Spectral Method for Detecting Inexact Repeats in Character Sequences

A. N. Pankratov[a],* and N. M. Pankratova[a],**

[a] *Institute of Mathematical Problems of Biology RAS – the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia*

** e-mail: pan@impb.ru*

*** e-mail: natpan1974@mail.ru*

Abstract—A method has been developed for detecting inexact repeats in character sequences that can use alphabets consisting of any number of randomly selected characters. Reduction of the problem of detecting repeats in character data to the problem of detecting repeats in continuous numerical sequences is proposed. A theorem is proved on the one-to-one representation of a character sequence by a bundle of continuous functions of a discrete argument. To detect and visualize repeats, a similarity dot matrix is used, which is calculated on the basis of a metric in the space of expansion coefficients of function bundle fragments. The problem of choosing the parameters of the method based on the estimation of the correlation integral is discussed.

Keywords: spectral-analytical method, Fourier series, similarity matrix, correlation integral

DOI: 10.1134/S1054661822030300

INTRODUCTION

Over the past decades, a huge number of extensive databases of various character sequences have been created in the world. In this connection, new tasks arise in linguistics (for example, anti-plagiarism), in bioinformatics (for example, evolutionary connections), in medicine (for example, genetic mutations and predisposition to any disease). Despite the rapid development of methods for analyzing large volume symbolic data, the task of creating an effective method for searching for inexact repeats in character sequences remains relevant. The disadvantage of many existing methods is the dependence on reading sequences that at least partially match the reference sequence. In particular, problems arise when long, relative to the reference genome, inserts are encountered in the genome. In [2], a method for searching for extended repeats and analyzing read sequences that do not match the standard is proposed. This method does not require prior knowledge of the genomic coordinates of extended repeats, detects changes in extended repeats, and can be applied to motifs of various lengths.

It is pointed out in [3] that many existing algorithms for searching for tandem repeats cannot cope with a high error rate when reading long sequences. The authors have developed new statistical methods for predicting the boundaries of a tandem repeat and determining the most likely repeating single copy – a fairly short k-measure. This algorithm aligns the representative unit with the input sequence using cyclic dynamic programming and evaluates the boundaries of repeats.

A computational method for detecting both exact and fuzzy tandem repeats was proposed by the authors of [7]. Their algorithm is based on the Ramanujan–Fourier transform and is used to detect periodicity in DNA sequences. A feature of this approach is that the transformation evaluates the period directly, without deriving it from the signal spectrum. In the opinion of the authors, this allows this method to be sensitive and effective.

Thus, despite the existence of a large number of existing methods for detecting repeats in character sequences, the task of constructing a fast and efficient method for analyzing long sequences remains relevant.

In this paper, we propose an algorithm for detecting inexact repeats in character sequences based on a generalized spectral-analytical method, which is a combination of numerical and analytical approaches to solving various problems associated with processing large data or image analysis. According to the chosen methodology, the original signals are represented by segments of orthogonal series, and data processing is carried out in the space of expansion coefficients. This approach has already been demonstrated in solving problems of searching for repeats in DNA sequences in [4–6].

1. THE THEOREM ON THE DECOMPOSITION OF A CHARACTER SEQUENCE

The basis for applying the spectral-analytical approach to character data is the transformation of a character sequence into some functional numerical

Received May 30, 2022; revised May 30, 2022; accepted May 30, 2022

representation suitable for applying data approximation methods. In this case, the following properties of the functional representation are important: (1) completeness and (2) continuity. The completeness of the description means that the original sequence can be reconstructed from the characteristic curves. Continuity is necessary to apply signal approximation methods. The fulfillment of these conditions in the case of character sequences is ensured by the fact that the content curves of character subsets are calculated in a window of a given length along the sequence. This type of curves includes the well-known and studied in bioinformatics GC-content curve. In this case, the window size, which is a parameter of such a description, actually introduces the concept of scale for the considered character sequence.

Theorem (on decomposition of a character sequence).

Let S be an arbitrary character sequence in the alphabet $A = \{a_1,\ldots,a_m\}$.

Then there exists a one-to-one representation of S as a bunch of $[\log_2 m]$ k-valued functions, where k is the scale parameter.

Proof. Let us encode the characters of the sequence with a numeric vector in the binary number system. It will take an integer number of bits, not less than the value of $\log_2 m$. Now consider a sliding window of width k, and sum up the number of ones of a certain bit of all characters in the sequence in this window. The function defined in this way, which depends on the beginning of the position of the window in the sequence, we will call the characteristic function of the sequence corresponding to the given bit of the binary character encoding. Each bit of each symbol of the sequence can be recovered from its corresponding characteristic function. In this case, the value of the characteristic function is equal to the number of characters in the encoding of which the corresponding bit contains one, i.e. is a function of the content of some subset of characters (at least half of the entire alphabet) in a window sliding along the sequence.

For example, in the case of genomic sequences given in the 4-letter alphabet {A, T, G, C}, 2-bit encoding can be used, thus achieving 4-fold compression of genomic files originally given in 8-bit encoding. In this case, the curves of the content of nucleotides {G, C} and {A, G} in a sliding window of length k can act as characteristic functions [4, 6]. For polypeptide sequences, a 20 letter alphabet is used {A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V}, including of the 26 letters of the English alphabet all except {B, J, O, U, X, Z}. To encode the character sequences of these alphabets, 5 characteristic functions will be required. In general, alphabets consisting of any number of arbitrarily chosen characters can be used.

2. SPECTRAL METHOD FOR DETECTING REPEATS

The characteristic curves that make up the description of the object are divided into overlapping fragments of length w with a step d. After that, a pairwise comparison of all fragments f_i, g_i, considered as discrete functions with numbering of samples within a window of length w, is performed based on the standard metric in Euclidean space:

$$\rho(f,g)^2 = (f-g, f-g) = \sum_{i=1}^{w} (f_i - g_i)^2.$$

To shorten the calculation of distances between fragments, the approximation of fragments of characteristic functions by segments of an orthonormal series is used. In this case, the distance is estimated using the formula:

$$\rho(f,g)^2 = \|c-d\|^2 = \sum_{i=1}^{n} (c_i - d_i)^2,$$

where $c = \{c_i\}$, $d = \{d_i\}$ are vectors of expansion coefficients in an orthonormal Fourier series, and n is their number (and, as a rule, $n \ll w$, but not necessarily). The use of spectral decomposition allows not only economical distance estimation, but also transformations to estimate inverted and complementary sequences in the space of decomposition coefficients, which means simultaneous recognition of all types of repeats without transforming the sequence itself [4].

To recognize repeats, a threshold decision rule is used: if $\rho < \varepsilon$, where ε is a threshold value, then the fragments are considered similar, and if $\rho \geq \varepsilon$, then the fragments are not similar. If there are several characteristic curves that make up a complete description of the object, recognition is carried out simultaneously, and the final result is a logical multiplication of the decision rules for each of the characteristic functions. This approach improves the stability of recognition to errors. This follows from the fact that the decision rule works in the region of the minima of the metric ρ, which is considered as a function of the fragment number. Thus, the set of minima determines the set of candidates for repeat. In the case of two features, for example, GC- and GA-curves, the set of repeats is taken as the intersection of sets of candidates for repeat, obtained for each of the features separately [4].

After carrying out these operations, the results of the comparison are displayed on a dot matrix, on which each point, however, corresponds to a comparison of two whole fragments, and not just sequence sites. The dot matrix is one of the visual standard representations of the results of comparing two sequences, which allows you to display the alignment of inexact repeats, as well as their relative position. The generalized dot matrix provides new opportunities for aligning inaccurate repeats. For example, it has been shown that an inexact extended tandem repeat can be displayed as a perfect square on a dot matrix. This is

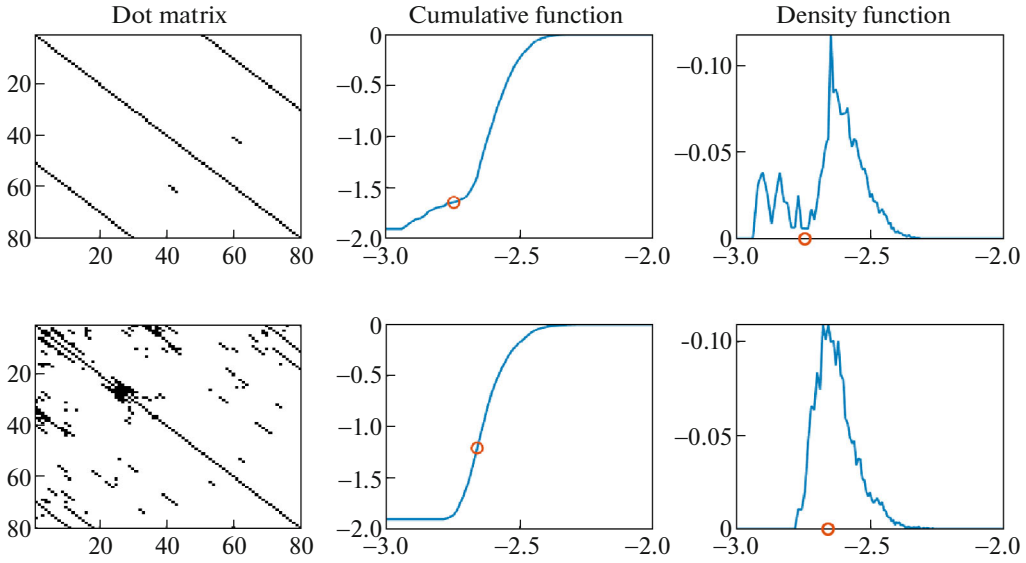

Fig. 1. Dot matrix and correlation integral as its cumulative function and density distribution from ε on a logarithmic scale for an inexact repeat (top row) and for a random sequence (bottom row). The marker denotes the chosen value of ε for which the dot matrix is constructed.

achieved due to the correct selection of the ratio between the window sizes and the approximation window step. Based on this important result, a fully automated method for recognizing tandem repeats was constructed and previously unknown repeats were found.

The structural scheme of the method looks like this:

(1) Preprocessing of a character sequence. At this stage, the formation of the initial alphabet takes place: the removal of unnecessary characters, the coding of the characters of the sequence.

(2) Transformation of a character sequence into a bundle of continuous characteristic functions.

(3) Transformation of characteristic functions into spectral representation. Unlike the previous steps, this stage may imply irreversible compression of information.

(4) Spectral comparison of fragments of the sequences.

(5) Display and analysis of the dot matrix in order to identify extended repeats, tandem repeats and study the relative position of repeats.

(6) Verification of repeats by one of the methods of discrete mathematics, for example, by dynamic programming or suffix tree.

Thus, an original method for detecting inexact repeats in signals of a character nature has been developed and presented. Compared to the classical approaches of discrete mathematics, this approach has a number of features:

(1) The transformation of a sequence into a signals bunch is performed by a sliding window, the width of which determines the scale at which the sequence will be processed. This allows the analysis of character sequences on a small scale, while the processing of sequences on a large scale must be carried out using methods of discrete mathematics.

(2) The spectral approach contains a threshold decision rule, i.e. based on finding inexact repeats. This determines the effectiveness of the method in relation to the methods of discrete mathematics, which are building inexact repeats through the finding of exact repeats.

3. ESTIMATE OF THE CORRELATION INTEGRAL

For an integral assessment of the quality of repeat detection, the correlation integral is used [1]:

$$C(\varepsilon) = \frac{1}{N^2}\sum_{i,j=1}^{N}\theta\left(\varepsilon - \left\|x_i - x_j\right\|\right),$$

where N is the number of fragments of functions for comparison, x_i is the vector of expansion coefficients of the ith fragment, and θ is the Heaviside step function.

In the example shown in the figure (Fig. 1), a sequence of 20000 characters in length in the alphabet of amino acid bases was considered. For the top row of pictures, the sequence consisted of 10000 random characters with the concatenation of the same sequence, but with 5000 mutations introduced into it. For the bottom row of pictures, all 20000 characters were random. The parameters of the calculations were the same in both cases: $k = 200$, $w = 20k$, $d = k$, $n = k$, with the exception of the value of ε. The value of ε was chosen as follows: for the first sequence, it corresponded to the minimum distribution density between two peaks of the distribution, and for the second sequence, it corresponded to the maximum distribution density. The evaluation of the correlation inte-

gral shows an almost perfect separation between the peaks corresponding to an inexact repeat (to the left of the optimal value of ε) and the bit peak responsible for the random component of the signal (to the right of the optimal value of ε). For the case of a completely random sequence, the distribution has a single peak. Thus, when searching for inexact repeats, a strategy is used that corresponds to a completely random sequence with random repeats displayed on a dot matrix, but if the parameters of the detected repeat are known, then the sensitivity selection strategy of the method can correspond to the first example, in which the optimal separating rule is found.

CONCLUSIONS

This paper shows the fundamental possibility of searching for repeats and the high computational efficiency of the proposed algorithms in the case of constructing dot matrices. The results matrices allow us to conclude that it is possible to further improve the quality of the method and its applicability to specific problems. It is possible to sign out significant advantages of the proposed solution to the problem of detecting inexact repeats in character sequences. Proposed method:

• smooths out local inexactness in signal repeats using integral estimation of these repeats;

• allows flexible alignment of inexact repeats by changing the size of the window and its step, i.e. scale selection;

• uses the spectral decomposition of signals, which leads to a significant reduction in calculations;

• possesses a high degree of vectorization and parallelization of calculations.

Thus, we can conclude that this method is effective for the analysis of long character sequences in order to detect or recognize extended inexact repeats.

COMPLIANCE WITH ETHICAL STANDARDS

This article is a completely original work of its authors, it has not been published before and will not be sent to other publications until the decision of the *PRIA* Editorial Board on not accepting it for publication is received.

Conflict of Interest

The authors declare that they have no conflicts of interest.

REFERENCES

1. P. Grassberger and I. Procaccia, "Characterization of strange attractors," Phys. Rev. Lett. **50**, 346–349 (1983). https://doi.org/10.1103/PhysRevLett.50.346
2. E. Dolzhenko, M. F. Bennett, P. A. Richmond, B. Trost, S. Chen, J. J. F. A. van Vugt, Ch. Nguyen, G. Narzisi, V. G. Gainullin, A. M. Gross, B. R. Lajoie, R. J. Taft, W. W. Wasserman, S. W. Scherer, J. H. Veldink, D. R. Bentley, R. K. C. Yuen, M. Bahlo, and M. A. Eberle, "ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data," Genome Biol. **21**, 102 (2020). https://doi.org/10.1186/s13059-020-02017-z
3. S. Morishita, K. Ichikawa, and E.W. Myers, "Finding long tandem repeats in long noisy reads," Bioinformatics **37**, 612–621 (2021). https://doi.org/10.1093/bioinformatics/btaa865
4. A. N. Pankratov, M. I. Pyatkov, R. K. Tetuev, N. N. Nazipova, and F. F. Dedus, "Search for extended repeats in genomes based on the spectral-analytical method", Math. Biol. Bioinf. **7**, 476–492 (2012). https://doi.org/10.17537/2012.7.476
5. A. N. Pankratov, R. K. Tetuev, M. I. Pyatkov, V. P. Toigildin, and N. N. Popova, "Spectral analytical method of recognition of inexact repeats in character sequences," Tr. Inst. Sist. Programm. Ross. Akad. Nauk **27**, 335–344 (2015). https://doi.org/10.15514/ISPRAS-2015-27(6)-21
6. M. I. Pyatkov and A. N. Pankratov, "SBARS: Fast creation of dotplots for DNA sequences on different scales using GA-, GC-content," Bioinformatics **30**, 1765–1766 (2014). https://doi.org/10.1093/bioinformatics/btu095
7. Y. Yadav, S. N. Sharma, and D. K. Shakya, "Detection of tandem repeats in DNA sequences using short-time Ramanujan Fourier transform," IEEE/ACM Trans. Comput. Biol. Bioinf. **19**, 1583–1591 (2021). https://doi.org/10.1109/TCBB.2021.3053656

Anton Nikolaevich Pankratov, born in 1972, graduated from the Faculty of Computational Mathematics and Cybernetics of Lomonosov Moscow State University (1994), Cand. Sci. (Computer Center of the Russian Academy of Sciences, 2004), Senior Researcher at the Institute of Mathematical Problems of Biology RAS – the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, specializes in the field of spectral methods and bioinformatics algorithms.

Natalia Mikhailovna Pankratova, born in 1974, graduated from the Faculty of Radiophysics of the Lobachevsky State University of Nizhny Novgorod (1996), PhD (Physics Faculty of Lomonosov Moscow State University, 2015), Researcher at the Institute of Mathematical Problems of Biology RAS – the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, specializes in the field of statistical methods and mathematical models in biology.