WEB-server for search of a periodicity in amino acid and nucleotide sequences

F E Frenkel¹, K G Skryabin¹ and E V Korotkov¹,²
¹Institute of Bioengineering, Research Center of Biotechnology of the Russian Academy of Sciences, Leninsky Ave. 33, bld. 2, 119071, Moscow, Russia.
²National Research Nuclear University “MEPhI”, Kashirskoe shosse, 31. Moscow 115409, Russia.

gene korotkov@gmail.com

Abstract. A new web server (http://victoria.biengi.ac.ru/splinter/login.php) was designed and developed to search for periodicity in nucleotide and amino acid sequences. The web server operation is based upon a new mathematical method of searching for multiple alignments, which is founded on the position weight matrices optimization, as well as on implementation of the two-dimensional dynamic programming. This approach allows the construction of multiple alignments of the indistinctly similar amino acid and nucleotide sequences that accumulated more than 1.5 substitutions per a single amino acid or a nucleotide without performing the sequences paired comparisons. The article examines the principles of the web server operation and two examples of studying amino acid and nucleotide sequences, as well as information that could be obtained using the web server.

1. Introduction
At present, extensive sequencing of diverse genomes and determination of amino acid sequences for a variety of proteins are carried out [1]. All this leads to the fact that development of a complete mathematical model for regulating the cell genetic activity is becoming more and more realistic. Creation of such a model in some sense transfers a significant portion of bioengineering into the application area, since it will not be required to carry out molecular and biological experiments in order to create new species of both plants and animals. All this could be performed using mathematical simulation accompanied by subsequent implementation of the results of such simulation in the form of new genomes. However, there are significant difficulties along this path. The point is that at first it is required to find all the genes in the eukaryotes genomes, determine (annotate) the biological function thereof and distinguish the entire regulatory elements (promoters, enhancers and binding sites for various proteins), as well as to classify the regulatory elements. A purely experimental solution to this problem could be possible, but it would take relatively much time and would be rather expensive. Therefore, mathematical and algorithmic methods are used to solve this problem. The purpose of these operations is aimed at finding all regulatory sequences, determining the functions thereof and annotating all the genes. However, mathematical methods developed up till now are far from solving these problems. We are able to reliably annotate less than half of all the known genes; and, as far as the regulatory sequences are concerned, situation in this sense is even less
programming, which allowed us to find local alignments. This means that the boundaries of a region, which we will denote as the corresponding was applied for the periods of the average value and the multiplicity, the value of the algorithm \( F \) sequence. Then this random matrix was optimized in order to achieve the maximum value of the local alignment characterized by the highest value of the transformed in such a way that the distribution of the generated random matrices matrices are used to construct a local alignment of the studied where \( s \) is the length of a period, and \( k \) is the size of the alphabet for the \( S \) sequence. Afterwards, these matrices are used to construct a local alignment of the studied \( S \) sequence with respect to each of the generated random matrices [6]. Before starting the alignment construction, the matrices were transformed in such a way that the distribution of the \( F \) similarity function for random sequences was the same for all the random matrices used [6]. Then, a random matrix was selected that possessed a local alignment characterized by the highest value of the \( F \) similarity function; and we obtain the optimized \( M \) matrix. To ensure this optimization, a genetic algorithm [6] or a special optimization procedure is used [5]. After this, using the random mixing of the \( S \) sequence a multiplicity of random sequences was created. For each sequence out of this multiplicity, the value of \( \text{max}F \) was calculated, which allowed us to determine for \( \text{max}F \) the \( \text{max}F \) average value and the \( D(\text{max}F) \) dispersion, as well as the \( Z = \frac{\text{max}F}{\sqrt{D(\text{max}F)}} \). The algorithm was applied for the periods of the \( n \) length from 2 to 100; and for each length of a period we calculated the corresponding \( Z \) value. As a result of applying the algorithm, we receive the dependence of \( Z \) on \( n \), which we will denote as \( Z(n) \). It should be noted that in this study we used the method of dynamic programming, which allowed us to find local alignments. This means that the boundaries of a region,

2. Mathematical method
Previously, we proposed an approach for multiple alignment and corresponding PWM, which was devoid of the aforementioned shortcomings [5,6]. This method was implemented to search for tandem repeats. Briefly, the essence of this approach is as follows. We are looking for tandem repeats in the \( S \) sequence; and for this purpose we first generate a multiplicity of random matrices of the \( k \times n \) size, where \( n \) is the length of a period, and \( k \) is the size of the alphabet for the \( S \) sequence. Afterwards, these matrices are used to construct a local alignment of the studied \( S \) sequence with respect to each of the generated random matrices [6]. Before starting the alignment construction, the matrices were transformed in such a way that the distribution of the \( F \) similarity function for random sequences was the same for all the random matrices used [6]. Then, a random matrix was selected that possessed a local alignment characterized by the highest value of the \( F \) similarity function; and we obtain the optimized \( M \) matrix. To ensure this optimization, a genetic algorithm [6] or a special optimization procedure is used [5]. After this, using the random mixing of the \( S \) sequence a multiplicity of random sequences was created. For each sequence out of this multiplicity, the value of \( \text{max}F \) was calculated, which allowed us to determine for \( \text{max}F \) the \( \text{max}F \) average value and the \( D(\text{max}F) \) dispersion, as well as the \( Z = \frac{\text{max}F}{\sqrt{D(\text{max}F)}} \). The algorithm was applied for the periods of the \( n \) length from 2 to 100; and for each length of a period we calculated the corresponding \( Z \) value. As a result of applying the algorithm, we receive the dependence of \( Z \) on \( n \), which we will denote as \( Z(n) \). It should be noted that in this study we used the method of dynamic programming, which allowed us to find local alignments. This means that the boundaries of a region,
where \( \max F \) was obtained, could differ from the beginning and end of the sequence studied. This also means that the \( Z(n) \) values for different \( n \) could be obtained for different fragments of the sequence studied. Therefore, we would separately provide the boundaries of fragments, for which significant \( Z(n) \) values were obtained.

3. Examples of sequence analysis using Web-server

The algorithm was implemented in the form of software on the http://victoria.biengi.ac.ru/splinter/login.php web site. Let us consider several examples of the amino acid and nucleotide sequences analysis on this web server. The first sequence encodes the CEP250 centrosome-associated protein (Q9BV73 in Swiss-prot). For this sequence, the server created the \( Z(n) \) spectrum, which is shown in Fig. This Figure demonstrates that the periodicity possesses a period equal to 7 amino acids in the sequence; and, what is more, it was identified with insertions and deletions of amino acids. Using all the previously developed algorithms and software programs (such as REP[7], PTRStalker [8], TREKS [9] and many others) it turned to be impossible to identify this period. Table 1 shows the alignment fragment; and the optimized weight matrix created by the server is presented in Table 2. It could be seen from Table 1 that a period of 7 amino acids in length could only be obtained with insertions and deletions of symbols.

![Figure 1](image_url)

Figure 1. \( Z(n) \) dependence for the CEP250 centrosome-associated protein sequence (Q9BV73 in Swiss-prot).

Table 2 demonstrates that for each position of the period there appears a certain multiplicity of amino acids, which possesses some positive weight and a certain number of amino acids having the negative weight. This suggests that an extremely indistinct periodicity with the length of 7 symbols, which could be associated with formation of the \( \alpha \) helices in the given protein, was identified. The presence of a period with the length of 7 amino acids in the \( \alpha \) helical structures was demonstrated previously [10].
Table 1. Alignment fragment for n=7 in the Q9BV73 sequence. Entire alignment could be obtained by analyzing the Q9BV73 sequence at the http://victoria.biengi.ac.ru/splinter/login.php web site.

|       | 1     | 2     | 3     | 4     | 5     | 6     | 7     |
|-------|-------|-------|-------|-------|-------|-------|-------|
| K     | 1.9   | 0.4   | 2.0   | 0.4   | 0.3   | 1.8   | 0.4   |
| N     | 2.1   | 1.4   | 1.8   | 1.1   | -1.4  | 1.1   | 1.1   |
| I     | -2.3  | 0.1   | 2.3   | -1.7  | -1.7  | 0.7   | -0.5  |
| M     | -0.8  | -1.4  | 2.0   | 2.3   | -2.7  | -0.8  | -3.4  |
| T     | -0.5  | -0.0  | -2.5  | 0.8   | 1.2   | 0.4   | 1.7   |
| R     | -1.1  | 0.3   | -0.8  | 0.3   | 0.0   | -2.2  | 2.7   |
| S     | 2.0   | 0.9   | -1.4  | 1.6   | 2.1   | -3.3  | -1.0  |
| L     | -5.5  | -8.0  | 3.6   | -6.5  | -4.4  | 3.8   | -6.7  |
| Y     | -1.6  | 1.6   | 1.8   | -1.6  | -0.2  | 1.8   | 0.2   |
| F     | 0.7   | -1.6  | -0.5  | -0.5  | -0.5  | 0.7   | -0.5  |
| C     | -1.1  | -1.1  | -1.1  | 0.9   | 1.8   | 0.9   |       |
| W     | -1.4  | -1.4  | 1.9   | 1.4   | -2.3  | 1.7   | -1.4  |
| P     | -0.6  | 1.2   | -1.5  | -0.6  | 2.2   | -2.4  | -0.6  |
| H     | -1.3  | -0.2  | -1.8  | -1.3  | 1.7   | -0.7  | 2.1   |
| Q     | 2.7   | 2.6   | -6.4  | 2.3   | -2.3  | -3.6  | 3.0   |
| V     | -2.0  | -0.8  | 2.5   | 0.0   | 0.8   | -0.4  | -1.2  |
| A     | 2.3   | 1.8   | -1.9  | 1.0   | 2.3   | 0.2   | -4.2  |
| D     | -1.5  | 2.4   | -4.1  | -0.2  | 2.3   | -3.2  | 0.2   |
| E     | 2.4   | 3.1   | -8.3  | 2.9   | 2.7   | -7.9  | 2.7   |
| G     | 2.5   | -2.8  | 1.6   | -1.0  | -0.1  | -1.9  | -1.5  |

Table 2. Optimized M position-specific matrix, which was obtained by aligning the Q9BV73 sequence at the http://victoria.biengi.ac.ru/splinter/login.php web site.

| 1     | 2     | 3     | 4     | 5     |
|-------|-------|-------|-------|-------|
| K     | 1.9   | 0.4   | 2.0   | 0.4   |
| N     | -2.1  | -1.4  | 1.8   | 1.1   |
| I     | -2.3  | 0.1   | 2.3   | -1.7  |
| M     | -0.8  | -1.4  | 2.0   | 2.3   |
| T     | -0.5  | -0.0  | -2.5  | 0.8   |
| R     | -1.1  | 0.3   | -0.8  | 0.3   |
| S     | 2.0   | 0.9   | -1.4  | 1.6   |
| L     | -5.5  | -8.0  | 3.6   | -6.5  |
| Y     | -1.6  | 1.6   | 1.8   | -1.6  |
| F     | 0.7   | -1.6  | -0.5  | -0.5  |
| C     | -1.1  | -1.1  | -1.1  | 0.9   |
| W     | -1.4  | -1.4  | 1.9   | 1.4   |
| P     | -0.6  | 1.2   | -1.5  | -0.6  |
| H     | -1.3  | -0.2  | -1.8  | -1.3  |
| Q     | 2.7   | 2.6   | -6.4  | 2.3   |
| V     | -2.0  | -0.8  | 2.5   | 0.0   |
| A     | 2.3   | 1.8   | -1.9  | 1.0   |
| D     | -1.5  | 2.4   | -4.1  | -0.2  |
| E     | 2.4   | 3.1   | -8.3  | 2.9   |
| G     | 2.5   | -2.8  | 1.6   | -1.0  |
Figure 2. $Z(n)$ dependence for nucleotide sequence taken from the chromosome 1 A.thaliana genome (14259442-14259804 b.p.) [11].

Table 3. Nucleotide sequence taken from the chromosome 1 A.thaliana genome (14259442-14259804 b.p.) alignment for $n=10$.

| Nucleotide sequence taken from the chromosome 1 A.thaliana genome (14259442-14259804 b.p.) alignment for $n=10$. |
|---------------------------------------------------------------|
| 56789a12**3456789a123456789a123456789a123456789a123456789a1* |
| TTCGTTAACCAATCGGTTA***TCGGTTCTATTTCG***ATTTT*GTATTTTTTGGGT |
| **23456789a123456*789a123456789a123456789a123456789a1**23456 |
| AGATTTTTATATCCATTCCGTTCGGTTGTTGTCGCGTCGTTGCTGTCGTTGCTGTTG |
| 789a123456789a123****456789a123456789a123456789a123456789a12 |
| AGTATCCATTCCGGTTATTATCGTCTCCGGTTATTGGATAT**GTATTTTTTTGAGTA |
| 3456789a123456789a123456789a123456789a123456789a123456789a12 |
| TTTCAGATATTTTGGATATTTTGTGTTATTTTTCGTGAGTTATATTATATATATTATA |
| 3456789a123456789a123456789*89a123456789*89a123456789a123456789a |
| TAT**GGAATTTTTTCAAGTAGTATATTCCGGATATTCGATATCGATATCGGTCTT |
| 123456789a123456789a123456789a123456789a123456789a123456789a |
| TAT*TTCCGTGATATTTTCAAGTAGTATATTCCGGATATTCGATATCGATATCGGTCTT |
| 123456789a12 |
| TTTTTTTGGT |
Let us also consider, as an example, the study of the periodicity in the nucleotide sequence developed by the method. For this purpose, we once again examined the sequence taken from the chromosome 1 A.thaliana genome (14259442-14259804 b.p.) for \( n = 10 \). This work was supported by a grant from the Russian Science Foundation (14-24-00175).

**Table 4.** Optimized \( M \) position-specific matrix, which was obtained by aligning the sequence taken from the chromosome 1 A.thaliana genome (14259442-14259804 b.p.) for \( n = 10 \).

|   | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| A | -5.7| 11.4| -5.0| -1.1| -5.0| -4.4| -3.7| -1.1| 0.2 | 0.9 |
| T | 3.7 | -9.4| -0.2| -1.5| 4.2 | 2.4 | -6.3| -8.9 | -7.6| -3.3|
| C | -2.2| 2.7 | -0.3| -3.2| -4.1| -1.2| 11.4| -2.2 | -4.1| -4.1|
| G | -5.6| -5.6| -2.0| -4.2| -5.6| -4.9| -1.3| 10.9| 8.0 | 0.1 |

Periodicity of 11-11 symbols (as well as multiple periods) could be associated with the DNA packaging [19]. We believe that sections, which periodicity with insertions and deletions constitutes 9-11 symbols, may be associated with the nucleus binding regions or with the formation of chromatin loops [20]. Periodic sections with the length of period 2 could be referred to micro- and minisatellite sequences [21]. Classical micro- and minisatellite with insertions and deletions correspond to the \( Z > 12.0 \) values; at the same time, the number of substitutions constitutes less than 0.5 per a single nucleotide. That means that micro- and minisatellites found at the \( 6.0 < Z < 12.0 \) level appear to be the ancient copies of sequences that later were subjected to significant evolutionary changes. Thus, the given server could be used to search for indistinct micro- and minisatellites with the length of more than 100 symbols. These sequences are of great importance in searching for new VNTR, SSR or STR and creating the DNA markers. Similar results were also obtained for the D.melanogaster, C.elegans and H.sapiens genomes, which are presented in the database at the [http://victoria.biengi.ac.ru/cgi-bin/indelper/index.cgi](http://victoria.biengi.ac.ru/cgi-bin/indelper/index.cgi) web site [4].

Thus, the [http://victoria.biengi.ac.ru/splinter/login.php](http://victoria.biengi.ac.ru/splinter/login.php) web server is able to identify the latent periodicity (where the average number of substitutions is equal to 1.5 per symbol or more) with insertions and deletions of symbols; and in accordance with this parameter, the given web resource has no analogues at the present moment.

**Acknowledgement**

This work was supported by a grant from the Russian Science Foundation (14-24-00175).

**References**

[1] Heather J M and Chain B 2016 The sequence of sequencers: The history of sequencing DNA. *Genomics* **107** 1–8

[2] Pellegrini M 2015 Tandem Repeats in Proteins: Prediction Algorithms and Biological Role. *Front. Bioeng. Biotechnol.* **3** 143

[3] Pellegrini M, Renda M E and Vecchio A 2010 TRStalker: an efficient heuristic for finding fuzzy tandem repeats. *Bioinformatics* **26** i358-66

[4] Frenkel F E, Korotkova M A and Korotkov E V. 2017 Database of Periodic DNA Regions in Major Genomes *Biomed Res. Int.* **2017** 1–9

[5] Korotkov E V and Korotkova M A 2015 [Developing the Mathematical Method in order to Search for Latent Periodicity in Protein Amino Acid Sequences with Deletions and Insertions]. *Biofizika* **60** 1057–68

[6] Pugacheva V.M., Korotkov A.E and Korotkov E.V. 2016 Search of latent periodicity in amino acid sequences by means of genetic algorithm and dynamic programming *Stat. Appl.*
Genet. Mol. Biol. 15

[7] Andrade M a, Ponting C P, Gibson T J and Bork P 2000 Homology-based method for identification of protein repeats using statistical significance estimates. J. Mol. Biol. 298 521–37

[8] Pellegrini M 2012 Using phylogenetic profiles to predict functional relationships ed J Helden, A Toussaint and D Thieffry Methods Mol. Biol. 804 167–77

[9] Jorda J and Kajava A V. 2009 T-REKS: identification of Tandem REPeats in sequences with a K-meanS based algorithm Bioinformatics 25 2632–8

[10] Trifonov E N 1998 3-, 10.5-, 200- and 400-base periodicities in genome sequences Phys. A Stat. Mech. its Appl. 249 511–6

[11] Theologis A, Ecker J R, Palm C J, Federspiel N A, Kaul S, White O, Alonso J, Altafi H, Araujo R, Bowman C L, Brooks S Y, Buehler E, Chan A, Chao Q, Chen H, Cheuk R F, Chin C W, Chung M K, Conn L, Conway A B, Conway A R, Creasy T H, Dewar K, Dunn P, Egu P, Feldblyum T V., Feng J, Fong B, Fujii C Y, Gill J E, Goldsmith A D, Haas B, Hansen N F, Hughes B, Huizar L, Hunter J L, Jenkins J, Johnson-Hopson C, Khan S, Khaykin E, Kim C J, Koo H L, Kremenetskaia I, Kurtz D B, Kwan A, Lam B, Langin-Hooper S, Lee A, Lee J M, Lenz C A, Li J H, Li Y, Lin X, Liu S X, Liu Z A, Luors J S, Maiti R, Marziali A, Militscher J, Miranda M, Nguyen M, Nierman W C, Osborne B I, Pai G, Peterson J, Pham P K, Rizzo M, Rooney T, Rowley D, Sakano H, Salzberg S L, Schwartz J R, Shinn P, Southwick A M, Sun H, Tallon L J, Tambunga G, Toriumi M J, Town C D, Utterback T, Van Aken S, Vaysberg M, Vysotskaia V S, Walker M, Wu D, Yu G, Fraser C M, Venter J C and Davis R W 2000 Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana Nature 408 816–20

[12] Benson G 1999 Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27 573–80

[13] Kolpakov R, Bana G and Kucherov G 2003 mreps: Efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res. 31 3672–8

[14] Wexler Y, Yakhini Z, Kashy Y and Geiger D 2005 Finding approximate tandem repeats in genomic sequences. J. Comput. Biol. 12 928–42

[15] Betley J N, Frith M C, Graber J H, Choo S and Deshler J O 2002 A ubiquitous and conserved signal for RNA localization in chordates. Curr. Biol. 12 1756–61

[16] Kravatskaya G I, Kravatsky Y V, Chechetkin V R and Tumanyan V G 2011 Coexistence of different base periodicities in prokaryotic genomes as related to DNA curvature, supercoiling, and transcription. Genomics 98 223–31

[17] Lobzin V V. and Chechetkin V R 2000 Order and correlations in genomic DNA sequences. The spectral approach Uspekhi Fiz. Nauk 170 57

[18] Sosa D, Miramontes P, Li W, Mireles V, Bobadilla J R and José M V 2013 Periodic distribution of a putative nucleosome positioning motif in human, nonhuman primates, and archaea: mutual information analysis. Int. J. Genomics 2013 963956

[19] Herzl H, Weiss O and Trifonov E N 1999 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics 15 187–93

[20] Kadauke S and Blobel G A 2009 Chromatin loops in gene regulation. Biochim. Biophys. Acta 1789 17–25

[21] Richard G-F, Kerrest A and Dujon B 2008 Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes Microbiol. Mol. Biol. Rev. 72 686–727

[22] Ulloa M 2005 Molecular Markers Learning Modules, Vol. 1 and 2/ Volume 1. Using Molecular Marker Technology in Studies on Plant Genetic Diversity Crop Sci.