УДК 57.01+577.4

ИССЛЕДОВАНИЕ ИДЕНТИФИКАЦИИ ИНФОРМАЦИОННЫХ КОМПОНЕНТ ГЕНОВ МЕТОДОМ СПЕКТРАЛЬНОГО И ВРЕМЕННОГО АНАЛИЗА

© 2008 Л.С. Бекасов, А.А. Тверетин, Ф.Ф. Буканов[1]

Статья посвящена проблеме распознавания структурной организации информационной части генов Mef2A, которая, как известно, состоит из перетасованных (чередующихся случайным образом) рабочих (экзонов) и "дремлющих" (интронов) подразделов [1]. В частности, проанализированы три нуклеотидные последовательности, кодирующие одинаковые белки трех разных организмов. Приведены результаты теоретических исследований.

Ключевые слова: *ген Mef2A, информационная компонента, методод спектрального анализа, кодирование.*

Введение

В общем случае генетический текст (далее текст) может быть сопоставлен с адекватным сложным временным сигналом с большой зашумленностью, который, кроме того, модулирован по частоте и амплитуде.

Еще одна особенность определяется тем, что структура текста не является однозначной из-за вырожденности кодонов по отношению к аминокислотам и мутационной динамике [2]. Более того, экзоны и интроны по своему объему имеют отличия на порядок и "перетасованы" между собой произвольным образом [3]. Между тем эти участки постоянны по своей структуре, но полностью не идентичны, в том числе и, возможно, имеют функциональные отклонения за счет большого количества шума.

Стремительное увеличение количества проектов по секвенированию геномов (иначе представлению их нуклеотидными последовательностями) человека, животных, растений, бактерий и вирусов привело к лавинообразному росту объема информации о таких последовательностях. Их анализ,

[1]Бекасов Лев Степанович, Тверетин Алексей Александрович, Буканов Федор Федорович, кафедра электронных систем и информационной безопасности Самарского государственного технического университета

обобщение и накопление знаний о структуре и функции генетических молекул относятся к числу наиболее важных проблем молекулярной генетики. Одним из подходов к решению этих проблем является функциональная аннотация новых генов с помощью компьютерных программ на основе анализа последовательностей ДНК и экспериментальной информации, накопленной в базах данных.

1. Методы и средства

Как известно, классическая теория спектрального анализа сигналов базируется на использовании комплексной экспоненциальной системы базисных функций:

$$\{e^{ik\omega_0 t}\} = \{..., e^{-i2\omega_0 t}, e^{-i\omega_0 t}, 1, e^{i\omega_0 t}, e^{i2\omega_0 t}, ...\}, \tag{1}$$

где $i = \sqrt{-1}$; k — номер базисной функции; ω_0 — угловая частота [4]. Если сигнал рассматривать как периодический несинусоидальный сигнал, отвечающий условиям Дирихле, то такой сигнал можно представить рядом Фурье как суперпозицию конечного или бесконечного числа базисных функций вида (1). Тогда спектральный состав разложения может быть охарактеризован дискретной спектральной функцией

$$S(k\omega_0) = \int_{-T/2}^{T/2} f(t)e^{ik\omega_0 t}dt. \tag{2}$$

Одним из замечательных свойств преобразования Фурье в экспоненциальном базисе является свойство инвариантности амплитудно-частотного спектра

$$S(k\omega_0) = |S(k\omega_0)| \tag{3}$$

к сдвигам сигнала $f(t)$, благодаря которому значительно упрощается проблема сопоставления различных спектров.

В то же время необходимо отметить, что при получении спектральных признаков с помощью экспоненциального базиса (2) теряется в явном виде структурная информация о сигнале, что только отрицательным образом может отразиться на вероятности правильного принятия решения [5]. Действительно, согласно (2), для каждого значения k фактически вычисляется значение взаимной энергии между сигналом $f(t)$ и k-й базисной функцией. При большом k анализ полученного спектра осложняется, особенно если данные имеют нечеткий характер, а при отбрасывании части гармоник теряется информация о сигнале. Очевидно, что в случае с нечеткими данными необходимо сжатие, причем оно должно учитывать самые небольшие изменения сигнала, что не выполняется при использовании (2).

Поэтому возникает задача получения такого спектра, который бы, с одной стороны, содержал частотную информацию, причем амплитудно-частотный спектр был бы инвариантен к сдвигам анализируемого сигна-

ла, а, с другой стороны, в явном виде содержал информацию о структуре анализируемого сигнала, а также для формирования которого требовалось бы минимальное время.

Для решения перечисленных проблем в [6] была предложена базисная комплексная система импульсных функций, с помощью которой можно получить спектр, отвечающий указанным требованиям. Предложенная система функций определяется на дискретном множестве

$$M = \{l : l = 0, 1, 2, ..., 2^n - 1\} \tag{4}$$

и имеет вид

$$Bah_u^k l = \overset{\wedge k}{c}_u(l) - i\overset{\wedge k}{s}_u(l), \tag{5}$$

где $u = 0, 1, 2, 3, \ldots, n-1$; 2^n — число подынтервалов, составляющих период некоторого подлежащего анализу дискретного сигнала $f(l)$.

Функции $\overset{\wedge k}{c}_u(l)$ и $\overset{\wedge k}{s}_u(l)$ формируются на основе вспомогательных функций $\overset{k}{c}_u(l)$ и $\overset{k}{s}_u(l)$ посредством их сдвигов на k подынтервалов, где $k = q, \ldots, 2^{n-u-1}$, q — позиция первого подынтервала.

Функции $\overset{k}{c}_u(l)$ и $\overset{k}{s}_u(l)$ определяются как:

$$c_0(l) = 1, \qquad s_0(l) = 0, \qquad l \in M. \tag{6}$$

В случае $u \neq 0$ и l, изменяющегося от 0 до 2^n с шагом 2^{n-u-1},

$$c_u(l) = \sum_{m=0}^{2^n-1} (\cos(2^{u-n}\pi m))e(l-m), \tag{7}$$

$$s_u(l) = \sum_{m=0}^{2^n-1} (\sin(2^{u-n}\pi m))e(l-m). \tag{8}$$

Если l принимает другие значения, то $c_u(l) = s_u(l) = 0$. Величина $e(l-m)$ представляет собой единичный импульс, определяемый из следующих условий:

$$e(l-m) = \begin{cases} 1, & l = m; \\ 0, & l \neq m. \end{cases} \tag{9}$$

Формирование амплитудно-частотного спектра анализируемого сигнала $f(l)$ осуществляется в соответствии с выражением

$$F_u = \sum_{k=q}^{2^{n-u-1}-1} F_u^k, \tag{10}$$

где $u = 0, 1, 2, 3, \ldots, n-1$, q — позиция первого подынтервала;

$$F_u^k = \sqrt{(a^k)^2 - (b^k)^2}; \quad a_u^k = \sum_{m=0}^{2^{u-1}-1} f(l_m)\overset{\wedge k}{c}_u(l); \quad b_u^k = \sum_{m=0}^{2^{u-1}-1} f(l_m)\overset{\wedge k}{s}_u(l); \tag{11}$$

$f(l_m)$ — значение анализируемого сигнала в точке l_m, где $l_m = 2^{n-u-1}m$.

Для получения структурной информации об анализируемом сигнале можно воспользоваться функциями $\overset{\wedge k}{c_u}(l)$ и $\overset{\wedge k}{s_u}(l)$, с помощью которых фактически генерируется последовательность единичных импульсов, сдвинутых относительно друг друга на один шаг дискретизации.

Этот метод авторы использовали применительно к генетическим текстам, заимствованным на странице GenBank (`http://www.ncbi.nlm.nih.gov`). Индексы базы GenBank соответственно AJ010072, U30823, DQ323505. Последовательности проанализированы с позиции кодона начала трансляции "ATG". Проанализированы нуклеотидные последовательности, кодирующие ген Mef2a трех организмов: "Gallus gallus", "Mus musculus domesticus", "Rattus norvegicus".

Ген под названием "Mef2a" играет роль в защите стенок артерии от появления закупорок, которые препятствуют притоку крови и вызывают сердечные приступы. Мутация этого гена вызывает многие болезни сердца.

Для применения метода представления данных с использованием спектрального анализа на основе комплексной системы импульсных функций требуется, чтобы данные имели числовой характер. Каждой букве генетического текста поставлено в соответствие весовое значение, определенное с помощью молекулярного веса [4].

Пусть

$$P_i = \begin{cases} 0, & X_i = C; \\ 1, & X_i = A; \\ 2, & X_i = T; \\ 3, & X_i = G, \end{cases} \tag{12}$$

где X_i — i-й нуклеотид в последовательности.

2. Результаты и их обсуждение

Проанализированы полученные последовательности нуклеотидов, кодирующие ген Mef2a, для трех организмов: "Gallus gallus", "Mus musculus domesticus", "Rattus norvegicus". Все три гена имеют разную длину, поэтому возникает вопрос выбора количества интервалов n. Предложено анализировать ген методом скользящего окна, где $n = 5$. Понятно, что с ростом n информация о структуре сигнала будет уменьшаться. Нахождение оптимальной величины n не является целью данной статьи и будет рассмотрено в дальнейших исследованиях.

Для всех трех последовательностей найдены числовые эквиваленты в соответствии с (12), где $i = 1..1487$. Выбор диапазона i обусловлен количеством нуклеотидов в самой короткой последовательности (Rattus norvegicus). Далее, найдены семейства $F_u^i, U = 0..4$ в соответствии с (10) для каждой позиции скользящего окна $q_i = (i-1) \cdot 2^n + 1$, где i — номер окна.

Проведено сравнение значений F_u^i для каждого i между генами всех

трех организмов попарно. Получены коэффициенты корреляции Пирсона K_i [7]. На рис. 1 по оси ординат представлены значения K_i между значениями F_u^i в соответствии с формулой (10).

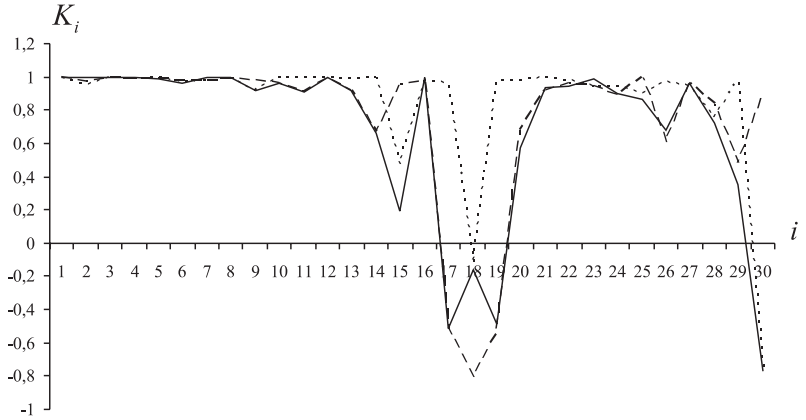

Рис. 1. Коэффициент корреляции Пирсона K_i, $i = 1..30$ между значениями F_u^i для гена Mef2a (--- Gallus gallus / Mus musculus domesticus, — Mus musculus domesticus / Rattus norvegicus, –– Gallus gallus / Rattus norvegicus)

Результаты показывают, что гены у всех трех организмов имеют ярко выраженные участки, на которых K_i стремится к единице, например, в диапазонах $i = 1..12$ и $i = 22..24$. Для оценки эффективности на рис. 2 приведены значения P_j, $j = 65..96$ в соответствии с (12). Из рис. 2 видно, что высокие значения K_i (0.997; 0.999; 0.995), где $i = 3$, соответствуют высоким коэффициентам корреляции K_i^P (0,733; 0,985; 0,738), где $i = 3$, между соответствующими значениями P_j, $j = 65..96$.

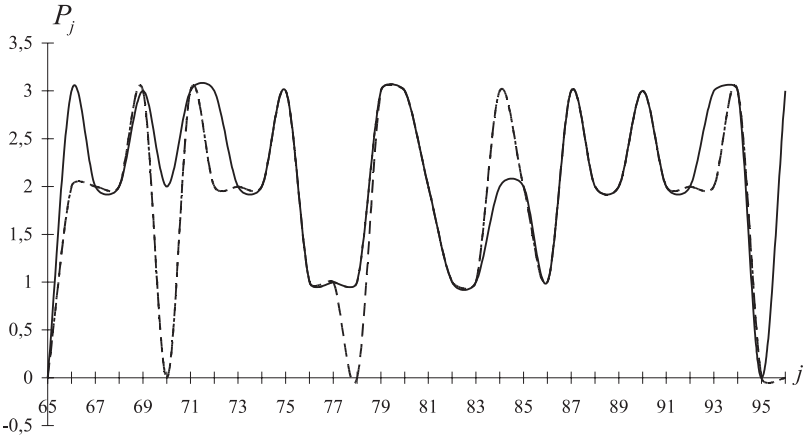

Рис. 2. Функция P_j, $j = 65..96$ для гена Mef2a (--- Gallus gallus, — Mus musculus domesticus, –– Rattus norvegicus)

Выводы

1) Применение метода представления данных с использованием спектрального анализа на основе комплексной системы импульсных функций позволяет получить амплитудно-частотный спектр конкретного генетического текста.
2) Применение скользящего окна с малым значением n позволяет выделить участки генетических текстов с разной степенью схожести, выраженной значением K_i.
3) Примененный метод чувствителен к многочисленным сдвигам внутри нуклеотидной последовательности и позволяет получить описание структуры сигнала даже при его большой зашумленности.
4) Применение данного метода позволяет выделять образы, которые довольно точно описывают функциональную принадлежность нуклеотидной последовательности.

Литература

[1] Франк-Каменецкий, М.Д. Компьютерный анализ генетических текстов / М.Д. Франк-Каменецкий. – М.: Наука, 1990. – 267 с.

[2] Сингер, М. Гены и геномы / М. Сингер, П. Берг. – М.: Мир, 1998. – 373 с.

[3] Писарчик, А.В. Простые повторяющиеся последовательности и экспрессия генов / А.В. Писарчик, Н.А. Картель // Молекулярная биология. – №34(3). – С. 357–362.

[4] Кристалинский, Р.Е. Преобразование Фурье и Лапласа в системах компьютерной математики: учебн. пособие для вузов / Р.Е. Кристалинский. – М.: Горячая линия – Телеком, 2005. – 216 с.

[5] Трахтман, А.М. Введение в обобщенную спектральную теорию / А.М. Трахтман. – М.: Сов. радио, 1972. – 352 с.

[6] Bahrushina, G.I. Development and Investigation of a New Retangular Orthogonal System Function for Invariant Object Recognition / G.I. Bahrushina, A.P. Bahrushin // Proceedings of the Sixth International Conference "Advanced Computer Systems"/ Szezecin–Poland. November 1999. – P. 64–67.

[7] Курникова, Е.Л. Основы статистики / Е.Л. Курникова, Л.В. Тарлецкая. – М.: МГИМО, 2008. – 144 с.

Поступила в редакцию 15/*II*/2008;
в окончательном варианте — 15/*II*/2008.

THE RESEARCH OF IDENTIFICATION GENE COMPONENTS INFORMATION BY SPECTRAL AND TIME ANALYSIS METHODS

© 2008 Bekasov L.S., Tveretin A.A., Bukanov F.F.[2]

The subject of this paper is structure organization recognizing problem of Mef2A gene information part, which, as known, consists of random mixed exon and intron segments [1]. Specifically, three nucleotide sequences, which encodes the proteins of three organisms was analyzed. The results of theoretical research were produced.

Keywords: *Mef2A gene, information component information, spectral analysis method, coding.*

Paper received 15/*II*/2008.
Paper accepted 15/*II*/2008.

[2]Bekasov Lev Stepanovich, Tveretin Alexey Alexsandrovich, Bukanov Fedor Fedorovich, Dept. of Electronic Systems and Informaton Security, Samara State Technical University, Samara, 443010, Russia.