Computational Biology and Chemistry 51 (2014) 12–21

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Investigation of phase shifts for different period lengths in the genomes of *C. elegans*, *D. melanogaster* and *S. cerevisiae*

Valentina Pugacheva[a,*], Felix Frenkel[a], Eugene Korotkov[a,b]

[a] *Bioengineering Centre of Russian Academy of Science, Moscow 117312, Russia*
[b] *National Research Nuclear University "MEPhI", Moscow 115409, Russia*

ARTICLE INFO

Article history:
Received 1 October 2013
Received in revised form 31 March 2014
Accepted 31 March 2014
Available online 13 April 2014

Keywords:
DNA sequence
Reading frame
Frameshift
Change point
Monte-Carlo method
Tandem repeat

ABSTRACT

We describe a new mathematical method for finding very diverged short tandem repeats containing a single indel. The method involves comparison of two frequency matrices: a first matrix for a subsequence before shift and a second one for a subsequence after it. A measure of comparison is based on matrix similarity. The approach developed was applied to analysis of the genomes of *Caenorhabditis elegans*, *Drosophila melanogaster* and *Saccharomyces cerevisiae*. They were investigated regarding the presence of tandem repeats having repeat length equal to 2 – 11 nucleotides except equal to 3, 6 and 9 nucleotides. A number of phase shift regions for these genomes was approximately 2.2×10^4, 1.5×10^4 and 1.7×10^2, respectively. Type I error was less than 5%. The mean length of fuzzy periodicity and phase shift regions was about 220 nucleotides.

The regions of fuzzy periodicity having single insertion or deletion occupy substantial parts of the genomes: 5%, 3% and 0.3%, respectively. Only less than 10% of these regions have been detected previously. That is, the number of such regions in the genomes of *C. elegans*, *D. melanogaster* and *S. cerevisiae* is dramatically higher than it has been revealed by any known methods. We suppose that some found regions of fuzzy periodicity could be the regions for protein binding.

1. Introduction

Fast development of sequencing methods led to a rapid accumulation of DNA sequences from genomes of various organisms (Liu et al., 2012; Xuan et al., 2012). Consequently, it gave rise to investigations of biological functions of DNA sequences and development of new mathematical methods for biological sequence analysis. It is essential to understand sequence properties and associate them with biological functions.

Effective mathematical methods have been developed to search for short tandem repeats including mini- and microsatellites (Gulcher, 2012; Merkel and Gemmell, 2008; Weber, 1990). Mini- and microsatellites play a key role in genome evolution by providing increased recombination rate (Myers et al., 2008; Richard and Pâques, 2000; Usdin, 2008). They are also related to many genetic diseases (Batra et al., 2010; Puri and Manku, 2010) and are supposed to be involved in adaptive evolution (Despons et al., 2011; Gemayel et al., 2010). From a practical standpoint, they are useful in genotyping and personal identification (Guichoux et al., 2011; Manasatienkij and Ra-ngabpai, 2012).

Many mathematical methods and algorithms for finding short tandem repeats have been developed in the last decade (Lim et al., 2013). Efficiency of tandem repeat detection depends on the sensitivity of these methods. Under sensitivity we mean the ability to detect highly diverged tandem repeats including mini- and microsatellites which have accumulated many point mutations. The methods for detecting short tandem repeats can be roughly divided into two classes (Merkel and Gemmell, 2008). The first class includes the methods that detect tandem repeats by using similarity between single periods within analyzed sequence. The second one consists of the methods that use spectral approaches for signal processing. The methods of the first class can efficiently detect periodicity with indels. However, they require all single periods in a sequence to have high similarity with each other. Usually this is related to the fact that a weight matrix for symbol pairs is used to estimate similarity between periods. So these methods may fail to find tandem repeats if similarity between single periods is low or absent. The methods of the first class are used in such programs as RepeatMasker (Chen, 2004), RECON (Bao and Eddy, 2002), REPuter (Kurtz et al., 2001), mreps (Kolpakov et al., 2003), Tallymer (Kurtz et al., 2008), TRF (Benson, 1999) and some other programs.

* Corresponding author at: prospekt 60-letiya Oktyabrya d.7 k.1, Moscow 117312, Russia. Tel.: +7 903 6133935; fax: +7 499 1350571.
E-mail address: virentis@gmail.com (V. Pugacheva).

http://dx.doi.org/10.1016/j.compbiolchem.2014.03.004
1476-9271/

The methods of the second class can reveal fuzzy periodicity having dissimilar single periods with some limitations. The methods of the second class are discrete Fourier transform, other spectral approaches (Leclercq et al., 2007; Saha et al., 2008; Sharma et al., 2004; Sussillo et al., 2004; Zhou et al., 2009) and the method of information decomposition (Korotkov et al., 2003a, 2003b).

Thus a gap appears to exist in current methods for finding tandem repeats (including mini- and microsatellites). Quite likely, there exist significantly more tandem repeats than are known currently. This gap is due to the fact that current methods cannot detect ancient tandem repeats that have accumulated large amount of both substitutions and indels.

Our objective is to develop a mathematical approach that will be able to find satellites in a case when both similarity between periods is low and the indels are present (fuzzy periodicity). The information decomposition method for finding periodicity in symbolic sequences was suggested earlier (Korotkov et al., 2003a, 2003b). This method is able to find periodicity in nucleotide or amino acid sequences containing large number of substitutions. Therefore, there is no statistically significant similarity between individual periods, but there is similarity for a set of more than two periods. Dynamic programming and many other algorithmic approaches usually cannot reveal fuzzy periodicity because they compare periods pairwise. Statistical significance of similarity between any two periods of fuzzy periodicity can be very low.

Spectral approaches such as Fourier or wavelet transform have several weaknesses for the detection of fuzzy periodicity. First, this is due to the fact that the spectral approaches can't find the periodicity in presence of indels. Second, periods with a length greater than the size of a sequence alphabet are represented incorrectly in the resulting spectral density (Korotkov et al., 2003a). For instance, the intensity of DNA periods those are four or more symbols long "spreads" into shorter periods. This drawback is also observed for amino acid sequences when period length is greater than 20. Moreover, it can also be significant for short periods if such periods contain the same amino acid in more than two positions (Korotkov et al., 2003a, 2003b).

The information decomposition method allows to find fuzzy periodicity in various genes and amino acid sequences (Korotkov et al., 1997, 1999) and to classify the periodicity types observed (Frenkel and Korotkov, 2008; Shelenkov et al., 2006). It was also found that fuzzy periodicity with a period length about 10–12 nucleotides was present in promoter sequences. Such periodicity may correspond to DNA bend near promoter region (Shelenkov and Korotkov, 2009). Further development of the information decomposition method was aimed to improve detection of fuzzy periodicity containing indels. The first step was to detect frameshifts in fuzzy triplet periodicity, i.e., to reveal fuzzy periodicity containing a single indel (Korotkov and Korotkova, 2010; Korotkova et al., 2011). This approach has shown that bacterial genomes contain about 4% of genes having frameshift mutations. These results almost do not include sequencing errors because the analyzed bacterial genomes have been sequenced dozens of times. At the same time, eukaryotic genomes contain over 9% of genes having frameshifts though some of these cases can be related to intron and exon identification errors.

In this paper we significantly modified this algorithm to reveal highly diverged tandem repeats containing single indel. We developed a new mathematical measure of similarity between frequency matrices. This measure was used to search for fuzzy tandem repeats in the genomes of *Caenorhabditis elegans*, *Drosophila melanogaster* and *Saccharomyces cerevisiae* by evaluating various period lengths from 2 to 11 bp. Detection of indel was modeled as detection of a phase shift in periodicity (Korotkov and Korotkova, 2010). The results show that there is considerably larger number of mini- and microsatellites in these genomes than were revealed earlier by other methods. Furthermore, a greater part of these genomes can be related to highly diverged tandem repeats.

2. Methods

2.1. The method for finding phase shifts between the periods of different length

2.1.1. Method description

Let $S = s(1), s(2), \ldots, S(l), \forall k = \overline{1, l} : s_k \in D$ be a sequence, where L is a length of the sequence S and $D = \{a, t, c, g\}$ is its alphabet. $d(1) = a$, $d(2) = t$, $d(3) = c$, $d(4) = g$. To search for tandem repeats (n is the length of a period) we take two segments of the same length l (l is divisible by n) in the sequence S to the left and to the right of some position x in S. x changes from $l+1$ to $L-l+1$ with a step equal to n. The first segment lies to the left of position x, and it is located from x_1 to x_2, where $x_1 = x - l$ and $x_2 = x - 1$. The second segment lies to the right of the position x and it is located from x_3 to x_4, where $x_3 = x$, $x_4 = x + l - 1$.

Let us introduce the indicator function for each element $s(k)$ of the sequence S as: $F_i(s(k)) = 1$ if $s(k) = d(i)$ and $F_i(s(k)) = 0$ if $s(k) \neq d(i)$. Then we calculate a matrix $V = v(i, j)$ for a subsequence from x_1 to x_2, where i varies from 1 to 4, j varies from 1 to n. $v(i, j) = \sum F_i(s(k))$ for such k from x_1 to x_2 that function $A(k, x_1, n) = j$. The function $A(k, x_1, n) = (k - x_1 + 1) \bmod n$ if $(k - x_1 + 1) \bmod n \neq 0$ and $A(k, x_1, n) = n$ if $(k - x_1 + 1) \bmod n = 0$. For k from x_1 to x_2 we can calculate function $A(k, x_1, n)$ and obtain a sequence of resulting values $1, 2, \ldots, n, 1, 2, \ldots, n, \ldots, 1, 2, \ldots, A(x_2, x_1, n)$ for k from x_1 to x_2. Thus an element $v(i, j)$ shows how many times we found a base $d(i)$ from the alphabet D in position j of a period. (Korotkov et al., 2003a, 2003b).

Then we calculate in a same way the matrices $W_k(x_3, x_4)$ for a subsequence from x_3 to x_4 as we did for V. $w_t(i, j) = \sum F_i(s(k))$ for such k from x_3 to x_4 that function $B(k, x_3, n, t) = j$. The function $B(k, x_3, n, t) = (k - x_3 + t) \bmod n$ if $(k - x_3 + t) \bmod n \neq 0$ and $B(k, x_3, n, t) = n$ if $(k - x_3 + t) \bmod n = 0$. Here t varies from 1 to n. A function $F_i(s(k))$ is calculated in a same way as above: $F_i(s(k)) = 1$ if $s(k) = d(i)$ and $F_i(s(k)) = 0$ if $s(k) \neq d(i)$. For k from x_3 to x_4 we can calculate function $B(k, x_3, n, t)$ and obtain a sequence of resulting values $t, \ldots, n, 1, 2, \ldots, n, \ldots, 1, 2, \ldots, B(x_4, x_3, n, t)$ for k from x_3 to x_4. Matrices $W_t(x_3, x_4)$ for $t = 2, 3, \ldots, n$ are equal to matrix $W_1(x_3, x_4)$ cyclically shifted by $t - 1$ rows. Thus $t - 1$ represents a phase shift between the matrices $W_1(x_3, x_4)$ and $W_t(x_3, x_4)$.

For a subsequence from x_1 to x_2 we calculate one matrix $V(x_1, x_2)$, while for a segment from x_3 to x_4 we calculate n matrices $W_t(x_3, x_4)$, $t = 1,2, \ldots, n$. Next we compare the matrix $V(x_1, x_2)$ to the matrices $W_t(x_3, x_4)$, $t = 1,2, \ldots, n$ using the similarity measure shown in Section 2.1.2. If there is an insertion of $t - 1$ nucleotides in a sequence at coordinate x and the subsequence from x_1 to x_4 contains tandem repeats, then the matrix $V(x_1, x_2)$ will be similar to the matrix $W_t(x_3, x_4)$, but not to $W_1(x_3, x_4)$. Otherwise the matrix $V(x_1, x_2)$ will be similar to the matrix $W_1(x_3, x_4)$. No similarity between the matrices $V(x_1, x_2)$ and $W_t(x_3, x_4)$, $t = 1,2, \ldots, n$ are observed in the absence of periodicity between x_1 and x_4.

Matrix similarity ensures detection of tandem repeats in a subsequence from x_1 to x_4. We can check whether this subsequence has an insertion or not by finding out which of the matrices $W_t(x_3, x_4)$, $t = 1,2, \ldots, n$ is the most similar to $V(x_1, x_2)$. When there is no such similarity for all values of t, we can conclude that there are no tandem repeats (for a given period length n) in a subsequence from x_1 to x_4. If the maximum similarity was found for $t = 1$, then sequence contains tandem repeats in a segment from x_1 to x_4 with no indels near x. Maximal similarity for $t > 1$ shows that the sequence analyzed contains tandem repeats in a segment from x_1 to x_4 with insertion of $t - 1$ bases (or deletion of $n - t + 1$ bases) near x.