Algorithms and software for obtaining dissimilar order and high order of symbolic sequences

N N Pozdnichenko, M S Stupin, A S Gumennyuk, M S Doroshenko and O P Shafeeva

Omsk State Technical University, 11, Mira ave., Omsk, 644050, Russia

Abstract. Generally accepted methods for processing symbolic sequences existing nowadays do not take into account the arrangement of components in those sequences or do it in an inefficient way. In this paper, we propose the development of previously presented means of formal order analysis (FOA). In particular, two new methods of order transformation are proposed here: the secondary order (the high order or the order derivative) and the dissimilar order. First method takes into account not only the intervals between the individual elements, but also the arrangement of the intervals themselves. And the second method takes into account the order of elements occurrences. Algorithms and software for implementing the above transformations are presented. Software approbation is carried out, and adequacy for study of the arrangement of components in symbolic sequences (on the example of nucleotide sequences) is shown.

1. Introduction

In a number of works by other authors [1, 2, 3, 4], the intervals between the closest identical elements (in form of inter-nucleotide distance) were used to study nucleotide sequences. However, after extracting the intervals, the authors used the statistical apparatus and compared the truncated rank distribution of intervals. In some works even intervals of congeneric sequences were described. Nevertheless, the concept of the order of elements as a separate entity has not been defined, and no integral characteristic that allows describing separate nucleotide sequence by a single number has been defined. All the above makes the application of such approaches not very convenient in practice.

In works [5, 6, 7, 8, 9] methods of formal order analysis were presented, they provide means for the analysis of symbolic sequences of different nature, in particular nucleotide sequences. Previously various approaches, methods and software tools for processing symbolic sequences were developed, in particular: functions of order (sliding window local characteristics)[9, 10], matrixes of accordance and binary characteristics [7], rank distributions of congeneric sequences characteristics [7, 8], statistical characteristics analogs, characteristics of order analogous to the moments of random variables.

In a number of works, the mathematical properties of the order were considered [5, 6, 11]. However, all considered approaches use only intervals between identical components. This paper proposes the approach using the arrangement of intervals for obtaining the secondary and higher orders, and the approach taking into account the intervals between the n-th (first, second, etc.) occurrences of different elements to obtain an order on the basis of dissimilar sequences, hereinafter referred to as a dissimilar order.
2. The Formulation of the problem

The subject of this work is development of algorithms and software implementing extraction of a high order and a dissimilar order as well as their numeric characteristics with the subsequent approbation of their adequacy and effectiveness for the study of nucleotide sequences.

To obtain a secondary order or a higher order from the existing one, it is necessary to construct a sequence where each element is replaced by an interval either to the next same element or from the previous one, depending on the binding (link), and then to extract the order of such a sequence in the standard way [8].

To obtain a dissimilar order it is necessary to replace all n-th occurrences of the sequence elements with the current numbers of these occurrences, and then to extract the order of such a sequence in the standard way [8].

3. Algorithms and software development

3.1. Dissimilar order extraction algorithm

Two new methods of order transformation were developed in this work. First method extracts dissimilar order from symbolic sequence. To obtain dissimilar order one needs to replace each element with number of its occurrence in congeneric sequence.

In pseudocode form the procedure of dissimilar order extraction looks as follows:

Chain function dissimilar (Chain source)
{
    result = new Chain(source.length);
    entries = new int[source.Alphabet.Length];
    for (j = 0; j < source.Length; j++)
    {
        index = source.Alphabet.IndexOf(source[j]);
        entry = ++entries[index];
        result[j] = entry;
    }
    return (result);
}

Where «source» is a representation of symbolic sequence as object, storing an alphabet and an order of this sequence; «entries» is an array storing current number of occurrences for each element from alphabet; «result» is a resulting dissimilar sequence.

3.2. High order extraction algorithm

In this work we define the original order of symbolic sequences as primary order (or just order); derivative from the primary order as secondary order (or order derivative); derivative from the secondary order as tertiary order; etc.

| A | C | C | T | G | A | C | T | G | C | T | A | T | C | G | G | A | T | T | G | A | T | A | C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 2 | 3 | 1 | 3 | 2 | 4 | 4 | 1 | 3 | 3 | 4 | 1 | 3 | 1 | 2 |
| 1 | 1 | 2 | 1 | 1 | 2 | 3 | 2 | 2 | 4 | 3 | 3 | 4 | 5 | 3 | 4 | 4 | 5 | 6 | 5 | 5 | 7 | 6 | 6 |

Figure 1. The example of initial sequence, its order and its dissimilar order.

In figure 1 on the first line is an example of a nucleotide (symbolic) sequence, on the second line is its order, and on the last line is the resulting dissimilar order sequence. The alphabet of the initial sequence being {A, C, T, G}. 

3.2. High order extraction algorithm

In this work we define the original order of symbolic sequences as primary order (or just order); derivative from the primary order as secondary order (or order derivative); derivative from the secondary order as tertiary order; etc.
Thus, the second order transformation method produces secondary order or, in general, high order. To obtain a secondary order, it is necessary to extract an array of intervals between the nearest identical elements from each congeneric sequence, where one of the following bindings can be used: to the beginning, to the end, the cyclic binding to the beginning and the cyclic binding to the end. Binding defines the reading direction (bindings to the beginning correspond to the reading from the right to the left and bindings to the end correspond to the reading from the left to the right). Further, the elements of the order are replaced by the lengths of the corresponding intervals of congeneric chains.

To obtain a higher order, the same steps are applied cyclically to the results of the previous transformation.

In figure 2 on the first line there is an example of a nucleotide (symbolic) sequence, on the second line is its order, and on the last line is the resulting secondary order sequence extracted from the congeneric sequences using cyclic binding to the end, and the alphabet of the initial sequence being \{A, C, T, G\}.

In pseudocode form the procedure of high order extraction looks as follows:

```csharp
Chain function high order (Chain source, Link link)
{
    result = new Chain(source. length);
    entries = new int[source. Alphabet .Length];
    intervals = new int[source. Alphabet. Length][1];
    for (j = 0; j < source .Alphabet. Length; j++)
    {
        Intervals[i] = Extract Intervals( source . Congeneric Chain[i], link);
    }
    for (j = 0; j < source .Length; j++)
    {
        index = source. Alphabet.IndexOf(source[j]);
        entry = entries[index]++;
        interval – intervals[index][entry];
        result[j] = entry;
    }
    return (result);
}
```

Where «source» is a representation of symbolic sequence as an object, storing an alphabet and an order of this sequence; «link» is a binding of the intervals (reading direction) for high order; «entries» is an array storing current number of occurrences for each element from the alphabet; «intervals» is a two-dimensional array of intervals between the nearest similar elements; «result» is a resulting high order sequence. Intervals array is filled by «Extract Intervals» method, which extracts all the intervals for given congeneric sequence (sequence containing elements of only one type).

| A | C | C | T | G | A | C | T | G | C | T | A | T | C | G | A | T | T | G | A | T | A | C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|1 | 2 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 2 | 3 | 1 | 3 | 2 | 4 | 4 | 1 | 3 | 3 | 4 | 1 | 3 | 1 | 2 |
|5 | 1 | 4 | 4 | 6 | 3 | 3 | 6 | 4 | 2 | 5 | 5 | 10 | 1 | 4 | 4 | 1 | 3 | 9 | 2 | 6 | 2 | 2 |

**Figure 2.** The example of initial sequence, its order and its secondary order.

3.3. Software development

A software module was developed to obtain a high order and a dissimilar order. Also this module calculates the characteristics of these sequences. The developed module was integrated into the
program complex Libiada [12] In figure 3, 4 there are the examples of the interface of this software module.

| Program | Description | Date 1 | Date 2 |
|---------|-------------|--------|--------|
| Bovine viral diarrhea virus genotype 2 | | 27.03.2016 18:14:16 | 27.03.2016 18:27:16 |
| Macrosiphum euphorbiae virus 1 isolate MOS polyprotein gene | | 27.03.2016 18:14:23 | 27.03.2016 18:27:03 |
| Soybean cyst nematode virus 5 strain SoCV polyprotein gene | | 27.03.2016 18:14:32 | 27.03.2016 18:26:36 |

**Figure 3.** The High order and dissimilar order extraction interface.
Order transformation result

higher order To the beginning

Transformations repeated 2 times

Figure 4. The tertiary order in the application interface.

4. The software approbation

In order to test the adequacy and the informativeness of the characteristics of new representations for the description and study of nucleotide sequences (and other data sets), as well as the approbation of the software complex, a study of complete genomes from the representatives of the family Rickettsiaceae was conducted. As a reference distribution of organisms, the classification obtained in the study [13] was used.

In Table 1 thirty six studied organisms from the family Rickettsiaceae are presented, 34 of which are Rickettsia and two are Orientia.

Table 1. The secondary order characteristics of complete procariotic genomes of representatives of family rickettsiaceae.

| Sequence name | n  | g  | SD | H     |
|---------------|----|----|----|-------|
| R. prowazekii str. Katsinyian | NC_017050.1 | 88 | 1.67598359 | 3.19711214 |
| R. prowazekii str. BuV67-CWPP | NC_017056.1 | 88 | 1.67601146 | 3.19712179 |
| R. prowazekii str. Madrid E | AJ235269.1 | 88 | 1.67582687 | 3.19712541 |
| R. prowazekii str. GvV257 | NC_017048.1 | 88 | 1.67592452 | 3.19712833 |
| R. prowazekii Rp22 | CP001584.1 | 88 | 1.67593515 | 3.19713386 |
| R. prowazekii str. RpGvF24 | NC_017057.1 | 88 | 1.67581072 | 3.19715095 |
| R. prowazekii str. Chernikova | CP003391.1 | 88 | 1.67590190 | 3.19719181 |
| R. prowazekii str. Breinl | NC_020993.1 | 88 | 1.67572321 | 3.19723555 |
| R. typhi str. TH1527 | NC_017066.1 | 85 | 1.67263713 | 3.19796214 |
| R. typhi str. Wilmington | NC_006142.1 | 85 | 1.67246556 | 3.19796671 |
| R. bellii OSU 85 | NC_009883.1 | 87 | 1.64546795 | 3.20362886 |
| R. bellii RML369-C | NC_00840.1 | 87 | 1.64473158 | 3.20374837 |
| R. canadensis str. CA410 | NC_016929.1 | 90 | 1.64759932 | 3.20380032 |
| R. canadensis str. McKiel | NC_009879.1 | 88 | 1.64680584 | 3.20424057 |
| R. felis URRWXCal2 | NC_007109.1 | 87 | 1.63000272 | 3.20732618 |
| R. australis str. Cutlack | NC_017058.1 | 84 | 1.62661138 | 3.20839627 |
| Rickettsia Species                  | GenBank Accession | D2-HVR Length | Square Root of Average Remoteness Standard Deviation |
|-----------------------------------|------------------|---------------|---------------------------------------------------|
| R. japonica YH                    | NC_016050.1      | 81            | 1.62614132                                       |
| R. parkeri str. Portsmouth        | NC_017044.1      | 84            | 1.62486505                                       |
| R. slovaca str. D-CWPP            | NC_017065.1      | 86            | 1.62410819                                       |
| R. conorii str. Malish 7          | NC_003103.1      | 84            | 1.62468248                                       |
| R. slovaca 13-B                   | NC_016639.1      | 86            | 1.62407562                                       |
| R. rickettsii str. Arizona        | NC_016909.1      | 88            | 1.62408424                                       |
| R. rickettsii str. Iowa           | NZ_CP000766.1    | 88            | 1.62393110                                       |
| R. rickettsii str. Brazil         | NC_016913.1      | 88            | 1.62407860                                       |
| R. rickettsii str. Hino           | NC_016914.1      | 88            | 1.62349341                                       |
| R. rickettsii str. Hlp#2          | NC_016915.1      | 86            | 1.62332926                                       |
| R. rickettsii str. Colombia       | NC_016908.1      | 88            | 1.62341171                                       |
| R. rickettsii str. "Sheila Smith"| NC_009882.1      | 89            | 1.62369530                                       |
| R. massiliae MTU5                 | NC_009900.1      | 84            | 1.62279389                                       |
| Candidatus R. amblyommii str. GAT-30V | NC_017028.1    | 83            | 1.62417867                                       |
| R. philippii str. 36D             | NC_016930.1      | 88            | 1.62311406                                       |
| R. peacockii str. Rustic          | NC_012730.1      | 82            | 1.61984729                                       |
| R. akari str. Hartford            | CP000847.1       | 82            | 1.61912378                                       |
| Orientia tsutsugamushi str. Ikeda | AP008981.1       | 90            | 1.61471078                                       |
| Orientia tsutsugamushi str. Boryong | NC_009488.1    | 90            | 1.61340758                                       |

**Figure 5.** The Rickettsiaceae family complete genomes classification using secondary order average remoteness standard deviation characteristic.
5. Discussion
Out of eight computed characteristics of the secondary order of complete genomes, only two characteristics were informative and they fully confirmed the previously constructed classification of rickettsia into three groups. Moreover, the characteristic of the average remoteness standard deviation (\(g\ SD\)) made it possible to determine more distinctly the boundaries of the groups (Figure 5). And according to the characteristic of the entropy (H), the genomes were located in exactly the same order as in the original study (Figure 6).

Thus, the developed tools and the obtained characteristics give vast opportunities for both studying real nucleotide sequences and other information chains, and for revealing the mathematical patterns of such transformations and the arrangement of components in sequences as a whole. A test of convergence of a sequence of transformations of different orders to a constant value can be an example of a mathematical study of the obtained transformations.

6. Conclusion
The software extracting a high order and a dissimilar order is designed and implemented in this work. The calculators of order characteristics for obtained sequences are also developed. The approbation of the developed software was carried out. The informativeness of the calculated characteristics was verified by validating the previously conducted research. The perspective of the proposed approaches for further studies, both real sequences, and mathematical regularities was noted.

Figure 6. The Rickettsiaceae family complete genomes classification using secondary order entropy characteristic.
References

[1] Achuth Sankar S. Nair, T.Mahalakshmi Visualization of genomic data using inter-nucleotide distance signals. *Proceedings of IEEE Genomic Signal Processing*. Romania: Bucharest. 2005.

[2] Vera Afreixo, Carlos A. C. Bastos, Armando J. Pinho, Sara P. Garcia, and Paulo J. S. G. Ferreira. Genome analysis with inter-nucleotide distances. *Bioinformatics*, 25(23): pp. 3064–3070, 2009.

[3] Chang G, Wang T. Genome analysis with the conditional multinomial distribution profile. *Journal of Theoretical Biology*, V. **271**, I. 1, 2011, pp. 44–50. 10.1016/j.jbi.2010.11.034.

[4] Vera Afreixo, Carlos A.C. Bastos, Armando J. Pinho, Sara P. Garcia, Paulo J.S.G. Ferreira. Genome analysis with distance to the nearest dissimilar nucleotide. *Journal of Theoretical Biology*, V. **275**, I. 1, 2011, pp. 52–58. 10.1016/j.jbi.2011.01.038.

[5] Rodionov I.N., Semenov A.V. On the comparison of the two approaches to the study of genetic texts. *Mathematical Biology and Bioinformatics: III international conf.*, Pushchino: Reports. Edited by Lahno V.D. OOO «MAKS Press», 2008. – pp. 125-126.

[6] Semenov A.V. On one algorithm for generation of all possible orders for sequences with given parameters and a procedure for calculating their number. Information technology and control automation. *Materials of III scientific-practical conference OmSTU (regional).* – Omsk: OmSTU, 2011. – pp. 166 – 168.

[7] Gumeniuk A.S., Klikushin Iu.N., Kobenko V.Iu., Tsiganenko V.N. *Algorithms for the analysis of structure of signals and data*. Omsk. 2010. 272 p.

[8] Gumenuk A.S., Pozdnichenko N.N., Shpynov S. N., Rodionov I.N. Formal analysis of structures of nucleotide chains. *Mathematical Biology and Bioinformatics*. 2013. vol. 8. № 1. pp. 373–397. URL: http://www.matbio.org/article.php?journ_id=15&id=158

[9] Gumenuk A.S., Pozdnichenko N.N., Shpynov S. N. Formal analysis of order in the local structure of the nucleotide sequences. *Tomsk State University Journal of Control and Computer Science*. 2014. №4(29). pp. 23–30. URL: http://journals.tsu.ru/informatics/en/ &journal_page=archive&id=1119&article_id=17725

[10] Tsymbal V.S., Pozdnichenko N.N. The study of mean values and spread of local characteristics of order of the 16S and 18S ribosomal RNA of organisms from the three the realm of life. *Problems of the development, manufacture and operation of rocket and space technology and training of engineering personnel for the aerospace industry: Materials of IX All-Russian scientific conference dedicated to the memory of the chief designer of PA "Poljot" A.S. Klinyshkov* (Omsk, 17 February 2015). – Omsk: OmSTU, 2015. pp. 324-336.

[11] Semenov A.V. On the investigation of combinatorial properties of the order of the symbolic chains. *Modern Problems of Mathematics, Informatics and Bioinformatics. International conference, dedicated to the 100th anniversary of the birth of Alexei Andreevich Lyapunov, corresponding member of the USSR Academy of Sciences: report theses.* – Novosibirsk: Akademgorodok, 2011. – pp. 66.

[12] Skiba A.A., Pozdnichenko N.N. On the development of a software package for computing the characteristics of the structure of ordered data sets. Information technology and mathematical modeling: *Materials of XII All-Russian scientific-practical conference with international participation of A.F. Terpugov* (29-30 Novmber 2013). – Tomsk: Tomsk University, 2013. – V.1. – pp. 50-55.

[13] S. Shpynov, N. Pozdnichenko, A. Gumenuk Application of Formal Order Analysis (FOA) to Higher Order Grouping of Bacteria in the Genera Rickettsia and Orientia // *Microbes and Infection* 17 (2015) pp. 839-844.