УДК 576.315.42

Г. М. Субоч, Ю. А. Сприжицкий, А. А. Александров

ВСТРЕЧАЕМОСТЬ ГОМОПУРИН-ГОМОПИРИМИДИНОВЫХ ЗЕРКАЛЬНЫХ ПОВТОРОВ В ПРИРОДНЫХ ДНК

Проведен детальный анализ встречаемости потенциальных сайтов образования H-*формы* (H-*палиндромов*) *и других локальных гомопурин-гомопиримидиновых зеркальных повторов в различных функциональных областях природных ДНК. Получены распределения встречаемости таких сочетаний нуклеотидов относительно точек инициации и терминации трансляции.*

Введение. Задача поиска возможной биологической роли неканонических структур, таких как кресты, Z-форма или H-форма [1, 3], является весьма актуальной. Одним из подходов к решению этой задачи может быть статистический анализ встречаемости в природных ДНК такого рода структур и определение районов с повышенной вероятностью их локализации относительно различных функциональных зон ДНК.

Недавно в гомопурин-гомопиримидиновых участках сверхспиральной ДНК был обнаружен pH-зависимый конформационный переход [2]. Наиболее вероятная структура, формирующаяся при этом, содержит трехнитчатую шпильку, образованную двумя гомопиримидиновыми и одной гомопуриновой нитью [2, 3]. Такая структура, названная H-формой, может образовываться лишь в области зеркального гомопурин-гомопиримидинового повтора типа 5′-..AGAGG....GGAGA..-3′ или 5′-..TCTCC....CCTCT..-3′ по комплементарной нити. Назовем участок последовательности в промежутке между повторами петлей, а сочетание нуклеотидов, соответствующее самому повтору,— стеблем. Следует подчеркнуть, что энергетические требования образования H-формы накладывают определенные ограничения на нуклеотидный состав и на длины петли и стебля. Будем исходить из следующего критерия причисления зеркального гомопурин-гомопиримидинового повтора к потенциальному сайту образования H-формы: GC-содержание повтора должно быть не меньше 75 %, а GC-содержание петли — меньше 50 %, причем длина стебля должна превышать три пары оснований, а длина петли должна лежать в пределах от 3 до 12 нуклеотидов. Структуры такого типа будем называть H-палиндромами, естественно, понимая при этом, что без участия белков такой участок перейти в H-форму не может.

В данной работе проведен детальный статистический анализ встречаемости H-палиндромов в различных функциональных областях природных ДНК. Используя модельные последовательности, методом Монте-Карло сделаны оценки статистической значимости результатов. Значения частот встречаемости H-палиндромов соотносятся с таковыми всех гомопурин-гомопиримидиновых зеркальных повторов с длиной стебля не меньше 4 нуклеотидов и длиной петли от 3 до 12 нуклеотидов. Получены распределения встречаемости таких сочетаний нуклеотидов относительно точек инициации и терминации трансляции.

Программы и методы. Для анализа встречаемости исследуемых структур нами был написан комплекс программ, позволяющих осуществить поиск локальных инвертированных и зеркальных несовершенных повторов, рассчитать их частоты встречаемости в различных областях генома, построить распределения локализации различных сайтов относительно точек инициации и терминации трансляции. Перекрывающиеся повторы, а также повторы, образованные многократно тандемно повторенными моно- или динуклеотидами, при расчете не учитывались. Все программы написаны на языке TURBO PASCAL и реализованы на персональном компьютере IBM PC.

Для решения вопроса, чем определяются наблюдаемые частоты встречаемости зеркальных повторов, мы генерировали и исследовали случайные последовательности трех видов: с заданным нуклеотидным составом, с заданными частотами тринуклеотидов и с заданным распределением блоков нуклеотидов различного типа по длинам. Выбор третьей модели обусловлен обнаруженными ранее [6] значительными отклонениями частот встречаемости блоков разной длины в природных последовательностях ДНК от ожидаемых для случайных последовательностей того же нуклеотидного состава. Преимущества этой модели продемонстрированы в [7]. Мы разработали алгоритм генерации случайной последовательности, для которой математическое ожидание числа блоков различной длины для нуклеотидов шести типов (пурины, пиримидины, A, T, G, C) равнялось бы заранее заданным величинам. Эти величины выбирались в соответствии с реально наблюдаемыми частотами блоков в кодирующих и некодирующих последовательностях ДНК из геномов различных организмов. Для каждого такого специфического набора параметров сблоченности генерировали случайную последовательность. Ранее было показано, что частоты блоков являются консервативными в пределах довольно больших таксономических групп [6], и отмечалась близость этих характеристик для регуляторных участков и интронов. Поэтому для оценки ожидаемого числа встречаемости повторов в интронах и в 5′, 3′-нетранслируемых участках последовательностей ДНК из организмов, принадлежащих одной таксономической группе, использовали одни и те же модельные последовательности. Алгоритм генерации описан в [7].

Мы генерировали для каждого типа областей модельные последовательности длиной по 100000 нуклеотидов. Для оценки разброса частот повторов, обусловленного статистически случайным характером генерации последовательности, применяли процедуру типа бутстреп [8, 9]. Методика определения средних значений частот (т. е. ожидаемых частот) и доверительного интервала отклонения от средних приводится в [7].

Выборки последовательностей получены из базы данных нуклеотидных последовательностей GenBank, содержащей около 8,5 млн нуклеотидов [10]. Длины выборок функциональных областей представлены в таблице. Следует подчеркнуть, что мы ис-

Размеры выборок функциональных участков ДНК из различных организмов
The sizes of the sets of DNA functional regions from the different species

Организм	Белок-кодирующие области	Интроны	5′- и 3′-нетранслируемые участки
Человек (*а*)	276,942	174,761	337,233
Грызуны (*б*)	321,807	150,253	291,800
Позвоночные без человека и грызунов (*в*)	118,825	62,150	113,290
Беспозвоночные (*г*)	117,855	15,035	117,034
Дрожжи (*д*)	131,826	—	127,034
Прокариоты (*е*)	549,988	—	292,021

Примечание. Буквы в скобках соответствуют аналогично обозначенным выборкам, приведенным на рис. 2.

ключили из рассмотрения последовательности, описания которых не содержат функциональных участков. Мы также не приводим результатов расчетов для областей, кодирующих тРНК, рРНК и мяРНК, из-за малой длины этих выборок, следствием чего является низкая статистическая значимость результатов.

Результаты и обсуждение. Мы рассчитали частоты встречаемости H-палиндромов в выборках последовательностей ДНК, указанных в таблице. Почти для всех выборок наблюдаемые частоты существенно отличаются от частоты, полученной на модельной последовательности с равновероятным случайным распределением нуклеотидов. Очевидно, что столь простая модель не может нас удовлетворить при оценке ожидаемых частот таких специфических структур, как гомопурин-гомопиримидиновые зеркальные повторы, поскольку, как уже отмечалось, сблоченность нуклеотидов может оказывать существенное влияние на эти

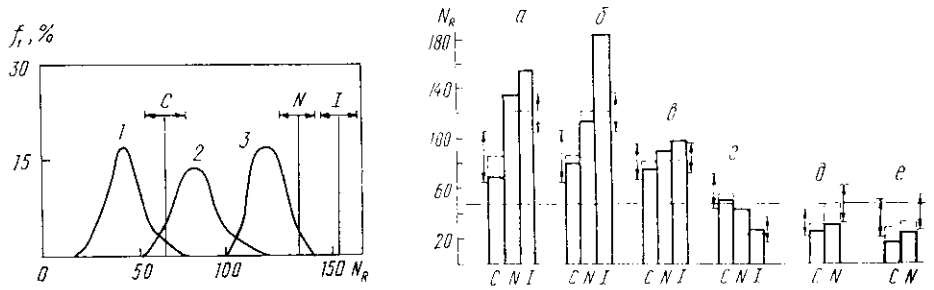

Рис. 1. Плотности распределения частот встречаемости H-палиндромов в модельных последовательностях с равновероятным случайным распределением оснований (*1*) и в последовательностях, сгенерированных с учетом параметров сблоченности нуклеотидов в различных функциональных областях ДНК млекопитающих: в белок-кодирующих областях (*2*); в некодирующих участках (*3*). По оси абсцисс отложено число H-палиндромов в последовательности длиной 100000 нуклеотидов; по оси ординат — относительная доля последовательностей, содержащих данное число H-палиндромов (f_1). Распределения получены путем применения процедуры бутстрепа к первоначально сгенерированным случайным последовательностям [7] и стандартной процедуры сглаживания экспериментальных данных с окном семь точек. Вертикальные линии отмечают на оси абсцисс средние значения частот H-палиндромов в интронах (*I*), кодирующих (*C*) и 5′, 3′-нетранслируемых (*N*) последовательностях ДНК человека. Горизонтальные отрезки демонстрируют разброс средних значений в подвыборках, полученных случайной разбивкой основной выборки на две части

Fig. 1. Distribution densities of the H-palindromes occurrence frequencies in model sequences with random distribution of nucleotides (*1*) and in sequences, generated by means of using frequencies of nucleotide runs in different functional regions of mammalian DNAs: in protein-coding areas (*2*), in noncoding regions (*3*). Abscissa: occurrence numbers of H-palindromes N; ordinate: the percentage counts of sequences, which contain the given number of H-palindromes f_1. The distributions are derived by means of application of bootstrap method to the model sequence and using standard smoothing procedure to the experimental points. The vertical lines correspond to the mean values of frequencies of H-palindromes in introns (*I*), coding areas (*C*) and in regulatory regions (*N*) of human DNA

Рис. 2. Частоты встречаемости H-палиндромов на 100000 пар нуклеотидов N_R в выборках последовательностей различных групп организмов (характеристики выборок *а* — *е* приведены в таблице): *C* — кодирующие области; *N* — 5′- и 3′-нетранслируемые участки; *I* — интроны. Штрих-пунктирной линией показан уровень, соответствующий ожидаемой частоте H-палиндромов в последовательности со случайным равновероятным распределением нуклеотидов. Отрезками штриховой линии отмечены средние частоты H-палиндромов в модельных последовательностях, сгенерированных с учетом параметров сблоченности, характеризующих данную выборку природных последовательностей. Стрелками обозначены доверительные интервалы для этих средних с 5 %-ным уровнем значимости, рассчитанные из распределений, полученных методом бутстрепа (см. рис. 1)

Fig. 2. The frequencies of occurrence of H-palindromes N_R in natural sequences from various groups of organisms (*a-e* characteristics, see tab. 1): *C* — coding areas; *N* — 5′- and 3′-noncoding regions; *I* — introns. The dashed-dots line corresponds to the level of expected frequency of H-palindromes in sequence with a random distribution of bases. The dashed line represents the mean frequencies of H-palindromes in model sequences generated using the nucleotide runs frequencies. The arrows limit a confidence intervals for these mean values with 5 % significance level

оценки [7]. На рис. 1 представлены распределения бутстрепа для модельных последовательностей и средние частоты встречаемости H-палиндромов для выборок ДНК человека. Видно, что учет сблоченности нуклеотидов приводит к значительному смещению ожидаемых значе-

ний, и что статистически достоверное превышение встречаемости Н-палиндромов в природных последовательностях над встречаемостью таких структур в модельных последовательностях наблюдается лишь для интронов, причем это превышение незначительно.

На рис. 2 приведены данные о частотах встречаемости Н-палиндромов в различных функциональных областях ДНК из геномов разных организмов. В каждом случае указывается ожидаемое число Н-палиндромов и доверительный интервал с 5 %-ным уровнем значимости двустороннего критерия принятия нулевой гипотезы (о равенстве наблюдаемых и ожидаемых значений этих частот). Приведенные результаты показывают, что наиболее значительное превышение числа Н-палиндромов над ожидаемым значением наблюдается для выборок ин-

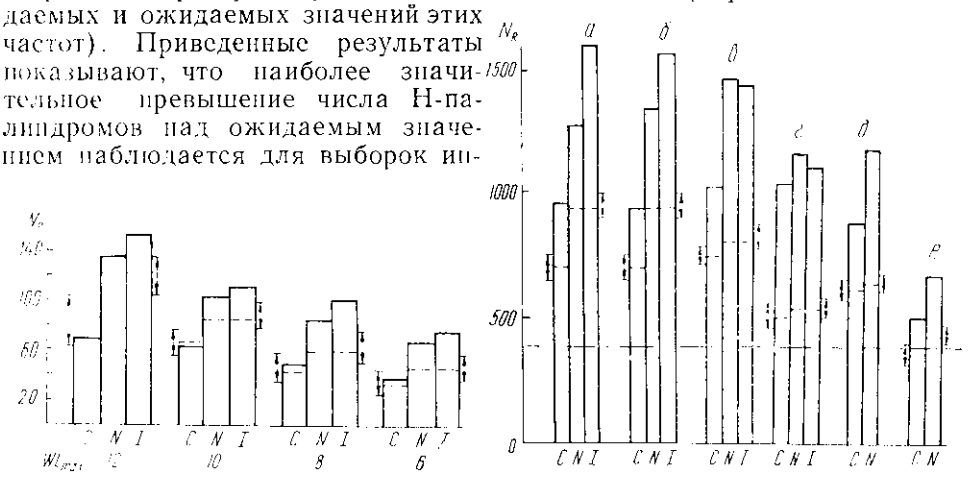

Рис. 3. Частоты встречаемости Н-палиндромов в последовательностях ДНК человека при различных значениях максимальной длины петли (см. также подпись к рис. 2)

Fig. 3. The frequencies of occurrence of H-palindromes in the human DNA sequences versus maximal length wl_{max} of gap (see legend to Fig. 2)

Рис. 4. Частоты встречаемости пуриновых повторов в различных функциональных областях разных организмов (см. подпись к рис. 2)

Fig. 4. Frequencies of occurrence of purine repeats in different functional regions of DNA of various organisms (see legend to Fig. 2)

тронов грызунов и человека. В остальных случаях различия результатов для выборок последовательностей разных групп могут объясняться специфическим набором параметров сблоченности. Однако существует устойчивая тенденция превышения встречаемости Н-палиндромов в интронах над 5′, 3′-нетранслируемыми областями (для последовательностей позвоночных) и в некодирующих областях над белок-кодирующими.

На рис. 3 представлены частоты встречаемости Н-палиндромов в ДНК человека в зависимости от максимальной длины петли wl. Видно, что результаты, полученные для $wl_{max}=12$, качественно не меняются при уменьшении этой величины. К такому же выводу мы пришли, проанализировав аналогичным образом другие выборки последовательностей.

Анализ последовательностей, составляющих Н-палиндромы, показал, что наиболее распространенными являются повторы, имеющие структуру типа 5′-..CCCC — CCCC..-3′. Они составляют 34 % общего числа Н-палиндромов. Другие типы Н-палиндромов встречаются в следующих пропорциях: ..CTCC — 22, ..CCTC — 18, ..TCCC — 17 и ..CCCT — 11 %.

Как отмечалось выше, Н-палиндромы — частный случай локальных зеркальных гомопурин-гомопиримидиновых повторов (для краткости будем называть последние просто пуриновыми повторами). Рис. 4 демонстрирует распределение встречаемости таких структур в различных выборках. В отличие от результатов, представленных на рис. 2 для Н-палиндромов, в данном случае можно отметить значительное превышение средних значений наблюдаемых частот над ожидаемыми для всех

без исключения выборок последовательностей, причем для млекопитающих встречаемость таких структур в интронах выше, чем в 5′- и 3′-нетранслируемых областях, а в тех в свою очередь выше, чем в белок-кодирующих. Для ДНК из организмов остальных таксономических групп также число повторов в некодирующих участках больше, чем в белок-кодирующих. Следует отметить, что наибольшая насыщенность пуриновыми повторами присуща ДНК млекопитающих, а наименьшая — ДНК прокариот.

Анализ процентного содержания Н-палиндромов среди всех пуриновых повторов показал, что эта величина в модельных последовательностях выше, чем в природных. Данное свойство может объяснять-

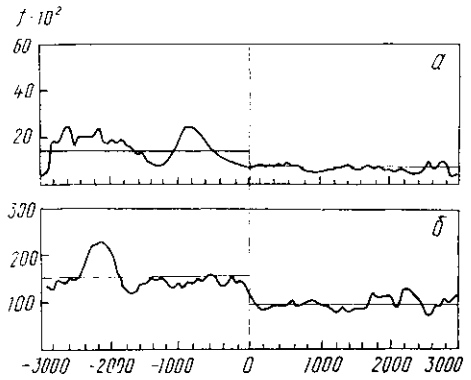

Рис. 5. Распределение нормированных частот встречаемости зеркальных повторов относительно точки инициации трансляции в последовательностях ДНК позвоночных: Н-палиндромов (*а*) и пуриновых повторов (*б*). По оси абсцисс отложено расстояние от инициирующего кодона в нуклеотидах; по оси ординат — нормированное число повторов в интервалах длиной 100 нуклеотидов, расположенных на данном расстоянии от инициирующего кодона — *f*. Нормировку производили на суммарную длину проанализированных для данного расстояния интервалов. Кривая получена посредством стандартной процедуры сглаживания экспериментальных данных с окном семь точек

Fig. 5. Distribution of the normalized frequencies of occurrence of mirror repeats f around translation initiation point in the sequences of vertebrate DNAs: H-palindromes (*a*); purine repeats (*б*). Abscissa: the distance from initiation codon; ordinate: the normalized number of repeats located in range of 100 bp. at a given distance from considered points

ся наличием в 5′- и 3′-нетранслируемых областях и интронах длинных АТ-богатых участков [6]. Однако только этим обстоятельством нельзя объяснить наблюдаемые отклонения, так как повышенная встречаемость пуриновых повторов наблюдается и в кодирующих областях. Биологическая роль зеркальных повторов пока неясна, однако, как отмечается в [11], высокая их встречаемость может быть выгодной из-за чрезвычайно низкой вероятности формирования вторичной структуры мРНК в области таких повторов.

Поскольку альтернативные структуры ДНК, в том числе и Н-форма, могут участвовать в регуляции генетических процессов, естественным является предположение о повышенной их встречаемости в конкретных местах, находящихся на определенном расстоянии от функциональных сайтов. Мы построили модель распределения встречаемости Н-палиндромов и всех гомопурин-гомопиримидиновых зеркальных повторов относительно точек инициации и терминации трансляции в последовательностях ДНК позвоночных. Известно, что исследуемая структура Н-формы может обусловливать гиперчувствительность к действию S1 нуклеазы [4, 5], которая часто наблюдается вблизи точек инициации транскрипции. Замена точек отсчета в данной работе вызвана недостаточной информацией в банке данных о точках инициации транскрипции и частично обоснована относительно небольшим разбросом расстояний между точками инициации транскрипции и трансляции в генах позвоночных.

На рис. 5, *а*, представлено распределение частот встречаемости вдоль ДНК центров Н-палиндромов в окне размером 100 нуклеотидов в зависимости от расстояния до инициирующего кодона. Сканирование последовательностей проводили с шагом 50 нуклеотидов. Видно, что распределение более гетерогенно в некодирующих областях, чем в бе-

лок-кодирующих. Эта же тенденция характерна для распределения относительно точки терминации трансляции. Нормированные частоты имеют в рассмотренных участках следующие средние значения: 15,9 ... ... [ATG 7,5 ... кодирующая область ... 7,9 TGA] ... 9,7, т. е. в 5′-нетранслируемых областях H-палиндромы встречаются чаще, чем в 3′-нетранслируемых. Рис. 5, *б*, демонстрирует аналогичные распределения для пуриновых повторов. Средние значения частот в данном случае распределены следующим образом: 148,2 ... [ATG 94,8 ... кодирующая область ... ... 95,4 TGA] ... 100,7. Интересно, что места расположения пиков в распределениях для H-палиндромов и пуриновых повторов не коррелируют друг с другом. Это свидетельствует о том, что распределение H-палиндромов не является простым следствием распределения пуриновых повторов, а указывает на возможность существования мест с повышенной вероятностью локализации H-палиндромов.

Авторы глубоко признательны М. Д. Франк-Каменецкому, С. М. Миркину, В. И. Лямичеву за предложение темы данной работы и множество полезных обсуждений в процессе ее выполнения.

СПИСОК ЛИТЕРАТУРЫ

1. *Лазуркин Ю. С.* ДНК: сверхспирализация и образование неканонических структур // Биополимеры и клетка.— 1986.—2, № 6.— С. 283—292.
2. *Лямичев В. И., Миркин С. М., Франк-Каменецкий М. Д.* pH-зависимый структурный переход в гомопурин-гомопиримидиновом блоке в сверхспиральной ДНК // Там же.— № 3.— С. 115—124.
3. *Lyamichev V. I., Mirkin S. M., Frank-Kamenetskii M. D.* Structures of homopurine-homopyrimidine tract in superhelical DNA // J. Biomol. Struct. and Dyn.— 1986.—3, N 4.— P. 667—699.
4. *DNA* H form requires a homopurine-homopyrimidine mirror repeat / S. M. Mirkin, V. I. Lyamichev, K. N. Drushlyak et al. // Nature.— 1987.—330, N 6147.— P. 495—497.
5. *Chemical* probing of homopurine-homopyrimidine mirror repeats is supercoiled DNA / O. N. Voloshin, S. M. Mirkin, V. I. Lyamichev et al. // Ibid.— 1988.—333, N 6172.— P. 475—476.
6. *Закономерности* сближенности нуклеотидов в кодирующих и некодирующих последовательностях ДНК из различных организмов / Ю. А. Сприжицкий, Ю. Д. Нечипуренко, А. А. Александров, М. В. Волькенштейн // Молекуляр. биология.— 1988.— 22, № 2.— С. 338—356.
7. *Субоч Г. М., Сприжицкий Ю. А.* Статистическая значимость встречаемости некоторых сложных сочетаний нуклеотидов: сравнение моделей ДНК // Биополимеры и клетка.— 1989.—5, № 4.— С. 30—37.
8. *Efron B.* Bootstrap methods: another look at the jacknife // Ann. Statist.— 1979.—7, N 1.— P. 1—26.
9. *Диаконис П., Эфрон Б.* Статистические методы с интенсивным использованием ЭВМ // В мире науки.— 1983.— № 7.— С. 60—73.
10. *GenBank* (1986). Genetic sequence data bank, R. 44.0. BBN laboratories, USA.
11. *Beckman J. S., Brendel V., Trifonov E. N.* Intervaning sequences exibit distinct vocabulary // J. Biomol. Struct. and Dyn.— 1986.—4, N 3.— P. 391—490.

Ин-т молекуляр. генетики АН СССР, Москва Получено 06.07.88

OCCURRENCE OF HOMOPURINE-HOMOPYRIMIDINE MIRROR REPEATS IN NATURAL DNAs

G. M. Suboch, Yu. A. Sprizhitsky, A. A. Alexandrov

Institute of Molecular Genetics, Academy of Sciences of the USSR, Moscow

Summary

Occurrence of the potential sites of the H-form (H-palindromes) formation in different functional regions of natural DNAs has been statistically analyzed in detail. Statistical significance of the results is estimated using model sequences and the Monte-Carlo method. It is shown that the unique frequencies of H-palindromes in all the samples, with the exception of the vertebrate introns, can be explained by the effect of the nonrandom

distribution of the nucleotide runs in natural DNAs. The values of frequencies of the H-palindromes occurrence are compared with frequencies of occurrence of all the homopurine and homopyrimidine mirror repeats with length of not less than 4 bp, and at a distance from 3 to 12 bp. It is shown that sites of picks' localization in distributions for H-palindromes and purine repeats do not correlate between themselves.

УДК 576.315.42

Г. М. Субоч, Ю. А. Сприжицкий

СТАТИСТИЧЕСКАЯ ЗНАЧИМОСТЬ ВСТРЕЧАЕМОСТИ НЕКОТОРЫХ СЛОЖНЫХ СОЧЕТАНИЙ НУКЛЕОТИДОВ: СРАВНЕНИЕ МОДЕЛЕЙ ДНК

Предложена схема моделирования цепочки ДНК как последовательности блоков нуклеотидов. Показано, что такая модель более адекватно описывает наблюдаемые частоты встречаемости локальных зеркальных гомопурин-гомопиримидиновых повторов, нежели марковская однородная модель второго порядка. Описывается методика оценки статистической значимости встречаемости в ДНК некоторых сложных сочетаний нуклеотидов.

Введение. Изучению встречаемости различных типов повторов в ДНК посвящено значительное число работ. В литературе описан ряд методов и компьютерных программ для поиска и оценки статистической значимости такого рода структур. Большинство исследователей сравнивают наблюдаемые значения частот встречаемости с ожидаемыми, рассчитанными аналитически на основе вектора частот олигонуклеотидов (такой подход приводится, например, в [1, 2]). Однако, учитывая неслучайный характер организации нуклеотидов в природных ДНК, который до конца не изучен, и то, что «словарь» ДНК нам известен лишь частично, закономерен вопрос об оптимальной в смысле набора и числа параметров модели цепочки ДНК, используя которую либо аналитически, либо методом Монте-Карло, можно получить оценки ожидаемых частот.

При анализе встречаемости в природных ДНК локальных гомопурин-гомопиримидиновых зеркальных повторов как потенциальных сайтов образования H-формы [3] возникла необходимость оценки ожидаемого числа таких структур. Ранее было показано [4], что в природных ДНК наблюдаемые частоты блоков типа поли(R), поли(Y), поли(A), поли(G) и т. д. значительно отличаются от ожидаемых, рассчитанных на основе нуклеотидного состава. Ясно, что учет этого эффекта может оказывать существенное влияние на оценку ожидаемого числа встречаемости таких специфических структур, как гомопурин-гомопиримидиновые повторы. Поэтому для получения подобных оценок мы генерировали случайные последовательности, в которых величины математического ожидания встречаемости блоков различных типов разной длины были равны полученным в [4] значениям для природных ДНК.

В данной работе описан алгоритм генерации такой последовательности. Показано, что моделирование цепочки ДНК как последовательности блоков нуклеотидов более адекватно описывает наблюдаемые частоты встречаемости локальных зеркальных гомопурин-гомопиримидиновых повторов, нежели марковская однородная модель второго порядка. Предложена методика оценки статистической значимости встречаемости в ДНК некоторых сложных сочетаний нуклеотидов (таких, например, как локальные гомопурин-гомопиримидиновые повторы) методом Монте-Карло, использующая процедуру бутстрепа и требующая сравнительно небольшого объема вычислений. Обсуждаются преимущества такого подхода и границы его применения.