ISSN 1064–5624, Doklady Mathematics, 2006, Vol. 74, No. 3, pp. 926–929. © Pleiades Publishing, Inc., 2006.
Original Russian Text © F.F. Dedus, L.I. Kulikova, S.A. Makhortykh, N.N. Nazipova, A.N. Pankratov, R.K. Tetuev, 2006, published in Doklady Akademii Nauk, 2006, Vol. 411, No. 5, pp. 599–602.

COMPUTER SCIENCE

Analytical Recognition Methods for Repeated Structures in Genomes

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

Presented by Academician Yu.I. Zhuravlev April 26, 2006

Received May 17, 2006

DOI: 10.1134/S1064562406060354

This paper aims to overcome serious difficulties encountered in the development of new approaches to the recognition of repeated structures in genomes. Modern molecular biology produces unprecedentedly large amounts of experimental data, which cannot be comprehended without resorting to present-day information technologies and effective mathematical methods designed for data recognition (analysis) and the simulation (prediction of the behavior) of biological systems and processes. Thus, modern biology and genetics need the participation of mathematicians, especially those experienced in signal recognition while dealing with large datasets.

SEARCH FOR REPEATED STRUCTURES

The genome of a biological organism can be represented as a linear symbol sequence specified over an alphabet consisting of four letters: A, T, G, and C. Previously, the search for structural and statistical periodicity in symbol sequences was based on Fourier analysis methods [1], autocorrelation functions, information theory [2], and various methods based on statistical criteria. However, these methods are poorly suited for periodic sequences distorted by symbol insertions or removals, which motivates the development of new simple and highly effective search techniques.

One technique for studying the statistical properties of complete genomes and their fragments is based on quantifying the amount of information (complexity) and redundancy. The amount of information in genome sequences was estimated by various methods: the classical measure of information, Kolmogorov algorithmic complexity [3], Shannon entropy [4], and complexity in the sense of compression algorithms (for example, the Lempel–Ziv algorithm [5]). However, from a practical point of view for the sake of computational gain, we used a simpler quantitative estimate of genome structures, namely, the content (composition) of certain nucleotides within a sliding window. For example, for a window of length N sliding along a DNA sequence $S = \{S_1, S_2, \ldots, S_n, \ldots, S_L\}$, the $(G + C)$- and $(G + A)$-compositions in each ith position are represented by the values of the functions

$$f_i^{(G,C)} = \frac{1}{N}\sum_{n=i+1}^{i+N} \left|(S_n = G) \cup (S_n = C)\right|,$$

$$f_i^{(G,A)} = \frac{1}{N}\sum_{n=i+1}^{i+N} \left|(S_n = G) \cup (S_n = A)\right|,$$

where $i = 1, 2, \ldots, L - N + 1$.

In this paper, repeated fragments are sought as correspondences between distinct but similar segments of functions of the form

$$f_i = f_i^{(G,C)} \cdot f_i^{(G,A)}.$$

The difficulties associated with large amounts of data are overcome within the framework of the generalized spectral-analytical method [6] due to its adaptive characteristics. To within the required accuracy, the

Institute of Mathematical Problems in Biology, Russian Academy of Sciences, Institutskaya ul. 4, Pushchino, Moscow oblast, 142290 Russia
e-mail: ffdedus@impb.ru, kulikova@impb.ru, nnn@impb.ru, pan@impb.ru, radja@impb.ru

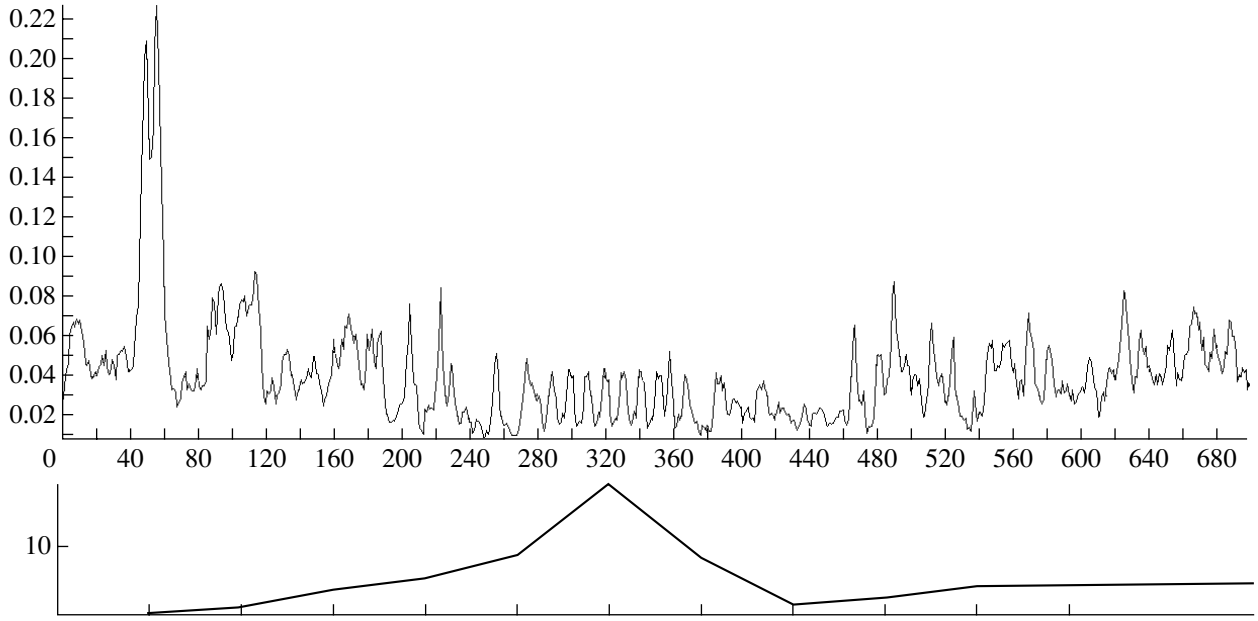

Fig. 1. Plot of f_i (upper panel) and the corresponding functional estimating the degree of recurrence (lower panel).

functions obtained are approximately expanded as series

$$f_i \approx \sum_{j=0}^{K} C_j \varphi_j(t_i)$$

in terms of a basis of classical orthogonal functions $\{\varphi_j(t)\}$ for $K \ll L - N + 1$.

Using this representation and transforming the expansion coefficients C_j, we can construct a function for estimating the similarity between different fragments of f_i.

As such a function, we can use a functional with expansion coefficients defined by C_j that reflects the degree of recurrence in f_i in the neighborhood of the ith position, as shown in Fig. 1.

Thus, we solve the problem of finding so-called tandem repeats, i.e., repeated symbol sequences in DNA that are directly adjacent to each other.

Compared with other approaches, this periodicity search strategy provides a more accurate technique that seeks not only tandem repeats but also dispersed approximate repeats. This is illustrated by two large dispersed fragments found in the Staphylococcus haemolyticus genome during a numerical experiment: from the 2495315th to 2498579th position and from the 2539025th to 2542289th position. For the search, we used the similarity between the expansion spectra of the corresponding segments in the curve shown in Fig. 2.

COMPUTATIONAL COMPLEXITY OF THE ALGORITHMS

As a typical example, we compare estimates for the algorithmic complexity in the problem of searching for tandem repeats in DNA. A simple exhaustive search through all possible versions of exact coincidence of repeats gives the quadratic algorithmic complexity $O(n^2)$ with respect to the sequence length n. When the number z of repeats in a genetic sequence is small as compared to n, this problem is frequently simpler to solve by the Landau–Schmidt method [7] with $O(n \ln n + z)$ complexity. However, if the number of possible mismatches in repeats is assumed to be k, the complexity of these algorithms increases to $O(kn^2)$ and $O(kn \ln n + z)$, respectively.

An implementation of the proposed approach leads to a linear algorithmic complexity that is independent of the assumed number of mismatches (this has an effect on the resulting estimates of repeats) and the results are stable with respect to various possible nonlinear distortions of input data, such as insertions and removals of text fragments. Moreover, the approach proposed implies can be parallelized, which would additionally accelerate the search for tandem repeats when the algorithms are implemented on computer clusters.

RESULTS

No satisfactory methods were previously available for detecting extended approximate tandem repeats. We have developed an effective promising method

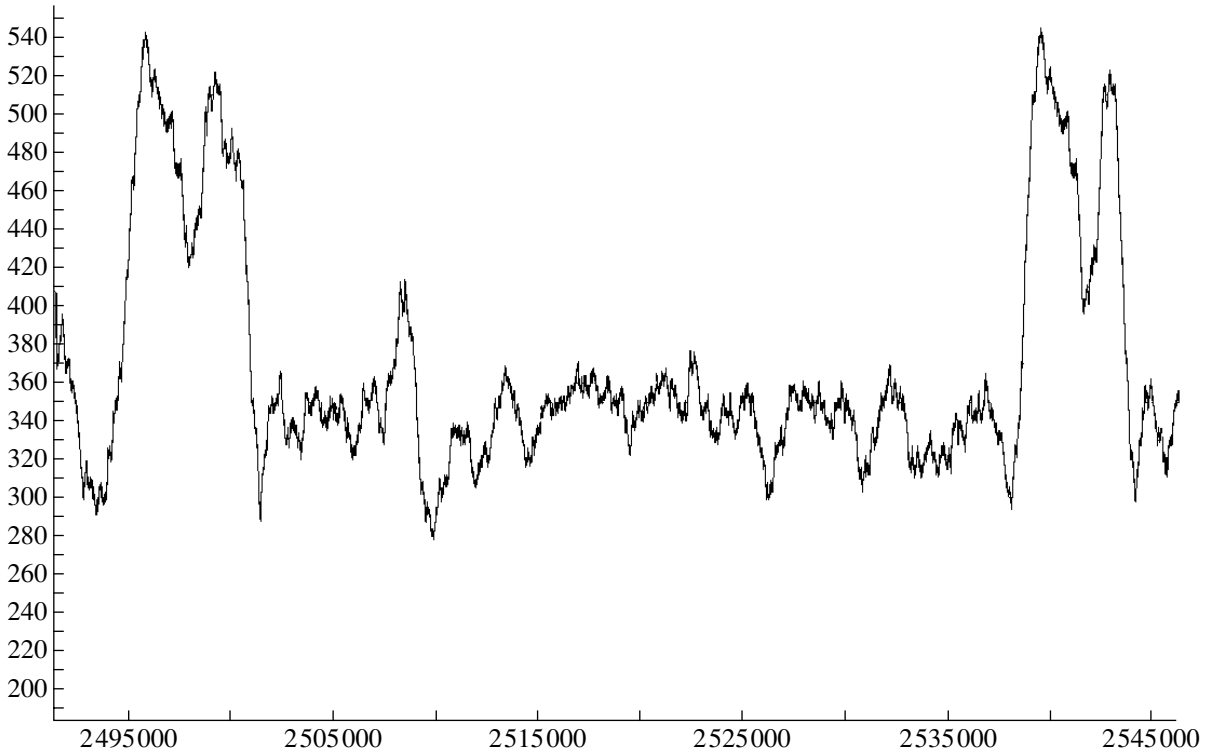

Fig. 2. Curve of change in the $(G + C)$-composition for a window length of 1000.

that combines numerical and analytical approaches. The advantages of this method are especially visual when it is applied to problems involving approximate concurrences and considerable distortions in genetic texts.

To test the method proposed, we implemented a software package based on the developed algorithms. As a test sequence, we used the Staphylococcus haemolyticus genome (GenBank AP006716), for which several tandem repeats have been previously detected (Table 1). The list of repeats produced by the software package was much more numerous than expected and included a number of new repeats in addition to those available in GenBank (Table 2). This suggests that our approach for analyzing genetic sequences is more effective in some problems than earlier developed methods.

The first experience gained in applying the combined computational technique allows us to hope for

Table 1. Tandem repeats for Staphylococcus haemolyticus available in GenBank

Coordinate of the beginning of repeat segment	Length of repeat	Multiplicity of repeat
328610	54	155
1182186	231	19
1502207	270	37

Table 2. Tandem repeats for Staphylococcus haemolyticus found by the proposed computational method

Coordinate of the beginning of repeat segment	Length of repeat	Multiplicity of repeat
44700	36	23
47386	108	15
96100	40	13
328610	54	155
1174866	231	16
1182186	231	19
1207631	141	3
1502207	270	37
1514160	270	8
2424573	72	7
2425484	258	8
2429172	90	7
2494500	274	3

a successful solution of the recognition problems posed.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project nos. 04-01-00756, 04-01-00814, 04-02-17368, 04-07-90402, 06-07-89274, 06-07-89303) and the Russian Science Support Foundation.

REFERENCES

1. D. Sharma, B. Issac, G. P. S. Raghava, and R. Ramaswamy, Bioinformatics **20**, 1405–1412 (2004).
2. E. V. Korotkov, M. A. Korotkova, and N. A. Kudryashov, Phys. Lett., No. 312, 198–210 (2003).
3. A. N. Kolmogorov, *Information Theory and Algorithmic Theory* (Nauka, Moscow, 1987) [in Russian].
4. C. E. Shannon, Bell Syst. Tech. J. **27**, 379–423 (1948).
5. A. Lempel and J. Ziv, IEEE Trans. Inf. Theory **24**, 530–536 (1976).
6. F. F. Dedus, S. A. Makhortykh, M. N. Ustinin, and A. F. Dedus, *Generalized Spectral–Analytical Method for Data Array Processing* (Mashinostroenie, Moscow, 1999) [in Russian].
7. D. Gusfield, *Algorithms on String, Trees, and Sequences* (Cambridge Univ. Press, New York, 1997), pp. 255–257.