ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2018 Управление, вычислительная техника и информатика № 43

УДК 004.8: 575
DOI: 10.17223/19988605/43/7

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

КАРТА ГЕНОВ – НОВОЕ СРЕДСТВО ПРЕДСТАВЛЕНИЯ МНОЖЕСТВА ОДНОХРОМОСОМНЫХ ГЕНОМОВ И ИХ КОМПОНЕНТОВ

Отмечена возможность однозначного отображения полных геномов и их компонентов информационными числовыми характеристиками строя, что позволяет быстро различать, сравнивать и осуществлять поиск длинных нуклеотидных последовательностей. На этой основе предложено наглядное картографическое представление аннотированных геномов, что позволяет облегчить неформальный экспертный анализ совокупностей организмов, а также их интерактивное и автоматическое исследование.
Ключевые слова: формальный анализ строя, межнуклеотидное расстояние, карта генов, хеширование характеристиками строя.

В работах [1,2] представлен новый подход, разрабатываемый на основе теории информации М. Мазура. При обработке упорядоченных массивов данных разной природы (так называемых информационных цепей) с использованием средств формального описания и анализа строя (ФОАС) непосредственно учитывается расположение компонентов в отдельных последовательностях. Для этих целей осуществляется предварительное преобразование массивов данных в строй цепи. Связи между компонентами строя – отдельные информации – определяются с помощью интервалов между ближайшими одинаковыми элементами (в случае нуклеотидных последовательностей – это межнуклеотидные расстояния [2–4]). Интервалы представляют описательные информации, а их произведение – число таких информаций. Двоичный логарифм от этого числа представляет количество идентифицирующих информаций. Полученные таким образом числовые информационные характеристики строя представляют полное оригинальное расположение элементов в целостном объекте.

Ранее формализм и характеристики строя при исследовании генетических последовательностей использовались для следующих целей: классификация организмов на высоких таксономических уровнях [2]; классификация прокариот на уровнях вида, рода, семейства [5]; определение сходства (близости) генетических последовательностей посредством сравнения распределений характеристик однородных последовательностей и вычисления матриц соответствий [2]; изучение локальной структуры нуклеотидных последовательностей; поиск различающихся фрагментов последовательностей с одинаковым строем [2].

1. Числовые характеристики строя как меры полных геномов и их компонентов

В данной работе предлагается использовать отдельные характеристики строя для двух новых задач: «хеширование» [6–8] генетических текстов и построение «карты генов». Использование одной характеристики не гарантирует отсутствия коллизий. Однако использование нескольких характеристик (в том числе частотных) позволяет значительно снизить их число.

Адекватное высокочувствительное числовое отображение расположения компонентов в длинных нуклеотидных последовательностях позволяет осуществлять «хеширование» как полных геномов, так и их фрагментов и составляющих частей. Хеширование нуклеотидных

последовательностей с помощью характеристик строя (т.е. их расположение в соответствии со значением той или иной характеристики) позволяет быстро различать, сравнивать и осуществлять поиск цепей без необходимости их поэлементного сравнения. При этом последовательности можно «расположить» и на числовой оси для экспертного либо автоматизированного анализа.

Среди разнообразных характеристик строя в результате обработки больших массивов данных были выделены наиболее емко отображающие целостную последовательность – глубина G и средняя удаленность g, представленные здесь в виде

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

$$G = n \cdot \log_2 \Delta_g, \qquad (1)$$

$$g = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j} \log_2 \Delta_{ij},$$

$$g = \sum_{j=1}^{m} \frac{n_j}{n} \log_2 \Delta_{gj}, \qquad (2)$$

$$g = \log_2 \Delta_g,$$

где m – мощность алфавита (m = 4); n – длина последовательности; n_j – число вхождений j-го нуклеотида; Δ_{ij} – интервал от i-го до (i + 1)-го вхождения j-го нуклеотида; Δ_g – среднее геометрическое всех интервалов последовательности; Δ_{gj} – средний геометрический интервал j-й однородной цепи.

Первая характеристика – глубина расположения нуклеотидов в полной последовательности, – как видно из формулы (1), оценивает целостный объект длиной n. Вторая характеристика представляет усредненную удаленность гомологичных нуклеотидов в последовательности, которая позволяет эффективно определять сходство геномов организмов независимо от длины последовательности. Эти же характеристики показали свою эффективность и при хешировании.

Несмотря на внешнее сходство, характеристики строя (в том числе формула (2)) не являются статистическими. Как правило, статистические оценки некоторой величины осуществляются суммированием ее отдельных значений без учета их расположения (с последующим делением на число слагаемых). Поэтому разные по порядку следования «выборки реализаций» могут давать одинаковые значения статистических характеристик. Напротив, все характеристики строя основаны на перемножении «межсобытийных» (межкомпонентных) интервалов (с последующим извлечением корня степени числа сомножителей). При изменении порядка следования событий в последовательности почти всегда меняются сомножители, поэтому числовые характеристики строя непосредственно учитывают расположение компонентов (строя) информационной цепи. В логарифмическом масштабе характеристики строя основаны на простом суммировании логарифмов межсобытийных интервалов.

Важно отметить, что характеристики строя можно использовать вместо хеш-суммы для компактного представления и быстрого сравнения символьных последовательностей (в том числе нуклеотидных). Однако они не могут непосредственно использоваться в криптографических целях, так как, в отличие от общепринятого хеша, не обладают свойствами динамического хаоса и лавинным эффектом (хеш-сумма двух схожих, но не идентичных, объектов будет сильно отличаться, и наоборот, имея хеш, практически невозможно найти, для какого объекта он вычислен). Характеристики строя для схожих объектов дают близкие значения.

2. Некоторые особенности организации данных в GenBank

В настоящее время наиболее крупной (международной) библиотекой нуклеотидных последовательностей и, в частности, полных геномов является GenBank (NCBI). Авторы последовательностей, представляющих полные геномы, при загрузке последовательности могут дать ее аннотацию и / или воспользоваться средством автоматического аннотирования [9]. Такая аннотация включает в себя информацию о «расположении» в полном геноме различных компонентов, таких как гены, различные РНК (рибосомальные, транспортные и т.д.), псевдогены, повторы, мобильные элементы и т.д. Поэтому для большинства геномов аннотации представлены в двух видах: загружаемые авторами или автоматические, выполненные инструментарием GenBank. Поэтому разные аннотации могут значительно отличаться, что затрудняет исследование и сравнение организмов по их компонентам.

Как правило, на сайте NCBI нуклеотидная последовательность полноразмерного генома, секвенированная одной группой авторов, доступна в двух базах данных: RefSeq (NCBI Reference Sequence Database) и INSDC (International Nucleotide Sequence Database Collaboration). Различия в номенклатуре и структуре аннотаций, а также отсутствие некоторых данных при депонировании последовательностей авторами зачастую усложняют процесс сравнения геномов по существующим аннотациям. Так, в связи с несовершенством способов автоматической аннотации для многих компонентов неизвестны их точные позиция и длина. Автоматическую обработку аннотаций также усложняет то, что любая кодирующая область (и большинство других типов компонентов) размечена дважды: как CDS (coding sequence – кодирующая последовательность) и как gene (ген), в то время как некоторые редкие типы компонентов размечены только один раз. В файле, соответствующем отдельном геному, в разделе «Особенности» (Features) не всегда заполнены рубрики: источник (source), организм (organism) и описание (description). Зачастую авторы последовательностей заполняют эти поля несогласованно, что затрудняет автоматическое извлечение имен и описаний геномов.

Вследствие этого и ряда других причин формат аннотаций, представленных в GenBank, является полуструктурированным и не приспособлен для полностью автоматической обработки [10, 11].

Отмеченные недостатки организации структуры данных в GenBank потребовали значительных усилий для разработки автоматизированных средств импорта и обработки больших совокупностей полных геномов и их компонентов и, в свою очередь, являются основанием для совершенствования представления данных в GenBank, а также проведения исследований с целью выявления естественных компонентов геномов.

3. Картографическое представление геномов по их компонентам

Логическим развитием идеи представления нуклеотидных последовательностей с помощью характеристик строя явилась идея «картографирования» компонентов полных геномов (генов и других областей) посредством характеристик строя. При построении карты генов заданная характеристика полных геномов откладывается по оси x, а заданная характеристика компонентов (генов) – по оси y. Отмеченная точка $<x_i, y_j>$ на плоскости карты генов представляет j-й компонент i-го генома. При этом координата x_i точки соответствует значению характеристики i-го полного генома, а ее координата y_j – значению характеристики соответствующего j-го фрагмента этого генома. Таким образом, каждый геном оказывается представлен «столбцом» точек, каждая из которых представляет определенный компонент этого генома.

Кроме того, разработанные программные средства позволяют отображать геномы на карте по их порядковому номеру, полученному при сортировке геномов по той же характеристике. Это необходимо, если на карте два или более генома, например близкородственных видов, или штаммы одного вида микроорганизмов визуально неразличимы, так как имеют слишком близкое значение характеристик.

В данной работе построены две карты генов для полных геномов и плазмид с заданными перечнями их компонентов (таких как гены, рибосомальные РНК, транспортные РНК, некодирующие области, псевдогены и т.д.). На рис. 1 и 2 представлены фрагменты этих карт генов. На обоих рисунках в качестве характеристики полных последовательностей выбрана средняя удаленность g, а в качестве характеристик компонентов – глубина G. Единицами измерения как удаленности, так и глубины являются биты. По оси x $g < 1{,}5$ бит, по оси y $G < 8\,000$ бит. На рисунках выбранные компоненты геномов обозначены эллипсами, а на черном фоне показаны всплывающие подсказки со всей информацией о них.

Картографическое представление геномов множества организмов позволяет осуществлять неформальный экспертный анализ на предмет сходства отдельных их компонентов и, как следствие, организмов и геномов в целом. В свою очередь, хешированное представление компонентов и организмов позволяет частично автоматизировать и упростить экспертный анализ. Кроме того, с помощью карты генов возможно выборочно автоматизировать процесс сравнения геномов и их компонентов.

Картографирование компонентов (кодирующих и некодирующих последовательностей) геномов позволяет сравнивать их по различным характеристикам как внутри одного генома, так и в массивах геномов близкородственных микроорганизмов. Предлагаемый подход актуален для обнаружения кодирующих последовательностей при сравнении характеристик фрагментов геномов «de novo» с помощью библиотеки нуклеотидных последовательностей, доступных по адресу: http://foarlab.org/. Другим возможным направлением его применения может стать определение структурного и функционального назначения различных областей геномов.

4. Программная реализация интерактивной карты генов

Разработанная программная реализация карты генов содержит также интерактивные функции. Среди них динамическая фильтрация геномов и типов отображаемых компонентов. Для фильтрации компонентов достаточно отметить или убрать соответствующие «галочки» в их списке перед картой. Чтобы скрыть или показать все компоненты отдельного генома, достаточно выбрать его название в «легенде».

Для выбранного фрагмента генома на карте может отображаться дополнительная информация, включающая название генома (со ссылкой на его страницу в GenBank), тип фрагмента, атрибуты фрагмента и их значения, его позицию в полном геноме и длину, в случае кодирующих последовательностей – ссылка на страницу этой последовательности в GenBank, а также значения характеристик текущего компонента и полного генома. Также возможен автоматизированный и автоматический поиск схожих компонентов (по их характеристике) в заданном диапазоне сходства (с заданной точностью). При выборе определенного компонента выполняется поиск других компонентов, значение характеристики которых отличается от данного не более чем на величину заданной погрешности. При этом все «совпадающие» компоненты на карте приобретают форму эллипса, а их данные отображаются во «всплывающей» подсказке. Процедуру поиска можно представить как построение горизонтальной полосы, ширина которой соответствует заданной точности, а центр проходит через данный элемент; все элементы, оказавшиеся в полосе, считаются подобными при подтверждении с помощью данных GenBank.

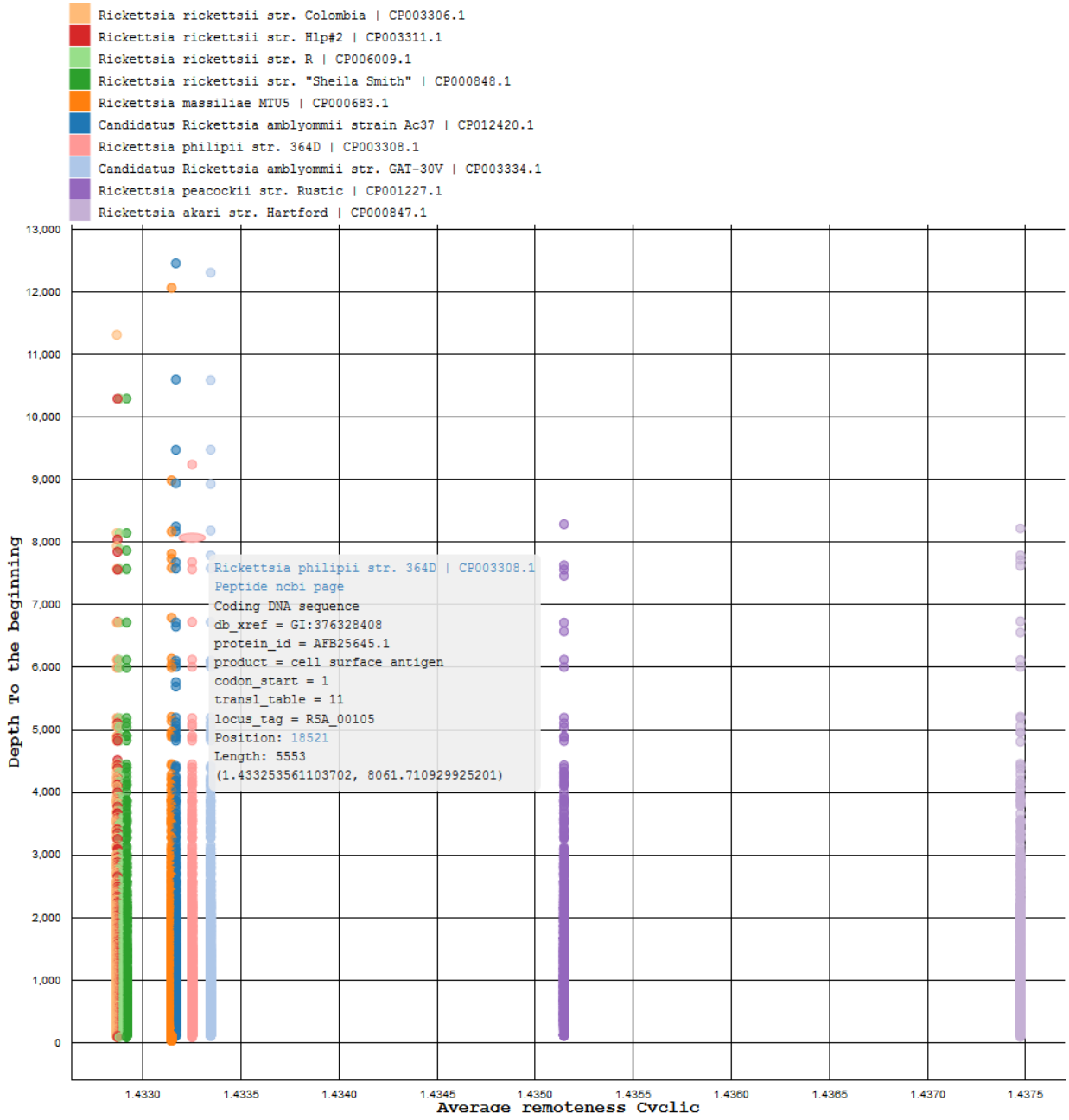

Рис. 1. Фрагмент карты генов полных геномов организмов семейства Rickettsia

Расшифровка всплывающей подсказки отдельного компонента (рис. 2):
Rickettsia felis URRWXCal2 plasmid pRF | CP000054.1 – название полной последовательности и ссылка на ее страницу в GenBank;
Peptide ncbi page – ссылка на страницу аминокислотной последовательности выбранного гена в GenBank;
Coding DNA sequence – тип выбранного фрагмента (кодирующая последовательность);
db_xref = GI:67005365 – номер выбранного фрагмента в GenBank;
protein_id = AAY62290.1 – id выбранного фрагмента в GenBank;
product = Conjugative transfer protein TraA_Ti – продукт выбранного фрагмента;
note = Possible nickase and helicase activities – пояснение;
codon_start = 1 – номер первого кодона фрагмента с которого начинается трансляция белка;
transl_table = 11 – таблица аминокислот;

locus_tag = RF_p39
Position: 36851 – позиция фрагмента в полной последовательности и ссылка на страницу фрагмента в GenBank;
Length: 2724 – длина выбранного фрагмента;
(1.4697611729042888, 4131.392884569981) – характеристика полной последовательности и выбранного фрагмента соответственно.

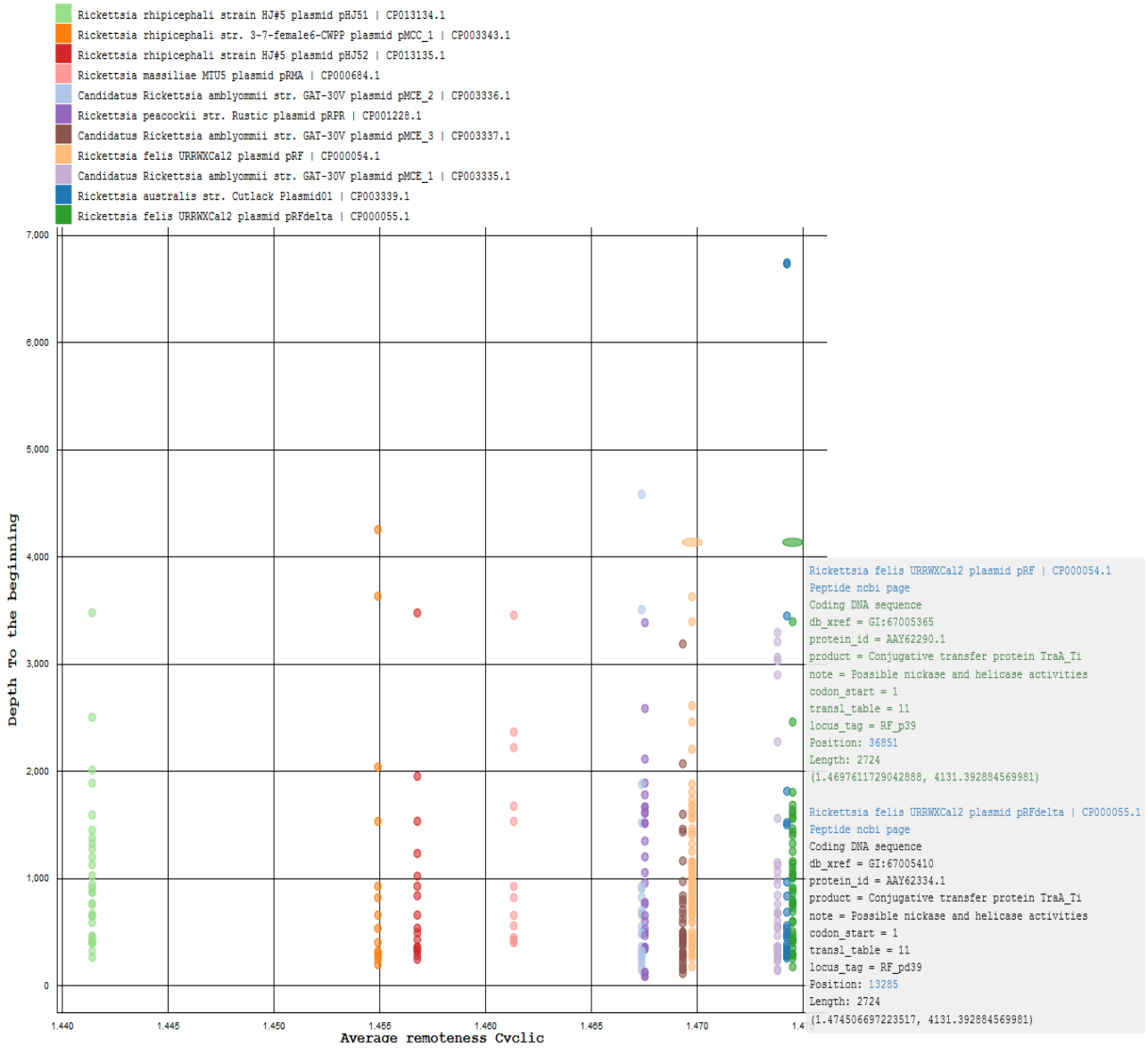

Рис. 2. Карта генов плазмид (внехромосомная ДНК) организмов семейства Rickettsia

Разработанное программное обеспечение позволяет формировать, сортировать и сравнивать компоненты изучаемой выборки геномов и плазмид по различным группам компонентов (кодирующая ДНК, рибосомальная РНК, транспортная РНК, псевдогены, некодирующие последовательности, повторяющиеся регионы и др.). Кроме того, для лучшего различения отдельных компонентов внутри генома карту можно масштабировать по вертикали. Анализ генов рибосомальной РНК в 38 геномах риккетсий и ориентий позволил получить представление о распределении нуклеотидов в последовательностях генов 5S, 16S и 23SРНК среди представителей рода Rickettsiaceae.

Заключение

В работе представлено два новых инструмента для описания и исследования нуклеотидных последовательностей. «Хеширование» с помощью числовых характеристик позволяет, во-первых, компактно и адекватно представлять длинные, в том числе полногеномные, последовательности, во-вторых, легко сравнивать множества таких последовательностей между собой и, наконец, быстро осуществлять их поиск.

Впервые удалось адекватно обозначить точками совокупности компонентов геномов (и плазмид) разных организмов и расположить их на плоскости «карты генов». Картографирование полных геномов по их компонентам (генам и др.) дает их наглядное представление и позволяет осуществлять неформальный экспертный и автоматизированный анализ, в том числе сравнивать и находить новые компоненты в полных геномах.

ЛИТЕРАТУРА

1. Gumenyuk A., Kostyshin A., Simonova S. An approach to the research of the structure of linguistic and musical texts // Glottometrics. 2002. No. 3. C. 61–69.
2. 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 (дата обращения: 15.04.2016).
3. Nair A.S.S., Mahalakshmi T. Visualization of genomic data using inter-nucleotide distance signals // Proceedings of IEEE Genomic Signal Processing. Bucharest, 2005. P. 11–13.
4. Afreixo V., Bastos C.A.C., Pinho A.J., Garcia S.P., Ferreira P.J.S.G. Genome analysis with inter-nucleotide distances. Bioinformatics. 2009. V. 25 (23). P. 3064–3070.
5. Shpynov S., Pozdnichenko N., Gumenuk A. Application of Formal Order Analysis (FOA) to Higher Order Grouping of Bacteria in the Genera Rickettsia and Orientia // Microbes and Infection. 2015. V. 17. P. 839–844.
6. Indyk P., Motwani R. Approximate nearest neighbors: towards removing the curse of dimensionality // Proc. of 30th STOC'98 Proceedings of the thirtieth annual ACM symposium on Theory of computing. 1998. P. 604–613. DOI: 10.1145/276698.276876.
7. Buldas A., Kroonmaa A., Laanoja R. Keyless Signatures' Infrastructure: How to Build Global Distributed Hash-Trees // Secure IT Systems. NordSec 2013 / N.H. Riis, D. Gollmann (eds.) // Lecture Notes in Computer Science. Berlin ; Heidelberg : Springer, 2013. V. 8208. P. 313–320.
8. Brinza D. et al. RAPID detection of gene–gene interactions in genome-wide association studies // Bioinformatics. 2010. V. 26 (22). P. 2856–2862. DOI:10.1093/bioinformatics/btq529.
9. NCBI Prokaryotic Genome Annotation Pipeline. URL: http://www.ncbi.nlm.nih.gov/genome/annotation_prok/ (access date: 15.04.2016).
10. GenBank Flat File Format. URL: http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html (access date: 15.04.2016).
11. The DDBJ/ENA/GenBank Feature Table Definition. URL: http://www.insdc.org/files/feature_table.html (access date: 15.04.2016).

Поступила в редакцию 4 ноября 2017 г.

Pozdnichenko N.N., Gumenyuk A.S. Shpynov S.N. (2018) MAP OF GENES – NEW TOOL FOR REPRESENTATION OF A SINGLE-CHROMOSOME GENOMES AND THEIR COMPONENTS. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie vychislitelnaja tehnika i informatika* [Tomsk State University Journal of Control and Computer Science]. 43. pp. 56–64

DOI: 10.17223/19988605/43/7

Previous works presented new approach – formal order analysis (FOA) based on Mazur's information theory and allowing describing and analyzing ordered data arrays of various nature (information chains). This approach directly takes into account arrangement of elements in sequences. Connections between elements of order (individual informations) are calculated as intervals between nearest similar elements (for nucleotide sequences this is inter-nucleotide distances). Multiplication of intervals gives number of descriptive informations. Binary logarithm of this value gives number of identifying informations. Characteristics calculated this way represents arrangement of elements in the whole object. Previously, order characteristics have been used in the study of genetic sequence for the following purposes: classification of

prokaryotes on levels of species, genus and family; classification of organisms at higher taxonomic levels; determination of the similarity of genetic sequences by comparison of the characteristics of distributions of congeneric sequences and using corresponding matrices; study of the local structure of the nucleotide sequences; search sequence fragments with the same order, etc.

The logical development of representation of the nucleotide sequences using the characteristics of order was the idea of a "mapping" of genomes and their components. In this case, the characteristic of whole genomes is plotted along the *x* axis, and along the *y* axis characteristic of their components is plotted; dot $<x_j, y_i>$ represents separate component of the genome. Cartographic representation of a set of organisms allows for expert analysis in order to search for the similarities of the individual components and consequently in the whole genomes.

Currently GenBank is the largest library of nucleotide sequences and in particular the whole genomes. Authors of sequences representing the complete genomes, upon upload of the sequence can give and its annotation or use the automatic annotation tool provided by GenBank. Such annotation includes information about the "location" of the different components in the genome. Two annotations presented for most genomes: one uploaded by authors and another automatically executed by GenBank annotation tool. Unfortunately, different annotations of one sequence can differ considerably, making it difficult to study and compare organisms by their components. Annotations presented in the GenBank are semi-structured and not adapted for completely automatic processing. In particular, each coding region is marked twice: as a CDS (coding sequence), and as a gene, while, for example, rRNA marked only once. Due to imperfections in both automatic and manual annotations, for many components their exact position and length are unknown and annotations of similar genomes are often marking very different lists of components, which also complicates the comparison of organisms using existing annotations.

Cartography of genome's components allows using of order characteristics as for comparison of components within the same genome, as within several genomes of closely related organisms. This approach is relevant for the detection and identification of (unnamed) coding regions and other important components. Another application of this approach can be definition of the functional and structural purpose of coding regions of the genome. This software allows one to filter, sort, and compare the sample of genomes and plasmids components into different groups.

Keywords: formal order analysis; inter-nucleotide distance; genes map; hashing with order characteristics.

POZDNICHENKO Nikolai Nikolaevich (Omsk State Technical University, Russian Federation).
E-mail: nick670@yandex.ru

GUMENYUK Alexander Stepanovich (Candidate of Technical Sciences, Associate Professor, Omsk State Technical University, Russian Federation).
E-mail: gumas45@mail.ru

SHPYNOV Stanislav Nikolaevich (Doctor of Medical Sciences, N.F. Gamaleya FRCEM, Moscow, Russian Federation).
E-mail: stan63@inbox.ru

REFERENCES

1. Gumenyuk, A., Kostyshin, A. & Simonova, S. (2002) An approach to the research of the structure of linguistic and musical texts. *Glottometrics*. 3. pp. 61–69.
2. Gumenuk, A.S., Pozdnichenko, N.N., Shpynov, S. N. & Rodionov, I.N. (2013) Formal analysis of structures of nucleotide chains. *Mathematical Biology & Bioinformatics – Mathematical Biology and Bioinformatics.* 8(1). pp. 373–397. [Online] Available from: http://www.matbio.org/article.php?journ_id=15&id=158. (Accessed: 15th April 2016). (In Russian). DOI: 10.17537/2013.8.373
3. Nair, A.S.S.S. & Mahalakshmi, T. (2005) Visualization of genomic data using inter-nucleotide distance signals. *Proceedings of IEEE Genomic Signal Processing*. DOI: 10.17537/2016.11.336
4. Afreixo, V., Bastos, C.A.C., Pinho, A.J., Garcia, S.P. & Ferreira, P.J.S.G. (2009) Genome analysis with inter-nucleotide distances. *Bioinformatics*. 25(23). pp. 3064–3070. DOI: 10.1093/bioinformatics/btp546
5. Shpynov, S., Pozdnichenko, N. & Gumenuk, A. (2015) Application of Formal Order Analysis (FOA) to Higher Order Grouping of Bacteria in the Genera Rickettsia and Orientia. *Microbes and Infection*. 17. pp. 839–844.
6. Indyk, P. & Motwani, R. (1998) Approximate nearest neighbors: towards removing the curse of dimensionality. *Proc. of 30th STOC'98 Proceedings of the thirtieth annual ACM symposium on Theory of computing*. pp. 604–613. DOI: 10.1145/276698.276876
7. Buldas, A., Kroonmaa, A. & Laanoja, R. (2013) Keyless Signatures' Infrastructure: How to Build Global Distributed Hash-Trees. In: Riis Nielson, H. & Gollmann, D. (eds) *Secure IT Systems. NordSec 2013. Lecture Notes in Computer Science*. Vol. 8208. Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-642-41488-6_21

8. Brinza, D. et al. (2010) RAPID detection of gene–gene interactions in genome-wide association studies. *Bioinformatics*. 26(22). pp. 2856–2862. DOI: 10.1093/bioinformatics/btq529
9. *NCBI Prokaryotic Genome Annotation Pipeline*. [Online] Available from: http://www.ncbi.nlm.nih.gov/genome/annotation_prok/. (Accessed: 15th April 2016).
10. *GenBank Flat File Format*. [Online] Available from: http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html. (Accessed: 15th April 2016).
11. *The DDBJ/ENA/GenBank Feature Table Definition*. [Online] Available from: http://www.insdc.org/files/feature_table.html. (Accessed: 15th April 201).