Complete spectra of periodicities in the problems of differentiation of closely related bacterial genomes

L A Miroshnichenko and V D Gusev
Sobolev Institute of mathematics SB RAS, Novosibirsk, Russia

E-mail: luba@math.nsc.ru, gusev@math.nsc.ru

Abstract. Tandem repeats (periodicities) constitute a significant part of the genomes of various organisms. They are often used as molecular markers in problems of differentiation of related objects. As such, bacteria of pseudotuberculosis and plague are considered. Genetically, the pathogen causing plague, *Yersinia pestis*, is very similar to *Yersinia pseudotuberculosis*. A comparison is made of the full spectra of periodicities in the genomes of bacteria of both species. Their general and species-specific features have been identified, confirming the high classification potential of the used object description.

1. Introduction
Sufficiently long repeating fragments are present in the genomes of various organisms [1]. They are important for the regulation of basic genetic processes during life and evolution [2]. Repeats are very diverse. So, for example, direct and symmetrical repeats are distinguished, repeats up to renaming elements of the alphabet, spread across the sequences and tandem repeating (periodicity). Often the tandem repetition is imperfect, periods may differ from each other by point mutations, distortions such as insertions and deletions are also possible [3,4]. In addition, fragments of sequences may possess the property of latent periodicity [5].

The object of the study in this work is the exact tandem repeats in DNA sequences with an arbitrary period length. The composition of tandem repeats in the genomes of many organisms (humans, mice, etc.) has been studied in sufficient detail. Unlike eukaryotic genomes, there are no extended periodicities in the bacterial genomes, such as satellites. So in the bacterial genomes under consideration, the longest perfect periodic fragments do not exceed 400 characters in length and there are almost no long periodicities with a period length shorter than 4.

Significant interest in tandem repeats is due to the fact that they are used as molecular markers in the intra- and interspecific classification of organisms [6–8]. In this regard, it seems relevant to obtain and study the full spectra of periodicities characterizing the genomes of different organisms. A significant factor is the visibility of these spectra for genomes of a sufficiently large length.

The aim of the article is to illustrate the possibilities of using the full spectra of exact periodicities for the differentiation of closely related objects, in particular, the bacteria *Yersinia pseudotuberculosis* and *Yersinia pestis*. *Yersinia pseudotuberculosis* is a gram-negative bacterium that causes Far East scarlet-like fever in humans. Genetically, the pathogen causing plague, *Yersinia pestis*, is very similar to *Yersinia pseudotuberculosis*. The plague appears to have evolved from *Y. pseudotuberculosis* more than 10000 years ago [9, 10]. The identification of even small structural and functional differences
between the genomes of these bacteria is of interest to specialists, since it can bring them closer to understanding the causes of such a high pathogenicity of the plague.

As the initial data, we used two collections of complete genomes characterizing the species *Yersinia pseudotuberculosis* and *Yersinia pestis* (14 genomes of each species) from GenBank (see table 1). The possibility of using a set of periodicities to describe and differentiate sufficiently long DNA sequences was first realized and confirmed by the authors of this article in [11]. We used periodicities in which monomers have the property of self-similarity [12]. In this paper, we use the full spectrum of periodicities, which enhances the capabilities of the approach and allows us to identify a number of interesting features of the description itself.

2. Local periodicities in bacterial genomes

Let $S$ be a DNA sequence composed of elements of the nucleotide alphabet $\Sigma = \{a, c, g, t\}$, $N = |S|$ is the length of $S$. By the local periodicity $P$ we mean the fragment $S$ formed by the tandem $k$-fold repetition of the basic (least possible of the acceptable variants) monomer $M$ ($k \geq 2$). For a brief presentation of the periodicity, it is convenient to use the notation $P = M^k$. For example, $P = tcagaagtcagaag = M^2$, where the monomer $M = tcagaag$, $p = |M| = 5$ and the repetition rate is $k = 3$. The notation $P = M^k$ can also be used when the periodicity ends with an incomplete monomer, and $k = |P|/|M|$ not an integer. So, $P = tcagaagtcagaag$ can be represented as $P = (tcagaag)^{13/5}$.

The periodicity detection algorithm is based on the construction of complexity decompositions in the sliding window mode. The complexity [13] of the sequence (text) $S$ is determined by the minimum number of steps ($c(S)$) required for the hypothetical process of “construction” $S$ using the operation of copying the maximum possible fragment from the already synthesized text prefix or generating a new character. The sequence of fragments formed at each step of the process, we call the "complexity decomposition" of the text $S$.

For example, let $S = ataatataaatatatatatatt$. The step-by-step process of constructing $S$ is as follows: $H(S) = a * t * a * ata * taatat * ttttt * aatatt$. Elements of the alphabet (in this case, the first two components “a” and “t”) are obtained using the operation of generating the character, the remaining components use the copy operation. The arrows in figure 1 explain the copying scheme. The beginning of each arrow indicates the starting position from which the next decomposition component is copied. The end of the arrow points to the first element of this component. The number of components in the decomposition determines the complexity of the sequence $c(S)$. In our example, $c(S) = 7$.

\[ S = a * t * a * ata * taatat * ttttt * aatatt \]

**Figure 1.** Scheme of complexity decomposition of the sequence $S = ataatataaatatatatatatt$.

The copying process, which began with any element of the synthesized prefix $S$, can be continued using the elements that have just been synthesized at this step. Thus, the ttttt component at figure 1 is constructed. This property underlies tandem repeat detection algorithms [14, 12]. In [12] the construction of complexity decomposition for detecting periodicities proceeds in the sliding window mode. The window size is calculated based on the length of the maximum repetition in the text, determined on the basis of the suffix tree [15]. This ensures that all periodicities are detected.

Each periodicity $P$ in the sequence $S$ is characterized by its position in the genome, the period (monomer) $M$, its length $p = |M|$ and the repetition factor $k$. The set of possible periodicities different in their positional bindings presented in $S$ will be called the complete spectrum of periodicities of the sequence $S$ and denoted by $\Pi(S)$. Moreover, monomers can coincide at different periodicities.

Bacterial genomes are about some million nucleotide pairs in length. Along with long nonrandom periodicities, they contain many short (random) tandem repeats. This "noise" background should be
filtered out. Only periodicities whose length exceeds a certain threshold \( |P| \geq r \) are relevant. Analysis of randomized analogues of bacterial genomes obtained as a result of random mixing of characters showed that there are practically no periodicities of length 20 or higher in them. In this article \( r = 40 \), i.e. we consider only periodicities with length \( |P| \geq 40 \), which guarantees their nonrandomness and visibility (on average, the genome accounts for about \( 10^2 \) such repeats).

Let \( n(p) \) be the number of periodicities in a DNA sequence with a periodicity length \( |P| \geq r \) and period (monomer length) \( p = |M| \). The full spectrum of periodicities \( \Pi(S) \) is characterized by the function \( n(p) \), where period \( p \) varies from 1 to the maximum value realized in a particular genome \( S \). In many cases, this information is already sufficient to differentiate the genomes of different species, in particular, \textit{Yersinia pseudotuberculosis} and \textit{Yersinia pestis}.

Let \( R = \sum_{p \geq 1} n(p) \) be the total number of periodicities in a DNA sequence which length \( |P| \geq r \). Table 1 presents the \( R \) values for 28 sequences.

| ID | strain                       | \( N \)   | \( R \) |
|----|------------------------------|----------|--------|
| 1  | NZ_CP009757 MD67             | 4729690  | 160    |
| 2  | NZ_CP009712 IP 32953         | 4743970  | 165    |
| 3  | NC_006155 IP 32953           | 4744670  | 174    |
| 4  | NC_010634 PB1/+              | 4695600  | 165    |
| 5  | NZ_CP009780 PB1/+            | 4695530  | 164    |
| 6  | NZ_CP009786 strain 1         | 4728290  | 173    |
| 7  | NZ_CP010067 PA3606           | 4742570  | 157    |
| 8  | NZ_CP031780 FDAARGOS_342    | 4905530  | 140    |
| 9  | NC_009708 IP 31758           | 4723248  | 145    |
| 10 | NC_010465 YPIII             | 4689440  | 137    |
| 11 | NZ_CP009792 YPIII           | 4689370  | 137    |
| 12 | NZ_CP008943 ATCC 6904        | 4806544  | 135    |
| 13 | NZ_CP009759 EP2/+           | 4706100  | 142    |
| 14 | NZ_LT596221 NZYP4713        | 4724230  | 152    |

| ID | strain                  | \( N \)   | \( R \) |
|----|-------------------------|----------|--------|
| 1  | NC_008149 Nepal516       | 4534590  | 96     |
| 2  | NC_003143 CO92           | 4653728  | 92     |
| 3  | NC_010159 Angola         | 4504254  | 64     |
| 4  | NZ_CP009704 Harbin35     | 4531804  | 94     |
| 5  | NZ_CP009785 El Dorado    | 4662886  | 90     |
| 6  | NZ_CP033690 FDAARGOS_603 | 4636496  | 96     |
| 7  | NZ_CP009906 Antiqua      | 4708021  | 89     |
| 8  | NZ_CP016273 Cadman       | 4602429  | 93     |
| 9  | NC_004088 KIM10+         | 4600755  | 88     |
| 10 | NC_017154 D106004        | 4640720  | 101    |
| 11 | NZ_CP009844 Dodson       | 4654228  | 91     |
| 12 | NZ_CP006783 1412         | 4524942  | 101    |
| 13 | NZ_CP009723 Shasta       | 4659530  | 94     |
| 14 | NZ_CP009492 PBM19        | 4685419  | 95     |

As an example, consider one sequences of each class: \textit{Yersinia pseudotuberculosis} strain MD67, complete genome (NZ_CP009757), \( N = 4729690 \) (\textit{Yersinia pseudotuberculosis}) and \textit{Yersinia pestis} Nepal516, complete genome, \( N = 4534590 \) (NC_008149). Table 2 shows the dependences \( n(p) \) for periodicities of length \( |P| \geq 40 \). Figure 1 shows the histograms \( n(p) \) for \( p = |M| \leq 23 \).
Table 2. The dependence of the number of long periodicities on the length of the period $p$ for Yersinia pseudotuberculosis strain MD67 and Yersinia pestis Nepal516. Here $n_1 = n(p)$ for the MD67 chain, $n_2 = n(p)$ for Nepal516.

| $p$ | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | >24 | R  |
|-----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| $n_1$| 0  | 0  | 9  | 15 | 11 | 1  | 1  | 2  | 15 | 6  | 12 | 34 | 19 | 11 | 4  | 3  | 2  | 2  | 4  | 160|
| $n_2$| 1  | 0  | 2  | 8  | 10 | 6  | 1  | 0  | 1  | 5  | 5  | 9  | 18 | 8  | 5  | 6  | 1  | 0  | 0  | 9  | 96 |

Table 2 and figure 2 shows that long periodicities most frequently encountered in the studied bacterial species are generated by monomers with lengths ranging from 6 to 22 nucleotides. There are no periodicities with a period length shorter than 4. Periodicities with a monomer length of $|M| > 22$ are found in most cases once and have a low repetition rate. So in Yersinia pestis Nepal516 two periodicities with $p = 124$ and periodicities with a period of 27, 30, 34, 36, 122, 141 and 143 are detected once. In Yersinia pseudotuberculosis MD67 there are periodicities with a period of 36, 38, 123 and 141. Tandem repeats with a period $p = 141$ in both genomes are located in the non-coding regions and are located between the genes encoding fructokinase and AAA family ATPase.

![Figure 2](image)

Figure 2. The dependence of the number of long periodicities on the length of the period for representatives of two types of bacteria

Not all monomer lengths are “equivalent” in terms of their representation in periodicities. The low values of $n(p)$ in Fig. 1 in the range of values $10 \leq p \leq 13$ indicates a non clear “avoidance” of the indicated periods by the known mechanisms of generating tandem repeats. The maximum lengths of monomers in the periodicities of both genomes are quite close: 141 (MD67) and 143 (Nepal516).

More than 20% of the periodicities from the MD67 genome (34/160) and about 15% of the periodicities from the Nepal516 genome (15/96) are part of the compounds. So, in the non-coding region of the MD67 genome, the fragment (ttataaa)$^{21}$(ttaaag)$^{15}$ was found, and in Nepal516: (caataagt)$^9$(caataagc)$^{24}$. The longest compound or imperfect periodicity with a period $p = 18$, was identified in both genomes in the BZ21_RS009960 gene encoding type VI secretion system tip protein VgrG. This fragment contains 7 perfect periodicities (MD67) and 5 (Nepal516) The amino acid sequence also forms an imperfect periodicity of the form SPPDPT-VPPDPT- (LPPDPT)$^2$-LPPETT-(APPETT)$^{27}$-APPEPT in the sequence MD67 and SPPDPT-VPPDPT- (LPPDPT)$^3$-LPPETT-(APPETT)$^{16}$-APPEPT in Nepal516.

With a relatively small difference in the lengths of the compared genomes, the difference in the number of periodicities found in them is very significant: 160 (in total) in the MD67 genome and only 94 in the Nepal516 genome. The $n(p)$ values given in the second row of Table 2 for the MD67 genome
exceed the similar (with the same \( p \)) values for the Nepal516 genome. The most significant differences are observed at \( p = 6, 15 \) and 19.

3. Comparison of the full spectra of periodicities of the genomes *Yersinia pseudotuberculosis* and *Yersinia pestis*

This section provides data on the number of periodicities of length \( |P| \geq 40 \) found in each of the 14 bacterial genomes of *Yersinia pseudotuberculosis* (Table 3) and *Yersinia pestis* (Table 4). Values \( p = |M| \) vary from 3 to 188. There are no periodicities of length \( |P| \geq 40 \) and period \( p < 3 \). A tandem repeat of 399 characters with a period of 188 was identified in only one sequence: *Yersinia pseudotuberculosis* strain ATCC 6904. In four genes of *Yersinia pseudotuberculosis*, periodicities with a length of 172 were found. For most sequences, the maximum value of period \( p \) is 141 or 143. The total number of periodicities with a period of \( p > 24 \) is quite small: from 4 to 21 in different sequences. Therefore, they are omitted in the presented tables. It should also be noted that in table 4 there are no rows with values \( p = 3, 5, 11, 13 \), present in the table. 3. This means that in the genomes of the plague bacterium no periodicities with a monomer of the indicated length were found. This fact can be used to differentiate genomes of both types.

### Table 3. The dependence \( n(p) \) of the number of periodicities of length \( |P| \geq 40 \) of the length of the period in the genomes of *Yersinia pseudotuberculosis*. The column numbers correspond to the species numbers listed in table 1.

| \( p \) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 1 | 1 | | | | | | | | | | | |
| 4 | | | | | | | | | | | | | | |
| 5 | 1 | 1 | | | | | | | | | | | |
| 6 | 9 | 4 | 4 | 4 | 4 | 2 | 4 | 2 | 2 | 3 | 3 | 8 | 3 | 3 |
| 7 | 15 | 15 | 14 | 17 | 17 | 19 | 11 | 9 | 15 | 10 | 10 | 11 | 5 | 17 |
| 8 | 11 | 8 | 9 | 11 | 11 | 7 | 9 | 10 | 9 | 6 | 9 | 10 | 13 | |
| 9 | 7 | 8 | 8 | 11 | 11 | 8 | 6 | 7 | 7 | 5 | 5 | 9 | 6 | 12 |
| 10 | 1 | | | | | | | | 2 | | | 3 | 2 | 1 |
| 11 | 1 | | | | | | | | 2 | | | 1 | 1 | |
| 12 | 1 | | | | | | | | | | | 1 | 1 | 1 |
| 13 | 1 | | | | | | | | | | | 1 | | |
| 14 | 2 | 1 | 1 | 2 | 2 | 3 | 2 | 1 | 1 | 1 | 1 | 2 | 4 | |
| 15 | 15 | 15 | 15 | 13 | 13 | 13 | 6 | 11 | 12 | 12 | 12 | 6 | 7 | 11 |
| 16 | 6 | 11 | 11 | 9 | 9 | 9 | 17 | 11 | 11 | 10 | 10 | 10 | 5 | 9 | 6 |
| 17 | 12 | 16 | 16 | 13 | 13 | 21 | 13 | 16 | 11 | 20 | 20 | 9 | 10 | 15 |
| 18 | 34 | 31 | 31 | 33 | 32 | 35 | 33 | 30 | 30 | 30 | 26 | 33 | 28 | |
| 19 | 19 | 16 | 17 | 16 | 16 | 10 | 10 | 9 | 8 | 9 | 9 | 10 | 9 | 14 |
| 20 | 11 | 11 | 11 | 13 | 13 | 14 | 15 | 6 | 13 | 9 | 9 | 8 | 7 | 10 |
| 21 | 4 | 8 | 8 | 4 | 4 | 8 | 4 | 7 | 3 | 4 | 4 | 9 | 7 | 5 |
| 22 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 1 | 1 | 4 | 4 | 5 | 6 | 3 |
| 23 | 2 | 3 | 3 | 2 | 2 | 3 | 3 | 2 | 3 | 1 | 1 | 1 | |
| 24 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 3 | 1 | |
| \( >24 \) | 4 | 12 | 21 | 8 | 6 | 17 | 14 | 14 | 11 | 11 | 11 | 10 | 16 | 4 |

\( R \) | 160 | 165 | 174 | 165 | 164 | 173 | 157 | 140 | 145 | 137 | 137 | 135 | 142 | 152 |
A comparative analysis of the periodicity spectra for a complete set of genomes (14 samples of each species) made it possible to identify both common characters and specific for each species. Common signs include:

- The dependence of the number of long periodicities in the genomes on the size of the repeating fragment (monomer) has two local maxima and a local minimum between them at $10 \leq p \leq 13$.
- The main maximum in the graphs of the dependence of the number of periodicities on the length of the period is observed at $p = |M| = 18$ for both species. The probable reason, perhaps not the only one, is the presence in the genomes of a rather extended zone, which once was an ideal periodicity with repeated repetition of a fragment of length 18. Over time, as a result of point distortions, it was broken into smaller periodicities with the same period length, but shorter repetition rate: from two to three. This tipped the balance towards the dominance of periodicities with $p = |M| = 18$.
- Among the periodicities with a long monomer length that are not covered by Tables 3 and 4, there are nevertheless separate instances present in almost all genomes of the initial selection. The periodicity with the period length $|M| = 36$, present in all genomes of both collections. Most bacterial genomes of Yersinia pseudotuberculosis and Yersinia pestis also have periodicities with a period of 124 and 141.
- Most periodicities are detected in non-coding regions.

**Table 4.** The dependence $n(p)$ of the number of periodicities of length $|P| \geq 40$ of the length of the period in the genomes of Yersinia pestis. The column numbers correspond to the species numbers listed in table 1.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 6 | 2 | 3 | 3 | 1 | 3 | 1 | 3 | 3 | 3 | 5 | 3 | 4 | 3 |
| 7 | 8 | 8 | 6 | 8 | 8 | 10 | 8 | 9 | 8 | 11 | 9 | 9 | 10 |
| 8 | 10 | 8 | 7 | 10 | 8 | 10 | 10 | 8 | 6 | 8 | 8 | 9 | 10 |
| 9 | 6 | 5 | 2 | 6 | 7 | 7 | 6 | 5 | 7 | 7 | 5 | 4 | 6 |
| 10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 12 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 14 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 |
| 15 | 5 | 5 | 5 | 4 | 5 | 4 | 5 | 5 | 6 | 5 | 7 | 5 | 5 |
| 16 | 5 | 4 | 2 | 6 | 3 | 5 | 2 | 6 | 4 | 2 | 4 | 5 | 5 |
| 17 | 9 | 7 | 4 | 9 | 7 | 9 | 7 | 9 | 9 | 7 | 11 | 7 | 8 |
| 18 | 18 | 20 | 12 | 19 | 19 | 17 | 20 | 20 | 14 | 18 | 19 | 16 | 19 |
| 19 | 8 | 6 | 2 | 7 | 6 | 9 | 6 | 6 | 6 | 7 | 6 | 10 | 6 |
| 20 | 5 | 4 | 7 | 5 | 5 | 5 | 4 | 5 | 5 | 5 | 4 | 3 | 4 |
| 21 | 6 | 6 | 4 | 6 | 7 | 5 | 6 | 7 | 6 | 6 | 8 | 6 | 6 |
| 22 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 1 | 1 | 1 |
| 23 | 0 | | | | | | | | | | | | |
| 24 | 0 | 1 | 1 | 1 | 1 | | | | | | | |
| >24 | 9 | 10 | 5 | 6 | 9 | 7 | 9 | 10 | 7 | 10 | 10 | 9 | 11 | 10 |

**R** 96 92 64 94 90 96 89 93 88 101 91 101 94 95
Of the features characteristic of only one of the two species of bacteria, we note:

- In the genomes of the bacterium Yersinia pseudotuberculosis, significantly more long (\(|P| \geq 40\)) periodicities were detected than in the genomes of the bacterium Yersinia pestis.
- There are periodicities characteristic of only one of the species. We point out, in particular, the periodicity with \(p = |M| = 41\) and 113, represented only in the genomes of Yersinia pestis (in 11 of the 14 genomes), as well as the tandem repeat with \(|M| = 123\), represented only in the genomes of Yersinia pseudotuberculosis (in 14 out of 14 cases). They are of particular interest for the differentiation of representatives of both species of bacteria.

4. Conclusion

Of great importance in the problems of intra- and interspecific classification of closely related microorganisms is the choice of the initial description of objects. The most reliable results are achieved using genome-wide sequencing of these objects (in particular, viruses and bacteria). But even with genome lengths of the order of \(10^4 - 10^6\) pairs of nucleotide residues, the problem arises of choosing an informative (in terms of classification) set of characters that allow us to divide the original objects into classes. It is proposed to use for these purposes the full and truncated spectra of tandem repeats (periodicities) represented in the genomes of the analyzed microorganisms. They provide a reasonable compromise between the complexity of processing and the visibility of the results. The presentation is illustrated by the example of a selection of the complete genomes of related bacteria - pathogens of pseudotuberculosis and plague. With the proximity of genomes in general, these objects fundamentally differ in the degree of pathogenicity for humans. The results can bring molecular biologists closer to a better understanding of the causes of the abnormally high pathogenicity of the plague pathogen.

Acknowledgments

The study was carried out within the framework of the state contract of the Sobolev Institute of Mathematics (project no. 0314-2019-0015).

References

[1] Britten R G and Korne D E 1968 Repeated sequences in DNA Science 161 pp 529–40
[2] Smirnov G V 2010 Repeats in bacterial genomes: evolutionary consideration Molecular Genetics, Microbiology and Virology 25(2) pp 56–65 (in Russian)
[3] Benson G 1999 Tandem repeats finder: a program to analyze DNA sequences, Nucleic Acids Research 27(2) pp 573–80
[4] Frenkel F E, Korotkova M A and Korotkov E V 2017 Database of periodic DNA regions in major genomes BioMed Research International 2017 ID 7949287
[5] Chaley M and Kutyrykin V 2016 Stochastic model of homogeneous coding and latent periodicity in DNA sequences. J. Theor. Biol. 390 pp 106–16. doi: 10.1016/j.jtbi.2015.11.014.
[6] Nakamura Y M, Leppert P, O’Connel R et al 1987 Variable number of tandem repeat (VNTR) markers for human gene mapping Science 235 pp 1616–22
[7] Usdin K 2008 The biological effects of simple tandem repeats: lessons from the repeat expansion diseases Genome Res. 18 (7) pp 1011–9.
[8] Belkum A, Scherer S, Alphen L and Vertrugh H 1998 Short-sequence DNA repeats in prokariotic genomes Microbiology and molecular biology reviews 62(2) pp 275–93
[9] Platonov M E, Evseeva V V, Dentovskaya S V and Anisimov A P 2013 Molecular typing of Yersinia pestis Molecular genetics, microbiology and virology 2 pp 2–12 (in Russian)
[10] Chain P S, Carniel E, Larimer F W et al 2004 Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotumerculosus Proc. Natl Acad Sci USA 101 (38) pp 13826–31
[11] Gusev V D, Miroshnichenko L A and Dzhioev Yu P 2019 Fractal characteristics of genomes as
an instrument of intra- and interspecific classification of organisms Proc. Int. Conf. Advanced Mathematics, Computations and Applications Mathematics (AMCA-2019) within Marchuk Scientific Readings (Novosibirsk) pp 122–6

[12] Gusev V D, Miroshnichenko L A and Chuzhanova N A 2009 Identification of fractal-like structures in DNA sequences Information Science & Computing. International Book Series 8: Classification, Forecasting, Data Mining (ITHEA, Sofia) pp 117–23

[13] Lempel A and Ziv J 1976 On the complexity of finite sequences IEEE Trans. Inform. Theory IT-22(1) pp 75–81

[14] Crochemore M and Rytter W 1994 Text Algorithms (Oxford University Press, New York) p 412

[15] Ukkonen E 1995 On-line constructing of suffix trees Algorithmica 14, pp 249–60