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.

Алгоритмы и методы. Генерация модельных последовательностей. Генерацию последовательностей, соответствующих марковским однородным моделям нулевого и второго порядка, проводили по обычной схеме, использующей стандартный генератор псевдослучайных чисел. Рассмотрим алгоритм получения последовательности, соответствующей «блочной» модели цепочки ДНК.

Введем обозначения. Под блоком длины l понимается цепочка из l нуклеотидов одного вида, фланкированная с обеих сторон нуклеотидами другого вида. Будем рассматривать два типа блоков: поли(N) и поли(D), где N обозначает один из нуклеотидов A, T, G или C, а D — пурин(R) или пиримидин(Y). Нуклеотидную цепочку можно представить в виде последовательности чередующихся блоков типа D, каждый из которых состоит из чередующихся блоков типа N. Частоты встречаемости этих блоков обозначим через $F^N(l)$ (N=A, T, G, C) и $F^D(l)$ (D=R, Y). Назовем такие величины параметрами сближенности.

Нашей задачей является получение случайной последовательности, характеризуемой заданным набором параметров сближенности. Рассмотрим такую последовательность. Примем, что в блоке D на k-м месте находится блок типа N в случае совпадения первого нуклеотида блока N с k-м нуклеотидом блока D. Пусть $P^N{}_{ikl}$ — вероятность того, что, если в блоке D длины l на k-м месте находится блок N, то его длина равна i ($i \leqslant l-k+1$); $Q^N{}_{ikl}$ — вероятность того, что произвольно выбранный блок N длины i находится на k-м месте в блоке D длины l. Тогда

$$P^N_{ikl} = \frac{Q^N_{ikl} \cdot F^N(i)}{\sum_{i=1}^{l-k+1} Q^N_{ikl} \cdot F^N(i)}. \qquad (1)$$

Значения Q^N_{ikl} определяются как

$$Q^N_{ikl} = P_{il} \cdot P_{ikl}, \qquad (2)$$

где P_{il} — вероятность того, что произвольно выбранный блок типа N длины i находится в блоке D определенной длины l ($l \geqslant i$), а P_{ikl} — вероятность того, что первый нуклеотид этого блока N окажется на k-м месте в данном блоке D. Эти величины определяются комбинаторно. Число размещений блоков типа N длины $i \leqslant l$ внутри всех возможных 2^l блоков типа D длины l равно

$$n_{il} = \begin{cases} 2^{(l-i)} & \text{при } l-i < 2; \\ 2^{(l-i)} + (l-i-1) \cdot 2^{(l-i-2)} & \text{при } l-i \geqslant 2. \end{cases} \qquad (3)$$

Тогда $n_{il}/2^l$ есть среднее число блоков N длины i в случайном блоке D длины l. Отсюда

$$P_{il} = \frac{F^D(l) \cdot n_{il}/2^l}{\sum_{l=i}^{l_{\max}} F^D(l) \cdot n_{il}/2^l}. \qquad (4)$$

Вероятность P_{ikl} имеет следующие значения:

$$P_{ikl} = \begin{cases} 2^{(l-i-1)}/n_{il} & \text{при } k=1,\ i<l; \\ 1 & \text{при } k=1,\ i=l. \\ 2^{(l-i-2)}/n_{il} & \text{при } k>1,\ i \leqslant l-k; \\ 2^{(k-2)}/n_{il} & \text{при } k>1,\ i=l-k+1. \end{cases} \qquad (5)$$

Таким образом, из системы (1)—(5) мы находим набор значений $P^N{}_{ikl}$.

Генерацию случайной последовательности с заданным математическим ожиданием встречаемости блоков различных типов разных длин производили по следующей схеме:

1) выбор вида (R или Y) первого блока типа D;

2) выбор длины l блока D в соответствии с $F^D(l)$;

3) $k=1$ (позиция блока N внутри блока D);

4) выбор вида (A или G для R и T или C для Y) блока N в соответствии с $P^N{}_{ikl}$ (для $k=1$, $i \leqslant l$);

5) выбор длины i блока N в соответствии с $P^N{}_{ikl}$ (для $i \leqslant l-k+1$);
6) запись поли(N) длины i;
7) $k=k+i$;
8) смена вида блока N (G→A, A→G или T→C, C→T);
9) если $k \leqslant l$, то выполнить (5)—(9);
10) смена вида блока D (R→Y, Y→R);
11) если длина последовательности меньше заданной, повторить (2)—(11).

Построение распределений бутстрепа. Используя модельные последовательности и рассчитывая для них частоты встречаемости некоторых структур, мы сталкиваемся с проблемой оценки разброса этих частот, обусловленного статистически случайным характером генерации последовательности. Для этой цели мы предлагаем использовать процедуру типа бутстреп [5], смысл которой заключается в том, что на основе одной выборки имитируется процесс получения большого числа выборок, для которых строится распределение значений частот. В данной работе при оценке разброса ожидаемого числа локальных зеркальных гомопурин-гомопиримидиновых повторов для всех наборов параметров генерировали по три модельные последовательности длиной 10^5 нуклеотидов, для каждой из которых составляли таблицу локализации указанных структур. Затем эти последовательности 100 раз разбивали случайным образом на 100 частей по 1000 нуклеотидов и каждый раз в полученной суммированием этих частей последовательности рассчитывали частоту встречаемости повторов на основе таблицы их локализации в исходной последовательности. Полученные таким образом распределения использовали для оценки доверительных интервалов принятия нулевой гипотезы о равенстве наблюдаемых и ожидаемых частот встречаемости повторов.

Такую же процедуру применяли для проверки правильности алгоритма генерации случайной последовательности с заданными параметрами сблоченности. Для этого случайную последовательность разбивали на участки, для каждого из которых параметры сблоченности были уже рассчитаны. Затем случайным соединением этих участков в одну последовательность исходной длины получалось множество последовательностей, на основе которого строили распределения бутстрепа для частот встречаемости блоков различных типов. Из табл. 1 видно, что во всех случаях за исключением двоек для поли(G) и поли(C) значения заданных параметров сблоченности попадают в 95 %-ный доверительный интервал для средних значений параметров сблоченности модельной последовательности. Исключения могут быть следствием малого числа бутстреп-выборок.

Применение нами процедуры бутстрепа обусловлено большими затратами машинного времени для поиска повторов и расчета параметров сблоченности нуклеотидов для такого числа сгенерированных случайных последовательностей, которое позволило бы нам сделать достоверные оценки границ доверительных интервалов принятия гипотез прямым методом. Следует подчеркнуть, что, хотя метод бутстрепа и является приближенным, в большинстве случаев он дает значения дисперсии и средние величины, близкие к реальным (ошибка около 10 %) [6].

Сравнение моделей. Одной из целей статистического моделирования является выбор из альтернативных моделей такой, которая наиболее адекватно отражала бы те или иные свойства реального объекта. Для любой модели мы имеем вектор входных переменных (параметров) $\overline{X}=[x_1, \ldots, x_n]$ и выходные переменные или отклик $Y=g(\overline{X})+r$, где $g(\overline{X})$ — некоторая функция от $\overline{X}$, называемая детерминированной составляющей отклика, а r — его случайная составляющая. Пусть Y_0 — характеристика реального объекта, соответствующая рассматриваемому отклику модели.

Тогда мерой адекватности модели может служить величина $C_y = \left[\sum_{\overline{x}} \sum_{\Omega} (Y_0-Y)^2\right]^{1/2}/N$, где первая сумма берется по всем возможным входным переменным, вторая — по множеству используемых модельных объектов для каждого $\overline{X}$; $1/N$ — нормировочный множитель. Также дополнительно можно рассматривать значения $C_r = \left[\sum_{\overline{x}} \sum_{\Omega} r^2\right]/N$, отражающие качество проведения численного эксперимента и зависящие от многих факторов, таких, например, как число модельных объектов или способ получения ряда псевдослучайных чисел.

В качестве отклика для моделей нуклеотидных последовательностей могут рассматриваться частотные характеристики различных структур. В данной работе такими

структурами являлись локальные зеркальные гомопурин-гомопиримидиновые повторы (типа AGAAG ... GAAGA, минимальная длина повтора равнялась 4 или 5, расстояние между ними — от 3 до 12 нуклеотидов) и H-палиндромы (аналогичные повторы, у которых GC-содержание повтора не ниже 75 %, а GC-содержание разделяющего участка менее 50 %). Сравнивались три модели: марковские однородные модели нулевого и второго порядков и «блочная» модель цепочки ДНК, входными переменными $\overline{X}$ кото-

Таблица 1

Частоты встречаемости блоков на 100000 нуклеотидов в выборке кодирующих областей млекопитающих и в случайной последовательности, сгенерированной по этим параметрам

The occurrence frequencies of the runs (per 100000 bp.) in mammalian coding regions and in random sequence with these parameters

Длина	Тип	Млекопитающие	Средние значения распределения бутстрепа	Границы доверительного интервала с 5 %-ным уровнем значимости	
				Нижняя	Верхняя
1	R	10460	10548	10255	10822
	Y	10581	10622	10368	10897
	A	13719	13460	13243	13778
	G	13756	13617	13379	13855
	C	12957	12785	12557	12999
	T	13889	13978	13744	14167
2	R	5454	5481	5299	5658
	Y	5718	5878	5701	6048
	A	3400	3450	3344	3563
	G	4489	4316	4162	4460
	C	4420	4214	4090	4335
	T	2481	2657	2482	2762
3	R	2721	2746	2613	2873
	Y	2750	2816	2688	2930
	A	847	928	847	1005
	G	930	965	909	1022
	C	1169	1215	1133	1291
	T	612	644	589	702
4	R	1653	1675	1569	1778
	Y	1653	1606	1522	1695
	A	239	252	213	292
	G	262	258	219	299
	C	348	318	283	353
	T	131	154	126	182
5	R	1086	1085	1017	1151
	Y	944	931	862	993
	A	63	70	50	89
	G	66	74	54	95
	C	92	72	54	93
	T	31	37	23	52

Примечание. Для блоков длиной более пяти нуклеотидов результаты аналогичны.

рых являлись либо вектор частот нуклеотидов ($n=4$), либо вектор частот тринуклеотидов ($n=64$), либо набор параметров сблоченности. (Мы выбрали $n=60$, так как блоки длины более 10 нуклеотидов встречаются редко, и значения параметров сблоченности для $l>10$ аппроксимировались экспоненциальной зависимостью частот блоков от их длины.)

Мы получали значения Y_0 (частот встречаемости повторов в природных ДНК), а значения Y рассчитывали по последовательностям, полученным методом бутстрепа из первоначально сгенерированных для каждой модели трех случайных последовательностей. Детерминированная составляющая отклика в нашем случае — среднее значение Y для множества бутстреп-выборок (т. е. математическое ожидание встречаемости изучаемых повторов в ДНК), а среднее значение случайной составляющей — его дисперсия.

Чтобы осуществить желаемый выбор модели, нужно рассчитать величину C_y по всему множеству входных переменных (набором частот блоков или олигонуклеотидов, наблюдаемых в природных ДНК) и по числу сгенерированных модельных последовательностей, дающему достоверный результат. Охватить все множество параметров невозможно, поэтому мы выбрали только два набора $\bar{X}$, соответствующих статистике

Таблица 2

Значения C'_y для различных множеств модельных последовательностей и разных откликов модели

The values of C'_y for different model sequences and for different Y (see text): 1 — for purine repeats of length 4 or more nucleotides; 2 — for purine repeats of length 5 or more nucleotides; 3 — for H-palindromes with length 4 or more

Источник	1			2			3		
	N	T	R	N	T	R	N	T	R
Фаг λ	57	109	19	13	21	7	46	38	4
Грызуны	962	686	518	231	171	120	73	42	8

Примечание. 1 — для пуриновых повторов длиной четыре и более нуклеотидов; 2 — то же пять и более нуклеотидов; 3 — для H-палиндромов длиной четыре и более нуклеотидов

нуклеотидов ДНК фага λ ($L=48502$ нуклеотида) и некодирующих областей ДНК грызунов ($L=291800$ нуклеотидов), и сгенерировали по три последовательности для каждой из трех рассматриваемых моделей. Значения C'_y для каждого набора параметров, приведенных в табл. 2, рассчитывали по формуле:

$$C'_y = \left[\left[\sum_{i=1}^{3} (Y_0 - g_i(X))^2\right]^{1/2}\right]/3. \qquad (6)$$

Результаты и обсуждение. На рис. 1 представлены распределения бутстрепа для частот встречаемости локальных зеркальных гомопурин-гомопиримидиновых повторов в ДНК фага λ. Видно, что для пуриновых повторов с $l \geqslant 4$ (рис. 1, *а*) нужно отвергнуть гипотезу о равенстве наблюдаемых и ожидаемых частот, рассчитанных на основе статистики моно- (модель N) и тринуклеотидов (модель T). Однако мы принимаем эту гипотезу для модели R, учитывающей параметры сблоченности. Иначе говоря, повышенное содержание повторов с $l \geqslant 4$ в природных ДНК фага λ по сравнению с ожидаемым для случайной последовательности с тем же моно- и тринуклеотидным составом может объясняться специфическим набором параметров сблоченности ДНК этого фага. Для повторов с $l \geqslant 5$ (рис. 1, *б*) статистически значимого отклонения наблюдаемых частот от ожидаемых не наблюдается ни для одной из рассматриваемых моделей. Рис. 1, *в*, так же как и рис. 1, *а*, демонстрирует преимущества блочной модели с той разницей, что частота H-палиндромов в ДНК фага λ значительно меньше, чем в модельных последовательностях N и T.

На рис. 2 изображены аналогичные распределения для модельных последовательностей с параметрами, соответствующими некодирующим областям ДНК грызунов. В данном случае повышенное содержание H-палиндромов может объясняться характерным набором параметров сблоченности, тогда как частоты пуриновых повторов (рис. 2, *а*, *б*) не могут быть объяснены ни одной из моделей. Однако видно, что модель R дает наиболее близкие оценки (значения C_y из табл. 2 численно отражают этот факт).

Таким образом, проведенные расчеты показывают, что при оценке ожидаемых частот встречаемости сложных структур, подобных рассмотренным локальным повторам, наряду с частотами олигонуклеотидов целесообразно учитывать и эффект сблоченности нуклеотидов, осо-

бенно для некодирующих последовательностей. Подчеркнем также эффективность применения метода Монте-Карло совместно с процедурой бутстрепа для оценки доверительных интервалов принятия гипотез для распределений, близких к нормальным. Такой подход освобождает нас от весьма трудоемких выводов формул для ожидаемых частот и диспер-

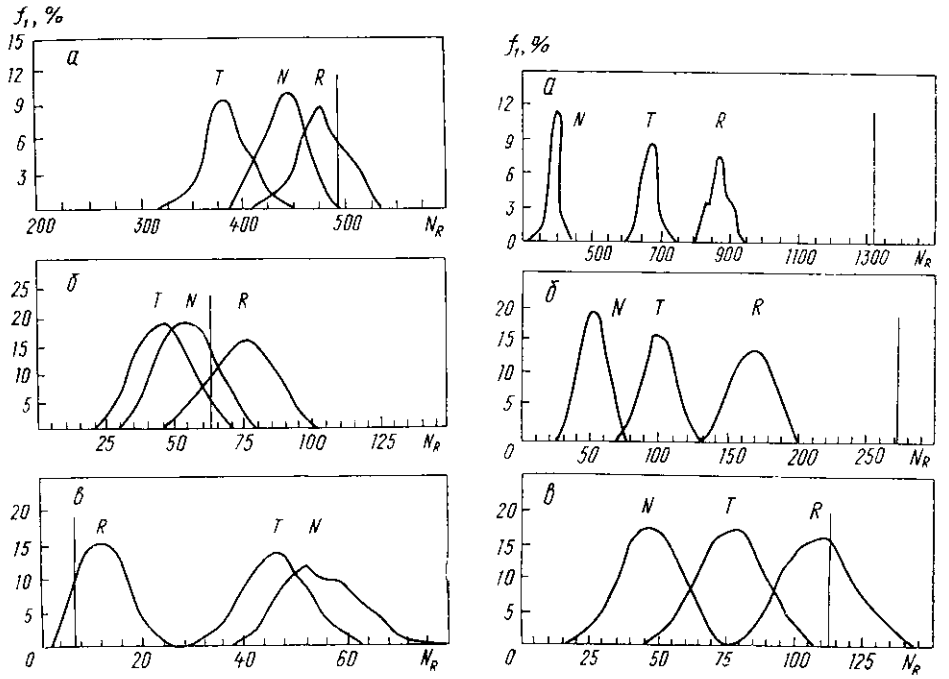

Рис. 1. Плотности распределения частот встречаемости локальных зеркальных гомопурин-гомопиримидиновых повторов: *а* — длиной не меньше 4 нуклеотидов; *б* — длиной не меньше 5 нуклеотидов; *в* — Н-палиндромов. Распределения получены для модельных последовательностей с заданными частотами нуклеотидов (N), тринуклеотидов (T) и блоков нуклеотидов (R), рассчитанными для ДНК бактериофага λ. По оси абсцисс отложено число повторов в модельной последовательности длиной 100 000 нуклеотидов; по оси ординат — относительная доля последовательностей, содержащих данное число повторов. Распределения получены путем применения процедуры бутстрепа к первоначально сгенерированным случайным последовательностям и стандартной процедуры сглаживания с окном семь точек. Вертикальные линии отмечают на оси абсцисс средние значения частот изучаемых повторов в исходной природной ДНК

Fig. 1. The probability densities of the frequencies of the local homopurine-homopyrimidine mirror repeats: *а* — with a length not less than 4 nucleotides, *б* — with a length not less than 5 nucleotides, *в* — H-palindromes; in the model sequences with given frequencies of bases — (N), with given frequencies of trinucleotides — (T) and in the sequences generated using the frequencies of nucleotide runs — (R). These parameters are calculated for phage DNA. Abscissa: the occurrence numbers of these structures; ordinate: the percentage of sequences containing a given number of such repeats, f_i. The distributions have been derived through the bootstrap method applied to the model sequence and using standard smoothing procedure to the experimental points. The vertical lines correspond to the average values of the frequencies of repeats in phage DNA

Рис. 2. Плотности распределения частот встречаемости локальных зеркальных гомопурин-гомопиримидиновых повторов в модельных последовательностях с параметрами, соответствующими некодирующим областям ДНК грызунов (см. подпись к рис. 1)

Fig. 2. The probability densities of the frequencies of the local homopurine-homopyrimidine mirror repeats in the model sequences with parameters of the rodent noncoding regions (see legend to Fig. 1)

сий, а также помогает избежать необходимости учета самопересечения как олигонуклеотидов, так и изучаемых структур, способного оказывать существенное влияние на результаты.

Задача выявления статистически значимых отклонений наблюдаемых частот встречаемости тех или иных сочетаний нуклеотидов может быть разделена на три части: 1) выбор модели, порождающей случайную последовательность; 2) построение для данной модели распреде-

ления вероятностей (аналитически, Монте-Карло или Монте-Карло в сочетании с бутстрепом); 3) сравнение наблюдаемых величин с доверительным интервалом модельного распределения. Уже на первом этапе возникает существенная неоднозначность, являющаяся следствием множественности возможных моделей случайной последовательности. Получаемые на втором и третьем этапах оценки существенно зависят от выбора модели: наблюдаемые частоты могут значимо отличаться от случайных для одной модели и попадать в выбранный доверительный интервал для другой. Следовательно, не имеет смысла ставить вопрос о неслучайном характере встречаемости тех или иных структур вообще, вне связи с конкретной моделью порождения генетического текста. Точнее, следует говорить о том, достаточна или нет данная модель, т. е. те или иные корреляции нуклеотидов в природных последовательностях ДНК, для описания наблюдаемых частот интересующих нас сочетаний нуклеотидов.

Биологическая задача в подобных исследованиях ставится несколько иначе: оказывает ли отбор давление на поддержание или, наоборот, на дискриминацию изучаемых структур или же их частоты встречаемости случайны в том смысле, что являются простым следствием других факторов. Очевидно, что эта и сформулированная ранее задачи о статистической значимости не эквивалентны, поскольку ни одна модель не может учесть всех факторов естественного отбора. Отсутствие статистически значимых отклонений от достаточно простой модели может служить веским основанием считать, что отбор не оказывает давления на число изучаемых структур. С другой стороны, наличие значимых отклонений по отношению к одной или даже нескольким моделям не является доказательством существования такого давления, а, следовательно, и функциональной значимости изучаемых структур.

Таким образом, статистический анализ в сочетании с моделированием может позволить только отвергнуть гипотезу о давлении отбора, оказываемом на число тех или иных нуклеотидных сочетаний в исследуемом наборе последовательностей. При этом отсутствие такого давления еще не является поводом для отрицания их биологической роли, которая может быть никак не связана со средней по последовательности частотой их встречаемости. Тем не менее статистический анализ, не являясь строгим в вышесказанном смысле методом, дает возможность исследователю понять многие закономерности в строении генома, формулировать гипотезы относительно особенностей его функционирования и эволюции.

Другая важная задача, к решению которой непосредственное отношение имеют рассматриваемые в данной работе подходы,— это распознавание функциональной структуры фрагмента генома по его нуклеотидной последовательности. Модели различных областей генома могут служить образами, на сравнении с которыми и строятся некоторые алгоритмы распознавания [8]. В связи с этим весьма существенным является выбор наиболее адекватной модели. Предложенная в настоящей работе «блочная» модель является, на наш взгляд, наиболее удачной для описания некодирующих областей и может быть использована в соответствующих алгоритмах.

СПИСОК ЛИТЕРАТУРЫ

1. *Day G. R., Blake R. D.* Statistical significance of symmetrical and repetitive segments in DNA // Nucl. Acids Res.— 1982.—**10**, N 24.— P. 8323—8339.
2. *Saurin W.* Repetitive palindromic sequences in *Escherichia coli*: detection and characterization with a new computer program // CABIOS.— 1987.—**2**, N 2.— P. 121—127.
3. *Субоч Г. М., Сприжицкий Ю. А., Александров А. А.* Встречаемость гомопурин-гомопиримидиновых зеркальных повторов в природных ДНК // Биополимеры и клетка.— 1989.— 5, № 4.— С. 24—30.
4. *Закономерности* сблоченности 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, № 22.— С. 338—356.

5. *Efron B.* Bootstrap methods: another look at the jacknife // Ann. Statist.— 1979.—**7**, N 1.— P. 1—26.
6. *Boos D. D., Monahan J. F.* Bootstrap methods using prior information // Biometrica.— 1986.—**73**, N 1.— P. 77—83.
7. *GenBank* (1986). Genetic sequence data bank, R. 44.0. BBN laboratories, USA.
8. *Статистические* закономерности в первичных структурах функциональных областей генома *Escherichia coli*. 3. Компьютерное распознавание кодирующих областей / М. Ю. Бородовский, Ю. А. Сприжицкий, Е. И. Голованов, А. А. Александров // Молекуляр. биология.— 1986.—**20**, № 5.— С. 1390—1398.

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

STATISTICAL SIGNIFICANCE OF THE OCCURRENCE OF SOME COMPLEX NUCLEOTIDE COMBINATIONS: COMPARISON OF THE DNA MODELS

G. M. Suboch, Yu. A. Sprizhitsky
Institute of Molecular Genetics, Academy of Sciences of USSR, Moscow

Summary

A scheme for modeling of the DNA chain as a sequence of the nucleotide runs of different length is presented. The advantages of such a method and range of its application are discussed. A procedure is suggested to estimate statistical significance of occurrence of some complex sequence structures in DNA by the Monte-Carlo method. It uses a bootstrap algorithm and necessitates comparatively small number of calculations. Three different models, used to derive such estimations of the frequencies of homopurine-homopyrimidine mirror repeats in the DNA of phage λ and rodentia noncoding regions are compared.

УДК 577.27

О. В. Рохлин, А. Соломон, Д. Вайсс, А. Р. Ибрагимов, Е. Л. Арсеньева, Г. Т. Богачева

ЭКСПРЕССИЯ ЭПИТОПОВ С-ДОМЕНА L-ЦЕПЕЙ КАППА И ЛЯМБДА ТИПОВ ИММУНОГЛОБУЛИНОВ ЧЕЛОВЕКА У БЕЛКОВ РАЗЛИЧНЫХ ВАРИАБЕЛЬНЫХ ПОДГРУПП

С помощью монАТ исследована экспрессия эпитопов С-домена L-цепей Ig *человека каппа и лямбда типов различных V-подгрупп. Установлено, что уровень экспрессии данного эпитопа С-домена зависит от принадлежности L-цепи к той или иной подгруппе.*

Введение. Легкие цепи (*L*-цепи) иммуноглобулинов (Ig) человека подразделяются на два типа, обозначаемые каппа (κ) и лямбда (λ) [1]. Примерно 60 % Ig человека содержат *L*-цепи κ-типа, 40 % — λ. Строение постоянного (*C*) домена κ-цепей контролируется единственным геном [1], тогда как строение *C*-домена λ-цепей определяется девятью генами, часть из которых относится к категории псевдогенов [2, 3]. Строение вариабельных (*V*) доменов *L*-цепей определяется несколькими десятками генов, и их разнообразие увеличивается в процессе дифференцировки В-лимфоцитов за счет генных перестроек и соматических мутаций [1]. Как по особенностям первичной структуры, так и по иммунохимическим свойствам *V*-доменов κ-цепи можно разбить на четыре *V*-подгруппы, а λ-цепи — на шесть *V*-подгрупп [1]. В ряде лабораторий получены моноклональные антитела (монАТ) к *C*-доменам κ- и λ-цепей [4, 5], однако экспрессия соответствующих антигенных детерминант (эпитопов) не изучена у индивидуальных белков различных *V*-подгрупп κ- и λ-цепей. Между тем существуют косвенные данные, указывающие на то, что *V*-домен способен влиять на активность определенных эпи-