УДК 57.01+577.4

СРАВНЕНИЕ КОНКАТЕНИРОВАННЫХ ДАННЫХ НА ОСНОВЕ ИХ СПЕКТРАЛЬНЫХ ХАРАКТЕРИСТИК

Тверетин А.А., Бекасов Л.С.

Самарский государственный технический университет, Самара, Россия

Подробная информация об авторах размещена на сайте «Учёные России» - http://www.famous-scientists.ru

Статья посвящена актуальной проблеме сравнения конкатенированных данных с целью создания образов для распознавания конкретных ситуаций. Типичным примером таких данных являются генетические тексты. Предложен метод преобразования данных использованием спектрального анализа на основе комплексной системы импульсных функций. Произведено сравнение спектральных характеристик для отдельно взятых трёх автономных текстов на основе такого представления.

Введение

Стремительное увеличение количества проектов по секвенированию геномов человека, животных, растений, бактерий и вирусов привело к лавинообразному росту объема информации о нуклеотидных последовательностях. Их анализ, обобщение и накопление знаний о структуре и функции генетических молекул относятся к числу наиболее важных проблем молекулярной генетики.

Системы хранения данных в живой природе построены по иным принципам, чем устройства подобного назначения технического характера [2]. Технологии использования этих данных весьма сложны, поскольку они предназначены в целом для воспроизводства (создания) себе подобных существ во всех представляемых смыслах толкования этих терминов. Этим объясняется большой интерес со стороны учёных всего мира, работающих в различных предметных областях. Успехи биологов выражаются в том, что большинство сверхсложных систем хранения данных (молекул ДНК, геномов и т.д.) секвенировано, что привело к лавинообразному росту объёма информации.

В общем случае, текст, записанный при помощи некоторого множества символов, семантика которого неизвестна представляется как конкатенированная последовательность элементов четырех разновидностей. Сложность такого текста состоит в том, что структура последовательности является непостоянной, участки произвольной длины в нем смещаются относительно друг друга случайным образом и, что семантика подтекстов (не исключено их наложение друг на друга) также неизвестна. Между тем известно [2], что конституция этих участков постоянна по своей структуре, но полностью эти участки неидентичны из-за присутствия большого количества шума.

Кроме этого, данные последовательности содержат в себе большое количество логических структур, которые вложены друг в друга, причем принцип логической организации пока неясен. Как известно, с точки зрения системного анализа, необходимо применять принцип «от частей к целому», т.е. поэтапно изучать логические уровни организации такой системы.

Методы и средства

Для решения проблем описания конкатенированных данных было предложено представлять их в виде спектра с использованием комплексной системы импульсных функций [3], с помощью которой можно получить спектр, отвечающий указанным требованиям. Предложенная система функций определяется на дискретном множестве

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

и имеет вид

$$Bah_u^k l = \hat{c}_u^k(l) - i\,\hat{s}_u^k(l)$$

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

Функции $\hat{c}_u^k(l)$ и $\hat{s}_u^k(l)$ формируются на основе вспомогательных функций $c_u^k(l)$ и $s_u^k(l)$ посредством их сдвигов на *k* подинтервалов, где $k = 0..2^{n-u-1} - 1$.

Функции $c_u^k(l)$ и $s_u^k(l)$ определяются как:

$$c_0(l) = 1, s_0(l) = 0, l \in M.$$

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

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

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

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

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

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

$$F_u = \sum_{k=0}^{2^{n-u-1}-1} F_u^k,$$

где $u = 0,1,2,3,...,n-1$;

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

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

Проанализированы нуклеотидные последовательности гена Mef2a трех организмов: «Gallus gallus», «Mus musculus domesticus», «Rattus norvegicus». Данные генетических текстов взяты на сайте GenBank (http://www.ncbi.nlm.nih.gov). Индексы базы GenBank соответственно AJ010072, U30823, DQ323505. Последовательности проанализированы с позиции кодона начала трансляции “ATG”.

Для применения метода представления данных с использованием спектрального анализа на основе комплексной системы импульсных функций, необходимо, чтобы анализируемые последовательности были представлены в виде весовых коэффициентов элементов последовательности. В [1] каждой букве генетического текста поставлено в соответствие весовое значение, определенное с помощью молекулярного веса.

Пусть:

$$F_i = \begin{cases} 0, X_i = C; \\ 1, X_i = A; \\ 2, X_i = T; \\ 3, X_i = G, \end{cases}$$

где *Xi – i-й* нуклеотид в последовательности. Однако, такой подход осложняет анализ схожести последовательностей в связи с ненормированным представлением данных, причем проблема остается при любом ранжировании.

Предлагается разделить последовательность на четыре составляющих,

$$F_i^j = \begin{cases} 0, X_i \neq j; \\ 1, X_i = j, \end{cases} \text{ где - } j \in \{C, A, T, G\}.$$

Данный подход позволяет проанализировать степень схожести последовательностей в разрезе их элементной составляющей.

Результаты и их обсуждение

Проанализированы полученные последовательности гена Mef2a, $F_i, i = 1..256$ для трех организмов: «Gallus gallus», «Mus musculus domesticus», «Rattus norvegicus». На рисунках 1,2,3,4 представлены значения F_u для элементов «C», «A», «T», «G» соответственно.

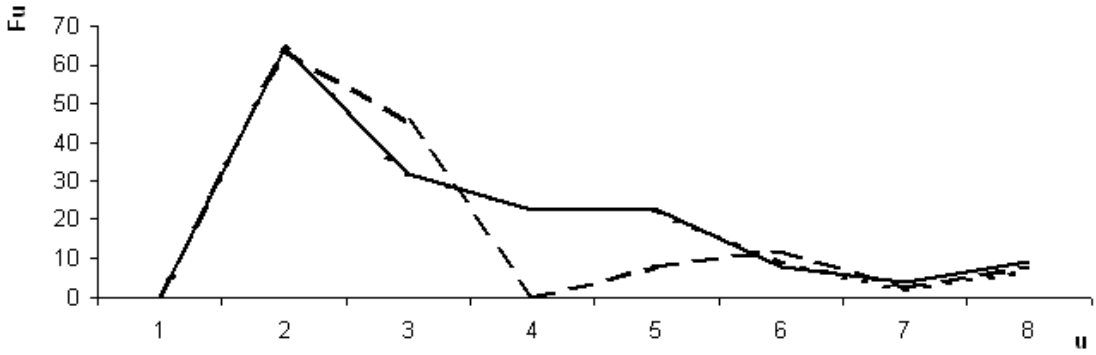

Рис. 1. Амплитудно-частотный спектр нуклеотидных последовательностей элемента «C» гена Mef2a. (- - Gallus gallus, - Mus musculus domesticus, --- Rattus novegicus)

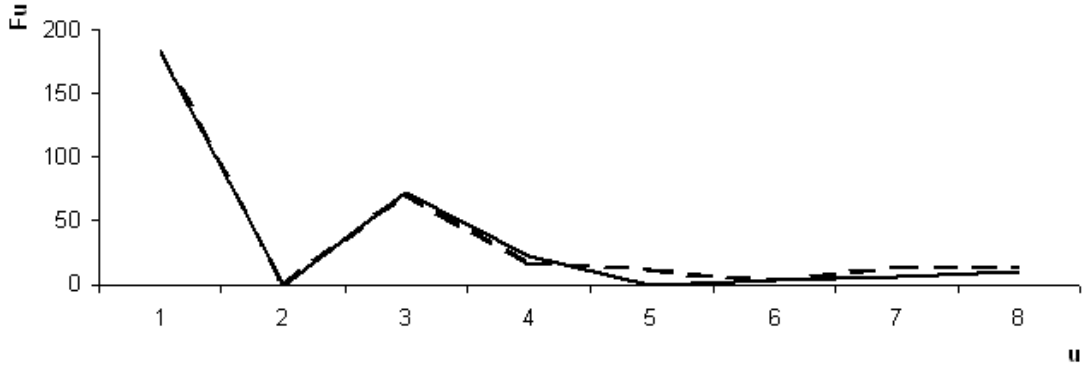

Рис. 2. Амплитудно-частотный спектр нуклеотидных последовательностей элемента «A» гена Mef2a. (- - Gallus gallus, - Mus musculus domesticus, --- Rattus novegicus)

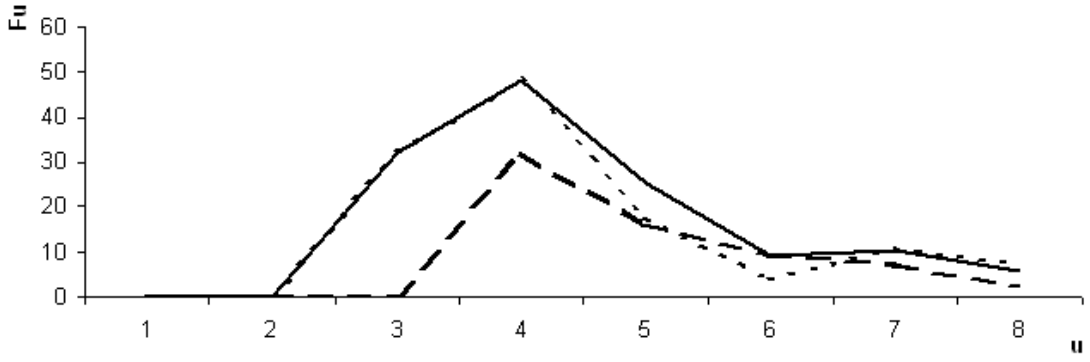

Рис. 3. Амплитудно-частотный спектр нуклеотидных последовательностей элемента «T» гена Mef2a. (- - Gallus gallus, - Mus musculus domesticus, --- Rattus novegicus)

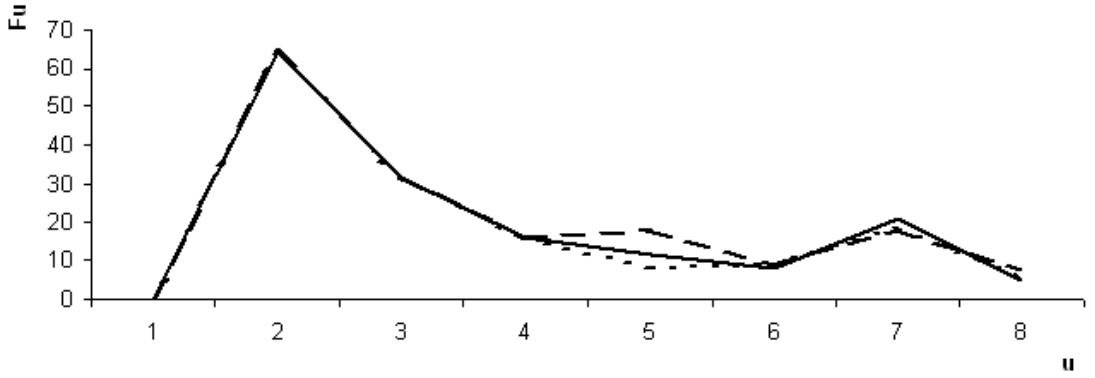

Рис. 4. Амплитудно-частотный спектр нуклеотидных последовательностей элемента «G» гена Mef2a. (- - Gallus gallus, - Mus musculus domesticus, --- Rattus novegicus)

Значения F_i, описывающие ген Mef2a слабо коррелируют между собой, несмотря на то, что данный белок является функционально идентичным у всех трех организмов. После получения амплитудно-частотного спектра F_u, корреляция Пирсона, вычисленная попарно, между значениями F_u, составила 0,88; 0,88; 0,99 соответственно для элемента «C»; 0,72; 0,76; 0,98 соответственно для элемента «T»; 0,99; 0,99; 0,99 соответственно для элемента «A»; 0,98; 0,99; 0,99 соответственно для элемента «G». Данные результаты говорят о том, что схожесть данных трех последовательностей для разных организмов неоднородна для различных элементов. Наблюдается высокое значение корреляции по всем элементам для «Mus musculus domesticus» и «Rattus norvegicus», однако высокая корреляция их и «Gallus gallus» наблюдается только для элементов «A» и «G».

Выводы

1. Метод представления данных с использованием спектрального анализа на основе комплексной системы импульсных функций, позволяет получить амплитудно-частотный спектр конкретного генетического текста.

2. Примененный метод чувствителен к многочисленным сдвигам внутри нуклеотидной последовательности, и позволяет получить описание структуры сигнала с учетом его зашумлености.

3. Разделение исходной последовательности на элементные составляющие дает информацию о схожести для каждого ее элемента.

СПИСОК ЛИТЕРАТУРЫ:

1. Бекасов, Л.С. Методы представления генетической информации

/ Л.С. Бекасов, А.А. Тверетин // Вестник Самарского гос. техн. ун-та, сер. "Физико-математические науки". – 2007. - № 14. - С. 129-134.

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

3. 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». - 1999. - P. 64-67.

COMPARISON OF APPEND DATA ON BASIS OF SPECTRAL CHARACTERISTICS

Tveretin A.A., Bekasov L.S.
Samara state technical university, Samara, Russia

This paper is dedicate to the urgent problem of comparison of append data for the purpose of making pattern for recognition specific situations. Characteristic example of such data is gene texts. The method of data converting with use of spectral analysis on basis of complex system of impulse functions is proposed. Comparison of spectral characteristics for three autonomous texts on basis of such view was made.