УДК 57.01+577.4

ОБЪЁМНЫЙ ПРИНЦИП СПЕКТРАЛЬНОГО ПРЕДСТАВЛЕНИЯ РЕГУЛЯТОРОВ КОНКАТЕНИРОВАННОГО ТИПА

Л.С. Бекасов

Самарский государственный технический университет,
443100, Самара, ул. Молодогвардейская, 244

E-mail: Levbekasov @Yndex.ru

Предложены методы повышения чувствительности при распознавании структурных изменений в сложных системах, представляемых текстами из четырёхбуквенных конкатенаций. Вводится понятие коэффициента насыщенности и преобразование плоскостного представления текстов в объёмное. Использование этих понятий в сочетании с Фурье-методом на базе модифицированных опорных ортогональных импульсных функций позволяет однозначно различать структурные изменения в конкатенированных текстах любой сложности.

Ключевые слова: *GATC-конкатенации, коэффициент насыщенности, модифицированная базисная система опорных импульсных ортогональных функций, расслоение данных, преобразование ранжирования, спектральные характеристики.*

Введение

Информационные процессы в живых средах, как известно, определяют работу систем, характеризующихся очень высоким уровнем сложности. Их элементарные компоненты – нуклеотиды четырёх типов аденин (A), гуанин (G), цитозин (C) и тимин (T) принято считать алфавитом, из которого составляются генетические тексты. Эти элементы при помощи водородных связей могут образовывать комплементарные пары, а именно GC и AT. Формальное описание таких текстов либо их фрагментов представляет большой интерес для биологов, например, для распознавания особенностей в организации отдельных подструктур в генетических конструкциях. В частности, при списывании инструкций, расположенных в генах, ошибка в распознавании хотя бы одной буквы может повлечь за собой исключение реализации непосредственно самого белка, изменение его аминокислотного состава, сбой в считывании самой инструкции, изменение пути сплайсинга (очистки программы синтеза белка от маскированных вставок) и т.д. [1, 2, 3, 4,]. Ещё одна особенность задачи такого распознавания определяется относительно малой величиной распознаваемой цепочки, поскольку именно такие цепочки малой протяжённости являются признаками (метками) либо начал, либо окончаний длины самой инструкции.

Эти факторы обусловливают жесткие требования к формальным методам описания таких текстов.

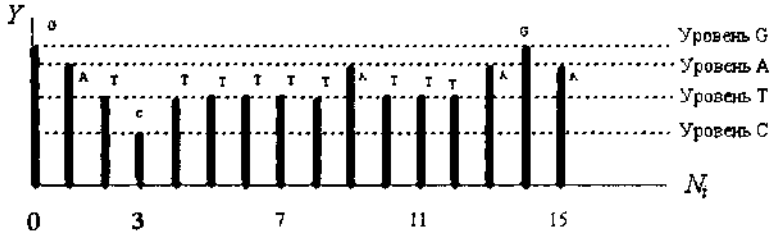
Исходные GATC-конкатенации формально могут быть описаны самыми различными математическими методами [5, 6, 7, 8, 9, 10]. Доминирующее место среди этих методов занимают методы, построенные на статистических и спектральных разделах математики, а также матричном исчислении. Однако данные методы в подавляющим большинстве не обладают возможностями как по чувствительности, так и по разрешающей способности по отношению к отдельно взятым буквам подобных текстов.

Бекасов Лев Степанович – доцент кафедры «Электронные системы и информационная безопасность», к.т.н., доцент.

Ниже приведённые методы формализации сцепленных элементов, в принципе, удовлетворяют таким требованиям распознавания с идеальным разрешением и применимы к анализу любых недетерминированных процессов.

Методы и инструменты

Сущность предлагаемого метода состоит в том, что анализируемая цепочка нуклеотидов *Ф(N)* = GATCTTTTTATTTAGA, как некоторая функция количества (рис. 1), рассматривается через объединение элементарных GATC-подпространств (непересекающихся вложений), т.е. $\{(F_A(N), F_G(N), F_T(N), F_C(N))\}$, следовательно, эти подпространства в целом можно представить в системе NYS декартовых координат. Причём в координате S эти подпространства сформированы послойно в виде отдельных система координат, у которых ось абсцисс *N* отражает однаковое количество *N* нуклеотидов, а оси ординат *Y* показывают ранжированный уровень нуклеотидов, соответствующих некоторому физическому критерию, в частности молекулярной массе. Такое ранжирование предусматривает замену нуклеотидов следующим образом: Гуанин (G)=4, Аденин (A)=3, Тимин T=2, Цитозин (C)=1 [11]. Далее условимся, что оси абсцисс находятся в одной плоскости *P* с одинаковыми расстояниями δ между ними. Упростим задачу и представим на оси ординат всех подпространств лишь бинарное значение уровня нуклеотидов, как показано на рис. 2.



Р и с. 1. Фрагмент анализируемой GATC-нуклеотидной последовательности длиной в 16 пар оснований (*oriC*) в области точки начала репликации хромосомы K-12 *E.coli* (заимствовано из работы Г.И. Кравацкой)

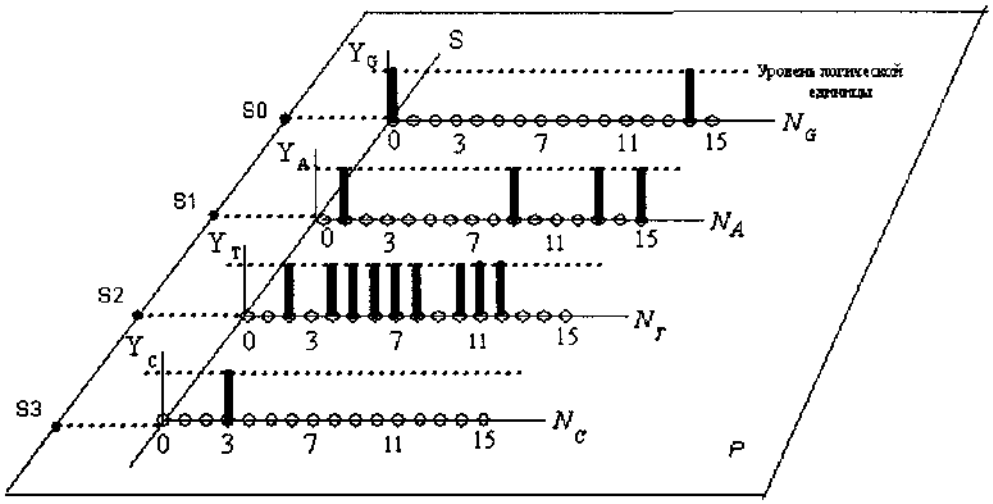
Из рис. 2 следует, что двумерное представление сгруппированных данных по рис. 1 преобразовалось в трёхмерную систему координат, т.е. образовались расслоённые GATC-пространства, находящиеся на равных расстояниях $\delta_i = S_{i+1} - S_i$ друг от друга по оси S. Причём за пределами GATC-пространств $\delta = 0$, а внутри этих пространств $\delta > 0$. На следующем этапе вводится дифференциальная оценка места положения каждой координатной системы на оси S. Будем различать два типа значений δ_i, а именно δ' и δ''. Символ δ' означает число (сумму) интервалов δ, находящихся между координатами S_i и $S_{\max}$ (в частности S3), а символ δ'' определяет число интервалов, находящихся между $S_{\min}$ (S0) и S_i. Принцип этого расслоения формулируется в следующем виде:

$$\Delta_Y = \delta' - \delta'' . \qquad (1)$$

Применительно к анализу фрагмента *Ф(N)* (см. рис. 1) по рис. 2 получим следующие дифференциальные оценки расслоённых систем координат:

$$\Delta_G = \delta_1' + \delta_2' + \delta_3' = 3 - 0 = 3 \qquad \Delta_A = \delta_2' + \delta_3' - \delta_1' = 2 - 1 = 1$$

$$\Delta_C = 0 - \delta_1' - \delta_2' - \delta_3' = 0 - 3 \qquad \Delta_T = \delta_3' - \delta_1' - \delta_2' = 1 - 2 = -1$$



Р и с. 2. Бинарное представление расслоённой GATC-конкатенации по рис. 1

Таким образом, единичные значения на расслоённых осях можно ретрансформировать в исходное состояние (по рис.1) следующими ранжированными категориями, заимствованными из координаты S: G-импульсы представятся уровнем, равным 3, а импульсы A-, T-, C- соответственно уровнями 1, -1 и -3. Множество 3; -3; 1; -1 уровней, полученное таким способом, имеет характер кольца. В работе [12] с целью оценки отношений сцепок из соседних нуклеотидов использовались иные критерии, в частности энергия водородной связи, предложенная Полтевым и его коллегами в [13]. Числовой эквивалент ранжирования этих отношений имел тот же самый кольцевой характер.

Нелишне напомнить, что целью работы является оптимизация (в смысле точности, чувствительности и динамической разрешающей способности) формального представления сцепленных элементов, но не их физиологических возможностей, так как смена изначального ранжирования по другому критерию даст численно иной результат. Ниже при ранжировании нуклеотидов в качестве критерия условимся использовать *молекулярную массу*. Ещё одна особенность состоит в том, что разность по модулю между двумя пуринами/пиримидинами равна двум, тогда как та же разность при ранжировании по молекулярной массе равна единице.

Другой способ формализации предусматривает применение к расслоённой GATC-конкатенации коэффициента насыщенности [14] γ_i раздельно для каждого нуклеотида в соответствии с выражением [14]:

$$\gamma_i = \frac{\rho}{2^r - \rho}, \qquad (2)$$

где ρ – сумма логических единиц в данном подмножестве; $F = 2^r$ – длина всей пачки (r – целое число).

Сумма таких коэффициентов для конкатенации в целом будет характеризовать её однозначно, а именно:

$$\gamma_G = 0{,}1428; \gamma_A = 0{,}3333; \gamma_T = 1{,}2857; \gamma_C = 0{,}0666; \gamma_\Sigma = 1{,}8284.$$

Поскольку известными методами (аналоговыми и дискретными) невозможно получить данные о спектральных свойствах и временной структуре сигнала, что утверждается в [15], ниже предлагается использовать модифицированную базисную комплексную систему импульсных функций, с помощью которой можно получить спектр, отвечающий указанным требованиям. Эта система функций определяется на дискретном множестве

$$M = \{l : l = 0,1,2,...,2^n - 1\} \tag{3}$$

и имеет вид

$$Bah_u^k l = \hat{c}_u^k(l) - i\hat{s}_u^k(l) \tag{4}$$

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

Функции $\hat{c}_u^k(l)$ и $\hat{s}_u^k(l)$ определяются как

$$c_0(l) = 1, s_0(l) = 0, l \in M\ . \tag{5}$$

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

$$c_u(l) = \sum_{m=0}^{2^n-1} (\cos(2^{u-n}\pi m))e(l-m)\,; \tag{6}$$

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

Если *l* принимает другие значения, то $c_u(l) = s_u(l) = 0$.

e(l-m) представляет собой единичный импульс, определяемый из следующих условий:

$$e(l-m) = \begin{cases} 1, l = m; \\ 0, l \neq m. \end{cases} \tag{8}$$

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

$$F_u = \sum_{k=0}^{2^{n-u-1}-1} F_u^k\ , \tag{9}$$

где $u = 0,1,2,3,...,n-1$;

$$F_u^k = \sqrt{(a^k)^2 - (b^k)^2}; a_u^k = \sum_{m=0}^{2^{u-1}-1} f(l_m)\hat{c}_u^k(l); b_u^k = \sum_{m=0}^{2^{u-1}-1} f(l_m)\hat{s}_u^k(l); \tag{10}$$

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

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

Амплитудно-частотный спектр, полученный в соответствии с выражением (10), является инвариантным к временным сдвигам сигнала $f(l)$.

Результаты и обсуждения

Оценка эффективности рассмотренного метода ниже представляется спектральными характеристиками (табл. 1) последовательности *Ф(N)* путём применения к ней преобразования Фурье с нелинейно модифицированными ортогональными базисными функциями применительно к плоскостному *Ф(N)* и объёмному $\Phi_0(N)$ принципам ранжирования.

Таблица 1

Таблица спектральных характеристик с различными форматами ранжирования конкатенированных элементов

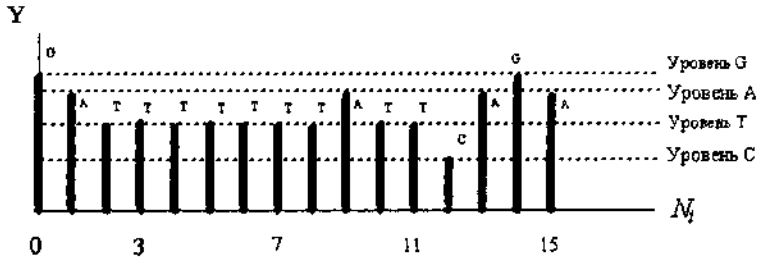
Плоскостное ранжирование *Ф(N)*		Объёмное ранжирование $\Phi_0(N)$	
F0	28,295	F0	16,5576
F1	6,4142	F1	12.8204
F2	5,9906	F2	10,1289
F3	3,0000	F3	6,0000

Далее следует обратить внимание на два типа принципиально различных структурных изменений в анализируемой по рис.1 конкатенации.

В первом случае количество G, A, T, и C элементов в пачке неизменно, но эти элементы перегруппировались. Теперь эта пачка записывается как

Ф(N)*=GATTTTTTATTCAGA. Этот случай показан на рис. 3.

По сравнению с рис.1 здесь поменялись местами элементы C и T. Элемент C переместился с третьей позиции на двенадцатую, а элемент T с двенадцатой позиции сместился на третью.



Р и с. 3. Модифицированный фрагмент анализируемой GATC-нуклеотидной последовательности длиной в 16 пар по рис. 1, в котором поменялись местами элементы C (занимал третью позицию) и T (занимал двенадцатую позицию)

Соответственно изменились спектральные характеристики (табл. 2), тогда как коэффициент насыщенности не изменился, поскольку количество элементов в каждом из подпространств осталось неизменным.

Во втором случае происходят структурные изменения иного характера, т.е. перераспределяется количество элементов в подпространствах, тогда как в целом число элементов в пачке остаётся постоянным (*N*-const.). Пусть исходная конкатенация имеет вид *Ф**(N)*=GATCTTTTGATTTAGA (на девятой позиции элемент T заместился элементом G).

Теперь эти подпространства количественно определяются следующим образом. Имеется три элемента G, четыре элемента A, восемь элементов T и один элемент C. В этом случае спектральные характеристики изменились (представлены табл. 3), и изменился также коэффициент насыщенности. Теперь коэффициенты имеют значения

$$\gamma_G = 0{,}2307; \gamma_A = 0{,}3333; \gamma_T = 1{,}0000; \gamma_C = 0{,}0666; \gamma_\Sigma = 1{,}6300$$

Таблица 2

Спектральные характеристики структурно модифицированных фрагментов регулятора

Объёмное ранжирование			
Ф(N)*		*Ф**(N)*	
F*0	16,5576	F**0	16,5576
F*1	12.8204	F**1	12,4721
F*2	10,1289	F**2	10,0305
F*3	6,0000	F**3	4,4721

Анализ полученных двух таблиц позволяет сделать следующие выводы.

1. Применение процедуры расслоения позволило использовать новый критерий ранжирования элементов конкатенации, а именно дифференциальную разность между отдельными координатами, которая имеет свойство кольцевой группы чисел. Использование этого критерия ранжирования при получении спектральных характеристик позволило удвоить чувствительность Фурье-метода, применяемого для этих целей при идеальном разрешении.

2. Предложенный выше коэффициент γ насыщенности для оценки расслоённого подмножества данных в сочетании с гармоническим анализом четко указывает на характер структурных изменений в исходной конкатенации. В частности, стабильное значение этого критерия указывает на количественную стабильность элементов данного подмножества в пределах анализируемой конкатенации, тогда как спектральная характеристика меняется, о чём свидетельствует сопоставление значений $F(N)$ и $\Phi(N)$ (см. табл. 1 и 2).

3. Перемещение элементов по двум координатам S и N привело соответственно к изменениям величины коэффициента γ и спектральных характеристик $\Phi^{**}(N)$. Первоначальное значение суммарного коэффициента γ_Σ составляло величину $\gamma_\Sigma = 1{,}8284$, а после модификации получило значение $\gamma_\Sigma = 1{,}6300$, что указывает на структурное изменение по двум составляющим. Анализируя коэффициенты насыщенности γ послойно, легко видеть конкретное значение модификации.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Лозовская Е.* От гена к белку через сплайсинг. http://www.inauka.ru/science/artikel69 210.html.
2. Информационный справочный ресурс по биологии. http://www. cellbiol. ru/book /molekulyarnaya_biologiya/processing_rnk/splajsing
3. *Марков А.* Почти все человеческие гены кодируют более одного белка. http://bio.fizteh.ru /student biotech /2008/altertative_splis_09112008.html?xsl:print=1
4. *Марков А.* Новый механизм генной регуляции без участи белков. http://www.science-time.com/novyi-mexanizm-gennoi-regulyacii-bez-uchastiya-belkov.html
5. *В.В. Налимов.* Теоретическая биология? Ее все еще нет… http://www.biometrica.tomsk.ru/index.htm.

6 Statistical Methods for DNA Sequence Segmentation. J.V. Braun. H-G.Muller // Statistical Science. – 1998. – Vol. 13. – №2. – 142-162.
7 *Петухов Н.Н.* Биериодическая таблица генетического кода и число протонов. – М., 2001 (http://members. tripod.com/vismath/sg/petoukhov.htm).
6. Статистические флуктуации в процессах регуляции экспрессии генов: рассмотрение с точки зрения статистической механики / Ю.Д. Нечипуренко, А.М. Вольф, Г.В. Гурский // Молекулярная биофизика. – 2003. – Том 48. – Вып. 6. – 986-997 с.
7. Сходство периодических структур в расположении нуклеотидов на участках начала репликации бактериальных геномов / Г.И. Кравацкая, Г.К. Франк, В.Ю. Макеев, Н.Г. Есипова // Биофизика. – 2002. – Том 47. – Вып. 4. – 595-599 с.
8. A novel Lie algebra of the gene3tic code jver the Galois field of four DNA bases Robersy Sanchez, Ricardo Grau, Eberto Morgado // Mathematical Biosciences. – 202 (2006). – 156-174.
9. *Бекасов Л.С.* Моделирование сайтов узнавания // Вестник Самар. гос. техн. ун-та. Сер. Технические науки. – 2009. – Вып. 41. – С. 59-64.
10. *Бекасов Л.С., Тверетин А.А.* Методы представления генетической информации // Вестник Самар. гос. техн. ун-та. Сер. Физико-математические науки. – 2007. – Вып. 14. – С. 129-134.
11. Взаимодействия между основаниями нуклеиновых кислот. Новые параметры потенциалов и новые минимумы энергии / В.И. Полтев, А.С. Дерябина, Э. Гонзалес, Т.И. Грохлина // Биофизика. – 2002. – Т. 47. – С. 996-1004.
12. *Бекасов Л.С., Васильева Л.П., Лапаева М.М.* Распознавание структурной организации конкатенированных текстов с минимальным синтаксисом и неопределённой семантикой //Актуальные проблемы информационной безопасности. Теория и практика использования программно-аппаратных средств: Тр. Всерос. конф. с международ. участием. – Самара, 2008. – С. 32-37.
13. *Bachrushina G.I., Bachrushin A.P.* Development and Investigation of a New Rectangular Orthogonal System Functions for Invariant Object Recognition // Proceedings of the Sixth International Conference «Advanced Computer Systems» / Szezecin-Poland. – 1999, November. – P. 64-67.

Статья поступила в редакцию 20 августа 2009 г.

UDC 57.01+577.4

VOLUME IDEA OF SPEKTRAS REPRESENTATION OF APPEND-TYPE CONTROLLERS

L.S. Bekasov

Samara State Technical University,
244, Molodogvardeyskaya str., Samara, 443100

The methods of sensitizing at distinction of structural modifications in multiple Systems, which are adduced by texts from four-letter Concatenations are proposed in this work. There are also introduced such ideas as "the index of saturation" and plane conception of texts into volume conception reformation. Using of these ideas in combination with the Fourier-method conception based on modify abutting orthogonal impulse functions allow to discriminate structural changes in appended texts of any complexity.

Key words: *GATC-concatenations, saturation coefficient, modified basic system of abutting impulse.*