Comparative analysis of periodicity search methods in DNA sequences

Yulia M. Suvorova a,*, Maria A. Korotkova b, Eugene V. Korotkov a, b

a Centre of Bioengineering Russian Academy of Sciences, Prospect 60-nya Oktyabrya 7/1, Moscow 117312, Russian Federation
b National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoe Shosse, 31, Moscow 115522, Russian Federation

A R T I C L E   I N F O

Article History:
Accepted 11 July 2014
Available online xxx

Keywords:
DNA sequence analysis
periodicity
information decomposition
Fourier analysis
correlation

A B S T R A C T

To determine the periodicity of a DNA sequence, different spectral approaches are applied (discrete Fourier transform (DFT), autocorrelation (CORR), information decomposition (ID), hybrid method (HYB), concept of spectral envelope for spectral analysis (SE), normalized autocorrelation (CORR_N) and profile analysis (PA). In this work, we investigated the possibility of finding the true period length, by depending on the average number of accumulated changes in DNA bases (PM) for the methods stated above. The results show that for periods with short length (<4 b.p.), it is possible to use the hybrid method (HYB), which combines properties of autocorrelation, Fourier transform, and information decomposition (ID). For larger period lengths (>4) with values of point mutation (PM) equal to 1.0 or more per one nucleotide, it is preferable to use information of decomposition method (ID), as the other spectral approaches cannot achieve correct determination of the period length present in the analyzed sequence.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

The studies of the first nucleotide sequences revealed that they contain various repeating motifs with periods of varying sizes. The period can be equal to a few base pairs (Fickett and Tung, 1992), several dozens of base pairs (for example, the period 10–11 base pairs associated with the structure of the DNA molecule (Herzel et al., 1999; Lobzin and Chechetkin, 2000) or it can be as long as repeating motifs in the genomes of warm-blooded vertebrates, the so-called isochor (Bernardi et al., 1985). The most famous periodicity is that where the period contains three DNA bases and such periods were found in coding DNA sequences (Fickett, 1982) Periodic properties of DNA sequences are often associated with biological properties (Lobzin and Chechetkin, 2000).

The study of periodicity of nucleotide sequences is hampered by the fact that most of the methods for determining periodicity, were traditionally designed to work with numerical sequences. This primarily refers to the spectral methods based on the Fourier transform (Silverman and Linsker, 1986; Tiwari et al., 1997) and wavelet transformations (Wang and Stein, 2010). In order to apply Fourier transform method to a DNA sequence, the nucleotide sequence must be converted to a numeric sequence. Different methods of achieving such a conversion have been proposed, for an overview, see (Afreixo et al., 2004). In addition, most methods offer one type of conversion of a DNA sequence. The concept of spectral envelope was created in order to overcome this drawback of application of the simple Fourier method, for the analysis of nucleotide sequences (Stoffer et al., 1993). In this case, recoding was chosen to optimize the separate performance of the transformation for each period p. This method was previously used for the analysis of DNA sequences (Howe and Song, 2013; Stoffer et al., 1993).

Another popular method applied in the analysis of DNA sequences is to search for periodicity in symbolic sequences, based on autocorrelation function (Bernaola-Galván et al., 2002; Herzel et al., 1998; Li, 1997). These methods may be applied with the help of translating the nucleotides into a numeric sequence or directly into the nucleotide sequence itself. In some cases, autocorrelation of the numeric sequence is calculated, then the Fourier transform is determined from the autocorrelation (de Sousa Vieira, 1999; Lee and Luo, 1997). Also, a hybrid method was proposed that combines the properties of Fourier transform and autocorrelation of nucleotide sequences (Epps et al., 2011; Epps, 2009).

To search for periodicity in the DNA sequence, methods based on dynamic programming were used (Benson, 1999). These approaches have been successfully used to find periodicity in DNA sequences, despite changes in the number of nucleotides resulting from deletions and insertions. To use these methods, the sequences should have a relatively high level of similarity between

* Corresponding author.
E-mail addresses: suvorovaya@gmail.com (Y.M. Suvorova), bioinfo@rambler.ru (M.A. Korotkova), geneekorotov@gmail.com (E.V. Korotkov).
the individual periods. This limitation is due to the fact that weight matrix for pairs of nucleotides is used by dynamic programming for periodicity search. The using of such matrix on the one hand gives to these methods the possibility of detection the periodicity in a presence of insertions or deletions of nucleotides, and on the other hand does not allow to find so-called latent periodicity (Korotkov et al., 2003). In this study, we are interested in the use of mathematical methods to detect periodicity with maximum degree of divergence in DNA bases. Methods based on dynamic programming will be worse than the spectral approaches; thus, we compared spectral methods only.

There are several terms that describe the DNA sequence periodicity. They can be roughly divided according to the number of point mutations (PM). The term “tandem repeats” (Benson, 1999) means that the same sequence pattern repeats several times. For example, it is the sequence (atgattc)30. In this case there is a complete similarity between any two periods of the sequence. When the number of PM is small the terms “diverged tandem repeats” or “periodicity” may be used (Lobzin and Chechetkin, 2000). In this case one still can find a similarity between any two periods but these periods are not the exact copies of each other. And finally in the case when statistically significant similarity can be found only between some of the periods (more than two) the term “latent periodicity” is used (Korotkov et al., 2003). In this paper we are interested in the search of the latent periodicity by various spectral methods.

The method of information decomposition (ID) was designed to identify latent periodicity in symbolic sequences. The distinguishing feature of information decomposition is that it does not require translation of symbols into a numeric sequence. A comparison of information decomposition with the Fourier transform (Korotkov et al., 2003) revealed the advantage of the former over the latter, in the search for latent periodicity. In this paper, we have investigated the possibility of using several spectral methods (discrete Fourier transform (DFT), autocorrelation approach (CORR), information decomposition (ID), a hybrid method (HYB), concept of spectral envelope (SE), normalized autocorrelation (CORR_N) and profile analysis (PA)) to search for latent periodicity in DNA sequences, having different number of substitutions of DNA bases. It is important to compare these approaches in order to understand the possibilities of correct identification of periods in a DNA sequence with different number of accumulated replacements of the bases. To solve this task, we modeled the latent periodicity of different DNA lengths and studied the ability of the methods to find the appropriate period at the various levels of replacements of the DNA bases (point mutations per one nucleotide – PM).

The result of this comparison shows that the method of information decomposition most correctly defines the length of the period, for sequences with high PM, in comparison to all other approaches. Similarly, Fourier transform and modified spectral envelope created advantages for short periods compared with long periods, as earlier discovered (Epps, 2009). Autocorrelation approach tends to find long periods compared to short periods which exist in the analyzed sequence. For short periods, (period length less than or equal to 4) only the hybrid method (HYB) is comparable with information decomposition (ID). This result shows that correct identification of latent periodicities in symbolic sequences is possible with the use of information decomposition.

2. Methods

2.1. Information decomposition

Let us consider a sequence of DNA \( S = s_1 s_2 \ldots s_n S_i \in \{a, t, c, g\} \). To define information decomposition of the sequence \( S \) the frequency matrix was calculated with size \( k \times p \), where \( k \) is the size of the alphabet of the considered sequence \( S \) (for DNA sequence \( k = 4 \), and \( p \) is the length of the period. The element of matrix \( M \) is \( m(i,j) \) and it indicates the number of symbols of type \( i \) which have the position \( j \). Position \( j = \text{mod}(ip) \) if \( \text{mod}(ip) \neq 0 \) and \( j = p \) if \( \text{mod}(ip) = 0 \). The value of the mutual information for the length of period \( p \) is defined as (Kullback, 1997):

\[
I_p = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{p} m(i,j) \log_2(m(i,j)) - \frac{1}{p} \sum_{j=1}^{p} y(j) \log_2(y(j))
\]

where \( x(i) = \sum_{j=1}^{p} m(i,j) \), \( y(j) = \sum_{i=1}^{k} m(i,j) \). Estimation of statistical significance of the information obtained with the use of formula (2) was carried out on a set of specially generated artificial sequences (see Section 2.5.). Summary statistics were calculated using the following formula (2):

\[
Z_p(ID) = \frac{I_p - I_{opt}}{\sqrt{D(I_p)}}
\]

where \( I \) and \( D(I) \) are the mean and variance values of \( I \) respectively, calculated for the set of random sequences. The values \( Z_p(ID) \) obtained using Eq. (2) for different periods \( p = 2, 3 \ldots n/2 \) represent the spectrum of information decomposition for the analyzed sequence \( S \) (Korotkov et al., 2003).

2.2. The autocorrelation approach to search for periodicity

Autocorrelation CORR for the sequence \( S \) was calculated for the period \( p \) directly from the sequence as follows (Afreixo et al., 2004)

\[
\text{CORR}(p) = \sum_{k=1}^{n-p} d(s_k, s_{k+p})
\]

where for any two nucleotides \( x \) and \( y \)

\[
d(x, y) = \begin{cases} 1, & x = y \\ 0, & x \neq y \end{cases}
\]

we also used the standardized measure:

\[
\text{CORR}_N(p) = \sum_{k=1}^{n-p} d(s_k, s_{k+p})
\]

and finally, the value of \( Z_p(CORR) \) was calculated using a set of artificial sequences (see Section 2.5.):

\[
Z_p(CORR) = \frac{\text{CORR}(p) - \text{CORR}(p)}{\sqrt{D(CORR)(p)}}
\]

where \( \text{CORR}(p) \) and \( D(CORR)(p) \) are the mean and variance of \( Z_p(CORR) \) for an artificial set of sequences for the period \( p \) (see Section 2.5.).

2.3. Hybrid method

Hybrid method combines the properties of autocorrelation and Fourier transformation and have recently been suggested in the search for periodicity (Epps et al., 2011; Epps, 2009):

\[
\text{Hybrid}(p) = \text{CORR}(p) \cdot X_p
\]

where \( \text{CORR}(p) \) is the autocorrelation and \( X(p) \) – integer conversion (Epps et al., 2008) for the period \( p \).

\[
X_p = \sum_{k=0}^{n-1} s_k \exp(-j\frac{2\pi k}{p}), p = 1, 2, \ldots, p_{\text{max}} \leq n
\]

The authors showed that for different periods, this method gave better results than just autocorrelation and integral Fourier
transformation methods. For calculations of \(\text{CORR}(p)\) and \(X(p)\) a DNA sequence is translated into a binary numeric sequence, where
1 represents the symbol(s) of interest, and zero (0) is assigned for
all other symbols. In this work, to study the periodicity of a DNA
sequence, we summed the results of the hybrid method obtained
in all four DNA bases for the respective period lengths. The
calculations used the code in “Python”, laid out by the authors in
the library PyCogent (Knight, 2007).

2.4. Concept of spectral envelope for spectral analysis

To obtain a spectrum of Fourier transform, the original
sequence \(S\) should be presented in a numeric form. A method
based on the representation of the sequence \(S\) in the form of four
binary indicator sequences was chosen (one for each type of
nucleotide \(j\)) (Silverman and Linsker, 1986; Tiwari et al., 1997).
Decomposition of the sequence \(S\) into four digital sequences takes
place using a single function:

\[
u_j(k) = \begin{cases} 1, & s_k = j \\ 0, & s_k \neq j \end{cases}
\]  

(9)

The original sequence \(S\) is converted as a result of a set of four
binary indicators of sequences, each of which makes the Fourier
transform. To calculate the resulting Fourier transformation of
the sequence \(S\), the function \(\text{fft}\) of the language \(R\) (R Development Core
Team, 2011) was used. The resulting value is the sum of the squares
of the module of the Fourier transformation of the four indicator
sequences.

\[R = |U_A|^2 + |U_T|^2 + |U_C|^2 + |U_G|^2\]  

(10)

For comparison of the methods used in the present work, the
statistical significance of \(R\) was estimated by calculating:

\[Z(\text{DFT}) = \frac{R - \bar{R}}{\sqrt{D(R)}}\]  

(11)

where \(\bar{R}\) and \(D(R)\) represent the mean and variance of the value \(R,
\) calculated using the set of random sequences (see Section 2.5).

The logic behind the concept of spectral envelope for spectral analysis (Stoffer et al., 1993) is that for each frequency \(\omega\), we find
the optimal encoding symbols in numbers. To do this, real vectors
are chosen \(\beta = (\beta_1, \beta_2, \ldots \beta_k) \in \mathbb{R}^k\) (which defines the encoding)
that the spectral density for the DNA sequence \(s_1 s_2 \ldots s_n, s_i \in \{c_1, c_2, \ldots c_k\}\) (if sequence DNA \(C = \{A,C,G\}\) for the con-
considered frequency \(\omega\) has a maximum. Spectral envelope is the
optimization of relations and value of the Fourier transform, for
the given frequency of dispersion (Stoffer et al., 1993):

\[\lambda(\omega) = \max_{\beta} \left\{ \frac{\{\hat{f}(\omega, \beta)\}}{\sigma^2(\beta)} \right\}\]  

(12)

To build a spectrum, the method of spectral envelope was used,
language code \(R\), presents the book (Shumway and Stoffer, 2011). In
order to compare the statistical significance of the results obtained
from different methods, the values obtained on a set of artificial
sequences was used (see Section 2.5).

\[Z_{\omega}(\text{SE}) = \frac{\lambda(\omega) - \bar{\lambda}(\omega)}{\sqrt{D(\lambda(\omega))}}\]  

(13)

where \(\bar{\lambda}(\omega)\) and \(D(\lambda(\omega))\) are the mean value and variance \(R,
\) calculated using a set of random sequences (see Section 2.5).

Fig. 1. Dependence of \(P_{MSO}\) from the number of periods \(N\) in the analyzed sequence for information decomposition (ID) - - - - - - , for the autocorrelation approach (CORR)
- - - - - - - - , a hybrid method (HYB) - - - - - - - - and spectral envelope (SE) - - - - - - - - - - - - . The dependencies are shown for \(p = 3, 10, 23\) and 45 in Fig. 1A–D, respectively.
2.5. Generating random sequences

Calculation of standardized Z-score for each method were conducted using a variety of random sequences. We randomly shuffled the original DNA sequence and for each DNA sequence, a set of 500 random sequences were generated.

2.6. Creating a periodic sequence with different levels of PM

We created artificial nucleotide sequences (alphabet = {a,t,c,g}) to explore the possibility of identifying the true period in the nucleotide sequence of different spectral methods. Each set originally contained 500 nucleotide sequences, each of which contained a short random nucleotide sequence of the same length, which was repeated N times. We denoted the set as a “Var”. Thus, we created Var3, Var10, Var23 and Var45, where the numbers 3, 10, 23 and 45 show the short order in which the length (p) was tandemly repeated.

Then, for sets Var3, Var10, Var23 and Var45, we created the sets Var3(PM), Var10(PM), Var23(PM) and Var45(PM), where PM shows that the point has accepted mutation for a single nucleotide in each sequence from a set. PM was changed from 0.1 to 4.0. All these sets were created for a different number of periods N, which were equal to 3, 10, 23, and 45 nucleotides. Mathematical methods and approaches described in paragraphs 2.1–2.4 were applied for these sets and as a result, we calculate a measure of the prediction performance W:

\[ W = \frac{Q_1}{Q_2} \]  

where \( Q_1 \) is the number of sequences from a variety of Var, where the length of the period was determined correctly. This means that the length of the period \( p_{\text{max}} \) that is maximum in the spectra defined by the formulas (2), (3), (7), and (10), was equal to the length \( p \). \( Q_2 \) is the number of sequences from a variety of Var for which \( p_{\text{max}} \) is not equal to \( p \). We calculated the value of PM, where

---

**Fig. 2.** Application of the ID method, autocorrelation approach (CORR) and normalized autocorrelation, (CORR_N), DFT, spectral envelope (SE) and hybrid method (HYB) to search for the periodicity of the region from the first chromosome of the Caenorhabditis elegans genome (NC_003279 15,072,434) from 11411926 to 11412,706 nucleotides. A: ID; B: CORR; C: CORR_N; D: DFT; E: SE; F: HYB.
$W = 0.5$ for each number of periods $N$; and called it $PM50$. Then, we calculated the dependence of $PM50$ on the number of periods $N$ for different lengths of period $p$ which was equal to 3, 10, 23 and 45 nucleotides. The value of $PM50$ is critical to the application of methods and approaches, a high value of $PM50$ favors the detection of divergent periodicity compared to methods with smaller values. The value of $PM50$ for different methods correspond to the Z-value which is equal to $3.0 \pm 0.5$, if we calculate Z-value by formulas (2), (6), (11) and (13).

3. Results and discussion

Fig. 1 shows the dependence of $PM50$ on the number of periods $N$ in the analyzed sequence, for the methods of information decomposition (ID), autocorrelation approach (CORR), hybrid method (HYB) and spectral envelope (SE). These dependencies are shown for $p = 3, 10, 23$ and 45 on Fig. 1a–d, respectively. Method DFT and normalized autocorrelation (CORR_N) gave $W$ value below 0.5 for all lengths of period $p$, for all studied $N$ and PM. Therefore, these dependencies are not shown in Fig. 1, as the dependencies correspond to a straight line $PM50 = 0.0$. For the method of spectral envelope (SE), values of $W$ above 0.5 were obtained for $p = 3$ only, so the graph for this method in Fig. 1b–d are not shown. We can assume that DFT, CORR_N and SE methods do not allow accurate determination of period length in the analyzed sequence for large values of PM.

As one can see from Fig. 1a, HYB method is a little better than ID method for $p = 3$. SE and CORR methods are worse than the ID and HYB methods for determination of period length. However, at large lengths of $p$, the ID method allows a more accurate determination of period length for much higher values of PM, in comparison to other methods. This is especially noticeable for the length of period $p = 45$. In this case the ID method $W$ is equal 0.5 for the PM = 2.0 and $W > 0.5$ for PM < 2.0.

We also examined profile analysis (Chaley and Kutyrklin, 2011) to predict the true length of the period in our analyzed sets. Profile analysis uses the criteria $\chi^2$ to search for periodicity in the sequence and select the $p$ that has the highest $PL(p)$. Here

$$PL(p) = \sum_{i=1}^{f} \max\{m(i)/Np\}.$$  Thus, $PL$ can lie in the interval from 0 to 1.  

Our calculations show that the profile analysis has $W$ below 0.5 for all investigated lengths of the period $p$. It means that the method of profile analysis cannot correctly identify the true length of the period.

Fig. 2 shows application of the method of information decomposition (ID), autocorrelation approach (CORR), hybrid method (HYB) and spectral envelope (SE) to the region of the DNA sequence from the first chromosome of *Caenorhabditis elegans* from 11,411,926 to 11,412,706 nucleotides. It is possible to see from Fig. 2a that the maximum value of $Z$ for ID method is achieved for $p = 31$ ($Z(31) = 9.2$). When we were using the autocorrelation approach (CORR) (Fig. 2B) and the hybrid method (HYB) (Fig. 2F) the length of the period (which has a maximum of the spectral functions) is equal to 20 nucleotides. If DFT or SE methods are applied (Fig. 2D–E) then we get the length of the period equal to 10.3 nucleotides. Previously, it was observed that DFT passes the intensity of the long periods to the shorter periods (Korotkov et al., 2003). The application of normalized autocorrelation (CORR_N)

---

**Fig. 3.** It shows the spectra of the Z statistics. They were obtained using formulas 6, 11 and 13 for normalized autocorrelation (CORR_N), DFT and spectral envelope (SE) for the sequence NC_003279 15,072,434 with 11411926 on 11,412,706 nucleotides. A: CORR; B: DFT; C: SE.

---

gives a period length equal to 217 nucleotides (Fig. 2C). If we take into account the results obtained from artificial sequences, it would be suggested that most probably this sequence has a period equal to 31 nucleotides.

Next, we tried to use the spectra Z (formula 6, 11 and 13) for DFT, and spectral envelope SE and normalized autocorrelation (CORR_N) for the region from the first chromosome Caenorhabditis elegans genome (NC_003279 15,072,434) from 11411926 to 11,412,706 nucleotides. The results of this analysis are shown in Fig. 3. Using Z statistics (formulas 6, 11 and 13) shows that these methods find the same length of the period, as it was done for the intensity of the signal by the Formulas (3) and (10).

Obtained results show that the periodicity of a short length of DNA (p less than or equal to the size of the alphabet of an analyzed sequence) are found by information decomposition and hybrid methods (Fig. 1a), as earlier noted (Korotkov et al., 2003). However, once the length of the period is more than 4 (size of the alphabet of an analyzed sequence) the DFT and all related with DFT approaches begin to “endure” intensity of the long periods in the intensity of the short periods. Therefore, for relatively longer periods, the DFT, SE and hybrid approaches reveal the length of the period that is different from the length, which is present in fact in the analyzed sequence.

It is also interesting to note that the normalized autocorrelation (CORR_N) worse in predicting the length of the period than the simple autocorrelation approach (CORR), (Fig. 1). This may be due to the fact that the intensity of harmonics (with periods multiples to the basic period) is almost always more than the intensity of the basic period. This does not permit the normalized autocorrelation which correctly detects the right length of the period, existing in the analyzed sequence.

In the present work p is less than 100 nucleotides. However, the method of information decomposition allows finding longer periods in nucleotide sequences. In this case it is necessary to fill a 4 × p matrix (Korotkov et al., 2003). If p < 100,000 DNA bases, all calculations can be done using a personal computer, and large values of p require a computer cluster.

In general, our results show that information decomposition method is better at finding the right period length among the studied spectral approaches for artificial DNA sequences. It allows the detection of the latent periodicity in the presence of a larger number of nucleotide substitutions than it is possible to do by other spectral methods. Especially this advantage of the ID method becomes evident when p is greater than three nucleotides. Such properties of the ID method can be very useful for finding of ancient tandem duplications in different genomes.

References


