ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2014 Управление, вычислительная техника и информатика № 4 (29)

УДК 004.8: 575

А.С. Гуменюк, Н.Н. Поздниченко, С.Н. Шпынов

ФОРМАЛЬНЫЙ АНАЛИЗ СТРОЯ ЛОКАЛЬНОЙ СТРУКТУРЫ НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Рассматриваются средства для анализа локальной структуры нуклеотидных последовательностей, которые в предыдущих публикациях использовались для оценки порядка расположения компонентов нуклеотидной цепи в целом. Определены функции некоторых характеристик строя и представлены соответствующие формулы для их вычисления. Рассмотрены возможности использования этих функций для описания и исследования локальной структуры нуклеотидных цепей.

Ключевые слова: строй цепи; нуклеотидная последовательность; характеристики строя; функции характеристик строя; L-граммы; локальная структура нуклеотидной цепи.

В опубликованных ранее работах [1, 2] дано определение строя цепи [3] и представлены интегральные характеристики строя знаковых, в том числе нуклеотидных последовательностей, которые показали высокую чувствительность к расположению компонентов. На основе введенных формализмов были рассмотрены возможности сравнения, классификации, хеширования последовательностей с помощью характеристик строя.

Разрабатываемый авторами подход позволяет отображать локальную структуру знаковых цепей произвольной природы числовыми последовательностями, которые представляют расположение компонентов. Это, в свою очередь, открывает возможности для применения разнообразных математических методов анализа числовых массивов данных, которые без такого преобразования неприменимы непосредственно к символьным последовательностям.

1. Интегральные характеристики строя цепи

Прежние публикации [4, 5] представляли средства и числовые характеристики для анализа строя целостных и полноразмерных нуклеотидных цепей, а также составных частей таких цепей. Приведем некоторые из этих характеристик:

$$\Delta_{ij} = x_{i+1j} - x_{ij};\ x_{i+1j}, x_{ij} \in [1, n], \tag{1}$$

$$G = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log_2 \Delta_{ij}, \tag{2}$$

$$g = {}^{1}/_{n}\, G = \log_2 \Delta_g = \sum_{j=1}^{m} {}^{n_j}/_{n} \log_2 \Delta_{gi}, \tag{3}$$

$$r = {}^{\Delta_g}/_{D}, \tag{4}$$

$$\Delta_g = \sqrt[n]{\prod_{j=1}^{m} \prod_{i=1}^{n_j} \Delta_{ij}}, \tag{5}$$

где x – номер места на позиции рассматриваемой полноразмерной цепи; x_{ij} – номер места i-го вхождения j-го элемента алфавита на позиции данной цепи; Δ_{ij} – интервал от i-го до $(i + 1)$-го вхождения j-го символа; n_j – число вхождений j-го символа; n – длина цепи (число мест на её позиции); m – мощность алфавита рассматриваемой цепи (для нуклеотидной последовательности $m = 4$); G – глубина расположения компонентов в цепи; g – средняя удалённость компонентов в цепи; r – регулярность расположения компонентов в цепи ($0 < r < 1$); Δ_g – среднегеометрический интервал между любыми соседними одинаковыми символами; D – число описательных информаций, вычисляемое по формуле М. Мазура [6].

2. Функции характеристик строя цепи

Общепринятым методом исследования больших массивов данных измерений, лингвистических текстов, нуклеотидных последовательностей и длинных цепей другой природы является «просмотр окном» [7]. В настоящей работе представлены средства для анализа локальной структуры целостных полноразмерных последовательностей на основе характеристик строя отдельных, но связных фрагментов (L-грамм) и возможности их использования.

Для формального определения функции характеристики строя введем ряд понятий, часть из которых дана в рамках «алгебры ментальных событий» [8].

Место – элементарная ячейка, предназначенная для хранения одного компонента цепи. **Позиция** – это упорядоченное множество мест. **Фрагмент** – участок полной цепи. **Окно** – позиция фрагмента (участок позиции полной цепи). **Размер окна** – количество мест на позиции окна. **Шаг** – это смещение окна на позиции полной цепи, позволяющее выделить следующий фрагмент цепи. **Размер шага** – размер смещения окна, измеряемый числом мест. **Функция характеристики строя цепи** – это упорядоченное множество значений характеристик строя, вычисленных для всех фрагментов, последовательно взятых на позиции полной цепи. **Отсчётное значение функции характеристики строя** – это значение функции характеристики строя, вычисленное для отдельного фрагмента, задаваемого его номером, длиной и размером шага.

Ниже представлены формулы для вычисления отсчётных значений некоторых функций характеристик строя.

$$\Delta_{ij} = x_{i+1j} - x_{ij};\ x_{i+1j}, x_{ij} \in [s*k, s*k+l], \tag{6}$$

$$f_G(k,l,s) = \sum_{j=1}^{m}\sum_{i=1}^{n_j} log_2\Delta_{ij}, \tag{7}$$

$$f_g(k,l,s) = f_G(k,l,s)/l\ , \tag{8}$$

$$f_{\Delta_g}(k,l,s) = \sqrt[l]{\prod_{j=1}^{m}\prod_{i=1}^{n_j}\Delta_{ij}}, \tag{9}$$

$$f_r(k,l,s) = f_{\Delta_g}(k,l,s)/f_D(k,l,s)\ , \tag{10}$$

где x_{ij} – номер места i-го вхождения j-го элемента алфавита на позиции данного фрагмента; k – номер фрагмента; s – размер шага (при s = 1 фрагменты являются L-граммами); l – размер окна; $f_G(k,l,s)$ – функция глубины; $f_g(k,l,s)$ – функция средней удалённости; $f_{\Delta_g}(k,l,s)$ – функция среднего геометрического интервала; $f_r(k,l,s)$ – функция регулярности; $f_D(k,l,s)$ – функция числа описательных информаций.

Общее количество фрагментов при заданных длине цепи, шаге и размере фрагмента определяется в виде

$$k_{max} =]^{n}/_{S} - l + s[. \tag{11}$$

Заметим, что мощность алфавита данного фрагмента m может быть меньше мощности алфавита всей цепи (минимум 1, если фрагмент полностью заполнен одинаковыми компонентами). Аргументы функций (k,l,s) – натуральные числа. Таким образом, данные функции являются функциями дискретных аргументов. Зная все три параметра, можно посчитать отдельное значение такой функции. Также возможно вычислить многомерную функцию, изменяя не только номер фрагмента, но и два других параметра. Так как все функции характеристик строя не имеют обратных, задача восстановления значений аргументов по отдельному значению функции является поисковой и требует больших вычислительных ресурсов.

3. Свойства функций характеристик строя

На рис. 1 представлены графики функции удалённости 18S рибосомальной РНК организма *Cricetulus griseus* (Хомячок китайский) (GenBank id DQ235090.1?from=11629&to=13499) [9], вычисленные с размером окна 100 (рис. 1, *a*) и 25 (рис. 1, *b*) и шагом 2. Сравнение графиков показывает, что при

уменьшении размера окна график функции характеристики оказывается всё менее «сглажен» и в нем проявляются всё большие колебания.

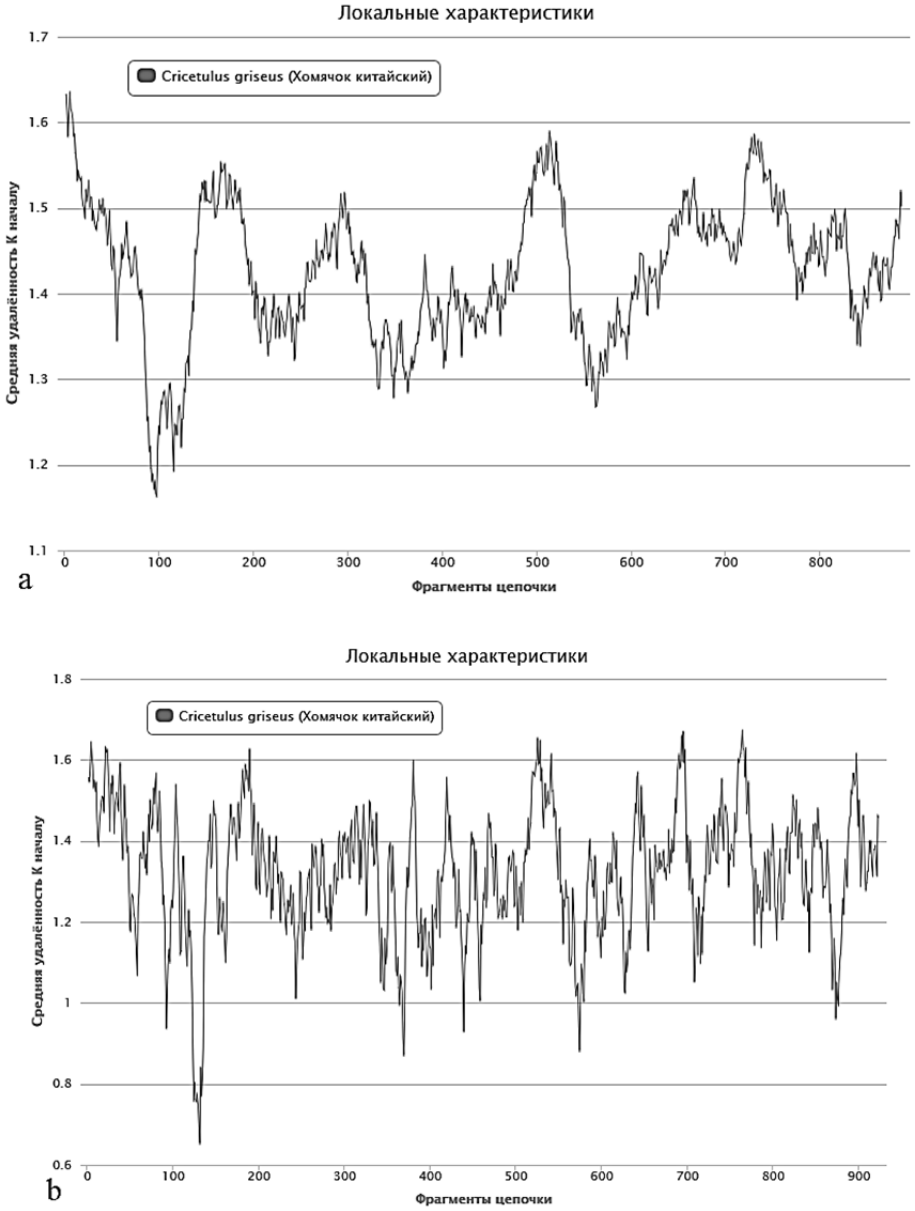

Рис. 1. Графики функции удалённости: *a* – при длине фрагментов 100 нуклеотидов; *b* – при длине фрагментов 25 нуклеотидов

Больший размер окна позволяет обнаруживать схожие по строю фрагменты большей длины. В пределе, при увеличении размера фрагментов, значения функции стремятся к значению соответствующей интегральной характеристики строя полноразмерной цепи. При уменьшении длины фрагментов отдельные значения функции позволяют выявлять всё более тонкие особенности расположения компонентов в пределах окна фрагмента. Предварительные исследования показали, что зависимость между размером окна и разбросом характеристики имеет гиперболический характер. Однако если размер окна уменьшается до мощности алфавита ($m = 4$), данная зависимость нарушается. Таким образом, как и предполагалось, неопределённость расположения фрагмента связана с неопределённостью значений функции, получаемых при заданном размере окна.

На рис. 2 представлены графики функции удалённости (рис. 2, *a*) и функции регулярности (рис. 2, *b*) той же нуклеотидной последовательности, что и на рис. 1, вычисленные с размером окна 50 и шагом 2. Из рисунка видно, что данные функции не являются функционально зависимыми и могут дополнять друг друга при комплексном описании локальной структуры нуклеотидных последовательностей.

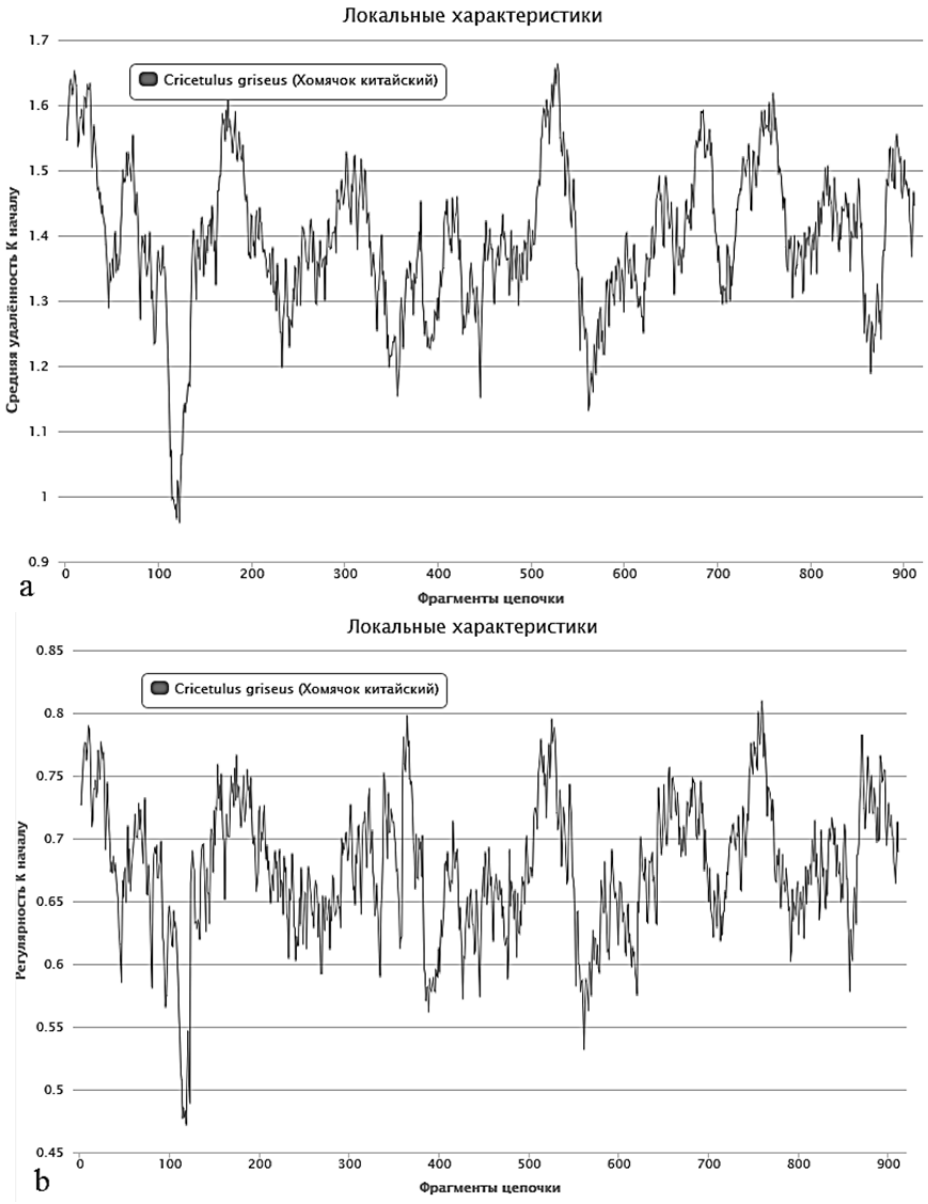

Рис. 2. Графики функций характеристик строя: *a* – функция удалённости; *b* – функция регулярности

Следует учитывать, что при увеличении длины шага уменьшается количество вычисляемых отсчётных значений, и для выявления совпадающих цепочек и фрагментов при таком разбиении требуется более сложная поисковая процедура, которая может потребовать дополнительных вычислительных ресурсов. Кроме того, при экспертном анализе сокращение количества отсчётных значений упрощает восприятие графического представления функций, вычисленных для длинных нуклеотидных последовательностей.

Дополнительных исследований требует выбор оптимального размера окна для решения различных задач, в том числе в зависимости от мощности алфавита исходной последовательности.

4. Применение функций характеристик строя для исследования нуклеотидных цепей

Ниже на представленных графиках видно, что особи одного вида (рис. 3) демонстрируют практически идентичную форму функции глубины f_G. Виды одного класса (рис. 4) имеют схожую форму этой функции, виды разных классов (рис. 5, 6) – сильно отличающуюся форму данной функции.

Рис. 3. Графики функции глубины 18S рибосомальных РНК двух особей одного вида (DQ235090.1, NR_045132.1)

Рис. 4. Графики функции глубины 18S рибосомальных РНК двух видов, принадлежащих к разным отрядам одного класса (DQ235090.1, AJ311675.1)

Рис. 5. Графики функции глубины 18S рибосомальных РНК двух видов, принадлежащих к разным классам одного типа (AJ311672.1, DQ235090.1)

Рис. 6. Графики функции глубины 18S рибосомальных РНК двух видов, принадлежащих к разным классам одного типа (AJ311672.1, EU637036.1)

Отметим, что значения интегральной характеристики этих же организмов *G* (таблица) соответствуют представленным графикам [4].

Значения интегральной характеристики исследуемых нуклеотидных последовательностей

№ пп	Название организма	*G*	№ пп	Название организма	*G*
1	*Cricetulus griseus*	2745,91	4	*Erinaceus europaeus*	2 671,10
2	*Cricetulus griseus* 2	2752,30	5	*Kareius bicoloratus*	2 722,66
3	*Crocodylus niloticus*	2625,63			

Перечислим другие применения функций характеристик строя: выделение повторяющихся одинаковых или схожих фрагментов; выделение разных фрагментов с одинаковым или схожим строем; поиск границ генов, интронов и экзонов, «слов»; установление функционального назначения некодирующих последовательностей; сравнение последовательностей; более надёжное хеширование, чем на основе интегральных характеристик.

Заключение

Сформулировано понятие функции характеристики строя цепи. Заданы формулы для вычисления значений некоторых функций характеристик строя и продемонстрированы возможности использования таких функций для анализа локальной структуры нуклеотидных последовательностей. Разработаны программные средства для вычисления и отображения функций характеристик строя [10]. Программные средства апробированы при сравнении рибосомальных РНК нескольких организмов. По результатам исследований выявлено влияние длины фрагментов (*L*-грамм) на форму функций характеристик строя.

Отмечены возможности использования функций характеристик строя для поиска схожих или совпадающих фрагментов в рамках одной или нескольких генетических цепочек, а также решения обратной задачи – поиск мест вхождений заданного фрагмента на позиции полноразмерной цепи.

Отображение строя нуклеотидных последовательностей функциями, кроме отмеченных средств, позволяет применять также классические методы математики, а именно математический анализ, спектральный анализ, корреляционный анализ и т.п., что было бы невозможно при непосредственном анализе самих знаковых последовательностей. Отмечено, что графическое представление функций характеристик строя делает возможным экспертный анализ длинных нуклеотидных цепей.

ЛИТЕРАТУРА

1. *Гуменюк А.С., Морозенко Е.В., Родионов И.Н.* Формализация анализа строя знаковых цепей // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2011. № 2(15). С. 15–23.
2. *Gumenyuk A., Kostyshin A., Simonova S.* An approach to the research of the structure of linguistic and musical texts // Glottometrics. 2002. No. 3. С. 61–69.
3. *Гуменюк А.С.* О средствах анализа взаимного расположения компонентов знаковой последовательности // Военная техника, вооружение и технологии двойного применения : материалы III Междунар. технолог. конгр. Омск : ОмГТУ, 2005. Ч. 2. С. 48–52.
4. *Гуменюк А.С., Поздниченко Н.Н., Шпынов С.Н., Родионов И.Н.* О средствах формального анализа строя нуклеотидных цепей // Математическая биология и биоинформатика. 2013. Т. 8, № 1. С. 373–397. URL: http://www.matbio.org/article.php?journ_id=15&id=158
5. *Гуменюк А.С., Поздниченко Н.Н.* Анализ строя нуклеотидных последовательностей // Материалы Всероссийской конференции с международным участием «Знания – онтологии – теории» (ЗОНТ-2013) 8–10 октября 2013 года. Новосибирск, 2013. Т. 2. С. 58–68.
6. *Мазур М.* Качественная теория информации. М. : Мир, 1974. 240 с.
7. *Садовский М.Г.* Информационно-статистический анализ нуклеотидных последовательностей : дис. ... д-ра физ.-мат. наук. Красноярск, 2004. 394 с.
8. *Гуменюк А.С.* О формализмах конструирования абстрактных объектов во внутреннем физическом пространстве информационной системы (Элементы алгебры ментальных событий) // Системный анализ в проектировании и управлении: труды X Междунар. науч.-практ. конф. СПб. : Изд-во Политех. ун.-та, 2006. Ч. 2. С. 172–181.
9. *National* Center for Biotechnology Information. URL: http://www.ncbi.nlm.nih.gov/nuccore/
10. *Цымбал В.С., Поздниченко Н.Н.* О разработке модуля для вычисления локальных характеристик строя нуклеотидных последовательностей // Материалы XII Всероссийской научно-практической конференции с международным участием «Информационные технологии и математическое моделирование (ИТММ-2013)». Томск : Изд-во Том. ун-та, 2013. С. 61–65.

Гуменюк А.С., канд. техн. наук, доцент. E-mail: gumas45@mail.ru
Омский государственный технический университет
Поздниченко Н.Н. E-mail: nick670@yandex.ru
Омский государственный технический университет
Шпынов С.Н., д-р мед. наук. E-mail: stan63@inbox.ru
НИИЭМ им. Н.Ф. Гамалеи (г. Москва)

Поступила в редакцию 8 июня 2014 г.

Gumenyuk Alexander S., Pozdnichenko Nikolay N. (Omsk State Technical University, Russian Federation)*, Shpynov Stanislav N.* (Gamaleya Institute of Epidemiology and Microbiology, Russian Federation).

Formal analysis of order in the local structure of the nucleotide sequences.

Keywords: chain's order, sequence, nucleotide sequence, order characteristics, functions of order characteristics, L-grams, local structure of nucleotide chain.

The definition of the chain order and integral characteristics of the order, in particular, for nucleotide sequences were presented in the previous papers. These characteristics showed the high sensitivity to the arrangement of components. The possibility of comparison, classification and hashing based on the introduced formalisms and using characteristics to order have been considered. The approach developed by the authors allows displaying the local structure of sign sequences of arbitrary nature by numerical sequences that represent the arrangement of their components.

The generally accepted method of studying large arrays of measurement data, linguistic texts, nucleotide sequences, and long sequences of another nature is the «window scan». This paper describes means for analysis of the local structure of complete full-length sequences based on the characteristics of the order of separate fragments (L-grams), named the functions of characteristics of order. The formulas for calculation of the values of these functions are of the following form:

$$\Delta_{ij} = x_{i+1j} - x_{ij};\ x_{i+1j}, x_{ij} \in [s*k, s*k+l],$$

$$f_G(k,l,s) = \sum_{j=1}^{m}\sum_{i=1}^{n_j} log_2 \Delta_{ij},$$

$$f_g(k,l,s) = f_G(k,l,s)/_l,$$

$$f_{\Delta_g}(k,l,s) = \sqrt[l]{\prod_{j=1}^{m}\prod_{i=1}^{n_j}\Delta_{ij}},$$

$$f_r(k,l,s) = f_{\Delta_g}(k,l,s)/f_D(k,l,s),$$

where x_{ij} is a number of position of i-th occurrence of j-th element of alphabet on position of current fragment; k is a number of fragment; s is a step size (when s=1 fragments become L-grams); l is a window length; $f_G(k,l,s)$ is a depth function; $f_g(k,l,s)$ is an average

remoteness function; $f_{\Delta_g}(k,l,s)$ is an average geometric interval function; $f_r(k,l,s)$ is a regularity function; $f_D(k,l,s)$ is a descriptive information function.

A larger window allows detecting fragments with similar order of greater length. Increasing the length of the fragments results in function values provides tending to a value of corresponding integral characteristic of the full length sequence. Reducing the length of the fragments allows using separate functions values for detection of more detailed features of the arrangement of components within the window. Preliminary studies showed that the relationship between the window length and the dispersion of the characteristic values is hyperbolic. However, if the window length is reduced to the cardinality of the alphabet (m = 4), this dependence is violated. Thus, as expected, the uncertainty of the location of the fragment is associated with the uncertainty of the function values obtained for a given window length. Selection of an optimal window length for various tasks, including, dependence on the cardinality of the alphabet of the original sequence, requires additional research.

Software for calculating and displaying the functions of characteristics of order is developed and tested on ribosomal RNA of several organisms.

Research revealed the influence of fragment (L-gram) length on the shape of functions of characteristics of order. The possibility of using the functions of the characteristics of order for finding similar or overlapping fragments in one or more sequences is considered, as well as – the inverse problem – finding of occurrences of the specified fragments in the complete genome sequence. Displaying the order of nucleotide sequences with functions, besides noted means also allows using the classical methods of mathematics, such as: mathematical analysis, spectral analysis, correlation analysis, etc., that would be impossible with the direct analysis of symbolic sequences themselves. It is noted that the graphical representation of functions of characteristics of order allows carrying expert analysis of long nucleotide sequences, including complete genome sequences.

REFERENCES

1. Gumenyuk A.S., Morozenko E.V., Rodionov I.N. Formalization of analysis of order of sign chains. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika – Tomsk State University Journal of Control and Computer Science*, 2011, no. 2(15), pp. 15-23. (In Russian).
2. Gumenyuk A., Kostyshin A., Simonova S. An approach to the research of the structure of linguistic and musical texts. *Glottometrics*, 2002, no. 3, pp. 61-69.
3. Gumenyuk A.S. [On the means of analysis of mutual arrangement of the components of the sign sequence]. *Voennaya tekhnika, vooruzhenie i tekhnologii dvoynogo primeneniya* : materialy III Mezhdunar. tekhnolog. Kongressa [Military equipment, arms and double purpose technologies. Procs. of the 3rd International Congress of Technology]. Omsk: OmSTU Publ., 2005, pt. 2, pp. 48-52. (In Russian).
4. Gumenyuk A.S., Pozdnichenko N.N., Shpynov S.N., Rodionov I.N. Formal Analysis of Structures of Nucleotide Chains. *Matematicheskaya biologiya i bioinformatika – Mathematical Biology and Bioinformatics*, 2013, vol. 8, no. 1, pp. 373-397. Available at: http://www.matbio.org/arti-cle.php?journ_id=15&id=158.
5. Gumenyuk A.S., Pozdnichenko N.N. [Analysis of order of the nucleotide sequences]. *Materialy Vserossiyskoy konferentsii s mezhdunarodnym uchastiem "Znaniya – ontologii – teorii" (ZONT-2013)* [The 4th All-Russian Conference "Knowledge – Ontology – Theory" (KONT-13).]. Novosibirsk, 2013, vol. 2, pp. 58-68. (In Russian).
6. Mazur M. *Kachestvennaya teoriya informatsii* [Qualitative information theory]. Moscow: Mir Publ., 1974. 240 p.
7. Sadovskiy M.G. *Informatsionno-statisticheskiy analiz nukleotidnykh posledovatel'nostey*: dis. d-ra fiz.-mat. nauk [Information and statistical analysis of nucleotide sequences. Physics and Mathematics Doc. Thesis]. Krasnoyarsk, 2004. 394 p.
8. Gumenyuk A.S. [On the formalism of constructing abstract objects in the inner physical space of information system (Elements of the algebra of mental events)]. *Sistemnyy analiz v proektirovanii i upravlenii: trudy X Mezhdunar. nauch.-prakt. konf.* [The system analysis in the design and management. Procs. of the 10th International Scientific and Practical Conference]. St. Petersburg: Polytechnic University Publ., 2006, pt. 2, pp. 172-181.
9. *National Center for Biotechnology Information*. Available at: http://www.ncbi.nlm.nih.gov/nuccore/
10. Tsymbal V.S., Pozdnichenko N.N. [About the development of the module for the calculation of local characteristics of order of the nucleotide sequences]. *Materialy XII Vserossiyskoy nauchno-prakticheskoy konferentsii s mezhdunarodnym uchastiem "Informatsionnye tekhnologii i matematicheskoe modelirovanie (ITMM-2013)"* [Proc. of 12th All-Russian research conference "Information technologies and mathematical modeling" (ITMM-2013)]. Tomsk: Tomsk State University Publ., 2013, pp. 61-65.