МЕТОД ФИКСИРОВАННЫХ ЧАСТОТ ДЛЯ АНАЛИЗА НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДНК

Федосеев С.В., Кудряшов Н.А., Коротков Е.В.

Московский Инженерно Физический Институт, г. Москва,

Мы разработали метод фиксированных частот для анализа скрытой периодичности в нуклеотидных последовательностях ДНК. Метод был применен для исследования 22-ой хромосомы человека. Периодичность была определена для периодов от T=2 до T=10. Ранее разработанные математические методы используют теорию информации [1], Фурье и вейвлет анализ [2]. Изучение скрытой периодичности в ДНК, по нашему мнению, позволит выявить новые свойства и функции ДНК.1. E. V. Korotkov, M. A. Korotkova, N. A. Kudryashov, Phys. Lett. A. (2003).2. G. Dodin, and al., J. Theor. Biol. 206, 323 (2000).

1. Введение.

В настоящее время в области изучения последовательностей оснований ДНК работает большое количество исследований в разных странах. Значительные усилия направлены на создание новых компьютерных банков данных, которые содержат как сами последовательности, так и результаты, полученные на основе изучения этих последовательностей. Наиболее известные из них: Genetic Sequence Data Bank (GENBANK) в США и банк Европейской молекулярно-биологической лаборатории EMBL (European Molecular Biology Laboratory Гейдельберг, ФРГ).

В данной работе мы предприняли попытку разработать математический метод, который бы позволил обнаруживать так называемую скрытую периодичность [1] в присутствии небольшого количества вставок и делеций символов. Ранее был развит метод информационного разложения символьных последовательностей, который позволяет обнаружить скрытую периодичность в любых символьных последовательностях без учета возможности делеций и вставок символов [1]-[3]. Также для поиска скрытой периодичности применяется методы Фурье и вейвлет анализа [4].

Мы применили развитый нами в настоящей работе подход к изучению полных геномов и некоторых полных хромосом человека. Полученные результаты показывают, что в хромосомах, как эукариот так и прокариот существуют протяженные районы с определенной структурированностью последовательностей оснований ДНК, которая не обусловлена кодирующими свойствами данной последовательности и таким образом не связана с триплетной периодичностью. Мы предполагаем, что данные последовательности могут представлять “горячие” районы для генетической рекомбинации, где мы видим в последовательности последствия множественных генетических перестроек. Можно также предполагать, что определенная структурированность последовательностей оснований ДНК может облегчать осуществление эволюционных преобразований геномов. В данной работе мы представляем результаты обработки нуклеотидной последовательности генома 22-й хромосомы человека взятой на сайте GENBANK. Предлагаемый метод обнаруживает скрытую периодичность в исследуемой последовательности. В приложении показаны схемы (рис. 1-4) для полной нуклеотидной последовательности 22-ой хромосомы, с расположенными поверх участками со скрытой периодичностью, обнаруженными предлагаемым методом.

2. Метод.

2.1 Метод фиксированных частот (М. Ф. Ч.).

Последовательность оснований ДНК имеет алфавит - из четырех символов {a,c,t,g}. Каждый символ означает определенный вид нуклеотида: a – адеин, c – цитозин, t – тимин, g – гуанин. В дальнейшем будем обозначать символьную последовательность символом G, а каждый определенный символ последовательности будет g_k, k=1,…,N – общее количество символов, k – порядковый номер символа в последовательности. Алфавит последовательности G обозначим как A={a,c,t,g} для ДНК (в общем случае алфавит может состоять из любого количества символов).

Сделаем вспомогательные определения. Рассмотрим символьную последовательность $G=\{g_1,\ldots,g_N\}$ нуклеотидов ДНК. Символ G[a,b] обозначает подпоследовательность $G[a,b]=\{g_a,g_{a+1},\ldots,g_b\}$.Частотой символа ‘a’ для G[a,b] обозначим количество символов ‘a’ содержащихся в последовательности G[a,b]. Эта частота есть n(a). Аналогичным образом определяются частоты других символов для G[a,b]: n(c),n(t) и n(g).

Пусть в последовательности G существуют два повторяющихся фрагмента: G[a,b]=G[c,d]. Тогда частоты для них совпадают по всем символам: n_1(‘a’)=n_2(‘a’); n_1(‘c’)=n_2(‘c’); n_1(‘t’)=n_2(‘t’); n_1(‘g’)=n_2(‘g’) (здесь n_1-обозначена частота для G[a,b] и n_2 частота для G[c,d]).

Дадим определение скрытой периодичности. Для этого рассмотрим последовательность символов G. Пусть в G присутствует подпоследовательность G[i,i+np], такая, что: $g_{j+rp+i}=g_{j+i}$, для любых i, r ($1\le i \le n$, $1 \le r < p$, $j \ge 0$, $n \ge 2$, $p \ge 2$). Тогда последовательность: G[i,i+n] называют кластером с периодом n. Если последовательность G[i,i+np] будет изменена случайным образом, так чтобы изменения были статистически маловероятны, тогда последовательность G[i,i+np] будет называться последовательностью со скрытой периодичностью.

Для последовательности со скрытой периодичностью равенства частот выполняться не будут. Поставим цель отыскания некоторой характеристики, позволяющей судить о приближенном совпадении фрагментов G[a,b] в последовательности G .

Рассмотрим последовательность G. Зафиксируем значение частоты для символа ‘a’: n(a)=const. Разобьем G на непересекающиеся последовательности, таким образом, чтобы каждая последовательность содержала ровно заданное количество символов ‘a’. Разбиение производим с начала последовательности G, включая по очереди в подпоследовательность все символы, до тех пор, пока символов ‘a’ не наберётся ровно n(a). (см. Приложение, пример 1).

Таким образом, получим несколько последовательностей с разными границами: $G[a_1,b_1]$, $G[a_2,b_2]$, $G[a_3,b_3]$, … , $G[a_m,b_m]$, где ($a_k=b_{k-1}+1$, k=2…m). Здесь m – их количество. Рассмотрим теперь длины этих последовательностей. Обозначим их соответственно L_k. Каждая L_k это общее количество символов в $G[a_k,b_k]$. Легко видеть, что $L_k=b_k-a_k+1$, (k =1..m). Так как длина последовательности не может быть меньше количества символов ‘a’ содержащихся в ней, то для любых k: $L_k \ge n(a)$.

В случае, когда длина G достаточно большая (например, более 500 символов) и количество последовательностей G[a,b] также велико, то удобно рассматривать спектр для величин $L=L_k$. (см. Приложение, пример 2).

Если последовательность G разыгрывается методом Монте-Карло и появления каждого (a,c,t,g) символа равновероятно (p = ¼), то спектр такой последовательности можно посчитать теоретически. Рассмотрим последовательность G из алфавита A={a,c,t,g}. Пусть n – частота символа ‘a’ из алфавита A. Тогда спектр P для длин L построенных с фиксированной частотой n=n(a) будет таким:

$$P(n,L)=\frac{1}{4}C_L^n 4^{-L}3^{L-n} \text{ , где } L \ge n \text{ (n=const).} \quad (2.1.1)$$

У каждой последовательности G[a,b] построенной на основе зафиксированной частоты последний символ является символом ‘a’. Следовательно, вероятность P(n,L) - есть вероятность обнаружить среди L символов n символов ‘a’ при условии, что последним будет обязательно символ ‘a’. Значит:

$$P(n,L) = p\, P_b(n,L) \qquad (2.1.2)$$

Здесь p – вероятность того, что последний символ именно символ ‘a’,
$P_b(n,L)$– вероятность обнаружить среди L символов n символов ‘a’.

Так как появление символов равновероятно и равно ¼ то p = ¼. $P_b(n,L)$ – биномиальная вероятность с p = ¼, и q=1-p (см. определение биномиальной плотности вероятности):

$$P_b(n,L) = C_L^n p^{-n} q^{L-n}$$

При подстановке в (2.1.2) величин $P_b(n,L)$ и p = ¼, получим выражение (2.1.1).
Это верно также для любого другого символа (не только для ‘a’).
На рис. 5 (приложение) для примера показан P(n=5,L).

2.2. Критерий согласия χ^2.

Применим теперь критерий согласия Пирсона (или критерий χ^2) для анализа последовательности $\{L_k\}$, для заданной символьной последовательности G. Разобьем пространство значений величины $L=L_k$ на конечное число r частей $S_1 \ldots S_r$ без обших точек, и пусть $p_1 \ldots p_r$ соответствующие значения заданной вероятностной функции P, так что $p_i=P(S_i)$ и $\sum_{i=1}^{r} p_i = 1$. Здесь $P(L_k)$ – это теоретическая вероятность обнаружить значение $L=L_k$, а $P(S_i)$ – вероятность обнаружить значения L принадлежащие пространству значений S_i.
Пусть соответствующие групповые частоты в выборке есть $f_1 \ldots f_r$, так что каждому множеству S_i принадлежит f_i выборочных значений и при этом $\sum_{i=1}^{r} f_i = n$. Теперь рассмотрим выражение:

$$\chi^2 = \sum_{i=1}^{r} x_i$$

где

$$x_i = \frac{1}{np_i}(f_i - np_i)^2 \qquad (2.2.1)$$

математическое ожидание и дисперсия:

$$M(\chi^2) = r-1$$

$$D(\chi^2)=2(r-1)+\frac{1}{n}\left(\sum_{i=1}^{r}\frac{1}{p_i} - r^2 - 2r + 2\right)$$

При $n\to\infty$ выборочное распределение величины χ^2 стремится к распределению, определяемому плотностью вероятности:

$$k_{r-1}(x) = \left\{2^{\frac{r-1}{2}}\Gamma\left(\frac{r-1}{2}\right)\right\}^{-1} x^{\frac{r-3}{2}} e^{-\frac{x}{2}}, \quad (x>0).$$

Распределение χ^2 с r-1 степенями свободы.
P-процентный уровень значимости определяется из выражения:

$$\beta=P(\chi^2>\chi_\beta^2)=\int_{\chi_\beta^2}^{\infty} k_n(x)dx. \quad (2.2.2)$$

Величина β есть вероятность того, что отклонения спектра P(L) от спектра определяемого вероятностной функцией P случайны.

2.3. Влияния триплетной периодичности.

Ранее, в многочисленных публикациях было показано, что кодирующие участки генома содержат так называемую триплетную периодичность. Ее присутствие в кодирующих областях связано с неравномерным использованием аминокислот в белковых молекулах. Чтобы триплетная периодичность не влияла на результаты нашего исследования по поиску скрытой периодичности со вставками и делециями, при расчете величин χ^2 и r её влияние нужно учитывать. Для учета влияния триплетной периодичности в критерии согласия мы строили вероятностную функцию P следующим образом. Мы разбиваем символьную последовательность G на триплеты символов, и строим следующую матрицу вероятностей:

	A	**T**	**C**	**G**
1	P_{A1}	P_{T1}	Pc_1	P_{G1}
2	P_{A2}	P_{T2}	Pc_2	P_{G2}
3	P_{A3}	P_{T3}	Pc_3	P_{G3}

$$P_{A1}=\frac{n_{A1}}{N_1},\ P_{T1}=\frac{n_{T1}}{N_1},\ P_{C1}=\frac{n_{C1}}{N_1},\ldots P_{G3}=\frac{n_{G3}}{N_3}.$$

Где:

n_{A1} - количество символов A на первых позициях триплетов; n_{A2} - на вторых позициях триплетов, и т.д.; N_1 - общее количество символов на первых позициях триплетов, N_2 - на вторых позициях триплетов, и т. д. Так как каждый элемент матрицы есть вероятность появления символа указанного в столбце на соответствующем месте(1,2 или 3), то должны выполнятся равенства:

$$P_{A1}+P_{T1}+P_{C1}+P_{G1}=1$$
$$P_{A2}+P_{T2}+P_{C2}+P_{G2}=1$$
$$P_{A3}+P_{T3}+P_{C3}+P_{G3}=1$$

Затем методом Монте-карло на основе этой матрицы необходимо сгенерировать несколько (например, 100) случайных последовательностей символов. Для каждой из этих последовательностей определяются спектры: P_{A1}, P_{C1}, P_{T1}, P_{G1}, ; P_{A2}, P_{C2}, P_{T2}, P_{G2},… P_{A100}, P_{C100}, P_{T100}, P_{G100}. Затем эти спектры усредняются:

$$P_A=\langle P_{Ai}\rangle_{i=1..100},\ P_C=\langle P_{Ci}\rangle_{i=1..100},\ P_T=\langle P_{Ti}\rangle_{i=1..100},\ P_G=\langle P_{Gi}\rangle_{i=1..100}.$$

Полученные плотности вероятностей используются в критерии согласия в качестве теоретической плотности для соответствующего символа. В частности, когда появление всех символов равновероятно, то:

$$P_A(n,L)=P_C(n,L)=P_T(n,L)=P_G(n,L)=\frac{1}{4}C_L^n 4^{-L}3^{L-n}\ ,\ \text{где } L\geq n\ (n=\text{const}).$$

Здесь n-фиксированная частота, L-длина последовательностей.

В качестве оценки отклонений спектра исследуемой последовательности для заданного символа, от полученного усреднённого спектра мы берём уровень доверия Пирсона.

3. Результаты.

3.1. Применение метода фиксированных частот для анализа 22-ой хромосомы человека.

Разработанный метод применялся для исследования нуклеотидной последовательности 22-й хромосомы человека, полная длина которой около 33 млн. нуклеотидов. В результате этого в исследуемой последовательности были обнаружены фрагменты со скрытой периодичностью. На рис. 3.1.1-3.1.3 приведены схемы расположения этих фрагментов. По горизонтальной оси отложен порядковый номер символа. Вертикальная ось разбита на 8

частей, каждая из которых соответствует фиксированной частоте, в диапазоне от 3 до 10. Цветами обозначены различные символы в последовательности. Красный – ‘a’, зеленый – ‘c’, синий – ‘t’, черный – ‘g’. Суммарная длина последовательности 10 млн. нуклеотидов.

Анализ фрагментов нуклеотидной последовательности полученных МФЧ показывает, что периодичность наиболее четко выражена для адеина (‘a’) и тимина (‘t’). По-видимому, это связано с тем, что данные символы наиболее часто встречаются в периодичных фрагментах. Спектры данных символов имеют четко выраженные отклонения от триплетных спектров. Таким образом, данные фрагменты не обусловлены кодирующими свойствами последовательности.

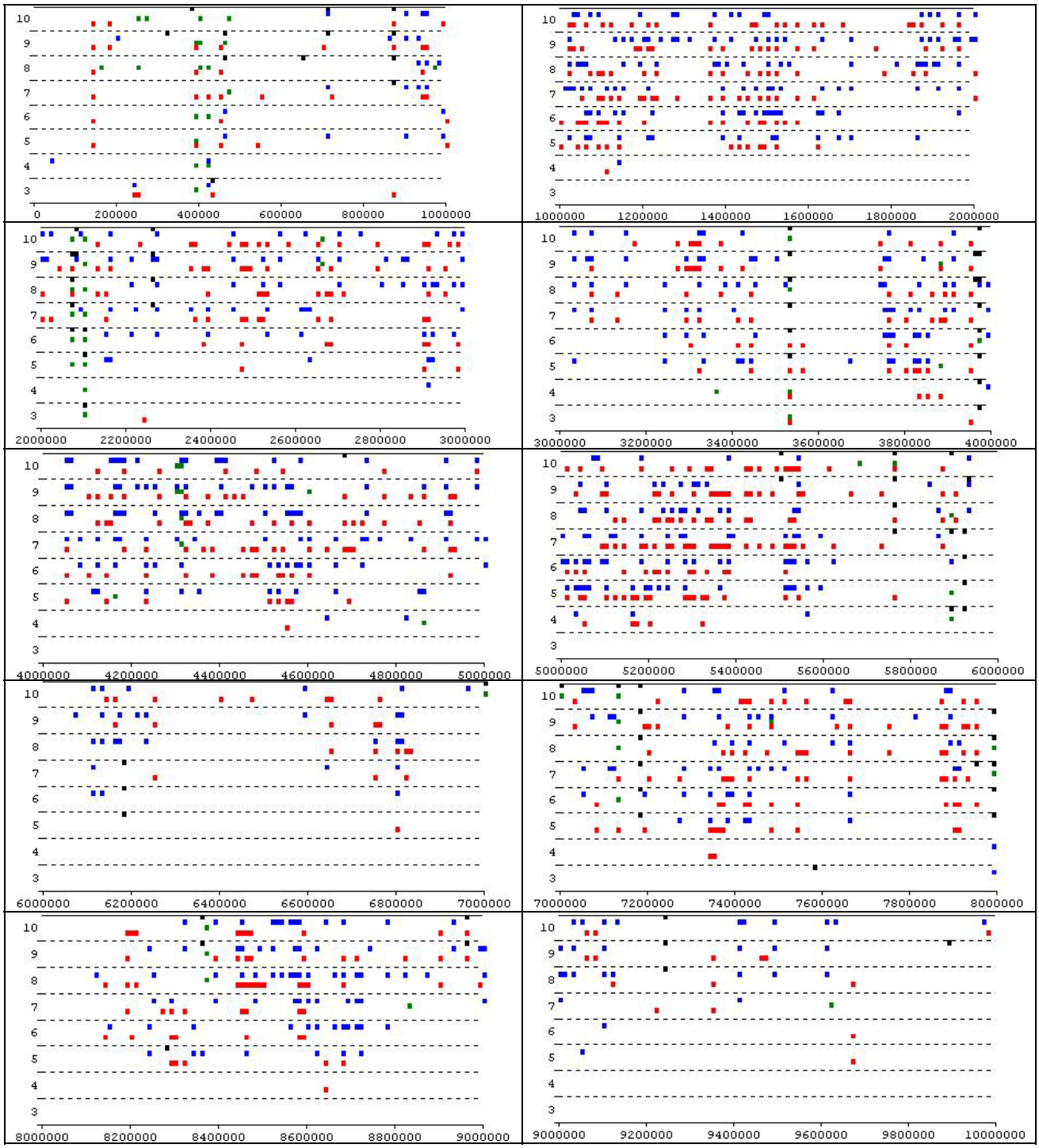


Рис. 1. Фрагменты со скрытой периодичностью полученные М.Ф.Ч. для участка [0,107] 22-ой хромосомы человека.

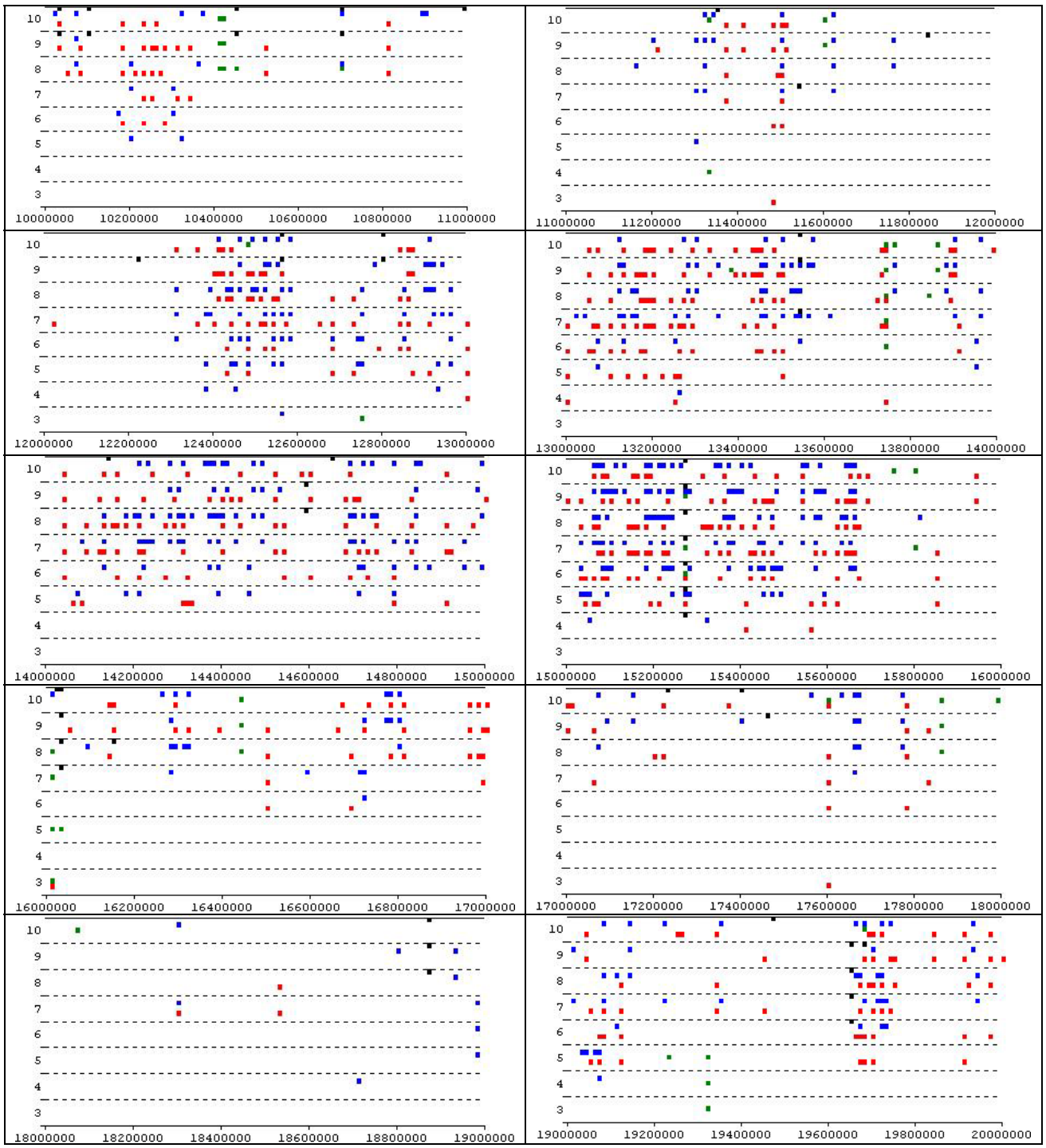


Рис. 2. Фрагменты со скрытой периодичностью полученные М.Ф.Ч. для участка [107, 2×107] 22-ой хромосомы.

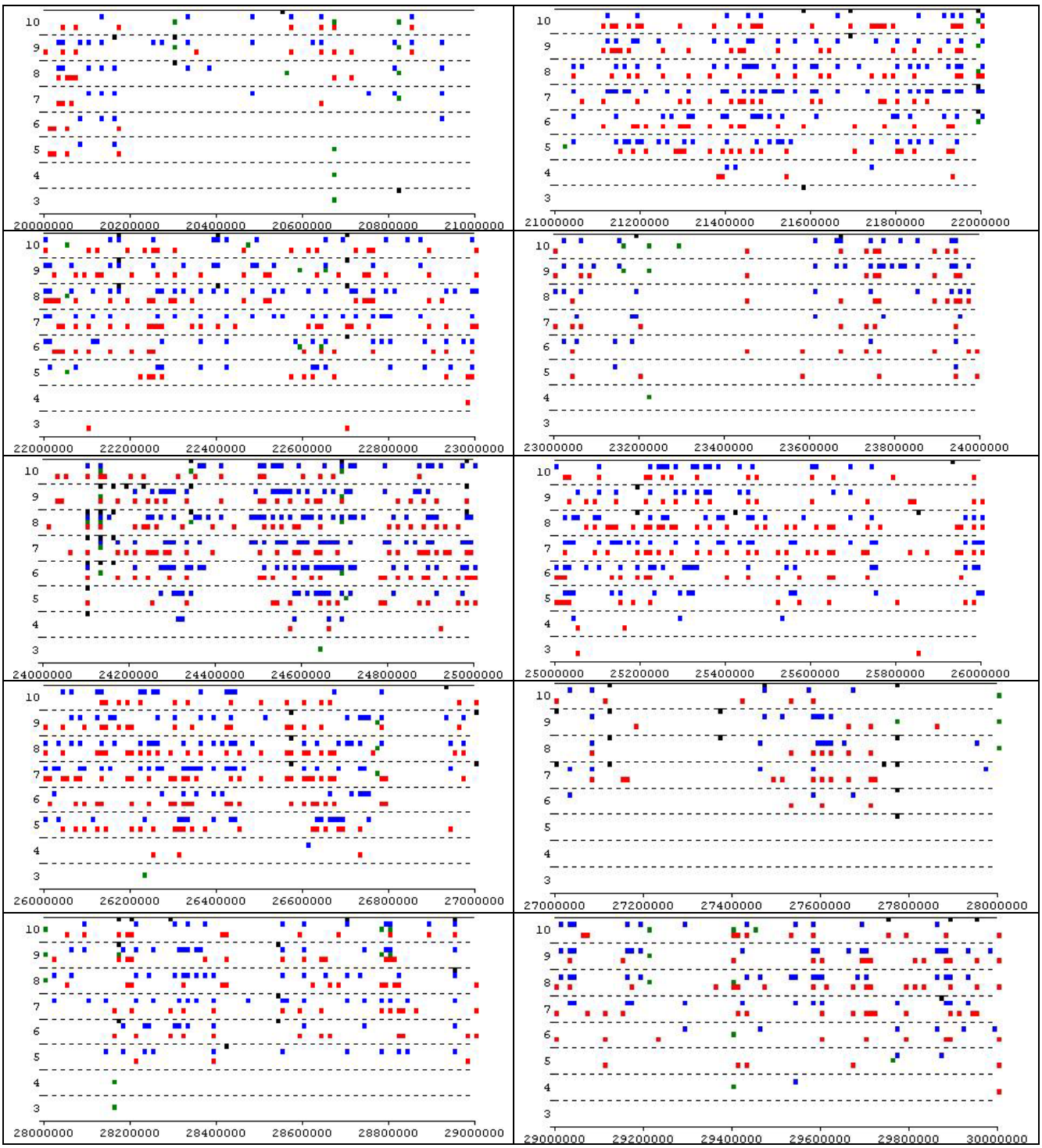


Рис. 3. Фрагменты со скрытой периодичностью полученные М.Ф.Ч. для участка $[2\times10^7, 3\times10^7]$ 22-ой хромосомы.

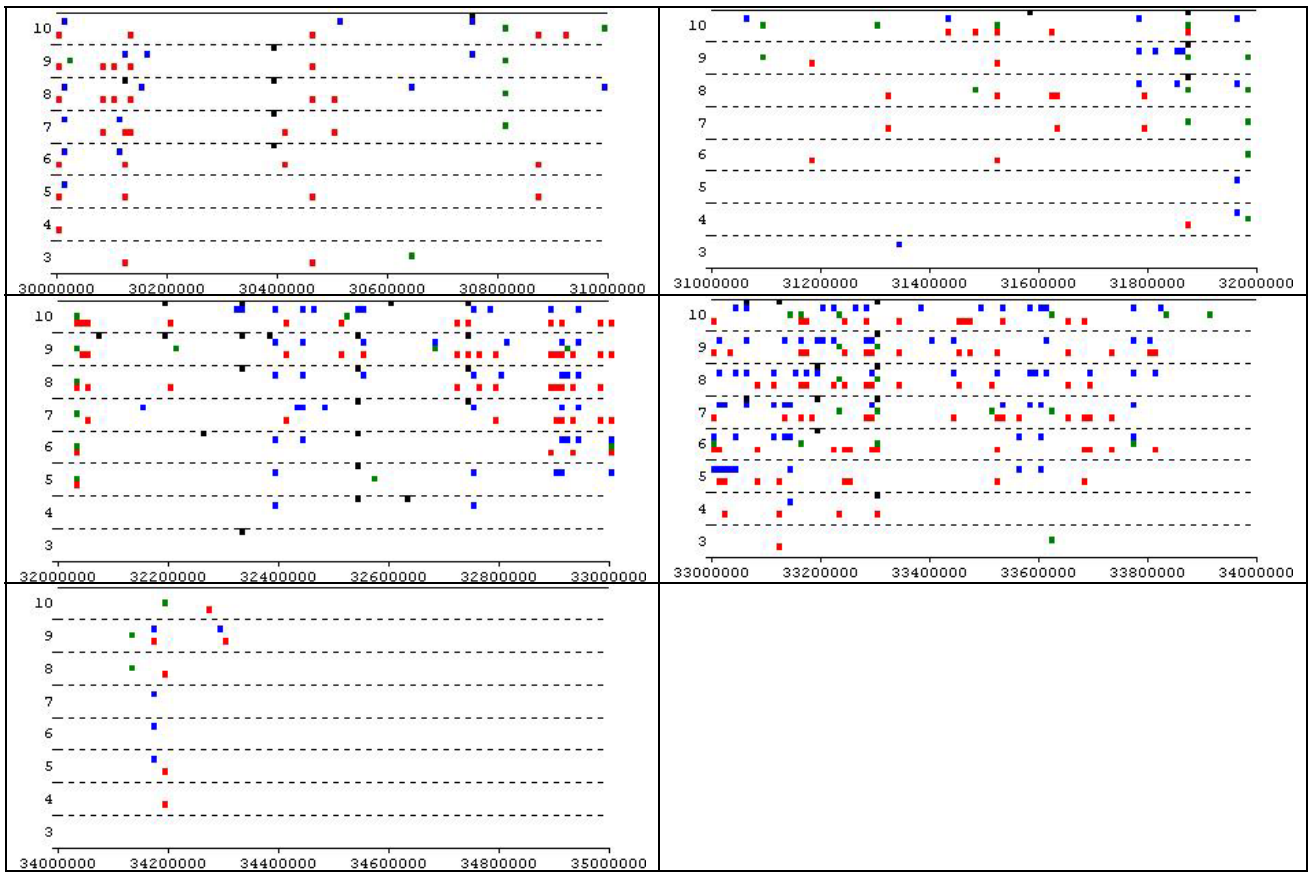


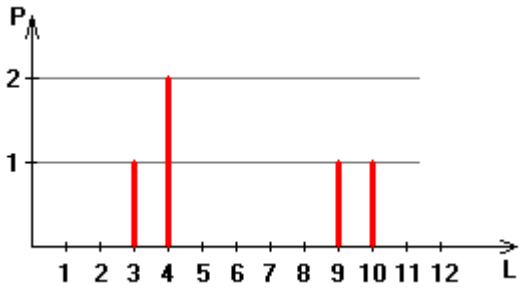
Рис. 4. Фрагменты со скрытой периодичностью полученные М.Ф.Ч. для участка [3×10^7, 3.4×10^7] 22-ой хромосомы.

Приложения.
Пример 1.

Пусть G={ a a a t t g g a c a t a a c a a t a a g t t c t g a a c a g} и n(a)=const=3*Разделив G[a,b] запятыми получим:*
G = { a a a, t t g g a c a t a, a c a a, t a a g t t c t g a, a c a g}
Получим следуюшие G[a,b]:
G[1,3]={a a a},
G[4,12]={t t g g a c a t a},
G[13,16]={ a c a a }
G[17,26]={ t a a g t t c t g a }
Видно, что эти последовательности имеют различную длину и равные частоты для символа ‘a’: n(a)=3.

Пример 2.
Пусть G={ a a a t t g g a c a t a a c a a t a a g t t c t g a a c a a c} и n(a)=const=3*Разделив G[a,b] запятыми получим:*
G[a,b] = { a a a, t t g g a c a t a, a c a a , t a a g t t c t g a, a c a a, c}
Получим следующие G[a,b] и L_k (k=1,2,3,4,5):
G[1,3]={a a a}, L_1 = 3;
G[4,12]={t t g g a c a t a}, L_2 = 9;
G[13,16]={ a c a a }, L_3 = 4;
G[17,26]={ t a a g t t c t g a }, L_4=10;
G[27,30]={a c a a}, L_5=4.
Видно, что эти последовательности имеют различную длину и равные частоты для символа ‘a’: n(a)=3.

Гистограмма для G[a,b] будет иметь вид:



Гистограмма для n(a)=3.

Пример 3:

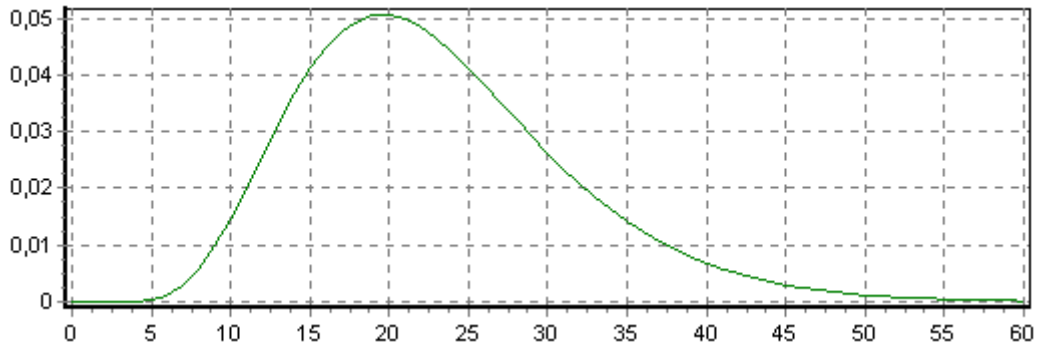


Рис 5. Плотность распределения для n(a) = 5 и случайной последовательности.

Литература

1. *Korotkov E. V., Korotkova M. A, Kudryashov N. A.* Information decomposition method to analyze symbolical sequence. Phys. Lett. A. (2003).
2. *Laskin A. A., Korotkov E. V., Kudryashov N. A.* Detecting of hidden periodicity in protein sequences and its correlation with structure and function of protein. Genome Informatics 12: 343-344 (2001).
3. *Korotkov E. V., Kudryashov N. A.,* Latent periodicity of many genes. Genome Informatics 12: 437-439 (2001).
4. *Dodin G., Vandergheynst P. and al.,* “Fourier and Wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences”, J. Theor. Biol. 206, 323-326 (2000).
5. *Gatherer D., McEwan N. R.* Analysis of sequence periodicity in E. coli proteins: Empirical Investigation of the ‘Duplication and Divergence’ Theory of protein evolution. J. Mol Evol 57: 149-158 (2003).
6. *Holste D., Grosse I., Herzel H.* Statistical analysis of the DNA sequence of human chromosome 22, Phys. Review E, vol. 64: 041917 (2001).
7. *Gatlin L. L.,* Information Theory and the Living System (Columbia University Press, New York, 1972).