LEPSCAN— a web server for searching latent periodicity in DNA sequences

Andrew Shelenkov and Eugene Korotkov

Submitted: 7th April 2011; Received (in revised form): 18th June 2011

Abstract

A web server for searching latent periodicity based on the method of modified profile analysis has been developed. This method allows searching latent periodicity in presence of insertions and deletions. During searching process, the periodicity classes are used which were found by us earlier for various groups of organisms. Period length belongs to the range $2^{20}$ nt, not including the triplet periodicity. The results obtained are subjected to various filtration steps to ensure their statistical significance. Availability: The use of web server is free for non-commercial users. No registration is required. URL of the server is http://victoria.biengi.ac.ru/lepscan. Current software version is 1.06.

Keywords: sequence analysis; periodicity; DNA; in silico prediction; website review

INTRODUCTION

The presence of repeated sequences is a common phenomenon for both eukaryotic and prokaryotic genomes. It has been suggested that the repeats produce unusual physical structures in the DNA, causing polymerase slippage and the resulting amplification [1]. The other potential role for tandem repeats is gene regulation. In the last few years, tandem repeats have been increasingly recognized as markers of choice for genotyping a number of pathogens [2]. More recently, a number of studies [3, 4] have confirmed the idea that tandem repeats reminiscent of mini and microsatellites are likely to be a significant source of very informative markers for the identification of pathogenic bacteria. The rapid evolution of these structures appears to contribute to the phenotypic flexibility of pathogens.

Microsatellite repeats have been extensively used for genetic mapping and population studies [5]. They are frequently polymorphic with the number of repeat units varying between organisms. The polymorphism associated with tandem repeats has been used in mammalian genetics for the construction of genetic maps and still is the basis of DNA fingerprinting in forensic applications.

Among the best programs for finding the tandem repeats are Tandem Repeats Finder [6] and mreps [7]. Although they overcome some limitations of other algorithms, they have their own, namely, the size of the pattern (Tandem Repeat Finder) and too rigid pattern definition (mreps). Since the results of these methods depend strongly on the homology of the DNA regions, they can not be used to identify the fuzzy periodicity and ancient minisatellites [8]. Here, we propose a web server implementing the method of Modified Profile Analysis (MPA) developed by us. As we have shown earlier [8, 9], our method can reveal the periodical sequences that were not found by the existing software packages. A remarkable feature of this method is its ability to identify fuzzy or loose repeats (e.g. possible ancient microsatellites) that can not be revealed by other methods, including the ones based on Fourier transform and dynamic programming. Although our method is able to find some perfect tandem repeats, its main purpose is to identify latent periodicity in presence of insertions and deletions.

Andrew Shelenkov, Bioinformatics Department, Bioengineering Centre, Russian Academy of Sciences, Moscow, Russia. Tel: +7 499 1352161; Fax: +7 499 1350571; E-mail: fallandar@gmail.com

Andrew Shelenkov is a post-doctoral fellow at Bioengineering Centre, Russian Academy of Sciences. His research interests include DNA sequence analysis, with focus on new algorithms for periodicity and motif searching.

Eugene Korotkov is an Associate Professor at the Department of Applied Mathematics in Moscow Engineering Physics Institute and Principal Investigator in Bioinformatics Department of Bioengineering Centre, Russian Academy of Sciences. His research interests are in developing new methods of searching periodicity and motifs in various sequences, such as DNA, protein and linguistic sequences.
periodicity, not the perfect one. So some other methods could perform better when searching for perfect tandem repeats.

METHODS
The method implemented in LEPSCAN (LatEnt Periodicity SCANner) consists of three parts—usage of Information Decomposition (ID) method [10] to obtain the initial data, classification of the results found and the usage of MPA [8] for searching the sequences with insertions and deletions.

The primary method used for scanning the sequence in order to find the latent periodicity is MPA that allows revealing the periodic subsequences of arbitrary length in the presence of insertions and deletions. Periodicity type is defined by the nucleotide frequencies in various period positions.

The web server allows a user to perform a search of the latent periodicity which type belongs to a set of periodicity classes defined beforehand. These classes have been obtained by the authors for various organism groups and period lengths. In order to obtain them, GenBank has been scanned by the ID method [11] and the periodic sequences found have been classified. The use of periodicity classes rather than all periodic sequences found decreases dramatically the searching time for the MPA method. The advantages of using the combination of these methods are: no limitations are placed on the size of the sequence containing the repeated pattern, so the search becomes more versatile; usage of the classification matrices allows the method to find even distantly related sequences; although we use the alignment matrices, we do not fill them entirely, which speeds up the calculations even more. And the main advantage is that by using this combination of methods we can find the distantly related repeats and ancient minisatellites possessing very fuzzy periodicity.

Since the methods used have been thoroughly described in our previous works, we give only brief description of them below. The description of these methods is also presented in ‘Method’ section of the web site. The links for downloading the papers describing the methods used are also provided there.

To obtain the classes of periodicity, we obtain the frequency matrices using the method of ID and classify these matrices. ID is a spectrum representing the statistical significance of mutual information for periods of various lengths in the analyzed symbolical sequence. Mutual information between the sequence of interest and artificial symbolical periodic sequences can be used to obtain an ID spectrum. Let the sequence under consideration has a length \( l \). We generate random sequences possessing the periodicity with a period length equal to from 2 to \( l/2 \) using numbers as symbols. The artificial sequence with period length equal to \( n \) symbols can be presented as: 1, 2, 3, \ldots n; 1, 2, 3, \ldots n \ldots. Further, we can determine the mutual information between the analyzed sequence and each of the artificial periodic sequences. To do this, we fill the matrix \( M \) of size \((n \times k)\), where \( n \) shows the period length of the artificial periodic sequence used and \( k \) is the size of the alphabet of the sequence under study. The elements of this matrix are equal to the numbers of coincidences of \( ij \) \((i = 1, 2, \ldots n; j = 1, 2, \ldots k)\) type between sequences compared. \( L \) is the length of the analyzed symbolical sequence, \( x(i), i = 1, 2, \ldots n \) are the frequencies of symbols 1, 2, \ldots, \( n \) in the artificial periodic symbolical sequence; \( y(j), j = 1, 2, \ldots, k \) are the frequencies of symbols in the analyzed symbolical sequence. The value of the mutual information is calculated using formula

\[
I(n,k) = \sum_i \sum_j \frac{M(i,j) \ln M(i,j)}{L} - \sum_i x(i) \ln x(i) - \sum_j y(j) \ln y(j) + L \ln L
\]

The statistical significance of \( I(n,k) \) is estimated using the Monte–Carlo method by means of \( Z(n,k) \) calculation with formula

\[
Z(n,k) = \frac{I(n,k) - \overline{I}(n,k)}{\sqrt{D(I(n,k))}}
\]

Where \( \overline{I}(n,k) \) and \( D(I(n,k)) \) show the average value and deviation of the \( I(n,k) \) value, for a set of random matrices with the same sums \( x(i) \) and \( y(j) \) as in the initial matrix \( M(n,k) \).

The region of the sequence under study is considered to be periodic if the statistical significance \( Z \) for this region is greater than some threshold value (e.g. 4.0).

For each periodic region found by ID method, there defined a corresponding position-frequency matrix \( M \). This matrix can be considered as identifier of the found region. To make a classification easier, each such matrix was represented as a vector.
For example, a vector for a period length = 2 looked like:

\[
\begin{array}{cccccccc}
A_1 & T_1 & C_1 & C_1 & A_2 & T_2 & C_2 & G_2 \\
\end{array}
\]

Here, \( K_i (K = \{a, t, c, g\}, i = \{1, 2\}) \) represents the frequency (number of times the symbol is observed) of symbol \( K \) in \( i \)th position of the period for the region found.

General comparison scheme between the vectors was as follows. At first, we performed the pairwise comparison of all the initial periodicity vectors. For each pair of vectors, we calculated the Pearson statistics for which values are distributed as \( \chi^2 \). This statistics was a measure of vectors’ dissimilarity. While making a comparison, we took into consideration all cyclic and complementary permutations of vectors’ columns.

The minimal statistics value from the set of values described above was chosen for a compared pair of vectors. On a second step, we chose the two vectors for which the value of the Pearson statistics was minimal. If this value was corresponded to accidental probability of \( \leq 5\% \), then these two vectors were combined via recapitulation of their elements. A cyclic permutation, fixed inverse and complementary transformation were considered while making the vectors, combination.

After this, we returned to the first step, but we have excluded two merged vectors from the set and replaced them by one vector representing their combination. The process of vector comparison and merging had been continued until the minimal value of the Pearson statistics for the set of vectors became greater than the \( \chi^2 \) value corresponding to 5\% level. The vectors that were left up to this moment were considered as the periodicity classes. To make the significance of the classes higher, we exclude the classes containing less than three vectors in them.

Our web server searches for the periodicity having the type defined by the classes obtained earlier (i.e. with a frequency matrix similar to the one of the periodicity class).

We have used the dynamic profile alignment approach [8, 9], which takes into account divergence of sequences due to spot mutations and also indels. The method combines algorithms of dynamic programming for finding the best alignment with analysis of position specific nucleotides as in a profile. Since optimal alignment of analyzed sequence has been built against class matrix, that represents distribution of base weights over consensus positions, here we will use a term of ‘alignment against position-specific weight matrix (PSWM)’.

We have used a dynamic algorithm for finding local similarity, also known as the Smith–Waterman algorithm, to align the GenBank sequence being analyzed against PSWM.

Scanning of PSWM through sequence is carried out with a step of 20 bases. The periodicity class consensus was reproduced as many times as it was required to match the length of maximal subsequence. At each step an analyzed sequence has been aligned against PSWM.

To calculate the statistical significance, the Monte–Carlo method is used. Two hundred sequences with the same symbol composition (i.e. symbol frequencies and triplet correlation) as in original sequence are generated. We defined a Z-score as a measure of statistical significance. Z-score is the normalized deviation of the found sequence alignment weight from the average weight of random sequences alignment against PSWM:

\[
Z = \frac{W - M(W_{\text{rnd}})}{\sigma(W_{\text{rnd}})}
\]  

where \( W \), found alignment weight, \( W_{\text{rnd}} \), random sequence alignment weight, \( M \) and \( \sigma \), mean value and standard variance of \( W_{\text{rnd}} \), respectively.

We set up the minimum length of a periodical sequence to be found equal to 40 bases. All significant results revealed are filtered in order to exclude the overlapping sequences. If an overlapped region exceeded 30\% of the length of smaller of two overlapped alignments, only the alignment with the greater Z-score is left.

Since the method of MPA ensures only the statistical significance of sequence similarity, not its periodicity, we have to perform additional statistical test for the sequences found by this method. To do this, we calculate the ID spectrum for all the sequences found by MPA. Then the Monte–Carlo method is used to calculate the statistical significance. For each sequence 200 sequences are generated by randomly shuffling its symbols and then the mean value, variance and, finally, Z-score are calculated as it was described above. The sequence considered to be periodic with period length equal to \( P \) if the value corresponding to this length in ID spectrum is maximal and also is \( \geq 7.0 \). This test ensures that the sequences showing the given parameters possess the periodicity with a period length equal to \( P \) at statistically significant level.
IMPLEMENTATION

The users access the web server through the simplified web interface. Available interface languages include English and Russian. To make a request, a user must enter (copy and paste or upload a file) the sequence for performing periodicity search and must specify the e-mail address. After this other parameters, such as period length and organism group (see below), can be specified (or left blank) and then ‘Send request’ button should be hit. Default values for lower and upper borders of the period length are 2 and 20, respectively. Default value for the organism group is ‘all’. Search results will be delivered to the e-mail address specified usually within 24 h. The results will be presented as zip-archive containing the data regarding the found periodic regions of the original sequence entered by user. Subsequences will be located in separate files (filename format is PER[period_length]_final) within archive corresponding to their period length. Original sequence is also stored within archive in the file P000 for user reference.

Input sequence format can be either FASTA or Genbank. There is a link to third-party web site for sequence format conversion near the form for entering initial data.

A search of potential micro- and minisatellite sequences can be performed using the following user-specified parameters—organism group (corresponding to the groups used in Genbank) and period length (2–20, not divisible by three). Organism groups include all (union of other groups), bacteria, invertebrates, plant, primates, rodent, vertebrates, viruses.

The search is performed using the periodicity classes obtained for the corresponding period lengths and organism group. Our investigations have shown the presence of periodicity type specificity for the organism groups whose sequences have been analyzed. Nevertheless, to make a rigorous investigation it is recommended to perform the search for different groups. However this may result in a delay of receiving the results because calculation time would increase. There are no restrictions on input sequence length, but since the search is performed using 2000-nt window, the total length of periodic sequences revealed is limited to 2000. Also, inputting very long sequence (e.g. with a length >1 million bases) may result in calculation delay, so that the results may not be available in 24 h. The minimal length of a periodic sequence to be revealed was set to 40 bases to ensure the statistical significance of periodicity.

Periodicity of a period length divisible by three is a special case which lies beyond the scope of our investigations. LEPSCAN focuses on revealing potential micro- and minisatellites. Usually there is no way to clearly state if the triplet periodicity origin is concerned with gene structure or with presence of satellites. This is why we have decided not to investigate the periodicity of this type, so it is not supported by our web server.

The data for each periodic subsequence found includes: locus identifier, coordinate in sequence, coordinate in consensus, period length, nucleotide frequency matrix (a matrix of periodicity class which elements are equal to nucleotide frequencies for the corresponding period positions), an alignment of the found periodic region of original sequence against the periodicity class, statistical significance of the alignment (obtained using simulation modeling) and some service information.

Complete description of the formats is given in the ‘Help’ section of the web server. It is possible to return to server homepage (language selection) by clicking program logo located at the top of any other page.

An example of a sequence possessing latent periodicity with insertions and deletions is given below. This sequence is taken from the complete sequence of Pseudomonas aeruginosa PA14 pathogenicity island PAPI-1 [12]. Other programs for searching periodicity [6, 7] have revealed no periods in this sequence or around it (±200 nt upstream and downstream of the sequence).

**LOCUS AY273869**

**COORDINATE IN SEQUENCE** = 20947, 21116

**COORDINATE IN CONSENSUS** = 1, 172

**PERIOD LENGTH** = 5

**STATISTICAL SIGNIFICANCE** = 7.21

**NUCLEOTIDE FREQUENCY MATRIX:**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATRIX_a</td>
<td>12</td>
<td>14</td>
<td>6</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>MATRIX_t</td>
<td>0</td>
<td>6</td>
<td>13</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>MATRIX_c</td>
<td>0</td>
<td>19</td>
<td>7</td>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td>MATRIX_g</td>
<td>27</td>
<td>0</td>
<td>13</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>
ALIGNMENT OF SEQUENCE VS. CONSENSUS:

```
gcg-atccgcgcgcggcgttctgccgggtgc-tatcggct
gcggagcgcgcgggagcggacggaacgggagcggagcgggagcgg
ccctggagccctctcgag-acggcatgcgc-gacgtcttggg
gcggagcgcgcgggagcggacggaacgggagcggagcggagcgg
ctcagtaacctgcgttgctccggccgcgcaggttccttggccgcgc
 gcgcacaggtat cgctgcggcgcggcgcggcgcacgccccgcgcggagccgcgc
 gcggagcgcgcgggagcggacggaacgggagcggagcggagcgg
 ggtcgacGCCCGGCTAC
gagc-gggagcgagcgc (consensus is given bold-faced)
```

The approach used in our web server has been previously applied to searching for periodicity in GenBank, e.g. for searching potential microsatellites (period length 2–5) in plant [9] and bacterial [8] genomes. The results provided in these papers show that the approach developed allows to reveal five times more (e.g. 80 396 in comparison with 14 766) periodic sequences than application of ID, dynamic programming or Fourier transform [8, 9]. All the sequences have passed strict filtration to ensure statistical significance of their periodicity.

The method for searching periodicity in symbolic sequences developed by us appeared to reveal latent periodicity with insertions and deletions of symbols. We applied the Tandem Repeat Finder [6], which uses the dynamic programming for revealing periodicity, to the sequence above and it was unable to reveal periodicity in the sequence from the locus AY273869. This may be caused by the fact that usage of weight matrix for symbol matches strictly limits the ability to reveal latent periodicity. Thus it seems to be evident that the programs based on dynamic programming can be useful for detection of periodicity caused by the homology between separate periods.

It is also interesting to compare the results obtained by us with results obtained by using Fourier transform [10] and ID [10]. The results of Fourier transform application to the fragment of AY273869 sequence are shown in the Figures 1 and 2. From the Figure 1 it is clear that Fourier transform can not find the statistically significant period in the unaligned sequence. But if we put the insertions for the fragment of AY273869 sequence as it was made above, then Fourier transform reveals the period with length 2.5 nt. This does not correspond to the period that really exists in the sequence and is caused by the general disadvantage of Fourier transform [10, 13]. Fourier transform, as it was shown earlier [10, 13], spreads the statistical significance of the longer periods over the shorter periods. Let us show the example. Let we have the symbolic sequence obtained by repeating the period YRTDF 50 times. For this sequence we have five numerical sequences consisting of zero and one (according to the alphabet used). In this case, for the letters Y, R, D and F there will be obtained Fourier harmonics showing the period length for the symbolic sequence equal to six, and for the letter T there will be obtained the period length equal to three. This will lead to

**Figure 1:** Application of Fourier transform to searching for the periodicity in a fragment of AY273869 sequence shown above. Fourier transform was applied to the nucleotide sequence in a same way as in [13]. Nucleotide sequence was decomposed into four binary sequences; Fourier transform was applied to each of them individually and the intensity \( V(i,n) \) was calculated. To calculate the statistical significance, a source sequence was shuffled 1000 times. We calculated, for each period length and each binary sequence shown in the figure, the mean intensity value \( V(i,n) \) and its dispersion \( D(i,n) \), where \( i \) lies in a range from one to four (four binary sequences for a, t, c, g), \( n \) is a period length. As a result, the statistical significance was calculated as

\[
Z(i) = \left( V(i,n) - V(i,n) \right) / \sqrt{D(i,n)}
\]

\( Z(i) \) may be considered approximately as an argument of standard normal distribution for which the mean value equals zero and dispersion equals 1.0 [13]. Full Fourier transform for the sequence of DNA bases is a sum of four individual spectra, and the total statistical significance was calculated as

\[
Z = \frac{Z(1) + Z(2) + Z(3) + Z(4))}{2.0}
\]

Thus in the graphs of Fourier transform we show the statistical significance for various periods calculated using the Monte-Carlo method, depending on the period length.
decreasing of statistical significance for six-letter period by the value of significance of the revealed three-letter period. Such an effect will be greater for the greater ratio between the period length and the size of the alphabet used. So the statistical significance of the longer period ‘spreads’ over the statistical significance of the shorter periods, i.e. there exist an effect of fading the harmonics with longer periods in favor of the harmonics with shorter periods. This effect will be greater when some substitutions are made in the periodic sequence, and the sequence is not formed by simple duplication of identical periods anymore. As the result, Fourier transform shows the presence of non-existent periods in a sequence or does not reveal the periods at all due to either the insertions or deletions of nucleotides or to the statistical significance ‘spreading’ effect [11, 13].

The application of the ID to the fragment of AY273869 sequence is shown in the Figures 3 and 4. The ID also does not reveal the periodicity in unaligned sequence (Figure 3), but it reveals the statistically significant periodicity for the length equal to 5 nt in the aligned sequence (Figure 4). This shows that the approach used for building our web server combines the advantages of ID and dynamic programming that allows it to reveal the latent periodicity in nucleotide sequence in presence of nucleotide insertions and deletions. This is the advantage of the developed methods in comparison with dynamic programming, Fourier transform and ID.

LEPSCAN server has been tested in local network of Bioengineering Center and has proved its utility for tandem repeats’ and mobile elements’ annotation. We believe that our server can be a useful tool supplementing the existing web sites for searching tandem repeats. Because of the combination of methods used for its creation, our web server allows revealing the fuzzy repeats and potential minisatellites that can be omitted by other algorithms [8].
LEPSCAN is a web server for searching latent periodicity based on the method of MPA. This method allows searching latent periodicity in presence of insertions and deletions. Period length belongs to the range 2–20 nt, not including the triplet periodicity. The results obtained are subjected to various filtration steps to ensure their statistical significance. This website can be a useful tool supplementing the existing websites for searching tandem repeats.

References