DNA Research **13**, 89–102 (2006) doi:10.1093/dnares/dsl004

Search and Classification of Potential Minisatellite Sequences from Bacterial Genomes

Andrew Shelenkov*, Konstantin Skryabin and Eugene Korotkov

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

(Received 12 April 2006; revised 1 June 2006; published online 16 September 2006)

Abstract

We used the method of Information Decomposition developed by us to identify the latent dinucleotide periodicity regions in bacterial genomes. The number of potential minisatellite sequences obtained at high level of statistical significance was 454. Then we classified the periodicity matrices and obtained 45 classes. We used the other new method developed by us—Modified Profile Analysis—to reveal more periodic sequences in the presence of indels using the classes obtained. The number of sequences found by combination of these two methods was 3949. Most of them cannot be revealed by other methods including dynamic programming and Fourier transformation.

Key words: information decomposition; modified profile analysis; minisatellites; bacterial genomes

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.[1,2] 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.[3,4] In the last few years, tandem repeats have been increasingly recognized as markers of choice for genotyping a number of pathogens.[5–7] The rapid evolution of these structures appears to contribute to the phenotypic flexibility of pathogens.

Tandem repeats are usually classified among satellites (spanning megabases of DNA, associated with heterochromatin), minisatellites (repeat units in the range 6–10 bp, spanning hundreds of base pairs) and microsatellites (repeat units in the range 2–5 bp, spanning a few tens of nucleotides).[8] Microsatellites or simple sequence repeats (SSRs) are tandemly repeated DNA sequences found in varying abundance in most genomes.[9,10] These repeats have been extensively used for genetic mapping and population studies.[11] Microsatellites 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. Polymorphic minisatelites are also found in bacterial genomes.[8]

The availability of whole-genome sequences has opened the way to the systematic evaluation of tandem repeats diversity and application to epidemiological studies.[8] More recently, a number of studies[12,13] have confirmed the idea that tandem repeats reminiscent of minisatellites and microsatellites are likely to be a significant source of very informative markers for the identification of pathogenic bacteria.

Once repeats are identified, the central task becomes the clustering of tandem repeats into families, i.e. repeats that occur in different locations in a genome but have identical or very similar underlying patterns. Grouping these repeats will facilitate identification and study of their common properties. Tandem repeat families have been detected in both prokaryotes and eukaryotes,

Communicated by Kenta Nakai

* To whom correspondence should be addressed. Tel. +7-495-135-2161. Fax +7-495-135-2161. E-mail: fallandar@mail.ru

including the *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and human genomes.[14] Analyzing the mutational history of tandem repeats requires utilizing the pattern of mutations among adjacent copies to describe the interwoven progression of substitutions, indels and duplication/excision events leading from a single copy of the pattern to the present day sequence. Such histories can suggest how the boundaries and size of the duplication unit vary and may reveal details about the duplication mechanism.[15]

Among the best programs for finding the tandem repeats are Tandem Repeats Finder[15] and mreps.[16] Although they overcome some limitations of other algorithms, they have their own, limitations, 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 cannot be used to identify the fuzzy periodicity and ancient minisatellites.

A number of public tandem repeat databases are available now. To name a few, a database of bacterial genomes,[8] a short tandem repeat DNA database,[17] TRDB[15] and TRbase.[18] The methods and programs described in the previous paragraphs are used to fill such databases in most cases. In Section 3 we will discuss in detail the advantages and drawbacks of the existing methods.

This paper has the following goals. The first goal is to identify the potential minisatellites with the method of Information Decomposition (ID),[19] to calculate the periodicity matrices and to classify them. The second goal is to find the new periodic regions (in the presence of insertions and deletions) with the Modified Profile Analysis (MPA)[20] approach using the periodicity class matrices found on the previous step.

We have chosen bacterial genomes as a target of our study for the following reasons. First, bacterial minisatellites can be used as markers for PCR, so there is a practical application for this work. Second, the mutational history of bacterial genomes is evidently more prolonged than that of eukaryotic genomes, so search for the fuzzy repeats is more complicated, but at the same time can give more insights concerning the biological role of 'ancient' minisatellites. Third, the comparatively small number of available sequences could result in less number of classes, thus making it easier to explore their properties and to reveal the biological principles standing behind the results of classification process.

As we will show in Section 3, our method can reveal the periodical sequences that were not found by the existing software packages.

2. Materials and methods

Our search for the potential minisatellites was focused on the bacterial genomes, since their size allows finishing the computations in a reasonable amount of time; since, the periodical sequences found yield the evident practical application, they can be used in PCR analysis. In bacterial genomes it is possible to find the very old or 'ancient' microsatellites possessing very fuzzy periodicity, so they can be passed through by the mathematical methods of tandem repeats finding. Yet these ancient sequences are of great biological interest since they are usually highly polymorphous and, thus, can be used as genetic markers.

The method proposed in this paper consists of three parts—usage of ID method[19] to obtain the initial data, classification of the results found and the usage of MPA for searching the sequences with indels.[21] The usage of periodicity classes rather than all periodic sequences found dramatically decreases 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 us 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 that cannot be revealed by other methods.

2.1. Information decomposition

The method of ID is described in detail in ref. (19). Since this method plays an important role in the present study, we also give its thorough description.

At first, let us define the concept of latent periodicity of a symbolical sequence.

Suppose that we have a sequence S, which consists of N subsequences S^i ($i = 1, 2, \ldots, N$) of equal length L:

$$S \equiv S^1 S^2 \ldots S^N \equiv \{s_1^1 s_2^1 \ldots s_L^1 s_1^2 s_2^2 \ldots s_L^2 \ldots s_1^N s_2^N \ldots s_L^N\}$$

where s_i^j are DNA bases. Suppose we are looking for a period of the length L in the sequence S ignoring possibilities of insertions and deletions of symbols. To find such a period we should evaluate global homology between subsequences S^i ($i = 1, 2, \ldots, N$). If this homology between subsequences S^i is statistically significant, we can conclude that symbol periodicity with period L exists in sequence S. Possibilities for finding periodicity in the sequence S will depend on the mode of insertion of the quantitative measure determining similarity of subsequences S^i. At the present time, the method of dynamic programming and Fourier transformation are often used for these purposes. To introduce the quantitative measure for global homology of subsequences S^i these methods use a search for homology between these subsequences. While using the method of

dynamic programming, the search for homology is set by BLAST or Identity matrices; in these matrices, weight of coinciding bases is always higher than the weight of non-coinciding ones, and, when using the Fourier transformation, the search for homology is set by laws of autocorrelation function construction.[22] Earlier we have shown[19] that the homology search that uses quantitative measures could miss a hidden periodicity of length L in sequence S because of the lack of statistically significant homology between subsequences S^i; also we have shown that periodicity can be recognized only if a sufficient number of S^i periods is available.

Let us consider a notion of latent periodicity in more detail. For introduction of the quantitative measure of subsequences S^i homology, we need to construct a multiple alignment without inserts and deletions of symbols and arrange these subsequences in sequential order. The total weight of this multiple alignment, which can be considered as the quantitative measure of similarity of subsequences S^i, can be introduced as the sum of weights for all positions:

$$W = \sum_{i=1}^{L} W_i \tag{1}$$

The traditional approach calculates position weight as the sum of weights of all possible pairs of DNA bases that could appear while comparing the sequences:

$$W_i = \sum_{\alpha} \sum_{\beta>\alpha} P\left(s_i^{\alpha}, s_i^{\beta}\right) \tag{2}$$

where α and β show numbers of subsequences S^i; P is some weight matrix such as BLAST or Identity matrices. This expression can also be introduced as the following sum:

$$W_i = \frac{1}{2} \sum_{l,k} m(i,l)\left(m(i,k) - \delta_l^k\right) P(l,k) \tag{3}$$

where δ_l^k is 1 at $l = k$ and 0 in all other cases. The function δ_l^k is introduced to exclude from consideration the similarity of the subsequence S^i to itself. Variable values l and k show a type of DNA bases; $m(i, l)$ represents amount of base type l in the position i of multiple alignment. Earlier we proposed another measure of similarity,[23–25] which may be defined as the 'information content'[26]:

$$W'_i = \sum_{l=1}^{4} m(i,l) \ln \frac{K m(i,l)}{x(i)\, y(l)} \tag{4}$$

where $K = NL$, $x_i = \sum_{l=1}^{4} m(i,l)$ and $y_l = \sum_{i=1}^{L} m(i,l)$.

It is clear that the measures of homology of subsequences S^i determined by formulas (3) and (4) are different and, thus, an alignment could have higher weight using formulas (1) and (4) and lower weight using formulas (1) and (3), and *vice versa*. However, the term 'high weight' is less informative especially during comparison of weights determined by means of different mathematical measures. For each of the introduced measures we should determine the probability P that the weight (higher or equal to W) would be found during alignment of purely random sequences. In the case of periodicity search for length L in R independent sequences (e.g. analysis of R sequences from the Genbank database) the probability f should be considered instead of probability P:

$$f = 1 - (1 - p)^R \tag{5}$$

For evaluation of probability P during a search for the period of length L in the sequence S using the measures determined by formulas (3) and (4), we can randomly mix initial sequence S and create many random sequences Q_i $(i = 1, 2, \ldots)$ with length equal to the length of sequence S. Using many such sequences Q_i, we can determine the value Z as:

$$Z = \frac{W - E(W)}{\sqrt{D(W)}} \tag{6}$$

where $E(W)$ and $D(W)$ are the mean and variance of the weight W, respectively; they are calculated for many random sequences Q_i. High values of Z correspond to all lower values of probability P; they suggest the existence of significant similarity between subsequences S^1, S^2, ..., S^N. For evaluation of only non-accidental similarity, some threshold value of Z corresponding to the probability $P < 0.05$ is usually defined. Homologies with Z-value exceeding threshold level are considered as non-random.

As it was mentioned above, different weights and different Z-values could result in differences in homology level. In some cases, periodicity could be evident using the information measure [formula (4)] and at the same time could be missed by the methods of homology search [formula (3)]. Let us define the probability P that corresponds to Z-value calculated by formulas (3), (1), and (6) as α and define the probability P that corresponds to Z-value calculated by formulas (4), (1), and (6) as β. Let us also assume that the sequence S contains hidden periodicity of length L provided that the probability $\alpha > 0.05$ and the probability $\beta < 0.05$. Suppose that the sequence S contains periodicity related to homology between subsequences S^i at $\alpha < 0.05$ irrespectively of the probability values β. Let us also suppose that the sequence S lacks periodicity of the length L at α and $\beta > 0.05$.

In practice such a difference in probabilities α and β can be found quite often while analyzing rather long sequence when some set of symbols, not only one, appears at each position of the alignment. In this case, the number of homologous coincidences can be relatively small for each position. Since the weights of homologous

coincidences in BLAST or Identity matrices are significantly higher than the weights of non-homologous coincidences, the final value of W can be relatively small for small number of homologous coincidences; this will provide rather high value of α. At the same time, β is determined on the basis of deviations of DNA base frequencies for each position i from the DNA base frequencies determined for the whole sequence S being analyzed. Such deviations can be significant and will result in a small and statistically significant value of β for a high and statistically insignificant value of α.

As an example, let us consider two DNA sequences; one of them possesses perfect periodicity whereas the other one has latent periodicity. We determine α and β probabilities using the method of Monte Carlo. For each DNA sequence, we generate 500 random sequences with the same base composition as in the initial ones by means of random shuffling of bases throughout the whole DNA sequence. Using formulas (3) and (1), we calculate the weight W for each of the 500 randomly generated sequences, then determine $E(W)$ and $D(W)$ and finally calculate Z-value by formula (6). We perform the same calculations for the weight determined by formulas (4) and (1). As the result, we will have two Z-values for each DNA sequence: one value has been calculated by formulas (3) and (1) and the other one by formulas (4) and (1).

Let us take a DNA sequence containing 20 ($N = 20$) tandem repeats {atcgagt} as the first sequence. In subsequent consideration, this sequence of 140 bases in length will play a role of the sequence S while the length of subsequence S^i is equal to seven DNA bases; this means that we are looking for the period of seven DNA bases (Table 1). For the sequence S, we generate multiple alignment of periods; here they play the role of subsequence S^i. These periods are positioned one under the other. For this multiple alignment of periods, Z-value calculated by formulas (3), (1) and (6) using BLAST matrix is equal to 68.5 ± 3.5, whereas Z-value calculated by formulas (4), (1) and (6) is 55.2 ± 1.9. These calculations show that in the case of perfect periodicity both methods of weight function calculation give very similar results that are indistinguishable within the error of calculation by the Monte Carlo method.

As a second example, we consider a DNA sequence given below, which is characterized by the existence of latent period of seven symbols in length ($L = 7$, $N = 20$). Let us assume that DNA bases listed in Table 2 could appear in each position of the period with equal probability. One of possible sequences in which such periodicity is actually observed is the following:

ACCTACATGGGTTTTAGTATGTTCTACTCGG-
ACTACACCCTATTCTCCGCCTCTTGTGGTTCTT-
GTGCCGTGCCCCTTCTTACTTACCCCTTTTGCC-
GTATGCTAAAGATCGAATCCTGTCTAACTTTG-
AATTAAGTATT

Since this period is a hidden one, it is impossible to reveal the periodicity 'visually' by homology between separate periods. Let us evaluate probability values α and β for this sequence. For this sequence, Z-values determined by the Monte Carlo method are 2.5 ± 0.6 and 8.0 ± 0.7. Using normal distribution for the evaluation of distribution of Z-value, we obtain the value of

Table 1. Matrix used for generation of artificial sequence with hidden periodicity of seven DNA bases in length

Position of period	A set of DNA bases that could appear in a given position of the period
1	atc
2	agct
3	ctg
4	agct
5	cta
6	ct
7	gat

In each position of the period probabilities of selection of any base are equal to each other.

Table 2. The distribution of periodic sequences found by the groups of organisms to which they belong

Category	Subcategory	Number of sequences
Archaea	Crenarchaeota	9
	Euryarchaeota	20
Bacteria	Actinobacteria	61
	Bacteroid	6
	Chlamydiae	4
	Chlorobi	2
	Cyanobacteria	7
	Deinococcus-Thermus	1
	Thermotogae	1
	Firmicutes	
	Firmicutes Bacillales	33
	Firmicutes Clostridia	8
	Firmicutes Lactobacillales	22
	Firmicutes Mollicutes	4
	Planctomyces	8
	Proteobacteria	
	Alphaproteobacteria	43
	Betaproteobacteria	26
	Deltaproteobacteria	2
	Epsilonproteobacteria	21
	Gammaproteobacteria	158
	Spirochaetes	14
	Uncultured bacterium	3
Plasmids	Plasmids	1

probability $\alpha > 0.05$ and the value of probability $\beta < 10^{-9}$. Thus, it is clear that such periodicity will be statistically insignificant if we use the weight introduced on the basis of BLAST matrix [formulas (3) and (1)] and that it is revealed at a statistically significant level using the information measure introduced by formulas (4) and (1).

To make a search of the latent periods in symbolical sequences we are going to use the idea of mutual information. This idea is the keystone of the ID method. 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 have 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,..., n, 1,2...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 $(n \times k)$ matrix M 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 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

$$I = \sum_{1}^{n}\sum_{1}^{k} M(i,j)\ln M(i,j) - \sum_{1}^{n} x(i)\ln x(i) - \sum_{1}^{k} y(j)\ln y(j) + L\ln L \tag{7}$$

For ID construction it is necessary to take into account that the value $2\mathrm{I}(n, k)$ is distributed as χ^2 with $(n-1)(k-1)$ degrees of freedom.[26] To estimate the statistical significance of $\mathrm{I}(n, k)$ the Monte Carlo method is used by means of $Z(n, k)$ calculation using formula

$$Z(n,k) = \frac{\{\mathrm{I}(n,k) - \overline{\mathrm{I}(n,k)}\}}{\sqrt{D(\mathrm{I}(n,k))}} \tag{8}$$

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

The spectrum $Z(n,k)$ is similar to a spectrum of Fourier transformation for numerical sequences but has the following advantages: (i) the calculation of the spectrum does not require any transformation of a symbolical sequence to numerical sequences; (ii) 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; (iii) the statistical significance of long periods is not spread onto the statistical significance of shorter periods; (iv) on the basis of the matrix M it is possible to determine the type of periodicity (Fig. 1).

The window with the size equal to 2000 nt was used for scanning the sequence. The length L of the sequence under consideration was varied in order to find the region of genomic sequence possessing the highest level of statistical significance. The maximum possible value of the length L was equal to the window size.

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.

2.2. *Algorithm of the latent periodicity classification*

Each periodic region found by ID corresponds to its positional frequency matrix M'. Since we consider the periods of length equal to 2, this matrix can be represented as a vector as in Fig. 2.

A block diagram for the algorithm of the latent periodicity classification is shown in Fig. 3.

Since the regions of the latent periodicity were of different length, all compared matrices have been normalized to unity. We denoted the period length as N ($N = 2$), each matrix of the latent periodicity has been represented as a vector of nucleotides' frequencies distributed over $4N$ ranks. We chose the Pearson statistics as a measure of the similarity of two vectors (or as a distance between them). To use this statistics, we build two matrices—M1 and M2.

Let us denote a matrix obtained from a combination of such two vectors as M1. It has the marginal frequencies $X(i) = \Sigma_{\mathrm{j}}\,\mathrm{M1}(i, j)$, and $Y(j) = \Sigma_{\mathrm{i}}\,\mathrm{M1}(i, j)$, where $\Sigma_i X(i) = \Sigma_j Y(j) = 2$.

	1	2
A	1/2	1/4
T	0	1/4
C	1/2	1/4
G	0	1/4

Figure 1. Matrix T shows the symbol frequencies of the sequence with the latent periodicity.

A_1	T_1	C_1	G_1	A_2	T_2	C_2	G_2

Figure 2. Periodicity matrix represented as a vector. K_i ($K = \{\mathrm{A,T,C,G}\}$, $i = \{1,2\}$) represents the frequency (number of times the symbol is observed) of symbol K in i-th position of the period for the region found.

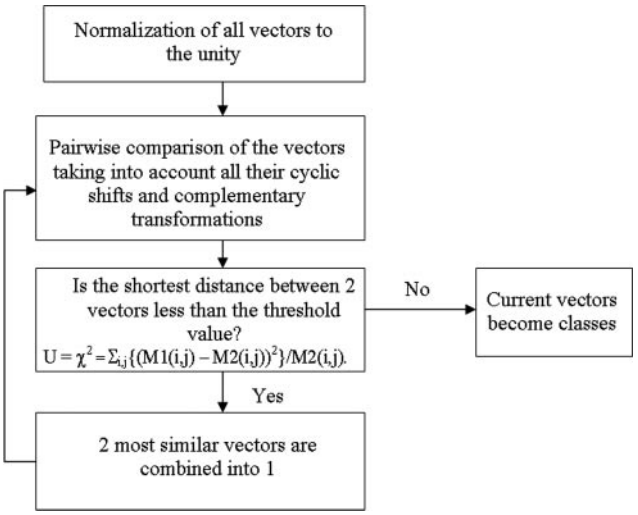

Figure 3. Block diagram of the latent periodicity matrices classification.

A matrix M2 has been constructed as the expected one over a set of the random matrices, having the same marginal quantities $X(i)$ and $Y(j)$ as M1. Its elements are defined as $\mathrm{M2}(i,j) = (1/2)X(i) \times Y(j)$. The Pearson statistics, whose values' distribution follows the χ^2, allows the estimation of the deviation of quantities in matrix M1 from expected ones in M2 matrix:

$$U = \chi^2 = \frac{\sum_{i,j}\{(\mathrm{M1}(i,j) - \mathrm{M2}(i,j))^2\}}{M2(i,j)} \qquad (9)$$

The number of the χ^2 degrees of freedom was equal to $(2 \times 4N) - 1$, that is, the number of comparison ranks (the number of matrix M1 or M2 elements) minus the number of independent linkages—a single claim on constancy of marginal elements: $X(1) = X(2) = 1$.

The pairwise comparison has been made between the vectors shown in Fig. 4. The lower index in Fig. 4 corresponds to the period position; the upper one reflects an ordinal number of the compared vector.

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. While making a comparison, we took into consideration all cyclic permutations of vectors' columns, which was necessary because of uncertainty of the period start position. A possibility of classic DNA inversions was also considered in this calculation. To achieve this, we replaced one of the vectors being compared by its complementary and inverse variant.

In a second step, we chose the two vectors for which the value of the Pearson statistics was minimal. If this value corresponded to accidental probability of less than or equal to 5%, then these two vectors were combined via recapitulation of their elements. The elements of a new

A_1^1	T_1^1	C_1^1	G_1^1	A_2^1	T_2^1	C_2^1	G_2^1	X(1)
A_1^2	T_1^2	C_1^2	G_1^2	A_2^2	T_2^2	C_2^2	G_2^2	X(2)
Y(1)	Y(2)	Y(3)	Y(4)	Y(5)	Y(6)	Y(7)	Y(8)	

Figure 4. A scheme of comparison between the two latent dinucleotide periodicity matrices is shown. Both matrices are presented as $4N$-dimensional vectors. X and Y are marginal frequencies, $X(i) = \Sigma_j\, \mathrm{M1}(i,j)$, and $Y(j) = \Sigma_i\, \mathrm{M1}(i,j)$, where $\Sigma_i X(i) = \Sigma_j Y(j) = 2$.

vector were calculated as weighted sums of the elements of two source vectors. The contribution of the specific vector to the sum was greater, the higher the number of vectors that had been already merged into it. A cyclic permutation, fixed inverse and complementary transformation were considered while making the vectors' combination. Such a new vector was then normalized again to unity. 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 χ^2 value corresponding to 5% level. The vectors that were left up to this moment were considered as the periodicity classes.

The classes obtained were then represented on a tree diagram in Fig. 5. The χ^2 value was chosen as the measure of dissimilarity between the class matrices in pairwise comparison. We used the minimum-variance method[27] to build the tree diagram of classes' similarity.

Let us note that critical level of the χ^2 value was estimated as the result of all $2N$ trials in searching for pairwise vector similarity. An accidental probability of similarity found in $2N$ trials: $\alpha = 1 - (1 - P)^{2N}$ should be less than or equal to 5%. From this point a critical level of accidental probability in one trial P was calculated by using the inverse χ^2 function.

2.3. Modified profile analysis

We have used the dynamic profile alignment approach,[21] which takes into account divergence of sequences as due to spot mutations and also indels. The method unites algorithms of dynamic programming for finding the best alignment[28] with analysis of position specific nucleotides as in a profile.[29] Since optimal alignment of analysed sequence has been built against a 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). The procedure for building the PSWM is similar in a way with the one described in ref. (20), but there are some differences that will be described below.

In order to find periodic sequences in GenBank we have built position-specific frequency matrixes of the

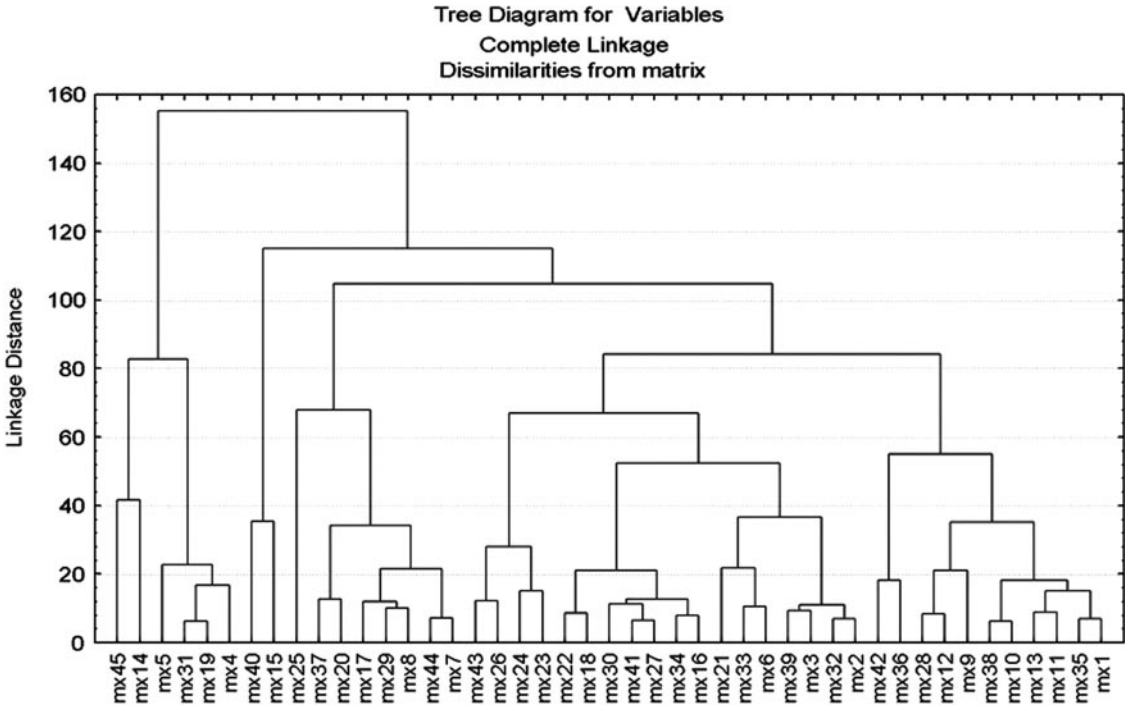

Figure 5. A tree diagram for the 45 classes of dinucleotide periodicity. The χ^2 value was chosen as the measure of dissimilarity between the class matrices in pairwise comparison. We used the minimum-variance method to build the tree diagram of classes' similarity.

nucleotides (PSMs) for each of the periodicity classes obtained in the previous step. To prevent distortion of base frequencies in isotypical periodic sequences, (effect of unequal representation) we have used the Sibbald and Argos algorithm[30] to weigh up each sequence inversely proportional to the number of sequences that are highly similar to this sequence.

Obtained PSMs have been converted to PSWMs in accordance with the formulae:

$$w'(S,j) = f(S,j)\ln\{f(S,j)/p(S)\} \qquad (10)$$
$$w(S,j) = w'(S,j) - \tilde{w}'(j)$$

where $S = \{\mathrm{A,T,C,G}\}$, $f(S,j)$—element of PSM matrix, $P(S) = \sum_j f(S,j)$, $\tilde{w}'(j) = 0.25\sum_S w'(S,j)$—the mean weight of the bases in the j-column of the PSM matrix, $w(S,j)$—a weight of base S in the PSWM matrix.

If we use w' only, rare nucleotides would have almost zero weight. To avoid this, we found mean weight for each position and subtracted it from w' to obtain the modified weight w. In this case, even very rare nucleotides would have negative weight, though the function w at small frequencies $f(S, j)$ is somewhat non-monotomic.

Therefore the transition to PSWM assists in assigning the higher weight to rarely occurring bases in the periodicity class when they have high occurrence frequency at the given position in the PSM and vice versa, decreasing the weight of bases with low occurrence frequency at the given position.[20]

2.4. *Algorithm of dynamic alignment of PSWM*

We have used a dynamic algorithm for finding local similarity,[28] also known as the Smith–Waterman algorithm,[31] to align the GenBank sequence being analysed against PSWM. Elements of the alignment matrix have been defined in accordance with formulae:

$$F(i,j) = \max\Big\{\max_{1\le k\le d\max}\{F(i-k,j) - v_d(1+\log(k))\};$$
$$\max_{1\le k\le d\max}\{F(i,j-l) - v_d(1+\log(l))\};$$
$$F(i-1,j-1) + w(S(i),j); 0.0\Big\};$$
$$F(0,0)F(0,0) = 0.0; F(i,0) = F(0,0) - v_d(1+\log(i));$$
$$F(0,j) = F(0,0) - v_d(1+\log(j)) \qquad (11)$$

where i is the position of the base in the analysed sequence, j is the position in consensus, $d\max = 40$ is the maximal analysed length of indels, $v_d = 1.0$ is the penalty for opening indel, and $w(S(i),j)$ is the PSWM element, calculated through formulae (10), where $S(i)$ is a base at position i of the sequence being analysed. Scanning of PSWM through GenBank has been carried out with a step of 20 bases while the scanning window size (and thus the maximal possible length of subsequence) was equal to 1000 nt. The periodicity class consensus was reproduced as many times as it was required to match the length of maximal subsequence. At each step an analysed sequence has been aligned against PSWM. We filled the matrix $F(k,j)$ completely and then found its maximal element $f_{\max}(k_m, j_m)$. Depending on the $f_{\max}$

position, we determined the optimal alignment as the way from the maximal element to the first zero with coordinates (k_0, j_0). This allows us to find the 'maximal subsequence'.[28] An alignment obtained shows this maximal subsequence S_m, and the location of related PSWM is marked by the nucleotides of higher weight. In order to speed up the calculation, we limited the maximal nucleotide subsequence to the range $|k - j| < 30$. The cumulative weight of the found alignment was defined as:

$$W = \sum_{j_0}^{j_m} w(S(i), j) \tag{12}$$

2.5. Statistical significance of similarity

To determine the statistical significance of the found alignment of the GenBank sequence, the probability of alignment of the given PSWM against random sequences with the same symbol composition needs to be estimated. To do this, the alignment matrix F' was filled using the formulae

$$\begin{aligned} F'(i,j) = \max\Big\{ & \max_{1\le k\le d\max}\{F'(i-k,j) - v_d(1+\log(k))\}; \\ & \max_{1\le l\le d\max}\{F'(i,j-l) - v_d(1+\log(l))\}; \\ & F'(i-1,j-1) + w(S(i),j)\Big\}; \end{aligned}$$

$$\begin{aligned} & F'(0,0) = 0.0; F'(i,0) = F'(0,0) - v_d(1+\log(i)); \\ & F(0,j) = F'(0,0) - v_d(1+\log(j)) \end{aligned} \tag{13}$$

e.g. in a same way as in formulae (11) but without considering zero values for choosing the maximum. We then calculated $d = f(k_m, j_m) - f(k_0, j_0)$. If we use formulae (11), the distribution of d is close to normal.[32] We confirmed this by using the imitation modeling (Monte Carlo method). We generated 100 sequences with the same symbol composition (i.e. symbol frequencies and triplet correlation) as in the sequence under consideration and then filled F' matrix for each such sequence. In other words, to determine the statistical significance of the found local alignment we have assumed it to be a global one (by fixing its co-ordinates) during imitation modeling. In such a case most calculations could be carried out before the databank scanning that greatly speeds up the calculations.

As a measure of statistical significance, we have determined the Z-score; that is, the normalized deviation of the found sequence alignment weight from the average weight of random sequences alignment against PSWM:

$$Z = \frac{(W - M(W_{\text{rnd}}))}{\boldsymbol{\sigma}(W_{\text{rnd}})}, \tag{14}$$

where W—found alignment weight, W_{rnd}—random sequence alignment weight, M and $\boldsymbol{\sigma}$—average value and standard variance of W_{rnd}, respectively.

We set up the minimum length of a periodical sequence to be found equal to 40 bases.

All significant results revealed were filtered in order to exclude the overlapping sequences. If an overlapped region exceeded 30% of the length of the smaller of two overlapped alignments, only the alignment with the greater Z-score has been left.

2.6. Statistical test for the periodicity

Since the method of MPA ensures only the statistical significance of sequence similarity, not its periodicity, we had to perform additional statistical test for the sequences found by this method. To do this, we have calculated the ID spectrum for all the sequences found by MPA in the same way as it was described in Section 2.1. We used the Monte Carlo method to calculate the statistical significance. For each sequence we have generated 200 sequences by randomly shuffling its symbols and then calculated mean, variance and, finally, Z-score as it was described above. The sequence was considered to be periodic with period length equal to 2 if the value corresponding to this length in ID spectrum was maximal and also was equal to or greater than 7.0. This test ensures that the sequences showing the given parameters possess the dinucleotide periodicity at statistically significant level.

3. Results and discussion

3.1. Classification results for bacterial genomes

Making a search for the latent dinucleotide periodicity in prokaryotic genomes from the GenBank-137 by using the ID method, we have found 454 sequences possessing the periodicity of such a type at level of Z-score ≥ 5. In order to find non-random sequences of maximal length it is essential to choose Z-score providing <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 ~10 times greater than the bacterial sequences presented in GenBank. So we can conclude that the number of 'noisy' sequences (i.e. the sequences that are not significantly periodic) in the set of the sequences found with $Z \geq 5$ is <5%. The distribution of the sequences found by the group of organisms they belong to is shown in Table 2. Rather a big number of sequences belongs to Gammaproteobacteria (158) and Actinobacteria (61). Together they account for more than a half of the periodical sequences. The organisms whose sequences include the biggest number of periodical sequences are as follows: *E. coli* (31), *Bradyrhizobium japonicum* (18), *Xanthomonas campestris* (17), *Helicobacter pylori* (16) and *Xanthomonas axonopodis* (16), from which the last three belong to the pathogenic bacteria and two others are symbiotic ones.

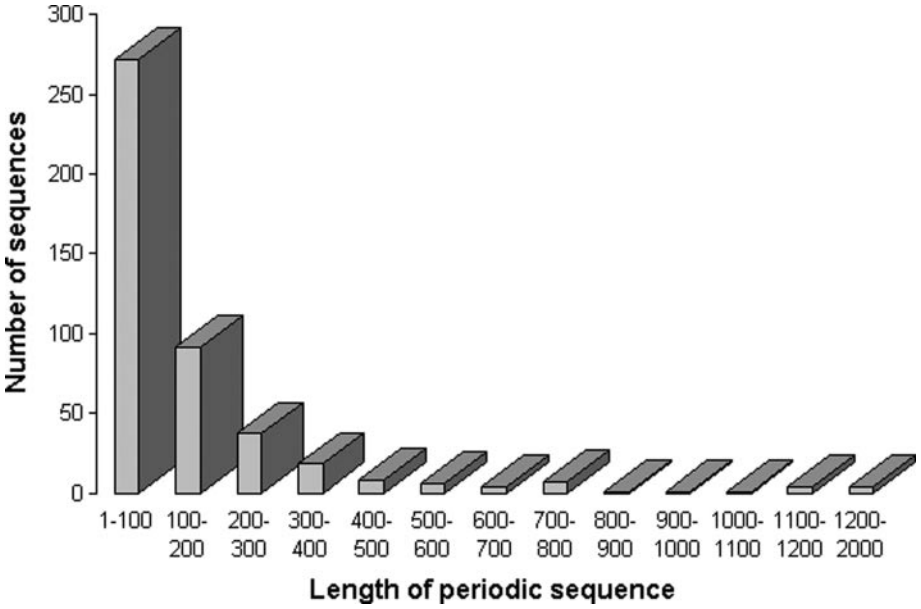

Figure 6. Length distribution of dinucleotide periodic sequences found in prokaryotic genomes.

Table 3. Distribution of the number of sequences in the classes obtained

Number of sequences	Number of classes containing the given number of sequences
3	22
4	5
5	5
6	5
7	2
8	2
9	2
11	1
19	1

A1	T1	C1	G1	A2	T2	C2	G2
17	9	35	18	12	15	21	32

	1	2
A	0,107	0,075
T	0,057	0,094
C	0,220	0,132
G	0,113	0,202

Figure 7. Non-normalized vector and normalized matrix of the latent periodicity class containing the biggest number of sequences which is equal to 19. Elements of the vector equal to the frequencies of appearance of a, t, c and g nucleotides in the first and second position of the period. The decimal numbers show the frequencies of corresponding nucleotides in period positions.

The length distribution of dinucleotide periodic sequences found in prokaryotic genomes is presented in Fig. 6. The length of most of the sequences (∼400) falls into the range 1–300, with the shortest one being equal to 28, but there are some examples of lengthy periodic sequences having length of more than 1000 nt.

All 454 sequences were classified by the latent period type as it was described above in Section 2. As a result, 45 classes have been discriminated; each of them combined three or more sequences of dinucleotide periodicity. The total number of sequences belonging to classes was 219. The distribution of the number of sequences contained in the classes is shown in Table 3.

We see that more than a half of the sequences that form the classes fall into two large classes, so the type of periodicity represented by the latter is quite common.

The tree diagram of class similarity is shown in Fig. 5.

The largest class combined 19 loci of latent dinucleotide periodicity (m × 10 on the tree diagram), next two largest classes contained 11 loci each (m × 32 and m × 38 on the tree diagram). The latent period type of the largest class is shown in Fig. 7. As we can see, cytosine and guanine are clearly predominated in both positions. Thus the period consensus may be conventionally described as {c,g}{c,g}. Classes of latent dinucleotide periodicity shown in Fig. 5 are combined according to the similarity in period type. Let us consider, for example, two extreme left and right groups of classes. The first of them combines the classes m × 45, m × 14, m × 5, m × 31, m × 19, m × 4, and the second combines m × 38, m × 10, m × 13, m × 11, m × 35, m × 1. The aggregation of classes in the first group takes place due to the significant frequency of adenine appearance in the first period position. The conventional consensus of combination is {}{n}, where n – any nucleotide from the set (a,t,c,g). The aggregation process in the second group is caused by the significant values of frequencies of cytosine and guanine in the first position of the

period. The combination conventional consensus in this case is $\{\}\{n\}$.

The presence of lengthy regions of latent dinucleotide periodicity in prokaryotic genes allows one to think that many prokaryotic genes show the high variability in those regions which have little (if any) influence on the functionality of the protein being coded. The mechanism of microsatellites origin seems to be connected with DNA strand sliding and with mispairing of neighboring repeats at the time of replication.[33] Owing to the short cycle of prokaryotic organism reproduction, the number of dinucleotide repeats grows rapidly, which leads to the appearance of the lengthy tracts. The nucleotide mutation rate should be high on the dinucleotide periodicity tracts because of the lack of selective constraints, and thus the dinucleotide periodicity, erodes rapidly becoming the latent periodicity. In addition, the existence of lengthy dinucleotide periodicity tracts can facilitate the improvement of the genome's physical properties, e.g. the raise of its flexibility, or, conversely, the rigidity, for certain DNA regions. Thus, the latent dinucleotide periodicity could be stabilized under the influence of yet other forces of natural selection.

All the class matrices obtained are available online as Supplementary Table 1 at www.dnaresearch.oxfordjournals.org.

3.2. Search with the MPA

We made the search using the method of MPA for 45 classes obtained in the previous step. Each class matrix was used for scanning through all bacterial loci in GenBank. Totally 27 087 corresponding to the frequency matrices of periodicity classes have been found. Among them there were both sequences possessing dinucleotide periodicity and many sequences possessing 6 nt periods of similar type (3 periods of 2 nt each). The presence of 6 symbol periodicity is caused by triplet periodicity of coding DNA regions (2 periods of 3 symbols each make a 6 symbol periodicity). Therefore, we filtered the overlapped sequences as it was described in the Section 2 and then use the statistical test described in the Section 2.6 to find whether the sequences found possess the dinucleotide periodicity at significant level. The number of filtered sequences that have passed the test was 3949. Such a big number in comparison with the 454 sequences that were found originally gives a reason to suggest that these sequences consist of strongly diverged tandem duplications. The periodicity of these possible ancient microsatellites is fuzzy, so it cannot be found by ID only.

All sequences found were aligned against the consensus sequence corresponding to the periodicity matrix. Examples of the alignments obtained are shown in Table 4. The ID spectra for the sequences from Table 4 are shown in Fig. 8. From this figure it is obvious that the maximal value corresponds to the period length equal to 2 and this value is >7. So these sequences possess the dinucleotide periodicity at statistically significant level. None of these sequences were found by other software packages.[15,16,34]

To get some biological insight, it is interesting to explore which functional elements of DNA contain periodicity and, inversely, which periodical sequences belong to which functional elements. We used the GenBank FEATURES field as a source of information for our study. Only the sequences with high statistical significance of their periodicity ($Z \geq 7.0$) were selected. The sequence was considered to belong to some functional element if its overlap with the region of corresponding feature was equal or greater than 30%.

The distribution of the number of periodical sequences overlapping with some functional elements of genome is shown in Table 5.

Since the sequences of interest were prokaryotic, it is not surprising that most of them belong to the gene regions, so these data are not shown in the Table 5. Also one can see that more than 100 sequences overlapped with sequence repeats were already detected empirically, so our method is proven to find such kind of repeats. What is really interesting is that periodical sequences were also found in promoters. The expected value for the number of promoters in the periodic sequences found is 32. It was calculated as a product of number of promoters in all bacterial genomes available and the fraction of periodic sequences in these genomes. Since the value obtained is 14, the abundance of promoters in periodic sequences is evidently not statistically significant. Nevertheless, the study of the periodic sequences found in promoters can be interesting since these sequences could appear to be ancient minisatellites.

3.3. Comparison with related works and discussion

A number of algorithms have already been proposed that either directly or indirectly detect tandem repeats. All of them suffer from significant limitations. One group of algorithms is based on computing the alignment matrices.[35–37] Their major drawback is excessive running time. But what is more important is that the methods that are based on using similarity matrices are able to identify only the repeats with high level of homology between repeat units. Similar phenomenon is described in ref. (38). The methods for searching the periodicity using Fourier transformation[39–41] are not able to identify the periodicity in presence of insertions and deletions in sequence and they do not produce the periodicity matrix that can be used for further analysis. Another group of algorithms finds tandem repeats indirectly using methods from the field of data compression. An algorithm by Milosavljevic and Jurka[42] detects 'simple sequences', i.e. mixtures of fragments that occur elsewhere. Simple sequences may or may not contain tandem repeats and this algorithm makes no attempt to deduce a repeated

Table 4. Examples of alignments of periodic sequences found

GenBank locus	Organism	Region description	Coordinates in sequence	Coordinates in consensus	Periodic sequence/consensus sequence	Z
AE012486	*Xanthomonas campestris*	Intergenic region	6393–6682	1–290	gagcgcgcttgcgcgcgatgaggcgttaccgggagagcttcatcgcgcgc gcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgc	8.9415
					gagcgcgctcctacgcggcatgtcttccggcctgatgccttcgactgccg gcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcg-cgcgcg	
					ggctttcccatgcacgcacgcgcgtcggagcaacaacacagac-cgcagg cgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcg	
					tgacatcacgcttgtgtaggagcgcgcttgcgcgcgatggggcgttatcg cgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcg	
					ggagagcttcatcgcgcgcgagcgcgctcctacgagacgggtcttccggc cgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcg	
					ctgaggccttcgattgccggcttgcccacgcacgcacgcgc cgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgc	
AF453480	*Burkholderia cepacia*	Gene	4166–4368	63–266	cgcgctcgcccgacgcgttcgcgcacgacggcatgacgccgctcgaacgc cgcgcgcgcgcg-cgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgc	7.4518
					gagctcgcgctcgcg-tcgttcgggccggcggccgcgcacgatcccgcgg gcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgc	
					acgacgccacgccgcacgccgcatccgcggctgccgacgccgccccgcgc gcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgc	
					gcgc-cgcacaatccgcatctcgcgcactggcgcacgcacctgaacgggc gcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgc	
					tcgc gcgc	
CJ11168X1	*Campylobacter jejuni*	Gene, polymorphism	176412–176535	163–287	gtgaatataaaaatgatcacataaaagagcgtatagaattgcgtttgcaa gcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgc	8.0281
					aataccatctcaaagcaaatcacagcaagcaaaaatg-gcttgagtttgc gcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgc	
					ttgcgagcataggtgctagcgcac gcgcgcgcgcgcgcgcgcgcgc	
CJ11168X1	*Campylobacter jejuni*	tRNA, Polymorphism	165728–165785	239–296	gcccattcgtctagcggttaggacatcgccctttcacg-gcggtaacacg gcgcgcgcg-cgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcgcg	7.5640
					agttcgag cgcgcgcg	

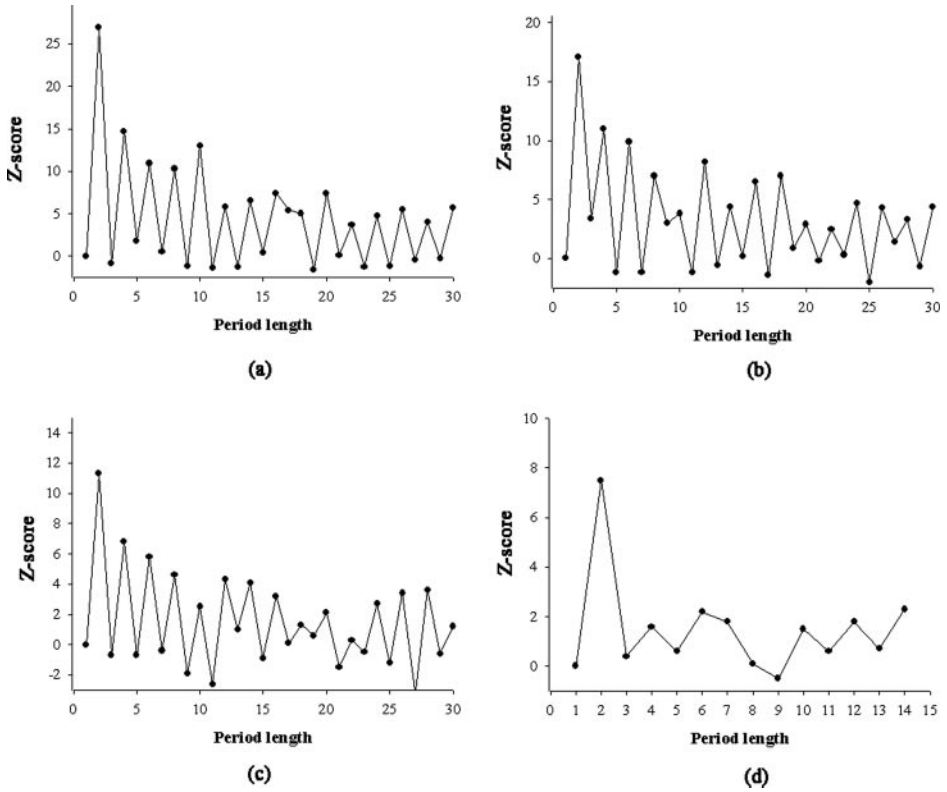

Figure 8. Information decomposition spectra for AE012486 (**a**); AF453480 (**b**); CJ11168X1, positions 176412–176535 (**c**); and CJ11168X1, positions 165728–165785 (**d**). The maximum at period length equal to 2 is clearly visible.

pattern. Some of the algorithms are sensitive dramatically to the insertions and deletions in sequence and, thus, can identify only the repeats obeying very strict rules.[43,44] The software programs for finding tandem repeats in genomic sequences available in the EMBOSS package[34] identify tandem repeats of a very limited type (certain microsatellites).

The major advantages of using the combination of our methods are that the size of the periodic sequence is not limited and is not specified beforehand; the methods are able to identify even very fuzzy repeats; the alignment matrices used are not filled entirely to speed the calculation. Though other methods developed share some of these advantages, to our knowledge, they are not able to identify the potential minisatellites found by using our approach.

The sequences shown in Table 4 have not been found to be periodic or to be minisatellites by other programs and methods available. We found 3949 possible minisatellite sequences in bacterial genomes; this is remarkably greater than the number of sequences already found. Because of the statistical threshold defined, we can conclude that >95% of the sequences found are periodic at statistically significant level.

As a first step, we use the ID method to find periodical sequences within bacterial genomes. Then we obtained classes of periodicity (defined by frequency matrix) to get some insight into the common patterns of periodicity for different organisms or group of organisms. The patterns, though rather fuzzy, do exist and we described this fact in details in the Results section. Since the repeats in bacterial genomes show high variability, most of them cannot be revealed by traditional methods and by the ID. MPA gives much better results, but it would take too much time to make a search for the periodicity with indels using all frequency matrices obtained using the ID.

Table 5. Distribution of some periodical sequences revealed in previously characterized DNA regions

GenBank feature	Number of sequences overlapping with the feature region
3′-UTR	1
mRNA	10
Promoter	14
rep_origin	26
repeat_region	115
repeat_unit	7
rRNA	51
sig_peptide	2
stem_loop	2
tRNA	60

If we take class matrices rather than all periodicity matrices, the time needed for the computations decreases dramatically. For bacterial genomes, using class matrices allows to speed up the calculation 10 times (45 class matrices versus 454 initial matrices). On the other hand, classes reflect more common properties of genome than individual matrices and, thus, can be used to reveal more periodical sequences. As we have shown in the Results section, some classes contain more than 10 sequences belonging to different organisms, so the genome properties they reflect are quite common. The number of sequences found by our method proves these suggestions.

Albeit the revealing of periodical sequences in genomes is an interesting and challenging task, the more important thing is to correlate the periodicity with possible biological functions and/or evolutional features. We have discussed above the difficulties in identification of ancient microsatellites and their importance for PCR analysis because of their highly polymorphic nature. In the Results section we showed some examples of periodical sequences found within the regions that had been already described as polymorphous. For each of these sequences we found the others defined by the same frequency matrix and, thus, possibly having the same nature. And it is possible to use the periodical sequences found with indels as a starting point for the PCR analysis, because they can turn out to be highly polymorphous ones. The study of possible ancient minisatellites may also be helpful for evolutionary analysis of genomes.

4. Conclusion

In this paper, we have presented a new method and algorithms for *de novo* identification of latent periodical sequences, which can be considered as potential microsatellites and minisatellites. A remarkable feature of this method is its ability to identify fuzzy or loose repeats (e.g. possible ancient microsatellites) that cannot be revealed by other methods.

Acknowledgements: The work presented in this paper was supported in part by the Russian Federal Ministry for Science and Education in the context of the 'System Biology' project (2005–2006).

Supplementary Data: Supplementary data is available online at http://dnaresearch.oxfordjournals.org.

References

1. Wells, R. 1996, Molecular basis of genetic instability of triplet repeats, *J. Biol. Chem.*, **271**, 2875–2878.
2. Weitzmann, M., Woodford, K., and Usdin, K. 1997, DNA secondary structures and the evolution of hypervariable tandem arrays, *J. Biol. Chem.*, **272**, 9517–9523.
3. Richards, R., Holman, K., Yu, S., and 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.
4. Lu, Q., Wallrath, L., Granok, H., and 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.
5. 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.
6. Frothingham, R. and Meeker-O'Connell, W. A. 1998, Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats, *Microbiology*, **144**, 1189–1196.
7. Supply, P., Mazars, E., Lesjean, S., Vincent, V., Gicquel, B., and Locht, C. 2000, Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome, *Mol. Microbiol.*, **36**, 762–771.
8. 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 Microbiology*, **1**, 2.
9. Toth, G., Gaspari, Z., and Jurka, J. 2000, Microsatellites in different eukaryotic genomes: survey and analysis, *Genome Res.*, **10**, 967–981.
10. Gur-Arie, R., Cohen, C. J., Eitan, Y., Shelef, L., Hallerman, E. M., and and Kashi, Y. 2000, Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism, *Genome Res.*, **10**, 62–71.
11. Dib, C., Faure, S., Fizames, C., et al. 1996, A comprehensive genetic map of the human genome based on 5,264 microsatellites, *Nature*, **380**, 149–152.
12. van Belkum, A., Scherer, S., van Leeuwen, W., Willemse, D., van Alphen, L., and Verbrugh, H. 1997, Variable number of tandem repeats in clinical strains of *Haemophilus influenzae*, *Infect. Immun.*, **65**, 5017–5027.
13. 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.
14. Benson, G. 2001, Tandem cyclic alignment, Proceedings of the 12th annual symposium on combinatorial pattern matching, *LNCS*, **2089**, 118–130.
15. Benson, G. 1999, Tandem repeats finder: a program to analyze DNA sequences, *Nucleic Acids Res.*, **27**, 573–580.
16. Kolpakov, R., Bana, G., and Kucherov, G. 2003, mreps: efficient and flexible detection of tandem repeats in DNA, *Nucleic Acids Res.*, **31**, 3672–3678.
17. Ruitberg, C. M., Reeder, D. J., and Butler, J. M. 2001, STRBase: a short tandem repeat DNA database for the human identity testing community, *Nucleic Acids Res.*, **29**, 320–322.
18. Boby, T., Patch, A.-M., and Aves, S. J. 2005, TRbase: a database relating tandem repeats to disease genes for the human genome, *Bioinformatics*, **21**, 811–816.
19. Korotkov, E. V., Korotkova, M. A., and Kudryashov, N. A. 2003, Information decomposition

method to analyze symbolical sequences, *Phys. Lett. A*, **312**, 198–210.
20. Frenkel, F. E., Chaley, M. B., Korotkov, E. V., and Skryabin, K. G. 2004, Evolution of tRNA-like sequences and genome variability, *Gene*, **335**, 57–71.
21. Korotkov, E. V., Korotkova, M. A., and Rudenko, V. M. 2000, MIR: family of repeats common for vertebrate genomes, *Mol. Biol. (Mosk)*, **34**, 553–559.
22. Dodin, G., Vandergheynst, P., Levoir, P., Cordier, C., and Marcourt, L. 2000, Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences, *J. Theor. Biol.*, **206**, 323–326.
23. Korotkova, M. A., Korotkov, E. V., and Rudenko, V. M. 1999, Latent periodicity of protein sequences, *J. Mol. Model.*, **5**, 103–115.
24. Korotkov, E. V., Korotkova, M. A., and Tulko, J. S. 1997, Latent sequence periodicity of some oncogenes and DNA-binding protein genes, *CABIOS*, **13**, 37–44.
25. Chaley, M. B., Korotkov, E. V., and Skryabin, K. G. 1999, Method revealing latent periodicity of the nucleotide sequences modified for a case of small samples, *DNA Res.*, **6**, 153–163.
26. Kullback, S. 1959, Information Theory and Statistics, John Wiley & Sons, New York.
27. Ward, J. H. 1963, Hierarchical grouping to optimize an objective function, *J. Amer. Stat. Assoc.*, **58**, 236–244.
28. Waterman, M. S. 1995, Introduction to Computational Biology. Map, Sequences and Genomes, Chapman & Hall, London.
29. Gribskov, M., McLachlan, A. D., and and Eisenberg, D. 1987, Profile analysis: detection of distantly related proteins, *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
30. Sibbald, P. R. and Argos, P. 1990, Weighting aligned protein or nucleic acid sequences to correct for unequal representation, *J. Mol. Biol.*, **216**, 813–818.
31. Smith, T. F. and Waterman, M. S. 1981, Identification of common molecular subsequences, *J. Mol. Biol.*, **147**, 195–197.
32. Webber, C. and Barton, G. J. 2001, Estimation of P-values for global alignments of protein sequences, *Bioinformatics*, **17**, 1158–1167.
33. Coggins, L. W. and O'Prey, M. 1989, DNA tertiary structures formed in vitro by misaligned hybridization of multiple tandem repeat sequences, *Nucleic Acids Res.*, **17**, 7417–7426.
34. Rice, P., Longden, I., and Bleasby, A. 2000, EMBOSS: The European molecular biology open software suite, *Trends Genet.*, **16**, 276–277.
35. Kannan, S. K. and Myers, E. W. 1996, An algorithm for locating nonoverlapping regions of maximum alignment score, *SIAM J. Comput.*, **25**, 648–662.
36. Benson, G. 1995, A space efficient algorithm for finding the best nonoverlapping alignment score, *Theor. Comput. Sci.*, **145**, 357–369.
37. Schmidt, J. P. 1998, All highest scoring paths in weighted grid graphs and their application to finding all approximate repeats in strings, *SIAM J. Comput.*, **27**, 972–992.
38. Laskin, A. A., Kudryashov, N. A., Skryabin, K. G., and Korotkov, E. V. 2005, Latent periodicity of serine–threonine and tyrosine protein kinases and other protein families, *Comput. Biol. Chem.*, **29**, 229–243.
39. Issac, B., Singh, H., Kaur, H., and Raghava, G. P. S. 2002, Locating probable genes using fourier transform approach, *Bioinformatics*, **18**, 196–197.
40. Chechetkin, V. R. and Lobzin, V. V. 1998, Nucleosome units and hidden periodicities in DNA sequences, *J. Biomol. Struct. Dyn.*, **15**, 937–947.
41. Jackson, J. H., George, R., and Herring, P. A. 2000, Vectors of Shannon information from Fourier signals characterizing base periodicity in genes and genomes, *Biochem. Biophys. Res. Commun.*, **268**, 289–292.
42. Milosavljevic, A. and Jurka, J. 1993, Discovering simple DNA sequences by the algorithmic significance method, *CABIOS*, **9**, 407–411.
43. Landau, G., Schmidt, J., and Sokol, D. 2001, An algorithm for approximate tandem repeats, *J. Comp. Biol.*, **8**, 1–18.
44. Subramanian, S., Mishra, R. K., and Singh, L. 2003, Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions, *Genome Biol.*, **4**, R13.