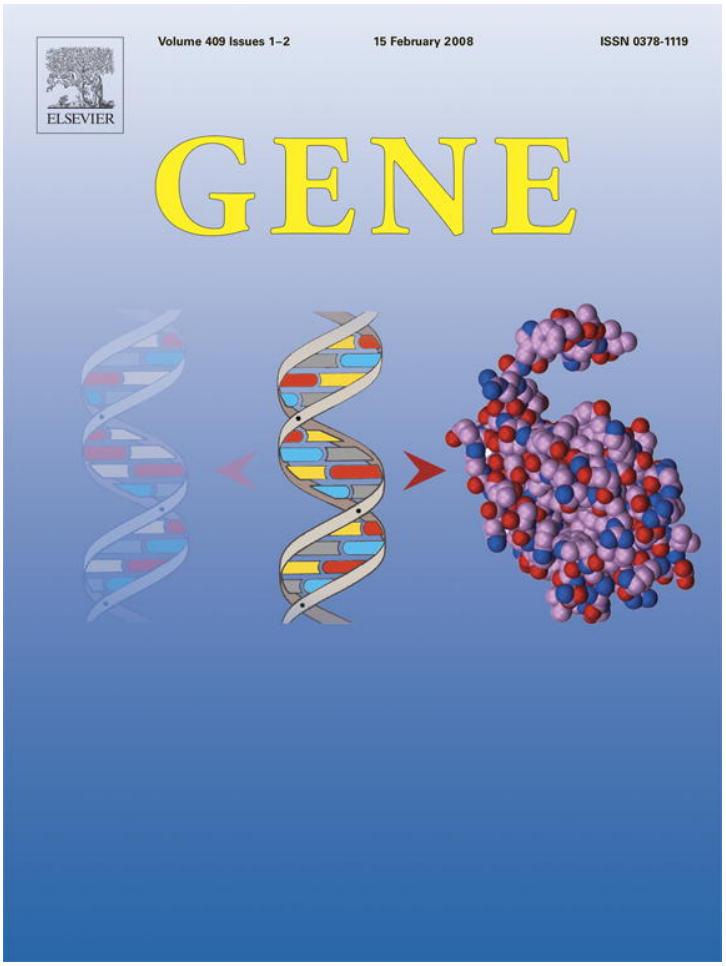

Available online at www.sciencedirect.com
ScienceDirect
Gene 409 (2008) 53–60
GENE
www.elsevier.com/locate/gene

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.

Perfect periodicity:

ACCTACAACCTACAACCTACAACCTACAACCTACA-ACCTACAACCTACAACCTACAACCTACAACCTACAAC-CTACAACCTACAACCTACAACCTACAACCTACAACCT-ACAACCTACAACCTACAACCTACAACCTACA

Latent periodicity:

ACCTACATGGGTTTTAGTATGTTCTACTCGGACTAC-ACCCTATTCTCCGCCTCTTGTGGTTCTTGTGCCGTGCC-CCTTCTTACTTACCCCTTTTGCCGTATGCTAAAGATCG-AATCCTGTCTAACTTTGAATTAAGTATT

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

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, $x(i)$, $i=1,2,\ldots, n$ are the frequencies of symbols 1,2,…, 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

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

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

estimate the statistical significance of $I(n,k)$ the Monte Carlo method is used by means of $Z(n,k)$ calculation using formula

$$Z(n,k) = \left\{I(n,k) - \overline{I(n,k)}\right\} / \sqrt{D(I(n,k))} \quad (2)$$

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.

In order to find non-random sequences of maximal length, it is essential to choose Z-score providing less than 5% probability of finding a random sequence with Z greater than this threshold score. We found such a value by applying the Monte Carlo method to a random set of symbols with a length about 2 times greater than the total length of sequences presented in GenBank. Our studies have shown that the threshold value equal to 5.0 ensures that the number of 'noisy' sequences (that is, the sequences that are not significantly periodic) in the set of the sequences found is less than 5%. So, all the sequences contained in the database have a score equal to or greater than 5.0.

Calculation of information decomposition spectra has the following advantages in comparison with methods based on Fourier transformation or dynamic programming:

- The calculation of the spectrum does not require any transformation of a symbolical sequence to numerical sequences.
- ID allows the revealing of both obvious periodicity and the latent periodicity of a symbolical sequence in which there is no statistically important similarity between any two periods.
- The statistical significance of long periods is not spread onto the statistical significance of shorter periods.
- It is possible to determine the type of periodicity on the basis of the nucleotide frequency matrix.

3. Results and discussion

In this paper, we describe the database of DNA sequences possessing latent periodicity at high level of statistical significance. These sequences were obtained using the information decomposition method by scanning available GenBank sequences (see Section 2.3). Such a scan has been made once and results have been inserted into database. We plan to update our database in the near future as new results become available. It should be noted that MMsat is a standalone database, i.e., all data are located on our server and no queries to NCBI website are made while making searches in our database. Nevertheless, there is a possibility to get additional information from NCBI for each of the sequences found by clicking corresponding links, as it is described below (see Section 3.1).

Each periodic region containing in the database possesses the following properties:

- locus code
- coordinates of the periodic region in locus sequence (begin and end)
- period length
- statistical significance of the periodicity (score calculated by using formula (Eq. (2))
- triplet significance (if this value is high for some sequences than they also possess triplet periodicity in addition to the periodicity with a period length specified)
- frequency matrix
- group of organisms to which the locus belongs to

Groups of organisms are the same as in GenBank (bacteria, invertebrates, mammals (not primates or rodent), plant, primates, rodent, vertebrates (not mammals, primates or rodent), viruses and phages). Search through all groups is also available. All of these properties (except triplet significance) may be specified as parameters while making queries to the database.

In addition, the data regarding the functional properties of the periodical sequences are also stored in the database. It is taken from GenBank 'feature' field which can include gene, promoter, repeat region, tRNA etc. The full list of supported features can be found on the web site in 'Help' section. A periodic region is considered to possess some functional property if its overlap with the functional region defined in GenBank is greater than 30% of this functional region's length. For example, if some periodic sequence had an overlap of more than 30% (e.g., 50%) with locus' region annotated as 'promoter' in GenBank, we consider this periodic region to have the same functional property ('promoter'). Of course, such assignment of a function is just useful approximation, not an accurate prediction.

The number of sequences and loci to which they belong to for different period lengths is shown in Table 2.

Table 2
Distribution of number of periodic sequences

Period length	Number of sequences	Number of loci
2	366,739	72,626
3	1,323,348	609,881
4	261,544	59,389
5	36,553	21,265
6	87,649	37,733
7	22,973	14,548
8	54,555	29,444
9	16,954	12,231
10	26,397	18,912
11	16,665	12,994
12	18,468	14,579
13	11,807	10,091
14	13,704	7787
15	13,704	10,904
16	11,337	9627
17	11,904	9746
18	11,273	9275
19	11,374	9540
20	13,101	10,492
21–50	269,863	75,699
51–80	196,295	60,285
81–100	60,085	35,551

The total number of periodic sequences contained in the database is 2,851,428 and the number of distinct GenBank loci in which they have been found is 672,024. Some loci contain more than one periodic sequence and these sequences, in turn, may have different period length.

3.1. Working with database

Users access the database through a simplified www interface. Access to the database is free for non-commercial purposes. No registration is required.

On the first page, the interface language can be chosen (English or Russian). Upon choosing the language, a brief description of the database loads together with navigation tab. All navigation is performed by clicking the corresponding fields on this tab. User can make requests to the database after choosing 'Database Query' field. The description of the method used for filling the database (information decomposition) can be found at the 'Method' page, database manual is located at the 'Help' page.

All the parameters that can be specified for the request are optional, but at least one parameter should be entered (either the organism group different from 'all' should be chosen or some other parameter should be entered). The following request fields are available: locus code, matrix, keywords, score (lower and upper bounds), period length (lower and upper bounds), position (begin and end), organism group (menu). Most of them have been described above (they correspond to the properties of periodic sequences stored in the database), so only the 'keywords' field is to be discussed. Here a user can specify the names of GenBank features separated by spaces. List of supported features is to be found at the 'Help' page. The results of database search will include only the sequences that possess at least one of the given functional properties though the coordinates of features in GenBank will not be given since this information can be obtained directly from a GenBank browser.

Search results are represented as the list of loci containing the periodic sequences satisfying the search conditions specified. Detailed information regarding the periodic regions of the

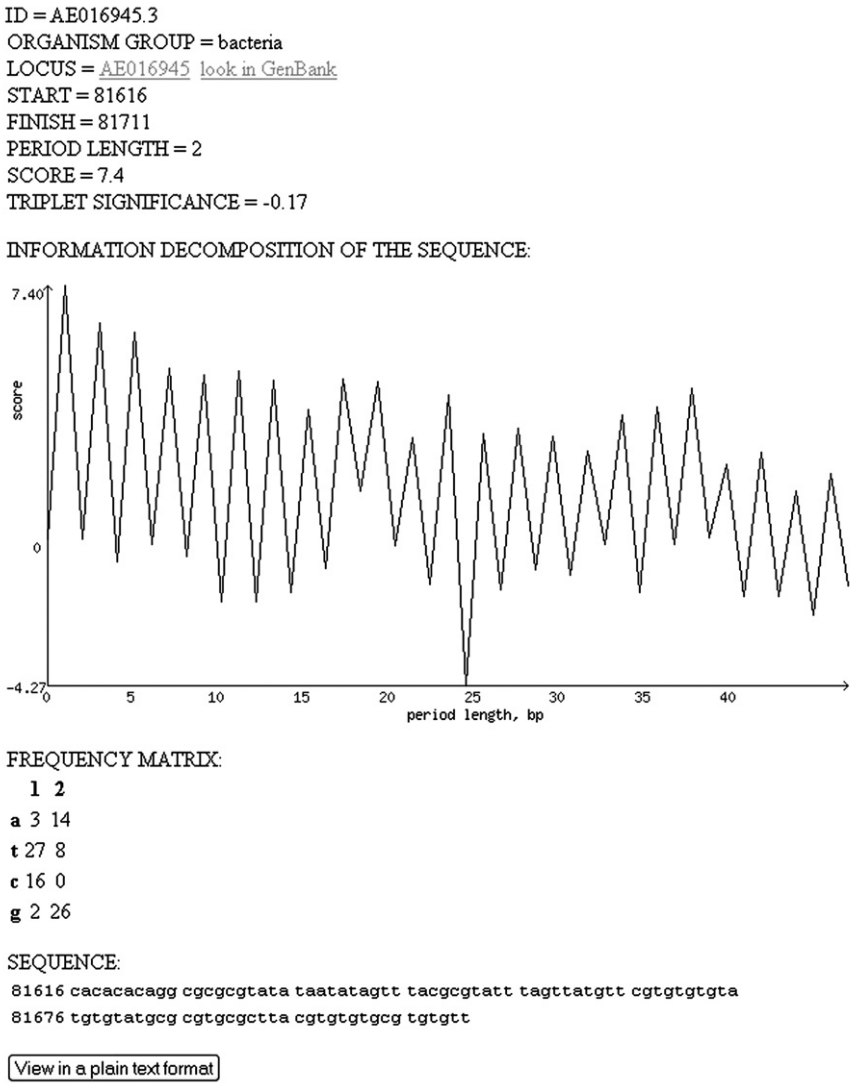

Fig. 1. Example of database query output. A screen with locus' detailed information is shown.

sequences can be obtained by clicking the code of the locus of interest. Some extra options are available on this page– information can be represented as plain text and the link to the GenBank browser for the corresponding sequence is given. If the query parameters include locus code and the number of found sequences is less than 5, then the page with detailed information is given right away, without loading the list of loci.

Detailed sequence information includes:

- ID–periodic region identifier, consist of locus code and internal database number of the region in locus
- Organism group
- Locus code
- Coordinates of periodic region in locus sequence (start and finish)
- Period length
- Score (statistical significance of region's periodicity, see Section 2.3)
- Triplet significance (see the description above)
- Graph of the information decomposition for the region
- Frequency matrix of the periodic region found
- The sequence of periodic region itself presented in Genbank format. Numbering of the nucleotides corresponds to the locus sequence.

An example of the output information is shown in Fig. 1.

It should be noted that the graph of information decomposition shows the statistical significance of various period lengths for the region found. This significance is calculated using the Monte Carlo method (see Section 2.3). The period length is assigned to the region if its value in the specter is maximal. For example, for the region in Fig. 1 the maximal value in the specter is 7.40 and it corresponds to the period length = 2.

Frequency matrix shows the number of appearances of each nucleotide (a, t, c and g) in each position of the period. Since our database contains sequences possessing latent periodicity, more than one nucleotide may appear in each position (see Sections 2.2 and 2.3). We use the frequency matrix, not the consensus sequence, since it gives more information regarding the structure of the periodic region. For example, the consensus for the matrix shown in the Fig. 1, could be either {t, c}{a, g} or {t}{g}. This matrix can be used for further investigations, for

Table 3
Examples of periodic sequences containing region annotated as polymorphic

GenBank locus code	Organism group	Organism	Region description	Coordinates in locus sequence	Period length	Periodical region	*Z*
AB010639	Pri	Homo sapiens	CDS, polymorphism	16–1305	3	gaagtccctgtctctgctcctcgctgtggctttgggcctggcgaccgccg tctcagcaggacccgcggtgatcgagtgttggttcgtggaggatgcgagc ggaaagggcctggccaagagacccggtgcactgctgttgcgccagggacc gggggaaccgccgccccggccggacctcgaccctgagctctatctcagtg tacacgaccccgcgggcgccctccaggctgccttcaggcggtatccccgg ggcgcccccgcaccacactgcgagatgagccgcttcgtgcctctccccgc ctctgcgaaatgggccagcggcctgacccccgcgcagaactgcccgcggg ccctggatggggcttggctgatggtcagcatatccagcccagtcctcagc ctctccagcctcttgcgaccacagccagagcctcagcaggagcctgttct catcaccatggcaacagtggtactgactgtcctcacccacacccctgccc ctcgagtgagactgggacaagatgctctgctggacttgagctttgcctac atgccccccacctccgaggccgcctcatctctggctccgggtccccctcc ctttgggctagagtggcgacgccagcacctgggtaagggacatctgctcc tggctgcaactcctgggctgaatggccaaatgccagcagcccaagaaggg gccgtggcatttgctgcttgggatgatgatgagccatggggcccatggac cggaaatgggaccttctggctgcctacagttcaaccctttcaggagggca cctatctggccaccatacacctgccatacctgcaaggacaggtcaccctg gagcttgctgtgtacaaaccccccaaagtgtccctgatgccagcaaccct tgcacgggccgccccaggggaggcacccccggaattgctctgccttgtgt cccacttctacccttctgggggcctggaggtggagtgggaactccggggt ggcccaggggccgctctcagaaggccgaggggcagaggtggctctcggc cctgcgccaccattccgatggctctgtcagcctctctgggcacttgcagc cgcccccagtcaccactgagcagcatggggcacgctatgcctgtcgaatt caccatcccagcctgcctgcctcggggcgcagcgctgaggtcaccctgga ggtagcaggtctttcagggccctcccttgaggacagcgtaggccttttcc tgtctgcctttcttctgcttgggctcttcaaggcactggg	5.79
AB050339	Vrt	Pseudobagrus ichikawai	Polymorphism	85–151	2	tgtgtgtgtgcatgagagagagagagagccagagagaaagagtgagatca acagagactcttactgc	6.44
AB006034	Rod	Mus musculus	CDS, polymorphism	769–1105	7	ctggaccctgggcccgcctctgccgagactgggatcagatgtttgccttt gcccagaggcacgtggagctgcgagaaggtgaagctgcgatgaggaacca gggaaagcctgaggaggatatgccgtctgggcatcacttaacccacttcc tttttcgggaaaaggtgtctgtccagtccatagtggggaatgtgacagag ctactactggctggagtggacacggtatccaatacgctctcctggacact ctatgagctttcccggcaccccgatgtccagactgcactccactctgaga tcacagctgggacccgtggctcctgtgcccacccccа	5.05

Table 4
The numbers of periodic sequences for the organisms of different categories (only the organisms with maximal number of sequences are shown)

Group	Organism	Number of periodic sequences	Percentage of organism's loci length in group[a]	Percentage of periodic sequences found in organism[b]
Bacteria	*Nostoc* sp.	649	10.71	1.03
	Escherichia coli	227	3.75	1.70
	Streptomyces avermitilis	206	3.40	1.35
Invertebrates	*Drosophila melanogaster*	48,543	66.89	48.03
	Plasmodium falciparum	6262	8.63	2.05
	Caenorhabditis elegans	4691	6.46	8.75
Mammals (other)	*Sus scrofa*	1708	20.45	14.74
	Equus caballus	1259	15.07	3.61
	Bos Taurus	1067	12.78	8.19
Plant	*Oryza sativa*	16,933	53.79	31.40
	Arabidopsis thaliana	5397	17.15	15.73
	Lotus corniculatus	1003	3.19	2.75
Primates	*Homo sapiens*	318,168	98.62	14.49
	Pan troglodytes	2551	0.79	0.77
	Papio anubis	993	0.31	0.31
Rodent	*Mus musculus*	271,451	96.57	88.98
	Rattus norvegicus	8756	3.12	2.61
	Mus sp.	82	0.03	0.01
Vertebrates (other)	*Danio rerio*	36,913	82.29	65.59
	Takifugu rubripes	768	1.71	1.50
	Gallus gallus	763	1.70	1.72
Viruses and phages	Human immunodeficiency	645	37.22	0.42
	Human papillomavirus	91	5.25	0.16
	Human herpesvirus	70	4.04	1.33

[a]"The percentage of organism's loci length in group" is the ratio of total length of all loci of the specific organism to the total length of all loci of this organism's taxonomic group in Genbank (e.g., bacteria, plant etc).
[b]The "percentage of periodic sequences found in organism" shows the fraction of periodic sequences found in the loci of the taxonomic group which appeared to belong to the loci of the specific organism.

example, for searching the periodicity with insertions and deletions of symbols. This matrix may be used as a query parameter. In this case, the periodic regions having similar frequency matrix will be found. Chi-square value is used as a matrix similarity measure. Threshold value corresponds to the accidental similarity probability of less than 5%.

There is a possibility to obtain the result in plain text format by pressing 'View in a plain text format' key located below each periodic region on the resulting page.

All information regarding the request and result formats can be also found at the 'Help' page.

3.2. Biological interpretation

We claim that the sequences contained in our database are potential micro- and minisatellites. To prove this, we are to find the regions which are polymorphic, but have not been annotated earlier as repeats. Since the experimental proof of polymorphism lies beyond the scope of our research, we used GenBank annotation ('features' field). Examples of such regions are shown in Table 3. All these regions were determined to be periodic by our method and are annotated as polymorphous ones, but not as repeats, in GenBank. Since some of the sequences contained in GenBank are poorly annotated, it is possible that some regions found by us will be found to be polymorphic in future.

To get some biological insights, it is interesting to investigate the distribution of periodic regions found among various organisms and groups of organisms.

At first, we should note that dinucleotide periodicity is the most widespread among almost all groups of organisms (except bacteria and viruses, for which such a period length equals 7 and 10, respectively). It is not a surprise since this type of periodicity is well-known and such microsatellites make a significant part of certain genomes (Gupta and Varshney, 2000). The predominance of periods with a length 10 in viruses is probably associated with large number of α-helical proteins that are coded in virus' sequences.

Our investigations have shown that periodicity phenomenon in general is not specific to some organisms or group of organisms. However, there is a possibility that certain types of periodicity (type is defined by frequency matrix) are specific to some organisms or functional regions. This could help in annotation of recently obtained sequences. Some data regarding the organism distribution of periodical sequences are shown in Table 4, whereas functional properties for some of these sequences (bacteria and plants) are summarized in Table 5.

The number of periodic regions found in some organism is not very informative itself since the length of loci differs dramatically for different organisms. This is why we put additional data to Table 4––fraction of organism loci's length

Table 5
Distribution of periodical sequences from some groups of organisms (bacteria and plant) revealed in previously characterized DNA regions

GenBank feature	Number of sequences overlapping with the feature region, bacteria	Number of sequences overlapping with the feature region, plant
3'UTR	3	65
mRNA	25	149
Promoter	3	36
Rep_origin	3	1
Repeat_region	185	11,505
Repeat_unit	15	117
Satellite	1	481
rRNA	306	146
Sig_peptide	13	1
Stem_loop	6	2
tRNA	1	5
Gene	3530	3595
Coding regions of the gene	3526	233

for the corresponding group and percentage of periodical regions found in this organism's loci. If the periodic regions are distributed evenly along genomes, then these two fractions would be approximately the same. However, we see that there exists significant difference in these parameters for some organisms. Thus we can conclude that the distribution of periodic sequences in genomes is not random, and there are some organisms in which it is more likely to found periodic sequences than expected by chance (e.g., *Nostoc* sp., *D. melanogaster*, *O. sativa*, *H. Sapiens*, Human immunodeficiency virus, etc). Also it should be noted that periodicity in eukaryotes is more common phenomenon than in bacteria and viruses. This fact has also been confirmed by other researchers (e.g., Marcotte et al., 1999; Fukushima et al., 2002). The most relatively widespread periodicity, according to our data, appears in rodent genomes.

It is also interesting to investigate the functional properties of the periodical sequences found. We take this information from GenBank's 'FEATURES' field. The periodical sequence was considered to have some functional property if its overlap with previously annotated region was greater than 30%. Because of space limits and some general reasons, this type of data is presented for bacterial and plant genomes only.

The database contains many regions with triplet periodicity. Their functional properties may include both potential microsatellite sequences and coding regions of DNA. They can be approximately distinguished by the number of repeats. If the region contains more than 100 repeats, then it most likely belongs to the coding region of DNA or to a pseudogene. If the number of repeats is less than 100, then this DNA sequence could be a microsatellite, possibly even evolutionary old.

One can see that more than 200 sequences for bacteria and more than 12,000 for plant overlapped with sequence repeats already detected empirically, so our method proves to find such kind of repeats. Since the annotation of bacterial genomes is much better than of plant genomes, these data could not be compared directly. However, it is interesting that many periodic sequences in bacteria lie in intergenic regions or in regions that have not been annotated previously (data not shown). Feature distribution supports both our statements––our method is able to identify repeats that were previously detected empirically and there are some periodic regions found by this method that were not previously revealed.

In summary, we have presented the database of GenBank sequences possessing latent periodicity. We used the method of information decomposition to reveal such sequences. We anticipate that it is possible to use the periodical sequences found as a starting point for the PCR analysis because some of them can turn out to be highly polymorphous ones. The study of possible ancient minisatellites may also be helpful for evolutionary analysis of genomes. However, minisatellites need not be ancient to be useful in this analysis because of their extremely high mutation rates (Bois, 2003). They often provide tools for studying variability over short time periods, which cannot be addressed by point mutations (e.g., Gaspari et al., 2007). Using our method, it is possible to identify fuzzy repeats that can be passed by other methods, so the database contains a lot of unique data. We plan to update the database in the future as the new data become available. We also plan to develop the web server for ID method in the future to make available the search of periodic sequences in recently obtained genomic sequences.

Acknowledgements

The authors wish to thank Felix Frenkel, Valentina Rudenko (Bioengineering Centre, RAS) and Anna Slavokhotova (Vavilov Institute of General Genetics, RAS) for useful suggestions that led to improvement of database web site's user interface.

References

Adair, D.M., Worsham, P.L., Hill, K.K., et al., 2000. Diversity in a variable-number tandem repeat from Yersinia pestis. J. Clin. Microbiol. 38, 1516–1519.

van Belkum, A., Scherer, S., van Leeuwen, W., Willemse, D., van Alphen, L., Verbrugh, H., 1997. Variable number of tandem repeats in clinical strains of Haemophilus influenzae. Infect. Immun. 65, 5017–5027.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580.

Blouin, M.S., Parsons, M., Lacaille, V., Lotz, S., 1996. Use of microsatellite loci to classify individuals by relatedness. Mol. Ecol. 5, 393–401.

Bois, P.R.J., 2003. Hypermutable minisatellites, a human affair? Genomics 81, 349–355.

Estoup, A., Wilson, I.J., Sullivan, C., Cornuet, J.M., Moritz, C., 2001. Inferring population history from microsatellite and enzyme data in serially introduced cane toads, Bufo marinus. Genetics 159, 1671–1687.

Fondon III, J.W., Garner, H.R., 2004. Molecular origins of rapid and continuous morphological evolution. PNAS USA 101, 18058–18063.

Frothingham, R., Meeker-O'Connell, W.A., 1998. Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats. Microbiology 144, 96–1189.

Fukushima, A., Ikemura, T., Kinouchi, M., et al., 2002. Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis. Gene. 300, 203–211.

Gaspari, Z., Ortutay, C., Toth, G., 2007. Divergent microsatellite evolution in the human and chimpanzee lineages. FEBS Lett. 581, 2523–2526.

Gupta, P.K., Varshney, R.K., 2000. The development and use of microsatellite markers for genetic analysis of plant breeding with emphasis on bread wheat. Euphytica 113, 163–185.

Jewell, E., Robinson, A., Savage, D., et al., 2006. SSRPrimer and SSR taxonomy tree: biome SSR discovery. Nucleic Acids Res. 34, W656–W6569.

Keim, P., Price, L.B., Klevytska, A.M., et al., 2000. Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within Bacillus anthracis. J. Bacteriol. 182, 2928–2936.

Kolpakov, R., Bana, G., Kucherov, G., 2003. mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res. 31, 3672–3678.

Korotkov, E.V., Korotkova, M.A., Kudryashov, N.A., 2003. Information decomposition method to analyze symbolical sequences. Phys. Let. A. 312, 198–210.

Le Fleche, P., Hauck, Y., Onteniente, L., et al., 2001. A tandem repeats database for bacterial genomes: application to the genotyping of Yersinia pestis and Bacillus anthracis. BMC Microbiol 1, 2.

Lu, Q., Wallrath, L., Granok, H., Elgin, S., 1993. (CT)n (GA)n repeats and heat shock elements have distinct roles in chromatin structure and transcriptional activation of the Drosophila hsp26 gene. Mol. Cell. Biol. 13, 2802–2814.

Marcotte, E.M., Pellegrini, M., Yeates, T.O., Eisenberg, D., 1999. A census of protein repeats. J. Mol. Biol. 293 (1), 151–160.

Microsat 2006. http://www.microsatellites.org/db_search.php.

Pellegrini, M., Yeates, T.O., 1999. Searching for frameshift evolutionary relationships between protein sequence families. Proteins 37, 278–283.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16, 276–277.

Richards, R., Holman, K., Yu, S., Sutherland, G., 1993. Fragile X syndrome unstable element, p(CCG)n, and other simple tandem repeat sequences are binding sites for specific nuclear proteins. Hum. Mol. Genet. 2, 1429–1435.

Supply, P., Mazars, E., Lesjean, S., Vincent, V., Gicquel, B., Locht, C., 2000. Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome. Mol. Microbiol. 36, 762–771.

Tompa, P., 2003. Intrinsically unstructured proteins evolve by repeat expansion. BioEssays 25, 847–855.

Weitzmann, M., Woodford, K., Usdin, K., 1997. DNA secondary structures and the evolution of hypervariable tandem arrays. J. Biol. Chem. 272, 9517–9523.

Wells, R., 1996. Molecular basis of genetic instability of triplet repeats. J. Biol. Chem. 271, 2875–2878.