Available online at www.sciencedirect.com
ScienceDirect
Gene 409 (2008) 53–60
GENE
www.elsevier.com/locate/gene

MMsat—a database of potential micro- and minisatellites

Andrew Shelenkov *, Alexander Korotkov, Eugene Korotkov

Bioinformatics Department of Bioengineering Centre of Russian Academy of Sciences, Prospect 60-tya Oktyabrya, 7/1, Room 303, 117312, Moscow, Russia

Received 20 July 2007; received in revised form 8 October 2007; accepted 16 November 2007
Available online 28 November 2007
Received by M. Di Giulio

Abstract

We present MMsat—a database of DNA sequences from GenBank possessing the latent periodicity at high level of statistical significance and having the period length in a range from 2 to 100 bases. The periodicity was found by analytical method of information decomposition. These sequences can be considered as potential micro- and minisatellites and thus can be useful for PCR analysis and evolutional studies. Distribution, properties, and potential functions of periodicity are discussed.

Availability: http://victoria.biengi.ac.ru/mmsat

Keywords: Bioinformatics; Databases; Nucleic acid; Information decomposition; Latent periodicity

1. Introduction

The presence of repeated sequences is a common feature for both eukaryotic and prokaryotic genomes. It has been suggested that the repeats themselves produce unusual physical structures in the DNA, causing polymerase slippage and the resulting amplification (Weitzmann et al., 1997; Wells, 1996). The other potential role for tandem repeats is gene regulation, in which the repeats may interact with transcription factors, alter the structure of the chromatin or act as protein binding sites (Richards et al., 1993; Lu et al., 1993). Also, repeat regions are often found in coding regions (Tompa, 2003), so they are directly involved in genome functioning. When they fall in regulatory regions, they may have direct influence on phenotype (Fondon et al., 2004). In the last few years, tandem repeats have been increasingly recognized as markers of choice for genotyping a number of pathogens (Keim et al., 2000; Frothingham and Meeker-O'Connell, 1998; Supply et al., 2000). The rapid evolution of these structures appears to contribute to the phenotypic flexibility of pathogens. Further, the studying of repeat regions is important for population genetic and forensic applications as well (Estoup et al., 2001; Blouin et al., 1996).

Tandem repeats are usually classified among satellites (spanning megabases of DNA, associated with heterochromatin), minisatellites (repeat units in the range 6–100 bp, spanning hundreds of base-pairs) and microsatellites (repeat units in the range 2–5 bp, spanning a few tens of nucleotides) (Le Fleche et al., 2001). Although the perfect tandem repeats can be found in genomes of different organisms, it is possible, especially in bacterial genomes, to find the very old or "ancient" microsatellites possessing very fuzzy periodicity, so they can be passed through by the mathematical methods of tandem repeat finding (Korotkov et al., 2003). Yet these ancient sequences are of great biological interest since they are usually highly polymorphous and thus can be used as genetic markers (Le Fleche et al., 2001; van Belkum et al., 1997; Adair et al., 2000).

We used the method of information decomposition (Korotkov et al., 2003) to make a search for periodic sequences through the whole GenBank database. That is, we have scanned GenBank using the software developed by us and put the results into our database. Information decomposition (ID) is a spectrum representing the statistical significance of mutual information

Abbreviations: ID, information decomposition; PCR, polymerase chain reaction.

* Corresponding author. Tel./fax: +7 499 135 2161.
E-mail address: fallandar@gmail.com (A. Shelenkov).

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.11.007

for periods of various lengths in the analyzed symbolical sequence. It is similar to a spectrum of Fourier transformation for numerical sequences, but has remarkable advantages (see Section 2.3).

In the current database we present totally 2,851,428 periodic sequence regions with a period length in a range 2–100 revealed at high level of statistical significance. The sequences possessing perfect periodicity were also included. The minimal length of a periodic sequence that can be revealed by information decomposition equals four repeats, i.e., 8 for period length=2, 12 for period length=3 etc. Most sequence regions contained in this database have not been revealed by other programs for finding tandem repeats (e.g., Benson, 1999; Kolpakov et al., 2003; Rice et al., 2000). Also, in comparison with existing microsatellite databases (e.g., Jewell et al., 2006; Microsat 2006) MMsat has some useful features, such as the possibility of making a search within certain functional properties (gene, intron etc.) and of searching for periodic sequences similar by their nucleotide composition (this can be done by specifying frequency matrix). Some of these sequences may appear to be highly polymorphous ones that allow their use as a starting point for PCR analysis. Also, the study of the sequences contained in our database can assist in evolutionary studies of various genomes. The evolutionary issue was previously considered to be important by other authors (e.g., Pellegrini and Yeates, 1999).

2. Materials and methods

2.1. Source sequences

DNA sequences in which the search for periodicity has been made were selected from GenBank (http://www.ncbi.nlm.nih.gov/Genbank). All sequences for eight groups of organisms (bacteria, invertebrates, mammals (not primates or rodent), plant, primates, rodent, vertebrates (not mammals), viruses +phages) have been taken into consideration.

2.2. Latent periodicity

The notion of latent periodicity is described in details in our previous papers (e.g., in (Korotkov et al., 2003)). In brief, latent periodicity differs from the perfect one by difference in appearance of symbols in period positions. Namely, only one symbol can appear in each position for perfect periodicity, while for latent periodicity more than one symbol can appear in any position. Below we show the examples of sequences possessing perfect and latent periodicities, respectively. Period length equals seven and number of repeats is 20, that is, total length of sequence is 140. Matrices of nucleotide appearance for the period positions are shown in Table 1. Let us assume that DNA bases listed in Table 1 for latent periodicity could appear in each position of the period with equal probability.

Table 1
Appearance of nucleotides in each position of period for two example sequences

Position of period	A set of DNA bases that could appear in a given position of the period, perfect periodicity	A set of DNA bases that could appear in a given position of the period, latent periodicity
1	A	ATC
2	C	AGCT
3	C	CTG
4	T	AGCT
5	A	CTA
6	C	CT
7	A	GAT

Perfect periodicity:

ACCTACAACCTACAACCTACAACCTACAACCTACA-ACCTACAACCTACAACCTACAACCTACAACCTACAAC-CTACAACCTACAACCTACAACCTACAACCTACAACCT-ACAACCTACAACCTACAACCTACAACCTACA

Latent periodicity:

ACCTACATGGGTTTTAGTATGTTCTACTCGGACTAC-ACCCTATTCTCCGCCTCTTGTGGTTCTTGTGCCGTGCC-CCTTCTTACTTACCCCTTTTGCCGTATGCTAAAGATCG-AATCCTGTCTAACTTTGAATTAAGTATT

2.3. Information decomposition

The method of information decomposition is described in details in the paper of Korotkov et al. (2003), so here we give only a brief description of it.

Information decomposition 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…n,1,2,3…n…. 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 ($i=1,2..,n$; $j=1,2...k$) type between sequences compared. L is the length of the analyzed symbolical sequence, $x(i)$, $i=1,2,...$, n are the frequencies of symbols 1,2,..., n in the artificial periodic symbolical sequence; $y(j)$, $j=1,2,...$, k are the frequencies of symbols in the analyzed symbolical sequence. The value of the mutual information is calculated using formula

$$I(n,k) = \sum_{1}^{n}\sum_{1}^{k} M(i,j)\ln M(i,j) - \sum_{1}^{n} x(i)\ln x(i) - \sum_{1}^{k} y(j)\ln y(j) + L\ln L \quad (1)$$

For ID construction it is necessary to take into account that the value $2I(n,k)$ is distributed as χ^2 with $(n-1)(k-1)$ degrees of freedom. However, for small sample statistics this approximation does not work well (Korotkov et al., 2003). So to