Spectral Analysis for Identification and Visualization of Repeats in Genetic Sequences

A. N. Pankratov, M. A. Gorchakov, F. F. Dedus, N. S. Dolotova, L. I. Kulikova, S. A. Makhortykh, N. N. Nazipova, D. A. Novikova, M. M. Olshevets, M. I. Pyatkov, V. R. Rudnev, R. K. Tetuev, and V. V. Filippov

Institute of Mathematical Problems of Biology, Russian Academy of Sciences,
ul. Institutskaia 4, Puschino, Moscow oblast, 142290 Russia
e-mail: pan@impb.ru

Abstract—The task of research on repeated segments in data sequences is considered in terms of genetic sequences. The principle of detection of repeats is offered based on comparison of specters of signal decomposition by classical orthogonal polynomials. The proposed approach can be applied in the search for extensive inexact repeats in different signals.

Key words: Direct and inverted repeats, orthogonal polynomials, spectral indexing.

DOI: 10.1134/S105466180904018X

INTRODUCTION

The search for repeats (homologies) in nucleotide sequences is one of the principal computational problems of bio-informatics. The textual similarity of genetic texts permits us to make hypotheses on their evolutional and functional similarity. Study of repeats can make a significant contribution to understanding of the structural–functional genetic organization, the most intriguing questions of which are the problem of information redundancy of genomes and the problem of information fragmentation in coding proteins.

Historically, the first method of determination of repeats in two sequences is the method of construction of a similarity dot matrix $M = (m_{ij})$ of two DNA-sequences (dot-matrices), where $m_{ij} = 0$, if the i-element of the first sequence is not equal to the j-element of the second sequence, and $m_{ij} = 1$, if the opposite is the case [1]. By means of this method, the researcher can identify parts of similarity of two sequences in the diagram. Continuous parts, consisting of units and parallels of the main diagonal, correspond to the parts of similarity sequences. Later, many different filters were invented to receive significant results and exclude single dots of noise in the picture. The simplest filter is scanning of diagonals by a window of specified length W and plotting of the diagonal segment with length W to the picture plane only when the specified number B of similarities was met within the window. Then it is referred to visualization of the parts with B/W-percent level of similarity [2]. A little different scheme of visualization is used for sequences specified in alphabets longer than four-letter alphabets, i.e., exact coincidence of characters becomes a rarer event. We apply weight matrices of character substitution for weighting of each possible character pairs, in unit fractions, but not only in zeroes and units. If it is rated in each replacement weight window so that they add up to W units, then it is possible to use the same filter; i.e., the window is considered suitable, if the sum of the character pairs rated by weight in this window exceeds B units. Thus, by implementing weight matrices of character replacement, it is not important to specify whether similar characters are on the intersection of the relevant vertical and horizontal. It is required just to put the relevant value from the weight matrix.

At one time this method was so popular that most of the multifunctional program packs for processing of genetic sequences obligatorily included construction of the homology dot matrix. For example, program pack SAMSON [3, 4] provided an additional service: direct, inverted, complementary–inverted, and complementary repeats were depicted in the same diagram with different colors. The parts of similarity were represented in a separate file in the form of pairwise equated sequences.

Some researchers [5–9] were engaged in solving of the problem of the picture noise with insignificant parts. The work of Reich and Meiske [10] is devoted to the problem of statistical significance of revealed hits (‘hit’ is the part of local similarity) in dot matrices. The function of distribution of the hit appearance random value is derived therein.

The amount of sequences subject to comparison has increased. Nowadays the improved methods of dot-matrix construction are applied for comparison of long (more than 100 kB) genome parts and whole genomes [11, 12]. Mainly, improvement consists in

Received March 18, 2009

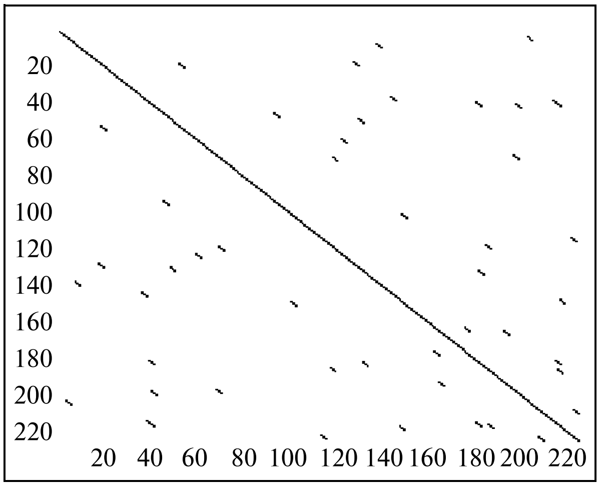

Fig. 1. Comparison of protein sequence using the dot matrix homology.

preliminary processing of compared sequences in order to reveal extensive blocks of similarity and use them in matrix construction.

In this work a method of construction of the similarity matrix of biological sequences is offered where similarity is estimated not by the similarity of characters, but by spectral characteristics of the G + C content profile. This method differs also in the fact that it is developed for extensive genome parts.

1. TASK SETTING

Nucleotide sequences represent texts of only four letters {A, T, G, C}. Complementary pairs are {A, T} and {G, C}. As the DNA-molecule consists of two antiparallel complementary strands, the same fragment is "registered" in DNA in two types. If a fragment is fixed in one strand, then this fragment is present in the complementary strand in complementary coding and inverted. For this reason during the search for repeats in genetic texts, complementary and inverted repeats can be present. Tandem and inverted repeats differ by location towards each other. The sequence inverted to itself is a palindrome.

The initial means of visualization of repeats is a dot-matrix homology, which can be represented in the form of a picture where the relevant matrix dot is filled if the letters in the sequence with coordinates thereof coincide. The repeat is the segment in the homology matrix parallel to the main matrix diagonal. Considering that the repeats, as a rule, are inexact, the task of repeat determination is significantly complicated. There are two types of correctable uncertainties in the compared sequences: mismatches of characters and character dropout of the sequence.

Let us consider the dot matrix homology in terms of the analysis of a protein amino acid sequence. To find concealed regularities in the primary structure of protein, a method is offered for creating the binary picture demonstrating how different segments of the protein are similar to each other. Any protein can be represented in the form of sequence specified in the alphabet of 20 amino acids marked as follows {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. To estimate the similarity of amino acids by characteristics, biologist use substitution matrices which can vary from task to task depending on the signs by which the amino acids are compared. There are many such weight substitution matrices. The most popular matrices are mutation data [13], block substitution matrices [14], and data matrices of space structure [15]. For each of these matrices, there are whole families reflecting extrapolations of protein evolutional consanguinity. The more similar amino acids x and z are, the larger the figure corresponding to the position of substitution matrix (x, z). In Fig. 1 the dot matrix homology is represented for a protein sequence with repeats of not less than three amino acids in succession according to some criterion of amino acid similarity.

From the mathematical point of view the concept of repeat appearing during comparison of two genetic sequences is formalized differently depending on the applied algorithm of sequences comparison. In this work the spectral principle of comparison of sequences' fragments is offered in order to reveal repeats, or, at least, candidates for repeat.

Spectral Method of Search for Repeats

The proposed algorithm for the search for repeats search is based on application of the generalized spectral-analytical method [16] to search for repeated structures in genomes [17]. The algorithm is conveniently divided into some relatively independent stages.

At the first stage of operation, the algorithm creates the function-profile of the DNA-sequence by the sliding window method wherein the calculation of the share of guanine G and cytosine C is carried out. This method of profile construction carries the physical sense of the connection of the DNA spiral, as the complementary connection of G and C arranged by three pairs of hydrogen bonds is stronger than the connection of A and T arranged by two pairs of hydrogen bonds. In addition, the constructed profile is invariant towards complementary substitutions in repeats. This solves the problem of the search for complementary repeats. The parameter of this stage is the width of statistical averaging of the sequence.

At the second stage, the function-profile is transferred to spectral representation using Chebyshev polynomial bases of the second kind or those of Legendre. This representation is used for comparison on the basis of some specially developed criterion at the next stage. It is necessary to note that use of approximate characteristics of the Chebyshev basis of the second kind in this task is appropriate as the weight func-

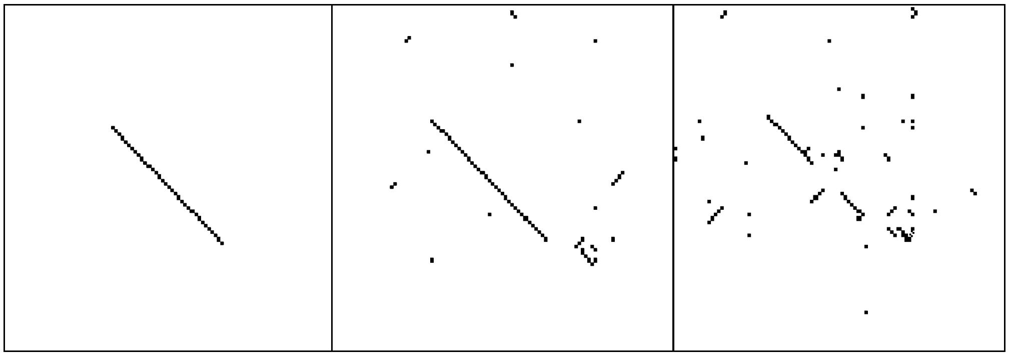

Fig. 2. Representation of direct repeat in the spectral homology matrix. Three matrices correspond to a sequential distortion of repeat of 0, 8, and 15% mutations. The repeat length is 1000 nucleotides.

tion of this basis is gradually tending to zero, which can compensate for wavering of repeat borders occurring at the stage of construction of the statistical profile. The parameters of this stage are the width of the window and the depth of spectral estimation. The calculations have shown that the method stably works if the width of the approximation window is two or more times greater than the window of statistical estimation.

At the third stage the criterion estimates the difference of vectors of the expansion coefficient. The criterion construction consists in selection of the metrics in the functional space and normalization thereof. Selection of the metrics is mainly conditioned by selection of the basis, its parameters, and decomposition depth. The rating is aimed at moving the characteristic invariant towards the signal value. Two variants of metric rating have been considered:

$$\theta_1(x(t), y(t)) = \frac{\|x(t) - y(t)\|}{\|x(t)\| + \|y(t)\|}$$

and

$$\theta_2(x(t), y(t)) = \frac{\|x(t) - y(t)\|}{1 + \|x(t) - y(t)\|},$$

where $x(t)$ and $y(t)$ are two compared functions corresponding to different parts of the functional profile, and $\|\cdot\|$ is the Euclidean norm in N-dimensional space of the signal decomposition coefficients. Both rates are suitable for identification of repeats as compared to the threshold values $\theta_1(x(t), y(t)) < \varepsilon_1$ and $\theta_2(x(t), y(t)) < \varepsilon_2$ becoming the method parameters.

At the fourth stage, on the basis of the obtained results of comparison, the matrix of spectral homologies (see Fig. 2) is constructed requiring analysis and interpretation of the results at the fifth stage.

So, five independent results of this algorithm are as follows:

profile construction;

its transfer to spectral representation (indexing);

construction of the decision rule for repeat identification;

comparison of indices for receiving the spectral homology matrix;

analysis of the matrix image.

Method Characteristics

Let us specify the main features of the offered algorithm characterizing its efficiency.

Application of indexing of the sequence permits us to speed up significantly the pairwise comparison of all fragments of the segment coverage.

Spectral representation permits us to get specters of the inverted pattern directly from the specter of the direct pattern, if the applied basis consists of even and uneven basis functions. So, the search for inverted repeats practically does not add computational complexity to the algorithm.

The metric is steady by the number of decomposition coefficients that provides the discriminant character to the comparison process. For example, if the comparison exceeds the threshold value by the first coefficients, then further calculation of the metrics is not obligatory.

The algorithm is completely based on calculations with the floating point and is vectorized and parallelized well.

RESULTS

For the simplest algorithm training, it is possible to use test sequences of different lengths containing known repeats. The program output is the matrix for which direct repeats are represented by diagonal segments parallel to the main diagonal, and inverted repeats are represented by diagonal segments located perpendicularly to the main diagonal. In order to differentiate the received repeats on the matrix of spectral homologies, they are depicted by different colors. Tandem repeats are marked by many repeats of both types fitted into the square.

It is noteworthy that due to the specific character of the method aimed at the search for extensive (more than 100 nucleotides long) divergent (broken as a result of mutation processes) repeats, the received results should be clearly represented. The most comfortable method of verification and visualization is equalization of repeated segments. It provides an idea of ways of their divergence and the degree of remaining similarity. The developed algorithm does not only look for exact repeats, but also candidates for repeat, because the algorithm deals with the function received from the nucleotide sequence. Though the DNA profile remains as conformity, verification of repeats represents the algorithm aimed at detection of literal conformity.

Testing of the algorithm has shown its stability during implementation of up to 15% of point mutations changing the profile to the researched sequence. In the case of increase in the number of mutations, noise and artifacts (see Fig. 2) appear conditioned by changing of the algorithm sensitivity.

A model of data representation is developed for the database of structural–functional elements of genomes, and a special database is designed and constructed. The database is implemented with a significant level of similarity that permits us to store genetic sequences and structural–functional elements of genomes of different organisms, both eukaryote and prokaryote. The modular, evolvable structure of the database permits storage of different types of structural elements of genomes and in perspective will provide the possibility to arrange the knowledge base. The format of input/output files is developed for the database on the basis of XML standards that provides possibility to supplement the database at once by a large number of various structural–functional elements. The database prototype is available on the Internet (http://www.jcbi.ru/bd/).

ACKNOWLEDGMENTS

The work was supported by the Russian Foundation for Basic Research, project nos. 07-01-00564, 08-01-12030, and 08-07-00353, and in part by Intel.

REFERENCES

1. A. J. Gibbs and G. A. McIntyre, "The Diagram, a Method for Comparing Sequences. Its Use with Amino Acid and Nucleotide Sequences," Ero. J. Biochem. **16**, 1–11 (1970).
2. R. Staden, "An Interactive Graphics Program for Comparing and Aligning Nucleic Acid and Amino Acid Sequences," Nucl. Asid Res. **10**, 2951–2961 (1982).
3. S. E. Vernoslov, A. S. Kondrashov, M. A. Roitberg, S. A. Shabalin, O. V. Yur'eva, and N. N. Nazipova, ""Samson" – a Software Package for Primary Biopolymer's Structures Analysis," Mol. Biologiya **24**, 524–529 (1990).
4. N. N. Nazipova, S. A. Shabalina, A. Yu. Ogurtsov, A. S. Kondrashov, M. A. Roytberg, G. V. Buryakov, and S. E. Vernoslov, "SAMSON: a Software Package for the Biopolymer Primary Structure Analysis," CABIOS **11** (4), 423–426 (1995).
5. A. D. McLachlan and D. R. Bosswell, "Confidence Limits for Homology in Protein or Gene Sequences. The c-*myc* Oncogene and Adenovirus Ela Protein," J. Mol. Biol. **185,** 39–49 (1985).
6. L. D. Brooks, B. S. Weir, and H. E. Schaffer, "The Probabilities of Similarities in DNA Sequence Comparisons," Genomics **3,** 207–216 (1988).
7. C. L. Queen and L. J. Korn, "Computer Analysis of Nucleic Acids and Proteins," in *Methods in Enzymology*, Ed. by L. Grossman and K. Moldave (Academic Press, New York, 1980), Vol. 65, pp. 595–609.
8. R. Arratia, L. Gordon, and M. S. Waterman, "An Extreme Value Theory for Sequence Matching," Ann. Stat. **14**, 971–993 (1985).
9. T. F. Smith, M. S. Waterman, and C. Burks, "The Statistical Distribution of Nucleic Acid Similarities," Nucleic Acids Res. **13**, 645–656 (1985).
10. J. G. Reich and W. Meiske, "A Simple Statistical Significance Test of Window Scores in Large Dot Matrices Obtained from Protein or Nucleic Acid Sequences," Comput. Appl. Biosci. **3** (1), No. 1, 25–30 (1987).
11. K. Szafranski, N. Jahn, and M. Platzer, "Tuple_Plot: Fast Pairwise Nucleotide Sequence Comparison with Noise Suppression," Bioinformatics **22** (15), 1917–1918 (2006).
12. A. Y. Ogurtsov, M. A. Roytberg, S. A. Shabalina, and A. S. Kondrashov, "OWEN: Aligning Long Collinear Regions of Genomes," Bioinformatics **18** (12), 1703–1704 (2002).
13. M. O. Dayhoff, W. C. Barker, and L. T. Hunt, "Establishing Homologies in Protein Sequences," Methods Ensymol. **91**, 524–545 (1983).
14. S. Heinkoff and J. Heinkoff, "Amino Acid Substitution Matrices from Protein Blocks," Proc. Natl. Acad. Sci. USA **89**, 10 915–10 919 (1992).
15. J. L. Risler, M. O. Delorme, H. Delacroix, and A. Henaut, "Amino Acid Substitutions in Structurally Related Proteins. A Pattern Recognition Approach. Determination of a New and Efficient Scoring Matrix," J. Mol. Biol. **204**, 1019–1029 (1988).
16. F. F. Dedus, S. A. Makhortykh, and M. N. Ustinin, "Generalized Spectral–Analytic Method in Information Processing Problems," Pattern Recognition and Image Analysis **12**, No. 4, 429–437 (2002).
17. F. F. Dedus, L. I. Kulikova, S. A. Makhortykh, N. N. Nazipova, A. N. Pankratov, and R. K. Tetuev, "Analytical Methods to Recognize the Recurring Structures in Genomes," Dokaldy Akademii Nauk **411** (5), 599–602 (2006).

Dar'ya Andreevna Novikova, born in 1985. Graduated from Faculty of Computational Mathematics and Cybernetics, Moscow State University in 2007, candidate; scientific interests: generalized spectral-analytical method, signal processing.

Natal'ya Sergeevna Dolotova, born in 1985. Graduated from Faculty of Computational Mathematics and Cybernetics, Moscow State University in 2007, candidate; scientific interests: identification and classification of biological objects.

Nafisa Nailovna Nazipova, born in 1960. Graduated from Faculty of Computational Mathematics and Cybernetics, Kazan State University in 1982, candidate of physical-mathematical science (2002), heads the laboratory of Bioinformatics in the Institute of Mathematical Problems of Biology, Russian Academy of Sciences. Scientific interests: bioinformatics, research on structural–functional arrangement of genetic sequences. Author of 57 publications including 12 articles in peer-reviewed journals, 2 chapters in collective monographs. She is the academic secretary of Scientific Board of Mathematical Biology, Russian Academy of Sciences, Secretary of the electron scientific periodical *Mathematical Biology and Bioinformatics.*

Maksim Ivanovich Pyatkov, born in 1984. Graduated from Ugra State University (2006), Master of Mathematics (2008), candidate, scientific interests: analysis of genome sequences, intellectual agents.

Sergei Aleksandrovich Makhortykh, born in 1963. Graduated from Faculty of Aerophysics and Space Research, Moscow Physical–Technical Institute in 1986, Candidate of Physical and Mathematical Science (1990), academic secretary of Institute of Mathematical Problems of Biology, Russian Academy of Sciences, head of the Data Processing Laboratory. Scientific interests: data processing, image identification, physical acoustics, biomedical annexes. Author of 70 articles, 2 monographs.

Mikhail Aleksandrovich Gorchakov, born in 1986. Graduated from Faculty of Computational Mathematics and Cybernetics, Moscow State University in 2008, candidate; scientific interests: applied mathematics, identification of the objects, artificial intelligence.

Anton Nikolaevich Pankratov, born in 1972. Graduated from Faculty of Computational Mathematics and Cybernetics, Moscow State University in 1994, Candidate of Physical and Mathematical Science (2004), senior research scientist of the Institute of Mathematical Problems of Biology, Russian Academy of Sciences. Scientific interests: spectral methods, mathematical modeling in biology, 13 publications.

Vitalii Vladimirovich Filippov, born in 1988. Graduated from Faculty of Computational Mathematics and Cybernetics, Moscow State University in 2008, candidate. Scientific interests: distributed, parallel, cluster systems.

Maksim Maksimovich Olshevets, born in 1972. Graduated from Faculty of Computational Mathematics and Cybernetics, Moscow State University in 1994. Scientific interests: image processing, processing of data on biomedical experiments, bioinformatics. Author of 7 publications.

Florents Fedorovich Dedus, born in 1927. Graduated in 1958 from Tool-Engineering Faculty, Mozhaiski Air Engineering Academy (Leningrad of the Order of the Red Banner), Doctor of Engineering Science (1991), professor of the Chair of Forecasting Mathematical Methods of the Faculty of Computational Mathematics and Cybernetics, Moscow State University. Scientific interests: image identification and signal analysis, processing of experimental data, combined numerical—analytical methods, control system.

Vladimir Removich Rudnev, born in 1983. Graduated from Tula Polytechnical University (2005). Candidate. Scientific interests: image identification, knowledge bases.

Lyudmila Ivanovna Kulikova, born in 1960. Graduated from Faculty of Computational Mathematics and Cybernetics, Kazan State University in 1982, candidate of physical—mathematical science (2007). Scientific interests: data processing, analytical decomposition, classical orthogonal polynomials, image identification.

Ruslan Kurmanbievich Tetuev, born in 1976. Graduated from Faculty of Applied Mathematics, Kabardino-Balkarian State University in 1999, candidate of physical—mathematical science (2007). Scientific interests: image identification and image analysis, experimental data analysis, spectral analysis.