ISSN 0278-6419, Moscow University Computational Mathematics and Cybernetics, 2007, Vol. 31, No. 2, pp. 49–53. © Allerton Press, Inc., 2007.
Original Russian Text © F.F. Dedus, L.I. Kulikova, S.A. Makhortykh, N.N. Nazipova, A.N. Pankratov, R.K. Tetuev, 2007, published in Vestnik Moskovskogo Universiteta. Vychislitel'naya Matematika i Kibernetika, 2007, No. 2, pp. 12–16.

Recognition of the Structural–Functional Organization of Genetic Sequences

F. F. Dedus, L. I. Kulikova, S. A. Makhortykh, N. N. Nazipova, A. N. Pankratov, and R. K. Tetuev

Department of the Methods of Mathematical Prediction, Faculty of Computational Mathematics and Cybernetics, Moscow State University, Leninskie Gory, Moscow, 119992 Russia

e-mail: radja@impb.ru

Received February 20, 2006

Abstract—The analysis problem for genetic sequences is formulated from the viewpoint of modern mathematical and information approaches. Statistical, correlative, entropic, and spectral approaches are considered. A brief introduction into the background of the problem is given. Genesis of a new interdisciplinary subject area, bioinformatics, a science considering application of computer methods in biological studies, is analyzed. New approaches to studying macromolecular systems on the basis of combined spectral–analytic technologies are proposed. Estimates of the algorithmic complexity of implementation of the proposed approaches are presented.

DOI: 10.3103/S0278641907020021

This paper is devoted to overcoming of serious obstacles arising during the development of new approaches to recognition of the structural–functional organization of genetic sequences.

Changes in molecular biology and genetics that occurred within a rather short period in the early XXI century, influenced the development of the subject areas related to biology. The starting event was discovery of the DNA duplex made by Watson and Crick in 1953 [1]. The next step was the development of the methods for decoding aminoacid and nucleotide sequences and determination of the spatial structures of these biopolymers. New investigation techniques have been developed and numerous discoveries have been made. Mass decoding of genomes started. This process was crowned with such an outstanding achievement as successful decoding of the human genome. Efficient methodical approaches that guarantee obtaining of fundamental knowledge of the molecular, genetic, cellular, and organismal life organization levels have been developed. Transformation of this knowledge for the needs of applied scientific fields has been realized.

It should be noted that the aforementioned technological breakthrough in molecular biology and genetics resulted from participation of physicists in biological investigations. The present state of these sciences is such that participation of mathematicians, especially mathematicians who have gained experience in recognition of signals in large data arrays, is required to understand and process the already obtained information and the new data that are currently gained with an increasing speed.

Total amounts of the source experimental data on only the molecular-genetic life organization level exceed several hundred terabytes (T = 1 tera units = 10^{12} units). In the last 20 years, molecular biology and genetics experience rapid information advance resulted from decoding of nucleotide sequences. Amounts of the obtained data are amazing.

For example, the length of the human genome is more than 3 billion base pairs and the entire genome contains more than 30 thousand genes. Decoding of the human genome delivers data on physical and citogenetic maps of chromosomes, chromosomes' nucleotide sequences, positions of individual genes, and mutations. The amount of these data is several tens of terabytes. At present, more than 1.5 million mutations, which introduce differences into individual human genomes, are revealed.

Structures of the genome DNA of several thousand viruses were decoded, for example, the genomes of yeast, fruit-fly, and some animals and plants. The aminoacid sequences of several million proteins and more than 15 thousand spatial structures of these sequences were also decoded. The technology of DNA chips can be used for simultaneous quantitative measurements of the expression of several ten thousand genes in a cell. Studies in the field of proteomics that are aimed at investigation of proteins and their interaction in living organisms are currently in progress. These investigations are based of the application of 2D gel electrophoresis, high-performance liquid chromatography, and mass spectroscopy. Experimental data are accumulated during investigation of the variety of human and

animal genomes. Similar amounts of experimental data are accumulated in such classical fields of biology as zoology, botany, taxonomy, and ecology.

Modern molecular biology became a source of unprecedentedly large amounts of experimental data whose understanding is impossible without application of modern information technologies and efficient mathematical data recognition (analysis) methods as well as modeling (prediction of behavior) of biological systems and processes.

All this explains the appearance of a new science, **bioinformatics**. The main subjects of investigation of bioinformatics are biological macromolecules (DNA, RNA, and proteins) and fundamental genetic processes (such as replication, transcription, and reparation). In recent years, studies on simulation of the operation of gene networks, which ensures procession of all functions of an organism, have appeared.

Bioinformatics belongs to high-technology fields of modern biology. Its main purpose is the development of information-computer and theoretical grounds of molecular biology, molecular genetics, genetic and protein engineering and selection, biotechnology, medical genetics, genodiagnostics, and genotherapy, the sciences whose outstanding achievements made biology one of the leading sciences of this century.

Bioinformatics occupies the key and extremely important position in modern biology. Its subject is investigation of biological systems at all levels of their organization: subcellular, cellular, supracellular, and organismal levels.

This paper is devoted to the analysis of one of bioinformatic problems: investigation of the structural–functional organization of genomes. The genome of a biological organism (the library of all genes of this organism) can be represented in the form of a linear sequence of symbols specified on an alphabet consisting of four letters. Each letter of this alphabet determines the type of the monomer appearing in the DNA molecules. It is known that genomes are strongly redundant, because the portion occupied by the segments coding the proteins and some products important for the organism is negligibly small as compared to the total length of the genome. Probably, the genome parts that do not encode the products currently familiar to biologists contain segments regulating various processes.

However, the function of non-encoding parts of genomes is still unknown. It is supposed that these segments are used in evolutional development of an organism and exactly these segments contain reserves for variability.

Redundancy of texts strongly correlates with such concept as repeatability. For a long period, the mechanisms and typical positions of simple and complex repeats in genomes, mechanisms used to keep homogeneity of these repeats, the character of selection of repeats in the non-transcribed (silent) part of the genome, and the rate of divergence inside the repeats of a particular type and inside taxonomic groups are widely discussed in the literature [2].

One of the methods used to study statistical properties of complete genomes and genome fragments is the measurement of the amount of information (complexity) and an associated parameter: redundancy. From the practical viewpoint, the values of complexity and redundancy can be used in the analysis of long symbol sequences.

The amount of information contained in genome sequences was estimated by calculating different characteristics: the classical information measure; the Kolmogorov's algorithmic complexity [3]; the Shannon entropy [4]; and the complexity considered in the sense of shrinking algorithms (for example, the Lempel–Ziv algorithm [5]). Such a characteristic as the length of the largest strictly repetitive sequence was also used [6]. Internal inhomogeneity of complete genomes has been demonstrated. The observed amounts of information in genome fragments are associated with the functional role of these fragments.

Recurring fragments of symbols (the so-called tandem repeats) in DNA and protein sequences occur rather frequently. Sometimes, these repeats, which have different lengths, are associated with particular functions (binding of the substratum and protein–protein interactions), spatial structures, or such features as flexibility of proteins or certain parts of the DNA. Short tandem repeats containing from 2 to 6 nucleotides (microsatellites) are probably used to regulate genes or to form the length polymorphism regions in a genome, which are used to identify a person or relationship. Longer repeats, which are referred to as variable number of tandem repeats (VNTRs), are potential recombination sites. Tandem repeats can also be associated with structural periodicity of symbol sequences. Structural periodicity in DNA and proteins undergoes evolutional divergence (variability) and with time looses the form of perfect repeats. It is not always possible to determine the cause of imperfect periodicity. The cause may be either duplication of individual fragments or convergence of the sequences directed by selection to attaining better functional properties. In the latter case, periodicity can be revealed due to not only the presence of dominating symbols in positions of the periodic unity but also on the basis of almost total absence of certain symbols in certain positions. Therefore, periodicity can often be revealed only after the analysis of the statistics of the symbol distribution over the period positions.

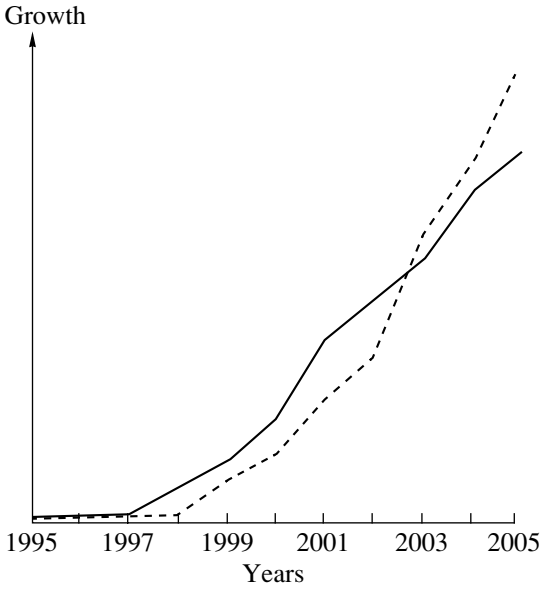
The search for the structural and statistical periodicity in the symbol sequences involves methods based on the Fourier analysis, correlation functions, and information theory. These methods require time-consuming computations and cannot be used to seek periodicity

distorted by insertions and deletions of symbols. Since analysis of the genome and protein sequences of different organisms and understanding of the obtained information are substantially slower processes than simple accumulation of decoded genetic sequences, development of simple and high-performance methods and algorithms for recognition of these sequences is still a very important problem.

Such methods are efficient for recognition of both microsatellite (containing from 2 to 6 symbols) repeats and longer tandem repeats. However, autocorrelation functions are weakly sensitive to imperfect (very diffuse) repeats. Therefore, the domain of applicability of these functions is limited by the structural or weakly diffuse periodicity. Application of the Fourier analysis requires expansion of the symbol sequence into a series reflecting the sequence order of the symbols of a certain type in the initial sequence. This procedure lowers sensitivity to longer periods if these periods contain periodicity in the distribution of the symbols of a particular type. The information decomposition (ID) method [7], which is based on calculation of the mutual information between an artificial periodic sequence and the analyzed sequence, has the best sensitivity to tandem repeats. This method is based on the analysis of the distribution of symbols in the $n \times m$ periodicity matrix, where n is the number of symbols in the alphabet and m is the length of the tested period. All mentioned methods require very time-consuming computations and cannot be used to seek periodicity distorted by insertions and deletions of symbols. At present, there are methods for seeking the tandem repeats with insertions and deletions; however, the basic principles of these methods either abandon the search for strongly diffuse tandem repeats [8] or require preliminary specification of the mask of coincident and noncoincident positions of the repeats [9].

It should be noted that, chronologically, the main stages of the history of formation and development of genetics and informatics are strongly correlated. It is also known that several important discoveries made in many fields of biology, especially in genetics, would be impossible without application of modern information technologies.

Tendencies in the development of modern computer technology demonstrate rapid (exponential) growth of main computational characteristics. This tendency is known as the Moore law [10]: each two years, capabilities of basic computer resources (the number of microchips, the speed, the memory capacity, etc.) are doubled. This law was the basis for several optimistic statements about solution in the nearest future of most bioinformatic problems with the use of available principles of processing and analysis of information collections.



Growths of (solid line) the number of microchips in modern microprocessors and (dashed line) the number of the GenBank's nucleotide bases shown on a common scale.

However, the growth of the currently obtained information collections that require processing and detailed analysis (genetic and protein sequences, etc.) is also exponential, and this exponent has larger index than that supposed by the researchers (see figure). In addition, the requirements for processing of biological data are liable to substantial qualitative and quantitative complications reflecting the development of the theory itself and the appearance of various working hypotheses in biology.

This circumstance caused a substantial lag of modern information technologies that placed them behind the demands made by some topical biological problems, especially problems related to processing of genetic sequences. A simple illustration may be the fact that, in spite of a substantial growth (as compared to the year 1997) of the throughput of communication channels and the capacity of the fixed memory, reception from the global network and storage of the genome of enteric bacillus *Escherichia coli* (about 4 MB), which was completely decoded by the year 1997, did not cause such problems as a similar operation performed five years later with the human genome (about 3 TB). This situation is further complicated by the fact that, at present, algorithms that have linear or sublinear complexity relative to the amount of input data have been proposed and implemented for only a very limited number of elementary information processing problems. As a typical example, we compare the estimates of algorithmic complexity in the problem of seeking the tandem repeats in a DNA. Calculations based on a simple enumeration of all possible variants of exact coinci-

dence of the repeats give quadratic algorithmic complexity $O(n^2)$, where n is the length of the sequence. If the genetic sequence contains a small number of repeats z relative to n, it is simpler to solve this problem with the Landau–Schmidt method [11], which has complexity $O(n\ln n + z)$. However, if we assume that the number of possible noncoincidences is equal to k, complexities of the aforementioned algorithms increase as $O(kn^2)$ and $O(kn\ln n + z)$, respectively.

The onset of investigations within the framework of the formulated project was accompanied by numerous discussions related to the choice of the mathematical apparatus that could be used as a basis for finding an acceptable solution of the problem of recognition of the structure of genetic sequences.

Continuous complication of most of applied problems (especially, microbiological problem, which operate with large data arrays of genetic sequences), digital data processing required in many cases, and overcoming of the cases of computational instability minimize advances in the field of increase of the computer speed. Therefore, introduction (in the appropriate cases) of analytic methods into the data processing procedures could substantially reduce and accelerate numerical calculations. Digital realization of analytic transformations during solution of problems is very difficult and inconvenient for the subsequent use.

Overcoming of these contradictions is possible only by means of application of combined digital–analytic technologies, which could ensure obtaining of acceptable (from the viewpoint of the speed and the accuracy of data processing) results for many applied problems.

The generalized spectral–analytic method (GSAM) described in [12] offers a natural way for integration of the discrete character of the source data arrays, the continuous nature of mathematical analysis and spectral representation, and the underlying variability and repeatability of the analyzed system. In the broad sense, reduction of the amounts of represented information and the amount of calculations required in solution of recognition problems is a consequence of the adaptive character of this approach. This adaptive character is based of the selective mathematical representation of information in a wide class of basis functions.

Investigations performed within the framework of this project are aimed at substantial lowering of the complexity of algorithms used for recognition of the structure and functional organization of genetic sequences. Attaining this goal requires development of new data processing principles that differ fundamentally from the existing approaches and, where possible, abandon application of the dynamic programming methods and combinatorial approaches. Advantages of the proposed methods are most pronounced in solution of the problems assuming inexact coincidences and substantial distortions of the genetic code.

Implementation of the proposed solution of this problem leads to an algorithmic complexity whose degree depends on the adopted number of noncoincidences (which influences the estimates of the obtained repeats). The results delivered by this algorithm are resistant to various possible nonlinear distortions of input data (such as insertions and omissions of the text fragments). Moreover, the proposed approach assumes the possibility of parallelization of calculations, a circumstance that results in further acceleration of the search for the tandem repeats in the case of implementation of algorithms with sublinear complexity on computational clusters. The strategy used here for the search for periodicity can be used for the search for not only tandem repeats but also spaced inexact repeats and is simpler than other approaches. An algorithm has been developed that seeks periodicity with the use of the spectral–analytic method during the analysis of the redundancy profiles and involves variation of the boundaries of the sought segment. The software implementation of this algorithm allows obtaining of acceptable results even at the initial stage of investigation. These results are not only comparable but, possibly, better than the results obtained with existing methods for recognition of the structural–functional organization of genetic sequences.

First experiments on the application of the combined computational technology instill confidence in the successful solution of the formulated recognition problems. However, the amount of the required investigations and development of the corresponding algorithms and the software require large time and labor efforts.

ACKNOWLEDGMENTS

This study was supported by the Russian Foundation for Basic Research, project nos. 04-07-90402, 06-01-08039, 06-07-89274, and 06-07-89303.

REFERENCES

1. J. D. Watson and F. H. C. Crick, "Molecular Structure of Deoxypentose Nucleic Acids," Nature, **171**, 737 (1953).
2. A. A. Aleksandrov, et al., *Computer Analysis of Genetic Texts* (Nauka, Moscow, 1990) [in Russian].
3. A. N. Kolmogorov, *Information Theory and Theory of Algorithms* (Nauka, Moscow, 1987) [in Russian].
4. C. E. Shannon, "A mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948).
5. A. Lempel and J. Ziv, "On the Complexity of Finite Sequences," IEEE Trans. Inf. Theory **22**, 75–81 (1976).

6. A. N. Gorban', E. M. Mirkes, T. G. Popova, and M. G. Sadovskii, "A New Approach to Studying the Statistical Properties of Genetic Sequences," Biofizika **38** (5), 762–767 (1993).
7. E. V. Korotkov and M. A. Korotkova, "Latent Periodicity of DNA Sequences from Some Human Gene Regions," DNA Seq. **5**, 353–358 (1995).
8. G. Benson, "Tandem Repeats Finder: a Program to Analyze DNA Sequences," Nucl. Acids Res. **27** (2), 573–580 (1999).
9. L. Noé and G. Kucherov, "Improved Hit Criteria for DNA Local Alignment," BMC Bionformatics **5**, 149 (2004) (http://www.biomedcentral.com/1471–2105/5/149).
10. G. E. Moore, "Cramming More Components Onto Integrated Circuits," Electron. Magazine **38** (8), 114–117 (1965).
11. D. Gusfield, *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology* (Cambridge Univ. Press, Cambridge, 1997; Nevskii Prospekt, St. Petersburg, 2003).
12. F. F. Dedus, S. A. Makhortykh, M. N. Ustinin, and A. F. Dedus, *Generalized Spectral–Analytical Method for Processing Information Collections* (Mashinostroenie, Moscow, 1999) [in Russian].