5. *Vogelstein B., Pardoll D. M., Coffey D. S.* Supercoiled loops and eukaryotic DNA replication // Ibid.— 1980.— **22**, N 1.— P. 79—85.
6. *Georgiev G. P., Nedospasov S. A., Bakayev V. V.* Supranucleosomal levels of chromatin organization // The Cell Nucleus.— 1978.— **6**, N 1.— P. 3—34.
7. *Mirkovitch J., Mirault M.-E., Laemmli U. K.* Organization of the higher-order chromatin loop: Specific DNA attachment site on nuclear scaffold // Cell.— 1984.— **39**, N 1.— P. 223—232.
8. *Cockerill P. N., Garrard W. T. C.* Chromosomal loop anchorage of the kappa immunoglobulin gene occurs next to the enhancer in a region containing topoisomerase II sites // Ibid.— 1986.— **44**, N 2.— P. 273—282.
9. *Gasser S. M., Laemmli U. K.* Cohabitation of scaffolding binding regions with upstream / enhancer elements of three developmentally regulated genes of *D. melanogaster* // Ibid.— **46**, N 3.— P. 521—530.
10. *Dijkwel P. A., Hamlin J. L.* Matrix attachment region are positioned near replication initiation sites, genes, and an interamlicon junction in the amplified dihydrofolate reductase domain of chinese hamster ovary cells // Mol. and Cell. Biol.— 1988.— **8**, N 12.— P. 5398—5409.
11. *Hamori E., Varga G.* DNA sequence (H) curves of the human immunodeficiency virus I and some related viral genomes // DNA.— 1988.— **7**, N 5.— P. 371—378.
12. *Cockerill P. N., Yuen M.-H., Garrard W. T.* The enhancer of the immunoglobulin heavy chain locus is flanked by presumptive chromosomal loop anchorage elements // J. Biol. Chem.— 1987.— **262**, N 11.— P. 5394—5397.
13. *Kas E., Chasin L. A.* Anchorage of the chinese hamster dihydrofolate reductase gene to the nuclear scaffold occurs in an intragenic region // J. Mol. Biol.— 1987.— **198**, N 4.— P. 677—692.
14. *Greenstein R. J.* Constitutive attachment of murine erythroleukemia cell histone depleted DNA loops to nuclear scaffolding is found in the β-major but not the α1-globin gene // DNA.— 1988.— **7**, N 9.— P. 601—607.
15. *Berrios M., Osheroff N., Fisher P. A.* In situ localization of DNA topoisomerase II, a major polypeptide component of the *Drosophila* nuclear matrix fraction // Proc. Nat. Acad. Sci. USA.— 1985.— **82**, N 13.— P. 4142—4146.
16. *Metaphase* chromosome structure. Involvement of topoisomerase II / S. M. Gasser, T. Laroche, J. Falquet et al. // J. Mol. Biol.— 1986.— **188**, N 3.— P. 613—629.
17. *Max E. E., Maizel J. V., Jr., Leder P.* The nucleotide sequence of a 5,5-kilobase DNA segment containing the mouse κ immunoglobulin *J* and *C* region genes // J. Biol. Chem.— 1981.— **256**, N 10.— P. 5116—5120.
18. *Emorine L., Max E. E.* Structural analysis of a rabbit immunoglobulin κ2 *J-C* locus reveals multiple deletions // Nucl. Acids Res.— 1983.— **11**, N 24.— P. 8877—8890.
19. *Sander M., Hsieh T.-S.* Double strand DNA cleavage by type II DNA topoisomerase from *Drosophila melanogaster* // J. Biol. Chem.— 1983.— **258**, N 15.— P. 8421—8428.

Ин-т биохимии и физиологии микроорганизмов АН СССР, Пущино

Получено 11.03.90

УДК 576.315.42

В. В. Волков, А. Ю. Леонтьев

ИССЛЕДОВАНИЕ СИММЕТРИИ ГЕНЕТИЧЕСКИХ ТЕКСТОВ МЕТОДОМ ФУРЬЕ-АНАЛИЗА

В работе предложен метод классификации симметричных структур одноцепочечной ДНК, использующий понятие цветной симметрии. Приводится систематическое перечисление цветных точечных и пространственных групп. Для поиска структур с инверсионной симметрией предлагается использовать фурье-анализ с фильтрацией, основанной на симметричной числовой кодировке нуклеотидной последовательности. Прямые повторы обнаруживаются с помощью фурье-анализа как первого этапа поиска, существенно сокращающего число операций сравнения. Методы опробованы на модельных последовательностях и областях инициации репликации прокариотического генома.

В настоящее время общеизвестно, что такие функциональные сайты, как точка начала репликации, сайты узнавания рестриктаз, сайты терминации транскрипции, энхансеры и другие являются симметричными структурами либо обогащены ими (см., например, [1—4]). Поэтому одним из этапов выявления функциональных сайтов в генетических текстах является поиск симметричных структур в молекуле ДНК.

Очевидно, что любые типы симметрии двухцепочечной ДНК в силу комплементарности обеих цепей закодированы в каждой из них, однако до последнего времени не существовало достаточно строгой классификации видов симметрии одноцепочечной ДНК. Так, в [5] приведены только три вида симметрии первичной структуры: прямой повтор, палиндром и комплементарный палиндром. Наиболее адекватный аппарат для описания симметрии одноцепочечной ДНК представляет собой теория так называемой цветной симметрии [6, 7], и первая часть настоящей работы посвящена описанию цветных групп симметрии генетических текстов.

Во второй части работы приведены данные по применению фурье-преобразования для поиска симметричных структур в одноцепочечной ДНК. Универсальным методом изучения симметрии веществ в различных агрегатных состояниях является анализ дифракции рентгеновских лучей, нейтронов, электронов. Эти методы позволяют по картине дифракции, представляющей собой фурье-образ изучаемого объекта, воссоздать его точечную и пространственную симметрию [6]. Перекодировав строку символов в алфавите $\alpha_0 = \{A, C, G, T\}$ в числовой ряд и подвергнув его фурье-преобразованию, мы фактически моделируем дифракционный эксперимент. В работе [8] применен метод функции Паттерсона для поиска гомологий в аминокислотных последовательностях, однако результаты такого поиска трудноинтерпретируемы. Ниже предложены различные варианты фурье-анализа при поиске различных видов симметрии.

Для анализа свойств симметрии одноцепочечной ДНК рассмотрим следующую ее модель. Со строкой символов в алфавите $\alpha = \{A, C, G, T, S, W, R, Y, M, K, B, D, H, V, N\}$ сопоставим одномерную систему равноотстоящих друг от друга точек, окрашенных в цвета, обозначаемые символами из α. В таком представлении комплементарный палиндром инвариантен по отношению к одновременному преобразованию пространственной координаты $x \rightarrow -x$ и изменению цвета по правилам $A \rightarrow T$, $C \rightarrow G$, $T \rightarrow A$, $G \rightarrow C$. Такое преобразование цветов в терминах теории цветной симметрии есть не что иное, как цветное отождествление, которое мы обозначим символом $1'$. Наряду с $1'$ можно ввести другое цветное отождествление, имеющее биологическое значение — $1''$ — преобразование пурина в пурин и пиримидина в пиримидин по правилам: $A \rightarrow G$, $C \rightarrow T$, $G \rightarrow A$, $T \rightarrow C$. Произведение этих двух цветных отождествлений есть новое цветное отождествление $1'''$, связывающее основания с амино- и кето-группами: $1'''A = 1''(1'A) = 1''T = C$. Три цветных преобразования представляют собой замкнутую систему в том плане, что образуют четвертую группу Клейна, и благодаря своему биологическому смыслу составляют основу для адекватного описания симметрии одноцепочечных молекул ДНК.

В одномерном пространстве возможны только две операции классической симметрии: тождественная — 1 и инверсия — $\bar{1}$. В сочетании с цветными операциями они образуют 16 точечных групп, из которых две (1 и $\bar{1}$) являются классическими, три $\bar{1}^i$ — цветными ($i = 1, 2, 3$ в соответствии с количеством штрихов у цветного отождествления), т. е. содержат цветное преобразование только вместе с пространственным, а остальные 11 групп являются нейтральными. Последнее означает, что в наборе операций симметрии, составляющих группу, имеется по крайней мере одно цветное отождествление. Из одиннадцати нейтральных групп пять соответствуют полносимметричным строкам N и не представляют интереса. Три группы с набором элементов $\{1, 1^i\}$ описывают симметрию произвольных строк в алфавитах {S, W, N}, {R, Y, N}, {M, K, N}, а другие три группы 1, 1^i, $\bar{1}^j$, 1^k описывают симметрию строк вида МККММК.

К функционально важным элементам первичной структуры ДНК относятся прямые повторы, которые на языке теории симметрии описываются операциями трансляции t: $x \rightarrow x + \Delta x$. Для рассматриваемой модели ДНК следует учесть как обычную, так и цветные трансляции,

т. е. одновременное преобразование координаты $x \to x + \Delta x$ и цвета в соответствии с одним из трех цветных отождествлений.

Сочетание классических и цветных точечных групп с обычной и цветными трансляциями дает 17 пространственных обычных и цветных групп t, t^i, $t : \bar{1}$, $t^i : \bar{1}$, $t ; 1^i$, $t^i : 1^i$, $t^i : \bar{1}^i$, из которых только для t надежно установлена функциональная значимость. Не исключено, что и другие симметрии играют существенную роль в функционировании генома, например, при взаимодействии нуклеиновой кислоты с белками, специфичными к пурин/пиримидиновому составу ДНК.

Дискретное фурье-преобразование есть разложение функции $\{u_i\} = \{x_i + jy_i\}$ в ряд $u_i = \sum_k v_k e^{-j\omega_k}$. Вычисление коэффициентов $v_k = x_k + jy_k$ представляет собой моделирование дифракционного эксперимента, а их набор несет информацию о симметрии функции u_i.

Таким образом, для поиска симметрии методом фурье-анализа ДНК необходимо представить числовым рядом. Поскольку цветные

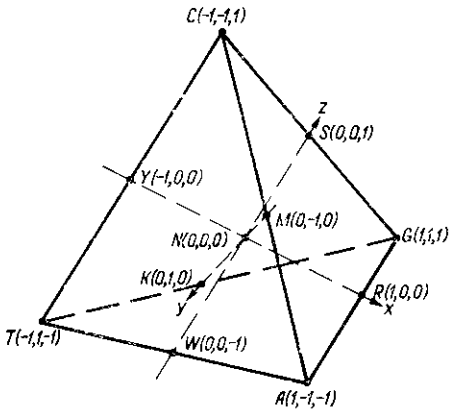

Рис. 1. Геометрическая интерпретация алфавита ДНК. Каждый нуклеотид представляется вектором в пространстве трех измерений, которым соответствуют следующие свойства: оси z — образовывать сильную или слабую комплементарную связь; оси x — быть пурином или пиримидином; оси y — иметь кето- или амино-группу. Центрам граней соответствуют символы *B*, *D*, *H*, *V*, а центру тетраэдра символ *N*

Fig. 1. A geometrical interpretation of the DNA alphabeth. Symbols correspond to the points in the three-dimensional space, the coordinational axes having the following sense β z-axis — an ability to take part in strong or weak complementary coupling, x-axis — a property to belong top urines or pyrimidines, y-axis — a property to have amino- or keto-group.

группы изоморфны обычным в пространстве с количеством измерений, увеличенным на число цветных отождествлений, каждому нуклеотиду естественно сопоставить вектор, трем проекциям которого отвечают свойства образовывать сильную (+I) и слабую (—I) комплементарную связь, быть пиримидином (+I) или пурином (—I), иметь кето- (+I) или амино-группы (—I). Определенные таким образом числовые представления нуклеотидов A, C, G, T соответствуют вершинам тетраэдра (рис. 1). Реальные последовательности представляют собой случайный или почти случайный набор символов из α_0 с вкрапленными в него регулярными структурами. Поэтому неупорядоченная часть молекулы создает шум в фурье-спектре, амплитуда которого может оказаться сравнимой или больше амплитуды полезного сигнала. Одним из способов выделения сигнала является построение кросс-спектров родственных последовательностей [9], однако даже для таких гомологичных сайтов, как точки начала репликации *Escherichia coli* и *Salmonella typhimurium*, распределение симметричных фрагментов неодинаково и при построении кросс-спектра информация об индивидуальных особенностях каждой из них теряется. В зависимости от типа симметрии, поиск которого проводится, можно использовать разные типы фильтрации.

Для обнаружения сайтов с симметрией $\bar{1}^i$ эффективным фильтром является предложенная выше кодировка символьной строки. Так, для сайта AGCT, анализируемого в произвольной числовой кодировке, в фурье-спектре присутствуют все четыре возможные амплитуды, в то время как в коде A = (I, —I), G = (I, I), C = (—I, I), T = (—I, —I) (проекция тетраэдра на плоскость, ортогональную оси y) спектр пред-

ставляет собой одну линию. Для каждого из фрагментов с симметрией $\bar{1}^i$ спектр Фурье будет выглядеть наиболее просто в описанной трехмерной кодировке, т. е. сигнал от такого фрагмента будет по крайней мере на $l/2$ превосходить шум (l — длина фрагмента).

Для поиска прямых повторов использовали кодировку A—(1,0), G—(2,0), C—(3,0), T—(4,0), N—(0,0). При выполнении фурье-преобразования последовательность дополняли нулями до длины 1 024 для лучшего разрешения сигналов. Спектр модулей $m_k^2 = x_k^2 + y_k^2$ преобразовывали в спектр периодов

$$T_l = \left(\sum_{i=1}^{N-1} m_{\mathrm{int}(N \cdot i/l+1)}\right)/(l+1),$$

где N — длина последовательности, l — длина периода. Для последующего анализа брали сигналы, превышающие средний шум на три

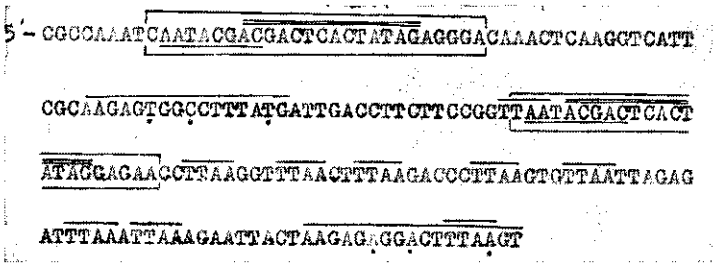

Рис. 2. Участок инициации репликации фага *T7*. Подчеркнуты и надчеркнуты повторы трех видов, обнаруженные с помощью предложенной методики, точками отмечены замены. Рамкой выделены промоторы РНК-полимеразы фага, практически совпадающие с найденными повторами

Fig. 2. The *T7* phage origin of replication with the repeats found by the suggested technique and indicated by the lines over and under the text. The promoters of the phage RNA-polymerase are shown by boxes.

стандартных отклонения и больше. Фрагменты найденной таким образом длины сравнивали между собой. Повторы длиной более семи нуклеотидов удавалось найти по присутствующим в спектре сигналам. Такая методика в отличие от представленной в [10] позволяет сократить до минимума число операций сравнения. На рис. 2 показана область инициации репликации фага *T7* с найденными повторами. Метод позволяет обнаруживать двухкратные совершенные повторы семи и более оснований в последовательности длиной 1 024. Для несовершенных повторов с заменами количество абсолютных совпадений также должно быть не менее семи при общей длине исследуемой последовательности 1 024. Если в гомологичных фрагментах имеются делеции, то такие повторы могут быть обнаружены при условии, что слева или справа от делеции имеется повторяющийся фрагмент длиной семь и более оснований.

AN INVESTIGATION OF THE SINGLE STRAND DNA SYMMETRY BY FOURIER TRANSFORMATION METHOD

V. V. Volkov, A. Yu. Leontyev

State University, Kazan

Summary

A system of the single strand DNA symmetry based on the theory of coloured symmetry is proposed. The search of the fragments with the inversion symmetry is carried out by means of the Fourier transformation, DNA being preliminary presented as a three-dimensional number series. For the search of the direct repeats the Fourier transformation is pro-

posed as the first step to decrease the number of comparisons. The both methods have been tested on the model sequences as well as on the prokaryotic origins of replication.

СПИСОК ЛИТЕРАТУРЫ

1. *Kornberg A.* Mechanisms of replication of the *Escherichia coli* genome // Eur. J. Biochem.— 1983.— **137**, N 3.— P. 377—382.
2. *Уотсон Дж., Туз Дж., Куртс Д.* Рекомбинантная ДНК.— М. : Мир, 1986.— 285 с.
3. *Pribnow D.* Biological regulation and development // Gene expression.— New York: Plenum press, 1979.— Vol. 1, N 4.— P. 219—277.
4. *The ccs* element, a 16 base pair palindrome essential for activity of the octopine synthase enhancer / J. G. Ellis, D. J. Llewellin, J. C. Walker et al. // EMBO J.— 1987.— **6**, N 11.— P. 3203—3208.
5. *Введение* в теорию генетических текстов / В. В. Соловьев, А. Э. Кель, И. В. Рогозин, Н. А. Колчанов.— Новосибирск: Изд-во Новосиб. ун-та, 1987.— 92 с.
6. *Вайнштейн Н. К.* Симметрия кристаллов. Современная кристаллография.— М. : Наука, 1979.— Т. 1.— 383 с.
7. *Заморзаев А. М., Галярский Э. И., Палистрант А. Ф.* Цветная симметрия, ее обобщения и приложения.— Кишинев : Штиинца, 1978.— 277 с.
8. *Pattern* recognition of sequence similarities in globular proteins by Fourier analysis. A novel approach to molecular evolution / A. M. Liquori, A. Ripamonti, C. Sadun et al. // J. Mol. Evol.— 1986.— **23**, N 1.— P. 80—87.
9. *Cosic I., Nesic D.* Prediction of not spots in SV40 enhancer and relation with experimental data // Eur. J. Biochem.— 1987.— **170**, N 1/2.— P. 247—252.
10. *Миронов А. А., Александров Н. Н.* Быстрый метод поиска гомологий нуклеотидных последовательностей // Биофизика.— 1988.— **33**, № 2.— С. 229—232.

Казан. гос. ун-т

Получено 10.05.90

УДК 577.112.5.0.87

П. В. Костецкий, И. В. Артемьев, О. И. Пожильцова, В. В. Ульяшин

ПАКЕТ ПРИКЛАДНЫХ ПРОГРАММ ДЛЯ АНАЛИЗА АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ НА ПЕРСОНАЛЬНОЙ МИКРО-ЭВМ «ИСКРА-226»

Приведено описание пакета программ для обработки аминокислотных последовательностей белков и пептидов. Входящие в пакет программы дают возможность: 1) ввода и редакции аминокислотных последовательностей; 2) распечатки выравненных аминокислотных последовательностей семейств гомологичных белков; 3) попарного сравнения гомологичных последовательностей; 4) расчета филогенетических деревьев гомологичных белковых последовательностей; 5) идентификации местоположения вариабельных и консервативных участков в аминокислотных последовательностях гомологичных белков; 6) поиска сходства двух сравниваемых белков с распечаткой точечной матрицы сравнения; 7) предсказания локализации B- и T-клеточных антигенных детерминант; 8) перенесения аминокислотных последовательностей банка белковых последовательностей PIR, хранящихся на магнитных лентах, на гибкий магнитный диск персональной ЭВМ «Искра-226» и обратно; 9) идентификации фрагментов структурного гена с помощью набора пептидов; 10) трансляции структурных генов, содержащих интроны. Пакет программ написан на алгоритмическом языке Бейсик. С целью ускорения работы отдельных наиболее трудоемких блоков программ использован автокод микро-ЭВМ. Программы реализованы в диалоговом режиме.

Введение. Расшифровка аминокислотных последовательностей белков и пептидов в настоящее время стала рутинной работой, и уже накоплено много данных о первичных структурах белков и пептидов. В организованном в США банке PIR (Protein Identification Resource) сейчас находится более 1,5 млн аминокислотных остатков (а. к. о) [1]. С накоплением знаний о первичной структуре белков и их функциональных особенностях закономерно встает вопрос о взаимосвязи биологической активности белков и их строения. Для исследования свойств белков и пептидов с использованием ЭВМ написан ряд пакетов про-