Январь 2000 г. **Том 170, № 1**

УСПЕХИ ФИЗИЧЕСКИХ НАУК

ОБЗОРЫ АКТУАЛЬНЫХ ПРОБЛЕМ

Порядок и корреляции в геномных последовательностях ДНК. Спектральный подход

В.В. Лобзин, В.Р. Чечеткин

В рамках спектрального подхода обсуждается проблема структурного анализа геномных последовательностей ДНК. Взаимная однозначность и взаимная дополнительность масштабов при преобразовании Фурье делают такой подход достаточно универсальным. Спектральные характеристики для случайных последовательностей с тем же нуклеотидным составом обладают свойством самоусредняемости для сравнительно коротких последовательностей длиной $M \geqslant 100-300$. Сравнение с характеристиками для случайных последовательностей определяет статистическую значимость структурных особенностей, наблюдаемых в геномных последовательностях ДНК. Кроме традиционных приложений для выявления скрытых периодичностей, спектральные методы можно эффективно применять для анализа взаимных корреляций в последовательностях ДНК. Совместное использование спектров для структурных факторов и корреляционных функций позволяет оценить не только интегральные корреляции, но и выяснить их источник. Общую количественную оценку для степени регулярности последовательности можно получить с помощью структурной спектральной энтропии. Обзор содержит также краткое введение в проблему и описание других основных методов анализа последовательностей ДНК.

PACS numbers: **02.50.–r**, **05.10.–a**, **87.10.+e**, 87.14.Gg

Содержание

В.В. Лобзин. Институт земного магнетизма, ионосферы и распространения радиоволн РАН,
142092 Троицк, Московская обл., Российская Федерация
Тел. (095) 334-01-13. Факс (095) 334-01-24
E-mail: lobzin@top.izmiran.troitsk.ru
В.Р. Чечеткин. Троицкий институт инновационных и термоядерных исследований,
142092 Троицк, Московская обл., Российская Федерация
Тел. (095) 334-50-57

Статья поступила 1 июня 1999 г., после доработки 30 июля 1999 г.

1. Введение

Быстрый рост числа расшифрованных геномных последовательностей ДНК и увеличение быстродействия и емкости памяти современных компьютеров, — пожалуй, две наиболее характерные черты науки конца XX века. На март 1999 г. банки данных насчитывали $3{,}3 \times 10^6$ последовательностей, содержащих в общей сложности $2{,}4 \times 10^9$ нуклеотидов. Предполагается, что тенденция к удвоению данных примерно каждые год-полтора сохранится и в ближайшем будущем (рис. 1). К 2003 г. планируется завершение расшифровки полной геномной последовательности ДНК человека, состоящей из $\sim 3 \times 10^9$ нуклеотидов. Обработка этой огромной информации требует совместных усилий не только генетиков, биологов, биохимиков, медиков, но и математиков, специалистов-компьютерщиков. С начала 90-х годов в этот процесс включились также и физики, и статьи по данной тематике стали регулярно печататься на страницах ведущих физических журналов. Стоит отметить, что лаборатории в Лос-Аламосе и Ливерморе, например, занимают ведущие позиции не только в области физических исследований, но и в компьютерном анализе последовательностей ДНК.

Информация, заключенная в геномной ДНК, должна воспроизводиться, распознаваться, считываться и слу-

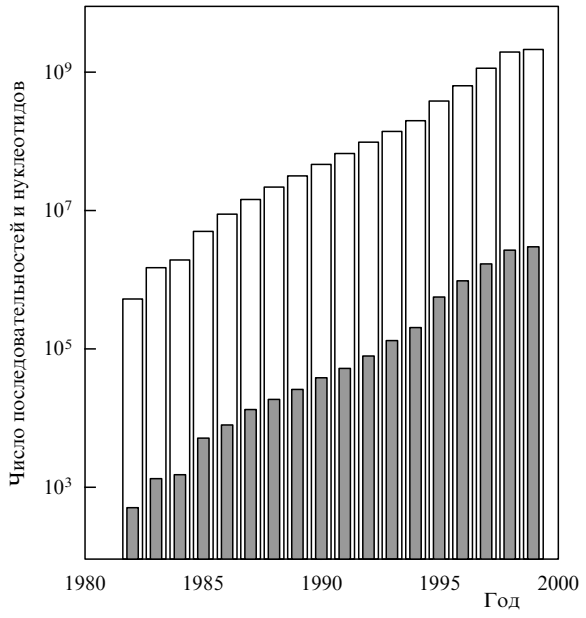

Рис. 1. Рост числа расшифрованных последовательностей ДНК (закрашенные столбики) и нуклеотидов в них.

жить своеобразной программой для запуска самых разнообразных молекулярных механизмов. Поэтому функции различных участков ДНК также могут быть существенно различными. Одна из стандартных задач компьютерного анализа последовательностей ДНК состоит в том, чтобы выделить воспроизводимые структурные особенности и связать их с соответствующими функциями. Многообразие путей молекулярной эволюции и различия в условиях существования живых организмов приводят к большому разнообразию функционально-структурных связей. Отбор и фиксация таких связей в ходе молекулярной эволюции позволяют использовать системный подход, основанный на сравнительном анализе последовательностей в банках данных.

Начнем с описания баз данных, тем более, что практически все данные (за исключением результатов некоторых исследований, выполненных по заказам фармацевтических корпораций) свободно доступны через международную электронную сеть Интернет. Основные данные по геномным последовательностям ДНК сосредоточены в "GenBank" (Лос-Аламос, США) и базах данных Европейской молекулярно-биологической лаборатории (ЕМБЛ) (Гейдельберг, Германия) и Европейского института биоинформатики (филиал ЕМБЛ в Кембридже, Англия). В сводных базах данных каждая последовательность ДНК помещена под своим кодом доступа (accession number). Кроме сводных баз данных, имеется свыше пятидесяти специализированных банков данных. Их подробное описание и электронные адреса можно найти в [1 – 3].

Литература на русском языке, которую можно использовать в качестве введения в предмет, сравнительно немногочисленна [4 – 7] и отвечает ситуации до 1990 г. Из большого потока относительно недавних публикаций можно рекомендовать [8 – 15]. Текущую сводку публикаций, начиная с 1992 г., представляющую интерес для физической аудитории, можно найти по электронному адресу [16].

Обзор посвящен, главным образом, спектральным методам структурного анализа последовательностей ДНК, что определяется кругом наших собственных исследований [17 – 27]. Тем не менее в разделе 2 мы постарались хотя бы кратко перечислить и другие основные методы анализа последовательностей ДНК. Общая схема спектрального анализа дается в разделе 3. В качестве приложений общей теории рассматривается выделение скрытых периодичностей (раздел 4) и изучение корреляций (раздел 5) в последовательностях ДНК. Разнообразие путей молекулярной эволюции приводит к значительным структурным различиям даже в функционально близких сегментах ДНК у различных организмов. Для иллюстрации мы отобрали сравнительно немного примеров, но попытались выделить воспроизводимые и до некоторой степени универсальные структурные особенности.

Прежде чем перейти к основному изложению, постараемся ответить на следующие вопросы: почему используется именно спектральный подход? в чем его достоинства и недостатки? Поскольку преобразование Фурье полностью обратимо, то с чисто формальной точки зрения речь идет о взаимно-однозначном отображении информации с помощью преобразования Фурье. Взаимная дополнительность масштабов при преобразовании Фурье позволяет перекрыть весь набор характерных длин. Такой подход позволяет также унифицировать описание характеристик не только последовательностей символов, но и пространственных линейных цепей (см. заключение). Поэтому в рамках спектрального подхода удается несколько дальше продвинуться в изучении связей между последовательностями ДНК и структурой белков.

Изучение структурных связей, как правило, неотделимо от проблем молекулярной эволюции. В [28] был рассмотрен сценарий блочно-иерархического объединения и отбора структурных единиц в ходе молекулярной эволюции. Спектральный структурный анализ позволяет в значительной мере детализировать подобный сценарий.

Статистическая значимость наблюдаемых регулярных связей нередко оказывается невысокой. Их проверка часто требует сочетания различных методов (см. заключение). Поэтому, несмотря на относительную универсальность спектрального подхода, на практике нельзя ограничиваться только им одним, а приходится использовать гораздо более широкий арсенал средств анализа (частично описанный в разделе 2).

2. Геномные последовательности ДНК и их анализ

С физической точки зрения наибольший интерес представляет изучение взаимных корреляций, периодичностей и эффектов соизмеримости – несоизмеримости в последовательностях ДНК. Для интерпретации этих эффектов необходимо представление о молекулярной структуре и молекулярно-биологических процессах, в которых участвует ДНК.

2.1. Краткий молекулярно-биологический обзор

Генетическая информация о развитии живых организмов заключена в длинных биополимерных макромолекулах — нуклеиновых кислотах [29 – 31]. Структурными

субъединицами нуклеиновых кислот являются нуклеотиды, сложные молекулярные образования, состоящие из азотсодержащего основания, пятиуглеродного сахара и фосфатной группы. Согласно двум видам сахаров различают дезоксирибонуклеиновую кислоту (ДНК) и рибонуклеиновую кислоту (РНК). В состав ДНК входят основания четырех видов: аденин (A), гуанин (G), цитозин (C) и тимин (T). В РНК тимин заменен урацилом (U). В пентозных кольцах сахаров положения углеродов нумеруются некоторым стандартным образом. При объединении в полимер происходит образование ковалентной связи между $3'$-группой остатка сахара одного нуклеотида с $5'$-группой другого. Поэтому начало нуклеотидной последовательности называют также $5'$-концом, а ее окончание — $3'$-концом. За исключением некоторых вирусов носителем генетической информации в организмах служит двухцепочечная ДНК, в которой аденин спарен двумя водородными связями с тимином, а гуанин спарен тремя водородными связями с цитозином.

При анализе различные нуклеотиды часто объединяются в бинарные группы: пурины $\mathrm{R} = (\mathrm{A}, \mathrm{G})$ и пиримидины $\mathrm{Y} = (\mathrm{C}, \mathrm{T}$ или $\mathrm{U})$, а также группы, различающиеся по силе связей, $\mathrm{W} = (\mathrm{A}, \mathrm{T})$ и $\mathrm{S} = (\mathrm{C}, \mathrm{G})$, и по физико-химическим свойствам, так называемое кето-аминное разбиение, $\mathrm{K} = (\mathrm{G}, \mathrm{T})$ и $\mathrm{M} = (\mathrm{A}, \mathrm{C})$.

Двухцепочечная ДНК может существовать в правоспиральных A-форме (шаг спирали $l \simeq 10{,}8$ пар, далее l везде измеряется в парах нуклеотидов для двухцепочечных молекул и в числах нуклеотидов для одноцепочечных молекул) и B-форме ($l \simeq 10{,}2-10{,}5$), а также в левоспиральной Z-форме ($l \simeq 12{,}0$), для которой характерно чередование пуринов и пиримидинов, RYRY... Стандартной формой является B-ДНК, однако локальные динамические переходы B – A и B – Z при определенных условиях (в зависимости от концентрации солей, уровня сверхспирализации и т.д.) могут происходить и в геномной ДНК.

Различают прокариотические клетки и стоящие на более высокой ступени эволюции эукариотические клетки. В отличие от прокариот эукариоты обладают оформленным клеточным ядром (по-гречески "карион"). Компактная упаковка ДНК в ядре достигается с помощью специальной иерархической укладки. На первом уровне компактизации ДНК эукариот собирается в нуклеосому, в которой 146 пар оснований намотаны вокруг молекулярного комплекса из белков-гистонов. Между закрепленными таким образом фрагментами имеется соединительная (линкерная) ДНК. Длина ДНК в нуклеосоме принимает значения в пределах $l = 200 \pm 40$. Затем нуклеосомы собираются в 30 нм фибриллу со структурой типа соленоида. На один шаг соленоида приходится примерно шесть нуклеосом. На третьем уровне компактизации фибрилла собирается в петли. Длина петель изменяется в широких пределах, $l \simeq 2 \times 10^4 - 10^5$. Наконец, на четвертом уровне компактизации происходит укладка петель. При квазирегулярной укладке можно выделить единицы, охватывающие примерно 10 петель.

Изгибная жесткость двухцепочечной ДНК характеризуется длиной Куна $l_{\text{Кун}}$ [32]. В типичных физиологических условиях $l_{\text{Кун}} \simeq 300-340$. Иногда вместо нее используют персистентную длину, $l_{\text{перс}} = 0{,}5 l_{\text{Кун}}$. Как видим, для средней длины ДНК в нуклеосоме $l_{\text{нукл}}$ выполняется условие $l_{\text{перс}} < l_{\text{нукл}} < l_{\text{Кун}}$. Длина Куна для одноцепочечной ДНК много меньше и равна примерно $l_{\text{Кун}} \simeq 12-14$.

У прокариот значительная часть ДНК связана с кодированием белков. При этом сначала по кодирующей ДНК снимается РНК-копия (матричная РНК или мРНК). Этот процесс называется транскрипцией. Затем по мРНК с помощью универсального генетического кода осуществляется синтез белков. Этот процесс называется трансляцией. В процессе трансляции триплетам нуклеотидов (или кодонам) сопоставляется одна аминокислота, а последовательность аминокислот образует белок. Обычно только одна из цепей ДНК является белок-кодирующей. Начало трансляции определяется инициирующим кодоном (как правило, ATG) и специфически узнаваемой областью, предшествующей инициирующему кодону. Конец трансляции определяется терминирующими кодонами (TAA, TAG или TGA). Поскольку имеется всего 20 различных аминокислот, то генетический код вырожден.

Трехмерная структура белков также имеет ярко выраженный иерархический характер (см., например, [33]). С первичным уровнем (или первичной структурой) отождествляется обычно сама последовательность аминокислот. Вторичная структура белка характеризует пространственное расположение атомов главной цепи молекулы на отдельных ее участках, а третичной структурой называется характерный для данного белка способ укладки всей цепи в пространстве [33, 34]. Основными элементами вторичной структуры являются α-спирали и приближенно плоские элементы β-структуры [32 – 34]. На один шаг α-спирали приходится примерно $3{,}6 \pm 0{,}2$ аминокислот, а для β-структур характерно чередование водородных связей. Элементы вторичной структуры состоят, как правило, из $\sim 8-20$ аминокислот и связываются между собой соединительными фрагментами. Полное число аминокислот в белковых молекулах $\sim 70-2000$ (в среднем $\sim 200-300$).

В геномах эукариот различают три класса последовательностей ДНК [29, 31, 35]: сателлитную ДНК, состоящую из многократно повторяющихся повторов, умеренно повторяющиеся последовательности, рассеянные по геному, и уникальные последовательности. Сателлитная ДНК отвечает за правильную структурную организацию хромосом. Умеренно повторяющиеся последовательности играют отчасти структурную, а отчасти регуляторную роль. Некоторые семейства повторов настолько специфичны, что их можно использовать в качестве "генетических отпечатков пальцев" [6]. Часть уникальной ДНК связана с кодированием белков. Для млекопитающих ее доля относительно невелика, $\sim 2-3\,\%$. Особенностью белок-кодирующей ДНК эукариот является ее прерывистый характер: кодирующие участки (экзоны) прерываются некодирующими (интронами). Соответствующая РНК-копия сначала собирается в рибонуклеопротеиновые частицы (РНП-частицы), которые состоят из фрагментов длиной ~ 600 нуклеотидов, намотанных на белковый кор, плюс соединительные фрагменты из ~ 100 нуклеотидов. Затем интроны вырезаются, а экзоны соединяются в мРНК. Этот процесс называется сплайсингом. Взаимное содержание ДНК в трех классах последовательностей для различных организмов меняется в широких пределах, $\sim 10-50\,\%$.

Таблица 1. Характерные длины в геномных последовательностях ДНК (в пáрах нуклеотидов для двухцепочечных молекул и в числах нуклеотидов для одноцепочечных молекул)

Механизм	Длина, l
Строение спирали ДНК	
Чередование пуринов и пиримидинов в Z-форме	2
Шаг спирали в B-форме	10,2 – 10,5
Шаг спирали в A-форме	10,8
Шаг спирали в Z-форме	12,0
Упаковка хроматина	
Нуклеосома	200 ± 40
Шаг спирали в 30 нм фибрилле	$(200 \pm 40) \times 6$
Петли	$2 \times 10^4 - 10^5$
Субъединица из петель	~ 10 петель
Изгибные характеристики	
Длина Куна для двухцепочечной ДНК	300 – 400
Длина Куна для одноцепочечной ДНК	12 – 14
Структура белков	
Кодон (отвечает одной аминокислоте)	3
Шаг α-спирали	$(3{,}6 \pm 0{,}2) \times 3$
Чередование водородных связей в элементах β-структуры	$2{,}0 \times 3$
Длины элементов вторичной структуры	$(8-20) \times 3$
Длины белков	$(70-2000) \times 3$
Прерывистая структура белок-кодирующих областей у эукариот	
Экзоны	50 – 400
Интроны	$50-5 \times 10^4$
Упаковка гетероядерной РНК в РНП-частицы	$600 + 100$
Полные длины геномов	
ДНК-вирусы	$5 \times 10^3 - 5 \times 10^5$
Бактерии	$7 \times 10^5 - 10^7$
Растения	$10^8 - 10^{11}$
Амфибии	$6 \times 10^8 - 10^{11}$
Птицы	$7 \times 10^8 - 2 \times 10^9$
Млекопитающие	$(2-3) \times 10^9$

Динамически связываясь с ДНК (часто кооперативно), регуляторные белки влияют на характер упаковки хроматина и "включают" (или "выключают") различные гены [29 – 31]. Длины контактов $l_{\text{контакт}} \sim 5-100$.

Часть характерных длин в геномах эукариот связана также с другими динамическими процессами: сайт-специфической рекомбинацией, когда фрагменты ДНК встраиваются и выщепляются из генома, и кроссинговером, когда происходит обмен участков между гомологичными хромосомами [29]. Эти длины принимают значения в очень широких пределах и требуют отдельного рассмотрения.

Полная сводка характерных длин представлена в табл. 1. Часть характерных длин весьма консервативна, а часть изменяется в довольно широких пределах. Существенно, что особенности молекулярного строения ДНК, механизмы ее упаковки и процессы с ее участием так или иначе отражаются на уровне последовательностей ДНК [36, 37]. Различия в структуре и физико-химических характеристиках разных оснований приводят к энергетической и функциональной неэквивалентности двухцепочечных молекул ДНК, которым отвечают разные последовательности нуклеотидов. Те из них, которые обладают преимуществами в энергетическом и/или функциональном отношении, в свою очередь получают преимущество в процессе эволюционного отбора. Например, регулярное относительно шага спирали включение фрагментов из нескольких аденинов приводит к изгибу двухцепочечной ДНК [38 – 40]. Такую ДНК было бы проще упаковывать в нуклеосому. С другой стороны, можно было бы ожидать, что квазирегулярная упаковка в нуклеосомы проявится в виде скрытых периодичностей в последовательностях ДНК, что действительно имеет место [26].

Функциональные различия для разных сегментов ДНК приводят к характерной неоднородности состава и структурных характеристик вдоль последовательности ДНК (далее мы будем называть их "эффектами сегментации"). На первом этапе задача состоит в разделении различных элементов такой мозаичной структуры. На следующем этапе необходимо выяснить, есть ли связь между различными элементами мозаичной структуры, и если есть, то какие молекулярные механизмы ответственны за нее. В разделе 2.2 мы кратко опишем основные методы анализа последовательностей ДНК.

2.2. Методы статистического анализа геномных последовательностей ДНК

Прежде всего условимся о терминологии. Длины $l < 10$ будем считать малыми, длины $10 \leqslant l < 10^3$ — промежуточными, а длины $l \geqslant 10^3$ — большими. Длина $l \sim 10$ — это шаг двойной спирали, $l \sim 10^3$ — средний размер гена, эта длина охватывает несколько нуклеосом и т.д. (см. табл. 1). Естественное подразделение масштабов задает также полная длина геномной последовательности. Такое разбиение будет использовано ниже в разделе 5. Термины "порядок" и "упорядоченность" используются в традиционном для физики смысле: они отражают наличие статистически значимых корреляций на данных расстояниях. В этом смысле "дальний порядок" охватывает и крупномасштабные вариации плотности, и незатухающие периодические осцилляции, и тандемное повторение длинной случайной последовательности.

Простейший подход к статистическому анализу последовательностей ДНК состоит в подсчете чисел появления или частот различных комбинаций нуклеотидов, $\{\mathrm{N}_1 \ldots \mathrm{N}_l\}$, $\mathrm{N} \in (\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T})$, при разбиениях последовательности на фрагменты длины l [5 – 9, 41 – 45] (в теории информации такие комбинации называются l-граммами [46]). При локальном анализе частоты подсчитываются в окне шириной W, перемещаемом вдоль последовательности, и обрабатываются с помощью статистических таблиц. Поскольку число различных l-грамм из четырех нуклеотидов равно 4^l и экспоненциально растет с длиной l, то в приложениях обычно l невелико, $l \leqslant 6$ ($4^6 = 4096$).

Если числа появления комбинаций нуклеотидов длины l разделить на полное число фрагментов разбиения (для последовательности длиной M это число равно $M - l + 1$), то таким частотам можно придать смысл вероятностей $P(\mathrm{N}_1 \ldots \mathrm{N}_l)$ и проанализировать их в рамках формализма марковских цепей [5 – 7, 47 – 50], используя известную формулу теории вероятностей [51]:

$$P(\mathrm{N}_1 \ldots \mathrm{N}_{l-1} | \mathrm{N}_l) = \frac{P(\mathrm{N}_1 \ldots \mathrm{N}_l)}{P(\mathrm{N}_1 \ldots \mathrm{N}_{l-1})}\,, \tag{2.1}$$

где $P(\mathrm{N}_1 \ldots \mathrm{N}_{l-1} | \mathrm{N}_l)$ — условная вероятность того, что после нуклеотидов $\mathrm{N}_1 \ldots \mathrm{N}_{l-1}$ следует N_l, а $P(\mathrm{N}_1 \ldots \mathrm{N}_l)$ и $P(\mathrm{N}_1 \ldots \mathrm{N}_{l-1})$ — вероятности комбинаций $\mathrm{N}_1 \ldots \mathrm{N}_l$ и

$\mathrm{N}_1 \ldots \mathrm{N}_{l-1}$ соответственно. Если существует такое число r, что при $l-1 \geqslant r$ вероятность (2.1) не зависит от r, то можно говорить о марковской цепи порядка r.

В рамках теории информации [46] вероятностям $P(\mathrm{N}_1 \ldots \mathrm{N}_l)$ можно сопоставить энтропию порядка l

$$H_l = -\sum_{\{\mathrm{N}_1 \ldots \mathrm{N}_l\}} P(\mathrm{N}_1 \ldots \mathrm{N}_l) \log_2 P(\mathrm{N}_1 \ldots \mathrm{N}_l)\,, \tag{2.2}$$

где суммирование выполняется по всем l-граммам, удельную энтропию

$$G_l = \frac{H_l}{l} \tag{2.3}$$

и избыточность

$$R_l = 1 - \frac{H_l}{2l}\,. \tag{2.4}$$

Величина R_l характеризует отклонения от полностью случайных последовательностей с равновероятной встречаемостью всех нуклеотидов. Расчетам этих величин посвящены работы [5, 52–57].

Как марковский подход, так и рассмотрение на основе теории информации требуют выполнения условия $M - l + 1 \gg 4^l$ и неявно предполагают статистическую однородность вдоль последовательности. В то же время в реальных геномных последовательностях четко видны эффекты сегментации уже для $M \geqslant 10^2 - 10^3$, что ограничивает область применимости этих подходов сравнительно малыми l (см. также критический разбор в [54, 56]). Поэтому данные методы анализа выявляют корреляции лишь на малых расстояниях.

Корреляции на промежуточных расстояниях можно проанализировать с помощью корреляционных функций [17, 18, 23, 58, 59] или функций взаимной информации [60, 61]. Эти функции вводятся далее в разделе 3, поэтому здесь мы не будем на них останавливаться.

При изучении крупномасштабных вариаций плотности наиболее популярным является бинарное разбиение на пурины $\mathrm{R} = (\mathrm{A}, \mathrm{G})$ и пиримидины $\mathrm{Y} = (\mathrm{C}, \mathrm{T})$. После этого исходная последовательность ДНК разбивается на сегменты длиной l и вычисляется среднеквадратичное отклонение между числами пуринов и пиримидинов $\langle [N_{\mathrm{R}}(l) - N_{\mathrm{Y}}(l)]^2 \rangle^{1/2}$ в зависимости от l [62–71]. Если сопоставить пуринам число $+1$, а пиримидинам — число -1, то такую модель можно отобразить на модель случайных блужданий. В [62] была высказана гипотеза о фрактальном характере подобных случайных блужданий. В этом случае $\langle [N_{\mathrm{R}}(l) - N_{\mathrm{Y}}(l)]^2 \rangle^{1/2} \propto l^{\alpha}$ с некоторым постоянным показателем α, отличным от 0,5. В настоящее время эта гипотеза отвергнута подавляющим большинством исследователей, поскольку было показано, что α также зависит от l [19, 63–65]. Существенно, что α сильно меняется при разбиении исходной последовательности на несколько крупных сегментов и вычислении α для отдельных сегментов [65, 69]. В [66] было показано, что в экзон-интронных последовательностях эффективное значение α слабо зависит от случайных перестановок интронов. Поэтому даже отклонение α от 0,5 оставляет открытым вопрос о характере дальних корреляций. В [64–66] было предложено объяснение в рамках крупномасштабной мозаичной структуры с отсутствием корреляций между элементами мозаики. В [20, 27] было, однако, показано, что нетривиальная структурная связь между элементами все-таки существует (см. также раздел 5). Крупномасштабную сегментацию в последовательностях ДНК можно изучать, используя технику с движущимися окнами [5] или вейвлетное преобразование [72–76].

Окраска хромосомной ДНК красителями, чувствительными к $\mathrm{S} = (\mathrm{G}, \mathrm{C})$ или $\mathrm{W} = (\mathrm{A}, \mathrm{T})$, обнаруживает крупномасштабные вариации состава на расстояниях $L \sim 10^5 - 10^6$ [29, 77]. В [78] для описания подобных эффектов была предложена модель скрытых случайных блужданий. В такой модели сначала выбирается случайная величина w, которая меняется в пределах $1/3 \leqslant w \leqslant 2/3$. Пусть в начальный момент времени случайно выбрано некоторое начальное значение w_0. Далее с вероятностью 1/2 рассматриваются случайные изменения w_0 на $w_0 - \Delta w$ или $w_0 + \Delta w$. На границах значение w остается постоянным. Последовательные случайные изменения w_0 определяют значение w_m на шаге m, которое, в свою очередь, определяет вероятность $P(w_m)$ случайного заполнения узла m нуклеотидом W. Функция $P(w)$ и инкремент Δw подгоняются в соответствии с наблюдаемыми данными. Если начальное значение w_0 лежит в интервале $1/2 < w_0 < 2/3$, то в течение примерно $L \sim [(w_0 - 1/2)/\Delta w]^2$ шагов будет сохраняться повышенная вероятность заполнения узлов нуклеотидами W. При переходе в область $1/3 < w_0 < 1/2$ ситуация изменится на противоположную. Данная модель задает средние размеры областей с повышенным содержанием W или S, $\langle L \rangle \approx [(1/3 - 1/2)/\Delta w]^2$, однако вариации плотности внутри различных областей с размерами $\sim L$ оказываются некоррелированными. Детальное исследование таких корреляций пока не проводилось.

Как видим, изучение корреляций на разных масштабах осуществляется совершенно различными методами. Поэтому возникает вопрос о разработке некоторой единой техники анализа на всех масштабах. В следующих разделах мы покажем, что такого рода программу можно реализовать в рамках спектрального подхода.

3. Спектральный анализ последовательностей ДНК

Спектральные методы анализа последовательностей ДНК традиционно применяются для выявления скрытых периодичностей [21, 26, 79–85], для изучения корреляций между различными последовательностями [23, 27, 86–89] и для исследования дальних корреляций [19, 27, 60, 90]. Общее введение в эти методы можно найти в [91]. Использование преобразования Фурье позволяет получить статистические критерии, которые обладают свойством самоусредняемости уже для сравнительно коротких последовательностей длиной $M \geqslant 100 - 200$. Как раз начиная с этих длин и происходит сегментация в геномных последовательностях ДНК (см. табл. 1). Поэтому в рамках такой техники можно изучать как отдельные элементы мозаичной структуры, так и связь между ними. Стратегия состоит в том, чтобы сравнивать наблюдаемые характеристики для реальных последовательностей с характеристиками для случайных последовательностей с тем же составом нуклеотидов. Последнее важно для выделения эффектов сегментации, а также для отделения эффектов вариации нуклеотидного состава, связанных с эволюцией и различием в условиях функционирования живых организмов. Как будет показано ниже, на этом пути удается получить достаточно удобные и

универсальные критерии структурной регулярности, не зависящие от состава, что позволяет сравнивать структурные характеристики последовательностей ДНК с разным составом нуклеотидов. Далее в этом разделе излагаются, в основном, результаты работ [17, 18].

3.1. Общая теория

Рассмотрим последовательность ДНК длиной M. Ее можно задать с помощью функции положения:

$$\rho_{m,\alpha} = \begin{cases} 1, & \text{если нуклеотид типа } \alpha \text{ занимает сайт } m\,, \\ 0 & \text{в другом случае}\,, \end{cases} \tag{3.1}$$

где $\alpha \in (\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T})$, $m = 1, \ldots, M$. Фурье-гармоники, соответствующие нуклеотидам типа α, определяются согласно

$$\rho_\alpha(q_n) = M^{-1/2} \sum_{m=1}^{M} \rho_{m,\alpha} \exp(-\mathrm{i} q_n m)\,,$$

$$q_n = \frac{2\pi n}{M}\,, \qquad n = 0, 1, \ldots, M-1\,, \tag{3.2}$$

а обратное преобразование имеет вид

$$\rho_{m,\alpha} = M^{-1/2} \sum_{n=0}^{M-1} \rho_\alpha(q_n) \exp(\mathrm{i} q_n m)\,, \qquad m = 1, \ldots, M\,. \tag{3.3}$$

Нулевая фурье-гармоника не содержит информации о распределении нуклеотидов и определяется лишь их общим числом N_α:

$$\rho_\alpha(0) = \frac{N_\alpha}{M^{1/2}}\,. \tag{3.4}$$

Ниже основные характеристики выражаются с помощью элементов матричного структурного фактора

$$F_{\alpha\beta}(q_n) = \rho_\alpha(q_n)\, \rho^*_\beta(q_n) \tag{3.5}$$

(здесь и далее звездочка означает комплексное сопряжение). С помощью соотношения Винера–Хинчина элементам структурного фактора $F_{\alpha\beta}(q_n)$ можно поставить в соответствие парные корреляционные функции

$$K_{\alpha\beta}(m_0) = M^{-1} \sum_{n=0}^{M-1} F_{\alpha\beta}(q_n) \exp(-\mathrm{i} q_n m_0)\,,$$

$$m_0 = 0, \ldots, M-1\,. \tag{3.6}$$

Используя определения (3.2) и (3.3), их можно представить также в виде

$$K_{\alpha\beta}(m_0) = M^{-1} \sum_{m=1}^{M} \rho^{\mathrm{c}}_{m,\alpha}\, \rho^{\mathrm{c}}_{m+m_0,\beta}\,, \tag{3.7}$$

где

$$\rho^{\mathrm{c}}_{m,\alpha} = \begin{cases} \rho_{m,\alpha}\,, & \text{если } 1 \leqslant m \leqslant M\,, \\ \rho_{m-M,\alpha}\,, & \text{если } M+1 \leqslant m \leqslant 2M-1\,. \end{cases} \tag{3.8}$$

Вещественность функций $\rho_{m,\alpha}$ приводит к условию

$$\rho^*_\alpha(q_n) = \rho_\alpha(2\pi - q_n) \tag{3.9}$$

для фурье-гармоник, что, в свою очередь, накладывает следующие условия симметрии на элементы структурного фактора и корреляционные функции:

$$F_{\alpha\beta}(q_n) = F_{\beta\alpha}(2\pi - q_n)\,,$$

$$K_{\alpha\beta}(m_0) = K_{\beta\alpha}(M - m_0)\,. \tag{3.10}$$

При $\alpha = \beta$ условия (3.10) позволяют ограничиться только левыми полуспектрами, $1 \leqslant n \leqslant N$, $1 \leqslant m_0 \leqslant N$,

$$N = \left[\frac{M}{2}\right], \tag{3.11}$$

где квадратные скобки обозначают целую часть числа.

Поскольку каждая позиция в последовательности заполняется только одним нуклеотидом, то справедливы равенства

$$\sum_\alpha \rho_{m,\alpha} = 1\,, \tag{3.12}$$

$$\sum_\alpha \rho_\alpha(q_n) = 0 \qquad (n \neq 0)\,. \tag{3.13}$$

Эти ограничения выражают эффекты исключенного объема и приводят к нетривиальным корреляциям даже для случайных последовательностей. Как следует из (3.13), для бинарных последовательностей выполняется равенство $F_{11}(q_n) = F_{22}(q_n)$.

Важным свойством теории является наличие различных точных правил сумм. Приведем те из них, которые нам понадобятся в дальнейшем:

$$\bar{F}_{\alpha\beta} = (M-1)^{-1} \sum_{n=1}^{M-1} F_{\alpha\beta}(q_n) = \frac{\delta_{\alpha\beta} N_\alpha - N_\alpha N_\beta / M}{M-1}\,, \tag{3.14}$$

$$\bar{K}_{\alpha\beta} = (M-1)^{-1} \sum_{m_0=1}^{M-1} K_{\alpha\beta}(m_0) = \frac{N_\alpha N_\beta - \delta_{\alpha\beta} N_\beta}{M(M-1)}\,, \tag{3.15}$$

$$\sum_{m_0=1}^{M-1} \left[K_{\alpha\beta}(m_0) - \bar{K}_{\alpha\beta}\right] \left[K_{\gamma\delta}(m_0) - \bar{K}_{\gamma\delta}\right] =$$

$$= M^{-1} \sum_{n=1}^{M-1} \left[F_{\alpha\beta}(q_n) - \bar{F}_{\alpha\beta}\right] \left[F^*_{\gamma\delta}(q_n) - \bar{F}^*_{\gamma\delta}\right]\,, \tag{3.16}$$

где $\delta_{\alpha\beta}$ — символ Кронекера. Как видно из (3.14) и (3.15), для последовательностей с фиксированным составом нуклеотидов средние характеристики $\bar{F}_{\alpha\beta}$ и $\bar{K}_{\alpha\beta}$ оказываются одинаковыми.

В силу взаимной однозначности преобразования Фурье каждая структурная особенность исходной последовательности ДНК находит свое отражение в спектрах для структурных факторов и корреляционных функций. Так, периодичности порождают периодические вариации в $K(m_0)$ и серии эквидистантных пиков для $F(q_n)$ (см. раздел 4). В качестве другого случая рассмотрим простейший пример сегментации, когда аденины занимают сайты от 1 до L, а остальные нуклеотиды расположены за этой областью. Тогда получаем

$$F_{\mathrm{AA}}(q_n) = \frac{\sin^2(q_n L/2)}{M \sin^2(q_n/2)}\,, \tag{3.17}$$

что дает характерный рост в области малых волновых чисел $q_n \leqslant 1/L$. Для корреляционных функций с $\alpha = \beta$ и

конечным радиусом корреляций r_c имеем

$$\Delta K(m_0) = K(m_0) - \bar{K} \propto \exp\left(-\frac{m_0}{r_c}\right) + \exp\left(-\frac{M - m_0}{r_c}\right), \tag{3.18}$$

что дает

$$\Delta F(q_n) \propto \left[1 - \exp\left(-\frac{M}{r_c}\right)\right] \frac{\cos q_n - \exp(-1/r_c)}{\cosh(1/r_c) - \cos q_n}. \tag{3.19}$$

Соответствующие полуспектры для $M = 100$ приведены на рис. 2 для случая корреляций. В случае антикорреляций знаки ΔK и ΔF следует изменить на противоположные.

В образовании геномов важную роль играют не только прямые дупликации фрагментов, но также дупликации инвертированных и комплементарных фрагментов [4, 41]. Если фурье-гармоники для исходной последовательности имеют вид (3.2), то для инвертированной последовательности надо поменять начало и конец, что приводит к соотношению [21]

$$\rho_\alpha^{\mathrm{I}}(q_n) = \exp(-\mathrm{i}q_n)\, \rho_\alpha(2\pi - q_n), \tag{3.20}$$

где $\rho_\alpha^{\mathrm{I}}(q_n)$ — фурье-гармоники для инвертированной последовательности. Если последовательность составлена из прямого и инвертированного фрагментов, то она инвариантна относительно инверсии. Тогда (3.20) немедленно приводит к условию $\rho_\alpha(\pi) = 0$. Комплементарная последовательность получается из исходной с помощью инверсии и замен $\mathrm{A} \leftrightarrow \mathrm{T}$, $\mathrm{G} \leftrightarrow \mathrm{C}$. Поэтому для последовательности, составленной из прямого и комплементарного фрагментов, выполняются равенства:

$$F_{\mathrm{AA}}(q_n) = F_{\mathrm{TT}}(q_n), \qquad F_{\mathrm{CC}}(q_n) = F_{\mathrm{GG}}(q_n). \tag{3.21}$$

Таким образом, спектральный анализ позволяет, в принципе, выделить все структурные особенности анализируемой последовательности ДНК.

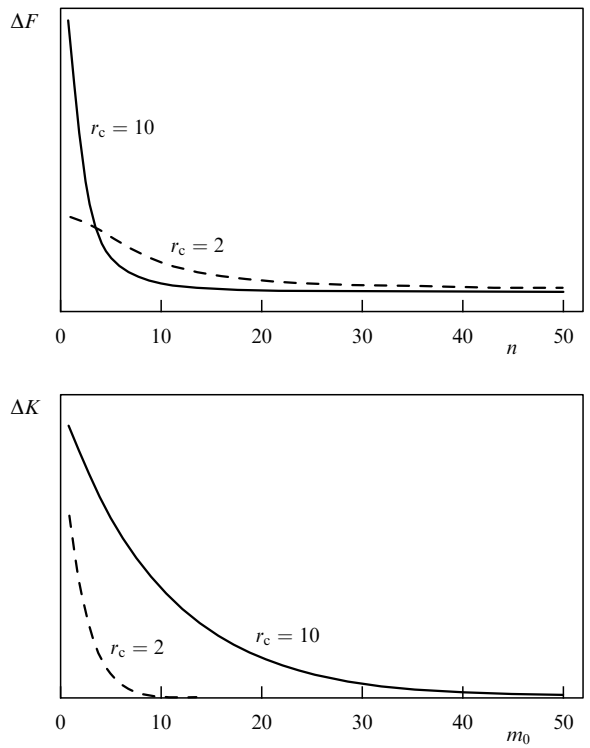

Рис. 2. Отклонения амплитуд фурье-гармоник ΔF и корреляционных функций ΔK от средних значений при различных радиусах корреляции.

3.2. Статистические характеристики для случайных последовательностей

В реальных геномных последовательностях ДНК все регулярные черты проявляются на сильном случайном фоне, обусловленном точечными мутациями (это основной источник случайных модификаций [4, 29]), вставками, делециями, транслокациями и т.д. Поэтому статистическая значимость наблюдаемых регулярных особенностей должна оцениваться относительно соответствующих характеристик для случайных последовательностей ДНК с тем же нуклеотидным составом.

Статистическое распределение для фурье-гармоник можно найти, усредняя характеристическую функцию (см., например, [51, 92])

$$Z = \exp\left[\mathrm{i} \sum_\alpha \sum_{n=1}^{M-1} u_\alpha(q_n)\, \rho_\alpha(q_n)\right] \tag{3.22}$$

по ансамблю случайных реализаций последовательностей с фиксированными полными числами нуклеотидов $\{N_\alpha\}$. На вспомогательные переменные $u_\alpha(q_n)$ удобно наложить условие, аналогичное (3.9):

$$u_\alpha^*(q_n) = u_\alpha(2\pi - q_n). \tag{3.23}$$

Различные произведения фурье-гармоник получаются путем дифференцирования Z по вспомогательным переменным $u_\alpha(q_n)$ и последующим приравниванием $u_\alpha(q_n)$ нулю, например,

$$\rho_\alpha(q_n) = \left.\frac{\partial Z}{\mathrm{i}\partial u_\alpha(q_n)}\right|_{u_\alpha = 0} \tag{3.24}$$

и т.д. Используя определения (3.1) и (3.2), перепишем Z в виде

$$Z = \prod_\alpha \prod_{m=1}^{M} (1 + \rho_{m,\alpha}\, z_{m,\alpha}), \tag{3.25}$$

$$z_{m,\alpha} = \exp\left[\mathrm{i}M^{-1/2} \sum_{n=1}^{M-1} u_\alpha(q_n) \exp(-\mathrm{i}q_n m)\right] - 1. \tag{3.26}$$

В результате проблема сводится к усреднению различных произведений $\rho_{m,\alpha}$.

Усреднение производится с помощью простых комбинаторных соображений, и результат имеет следующий вид:

$$\left\langle \prod_{k_1=1}^{n_{\mathrm{A}}} \rho_{m_{k_1},\mathrm{A}} \prod_{k_2=1}^{n_{\mathrm{C}}} \rho_{m_{k_2},\mathrm{C}} \prod_{k_3=1}^{n_{\mathrm{G}}} \rho_{m_{k_3},\mathrm{G}} \prod_{k_4=1}^{n_{\mathrm{T}}} \rho_{m_{k_4},\mathrm{T}} \right\rangle = \frac{C^{M-n_{\mathrm{A}}-n_{\mathrm{C}}-n_{\mathrm{G}}-n_{\mathrm{T}}}_{N_{\mathrm{A}}-n_{\mathrm{A}},\, N_{\mathrm{C}}-n_{\mathrm{C}},\, N_{\mathrm{G}}-n_{\mathrm{G}},\, N_{\mathrm{T}}-n_{\mathrm{T}}}}{C^{M}_{N_{\mathrm{A}},\, N_{\mathrm{C}},\, N_{\mathrm{G}},\, N_{\mathrm{T}}}}, \tag{3.27}$$

$$C^{m}_{n_{\mathrm{A}},\, n_{\mathrm{C}},\, n_{\mathrm{G}},\, n_{\mathrm{T}}} = \frac{m!}{n_{\mathrm{A}}!\, n_{\mathrm{C}}!\, n_{\mathrm{G}}!\, n_{\mathrm{T}}!},$$

$$n_{\mathrm{A}} + n_{\mathrm{C}} + n_{\mathrm{G}} + n_{\mathrm{T}} = m, \qquad 0! = 1. \tag{3.28}$$

Здесь угловые скобки обозначают усреднение по ансамблю случайных реализаций, все индексы $\{m_{k_1}\}, \ldots, \{m_{k_4}\}$ предполагаются различными в силу эффектов исключенного объема, n_α — полное число функций положения, отвечающих нуклеотидам типа α, а N_α — полное число нуклеотидов типа α в последовательности длиной M. Правая часть (3.27) равна отношению двух комбинаторных факторов: $C^M_{N_\mathrm{A},N_\mathrm{C},N_\mathrm{G},N_\mathrm{T}}$ — полное число различных случайных реализаций,

$$C^{M-n_\mathrm{A}-n_\mathrm{C}-n_\mathrm{G}-n_\mathrm{T}}_{N_\mathrm{A}-n_\mathrm{A},N_\mathrm{C}-n_\mathrm{C},N_\mathrm{G}-n_\mathrm{G},N_\mathrm{T}-n_\mathrm{T}}$$

— полное число реализаций при условии, что n_A позиций заняты нуклеотидами A, и аналогично для n_C, n_G и n_T (это позиции, которые входят в левую часть (3.27)).

Прямое вычисление низших средних с $q_n \neq 0$ дает

$$\big\langle \rho_\alpha(q_n)\big\rangle = 0\,, \tag{3.29}$$

$$\big\langle \rho_\alpha(q_n)\rho_\alpha(q_{n'})\big\rangle = \begin{cases} \bar F_{\alpha\beta}\,, & q_n + q_{n'} = 2\pi\,,\\ 0, & q_n + q_{n'} \neq 2\pi\,,\end{cases} \tag{3.30}$$

где $\bar F_{\alpha\beta}$ определяется соотношением (3.14). Равенство (3.30) отражает важное квазиэргодическое свойство: усреднение по ансамблю асимптотически эквивалентно усреднению по спектру. В пределе $M \gg 1$, $N_\mathrm{A} \gg 1, \ldots, N_\mathrm{T} \gg 1$ с учетом (3.23) в главном порядке по $M^{-1/2}$ можно использовать следующее приближенное выражение для $\langle Z\rangle$ [17, 18]:

$$\langle Z\rangle \approx \exp\Bigg[-\sum_{\alpha,\beta}\sum_{n=1}^{N} \bar F_{\alpha\beta}\, u_\alpha(q_n)\, u^*_\beta(q_n)\Bigg]\,, \tag{3.31}$$

где N задается выражением (3.11). С помощью (3.31) можно получить ряд полезных конкретных критериев регулярности.

Взаимные корреляции. Корреляции положений различных нуклеотидов характеризуются коэффициентом взаимной корреляции [51, 92]:

$$k(F_{\alpha\beta}|F_{\gamma\delta};\, M-1) = \\ = \sum_{n=1}^{M-1} \frac{\big[F_{\alpha\beta}(q_n) - \bar F_{\alpha\beta}\big]\big[F^*_{\gamma\delta}(q_n) - \bar F^*_{\gamma\delta}\big]}{(M-1)\,\sigma(F_{\alpha\beta})\,\sigma(F_{\gamma\delta})}\,, \tag{3.32}$$

$$\sigma^2(F_{\alpha\beta};\, M-1) = \sum_{n=1}^{M-1} \frac{\big[F_{\alpha\beta}(q_n) - \bar F_{\alpha\beta}\big]\big[F^*_{\alpha\beta}(q_n) - \bar F^*_{\alpha\beta}\big]}{M-1}\,. \tag{3.33}$$

Если k стремится к единице, то положения нуклеотидов полностью коррелированы, если же $k \approx 0$, то корреляции отсутствуют. Аналогичные величины можно ввести и для корреляционных функций (3.7), однако точное правило сумм (3.16) дает

$$\sigma(K_{\alpha\beta};\, M-1) = \frac{\sigma(F_{\alpha\beta};\, M-1)}{M^{1/2}}\,, \tag{3.34}$$

$$k(K_{\alpha\beta}|K_{\gamma\delta};\, M-1) = k(F_{\alpha\beta}|F_{\gamma\delta};\, M-1)\,. \tag{3.35}$$

Используя асимптотическую эквивалентность усреднений по ансамблю и по спектру, для случайных последовательностей получаем

$$\big\langle k(F_{\alpha\beta}|F_{\gamma\delta};\, M-1)\big\rangle = \frac{\bar F_{\alpha\gamma}\bar F_{\delta\beta}}{(\bar F_{\alpha\alpha}\bar F_{\beta\beta}\bar F_{\gamma\gamma}\bar F_{\delta\delta})^{1/2}}\,, \tag{3.36}$$

в частности, для $\alpha \neq \beta$

$$\big\langle k(F_{\alpha\alpha}|F_{\beta\beta};\, M-1)\big\rangle \equiv \langle k_{\alpha\beta}\rangle = \frac{N_\alpha N_\beta}{(M-N_\alpha)(M-N_\beta)}\,. \tag{3.37}$$

Уравнение (3.37) имеет простой физический смысл. Коэффициент корреляции $\langle k(F_{\alpha\alpha}|F_{\beta\beta};\, M-1)\rangle$ равен вероятности одновременно найти нуклеотиды типа α в позициях, свободных от нуклеотидов типа β, и наоборот. Как видим, эффекты исключенного объема приводят к ненулевым корреляциям между различными нуклеотидами даже в случайных последовательностях. Для среднеквадратичных отклонений от среднего значения коэффициента корреляции можно воспользоваться стандартной оценкой [93]:

$$\big\langle \big[\Delta k(F_{\alpha\alpha}|F_{\beta\beta};\, M-1)\big]^2\big\rangle \equiv \big\langle (\Delta k_{\alpha\beta})^2\big\rangle \approx \frac{1}{N}\,. \tag{3.38}$$

Мы не приводим здесь поправочных вкладов в (3.38), связанных с конечностью средних коэффициентов корреляции [23]. Существенно, что эти поправки отрицательны, поэтому (3.38) может лишь усилить оценку для статистической значимости нормированного отклонения, $\big[k(F_{\alpha\alpha}|F_{\beta\beta};\, M-1) - \langle k_{\alpha\beta}\rangle\big]/\big[2\langle(\Delta k_{\alpha\beta})^2\rangle\big]^{1/2}$, которое для случайных последовательностей имеет гауссово распределение.

Распределение амплитуд гармоник. С учетом того, что характеристическая функция и многомерная функция распределения вероятностей связаны между собой преобразованием Фурье [51, 92], из (3.31) можно получить распределение вероятностей для амплитуд гармоник в случайных спектрах (им соответствуют диагональные элементы структурного фактора (3.5)). Вероятность того, что амплитуда какой-то выбранной n-й гармоники находится в интервале между $F_{\alpha\alpha}(q_n)$ и $F_{\alpha\alpha}(q_n) + \mathrm{d}F_{\alpha\alpha}(q_n)$ равна

$$p\big(F_{\alpha\alpha}(q_n)\big)\,\mathrm{d}F_{\alpha\alpha}(q_n) = \exp(-f_{n,\alpha\alpha})\,\mathrm{d}f_{n,\alpha\alpha}\,, \tag{3.39}$$

$$f_{n,\alpha\alpha} = \frac{F_{\alpha\alpha}(q_n)}{\bar F_{\alpha\alpha}}\,. \tag{3.40}$$

Распределение (3.39) отвечает распределению Рэлея [51], которое в спектральном анализе играет столь же универсальную роль, как и распределение Гаусса для вещественных переменных.

Из (3.39) находим, что вероятность превышения некоторого выбранного значения $F_{\alpha\alpha}$ имеет вид

$$P\big\{F_{\alpha\alpha}(q_n) > F_{\alpha\alpha}\big\} = \int_{F_{\alpha\alpha}}^{\infty} p(F'_{\alpha\alpha})\,\mathrm{d}F'_{\alpha\alpha} = \exp\Bigg(-\frac{F_{\alpha\alpha}}{\bar F_{\alpha\alpha}}\Bigg)\,. \tag{3.41}$$

Это означает также, что среднее число гармоник в полуспектре с амплитудами, превышающими $F_{\alpha\alpha}$, равно

$$\langle n_\alpha\rangle = N\exp\Bigg(-\frac{F_{\alpha\alpha}}{\bar F_{\alpha\alpha}}\Bigg) \tag{3.42}$$

(здесь учтено, что только N гармоник являются статистически независимыми согласно (3.10) и (3.11)). Условие $\langle n_\alpha\rangle = 1$ определяет характерное значение амплитуд сингулярных выбросов в случайных спектрах:

$$F_{\alpha\alpha,\max} \approx \bar F_{\alpha\alpha}\ln N\,. \tag{3.43}$$

Структурная спектральная энтропия последовательности. Коэффициенты взаимной корреляции (3.32) характеризуют лишь взаимные положения нуклеотидов, но не степень упорядоченности последовательности. Действительно, две почти совпадающие случайные последовательности были бы сильно коррелированными, оставаясь при этом случайными. Интегральную степень упорядоченности последовательности можно оценить с помощью структурной спектральной энтропии

$$S_\alpha = -\sum_{n=1}^{M-1} f_{n,\alpha\alpha} \ln f_{n,\alpha\alpha}\,, \tag{3.44}$$

где $f_{n,\alpha\alpha}$ определены в (3.40). С учетом того, что $\{f_{n,\alpha\alpha}\}$ подчиняются правилу сумм (3.14)

$$\sum_{n=1}^{M-1} f_{n,\alpha\alpha} = \mathrm{const}\,, \tag{3.45}$$

нетрудно видеть, что при условии (3.45) S_α достигает максимума при строго равномерном распределении амплитуд по спектру. Поскольку для случайных последовательностей гармоники распределены по спектру более равномерно, чем для упорядоченных (см. ниже), то S_α приближенно характеризует степень упорядоченности.

Усредняя S_α с помощью функции распределения (3.39), для средней величины спектральной энтропии в случайных последовательностях получаем

$$\langle S_\alpha\rangle_{\mathrm{random}} = -(1-C)(M-1)\,, \tag{3.46}$$

где $C = 0{,}577215\ldots$ — постоянная Эйлера. Расчет среднеквадратичных отклонений требует учета поправок на корреляции между гармониками с разными q_n и дает приближенно $\left\langle(\Delta S_\alpha)^2\right\rangle_{\mathrm{random}} \simeq (0{,}5797\ldots)(M-1)$.

Полная спектральная энтропия равна

$$S = \sum_\alpha S_\alpha\,. \tag{3.47}$$

Степень упорядоченности последовательностей с разной длиной удобно описывать в терминах относительной энтропии $\Delta S_{\mathrm{rel}} = (\langle S\rangle_{\mathrm{random}} - S)/|\langle S\rangle_{\mathrm{random}}|$, поскольку для регулярных отклонений от случайного распределения, $\langle S\rangle_{\mathrm{random}} - S \propto (M-1)$, это отношение не зависит от M. Для случайных последовательностей $\Delta S_{\mathrm{rel}} \propto (M-1)^{-1/2}$ и мала при достаточно больших M. Относительная независимость ΔS_{rel} от длины последовательности и нуклеотидного состава позволяет использовать ΔS_{rel} в качестве универсальной меры структурной регулярности для различных последовательностей. Конкретные расчеты с использованием этой величины можно найти в [19, 24].

3.3. Примеры спектров для геномных последовательностей ДНК

В качестве примера рассмотрим спектры для геномной последовательности бактериофага PHIX174 (коды доступа в базе данных ЕМБЛ V01128 и J02482). С одной стороны, молекулярно-биологическая структура этого генома хорошо изучена [29, 94, 95], с другой стороны, данный геном обладает рядом универсальных черт, которые будут обсуждаться в этом и следующем разделах. Геном PHIX174 представляет собой одноцепочечную кольцевую ДНК длиной $M = 5386$, имеющую состав $N_{\mathrm{A}} = 1291$, $N_{\mathrm{C}} = 1157$, $N_{\mathrm{G}} = 1254$, $N_{\mathrm{T}} = 1684$. Геном содержит 11 различных генов, два из которых являются составными, а некоторые другие гены частично перекрываются.

Далее ограничимся только диагональными элементами структурного фактора (3.5). Соответствующие полуспектры для нормированных гармоник (3.40) показаны на рис. 3. Все фурье-спектры представляются в зависимости от номеров n волновых чисел $q_n = 2\pi n/M$ по причинам, которые будут пояснены в разделе 4. Пересчет в формальные периоды производится по формуле $p = M/n$. На вставках рис. 3 отдельно показана область малых волновых чисел, отражающая крупномасштабные эффекты сегментации (ср. (3.17)), и окрестность высоких пиков, отвечающих периодичности $p = 3$, которая является универсальной чертой белок-кодирующих областей (см. раздел 4.3). Пики имеют малые полуширины $\sim 1-2$ гармоники, указывающие на незатухающий характер периодичности, а пики для f_{AA} и f_{TT} имеют характерную расщепленную структуру.

В дальнейшем мы будем использовать следующий результат, полученный в [18]. Пусть задана последовательность целых чисел $k = 2, 3, \ldots$ Если n/M лежит вблизи $1/k$, то число модуляций в огибающей максимумов для $\cos(q_n m)$ (где $q_n = 2\pi n/M$) при изменении m от 1 до M равно модулю разности

$$n_{\mathrm{s}} = |kn - M|\,. \tag{3.48}$$

Переход от периода k к $k+1$ происходит при n, определяемом соотношением

$$(k+1)n - M = M - kn\,. \tag{3.49}$$

Модуляционные сверхпериоды можно наблюдать только в том случае, когда внутри интервала $\big(2\pi/(k+1),\ 2\pi/k\big)$ находятся хотя бы два различных волновых числа q_n, что приводит к условию

$$\frac{1}{k} - \frac{1}{k+1} > \frac{1}{M}\,. \tag{3.50}$$

В частности, наивысшая гармоника с $n = 1795$ для f_{AA} порождает один модуляционный сверхпериод, а наивысшие гармоники с $n = 1797$ для f_{GG} и f_{TT} порождают пять модуляционных сверхпериодов.

Зависимости доли гармоник $N(f)/N$, амплитуды которых превышают пороговое значение f, от величины этого порога f приведены на рис. 4. Как видно из рис. 4, лишь немногим более десятка гармоник отклоняется от экспоненциального распределения (3.42) для случайных последовательностей.

Для корреляционных функций (3.6), (3.7) удобно использовать переменные

$$\varkappa_{\alpha\alpha}(m_0) = \frac{K_{\alpha\alpha}(m_0) - \bar{K}_{\alpha\alpha}}{\left[2\left\langle(\Delta K_{\alpha\alpha})^2\right\rangle\right]^{1/2}}\,, \tag{3.51}$$

где $\bar{K}_{\alpha\alpha}$ определяется формулой (3.15) и

$$\left\langle(\Delta K_{\alpha\alpha})^2\right\rangle = \frac{\bar{F}^2_{\alpha\alpha}}{M}\,. \tag{3.52}$$

Для случайных последовательностей можно приближенно считать $\varkappa_{\alpha\alpha}(m_0)$ с различными m_0 независимыми гауссовыми переменными. Общий вид соответствующих

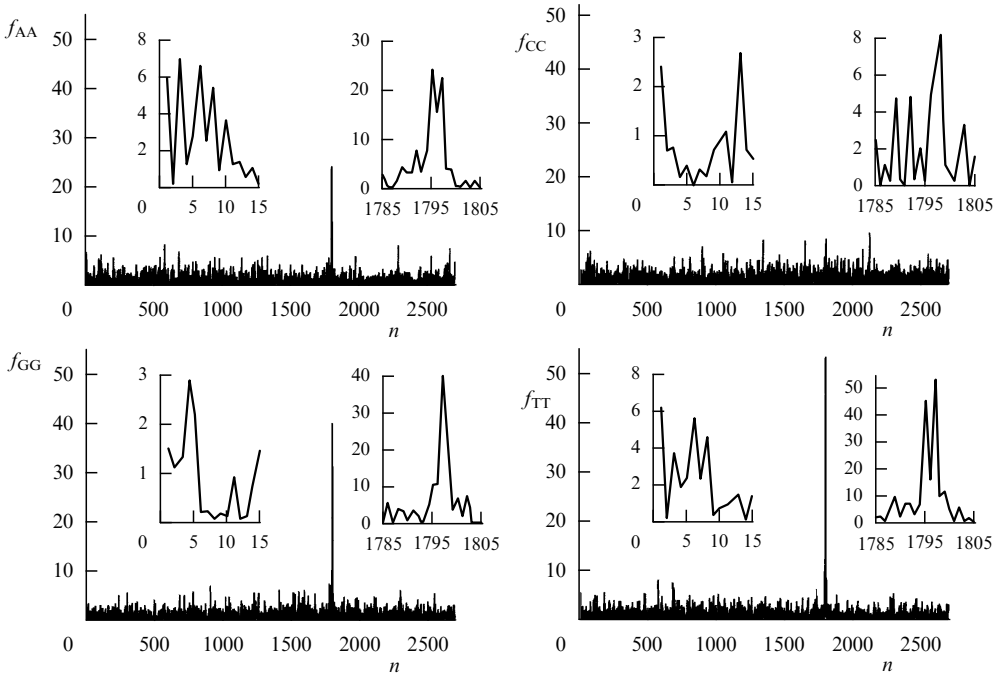

Рис. 3. Полуспектры для нормированных диагональных элементов структурного фактора (формулы (3.5) и (3.40)) для геномной последовательности PHIX174.

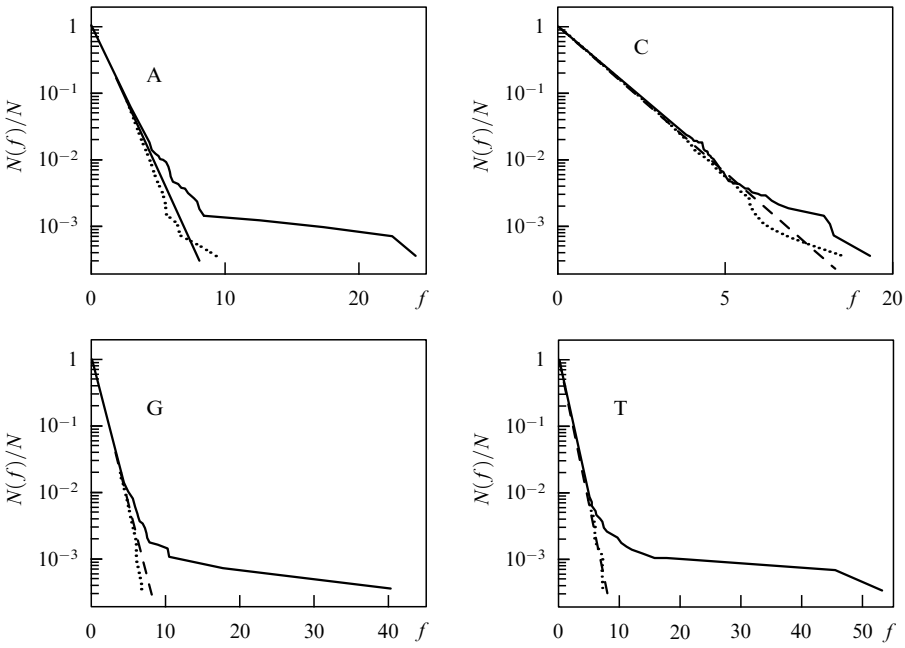

Рис. 4. Зависимости для относительных чисел гармоник с амплитудой, превышающей заданную величину f. Штриховые линии отвечают средней зависимости (3.42) для случайных последовательностей с тем же нуклеотидным составом. Пунктирные кривые отвечают одной из случайных реализаций.

полуспектров приведен на рис. 5 и более подробно на рис. 6. Осцилляции с периодом $p = 3$ хорошо видны в левой части рис. 6. В правой части рис. 6 показаны текущие дисперсии, усредненные по 100 сайтам:

$$\tilde{\sigma} = \left[100^{-1} \sum_{m_0'=m_0}^{m_0+99} \varkappa_{\alpha\alpha}^2(m_0)\right]^{1/2}, \qquad (3.53)$$

где $m_0 = 1, 101, 201, \ldots, M/2$. Дисперсии обнаруживают характерное спадание в области порядка средних размеров генов, $m_0 \sim 500$, однако для T четко видны также крупномасштабные вариации, сравнимые с полной длиной M.

Относительные структурные спектральные энтропии (см. определение в конце раздела 3.2) для генома PHIX174 равны

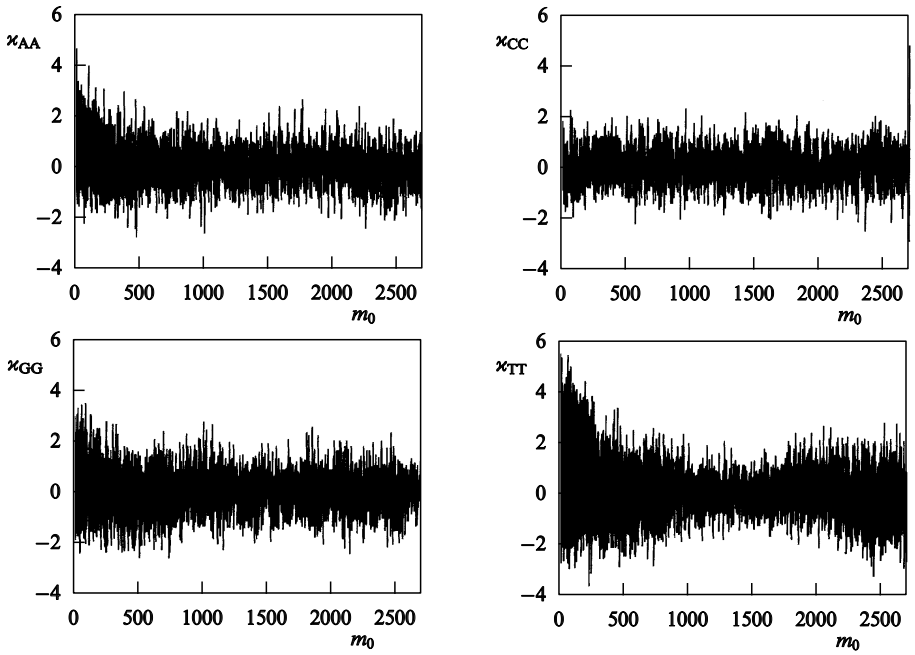

Рис. 5. Полуспектры для нормированных корреляционных функций (3.51) для геномной последовательности PHIX174.

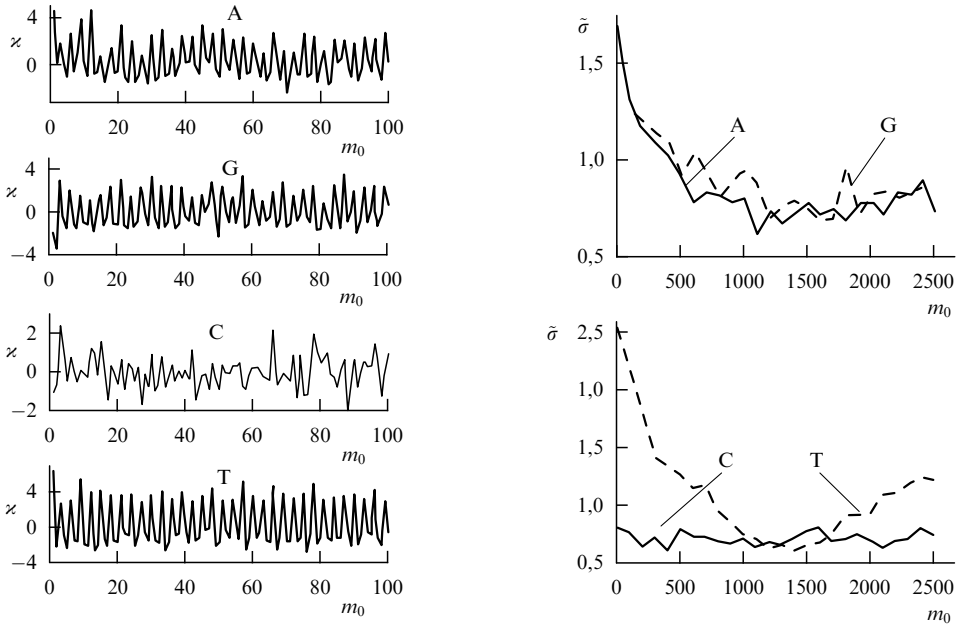

Рис. 6. Начальные участки спектров, приведенных на рис. 5 (слева). Сглаженные дисперсии для нормированных корреляционных функций (3.53) (справа).

$$\Delta S_{\mathrm{rel,\,A}} = 1{,}41 \times 10^{-1}\,; \qquad \Delta S_{\mathrm{rel,\,C}} = -1{,}75 \times 10^{-3}\,;$$
$$\Delta S_{\mathrm{rel,\,G}} = 1{,}20 \times 10^{-1}\,; \qquad \Delta S_{\mathrm{rel,\,T}} = 2{,}25 \times 10^{-1}\,;$$
$$\Delta S_{\mathrm{rel}} = 1{,}28 \times 10^{-1}\,.$$

Только для A, G и T отклонения от средних случайных значений оказываются статистически значимыми. Значения ΔS_{rel} сильно отличаются в различных генах [19].

Этот пример показывает, что структурные характеристики для разных нуклеотидов могут сильно отличаться в геномных последовательностях ДНК. Поэтому анализ всегда следует начинать на четырехнуклеотидном уровне. Лишь на следующем этапе исследования допустимы различные объединения нуклеотидов в бинарные последовательности.

4. Скрытые периодичности в геномных последовательностях ДНК

Выше, в разделе 1.1, уже отмечалась важная роль, которую играют повторяющиеся фрагменты ДНК в геномах высших организмов [29, 31, 35]. В скрытой форме повторы, модифицированные случайными мутациями, вставками, делециями и т.д., оказываются типичными практически для всех геномных ДНК. Отчасти это

связано с периодичностью структуры двойной спирали, приближенно повторяющейся структурой нуклеосом, 30-нм фибрилл и т.д. [38–40]. Ряд периодичностей, такие как период $p = 3$, присущий белок-кодирующим областям [5, 41, 45], имеет, возможно, эволюционное происхождение, связанное с генетическим кодом [96, 97]. Другие периодичности в белок-кодирующих областях могут быть связаны с регулярностью и сегментацией элементов вторичной структуры белков [98, 99]. Некоторые приложения этих идей можно найти в [22]. Часть периодичностей может быть обусловлена наличием повторяющихся участков кооперативного связывания ДНК с регуляторными белками [29]. Наконец, отметим периодичности, связанные с эволюционным отбором и последующей дупликацией отдельных структур [4, 28, 29]. В данном разделе мы опишем общую технику выявления скрытых периодичностей в рамках спектрального подхода [21, 26].

4.1. Структурные особенности для скрытых периодичностей в геномных последовательностях ДНК

Рассмотрим сначала идеальный случай, когда фрагмент последовательности длиной L повторяется N_p раз, так что $M = LN_p$. Фурье-гармоники (3.2) для такой последовательности можно представить в виде

$$\rho_\alpha(q_n) = \frac{\rho_{\alpha,\,\mathrm{repeat}}(q_n)\{1+\exp(-\mathrm{i}q_nL)+\ldots+\exp[-\mathrm{i}(N_p-1)q_nL]\}}{N_p^{1/2}}, \tag{4.1}$$

$$\rho_{\alpha,\,\mathrm{repeat}}(q_n) = L^{-1/2}\sum_{m=1}^{L}\rho_{m,\,\alpha}\exp(-\mathrm{i}q_nL)\,,$$

$$q_n = \frac{2\pi n}{M}\,, \qquad n = 0, 1, \ldots, M-1\,. \tag{4.2}$$

Отсюда для диагональных элементов структурного фактора (3.5) получаем

$$F_{\alpha\alpha}(q_n) = \frac{F_{\alpha\alpha,\,\mathrm{repeat}}(q_n)\sin^2(q_nLN_p/2)}{N_p\sin^2(q_nL/2)}\,. \tag{4.3}$$

Как видно из (4.3), повторы длиной L порождают серию из $L-1$ эквидистантных пиков при $q_n = \pi k/L$, $k = 1, \ldots, L-1$. Именно потому, что в общем случае периодичности связаны с сериями эквидистантных пиков, полуспектры Фурье представляются в терминах номеров n, а не формальных периодов $p = M/n$.

На рисунке 7 слева показаны нормированные полуспектры (см. (3.40)) для 10^2 повторов ATAAACT в геноме Drosophila virilis. Справа показаны соответствующие спектры для последовательности, полученной после 45 % случайных замен с вероятностями, пропорциональными содержанию соответствующих нуклеотидов в исходной последовательности.

Пусть теперь в последовательности из 10^2 повторов ATAAACT повторы последовательно перенумерованы и разделены на четные и нечетные. Пусть далее в части четных повторов произведены случайные замены C на G. Такие замены приводят к характерной модификации спектров для нуклеотидов C, когда внутри серии из исходных высоких эквидистантных пиков появляются

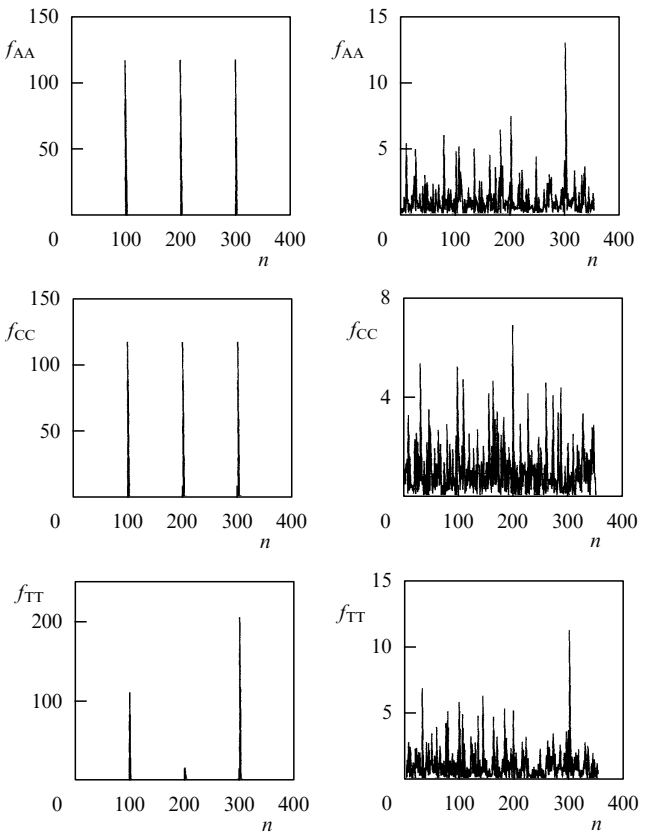

Рис. 7. Полуспектры для нормированных амплитуд гармоник для последовательности из 10^2 повторов ATAAACT (слева). То же после $\sim 45\,\%$ случайных замен с вероятностями, пропорциональными встречаемости нуклеотидов в исходной последовательности (справа).

эквидистантные серии с меньшей высотой (рис. 8). Их можно также приближенно рассматривать как частичное образование двойных комплексов ATAAACTATAAAGT или частичное удвоение периода. Процесс можно продолжить и дальше, т.е. разделить удвоенные фрагменты из 14 нуклеотидов на четные и нечетные и для части удвоенных четных фрагментов произвести случайные замены C на G, и т.д. Такие замены порождают характерную иерархическую систему эквидистантных пиков и могут рассматриваться как каскад частичных удвоений периода. Далее такой процесс будет называться умножением периода или образованием структурных субгармоник. Он оказывается весьма характерным для геномных последовательностей ДНК. Можно сказать, что практически любой резко выраженный период в

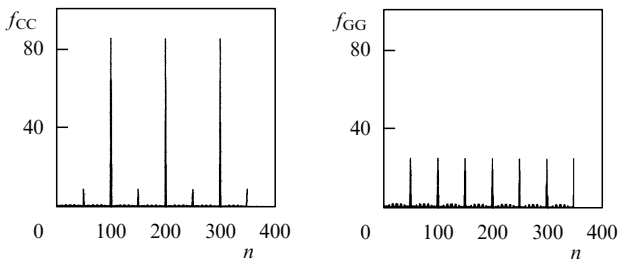

Рис. 8. Полуспектры для нормированных амплитуд гармоник для последовательности из 10^2 повторов ATAAACT после случайных замен C на G примерно в половине четных повторов.

последовательностях ДНК сопровождается образованием структурных субгармоник. Умножение периода, вообще говоря, не обязательно является удвоением и специфично проявляется в разных геномах. При этом по характеру распределения пиков можно выяснить, какие замены являются ответственными за умножение периода (ср. левую и правую части рис. 8).

Другой характерной чертой геномных последовательностей является одновременное существование различных периодичностей, что приводит к эффектам взаимной модуляции. Поясним это на простейшем примере вариаций плотности типа

$$\Delta\rho_m \propto \cos\left(\frac{2\pi m}{P}\right)\cos\left(\frac{2\pi m}{p_0}\right) \tag{4.4}$$

с $P \gg p_0$. Такие модуляции приведут к расщеплению периода p_0 на

$$\frac{1}{p_{1,2}} = \frac{1}{p_0} \pm \frac{1}{P}\,. \tag{4.5}$$

Подобные эффекты взаимной модуляции приводят к необходимости рассматривать не только серии строго эквидистантных гармоник, но и серии, несколько отличающиеся от строгой эквидистантности. Для этого каждая гармоника $F_{\alpha\alpha}(q_n), \ldots, F_{\alpha\alpha}(rq_n)$ в исходной эквидистантной серии окружается окном шириной w:

$$F_{\alpha\alpha}(kq_n) = F_{\alpha\alpha}(q_{kn}) \to \\ \to F_{\alpha\alpha}(q_{kn-w}), \ldots, F_{\alpha\alpha}(q_{kn}), \ldots, F_{\alpha\alpha}(q_{kn+w}) \tag{4.6}$$

и из $2w+1$ гармоник в каждом окне выбирается гармоника с наибольшей амплитудой. Оптимальный выбор окна w зависит от характера модуляций.

Введение окна w ухудшает разрешение исходного периода. Поскольку

$$\frac{2\pi(n_p \pm w)}{M} = \frac{2\pi}{p \pm \Delta p} \approx \frac{2\pi}{p}\left(1 \mp \frac{\Delta p}{p}\right),$$

то для неопределенности периода получаем оценку $\Delta p \sim 2wp^2/M$. Поэтому более длинные периоды должны изучаться с использованием более узких окон. Если окном окружаются эквидистантные гармоники, начиная только с k-й, то с учетом соотношения $2\pi(kn_p \pm w)/M = 2\pi k/(p \pm \Delta p)$ получаем некоторое уменьшение неопределенности $\Delta p \sim 2wp^2/(kM)$.

Основным источником случайных модификаций повторов являются точечные мутации, которые создают случайный фон, не изменяя длины периода. Случайные вставки и делеции приводят к вариациям длины периода. Чтобы качественно понять роль этих эффектов, усредним гармонику $\exp(\mathrm{i}q_k p)$ со случайной величиной p по гауссовой функции распределения со средним p_0 и среднеквадратичным отклонением $\langle(\Delta p)^2\rangle$,

$$\langle\exp(\mathrm{i}q_k p)\rangle \approx \\ \approx \left[2\pi\langle(\Delta p)^2\rangle\right]^{-1/2}\int_{-\infty}^{\infty}\mathrm{d}p\,\exp\left[-\frac{(p-p_0)^2}{2\langle(\Delta p)^2\rangle} + \mathrm{i}q_k p\right] = \\ = \exp\left[\mathrm{i}q_k p_0 - \frac{q_k^2\langle(\Delta p)^2\rangle}{2}\right]. \tag{4.7}$$

Из (4.7) получаем, что случайные вариации периода приводят к более сильному затуханию гармоник с более высокими волновыми числами. Поэтому, в частности, в серии эквидистантных гармоник могут оказаться статистически значимыми только несколько первых или лишь одна первая гармоника.

Таким образом, скрытые периодичности в последовательности могут быть выявлены с помощью отдельных высоких пиков, с помощью сумм эквидистантных гармоник, или с помощью комбинации обоих этих методов.

4.2. Критерии статистической значимости для скрытых периодичностей

Далее всегда предполагается, что изучаются только гармоники $F_{\alpha\alpha}(q_n)$ и амплитуды гармоник нормированы согласно (3.40). Для вывода критериев статистической значимости при поиске скрытых периодичностей следует использовать полную многомерную плотность распределения вероятностей для всего полуспектра (см. (3.11)):

$$p(f_1, \ldots, f_N) = \exp(-f_1 - \ldots - f_N)\,. \tag{4.8}$$

При многократных выборках следует принимать во внимание влияние статистики выбросов [100], так как в процессе все более и более длинных переборов для случайных величин можно встретить все более и более высокие значения.

Начнем с критериев для отдельных высоких пиков и будем последовательно усложнять ситуацию, следуя работе [21]. Вероятность того, что амплитуды всех гармоник не превышают порогового значения f, $0 \leqslant f_1 \leqslant f, \ldots, 0 \leqslant f_N \leqslant f$, равна

$$P(f_n \leqslant f;\, N) = \left[1 - \exp(-f)\right]^N, \tag{4.9}$$

а вероятность того, что хотя бы одна из N гармоник превысит f, является дополнительной к (4.9):

$$P(f_n > f;\, N) = 1 - \left[1 - \exp(-f)\right]^N. \tag{4.10}$$

В приложениях задают обычно 10%- или 5%-ный пороги статистической значимости для $P(f_n > f; N)$. Для приближенных оценок можно также использовать средние значения и характерные дисперсии. Определяя моменты $\langle f_{\max}^m\rangle$ как

$$\langle f_{\max}^m\rangle = \int_0^{\infty}\mathrm{d}f\,f^m\,\frac{\mathrm{d}P(f_n \leqslant f; N)}{\mathrm{d}f}\,, \tag{4.11}$$

получаем

$$\langle f_{\max}\rangle = \sum_{k=1}^{N}\frac{1}{k}\,, \tag{4.12}$$

$$\left\langle(\Delta f_{\max})^2\right\rangle = \langle f_{\max}^2\rangle - \langle f_{\max}\rangle^2 = \sum_{k=1}^{N}\frac{1}{k^2}\,. \tag{4.13}$$

Уравнение (4.12) отвечает известному результату для порядковых статистик для случайных величин с распределением Рэлея [51]. При $N \gg 1$ имеем $\langle f_{\max}\rangle \approx \ln N$ в согласии с оценкой (3.43). При этом величина $\left\langle(\Delta f_{\max})^2\right\rangle$ стремится к постоянному пределу $\pi^2/6$.

Сумма из r гармоник, S_r, имеет плотность распределения вероятностей [51]

$$p_r(S) = \int_0^\infty \mathrm{d}f_1 \ldots \int_0^\infty \mathrm{d}f_r\, \delta(S - f_1 - \ldots - f_r) \times \\ \times \exp(-f_1 - \ldots - f_r) = \frac{S^{r-1}}{(r-1)!} \exp(-S)\,, \quad (4.14)$$

а вероятность того, что S_r превысит порог S, равна

$$P_r(S) \equiv P(S_r > S) = \int_S^\infty \mathrm{d}S'\, p_r(S') = \exp(-S) \sum_{k=0}^{r-1} \frac{S^k}{k!}\,. \quad (4.15)$$

Из (4.14) находим среднее значение и дисперсию:

$$\langle S_r \rangle = r\,, \qquad \left\langle (\Delta S_r)^2 \right\rangle = r\,. \quad (4.16)$$

В случае N_r переборов для сумм S_r из r гармоник, таких, что ни одна из сумм S_r не содержит общих гармоник с любыми другими суммами, в соответствии с (4.9) и (4.10) получаем вероятность превышения S хотя бы в одном из переборов:

$$P(S_r > S; N_r) = 1 - \left[1 - P_r(S)\right]^{N_r}\,. \quad (4.17)$$

Выражения для среднего значения и дисперсии суммы при N_r различных выборках оказываются довольно громоздкими и проще использовать непосредственно (4.17).

Вероятность того, что сумма из r гармоник в схеме с окнами (см. (4.6)) не превысит порога S, определяется выражением

$$P_w(S_r \leqslant S) = \int_0^S \mathrm{d}f_1\, p_w(f_1) \int_0^{S-f_1} \mathrm{d}f_2\, p_w(f_2) \ldots \\ \ldots \int_0^{S-f_1-\ldots-f_{r-1}} \mathrm{d}f_r\, p_w(f_r)\,, \quad (4.18)$$

$$p_w(f) = \frac{\mathrm{d}P(f_n \leqslant f; 2w+1)}{\mathrm{d}f}\,,$$

$$P(f_n \leqslant f; 2w+1) = \left[1 - \exp(-f)\right]^{2w+1}. \quad (4.19)$$

Для выражений типа (4.18) удобно воспользоваться преобразованием Лапласа:

$$P_w(S_r \leqslant S) = \int_{a-\mathrm{i}\infty}^{a+\mathrm{i}\infty} \frac{\mathrm{d}\lambda}{2\pi\mathrm{i}} \frac{\exp(\lambda S)}{\lambda} \left(\prod_{k=1}^{2w+1} \frac{k}{\lambda + k} \right)^r, \quad (4.20)$$

где $a > 0$. С его помощью получаем

$$\langle S_{r,w} \rangle = r \sum_{k=1}^{2w+1} \frac{1}{k}\,, \quad (4.21)$$

$$\left\langle (\Delta S_{r,w})^2 \right\rangle = r \sum_{k=1}^{2w+1} \frac{1}{k^2}\,, \quad (4.22)$$

что находится в прямом соответствии с (4.12) и (4.13). Действительно, среднее значение для максимальной гармоники в окне шириной w определяется (4.12), а средняя сумма из r максимальных гармоник равна (4.21). Выражения (4.21) и (4.22) легко обобщаются на случай, когда разные эквидистантные гармоники окружаются окнами разной ширины.

В заключение этого раздела остановимся на статистической значимости для структурных субгармоник. Пусть заданы такие суммы из r_1 и r_2 гармоник, что они не содержат общих гармоник и величина $S_{r_1} + S_{r_2} = S$ фиксирована. Тогда вероятность того, что $S_{r_1} \geqslant S_1$ при условии $S_{r_1} + S_{r_2} = S$, определяется как [51]

$$P(S_{r_1} \geqslant S_1 | S_{r_1} + S_2 = S) = \frac{\int_{S_1}^S \mathrm{d}S_1'\, p_{r_1}(S_1') p_{r_2}(S - S_1')}{\int_0^S \mathrm{d}S_1'\, p_{r_1}(S_1') p_{r_2}(S - S_1')}\,, \quad (4.23)$$

где $p_{r_1}(S_1)$ и $p_{r_2}(S_2)$ определяются выражением (4.14). Средние $\langle S_1^m \rangle$ вычисляются стандартным образом:

$$\langle S_1^m \rangle = -\int_0^S \mathrm{d}S_1\, S_1^m \frac{\mathrm{d}P}{\mathrm{d}S_1}\,, \quad (4.24)$$

в частности,

$$\langle S_1 \rangle = \frac{r_1}{r_1 + r_2} S\,, \quad (4.25)$$

$$\left\langle (\Delta S_1)^2 \right\rangle = \frac{r_1 r_2}{(r_1 + r_2)^2 (r_1 + r_2 + 1)} S^2\,. \quad (4.26)$$

Рассмотрим набор эквидистантных гармоник $\{f_{n'k'}\}$, $k' = 1, \ldots, r$, с суммой S_r, статистически значимой в смысле вероятности (4.17), и выделим в нем группу с $k' = k, 2k, \ldots, r_1 k \leqslant r$, имеющую сумму S_{r_1}. Тогда набор $\{f_{n'k'}\}$ можно считать k-й субгармоникой, если выполняется неравенство

$$\frac{S_{r_1}}{S_r} > \frac{r_1}{r} + 2\left[\frac{r_1(r - r_1)}{r^2(r+1)}\right]^{1/2}, \quad (4.27)$$

а $S_r - S_{r_1}$ является статистически значимой суммой в смысле вероятности (4.15) для $r - r_1$ гармоник. Эти условия являются, вообще говоря, лишь необходимыми и предполагают приблизительно равномерное распределение амплитуд внутри каждой из групп для r_1 и $r - r_1$ гармоник (ср. рис. 8).

4.3. Периодичность $p = 3$ — фундаментальное свойство белок-кодирующих областей

Детальный анализ всего набора скрытых периодичностей в геномных ДНК оказывается довольно громоздким [21, 26]. Как уже отмечалось выше, важную роль в образовании периодичностей играет механизм умножения периодов. Другим важным фактором оказываются эффекты соизмеримости – несоизмеримости, когда несоизмеримые периоды p_1 и p_2 порождают статистически значимые периоды $P = n_1 p_1 \approx n_2 p_2$, где n_1 и n_2 — целые числа. Обычно эти эффекты проявляются лишь приближенно и требуют совместного анализа всех четырех спектров $f_{\alpha\alpha}(q_n)$, $\alpha \in (\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T})$. Поэтому поиск скрытых периодичностей с помощью суммы $\sum_\alpha f_{\alpha\alpha}(q_n)$ [82, 83, 85] или произведения $\prod_\alpha f_{\alpha\alpha}(q_n)$ [101] требует известной осторожности.

В качестве примера рассмотрим периодичность $p = 3$, которая является универсальным свойством белок-кодирующих областей [5, 41, 45] (это свойство было проверено путем сканирования по всем кодирующим последовательностям в базе данных). Поскольку для PHIX174 почти вся последовательность ДНК связана с кодирова-

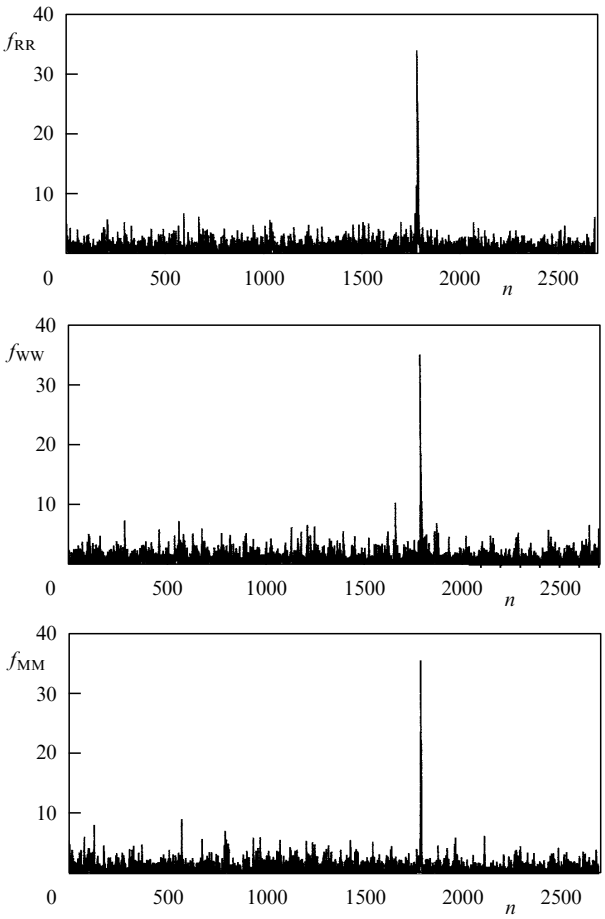

Рис. 9. Полуспектры для нормированных амплитуд гармоник для геномной последовательности PHIX174 при бинарных разбиениях R – Y, W – S и M – K.

нием белков, то соответствующие пики при $n = 1795$ хорошо видны на рис. 3. За немногими исключениями (см. ниже) эти пики сохраняются для любого из трех бинарных разбиений, R – Y, W – S, K – M (рис. 9). При этом учтено, что согласно (3.13) для бинарных последовательностей имеют место равенства $F_{RR}(q_n) = F_{YY}(q_n)$, $n \neq 0$, и т.д. Соответствующие максимальные относительные высоты (3.40) и номера гармоник для различных бинарных разбиений оказываются примерно равными: $f_{RR,\max} = 33{,}8$, $n = 1795$; $f_{WW,\max} = 35{,}0$, $n = 1797$; $f_{MM,\max} = 35{,}7$, $n = 1797$. Ближе всех к идеальному периоду $p = 3$ оказывается гармоника с $n = 1795$ для R – Y разбиения.

Как практически любая сильно выраженная периодичность, $p = 3$ в геноме PHIX174 порождает каскад структурных субгармоник [21]. На рисунке 10 для периодов, кратных трем, показаны максимальные отношения $(S_{r,w} - \langle S_{r,w}\rangle)/\langle(\Delta S_{r,w})^2\rangle^{1/2}$, которые вычислялись следующим образом. Высокие гармоники с номерами $1787 \leqslant n \leqslant 1803$ вблизи периода $p = 3$ исключались из спектра. Субгармоники с периодами $3k$ $(k = 2, \ldots, 20)$ задавались с помощью серий эквидистантных гармоник, начиная с номера n, дающего отношение M/n, наиболее близкое к $3k$, которые окружались затем окнами одинаковой ширины w (4.6). Средние значения для сумм и дисперсий сумм гармоник в случайных последовательностях определялись по формулам (4.21) и (4.22). Затем ширины окон менялись в пределах

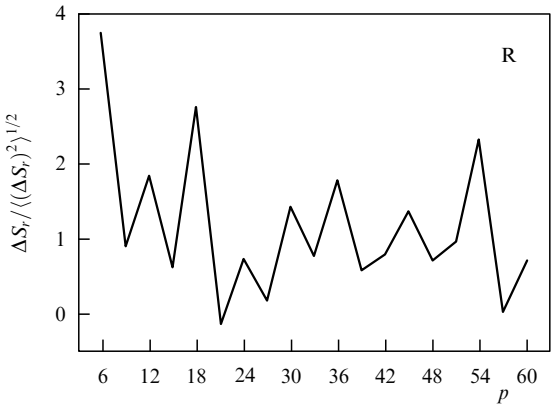

Рис. 10. Нормированные отклонения для сумм эквидистантных гармоник, отвечающих субгармоникам с периодами, кратными трем, для разбиения R – Y в геноме PHIX174.

$w = 0-10$ $(k = 2-9)$; $0-8$ $(k = 10)$; $0-5$ $(k = 11)$; $0-4$ $(k = 12)$; $0-3$ $(k = 13, 14)$; $0-2$ $(k = 15-20)$ и отбиралось значение, которое давало максимум для $(S_{r,w} - \langle S_{r,w}\rangle)/\langle(\Delta S_{r,w})^2\rangle^{1/2}$. Применение критериев раздела 4.2 приводит к выводу о статистической значимости субгармоник с периодами $p = 6, 18, 54$. Выделенная роль периода $p = 6$ возможно связана с элементами β-структуры (см. табл. 1).

Существует несколько объяснений происхождения и универсальности периодичности $p = 3$. В [96, 97] фундаментальная роль периодичности $p = 3$ связывается с эволюционным происхождением генетического кода и доминированием кодонов RRY [96] и RNY [97] (R — пурин, Y — пиримидин, N — произвольный нуклеотид) на ранней стадии молекулярной эволюции. Существенно, что периодичность $p = 3$ наблюдается также в транспортной РНК (или тРНК) [102], которая играет важную роль в синтезе белков по матричной РНК [29]. В [103] была высказана гипотеза о выделенной роли кодонов GCT в определении правильной рамки считывания при синтезе белков по мРНК, однако она не подтверждается экспериментально [104].

Очень часто оказывается, что вырождение генетического кода приводит к тому, что в третьих (менее значимых) позициях кодонов доминируют нуклеотиды одного типа [57, 105 – 109]. Приведем пример, когда этот эффект выражен особенно ярко. Для белка коллагена характерно повторение субъединиц из трех аминокислот [29, 33, 34], причем в первой позиции обычно находится глицин, кодируемый триплетами GGN (G — гуанин), во второй позиции доминирует пролин, кодируемый триплетами CCN (C — цитозин), а третья позиция заполняется относительно произвольно. На уровне кодирующей ДНК это означает, что для нуклеотидов G и C должна наблюдаться периодичность $p = 9$. На рисунке 11 показан соответствующий полуспектр для $f_{CC}(q_n)$ (код доступа для последовательности в базе данных ЕМБЛ X52046) с типичной серией эквидистантных пиков. Тем не менее спектр для $f_{TT}(q_n)$ опять имеет очень высокий пик для $p = 3$. Это показывает, что в третьих позициях кодонов GGN и CCN доминирует тимин.

Последовательность для PHIX174 кодирует, в основном, глобулярные белки, в то время как на рис. 11

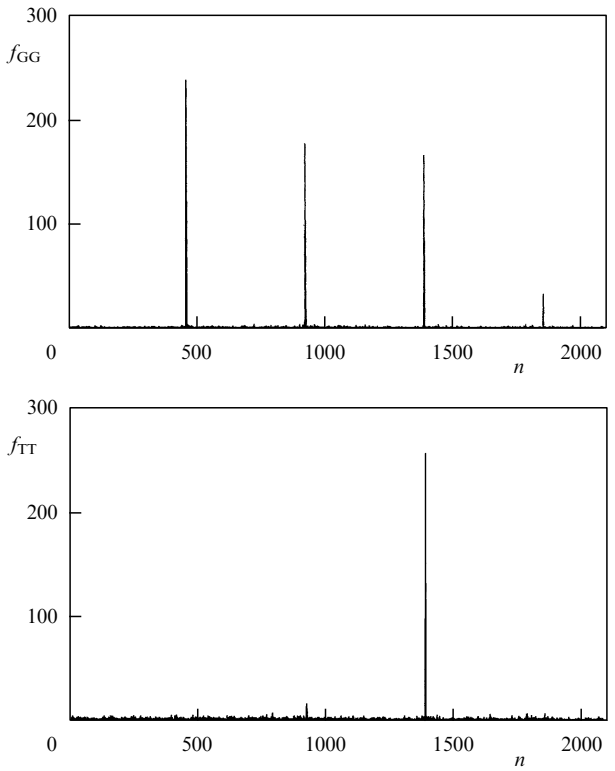

Рис. 11. Полуспектры для нормированных амплитуд гармоник для последовательности ДНК, кодирующей фибриллярный белок коллаген.

представлены спектры для последовательности, кодирующей фибриллярный белок с сильно вытянутой структурой [33, 34]. Для полноты приведем также спектр для последовательности, кодирующей трансмембранный белок, в котором полипептидная цепь несколько раз пронизывает мембрану, так что часть цепи находится во внеклеточном пространстве, часть — в цитоплазме, а часть — в мембране. Соответствующий полуспектр для последовательности (код доступа U59464), кодирующей трансмембранный белок PTCH, представлен на рис. 12. Приведен только полуспектр для A, поскольку остальные полуспектры имеют примерно такой же вид. В геноме человека этот ген представлен единственной копией, а мутации в нем приводят к заболеванию базально-клеточным раком [110, 111]. Данные о распределении гидрофобности вдоль цепи вместе с выделенными индикаторными последовательностями приводят к выводу о наличии шести цитоплазматических доменов, пяти внеклеточных доменов и десяти трансмембранных фрагментов в этом белке [112]. Поэтому, кроме характерного пика для периода $p = 3$, в спектре на рис. 12 можно видеть высокие гармоники с малыми волновыми числами, отражающие сегментацию и многодоменную структуру белка.

В [18, 21, 27] был указан еще один механизм генерации периодичности $p = 3$. Рассмотрим двойную цепь, в которой последовательность из пуринов R комплементарно располагается напротив последовательности из пиримидинов Y и пиримидины проникают в междоузлия между пуринами по механизму типа застегивания молнии:

$$\begin{matrix} \mathrm{R}\ldots\mathrm{R} \\ \mathrm{Y}\ldots\mathrm{Y} \end{matrix} \to \mathrm{RY}\ldots\mathrm{RY}\,,$$

и пусть этот процесс далее итерационно повторяется:

$$\begin{matrix} \ldots\mathrm{R}\ldots\mathrm{Y}\ldots \\ \ldots\mathrm{Y}\ldots\mathrm{R}\ldots \end{matrix} \to \ldots\mathrm{RY}\ldots\mathrm{YR}\ldots$$

и т.д. Легко видеть, что эти итерации можно представить в виде последовательных подстановок $\mathrm{R} \to \mathrm{RY}$, $\mathrm{Y} \to \mathrm{YR}$. В результате получаем одну из самых известных подстановочных моделей, так называемую последовательность Туэ–Морса, впервые введенную в теории чисел в 1906 г. Спектральные характеристики для этой последовательности хорошо изучены [113, 114]. В частности, можно строго показать, что доминирующей является периодичность $p = 3$. В такой модели реализуется механизм умножения периода. На рисунке 13 показан спектр для последовательности, полученной после 12 итераций, начиная с R ($M = 4096$). Отметим также ряд характерных пиков в области между периодами $2 < p < 3$, часто наблюдаемыми и в последовательностях ДНК. Вероятностная реализация этой модели [115] позволяет получить в области малых волновых чисел спектры типа шума $1/q^{\nu}$, имитирующие эффекты сегментации. Таким образом, в рамках данного меха-

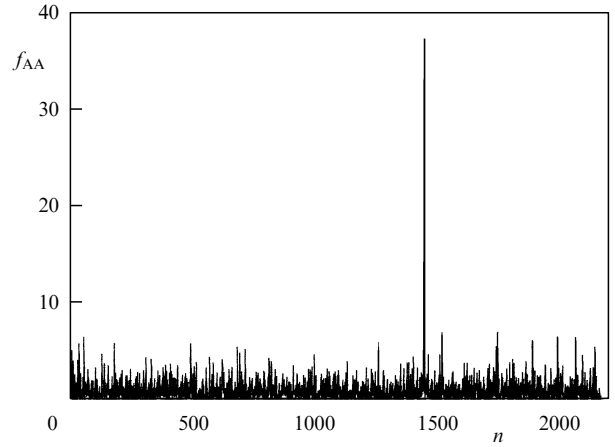

Рис. 12. Полуспектры для нормированных амплитуд $f_{AA}(q_n)$ для последовательности ДНК, кодирующей трансмембранный белок PTCH.

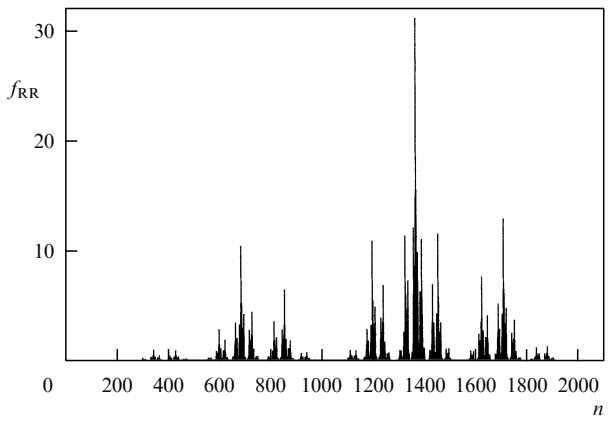

Рис. 13. Полуспектр для нормированных амплитуд гармоник для последовательности Туэ–Морса, полученной после 12 итераций $\mathrm{R} \to \mathrm{RY}$, $\mathrm{Y} \to \mathrm{YR}$, начиная с R.

низма периодичность $p = 3$, умножение периода и другие особенности оказываются прямым следствием комплементарности.

Фрагменты из триплетных повторов играют также важную структурную и регуляторную роли в геноме человека. Пролиферация повторов (увеличение длины фрагментов) может привести к тяжелым генетическим заболеваниям [116–118]. Однако пока трудно сказать, насколько выделенными именно в данном контексте (в отличие от их функции в белок-кодирующих областях) являются фрагменты из триплетных повторов среди других многочисленных семейств повторов.

4.4. Другие периодичности

Из других скрытых периодичностей в геномных последовательностях ДНК наиболее характерными являются, по-видимому, периодичности, связанные с шагом двойной спирали [38, 85, 119]. В [120] было показано, что в экзон-интронных последовательностях в геномах человека и мыши гистограммы для расстояний между 5′-концом для одного из интронов и 3′-концами первого и всех последующих экзонов проявляют квазипериодические вариации с периодом $p \approx 204-205$. Анализ скрытых периодичностей в экзон-интронных последовательностях [26] показывает, что эта периодичность действительно относительно устойчиво воспроизводится. Однако более правильно говорить о наборе приближенно воспроизводимых периодичностей в этом диапазоне длин. Эти периодичности, вероятно, связаны с упаковкой ДНК в нуклеосомы. Разнообразие путей молекулярной эволюции приводит к тому, что в разных геномах периодичность $p \approx 204-205$ по-разному проявляется в спектрах для нуклеотидов разных типов и может реализовываться, в том числе, по механизму умножения периода. Стоит отметить также роль эффектов соизмеримости–несоизмеримости: периодичность $p \approx 204-205$ охватывает примерно 20 шагов двойной спирали в B-форме.

В [121] была отмечена довольно частая встречаемость скрытой периодичности $p = 2$ в интронах, указывающая на модификацию Z-формы и возможные динамические B–Z переходы (см. табл. 1). Фрагменты из строго чередующихся пуринов и пиримидинов были также идентифицированы только в интронах [122]. Для крайних граничных 5′ и 3′ фрагментов в экзон-интронных областях нередко наблюдается совместное существование периодичностей $p = 2$ и $p = 3$ [123]. Эти скрытые периодичности выявляются, однако, лишь статистически и воспроизводятся гораздо менее устойчиво, чем периодичность $p = 3$ в белок-кодирующих областях. Интересно также отметить, что периодичность $p = 2$ появляется на первом шаге итераций в модели с последовательностью Туэ–Морса.

Вернемся к примеру генома бактериофага PHIX174. Вставки на рис. 3 и формула (3.48) указывают на существование пяти больших сверхпериодов в этой последовательности. На рисунке 14 представлены данные для отношений $(S_{r,w} - \langle S_{r,w}\rangle)/\langle(\Delta S_{r,w})^2\rangle^{1/2}$ для эквидистантных серий, последовательно генерируемых, начиная с $n = 3-51$. При этом высокие гармоники в области $1787 \leqslant n \leqslant 1803$ были удалены из спектра, а эквидистантные гармоники, начиная с третьей, окружались окнами шириной $w = 1$. На рисунке 14 показаны результаты только для C (более подробно см. [21]). Как видно из

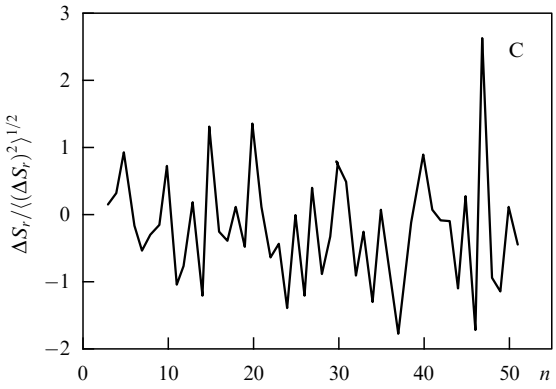

Рис. 14. Нормированные отклонения для сумм эквидистантных гармоник $f_{CC}(q_n)$, генерируемых, начиная с номера n, для геномной последовательности PHIX174.

рис. 14, практически для каждого n, кратного пяти, наблюдается пик, что подтверждает вывод о наличии пяти больших сверхпериодов в этой последовательности. Хорошо известно [95], что белковая оболочка (капсид), покрывающая кольцевую ДНК бактериофага PHIX174, имеет форму правильного додекаэдра. Возможно, что сверхпериоды $p = M/5$ связаны с упаковкой ДНК и осями симметрии пятого порядка, характерными для додекаэдра [124].

Эти примеры показывают, что в скрытых периодичностях для последовательностей ДНК заключена самая разнообразная эволюционная и структурная информация. Методы явной реконструкции случайно модифицированных скрытых повторов можно найти в [22, 25].

5. Анализ корреляций в геномных последовательностях ДНК

На первом этапе анализа последовательностей ДНК задача состоит в разбиении последовательности на сегменты с функционально разными свойствами (белок-кодирующие сегменты; области, кодирующие тРНК; сателлитная ДНК и т.д.). После этого на втором этапе возникает задача о структурной связи между различными сегментами. В этом разделе мы обсудим применение спектральных методов к анализу корреляций между положениями нуклеотидов в последовательности.

5.1. Метод "заметания" в спектрах для структурных факторов и корреляционных функций

Рассмотрение начнем с более простого случая без специального предварительного выделения эффектов сегментации. Крупномасштабные вариации плотности на длинах, сравнимых с полной длиной последовательности, можно изучать с помощью сглаженных среднеквадратичных отклонений для корреляционных функций (ср. (3.53) и рис. 6) или сглаженных фурье-спектров:

$$\widetilde{f}_{\alpha\alpha}(q_n) = (2k+1)^{-1} \sum_{n'=n-k}^{n+k} f_{\alpha\alpha}(q_{n'})\,, \tag{5.1}$$

где $f_{\alpha\alpha}(q_{n'})$ определяются в (3.40). Статистические критерии для $\widetilde{f}_{\alpha\alpha}(q_n)$ можно найти в [17, 18]. Здесь мы ограничимся более качественным уровнем.

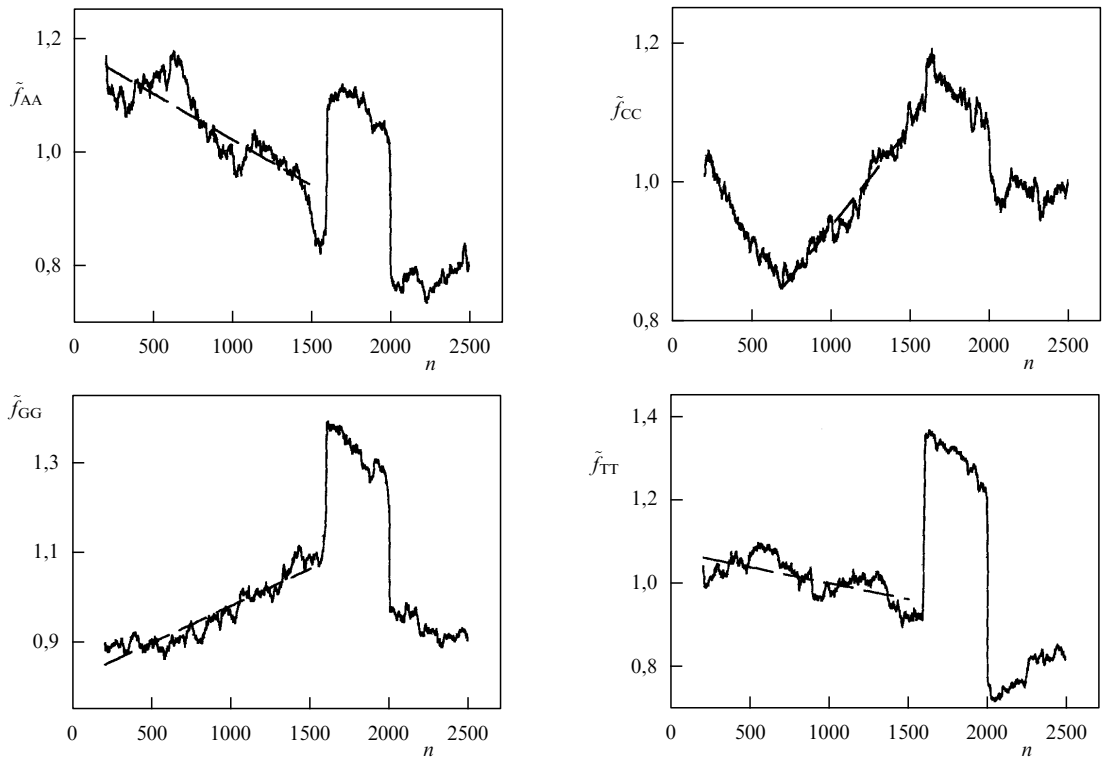

Рис. 15. Сглаженные полуспектры с $k = 200$ (формула (5.1)) для генома PHIX174.

Поясним основные идеи на нашем стандартном примере генома PHIX174. Сглаженные полуспектры с $k = 200$ в (5.1) для PHIX174 изображены на рис. 15. В области малых волновых чисел спектры для A и T обнаруживают устойчивую тенденцию к росту при уменьшении q_n, а спектры для C и G — тенденцию к убыванию (не вполне регулярную). Следуя стандартной терминологии [125], будем называть такое поведение персистентным в первом случае и, соответственно, антиперсистентным во втором. Если выделить окно шириной $W \sim 1/q_n$ и перемещать его вдоль последовательности, то в первом случае наблюдается тенденция к сохранению характера вариаций, а во втором — к его изменению.

Одним из стандартных методов анализа вариаций является метод кривых Херста (или метод нормированного размаха) [125]. Покажем, как поведение сглаженных спектров согласуется с поведением кривых Херста [19]. Выберем произвольный сайт m и значение для ширины интервала m_0. Определяя среднюю плотность нуклеотидов в интервале от m до $m + m_0 - 1$

$$\bar{\rho}_\alpha = m_0^{-1} \sum_{m'=m}^{m+m_0-1} \rho_{m',\alpha}\,, \tag{5.2}$$

введем отклонение

$$\Delta_\alpha(m, \tilde{m}) = \sum_{m'=m}^{\tilde{m}} (\rho_{m',\alpha} - \bar{\rho}_\alpha)\,, \qquad m \leqslant \tilde{m} \leqslant m + m_0 - 1 \tag{5.3}$$

и вычислим размах

$$R_\alpha(m, m + m_0 - 1) = \max_{m \leqslant \tilde{m} \leqslant m+m_0-1} \Delta_\alpha(m, \tilde{m}) - \\ - \min_{m \leqslant \tilde{m} \leqslant m+m_0-1} \Delta_\alpha(m, \tilde{m}). \tag{5.4}$$

Далее вычисляется отношение размаха $R_\alpha(m, m + m_0 - 1)$ к стандартному отклонению

$$\sigma(\rho_\alpha) = \left[m_0^{-1} \sum_{m'=m}^{m+m_0-1} (\rho_{m',\alpha} - \bar{\rho}_\alpha)^2 \right]^{1/2}, \tag{5.5}$$

и отношение $R_\alpha/\sigma(\rho_\alpha)$ при заданном текущем значении m_0 усредняется по всем сайтам m. Среднее отношение $\overline{R_\alpha/\sigma(\rho_\alpha)}$ сравнивается с аналогичной величиной $\langle \overline{R_\alpha/\sigma(\rho_\alpha)} \rangle$, полученной в результате усреднения по ансамблю случайных последовательностей с тем же нуклеотидным составом.

Соответствующие зависимости для $\overline{R_\alpha/\sigma(\rho_\alpha)}$, $\langle \overline{R_\alpha/\sigma(\rho_\alpha)} \rangle$ и

$$\Delta \ln\left[\overline{R_\alpha/\sigma(\rho_\alpha)} \right] = \ln\left[\overline{R_\alpha/\sigma(\rho_\alpha)} \right] - \ln\left[\left\langle \overline{R_\alpha/\sigma(\rho_\alpha)} \right\rangle \right]$$

для PHIX174 и $10 \leqslant m_0 \leqslant 1000$ показаны на рис. 16. Мы видим, что отклонения $\Delta \ln\left[\overline{R_\alpha/\sigma(\rho_\alpha)} \right]$ для A и T оказываются положительными, а для G и C, в основном, отрицательными в соответствии с поведением сглаженных спектров на рис. 15. Качественно такое поведение можно понять, используя спектры типа $F(q_n) \propto 1/q^\nu$ с постоянным показателем ν. Тогда для нормированного размаха получаем $\overline{R/\sigma} \propto m_0^{(1+\nu)/2}$, и отклонения от зависимости для случайных последовательностей определяются знаком ν. Пунктирные кривые на рис. 16 были построены при усреднении по 20 случайным реализациям. При усреднении по m учитывалось, что геном PHIX174 является кольцевым, и функции положения $\rho_{m,\alpha}$ заменялись на циклически продолженные функции $\rho^{\mathrm{c}}_{m,\alpha}$ (3.8). Разброс в значениях $\ln\left[\overline{R_\alpha/\sigma(\rho_\alpha)} \right]$ для случайных последовательностей растет с ростом m_0 и составляет примерно $\left\langle \left\{ \Delta \ln\left[\overline{R_\alpha/\sigma(\rho_\alpha)} \right] \right\}^2 \right\rangle^{1/2} \approx 0{,}01$ при $m_0 =$

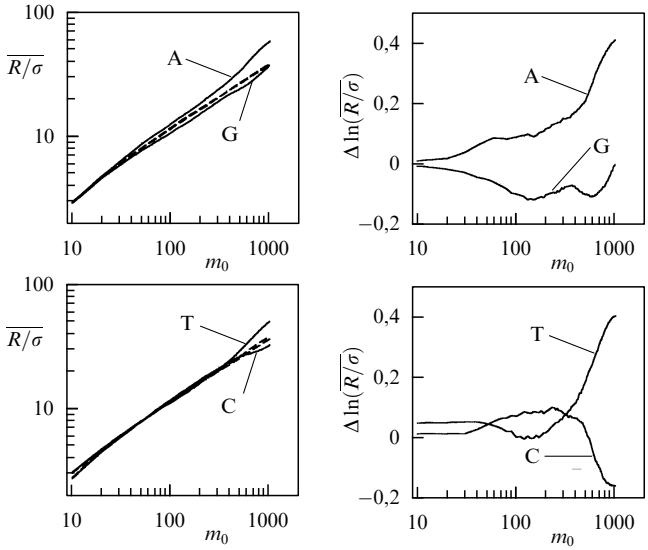

Рис. 16. Кривые Херста для генома PHIX174. Штриховыми линиями показаны кривые Херста, полученные в результате усреднения по 20 реализациям для случайных последовательностей с тем же нуклеотидным составом.

$= 100$ и 0,05 при $m_0 = 1000$, что позволяет сделать вывод о статистической значимости наблюдаемых отклонений для $\Delta\ln\left[\overline{R_\alpha/\sigma(\rho_\alpha)}\right]$. Антиперсистентное поведение кривой Херста для C хорошо согласуется с выводом о пяти больших сверхпериодах для C (см. рис. 14), поскольку квазипериодические вариации дают одну из реализаций антиперсистентного поведения. Немонотонный характер зависимости $\Delta\ln\left[\overline{R_\alpha/\sigma(\rho_\alpha)}\right]$ от m_0 означает наличие набора характерных длин (в данном случае эти длины порядка размеров генов). Такое немонотонное поведение типично и для других геномных последовательностей ДНК [15, 19, 63–65]. Кривые на рис. 16 показывают еще раз, что изучение должно начинаться на четырехнуклеотидном уровне. Так, для бинарного объединения $\mathrm{R} = (\mathrm{A}, \mathrm{G})$ и $\mathrm{Y} = (\mathrm{C}, \mathrm{T})$ персистентное и антиперсистентное поведение вариаций для A и G, и, соответственно, для T и C могут взаимно компенсировать друг друга и ослаблять вариации на больших масштабах.

Перейдем теперь к анализу взаимных корреляций между нуклеотидами разных типов. Интегральную оценку для значимости корреляций между различными нуклеотидами можно получить с помощью (3.32), (3.36) и (3.38). Для случайных последовательностей вероятность того, что отклонение

$$x_{\alpha\beta} = \frac{k(F_{\alpha\alpha}|F_{\beta\beta};\, M-1) - \langle k_{\alpha\beta}\rangle}{\left[2\left\langle(\Delta k_{\alpha\beta})^2\right\rangle\right]^{1/2}} \tag{5.6}$$

заключено в пределах $-x \leqslant x_{\alpha\beta} \leqslant x$ (или $|x_{\alpha\beta}| \leqslant x$), равна

$$P\big(|x_{\alpha\beta}| \leqslant x\big) = 2\pi^{-1/2}\int_0^x \mathrm{d}x'\,\exp(-x'^2) \equiv \mathrm{erf}\,(x)\,, \tag{5.7}$$

и, соответственно, для вероятности превышения $|x_{\alpha\beta}| > x$ получаем:

$$P\big(|x_{\alpha\beta}| > x\big) = 1 - \mathrm{erf}\,(x)\,. \tag{5.8}$$

Задавая пороги $P\big(|x_{\alpha\beta}| > x\big) = 0{,}1$ и 0,05, находим $x = 1{,}16$ и 1,39, что позволяет оценить значимость корреляций, наблюдаемых в геномных последовательностях ДНК.

В приложениях важно установить не только наличие или отсутствие значимых корреляций, но и выяснить их источник. Так, со структурной точки зрения корреляции могут быть обусловлены совпадающими скрытыми периодичностями, крупномасштабными вариациями плотности, короткодействующей связью или когерентными точечными мутациями. Происхождение корреляций можно установить с помощью метода "заметания" в спектрах для элементов структурного фактора и корреляционных функций [23, 27]. В этом методе вычисляются коэффициенты корреляции для части спектра $k(F_{\alpha\alpha}|F_{\beta\beta};\, 1 \leqslant n' \leqslant n)$ и $k(K_{\alpha\alpha}|K_{\beta\beta};\, 1 \leqslant m_0' \leqslant m_0)$, и значения n и m_0 последовательно увеличиваются от 1 до N (см. (3.11)) ("заметание" слева направо). Аналогично вычисляются коэффициенты $k(F_{\alpha\alpha}|F_{\beta\beta};\, n \leqslant n' \leqslant N)$ и $k(K_{\alpha\alpha}|K_{\beta\beta};\, m_0 \leqslant m_0' \leqslant N)$, где n и m_0 последовательно уменьшаются от N до 1 ("заметание" справа налево). Напомним, что для интегральных коэффициентов корреляции справедливо равенство (3.35), однако оно нарушается при вычислении для части спектра. Сравнивая зависимости текущих коэффициентов корреляции от n и m_0 при "заметании" слева направо и наоборот, можно выделить вклады от различных частей спектра и установить основные источники корреляций. Так, например, если наблюдаются скачки в зависимостях $k(F_{\alpha\alpha}|F_{\beta\beta};\, 1 \leqslant n' \leqslant n)$ и $k(F_{\alpha\alpha}|F_{\beta\beta};\, n \leqslant n' \leqslant N)$, то это свидетельствует о совпадающих (в случае корреляций) или сдвинутых (в случае антикорреляций) периодичностях. Используя взаимную дополнительность между волновыми числами q_n и масштабами m_0 для преобразования Фурье (см. (3.6)), $m_0 \propto 1/q_n$, корреляции на больших расстояниях можно детально исследовать с помощью $k(K_{\alpha\alpha}|K_{\beta\beta};\, m_0 \leqslant m_0' \leqslant N)$, поскольку такие коэффициенты корреляции содержат только корреляционные функции $K(m_0')$ с $m_0' \geqslant m_0$, а для изучения корреляций на малых расстояниях удобно использовать $k(F_{\alpha\alpha}|F_{\beta\beta};\, n \leqslant n' \leqslant N)$, так как в них входят только структурные факторы с волновыми числами $q_{n'} \geqslant q_n \propto 1/m_0$.

На рисунке 17 слева показаны графики для $k_{\mathrm{rel}} = (k_{\alpha\beta} - \langle k_{\alpha\beta}\rangle)/\langle k_{\alpha\beta}\rangle$ (см. (3.37)) в случае корреляций A–T в геноме PHIX174. "Заметания" для $k_{\mathrm{rel}}(F_{\mathrm{AA}}|F_{\mathrm{TT}})$ четко выявляют важную роль совпадающих периодичностей с $p = 3$ (ср. рис. 3 и скачки при $n \approx 1795$ на рис. 17). Зависимость $k_{\mathrm{rel}}(K_{\mathrm{AA}}|K_{\mathrm{TT}};\, m_0 \leqslant m_0' \leqslant N)$ от m_0 свидетельствует о значимости корреляций на больших расстояниях и определенной структурной целостности всего генома. Действительно, если разбить геном на фрагменты с $l = 500$ (средний размер гена для PHIX174) и случайно их перетасовать, то мы получим зависимости, изображенные на рис. 17 справа. Дополнительный рост для $k_{\mathrm{rel}}(K_{\mathrm{AA}}|K_{\mathrm{TT}};\, m_0 \leqslant m_0' \leqslant N)$ при $m_0 \leqslant 500$ свидетельствует об иерархической структуре корреляций: корреляции внутри генов оказываются сильнее, чем корреляции между генами. В этом смысле можно говорить о гене как о структурной единице.

Зависимости на рис. 17 свидетельствуют о том, что периодичность $p = 3$ когерентно пронизывает кластеры из нескольких генов. В этом можно убедиться и более прямым образом [20], изучая масштабные зависимости для $f_{\alpha\alpha}(q_n)$ (см. (3.40)) при q_n, отвечающем $p = 3$, для фрагментов последовательности от 1 до L по мере

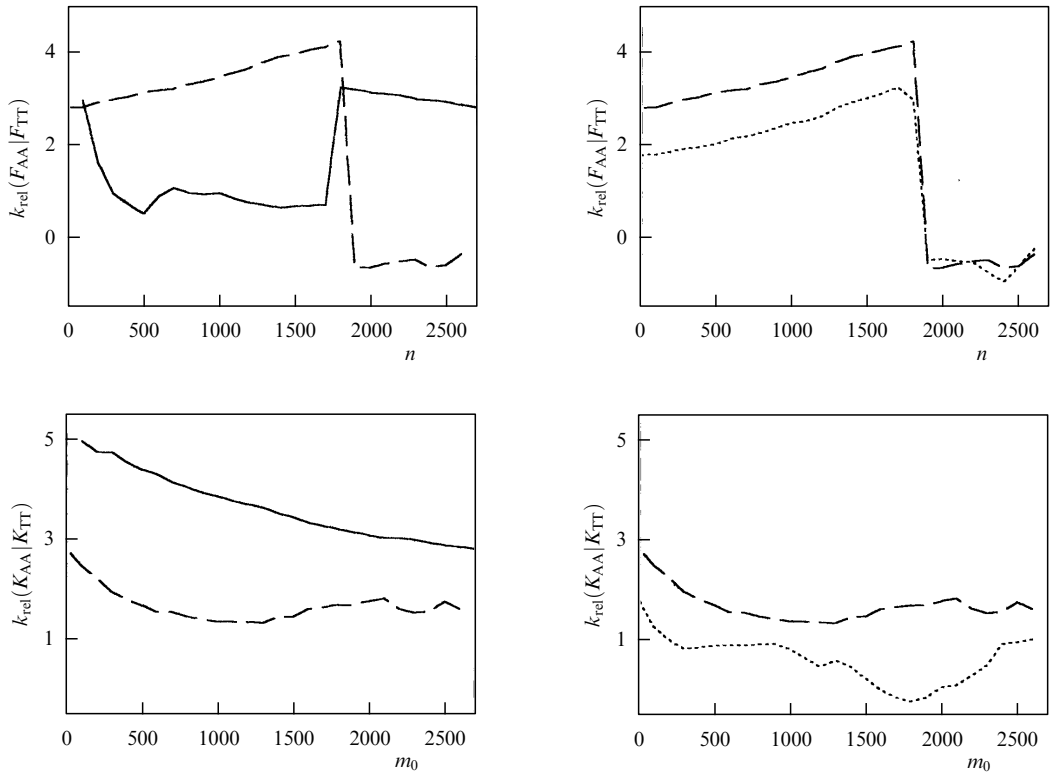

Рис. 17. Нормированные относительные текущие коэффициенты корреляции при расширении спектра слева направо (сплошные кривые) и справа налево (штриховые кривые) для корреляций A – T в геноме PHIX174 (слева). Кривые для "заметания" справа налево в спектрах для генома PHIX174 (штриховые кривые) и для последовательности, полученной после разбиения генома PHIX174 на сегменты длиной $l = 500$ с последующей случайной перетасовкой (пунктир) (справа).

увеличения L. При отсутствии структурной связи будут иметь место квазислучайные вариации в окрестности постоянного значения, а при наличии когерентной связи должен наблюдаться рост амплитуды $f_{\alpha\alpha}(q_n)$ при увеличении L. Если периодичность $p = 3$ является идеальной, то $f \propto L$. Если имеются два сегмента с длинами L_1 и L_2, имеющие периодичности $p = 3$ со сдвигом в начале отсчета, то $f \propto (L_1^2 + L_2^2 - L_1 L_2)/(L_1 + L_2)$. В этом случае получим минимум при $L_2 = L_1(\sqrt{3} - 1)$ и характерную зависимость типа зубца. Наконец, если сегмент длиной L_c с идеальной периодичностью $p = 3$ примыкает к сегменту длиной L_r со случайным распределением нуклеотидов, то $f \propto (L_c^2 + L_r)/(L_c + L_r)$.

На рисунке 18 показаны соответствующие зависимости для бинарных последовательностей R – Y для бактериофагов PHIX174 и MIG4XX (код доступа для последовательности J02454) и РНК-вируса TOEAV (код доступа X53459). Эти кривые свидетельствуют о структурной связи и модульном объединении нескольких генов (см. также общее обсуждение аналогичных механизмов в [4, 5, 28, 29]). Эффект частично обусловлен перекрытием некоторых генов в данных геномах. Однако модули охватывают и неперекрывающиеся гены. Разбиение генома PHIX174 на сегменты длиной $l = 500$ и 100 и случайная перетасовка сегментов разрушает структурную связь (рис. 19).

Геном MIG4XX был выбран отчасти потому, что PHIX174 и MIG4XX занимают близкие узлы на эволюционном древе. Оба бактериофага нападают на бактерию Escherichia coli, и имеется взаимно однозначное соответствие между генами PHIX174 и MIG4XX. Тем не менее и структурные характеристики [19], и нуклеотидный состав соответствующих последовательностей ДНК оказываются весьма различными. Согласно общей концепции [95], вирусы (а бактериофаги представляют

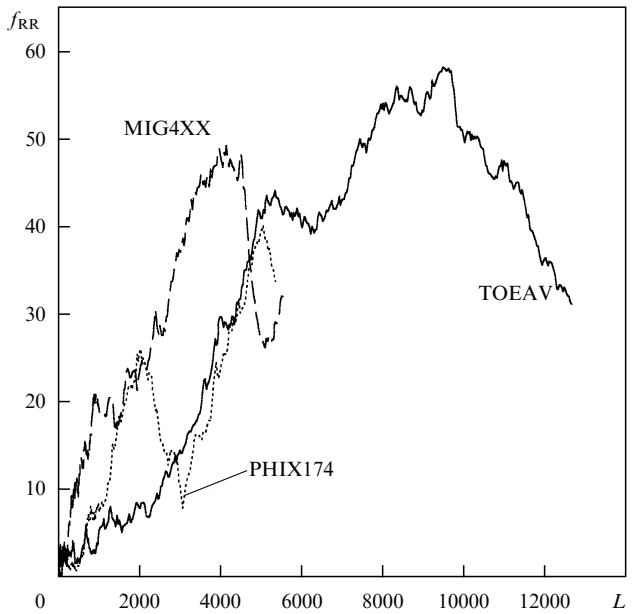

Рис. 18. Масштабная зависимость для нормированной амплитуды $f_{RR}(q_n)$ с q_n, отвечающим периодичности $p = 3$, при увеличении начального участка последовательности с $1 \leqslant m \leqslant L$ в геномах PHIX174, MIG4XX и TOEAV.

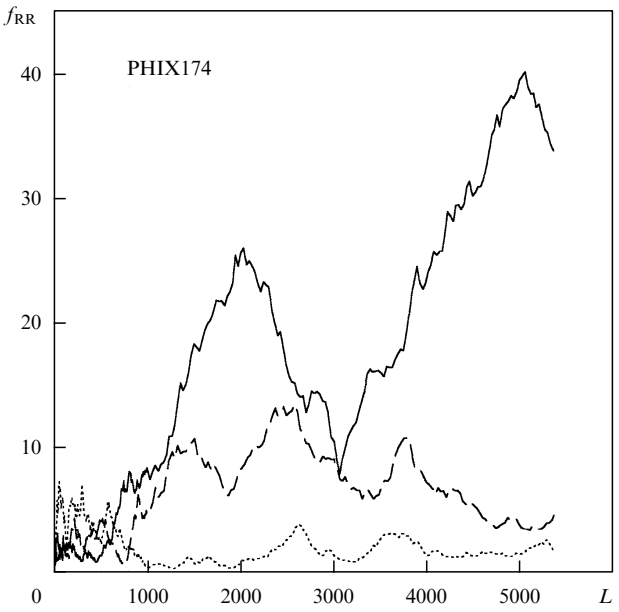

Рис. 19. Масштабная зависимость для нормированной амплитуды $f_{RR}(q_n)$ с q_n, отвечающим периодичности $p = 3$, при увеличении начального участка последовательности с $1 \leqslant m \leqslant L$ в геноме PHIX174 (сплошная кривая) и для последовательностей, полученных после разбиения генома на сегменты длиной $l = 500$ и 100 с последующей случайной перетасовкой (штриховая кривая и пунктир соответственно).

отдельную подгруппу вирусов) можно рассматривать как "сбежавшую" ДНК. Это короткие фрагменты ДНК, которые могут существовать относительно самостоятельно, хотя для их репликации и требуются белки клетки-хозяина. Дополнительный анализ показывает, что большие фрагменты ДНК для PHIX174 и MIG4XX являются взаимно комплементарными [22, 25]. Так как данный эволюционный механизм является довольно общим [4, 41], его полезно иметь в виду при изучении соответствия между последовательностями ДНК для эволюционно близких организмов, поскольку близость между последовательностями не обязательно может оказаться прямой.

5.2. Анализ корреляций в сегментированных последовательностях ДНК

Вопрос о наличии или отсутствии структурной связи между различными элементами мозаичной структуры генома является одним из наиболее важных и принципиальных при анализе последовательностей ДНК. В этом разделе мы опишем методы анализа корреляций, когда эффекты сегментации учитываются с самого начала [27].

Пусть последовательность ДНК длиной M состоит из K неперекрывающихся сегментов с длинами $\{L_k\}$, $(k = 1, \ldots, K)$,

$$M = \sum_{k=1}^{K} L_k \,. \tag{5.9}$$

Определим средние плотности нуклеотидов внутри каждого сегмента

$$\bar{\rho}_\alpha^{(k)} = \frac{N_\alpha^{(k)}}{L_k} \,, \tag{5.10}$$

и введем дифференциальные функции положения (ср. (3.1))

$$\tilde{\rho}_{m,\alpha} = \rho_{m,\alpha} - \bar{\rho}_{m,\alpha} \,, \tag{5.11}$$

где $\bar{\rho}_{m,\alpha} = \bar{\rho}_\alpha^{(k)}$, если сайт m находится внутри k-го сегмента. Фурье-гармоники определяются согласно

$$\rho_\alpha(q_n) = M^{-1/2} \sum_{m=1}^{M} \tilde{\rho}_{m,\alpha} \exp(-\mathrm{i} q_n m) \,,$$

$$q_n = \frac{2\pi n}{M} \,, \qquad n = 0, 1, \ldots, M-1 \,, \tag{5.12}$$

а все остальные соотношения определяются так же, как в (3.5), (3.6), (3.10) и (3.11).

При вычислении средних значений для диагональных элементов структурного фактора $\langle F_{\alpha\alpha}(q_n)\rangle$ и корреляционных функций $\langle K_{\alpha\alpha}(m_0)\rangle$ усреднение производится по ансамблю случайных последовательностей с теми же длинами сегментов $\{L_k\}$ и тем же составом нуклеотидов внутри каждого сегмента, как и в исходной последовательности. Усреднение внутри каждого сегмента производится независимо и определяется формулами (3.27) и (3.28). В результате получаем:

$$\left\langle F_{\alpha\alpha}(q_n)\right\rangle = M^{-1} \sum_{k=1}^{K} L_k \bar{F}_{\alpha\alpha}^{(k)} \left[1 - \frac{\sin^2(q_n L_k/2)}{L_k^2 \sin^2(q_n/2)}\right] , \tag{5.13}$$

$$\bar{F}_{\alpha\alpha}^{(k)} = \frac{N_\alpha^{(k)}(L_k - N_\alpha^{(k)})}{L_k(L_k - 1)} \tag{5.14}$$

и

$$\left\langle K_{\alpha\alpha}(m_0)\right\rangle = -\sum_{k=1}^{K} \frac{\widetilde{K}(m_0, L_k)\bar{F}_{\alpha\alpha}^{(k)}}{L_k} \,, \tag{5.15}$$

$$\widetilde{K}(m_0, L) = \theta(L - m_0)\,\frac{L - m_0}{M} + \\ + \theta(L + m_0 - M)\,\frac{L + m_0 - M}{M} \,, \tag{5.16}$$

где $\theta(x)$ — функция Хевисайда, $\theta(x) = 1$ при $x > 0$ и $\theta(x) = 0$ при $x \leqslant 0$. Далее вычисляются коэффициенты корреляции для части спектра для разностей

$$\Delta F_{\alpha\alpha}(q_n) = F_{\alpha\alpha}(q_n) - \left\langle F_{\alpha\alpha}(q_n)\right\rangle \,,$$
$$\Delta K_{\alpha\alpha}(m_0) = K_{\alpha\alpha}(m_0) - \left\langle K_{\alpha\alpha}(m_0)\right\rangle \tag{5.17}$$

и используется метод "заметания", как было описано в разделе 5.1. Для случайных сегментированных последовательностей средние значения для интегральных коэффициентов корреляции равны

$$\left\langle k(F_{\alpha\alpha}|F_{\beta\beta};\, 1 \leqslant n' \leqslant N)\right\rangle \equiv \langle k_{\alpha\beta}\rangle = \frac{\bar{F}_{\alpha\beta}^2}{\bar{F}_{\alpha\alpha}\bar{F}_{\beta\beta}} \,, \tag{5.18}$$

$$\bar{F} = M^{-1} \sum_{k=1}^{K} \bar{F}^{(k)} L_k \,, \tag{5.19}$$

$$\left\langle (\Delta k_{\alpha\beta})^2\right\rangle \approx \frac{1}{N} \,, \tag{5.20}$$

где средние $\bar{F}^{(k)}$ для отдельных сегментов определяются аналогично (3.14) с заменами M на L_k и N_α на $N_\alpha^{(k)}$ (ср. также (5.14)).

Метод легко обобщается на случай, когда сегменты разбиваются на различные классы. Для определенности будем рассматривать чередующиеся экзон-интронные фрагменты [29, 31]. Исходную последовательность будем обозначать как eiei... Далее введем дифференциальные функции положения двух типов, $\tilde{\rho}^{(e)}_{m,\alpha}$ и $\tilde{\rho}^{(i)}_{m,\alpha}$, где $\tilde{\rho}^{(e)}_{m,\alpha}$ определяется по формуле (5.11), если m находится внутри экзонов, и равна нулю в другом случае. Такая функция $\tilde{\rho}^{(e)}_{m,\alpha}$ определяет последовательность e0e0... с нулями на месте интронных фрагментов. Аналогично определяется функция $\tilde{\rho}^{(i)}_{m,\alpha}$ и последовательность 0i0i... Далее по функциям $\tilde{\rho}^{(e)}_{m,\alpha}$ и $\tilde{\rho}^{(i)}_{m,\alpha}$ вычисляются фурье-гармоники $\tilde{\rho}^{(e)}_{\alpha}(q_n)$ и $\tilde{\rho}^{(i)}_{\alpha}(q_n)$ (ср. (5.12)) и т.д. Для средних $\langle F^{(e)}_{\alpha}(q_n)\rangle$ и $\langle K^{(e)}_{\alpha}(m_0)\rangle$ суммирование в (5.13) и (5.15) производится теперь только по экзонным сегментам, и, аналогично, для $\langle F^{(i)}_{\alpha}(q_n)\rangle$ и $\langle K^{(i)}_{\alpha}(m_0)\rangle$ учитываются только интронные сегменты. Разбиение на последовательности e0e0... и 0i0i... позволяет изучать корреляции не только между отдельными сегментами, но и выделить роль позиционных эффектов, связанных с распределением длин $\{L_k\}$.

Наибольший интерес представляют коэффициенты корреляции

$$k\big(F^{(e)}_{\alpha\alpha}|F^{(i)}_{\beta\beta};\ 1\leqslant n\leqslant N\big)=k\big(K^{(e)}_{\alpha\alpha}|K^{(i)}_{\beta\beta};\ 1\leqslant m_0\leqslant N\big)$$

и их аналоги в методе "заметания". Для случайных сегментированных последовательностей справедливы оценки

$$\big\langle k(F^{(e)}_{\alpha\alpha}|F^{(i)}_{\beta\beta};\ 1\leqslant n\leqslant N)\big\rangle\equiv\big\langle k^{(e-i)}_{\alpha\beta}\big\rangle=0\,, \tag{5.21}$$

$$\big\langle(\Delta k^{(e-i)}_{\alpha\beta})^2\big\rangle\approx\min\left(\frac{2}{L_e}\,,\ \frac{2}{L_i}\right), \tag{5.22}$$

$$L_e=\sum_{\text{экзоны}}L^{(e)}_k\,,\qquad L_i=\sum_{\text{интроны}}L^{(i)}_k\,. \tag{5.23}$$

Дисперсия (5.22) учитывает, что полное число случайных степеней свободы в данном случае равно приближенно $\max(L_e, L_i)$. Статистическая значимость взаимных корреляций оценивается в терминах гауссовых переменных

$$x^{e-i}_{\alpha\beta}=\frac{k\big(F^{(e)}_{\alpha\alpha}|F^{(i)}_{\beta\beta};\ 1\leqslant n\leqslant N\big)}{\big[2\big\langle(\Delta k^{(e-i)}_{\alpha\beta})^2\big\rangle\big]^{1/2}}\,. \tag{5.24}$$

Вероятность того, что хотя бы для одной из s независимых гауссовых переменных будет выполняться неравенство $|x_{\alpha\beta}|>x$, равна

$$P\big(|x_{\alpha\beta}|>x;\ s\big)=1-\big[\operatorname{erf}(x)\big]^s\,. \tag{5.25}$$

Отсюда для $s=16$ получаем, что 10%- и 5%-ному порогам статистической значимости отвечают значения $x=1{,}92$ и $2{,}04$, соответственно.

В качестве примера рассмотрим корреляции между последовательностями e0e0... и 0i0i... для коллагенового гена (код доступа X52046) без крайних 5′ и 3′ пар экзонов и интронов. Заметим, что в этих обозначениях спектры на рис. 11 отвечают последовательности ee..., полученной после сплайсинга. Данная последовательность содержит 49 экзонов (с длинами $\{L^{(e)}_k\}$ от 54 до 295) и 48 интронов (с длинами $\{L^{(i)}_k\}$ от 82 до 1561), $L_e=4163$, $L_i=22448$, $M=26611$ (см. (5.23)). Сводные данные для интегральных коэффициентов корреляции представлены в табл. 2. В скобках указаны значения для нормированных переменных (5.24).

Таблица 2. Коэффициенты корреляции между экзонами и интронами в коллагеновом гене (X52046)

	Ae	Ge	Te	Ce
Ai	−0,002	−0,003	−0,018	0,006
	(0,12)	(0,23)	(1,36)	(0,46)
Gi	0,018	0,014	0,008	−0,002
	(1,33)	(1,05)	(0,60)	(0,17)
Ti	0,015	0,006	−0,001	0,018
	(1,15)	(0,48)	(0,08)	(1,37)
Ci	0,034	0,022	0,045	0,021
	(2,57)	(1,68)	(3,38)	(1,57)

На рисунке 20а представлены зависимости для метода "заметания", соответствующие наиболее сильным корреляциям Te–Ci. Зависимости для $k(F^{(e)}_{\mathrm{TT}}|F^{(i)}_{\mathrm{CC}})\equiv k(F;\ \mathrm{Te-Ci})$ опять обнаруживают структурную связь через периодичность $p=3$ (см. скачки при $n\approx 8870$). Зависимость для $k(K;\ \mathrm{Te-Ci})$ при "заметании" слева направо обнаруживает быстрое спадание корреляций при $m_0\geqslant 10^3$ и относительно слабые остаточные дальние корреляции, которые четко видны на штриховой кривой, отвечающей расширению спектра для корреляционных функций со стороны больших m_0 (рис. 20б). Случайные перетасовки экзонов и интронов с сохранением их чередования разрушают остаточные дальние корреляции при $m_0\geqslant 10^3$ и приводят к уменьшению интегрального коэффициента корреляции примерно в

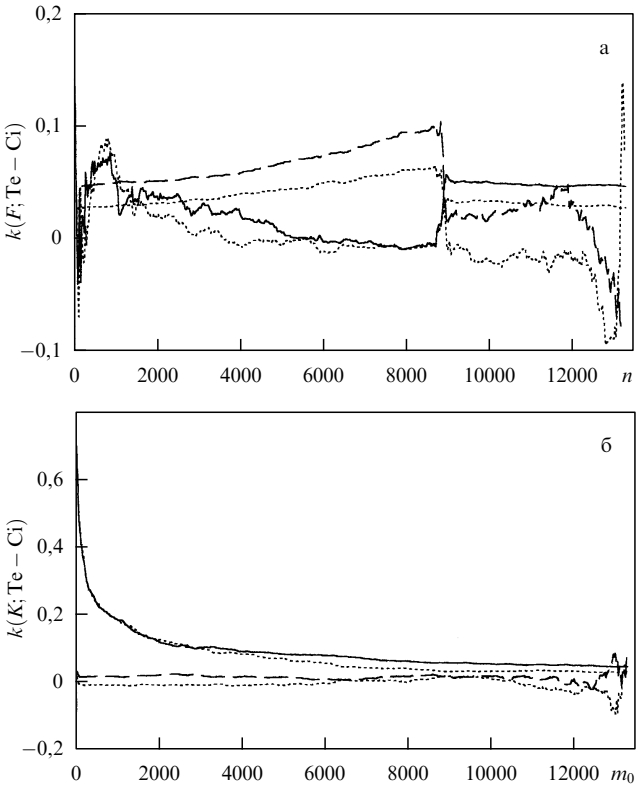

Рис. 20. Текущие коэффициенты корреляции при расширении спектра слева направо (сплошные кривые) и справа налево (штриховые кривые) для Te–Ci корреляций в коллагеновом гене. Соответствующие зависимости после случайной перетасовки экзонов и интронов при сохранении чередования изображены пунктиром.

полтора-два раза (пунктир на рис. 20). К аналогичным эффектам приводят также и случайные перетасовки интронов и экзонов в отдельности. Зависимости на рис. 20 свидетельствуют о скрытой периодичности $p = 3$ в некоторых интронах (явная проверка действительно обнаруживает ее в нескольких интронах). Такие интронные фрагменты могли бы динамически прилипать к мРНК и влиять на интенсивность синтеза белка [126].

Анализ корреляций в различных экзон-интронных последовательностях показывает, что механизмы структурной связи могут быть весьма разнообразными [27]. Кроме связи через периодичность $p = 3$, в корреляциях могут участвовать $p = 2$ и другие периодичности, причем сдвиги периодичностей в экзонах и интронах могут привести также и к антикорреляциям (в этом случае скачки, типа изображенных на рис. 20, оказываются отрицательными). В ряде случаев на фоне медленного спадания корреляций на масштабах $m_0 \sim 10^3$ наблюдаются также характерные корреляции на масштабах $m_0 \sim 10^2$, что свидетельствует о более сильном сходстве структурных характеристик для отдельных более коротких сегментов.

Точное вырезание интронов при сплайсинге обеспечивается специфическим узнаванием экзон-интронных границ [29, 31]. Так, для границы между $3'$ концом экзона и $5'$ концом интрона характерно наличие последовательности ${}^{\mathrm{A}}_{\mathrm{C}}\mathrm{AG|GT}^{\mathrm{A}}_{\mathrm{G}}\mathrm{AGT}$ в то время как для $3'$ конца интрона и $5'$ конца экзона специфична последовательность $(\mathrm{Y})_{11}\mathrm{NYAG|N}$, где Y — пиримидин, N — любой нуклеотид, а вертикальная разделительная черта соответствует границе. Характерные длины сохраняющихся фрагментов составляют $\sim 10-30$ нуклеотидов. Мы видим, однако, что структурная связь между экзонами и интронами простирается значительно дальше этой области. Такая связь может быть обусловлена упаковкой ДНК в нуклеосомы, механизмами сплайсинга, регуляции синтеза белка и др. [26, 27, 37, 120, 126].

Как видно из примеров, приведенных в разделах 5.1 и 5.2, совместное использование спектров для структурных факторов и корреляционных функций позволяет выделить набор характерных длин и установить механизмы корреляций. При этом спектральные методы оказываются практически единственными из всех известных, позволяющими выделить и количественно оценить значимость относительно слабых эффектов структурной связи между различными элементами мозаичной структуры генома.

6. Заключение

Взаимная однозначность (см. (3.2) и (3.3)) и взаимная дополнительность масштабов, $q_n \propto m^{-1}$, m_0^{-1} (см. (3.3) и (3.6)) при преобразовании Фурье обеспечивают целостное представление структурных характеристик последовательности ДНК. В отличие, например, от теории марковских цепей (раздел 2.2), спектральный подход не предполагает какого-либо априорного представления о характере упорядоченности. Напротив, сам характер упорядоченности и взаимных корреляций выявляется в процессе спектрального анализа, основанного на совместном использовании структурных факторов (3.5) и корреляционных функций (3.6). Интегральную степень регулярности последовательности ДНК можно оценить с помощью структурной спектральной энтропии (3.44), (3.47). Спектральный анализ выявляет широкий набор характерных длин, часть из которых связана с эффектами сегментации. При необходимости эффекты сегментации можно учесть с самого начала (раздел 5.2).

Локальные вариации структурных характеристик можно изучать с помощью стандартной техники с окнами, перемещаемыми вдоль исходной последовательности [5, 7, 8]. Используя методы теории фильтрации [91] и выделяя с помощью фильтров различные характерные участки в спектрах Фурье, можно решить обратную задачу о нахождении структурных особенностей, связанных с сингулярными гармониками (ср. [19]). Такой подход можно использовать для грубого выделения эффектов сегментации, в то время как точные границы сегментов можно определить с помощью каталога специфических сигнальных последовательностей.

В отличие от чисто статистического подхода, основанного на подсчете встречаемости различных комбинаций нуклеотидов [5–9], спектральные методы позволяют выявить ряд устойчивых структурных признаков в сегментах ДНК с различными функциональными свойствами. Так, в [45] был проведен сравнительный анализ эффективности более чем двадцати методов идентификации белок-кодирующих областей. Для окон шириной $W = 120$ наиболее эффективным оказался метод Фурье. За ним следует метод, основанный на подсчете комбинаций гексануклеотидов. Во втором случае мы сталкиваемся с необходимостью анализа $4^6 = 4096$ величин, в то время как метод Фурье требует выделения лишь нескольких гармоник в окрестности периодичности $p = 3$. Существенно также, что выделение устойчивых структурных признаков позволяет понять общие принципы формирования геномных последовательностей ДНК и дает основу для построения различных эволюционных теорий [4, 41, 96, 97]. В этом смысле ДНК является уникальным объектом: с одной стороны, она содержит информацию о вполне современных белках, представляющих интерес для медицинских и биохимических приложений [29] или даже для создания нанотехнологий нового типа [127]; с другой стороны, в последовательностях ДНК отражена история молекулярной эволюции приблизительно за ~ 4 млрд лет [4, 6, 9, 29, 41, 96, 97].

При анализе последовательностей по всей базе данных не следует забывать, что пока еще не закончен этап накопления начальных данных. Лишь в ближайшем будущем появятся полные расшифрованные последовательности для геномов высших организмов. Набор последовательностей в современном банке данных во многом определяется медицинскими, биохимическими и другими приложениями. Он является достаточно специфическим и далек от требований статистической однородности. Поэтому для сравнительного видового и эволюционного анализа необходим строгий предварительный отбор аналоговых данных [4, 6, 8, 9, 29, 128]. Не следует также забывать, что ДНК организована в сложную иерархическую систему, каждый уровень которой связан со своими генетическими и биологическими механизмами (и в этом смысле ДНК существенно отличается от однородных фракталов [125]). Чисто статистическая оценка сложности последовательности ДНК без предварительной реконструкции иерархии может оказаться подобной оценке сложности электронной схемы по длине соединительных проводов.

В рамках спектрального подхода удается добиться формальной унификации цифрового представления данных и сходства в статистиках для случайных аналогов для пространственных характеристик C_α-остовов белков, распределения физико-химических параметров вдоль полипептидной цепи и структурных характеристик соответствующей белок-кодирующей последовательности ДНК [24, 129], что дает практически уникальную возможность изучения прямых корреляций между этими величинами. Общая математическая теория структурного анализа пространственных линейных цепей в рамках спектрального подхода изложена в [130]. В [129] изучалась связь между пространственной структурой C_α-остовов белков по данным рентгеноструктурного анализа и распределением физико-химических параметров вдоль цепи, а в [24] была качественно показана зависимость $\Delta S_{\rm rel}$ для белок-кодирующей ДНК от степени пространственной регулярности структуры белка. Детальное изучение связи между всеми этими характеристиками представляло бы фундаментальный интерес и позволило бы понять, как пространственная структура белка отражается на уровне белок-кодирующей последовательности ДНК.

Результаты структурного анализа свидетельствуют об определенной структурной целостности геномных последовательностей ДНК. Несмотря на огромный накопленный материал, общие принципы формирования таких последовательностей остаются во многом не вполне ясными. Осмысление языка, на котором написаны генетические тексты, далеко еще не закончено. Систематизация и отбор устойчивых признаков выдвигают на следующем этапе исследований вопрос о реализации физических механизмов, ответственных за наблюдаемые структурные особенности.

Авторы выражают благодарность А.А. Веденову за поддержку программы исследований. Мы признательны А.М. Камчатнову, Е.Б. Левченко, М.А. Лифшицу, В.Ю. Макееву и А.Л. Чернякову за обсуждение различных аспектов проблемы, а также А.Ю. Турыгину за помощь при проведении численных расчетов и А.А. Ежову, совместно с которым разрабатывались методы реконструкции скрытых повторов. Мы признательны также нашим многочисленным коллегам за предоставление результатов своих работ часто до опубликования в печати.

Список литературы

1. Rodriguez-Tome P et al. *Nucl. Acids Res.* **24** 6 (1996)
2. Benson D A et al. *Nucl. Acids Res.* **26** 1 (1998)
3. Stoesser G et al. *Nucl. Acids Res.* **27** 18 (1999)
4. Ратнер В А и др. *Проблемы теории молекулярной эволюции* (Новосибирск: Наука, 1985)
5. Александров А А и др. *Компьютерный анализ генетических текстов* (М.: Наука, 1990)
6. Вейр Б *Анализ генетических данных* (М.: Мир, 1995)
7. *Математические методы для анализа последовательностей ДНК* (Под ред. М С Уотермана) (М.: Мир, 1999)
8. *DNA and Protein Sequence Analysis: A Practical Approach* (Eds M J Bishop, C J Rawlings) (Oxford: IRL Press, 1997)
9. Kolchanov N A, Lim H A *Computer Analysis of Genetic Macromolecules: Structure, Function and Evolution* (Singapore: World Scientific, 1994)
10. *Internet for the Molecular Biologist* (Eds S R Swindell, R R Miller, G S A Myers) (Oxford: Horizon Scientific, 1996)
11. Baxevanis A, Quellette B F F *Bioinformatics: A Practical Guide to the Analyses of Genes and Proteins* (New York: Wiley, 1998)
12. Vingron M, Waterman M S *J. Mol. Biol.* **235** 1 (1994)
13. Gelfand M S *J. Comp. Biol.* **2** 87 (1995); Гельфанд М С *Мол. биол.* **32** 103 (1998)
14. Fickett J W *Comput. Chem.* **20** 103 (1996)
15. Li W *Comput. Chem.* **21** 257 (1997)
16. http://linkage.rockefeller.edu/wli/dna_corr
17. Турыгин А Ю, Чечеткин В Р *ЖЭТФ* **106** 335 (1994)
18. Chechetkin V R, Turygin A Y *J. Phys. A* **27** 4875 (1994)
19. Chechetkin V R, Knizhnikova L A, Turygin A Y *J. Biomol. Struct. Dyn.* **12** 271 (1994)
20. Chechetkin V R, Turygin A Y *Phys. Lett. A* **199** 75 (1995)
21. Chechetkin V R, Turygin A Y *J. Theor. Biol.* **175** 477 (1995)
22. Chechetkin V R et al. "Reconstruction of hidden repeats in DNA sequences, their significance, and applications", Препринт ТРИНИТИ 0033-А (М.: ЦНИИАТОМИНФОРМ, 1996) (на англ. яз.) с. 42
23. Chechetkin V R, Turygin A Y *J. Theor. Biol.* **178** 205 (1996)
24. Chechetkin V R, Lobzin V V *Phys. Lett. A* **222** 354 (1996)
25. Ежов А А, Чечеткин В Р *Мат. моделирование* **10** 83 (1998)
26. Chechetkin V R, Lobzin V V *J. Biomol. Struct. Dyn.* **15** 937 (1998)
27. Chechetkin V R, Lobzin V V *J. Theor. Biol.* **190** 69 (1998)
28. Иваницкий Г Р и др. *Биофиз.* **30** 418 (1985)
29. Албертс Б и др. *Молекулярная биология клетки* (М.: Мир, 1986)
30. Рис Э, Стернберг М *От клетки к атомам. Иллюстрированное введение в молекулярную биологию* (М.: Мир, 1988)
31. Георгиев Г П *Гены высших организмов и их экспрессия* (М.: Наука, 1989)
32. Гросберг А Ю, Хохлов А Р *Статистическая физика макромолекул* (М.: Наука, 1989)
33. Creighton T *Proteins: Structures and Molecular Properties* (New York: Freeman, 1993)
34. Степанов В М *Структура и функции белков* (М.: Высшая школа, 1996)
35. Britten R J, Kohne E D *Science* **161** 529 (1968)
36. Trifonov E N *Bull. Math. Biol.* **51** 417 (1989)
37. Trifonov E N *Comput. Chem.* **17** 27 (1993)
38. Trifonov E N, Sussman J L *Proc. Natl. Acad. Sci. USA* **77** 3816 (1980)
39. Hagerman P J *Annu. Rev. Biochem.* **59** 755 (1990)
40. Olson W K, Zhurkin V B, in *Biological Structure and Dynamics* (Eds R H Sarma, M H Sarma) (New York: Adenine Press, 1996) p. 341
41. Nussinov R *J. Theor. Biol.* **125** 219 (1987)
42. Staden R *Comput. Applic. Biosci.* **4** 53 (1988)
43. Сприжицкий Ю А и др. *Мол. биол.* **22** 338 (1988)
44. Wada K et al. *Nucl. Acids Res.* **19** 1981 (1991)
45. Fickett J W, Tung C-S *Nucl. Acids Res.* **20** 6441 (1992)
46. Шеннон К *Работы по теории информации и кибернетике* (М.: ИИЛ, 1963)
47. Bleisdell B E *J. Mol. Evol.* **21** 278 (1985)
48. Бородовский М Ю и др. *Мол. биол.* **20** 1014, 1024, 1390 (1986)
49. Phillips G J, Arnold J, Ivarie R *Nucl. Acids Res.* **15** 2611, 2627 (1987)
50. Stückle E E et al. *Nucl. Acids Res.* **18** 6641 (1990)
51. Феллер В *Введение в теорию вероятностей и ее приложения* (М.: Мир, 1984)
52. Herzel H, Ebeling W, Schmitt A O *Phys. Rev. E* **50** 5061 (1994); *Chaos, Solitons and Fractals* **4** 97 (1994)
53. Mantegna R N et al. *Phys. Rev. Lett.* **73** 316 (1994)
54. Israeloff N E, Kagalenko M, Chan K *Phys. Rev. Lett.* **76** 1976 (1996)
55. Schmitt A O, Ebeling W, Herzel H *Biosystems* **37** 199 (1996)
56. Chatzidimitriou-Dreismann C A, Streffer R M F, Larhammar D *Nucl. Acids Res.* **24** 1676 (1996)
57. Luo L F *Collected Works on Theoretical Biophysics* (Hohhot: Inner Mongolia University Press, 1997)
58. Herzel H, Grosse I *Physica A* **216** 519 (1995)
59. Mrazek J, Kypr J *Comp. Applic. Biosci.* **11** 195 (1995)
60. Li W, Kaneko K *Europhys. Lett.* **17** 655 (1992)
61. Luo L F et al. *Phys. Rev. E* **58** 861 (1998)
62. Peng C-K et al. *Nature* (London) **356** 168 (1992)
63. Nee S *Nature* (London) **357** 450 (1992)
64. Karlin S, Brendel V *Science* **259** 677 (1993)

65. Chatzidimitriou-Dreismann C A, Streffer R M F, Larhammar D *Eur. J. Biochem.* **224** 365 (1994); *Biochim. Biophys. Acta* **1217** 181 (1994)
66. Капитонов В В, Титов И И *Доклады РАН* **337** 810 (1994)
67. Borovik A S, Grosberg A Y, Frank-Kamenetskii D F *J. Biomol. Struct. Dyn.* **12** 655 (1994)
68. Shnerb N, Eisenberg E *Phys. Rev. E* **49** R1005 (1994)
69. Stanley H E et al. *Nuovo Cimento D* **16** 1339 (1996)
70. Allegrini P et al. *Phys. Rev. E* **57** 4558; **58** 3640 (1998)
71. Viswanathan G M et al. *Physica A* **249** 581 (1998)
72. Астафьева Н М *УФН* **166** 1145 (1996)
73. Arneodo A et al. *Phys. Rev. Lett.* **74** 3293 (1995)
74. Altaiskii M, Mornev O, Polozov R *Genetic Analysis: Biomol. Engineer.* **12** 165 (1996)
75. Tsonis A A et al. *Phys. Rev. E* **53** 1828 (1996)
76. Arneodo A et al. *Physica A* **249** 439 (1998)
77. Bernardi G *Annu. Rev. Genet.* **29** 445 (1995)
78. Fickett J W, Torney D C, Wolf D R *Genomics* **13** 1056 (1992)
79. McLachlan A D, Stewart M *J. Mol. Biol.* **103** 271 (1976)
80. Silverman B D, Linsker R *J. Theor. Biol.* **118** 295 (1986)
81. Деев А А и др. *Биофиз.* **34** 564 (1989)
82. McLachlan A D *J. Phys. Chem.* **97** 300 (1993)
83. Makeev V J, Tumanyan V G *Comp. Applic. Biosci.* **12** 49 (1995)
84. Lee W J, Luo L F *Phys. Rev. E* **56** 848 (1997)
85. Кутузова Г И и др. *Биофиз.* **42** 354 (1997)
86. Felsenstein J, Sawyer S, Kochin R *Nucl. Acids Res.* **10** 133 (1982)
87. Benson D C *Nucl. Acids Res.* **18** 3001, 6305 (1990)
88. Cheever E A, Overton G C, Searls D B *Comp. Applic. Biosci.* **7** 143 (1991)
89. Arques D G, Michel C J, Oriex K *Comp. Applic. Biosci.* **8** 5 (1992)
90. Voss R F *Phys. Rev. Lett.* **68** 3805 (1992); *Fractals* **2** 1 (1994)
91. Марпл С Л *Цифровой спектральный анализ и его приложения* (М.: Мир, 1990)
92. Ван Кампен Н Г *Стохастические процессы в физике и химии* (М.: Высшая школа, 1990)
93. Андерсон Т *Введение в многомерный статистический анализ* (М.: Физматгиз, 1963)
94. Low R L, Arai K, Kornberg A *Proc. Natl. Acad. Sci. USA* **78** 1436 (1981)
95. Грин Н, Стаут У, Тейлор Д *Биология* (М.: Мир, 1993)
96. Crick F H C et al. *Origins Life* **7** 389 (1976)
97. Eigen M et al. *Sci. Am.* **244** 88 (1981)
98. Zhurkin V B *Nucl. Acids Res.* **9** 1963 (1981)
99. Kypr J, Mrazek J *Int. J. Biol. Macromol.* **9** 49 (1987)
100. Leadbetter M R, Lindgren G, Rootzen H *Extremes and Related Properties of Random Sequences and Processes* (New York: Springer, 1983)
101. Mani G A *J. Theor. Biol.* **158** 447 (1992)
102. Eigen M, Winkler-Oswatitsch R *Naturwissenschaften* **68** 282 (1981)
103. Lagunez-Otero J, Trifonov E N *J. Biomol. Struct. Dyn.* **10** 451 (1992)
104. Curran J F, Gross B L *J. Mol. Biol.* **235** 389 (1994)
105. Almagor H *J. Theor. Biol.* **117** 127 (1985)
106. Zhang C-T, Chou K-C *J. Mol. Biol.* **238** 1 (1994)
107. Mrazek J, Kypr J *J. Mol. Evol.* **39** 439 (1994)
108. Lio P, Ruffo S, Buiatti M *J. Theor. Biol.* **171** 215 (1994)
109. Frank G K, Makeev V J *J. Biomol. Struct. Dyn.* **14** 629 (1997)
110. Johnson R L et al. *Science* **272** 1668 (1996)
111. Gailani M R et al. *Nature Genetics* **14** 78 (1996)
112. Ежов А А и др. *Росс. журнал кожных болезней* (1) 17 (1999)
113. Godreche C, Luck J M *J. Phys. A* **23** 3769 (1990)
114. Cheng Z, Savit R *Phys. Rev. A* **44** 6379 (1991)
115. Li W *Phys. Rev. A* **43** 5240 (1991)
116. Wang Y-H et al. *Science* **265** 669 (1994)
117. Bates G, Lehrach H *Bioessays* **16** 277 (1994)
118. *Analysis of Triplet Repeat Disorders* (Eds D Rubinsztein, M Hayden) (Oxford: BIOS, 1998)
119. Bina M *J. Mol. Biol.* **235** 198 (1994)
120. Beckmann J S, Trifonov E N *Proc. Natl. Acad. Sci. USA* **88** 2380 (1991)
121. Arques D G, Michel C J *Nucl. Acids Res.* **15** 7581 (1987)
122. Vogt P *Human Genet.* **84** 301 (1990)
123. Arques D G, Michel C J *J. Theor. Biol.* **143** 307 (1990)
124. Хамермеш М *Теория групп и ее применение к физическим проблемам* (М.: Мир, 1966)
125. Федер Е *Фракталы* (М.: Мир, 1991)
126. Novak R *Science* **263** 608 (1994)
127. Drexler K E *Annu. Rev. Biophys. Biomol. Struct.* **23** 377 (1994)
128. Айала Ф *Введение в популяционную и эволюционную генетику* (М.: Мир, 1984)
129. Chechetkin V R, Lobzin V V *J. Theor. Biol.* **198** 197, 219 (1999)
130. Лобзин В В, Чечеткин В Р *ЖЭТФ* **116** 620 (1999)

Order and correlations in genomic DNA sequences. The spectral approach

V.V. Lobzin
Institute of Terrestrial Magnetism, Ionosphere, and Radiowave Propagation, Russian Academy of Sciences
142092 Troitsk, Moscow Region, Russian Federation
Tel. (7-095) 334-01 13. Fax (7-095) 334-01 24
E-mail: lobzin@top.izmiran.troitsk.ru
V.R. Chechetkin
Troitsk Institute for Innovation and Thermonuclear Investigations
142092 Troitsk, Moscow Region, Russian Federation
Tel. (7-095) 334-50 57

The structural analysis of genomic DNA sequences is discussed in the framework of the spectral approach, which is sufficiently universal due to the reciprocal correspondence and mutual complementarity of Fourier transform length scales. The spectral characteristics of random sequences of the same nucleotide composition possess the property of self-averaging for relatively short sequences of length $M \geqslant 100-300$. The comparison with the characteristics of random sequences determines the statistical significance of the structural features observed. Apart from traditional applications to the search of hidden periodicities, spectral methods are also efficient in studying mutual correlations in DNA sequences. By combining spectra for structure factors and correlation functions, not only integral correlations can be estimated but also their origin identified. Using the structural spectral entropy approach, the regularity of a sequence can be quantitatively assessed. A brief introduction to the problem is also presented and other major methods of DNA sequence analysis described.

PACS numbers: **02.50.–r**, **05.10.–a**, **87.10.+e**, 87.14.Gg

Bibliography — 130 references

Received 1 June, revised 30 July 1999