УДК 576.12

И. А. Шахмурадов, В. А. Гасумов

КОМПЬЮТЕРНЫЙ МЕТОД ПОИСКА В НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ УЧАСТКОВ ГОМОЛОГИИ С ВОЗМОЖНЫМИ ВСТАВКАМИ/ДЕЛЕЦИЯМИ И ОЦЕНКА ИХ СТАТИСТИЧЕСКОЙ ЗНАЧИМОСТИ

Разработан компьютерный метод, позволяющий выявить все участки гомологии с заданными характеристиками (длина участка, число несовпадений, число и размер делеций/вставок) в одной нуклеотидной последовательности или между двумя последовательностями ДНК (РНК) и оценить статистическую значимость найденных гомологий. Этот метод особо эффективен для поиска потенциальных шпилечных структур в нуклеотидных последовательностях, а также может применяться при выравнивании двух последовательностей ДНК (РНК).

Введение. В последние годы в связи с нарастающим потоком информации по первичной структуре и функционированию биополимеров (ДНК, РНК, белки) встала проблема систематизации, анализа и обобщения этих данных для выявления особенностей структурно-функциональной организации и эволюции генетического аппарата. В этой связи особо важное значение имеет разработка эффективных компьютерных методов анализа различных структурно-функциональных и эволюционных характеристик последовательностей ДНК (РНК) и белков [1—8]. Основная проблема, возникающая при поиске разнообразных структурных особенностей реальных последовательностей (повторяющихся участков, функциональных сайтов и т. д.), состоит в оценке их статистической значимости. Так, разработан ряд компьютерных методов поиска участков гомологии — повторов различных классов внутри или между нуклеотидными последовательностями [3—6]. Наиболее универсальным в этом плане является метод контекстного анализа, разработанный Соловьевым и соавт. [6], позволяющий выявить различные типы повторяющихся фрагментов внутри или между нуклеотидными последовательностями и, что важно, оценить статистическую значимость найденных гомологий. Единственным недостатком указанного метода является то, что он не допускает наличия в сравниваемых фрагментах ДНК (РНК) возможных вставок/делеций, а последнее довольно часто встречается при сравнении различных последовательностей.

Нами разработан компьютерный метод, учитывающий это обстоятельство при поиске гомологичных участков внутри или между нуклеотидными последовательностями, т. е. определяющий уровень статистической значимости выявляемых гомологий. Ниже описывается этот метод.

Метод. Статистический критерий. Будем рассматривать повторяющиеся участки (l, m_1, m_2, k_1, k_2, k_3) нескольких типов (рис. 1). Здесь l — число совпадающих (комплементарных) нуклеотидов, m_1 и m_2 — число первых и последних обязательных совпадений, k_1 — число симметрических несовпадений, k_2 — общее число нуклеотидов во вставках / делециях по двум сравниваемым фрагментам, k_3 — общее число вставок / делеций по двум сравниваемым фрагментам. Очевидно, что число (k) всевозможных мест вставок / делеций в двух сравниваемых сегментах равно $l-m_1-m_2+1$ и $k_3 \leqslant k$. Естественно предположить, что m_1, $m_2 > 0$; k_1, k_2, $k_3 = 0, 1, \ldots$; $k_2 \ll l$; $k_3 \ll k$. Пусть исследуется нуклеотидная последовательность длины N, образованная нуклеотидами 4 типов (A, G, T/U, C) с соответствующими частотами их встречаемости p_1, p_2, p_3, p_4 и со случайным их расположением.

Вероятность совпадения двух неперекрывающихся (т. е. статистически независимых в случайной последовательности) участков длины $l+k_1$ в l позициях равна:

$$p(l, k) = C^{k_1}_{l+k_1} p^l (1-p)^{k_1}, \qquad (1)$$

где p для различных типов повторов вычисляется следующим образом:

$$p = \sum_{i=1}^{4} p_i^2 \tag{2}$$

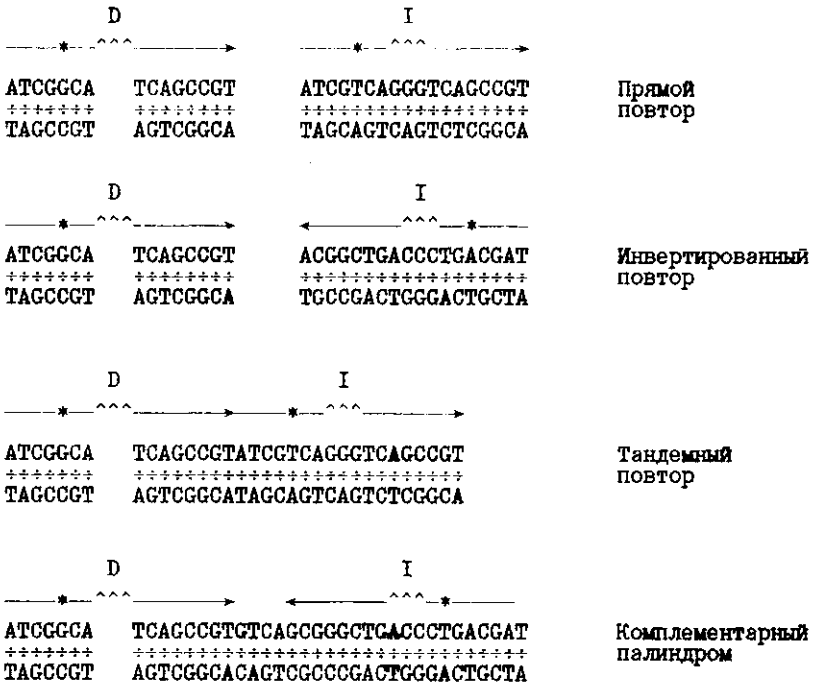
для прямых повторов в одной последовательности;

$$p = 2(p_1p_3 + p_2p_4) \tag{3}$$

для инвертированных повторов и комплементарных палиндромов в одной последовательности;

$$p = 2(p_1p_3 + p_2p_4 + p_2p_3) \tag{4}$$

для комплементарных палиндромов с учетом пар G—T(U).



Различные типы повторов. Вставка (I), делеция (D) и несовпадение (*) отмечены: $l=15$, $k_1=1$, $k_2=3$, $k_3=1$

Different types of repeats. The insertion (I), deletion (D) and distinction (*) are marked: $l=15$, $k_1=1$, $k_2=3$, $k_3=1$

Если сравниваются две последовательности длины N_1 и N_2 с частотами нуклеотидов p_{11}, p_{12}, p_{13}, p_{14} и p_{21}, p_{22}, p_{23}, p_{24} соответственно, то формулы (2) и (3) выглядят так:

$$p = \sum_{i=1}^{4} p_{1i}p_{2i}; \tag{2'}$$

$$p = p_{11}p_{23} + p_{13}p_{21} + p_{12}p_{24} + p_{14}p_{22}. \tag{3}$$

Вычислим число способов размещения k_3 вставок/инсерций, содержащих всего k_2 нуклеотидов, в пределах двух сравниваемых сегментов. Пусть в первом сегменте размещаем j_1 вставок (делеций во втором), содержащих всего i_1 нуклеотидов, а во втором сегменте — $j_2=k_3-j_1$ вставок (делеций в первом) из $i_2=k_2-i_1$ нуклеотидов, причем для i_1, i_2, j_1 и j_2 выполняются следующие условия:

$$0 \leqslant j_1 \leqslant i_1 \leqslant k_2, \quad j_1 \leqslant k_3, \quad 0 \leqslant j_2 \leqslant i_2 \leqslant k_2, \quad j_2 \leqslant k_3, \quad i_1 + i_2 = k_3, \; j_1 + j_2 = k_2; \tag{5}$$

i_1 и i_2 ($i_1 > 0$, $i_2 > 0$) нуклеотидов можно разместить по j_1 и j_2 ($j_1 > 0$, $j_2 > 0$) вставкам $C_{i_1-1}^{j_1-1}$ и $C_{i_2-1}^{j_2-1}$ способом соответственно. Кроме того, j_1 и j_2 ($j_1 > 0$, $j_2 > 0$) вставок

можно разместить по k допустимым местам $C_k^{j_1}$ и $C_{k-j_1}^{j_2}$ способом соответственно. Если же все k_2 нуклеотидов разместить по k_3 вставкам только в одном из двух сравниваемых сегментов, то число всевозможных расположений вставок будет $C_k^{k_3}$. Таким образом, общее число (φ_1) размещений k_3 вставок, содержащих всего k_2 нуклеотидов, по двум сегментам равно:

$$\varphi_1 = \sum_{i_1=0}^{k_2} \sum_{j_1=0}^{k_3} Z(i_1, j_1) C_k^{j_1} \sum_{i_2=0}^{k_2-i_1} \sum_{j_2=0}^{k_3-j_1} Z(i_2, j_2) C_{k-j_1}^{j_2}, \quad (6)$$

где i_1, i_2, j_1, j_2 выполняют условия (5) и

$$Z(i, j) = \begin{cases} 1, \text{ если } j = 0,1;\ i \geqslant j; \\ C_{i-1}^{j-1}, \text{ если } i \geqslant j > 1. \end{cases} \quad (7)$$

Наконец, определим число (φ_2) всевозможных размещений двух сравниваемых сегментов по одной или двум исследуемым последовательностям. В случае поиска прямых и инвертированных повторов в пределах одной последовательности

$$\varphi_2 = (N - 2l - k_2 + 1)(N - 2l - k_2 + 2)/2. \quad (8)$$

Для тандемных прямых повторов и комплементарных палиндромов

$$\varphi_2 = N - 2l - k_2 + 1. \quad (9)$$

В случае комплементарных палиндромов, образующих предполагаемую шпилечную структуру, минимальное расстояние между двумя инвертированными повторами, входящими в комплементарный палиндром, должно быть, по физико-химическим соображениям, не менее трех нуклеотидов. Поэтому, если d_1 и d_2 являются заданным минимальным и максимальным расстоянием соответственно между двумя инвертированными повторами палиндрома, то в этом случае

$$\varphi_2 = (d_2 - d_1 + 1)(2N - 4l - 2k_3 - 3d_2 + 3d_1), \quad (10)$$

где $d_2 \leqslant N - 2l - k_3$.

В случае сравнения двух разных последовательностей максимальное число ($\varphi_2^{\max}$) всевозможных размещений двух сегментов по этим последовательностям равно

$$\varphi_2^{\max} = (N - 2l - k_3 + 1)(N - 2l + 1), \text{ если } N_1 \leqslant N_2; \quad (11)$$

$$\varphi_2^{\max} = (N - 2l + 1)(N - 2l - k_3 + 1), \text{ если } N_1 > N_2. \quad (11')$$

Таким образом, общее число (φ) возможных пар повторяющихся участков (l, m_1, m_2, k_1, k_2, k_3) по одной или двум нуклеотидным последовательностям с учетом всех различных размещений допускаемых вставок / делеций равно:

$$\varphi = \varphi_1 \varphi_2, \quad (12)$$

где φ_1 и φ_2 вычисляются с помощью формул (6) — (11), (11′).

Поэтому среднее ожидаемое число повторов (l, m_1, m_2, k_1, k_2, k_3) в случайной последовательности ($t_{\text{сп}}$) равно

$$t_{\text{сп}} = \varphi \cdot p(l, k_1), \quad (13)$$

где φ и $p(l, k_1)$ определяются по формулам (1) и (12) соответственно. Пусть среднее число повторов (l, m_1, m_2, k_1, k_2, k_3) в случайной последовательности близко к 1 (это означает отсутствие значительного числа таких участков в случайных последовательностях). Тогда можно применить биномиальное распределение для оценки вероятности ($P(t)$) наличия в этой последовательности t искомых повторов:

$$P(t) = C_\varphi^t p(l, k_1)^t (1 - p(l, k_1))^{\varphi - t}. \quad (14)$$

Верхней границей доверительного интервала с уровнем значимости Q (обычно $0{,}95 \leqslant Q < 1$) для превышения среднего числа найденных повторов над ожидаемым по случайным причинам будет такое t_0, что

$$\sum_{t=0}^{t_0-1} P(t) < Q \text{ и } \sum_{t=0}^{t_0} P(t) \geqslant Q. \quad (15)$$

Если число (t) выявленных повторов (l, m_1, m_2, k_1, k_2, k_3) в реальной последовательности равно или превышает верхнюю границу доверительного интервала (t_0), рассчитанную для случайной последовательности той же длины и с теми же частотами нуклеотидов, мы называем их неслучайными с уровнем значимости Q.

Алгоритм поиска повторов. Для сравнения двух участков анализируемой последовательности разработан подход, суть которого заключается в использовании матрицы точечной гомологии с применением модифицированного нами метода оптимального выравнивания Нидлмана и Вунша [2]. При этом из всех допустимых вариантов (с точки зрения параметров анализа) выбирается самый оптимальный (по критерию Нидлмана и Вунша). При таком подходе очень часто встречается ситуация, когда реально найденное число симметрических несовпадений, общее число нуклеотидов во вставках и общее число вставок в пределах сравниваемых фрагментов оказываются меньше изначально заданных чисел k_1, k_2 и k_3. Поэтому указанные параметры анализа фактически являются максимально допустимыми числами несовпадений при фиксированных параметрах l, m_1 и m_2 а также d_1 и d_2 в случае поиска шпилечных структур. По этой причине после полного просмотра анализируемой последовательности осуществляется группировка выявленных повторяющихся участков по параметрам l_1, m_1, m_2, k_1, k_2 и k_3 и оценка их статистической значимости по вышеизложенному критерию.

Описанный метод реализован в диалоговом и файловом (с заданием параметров анализа заранее с помощью некоторого файла) режимах на персональном компьютере типа IBM и является достаточно эффективным при исследовании длинных нуклеотидных последовательностей (нескольких тысяч нуклеотидов).

Обсуждение метода. Описанный метод позволяет выявить все участки гомологии с заданными характеристиками (число совпадающих нуклеотидов, число несовпадений, число и размер делеций/вставок) в одной нуклеотидной последовательности или между двумя последовательностями ДНК (РНК) и оценить статистическую значимость найденных гомологий. Этот метод особо эффективен при поиске элементов (т. е. потенциальных участков образования шпилек) вторичных структур нуклеотидных последовательностей, а также для выравнивания двух последовательностей ДНК (РНК), так как в этих случаях наличие вставок/делеций в сравниваемых последовательностях практически неизбежно.

METHOD OF COMPUTER SEARCH OF HOMOLOGOUS SITES WITH THE POSSIBLE INSERTIONS / DELETIONS IN THE NUCLEOTIDE SEQUENCES AND EVALUATION OF THEIR STATISTICAL SIGNIFICANCE

I. A. Shakhmuradov, V. A. Gasumov

Institute of Botany, Academy of Sciences
of the Azerbaijan SSR, Baku, 370073

Summary

A computer method which allows revealing all the homologous sites with the given characteristics (the site's length, number of distinctions, number and size of deletions / insertions) in one nucleotide sequence or between two DNA (RNA) sequences and evaluating the statistical significance of the found homologies was developed. This method is especially effective for the search of potential stem-loop structures in the nucleotide sequences and can be used to align the two DNA (RNA) sequences.

СПИСОК ЛИТЕРАТУРЫ

1. *Fitch W. M.* An improved method of testing for evolutionary homology // J. Mol. Biol.— 1966.— **16**, N 1.— P. 9—16.
2. *Needleman S. B., Wunsch C. D.* A general method applicable to the search for similarities in the amino acid sequences of two proteins // Ibid.— 1970.— **48**, N 3.— P. 443—453.
3. *Wachter R.* The number of repeats expected in random nucleic acid sequences and found in genes // J. Theor. Biol.— 1981.— **91**, N 1.— P. 71—98.
4. *Brezinski D. P.* Statistical significance of DNA sequence symmetries // Nature.— 1975.— 253.— P. 128—130.
5. *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.
6. *Колчанов Н. А., Соловьев В. В., Жарких А. А.* Высокая насыщенность прямыми повторами в генах РНК-полимераз по данным контекстного анализа // Докл. АН СССР.— 1983.— **273**, № 3.— С. 741—744.

7. *Use* of «Perseptron» algorithm to distinguish translational initiation sites in *E. coli* / G. D. Stormo, T. D. Scheider, L. Gold, A. Ehrenfucht // Nucl. Acids Res.— 1985.— 13, N 8.— P. 2997—3011.
8. *Энхансероподобные* структуры в умеренно повторяющихся последовательностях эукариотических геномов / И. А. Шахмурадов, Н. А. Колчанов, В. В. Соловьев, В. А. Ратнер // Генетика.— 1986.— 22, № 3.— С. 357—367.

Ин-т ботаники им. В. Л. Комарова АН АзССР, Баку Получено 10.05.90

УДК 576.315.42

В. В. Шматченко, А. Б. Бережнев

КАРТИРОВАНИЕ МЕСТ ПРИКРЕПЛЕНИЯ ДНК К ЯДЕРНОМУ СКЕЛЕТУ МЕТОДОМ ГРАФИЧЕСКОГО ПРЕДСТАВЛЕНИЯ ПРОТЯЖЕННЫХ НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Методом графического представления нуклеотидных последовательностей в виде кривых линий, отражающих распределение AT- и GC-оснований по длине последовательности, выявлены характерные (S-образные) профили кривых, соответствующие местам крепления ДНК к ядерному скелету эукариот. Впервые показана применимость использованного метода для картирования участков связывания ДНК со скелетными структурами ядра. Тем самым продемонстрирована возможность обнаружения функционально однотипных негомологичных участков ДНК, что не представляется возможным с помощью других известных компьютерных методов анализа протяженных нуклеотидных последовательностей.

Введение. К настоящему времени установлена первичная структура большого числа генов и прилегающих к ним некодирующих областей ДНК. Поток такой информации нарастает и требует осмысления. Анализ же структурно-функциональной организации генома сегодня несколько отстает от процесса накопления данных о первичных последовательностях. Поэтому определенную ценность представляет всякий свежий подход, проливающий свет на выяснение функциональной значимости тех или иных участков последовательностей.

Широкие возможности в этом плане открывают разнообразные компьютерные методы анализа последовательностей, позволяющие обрабатывать значительные массивы информации и представлять результаты в удобной для осмысления форме, например в виде графиков. Одним из перспективных подходов такого рода является метод графического представления нуклеотидных последовательностей, предложенный Хаморн [1], где для визуального анализа предлагается более детальная картина, чем та, которая получается при глобальном анализе содержания GC-оснований. В последнем случае фиксируется лишь суммарный уровень GC- и AT-оснований, имеющий биологический смысл (см., например, [2]). Кроме того, данный метод в нашем исследовании позволяет судить о функциональном подобии негомологичных последовательностей ДНК путем установления схожести профилей получаемых с его помощью кривых линий, являясь своего рода графическим аналогом метода выравнивания. От выравнивания метод построения профилей нуклеотидных последовательностей отличает возможность эффективного сопоставления протяженных участков последовательностей, не обладающих к тому же близостью первичных структур, что бывает актуально в ряде случаев. В частности, при анализе хромосомной ДНК на предмет наличия мест прикрепления к ядерному скелету нами был использован данный метод как наиболее адекватный решаемой проблеме.