался для анализа символических последовательностей и геномных последовательностей ДНК [16]. Такой подход дает эффективные критерии для оценки интегральной регулярности [17], позволяет выявить скрытые периодичности в системе [18] и проанализировать характер корреляций [19]. Как будет показано ниже, несмотря на различие в постановках задачи, многие результаты [16–19] можно распространить и на случай пространственных линейных цепей.

Статья имеет следующий план. В разд. 2 дается общая постановка задачи и строится система инвариантов, характеризующих форму линейных цепей. Такие инварианты позволяют эффективно выявить возможные скрытые периодичности и симметрии в системе (разд. 3). Статистические характеристики для системы скалярных и псевдоскалярных инвариантов, введенных в разд. 2, исследуются в разд. 4 для случайных свободно-сочлененных цепей. Сравнение с характеристиками для случайных свободно-сочлененных цепей позволяет ввести ряд различных критериев регулярности (разд. 5). Общая теория иллюстрируется в разд. 6 на конкретных примерах цепей, генерируемых случайными блужданиями на кубической решетке, и C_α-остовов белковых структур. Заключительный разд. 7 содержит некоторые замечания о возможном обобщении результатов.

2. СПЕКТРАЛЬНОЕ ПРЕДСТАВЛЕНИЕ И СИСТЕМА ИНВАРИАНТОВ ДЛЯ ЛИНЕЙНОЙ ЦЕПИ

Пусть заданы декартовы координаты узлов цепи (или фиксированных атомов в полимерной молекуле) $\{\mathbf{r}_m\}$, $m = 1, \ldots, L$. Поскольку форма фиксированной конфигурации цепи не изменяется при сдвигах и поворотах системы как целого, внутренние структурные характеристики цепи, связанные с формой, следует описывать в терминах системы инвариантов, не меняющихся при сдвигах и поворотах. В принципе, для опи-

сания формы можно использовать любую систему инвариантов, лишь бы число независимых инвариантов было достаточно велико (сравнимо с числом внутренних степеней свободы $3L-6$) и позволяло бы детально охарактеризовать форму. Система инвариантов, которая будет введена в этом разделе, обладает тем дополнительным преимуществом, что позволяет использовать стандартные методы спектрального анализа [20]. Как будет показано ниже, такая система инвариантов допускает достаточно полное статистическое исследование и дает удобные количественные критерии регулярности цепи.

Введем векторные фурье-гармоники

$$\begin{aligned} &\boldsymbol{\rho}(q_n) = (L-1)^{-1/2} \sum_{m=1}^{L-1} \Delta\mathbf{r}_m \exp(-iq_n m), \\ &q_n = 2\pi n/(L-1), \qquad n = 0, 1, \ldots, L-2, \end{aligned} \tag{2.1}$$

$$\Delta\mathbf{r}_m = \mathbf{r}_{m+1} - \mathbf{r}_m. \tag{2.2}$$

Образуя из гармоник $\boldsymbol{\rho}(q_n)$ с различными q_n всевозможные скалярные и смешанные произведения, получаем систему скалярных и псевдоскалярных инвариантов, не меняющихся при сдвигах и поворотах.

Для того чтобы подобная система инвариантов характеризовала внутреннюю регулярность цепи, на систему необходимо наложить еще одно дополнительное условие. Поясним его смысл на конкретном примере. Выделим связи от 1 до Δm и перенесем фрагмент цепи между 1 и Δm параллельно из начала в конец. Тогда векторные фурье-гармоники для цепи с перенесенным фрагментом, $\boldsymbol{\rho}^{\mathrm{tr}}(q_n)$, и исходной цепи, $\boldsymbol{\rho}^{\mathrm{in}}(q_n)$, связаны между собой соотношением

$$\boldsymbol{\rho}^{\mathrm{tr}}(q_n) = \exp(-iq_n\Delta m)\boldsymbol{\rho}^{\mathrm{in}}(q_n). \tag{2.3}$$

Поскольку внутренняя регулярность обеих цепей одинакова, из (2.3) следует, что сумма волновых векторов для гармоник $\boldsymbol{\rho}(q_n)$ в скалярных и смешанных произведениях должна равняться $2\pi k$, где k — целое число. Окончательно для системы инвариантов получаем

$$\begin{aligned} &I(q_{n_1}, \ldots, q_{n_r}) = \mathrm{Inv}\left\{\rho_{\alpha_1}(q_{n_1}) \ldots \rho_{\alpha_r}(q_{n_r})\right\}, \\ &q_{n_1} + \ldots + q_{n_r} = 0 \bmod 2\pi. \end{aligned} \tag{2.4}$$

В (2.4) греческий индекс для $\rho_\alpha(q_n)$ отвечает проекции в декартовой системе координат, $\alpha \in \{x, y, z\}$. Предполагается, что из различных компонент $\rho_\alpha(q_n)$ с помощью символов Кронекера $\delta_{\alpha\beta}$ и Леви-Чивиты $\varepsilon_{\alpha\beta\gamma}$ устраиваются всевозможные свертки. Свертки с нечетным числом $\varepsilon_{\alpha\beta\gamma}$ дадут псевдоскалярные инварианты, меняющие знак при зеркальных отражениях, а свертки с четным числом символов $\varepsilon_{\alpha\beta\gamma}$ дадут скалярные инварианты, не меняющиеся при отражениях.

Учитывая, что вещественность $\Delta\mathbf{r}_m$ приводит к соотношению

$$\boldsymbol{\rho}(q_n) = \boldsymbol{\rho}^*(2\pi - q_n) \tag{2.5}$$

(здесь и далее звездочка означает комплексное сопряжение), получаем простейшие скалярные и псевдоскалярные инварианты:

$$F(q_n) = \boldsymbol{\rho}(q_n)\boldsymbol{\rho}^*(q_n), \tag{2.6}$$

$$H(q_n) = i\left[\boldsymbol{\rho}(q_n)\boldsymbol{\rho}^*(q_n)\right]\boldsymbol{\rho}(0). \tag{2.7}$$

Инварианты $F(q_n)$ далее будут называться структурными факторами.

Непосредственно из определения (2.1) можно получить точное правило сумм:

$$\sum_{q_{n_1}+\ldots+q_{n_r}=0 \bmod 2\pi} \rho_{\alpha_1}(q_{n_1})\ldots\rho_{\alpha_r}(q_{n_r}) = (L-1)^{(r-2)/2}\sum_{m=1}^{L-1}\Delta r_{m,\alpha_1}\ldots\Delta r_{m,\alpha_r}. \tag{2.8}$$

Из (2.8) немедленно следует, что все псевдоскалярные инварианты имеют нулевую среднюю спектральную плотность, так как свертка любого из символов $\varepsilon_{\alpha\beta\gamma}$ с правой частью даст нуль. Для структурных факторов $F(q_n)$ гармоника с $q_n = 0$ пропорциональна квадрату расстояния между концами цепи, а средняя величина гармоник с $q_n \neq 0$ равна

$$\overline{F} = \frac{1}{L-2}\sum_{n=1}^{L-2}F(q_n) = \frac{1}{L-2}\sum_{m=1}^{L-1}(\mathbf{r}_{m+1}-\mathbf{r}_m)^2 - \frac{(\mathbf{r}_L-\mathbf{r}_1)^2}{(L-1)(L-2)}. \tag{2.9}$$

Поскольку длины связей в цепи обычно примерно равны, $\Delta\mathbf{r}_m^2 \approx \text{const}$, заданная средняя спектральная высота $\overline{F}$ отвечает фиксированному расстоянию между концами.

Структурные факторы $F(q_n)$ с учетом соотношения Винера—Хинчина можно привести в соответствие с корреляционной функцией

$$K(m_0) = \frac{1}{L-1}\sum_{n=0}^{L-2}F(q_n)\exp(-iq_nm_0), \qquad m_0 = 0,\ldots,L-2. \tag{2.10}$$

Корреляционную функцию $K(m_0)$ можно представить также в виде

$$K(m_0) = \frac{1}{L-1}\sum_{m=1}^{L-1}\Delta\mathbf{r}_m^c\Delta\mathbf{r}_{m+m_0}^c, \tag{2.11}$$

где $\Delta\mathbf{r}_m^c$ — циклически продолженные разности координат:

$$\Delta\mathbf{r}_m^c = \begin{cases}\Delta\mathbf{r}_m, & 1\le m\le L-1,\\ \Delta\mathbf{r}_{m-L+1}, & L\le m\le 2L-2.\end{cases} \tag{2.12}$$

Определение (2.10), в свою очередь, также приводит к различным правилам сумм, наиболее важным из которых является [16]

$$\sum_{m_0=1}^{L-2}\left(K(m_0)-\overline{K}\right)^2 = \frac{1}{L-1}\sum_{n=1}^{L-2}\left(F(q_n)-\overline{F}\right)^2, \tag{2.13}$$

где $\overline{K}$ — среднее значение для $K(m_0)$ с $m_0 \neq 0$ (ср. (2.9)).

Условие (2.5) и определения (2.6), (2.10) приводят к симметрии соответствующих спектров:

$$F(q_n) = F(2\pi - q_n), \qquad K(m_0) = K(L-1-m_0). \tag{2.14}$$

Поэтому для $F(q_n)$ и $K(m_0)$ можно ограничиться только левыми полуспектрами, $1 \le n \le N$ и $1 \le m_0 \le N$, где

$$N = [(L-1)/2]. \tag{2.15}$$

Здесь квадратные скобки обозначают целую часть числа.

Аналогичные соотношения нетрудно получить также для псевдоскалярных инвариантов $H(q_n)$ и других высших инвариантов.

3. СТРУКТУРНЫЕ ФАКТОРЫ, СКРЫТЫЕ ПЕРИОДИЧНОСТИ И СКРЫТЫЕ СИММЕТРИИ

Рассмотрим сначала, какую физическую информацию о структуре можно получить непосредственно из спектров построенных наборов инвариантов. Хорошо известно, что фурье-преобразование используется обычно для выделения скрытых периодичностей на фоне случайных вкладов [20]. В самом деле, рассмотрим цепь, составленную из повторяющихся фрагментов из l звеньев. Тогда векторные гармоники $\boldsymbol{\rho}(q_n)$ будут иметь вид

$$\boldsymbol{\rho}(q_n) = \mathbf{b}(q_n) + \exp(-ilq_n)\mathbf{b}(q_n) + \ldots + \exp(-i(M-1)lq_n)\mathbf{b}(q_n), \tag{3.1}$$

где $Ml = L - 1$ и фурье-преобразование для $\mathbf{b}(q_n)$ осуществляется только с помощью суммирования по длине повтора:

$$\begin{aligned} &\mathbf{b}(q_n) = (L-1)^{-1/2} \sum_{m=1}^{l} (\mathbf{r}_{m+1} - \mathbf{r}_m) \exp(-iq_n m), \\ &q_n = 2\pi n/(L-1), \qquad n = 0, \ldots, L-2. \end{aligned} \tag{3.2}$$

Соответствующие скалярные структурные факторы равны

$$F(q_n) = \mathbf{b}(q_n)\mathbf{b}^*(q_n) \frac{\sin^2(Mlq_n/2)}{\sin^2(lq_n/2)}, \tag{3.3}$$

и их спектр состоит из эквидистантных пиков при $lq_n = 2\pi k$, $k = 0, \ldots, l-1$. Квазислучайные вариации длины l могут вызвать расщепление пиков и подавление пиков с $k \ge 2$. Поэтому скрытые периодичности в общем случае можно выявить с помощью оценки статистической значимости для суммы эквидистантных гармоник или по отдельным пикам, или сочетая оба этих критерия [18].

Существенно, что пространственные фурье-гармоники выявляют не только скрытые периодичности, но и скрытые симметрии. Покажем это на примере простейшей поворотной симметрии M-го порядка, отвечающей группе C_M [21, 22] (M-звездные конфигурации). Для этого случая векторные фурье-гармоники имеют вид

$$\boldsymbol{\rho}(q_n) = \mathbf{b}(q_n) + \exp(-ilq_n)\widehat{R}\mathbf{b}(q_n) + \ldots + \exp(-i(M-1)lq_n)\widehat{R}^{M-1}\mathbf{b}(q_n). \tag{3.4}$$

Здесь $\mathbf{b}(q_n)$ опять определяются формулой (3.2), $Ml = L - 1$, и 3×3-матрица поворота $\widehat{R}$ удовлетворяет условию

$$\widehat{R}^M = \underbrace{\widehat{R} \ldots \widehat{R}}_{M} = \widehat{I}, \tag{3.5}$$

где $\widehat{I}$ — единичная матрица. Матрица $\widehat{R}$ описывает поворот вокруг некоторой оси $\mathbf{n}$ на угол φ, $\widehat{R}^2$ означает вращение на 2φ и т.д. Тогда условие (3.5) дает $\varphi = 2\pi/M$.

Пусть $\mathbf{n}$ направлена вдоль оси $\mathbf{z}$ декартовой системы координат. Вводя комбинации

$$\mathbf{b}_\pm = \mathbf{b}_x \pm i\mathbf{b}_y, \tag{3.6}$$

можно показать, что

$$\widehat{R}\mathbf{b}_z = \mathbf{b}_z, \qquad \widehat{R}\mathbf{b}_\pm = \exp(\pm 2\pi i/M)\mathbf{b}_\pm. \tag{3.7}$$

Принимая во внимание, что

$$\mathbf{b} = \mathbf{b}_x + \mathbf{b}_y + \mathbf{b}_z = \frac{1}{2}(1-i)\mathbf{b}_+ + \frac{1}{2}(1+i)\mathbf{b}_- + \mathbf{b}_z, \tag{3.8}$$

и учитывая формулы (3.4), (3.7), (3.8), получаем для $F(q_n)$ серию эквидистантных пиков при $lq_n = 2\pi k$, $k = 0,\ldots,l-1$, отвечающих $\mathbf{b}_z$, серию при $lq_n = 2\pi k + 2\pi/M$, $k = 0,\ldots,l-1$, отвечающую $\mathbf{b}_+$, и серию при $lq_n = 2\pi k - 2\pi/M$, $k = 1,\ldots,l$, отвечающую $\mathbf{b}_-$. Поскольку $(L-1)/l = M$ и $q_n = 2\pi n/(L-1)$, поворотной симметрии M-го порядка соответствуют пиковые гармоники с номерами $n = Mk, Mk \pm 1$ (здесь k — целое число). Если $\mathbf{b}_z = 0$, то остаются только расщепленные пики при $n = Mk \pm 1$. Аналогичные соотношения нетрудно вывести и для других подгрупп симметрии при выделении соответствующих неприводимых представлений [21,22].

Если повторяющиеся элементы отвечают, например, спиралям, то анализ псевдоскалярных инвариантов позволит различить правые и левые спирали. Отметим также, что в ряде приложений может оказаться более удобной система псевдоскалярных инвариантов, связанная с топологическими характеристиками остова цепи [23].

4. СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ ДЛЯ СЛУЧАЙНЫХ СВОБОДНО-СОЧЛЕНЕННЫХ ЦЕПЕЙ

Поскольку преобразование (2.1) является обратимым, его можно рассматривать как более удобное отображение конфигурации цепи. Сравнивая различные структурные характеристики для фиксированных конфигураций цепи со средними характеристиками для случайной свободно-сочлененной цепи, можно получить ряд количественных критериев регулярности для цепи произвольной формы. В этом разделе мы рассмотрим общую теорию для анализа статистических характеристик случайных цепей.

Далее для простоты все связи предполагаются равными и имеющими единичную длину, $\Delta\mathbf{r}_m \equiv \mathbf{n}_m$ (где $\mathbf{n}_m$ — единичный вектор). Случайным аналогом для цепи с произвольной фиксированной формой является случайная свободно-сочлененная цепь с тем же числом звеньев и тем же расстоянием между концами, $|\mathbf{r}_L - \mathbf{r}_1|$ (см. (2.9)). Статистические распределения для инвариантов такой цепи можно получить с помощью характеристической функции (ср. [1,24,25])

$$Z = \frac{1}{\Omega}\int d\mathbf{n}\int d\mathbf{n}_1 \ldots \int d\mathbf{n}_{L-1}\,\delta(\mathbf{n}_1 + \ldots + \mathbf{n}_{L-1} - R\mathbf{n})\exp\left(i\sum_{n=0}^{L-2}\mathbf{u}(q_n)\boldsymbol{\rho}(q_n)\right). \tag{4.1}$$

Величина R при интегрировании с δ-функцией считается фиксированной, гармоники $\boldsymbol{\rho}(q_n)$ определяются уравнением (2.1) с $\Delta\mathbf{r}_m = \mathbf{n}_m$, а на вспомогательные векторные переменные $\mathbf{u}(q_n)$ удобно наложить условие, аналогичное (2.5):

$$\mathbf{u}(q_n) = \mathbf{u}^*(2\pi - q_n). \tag{4.2}$$

Нормировочный множитель Ω равен

$$\Omega = \int d\mathbf{n} \int d\mathbf{n}_1 \ldots \int d\mathbf{n}_{L-1}\, \delta(\mathbf{n}_1 + \ldots + \mathbf{n}_{L-1} - R\mathbf{n}). \tag{4.3}$$

Различные средние можно получить путем дифференцирования Z:

$$\langle \rho_{\alpha_1}(q_{n_1})\, \rho_{\alpha_2}(q_{n_2}) \ldots \rangle = \left. \frac{\partial^{\cdots} Z}{i\partial u_{\alpha_1}(q_{n_1})\, i\partial u_{\alpha_2}(q_{n_2}) \ldots} \right|_{\{\mathbf{u}(q_n)=0\}} \tag{4.4}$$

(здесь и далее угловые скобки означают усреднение по ансамблю случайных реализаций).

Переписывая экспоненту в (4.1) в виде

$$\exp\left(i \sum_{n=0}^{L-2} \mathbf{u}(q_n)\boldsymbol{\rho}(q_n) \right) = \exp\left(i\mathbf{g}_0 \mathbf{n} R + i \sum_{m=1}^{L-1} \mathbf{n}_m \mathbf{g}_m \right), \tag{4.5}$$

где

$$\mathbf{g}_m = (L-1)^{-1/2} \sum_{n=1}^{L-2} \mathbf{u}(q_n) \exp(-iq_n m), \tag{4.6}$$

$$\mathbf{g}_0 = \mathbf{u}(0)/(L-1)^{1/2}, \tag{4.7}$$

нетрудно видеть также, что при $m \neq 0$

$$\langle n_{m_1,\alpha_1}\, n_{m_2,\alpha_2} \ldots \rangle = \left. \frac{\partial^{\cdots} Z}{i\partial g_{m_1,\alpha_1}\, i\partial g_{m_2,\alpha_2} \ldots} \right|_{\{\mathbf{g}_m=0\}}. \tag{4.8}$$

Используя интегральное фурье-преобразование для δ-функции в (4.1), можно свести выражение для Z к трехкратному интегралу:

$$Z = \frac{1}{\Omega} \int d^3\mathbf{g}\, \frac{\sin\left(|\mathbf{g} - \mathbf{g}_0|R\right)}{|\mathbf{g} - \mathbf{g}_0|R} \prod_{m=1}^{L-1} \frac{\sin\left(|\mathbf{g} + \mathbf{g}_m|\right)}{|\mathbf{g} + \mathbf{g}_m|}. \tag{4.9}$$

Из выражений (4.1)–(4.9), в частности, получаем

$$\langle \boldsymbol{\rho}(q_n)\boldsymbol{\rho}^*(q_n) \rangle = \langle F(q_n) \rangle = \overline{F}, \tag{4.10}$$

где $\overline{F}$ — среднее по спектру, определяемое уравнением (2.9). Равенство (4.10) отражает важное квазиэргодическое свойство: усреднение по ансамблю случайных реализаций эквивалентно усреднению по спектру. В частности, все средние значения для псевдоскалярных инвариантов оказываются равными нулю (ср. разд. 2).

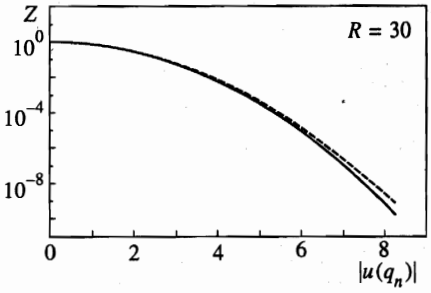

Рис. 1. Сравнение точного выражения (4.9) (сплошная линия) с приближенным (4.11) (штриховая линия) для цепи из $L - 1 = 100$ звеньев при фиксированном расстоянии R между концами цепи

При $L \gg 1$ аналогично [16] можно получить асимптотическое кумулянтное разложение для Z на основе точного правила сумм (2.8). В главном приближении по $\sim L^{-1}$ получаем

$$Z \approx \frac{\sin\left(|\mathbf{u}_0|R/(L-1)^{1/2}\right)}{|\mathbf{u}_0|R/(L-1)^{1/2}} \exp\left(-\frac{\overline{F}}{3}\sum_{n=1}^{N}\mathbf{u}(q_n)\mathbf{u}^*(q_n)\right). \tag{4.11}$$

Здесь учтено условие (4.2), и поэтому суммирование в (4.11) ограничено полуспектром (см. (2.15)).

На рис. 1 дается сравнение точного выражения (4.9) с приближенным (4.11) для случая, когда все переменные $\mathbf{u}(q_n)$ равны нулю, кроме одной пары комплексно-сопряженных переменных для произвольно выбранного значения q_n. Полное число звеньев равнялось $L - 1 = 100$. Как видно из рис. 1, точное выражение почти совпадает с приближенным в области $Z \geq 10^{-6}$. Поэтому для большинства практических приложений можно ограничиться приближением (4.11). Однако при больших $|\mathbf{u}(q_n)|$ асимптотики оказываются различными и при $Z \leq 10^{-15}$ расхождение между асимптотиками составляет примерно порядок.

5. КРИТЕРИИ РЕГУЛЯРНОСТИ ДЛЯ ЛИНЕЙНЫХ ЦЕПЕЙ

Рассмотрим некоторые конкретные следствия из общего выражения (4.11). Как известно, характеристическая функция Z связана с многокомпонентной функцией распределения вероятностей преобразованием Фурье [24, 25]. Далее нас будет интересовать только распределение амплитуд для гармоник с $q_n \neq 0$. В главном приближении по $\sim L^{-1}$ соответствующую многокомпонентную плотность распределения вероятностей можно представить в виде

$$p(F_1, \ldots, F_N) = p(F_1) \ldots p(F_N), \tag{5.1}$$

где плотность распределения для отдельных структурных факторов, $p(F_n)$, определяется выражением

$$p(F)\,dF = \frac{1}{2} f^2 \exp(-f)\,df, \tag{5.2}$$

$$f_n \equiv 3F(q_n)/\overline{F}. \tag{5.3}$$

Вероятность того, что амплитуда отдельного структурного фактора F_n не превысит величины F, равна

$$P(F_n \le F) = \int_0^F p(F')\,dF' = 1 - \left(1 + f + \frac{f^2}{2}\right)\exp(-f), \tag{5.4}$$

и, соответственно, вероятность превышения величины F является дополнительной по отношению к (5.4):

$$P(F_n > F) = 1 - P(F_n \le F). \tag{5.5}$$

Вероятность $P(F_n > F)$ определяет долю структурных факторов в полуспектре, $N(F)/N$, с высотами, превышающими заданную величину F. Поэтому, сравнивая зависимости $N(F)/N$ от F для конкретных полуспектров с зависимостью для случайных аналогов (5.5), можно оценить близость к случайному распределению. Статистическую значимость отклонений от зависимости (5.5) можно оценить с помощью теста Колмогорова—Смирнова [24].

Выражения (5.2), (5.4) отвечают распределению вероятности для суммы трех независимых случайных величин с одинаковым рэлеевским распределением [24]. Нетрудно видеть, что это является прямым следствием определения (2.6), так как скалярное произведение $\boldsymbol{\rho}(q_n)\boldsymbol{\rho}^*(q_n)$ включает в себя сумму по трем компонентам в декартовой системе координат.

Статистическую значимость отдельных высоких пиков можно оценить, сравнивая их с выбросами в случайных полуспектрах [26]. Эта задача важна для поиска скрытых периодичностей и симметрий в конфигурациях общего вида (разд. 3). Вероятность того, что амплитуды всех структурных факторов в полуспектре не превысят величины F, равна

$$P(F_n \le F;\ N) = \left[P(F_n \le F)\right]^N, \tag{5.6}$$

где $P(F_n \le F)$ определяется (5.4), а вероятность того, что амплитуда хотя бы одного из N структурных факторов превысит F, является дополнительной к (5.6):

$$P(F_n > F;\ N) = 1 - \left[P(F_n \le F)\right]^N. \tag{5.7}$$

Пороговые значения для относительных амплитуд (5.3) при различных N, определяемые условиями $P(F_n > F;\ N) = 0.1$ и 0.05, приведены в таблице.

Интегральную регулярность конфигурации цепи удобно оценить с помощью спектральной энтропии (см. подробное обсуждение в [16]):

$$S = -\sum_{n=1}^{L-2}\left(\frac{F(q_n)}{\overline{F}}\right)\ln\left(\frac{F(q_n)}{\overline{F}}\right). \tag{5.8}$$

Поскольку высоты структурных факторов в случайных полуспектрах распределены относительно равномерно, величина S для случайных конфигураций с заданным расстоянием между концами цепи (или с заданным $\overline{F}$) достигает максимума (с точностью до

Таблица

Критерии для сингулярных гармоник при различных порогах значимости в полуспектре из N гармоник

N	f_{thr} $P = 0.1$	f_{thr} $P = 0.05$
50	10.33	11.20
100	11.17	12.02
150	11.65	12.50
200	11.99	12.84
250	12.25	13.10
300	12.47	13.31
350	12.65	13.49
400	12.80	13.64
450	12.94	13.78
500	13.06	13.90

относительно малых случайных отклонений). Усредняя (5.8) с помощью функции распределения (5.1), получим

$$\langle S\rangle_{\text{random}} = -\left[\ln(1/3) + \Gamma'(4)/\Gamma(4)\right](L-2) \simeq -0.1575\ldots(L-2), \qquad (5.9)$$

здесь Γ и Γ' — гамма-функция и ее производная. Аналитическая оценка для случайных отклонений $\langle(\Delta S)^2\rangle_{\text{random}}$ требует выхода за рамки приближения (5.1) и учета корреляций между структурными факторами с разными волновыми числами. В следующем разделе мы получим численную оценку для $\langle(\Delta S)^2\rangle_{\text{random}}$ на основании результатов численного моделирования. Считая отклонения от $\langle S\rangle_{\text{random}}$ гауссовыми, по этим данным нетрудно оценить статистическую значимость отношения $(\langle S\rangle_{\text{random}} - S)/\left[2\langle(\Delta S)^2\rangle_{\text{random}}\right]^{1/2}$ для спектральной энтропии S, отвечающей произвольной фиксированной конфигурации. В приложениях удобно также использовать величину $\Delta S_{\text{rel}} = \left(\langle S\rangle_{\text{random}} - S\right)/|\langle S\rangle_{\text{random}}|$, где $|\langle S\rangle|$ — абсолютная величина средней энтропии для случайных конфигураций [17]. Для случайных отклонений $\Delta S_{\text{rel}} \propto (L-2)^{-1/2}$ и мала для длинных случайных цепей с $L \gg 1$, в то время как для регулярных отклонений имеем $\left(\langle S\rangle_{\text{random}} - S\right) \propto (L-2)$. Поэтому отношение ΔS_{rel} можно использовать для сравнения интегральной степени регулярности для цепей с разным числом звеньев. Спектральная энтропия (5.8) отражает также и некоторые информационные характеристики, связанные с заданной конфигурацией цепи.

6. ПРИЛОЖЕНИЯ РЕЗУЛЬТАТОВ

6.1. *Цепи, генерируемые случайными блужданиями на кубической решетке*

Проиллюстрируем приложения результатов на двух конкретных примерах. В качестве первого примера рассмотрим цепи, генерируемые случайным блужданием на куби-

Рис. 2. Полуспектры для корреляционных функций (в переменных (6.1)) и структурных факторов (в переменных (5.3)) для цепи из $L-1=200$ звеньев, полученной в результате случайного блуждания на кубической решетке, и зависимость от f для доли структурных факторов $N(f)/N$, превышающих заданное значение f. Штриховая кривая отвечает уравнению (5.5)

ческой решетке. На рис. 2 приведены полуспектры для конкретной случайной реализации с числом звеньев $L-1=200$. Для структурных факторов используются безразмерные отношения (5.3), а для корреляционных функций (2.10) используются переменные

$$\kappa(m_0) = \big(K(m_0) - \overline{K}\big) \big/ \left[2\langle(\Delta K)^2\rangle\right]^{1/2}, \tag{6.1}$$

$$\langle(\Delta K)^2\rangle = \overline{F^2}/3(L-1). \tag{6.2}$$

Здесь $\overline{K}$ — среднее значение для $K(m_0)$ с $m_0 \neq 0$, а для вычисления среднего квадратичного отклонения для случайных конфигураций $\langle(\Delta K)^2\rangle$ было использовано правило сумм (2.13) и распределение (5.2). Для случайных конфигураций отношения $\kappa(m_0)$ с разными m_0 можно приближенно считать независимыми гауссовыми переменными. В нижней части рис. 2 приведен график зависимости от f для доли структурных факторов, $N(f)/N$, превышающих заданное значение f. Штриховая кривая отвечает уравнению (5.5). Для спектральной энтропии (5.8) получаем величину $S/(L-2) = -0.155$ в соответствии с предсказанием (5.9).

На рис. 3 представлены зависимости от $(L-2)$ для спектральной энтропии (5.8) и средних квадратичных отклонений, полученных в результате усреднения по 200 случайным реализациям с $L-1 = 100, 150, \ldots, 400$. Для средней энтропии и средних квадратичных отклонений получаем соответственно

$$\langle S\rangle = (-0.1567 \pm 0.0003)(L-2), \tag{6.3}$$

$$\langle(\Delta S)^2\rangle = (0.092 \pm 0.015)(L-2), \tag{6.4}$$

в хорошем согласии с (5.9), $\langle S\rangle_{\text{random}} \approx -0.1575(L-2)$.

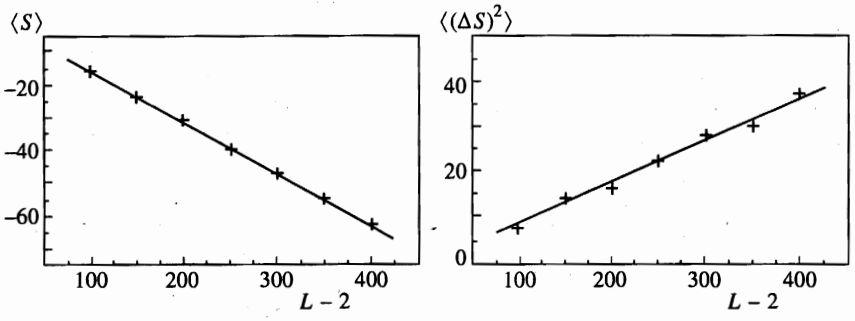

Рис. 3. Зависимости от $L-2$ для спектральной энтропии (5.8) и средних квадратичных отклонений, полученных в результате усреднения по 200 случайным реализациям при каждом значении L

6.2. *Структурные характеристики для C_α-остовов белков*

В качестве второго примера рассмотрим анализ структурных характеристик для C_α-остовов (по положению фиксированного атома углерода) белков. Напомним, что складывание трехмерной белковой структуры (или третичной структуры) осуществляется из так называемых элементов вторичной структуры [1, 3, 27, 28]. Элементы вторичной структуры универсальны и представляют собой либо фрагменты правых α-спиралей с периодом $p \approx 3.6$ (в единицах числа звеньев), либо приближенно плоские фрагменты β-элементов, соединенных между собой соединительными фрагментами, которые составляют отдельную группу элементов.

На рис. 4 приведены данные для представителей двух из четырех основных структурных классов белковых молекул [27, 28]. Для остовов молекул используется диаграммное представление [29], в котором α-элементам соответствуют ленточные спирали, β-элементам — ленты со стрелками, а соединительным элементам — фрагменты, подобные кускам проволоки. Полуспектры для корреляционных функций и структурных факторов приводятся в переменных (6.1) и (5.3). Буквенные обозначения соответствуют коду, под которым данная структура расположена в Брукхэйвенской базе данных [4].

Поскольку обе структуры содержат α-спирали, то, как видно из рис. 4, им соответствуют пики в спектрах для структурных факторов при значениях периодов $p = (L-1)/n \approx 3.6$, отвечающих скрытым периодичностям. Структура 256BA обладает приближенной симметрией второго порядка, что немедленно порождает высокий пик при $n = 2$ с амплитудой $f = 17.9$ (ср. таблицу в разд. 5). В свою очередь, структура 7TIMA обладает приближенной поворотной симметрией восьмого порядка, что порождает характерные расщепленные пики: $n = 7$, $f = 12.3$; $n = 8$, $f = 22.9$; $n = 9$, $f = 17.6$.

Отдельные высокие пики в спектрах для структурных факторов не всегда отвечают только лишь скрытым периодичностям и симметриям. Часть пиков при малых волновых числах q_n может отражать специфический дальний порядок, связанный с укладкой белковых структур [1, 27, 28], так как их складывание связано с выполнением ряда стерических и энергетических ограничений и имеет кооперативный характер.

Зависимость от f для доли числа структурных факторов, $N(f)/N$, превышающих заданное значение f, приведена на рис. 5 для структуры 7TIMA (график слева). Чтобы отделить влияние скрытых периодичностей и симметрий в этой структуре, мы отброси-

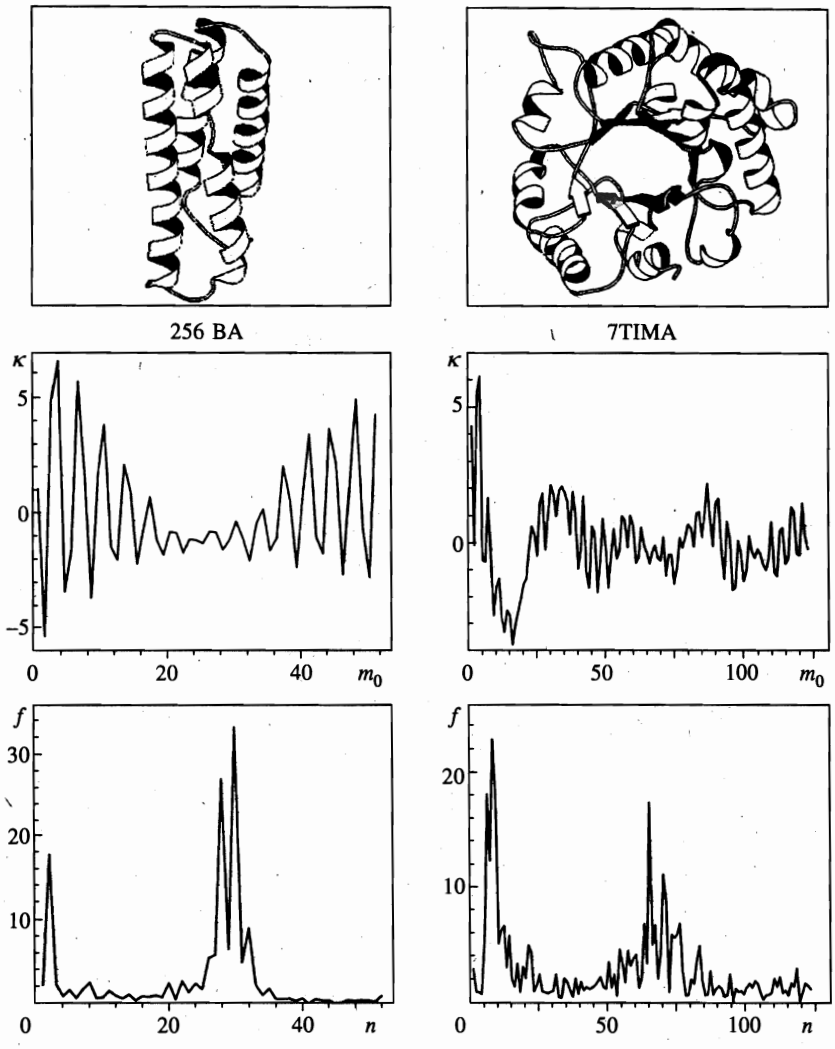

Рис. 4. Диаграммные представления структур [29] и полуспектры для нормированных корреляционных функций κ и структурных факторов f для цитохрома b562 (256BA) и триозофосфатизомеразы (7TIMA)

ли все высокие структурные факторы с $f \geq 10.0$ (превышающие типичные случайные выбросы, как следует из таблицы), а для оставшихся структурных факторов переопределили среднюю спектральную величину. Полученная зависимость уже достаточно неплохо описывается уравнением (5.5) (график справа на рис. 5). Согласие, естественно, улучшится для более нерегулярных структур. Как показывает анализ структур с отсутствием явно выраженных симметрий и относительно низкими значениями ΔS_{rel} (ср. ниже), зависимость (5.5) для них неплохо выполняется даже без предварительного отделения эффектов регулярности.

В силу высокой степени регулярности структур на рис. 4 отвечающие им значения

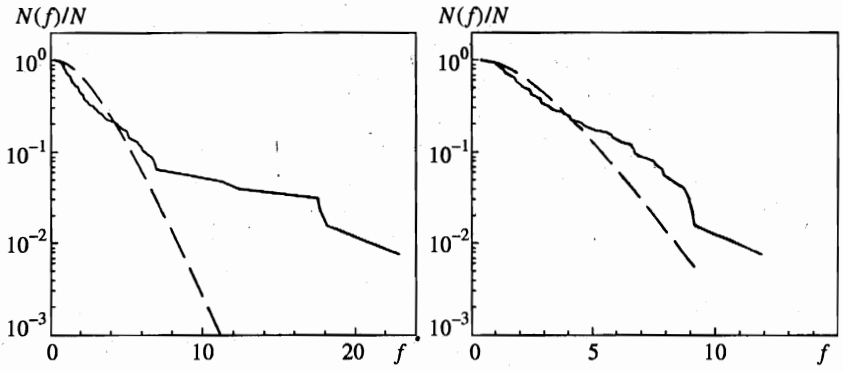

Рис. 5. Зависимости от f для доли структурных факторов, $N(f)/N$, превышающих заданное значение f, для триозофосфатизомеразы (7TIMA). График справа получен при отбрасывании гармоник с $f \geq 10$ с последующим переопределением средних спектральных значений (см. текст). Штриховые кривые отвечают уравнению (5.5)

$\Delta S_{\text{rel}} = \left(\langle S\rangle_{\text{random}} - S\right) / \left|\langle S\rangle_{\text{random}}\right|$ также оказываются большими: 5.470 (256BA); 1.996 (7TIMA). Как видим, значения ΔS_{rel} уменьшаются с повышением порядка симметрии и усложнением структуры.

Формированию третичной структуры белков из элементов вторичной структуры отвечает характерное убывание корреляционных функций на расстояниях $m_0 \sim$ 10–20 (см. рис. 4). Такое же поведение наблюдается и для корреляций других структурных и физико-химических характеристик вдоль белковой цепи [30, 31].

7. ЗАКЛЮЧЕНИЕ

Из полученных результатов следует, что спектральный подход позволяет достаточно просто и эффективно количественно исследовать структурные характеристики линейных цепей произвольной заданной формы. Метод позволяет выявить возможные скрытые периодичности и симметрии в системе, а оценка интегральной регулярности цепи в терминах спектральной энтропии (5.8) применима, уже начиная с относительно коротких цепей с $L \geq 100$. В схему просто включаются данные рентгеноструктурного анализа для полимерных цепей, собранные в имеющихся базах данных.

В качестве исходной случайной модели в работе использовалась свободно-сочлененная цепь. В следующем приближении эта модель должна учитывать эффекты исключенного объема [1–3]. К сожалению, их аналитическое рассмотрение довольно затруднительно и связано с громоздким численным счетом. Существенно, однако, что после отделения эффектов, связанных со скрытыми периодичностями и симметриями, статистика (5.2)–(5.5) приближенно описывает влияние нерегулярного структурного фона в реальных структурах (ср. рис. 5).

Формальная унификация цифрового представления данных и близость статистик для случайных аналогов делает метод удобным для изучения возможных корреляций и влияния физико-химических характеристик на структуру остовов полимерных молекул, что позволяет более детально оценить роль различных факторов в складывании

трехмерных структур.

В рамках спектрального подхода естественно учитываются структурные характеристики как на малых, так и на больших масштабах, и естественно выделяются эффекты дальнего порядка.

Уже для простейшего набора инвариантов типа структурных факторов (2.6) полное число независимых инвариантов $(L-1)/2$ оказывается сравнимым с полным числом степеней свободы $3L-6$. Поэтому с их помощью можно идентифицировать практически любую фиксированную структуру (с точностью до циклической перестановки фрагментов). Используя стандартную технику дополнения нулями [20], можно сравнивать наборы инвариантов для цепей с неравными длинами.

Таким образом, метод допускает достаточно полное исследование цепей с произвольной фиксированной формой.

Авторы выражают благодарность А. А. Веденову и участникам семинара под руководством М. А. Лифшица за обсуждение результатов работы и полезные замечания. Мы признательны также Ю. П. Лысову за любезную помощь в получении данных по белковым структурам.

Литература

1. П. Флори, *Статистическая механика цепных молекул*, Мир, Москва (1971).
2. П. Де Жен, *Идеи скейлинга в физике полимеров*, Мир, Москва (1982).
3. А. Ю. Гросберг, А. Р. Хохлов, *Статистическая физика макромолекул*, Наука, Москва (1989).
4. R. A. Laskowski, E. G. Hutchinson, A. D. Michie et al., Trends Biochem. Sci. **22**, 488 (1997).
5. C. A. Orengo, D. T. Jones, and J. M. Thornton, Nature **372**, 631 (1994).
6. L. Holm and C. Sander, Science **283**, 595 (1996).
7. T. J. P. Hubbard, A. G. Murzin, S. E. Brenner, and C. Chothia, Nucleic Acids Res. **25**, 236 (1997).
8. P. Biswas, A. Paramekanti, and B. J. Cherayil, J. Chem. Phys. **104**, 3360 (1996).
9. X. Yi, Phys. Rev. E **49**, 5903 (1994).
10. G. A. Arteca, Phys. Rev. E **54**, 3044 (1996).
11. G. Jumarie, Physica A **184**, 499 (1992).
12. T. G. Dewey, Phys. Rev. E **54**, R39 (1996).
13. A. Fernandez and A. Belinky, J. Phys. A: Math. Gen. **29**, L433 (1996).
14. L. F. Luo, *Collected Works on Theoretical Biophysics*, Inner Mongolia University Press, Hohhot (1997).
15. К. Шеннон, *Работы по теории информации и кибернетике*, ИИЛ, Москва (1963).
16. А. Ю. Турыгин, В. Р. Чечеткин, ЖЭТФ **106**, 335 (1994). V. R. Chechetkin and A. Y. Turygin, J. Phys. A: Math. Gen. **27**, 4875 (1994).
17. V. R. Chechetkin, L. A. Knizhnikova, and A. Y. Turygin, J. Biomol. Struct. Dyn. **12**, 271 (1994). V. R. Chechetkin and V. V. Lobzin, Phys. Lett. A **222**, 354 (1996).
18. V. R. Chechetkin and A. Y. Turygin, J. Theor. Biol. **175**, 477 (1995). V. R. Chechetkin and V. V. Lobzin, J. Biomol. Struct. Dyn. **15**, 937 (1998).
19. V. R. Chechetkin and A. Y. Turygin, J. Theor. Biol. **178**, 205 (1996). V. R. Chechetkin and V. V. Lobzin, J. Theor. Biol. **190**, 69 (1998).
20. С. Л. Марпл, *Цифровой спектральный анализ и его приложения*, Мир, Москва (1990).
21. М. Хамермеш, *Теория групп и ее применение к физическим проблемам*, Мир, Москва (1966).
22. Р. Фларри, *Группы симметрии. Теория и химические приложения*, Мир, Москва (1983).
23. V. R. Chechetkin and V. V. Lobzin, Phys. Lett. A **250**, 443 (1998).

24. В. Феллер, *Введение в теорию вероятностей и ее приложения*, Мир, Москва (1984).
25. Н. Г. Ван Кампен, *Стохастические процессы в физике и химии*, Высшая школа, Москва (1990).
26. M. R. Leadbetter, G. Lindgren, and H. Rootzen, *Extremes and Related Properties of Random Sequences and Processes*, Springer, Berlin (1983).
27. T. Creighton, *Proteins: Structures and Molecular Properties*, Freeman, New York (1993).
28. В. М. Степанов, *Структура и функции белков*, Высшая школа, Москва (1996).
29. P. Kraulis, J. Appl. Crystallogr. **24**, 946 (1991).
30. N. D. Socci, W. S. Bialek, and J. N. Onuchic, Phys. Rev. E **49**, 3440 (1994).
31. O. Weiss and H. Herzel, Zs. Phys. Chemie **204**, 183 (1998).