МЕТОД ФИКСИРОВАННЫХ ЧАСТОТ ДЛЯ АНАЛИЗА НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДНК.

Федосеев С.В., Кудряшов Н.А., Коротков Е.В.
Московский Инженерно Физический Институт, г. Москва

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

В данной работе предлагается математический метод для поиска скрытой периодичности в последовательностях нуклеотидов ДНК. Известны, также методы информационного разложения [1], и Фурье [2]. Предлагаемый метод отличается от указанных тем, что позволяет учесть небольшое количество вставок и делеций символов в символьных последовательностях. Это достигается тем, что вводится понятие фиксированной частоты.

Мы применяли предлагаемый метод для анализа 22-ой хромосомы человека, и было показано, что в ней присутствуют фрагменты, обладающие свойствами скрытой периодичности. Причём эта периодичность не связана с кодирующими свойствами ДНК (исключение триплетной периодичности). Мы предполагаем, что данные фрагменты могут представлять “горячие” районы для генетической рекомбинации, где мы видим в последовательности последствия множественных генетических перестроек.

Метод.

Нуклеотидная последовательности G (символьная последовательность оснований ДНК) разбивается на фрагменты, таким образом, чтобы каждый их них содержал ровно заданное количество n символов одного типа - α. Фиксированной частотой по определению является n. Далее, при помощи критерия согласия Пирсона определяются отклонения спектра длин (L=L1…Lm) этих фрагментов, от триплетного спектра P(L), который строится таким образом, чтобы исключить влияние триплетной периодичности (см. исключение триплетной периодичности). По величине уровня доверия (в критерии согласия) можно судить о значительности отклонения спектров, и следовательно сделать вывод о присутствии периодичности в последовательности.

Результаты.

Мы применили метод для анализа скрытой периодичности для полной нуклеотидной последовательности 22-ой хромосомы человека (около 33 млн. символов). В результате этого были получены фрагменты этой последовательности с периодичным поведением по отдельности для каждого символа: α = (a,c,t и g) и фиксированной частоты в n=3..10.

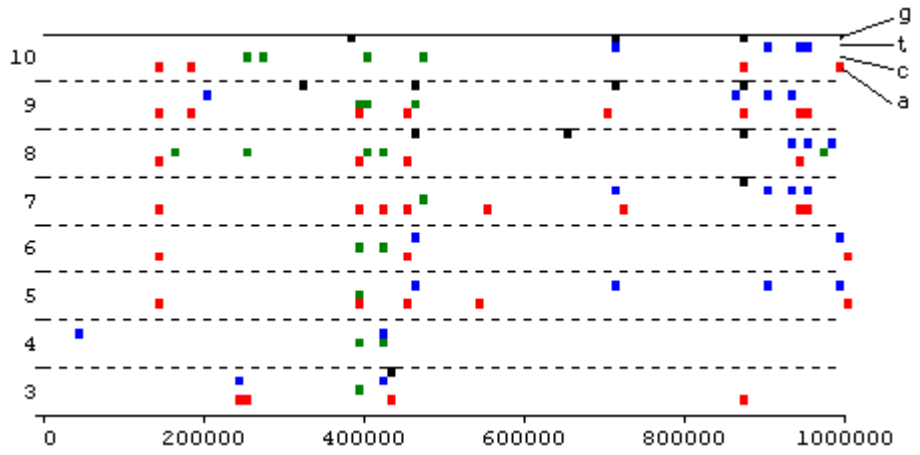

Рис. 1. Относительное расположение фрагментов со скрытой периодичностью для первого миллиона символов нуклеотидной последовательности ДНК. Горизонтальная ось – порядковый номер нуклеотида. Вертикальная – разбита на 7 участков (n=3…10), в каждом из которых периодичные фрагменты с длинной более 10 тыс. для 4-х символов (a,c,t,g).

Исключение триплетной периодичности.

Разобьем G на триплеты (тройки) символов и построим матрицу вероятностей:

$M = p\alpha k$, где $\alpha = \{a,c,t,g\}$, $k=1,2,3$.
$p\alpha k$ – вероятность обнаружить символ α на позиции k.

Методом Монте-Карло построим несколько символьных последовательностей, с такой же триплетной матрицей M. Усреднив спектры длин этих последовательностей, и получим спектр P(L).

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).