Computers in Biology and Medicine 69 (2016) 298–307

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Use of whole genome DNA spectrograms in bacterial classification

Vladimira Kubicova [a,*], Ivo Provaznik [a,b]

[a] Department of Biomedical Engineering, Brno University of Technology, Technicka 12, Brno 61600, Czech Republic
[b] International Clinical Research Center—Center of Biomedical Engineering, St. Anne's University Hospital Brno, Pekarska 53, Brno 65691, Czech Republic

ARTICLE INFO

Article history:
Received 14 November 2014
Accepted 29 April 2015

Keywords:
Relationships among bacteria
Whole genome comparison
SpectCMP method
DNA spectrogram
Spectrogram comparison

ABSTRACT

A spectrogram reflects the arrangement of nucleotides through the whole chromosome or genome. Our previous study suggested that the spectrogram of whole genome DNA sequences is a suitable tool for the determination of relationships among bacteria. Related bacteria have similar spectrograms, and similarity in spectrograms was measured using a color layout descriptor. Several parameters, such as the mapping of four bases into a spectrogram, the number of considered elements in the color layout descriptor, the color model of the image and the building tree method, can be changed. This study addresses the use of parameter selection to ensure the best classification results. The quality of the classification was measured by Matthew's correlation coefficient (MCC). The proposed method with optimal parameters (called SpectCMP—Spectrogram CoMParison method) achieved an average MCC of 0.73 at the phylum level. The SpectCMP method was also tested at the order level; the average MCC in the classification of class *Gammaproteobacteria* was 0.76. The success of a classification with respect to the correct phyla was compared to three methods that are used in bacterial phylogeny: the CVTree method, OGTree method and moment vector method. The results show that the SpectCMP method can be used in bacterial classification at various taxonomic levels.

1. Introduction

Identification of bacteria in medicine is essential for correct disease diagnosis, the treatment of infection and the trace-back of disease outbreaks that are associated with bacterial infections. Bacterial identification is used in the food industry for identification of a microbial contaminant that could be responsible for food spoilage, and it plays a key role in a wide variety of applications, including microbial forensics, bioterrorism threats and environmental studies [1,3]. It is generally agreed that the most useful way for scientists to organize organisms is to group them according to shared evolutionary history.

Today, the universally accepted DNA sequence-based method for phylogeny reconstruction is a method that is based on a 16S rRNA gene comparison [4,5]. The 16S rRNA genes are found in all organisms, and they are highly conserved among different species. However, bacteria can have multiple copies of this gene, and this circumstance can make interpretation difficult when base pair changes exist among copies. Moreover, some bacterial taxa have identical 16S rRNA genes [6]. Therefore, alternative genes were used as phylogenetic markers [7]. If the single genes are used to construct the phylogeny, the single gene trees are often inconsistent to each other [8]. Multi-gene approaches to phylogenetic analysis are presented in [9,10].

Phylogenetic information can be extracted from whole genome DNA sequences because single gene sequences do not contain sufficient information to construct an evolutionary history of organisms. If the entire genome is used instead of single genes, the data reflect the organism evolution and not the evolution of single genes. To compare single genes, sequence alignment is used. It does not make sense to align two complete genomes because every species has its own gene content and gene order in addition to different sizes of the genomes. In whole genome comparisons, identification of single nucleotide polymorphisms and insertion and deletion regions is used [11]. If the genomes share a set of common genes, those genomes can be compared by their gene content [8,12–14], gene order [15] or a combination of the gene order and gene content, e.g., the overlapping gene tree method (OGTree method) [16]. Previously mentioned methods are suitable for comparing closely related organisms. Two genomes can be compared by k-mer frequencies, e.g., the composition vector method (CVTree method) [17], feature frequency profile method [18] or return time distribution method [19]. A comparison of genomes by DNA graphical representation is presented in the moment vector method (MV method) [20] or frequency chaos game representation method [21].

* Corresponding author.
E-mail address: kubicova@feec.vutbr.cz (V. Kubicova).

http://dx.doi.org/10.1016/j.compbiomed.2015.04.038

In our previous work [22], a new approach to whole genome comparison was introduced; this method is based on spectrogram comparisons. A spectrogram gives a combined view of the local periodicity throughout the nucleotide sequence, and it was introduced as a tool for the visualization of DNA sequences in [23]. In spectrograms, some patterns, which are often related to the sequence function or structure, are observed. A periodicity of 10 bp reflects the DNA folding of bacteria [24]. A DNA sequence is often built from tandem repetitive regions (mostly in noncoding regions), which are clearly visible in a spectrogram as a series of horizontal lines. Therefore, the spectrogram was used as a tool for a tandem repetition search [25,26]. Furthermore, a spectrogram was used for the detection of protein-coding regions [27] because a strong periodicity at 3 bp exists. A suggested method that compares whole genome spectrograms is an alignment-free method, which could be used for comparing organisms that could be distantly related and do not share a common set of genes.

This work addresses parameter selection, which ensures the best possible results in bacterial classification. A method with an optimal parameter setting is called the Spectrogram CoMParison method (SpectCMP). The SpectCMP method was compared to three methods used in bacterial phylogeny: the CVTree method, OGTree method and MV method. Classification at the phylum and order level was studied.

2. Methods

A spectrogram computed from a genome visualizes the internal arrangement of bases and reveals structural characteristics. The coloring and patterns in a spectrogram reflect the properties of the DNA genome sequences.

In the method suggested in [22], a spectrogram was computed from a whole genomic DNA sequence; the DNA coding sections as well as the noncoding sections in the genome were considered. The distance between two bacteria was derived from the similarity of the spectrograms. A spectrogram from each studied genome was represented by its color layout descriptor (CLD), and the similarity in the CLD was counted by the Euclidean distance. The Euclidean distance of CLD created the elements of the distance matrix, and the building tree method was used to obtain a distance tree (Fig. 1). In a distance tree, similar bacteria (bacteria from the same taxon) should be placed in the same branch of the distance tree. In this method, several parameters can be changed.

2.1. Spectrogram as a tool for the visualization of genome properties

We investigated the properties of sequences by observing those properties that are readable from the whole genome spectrograms; these properties contribute to bacterial taxa differentiation. The relevant properties are the GC content, codon usage bias and strand asymmetry.

The GC content is represented in a spectrogram by the coloring of the spectrogram image. We observed that if the occurrence of guanine and cytosine is represented in a red and green layer in the color spectrogram, the dependence of the percentage of red and green sections of the pixels on the real GC content in the DNA sequence is almost linear (Fig. 2). The GC content could distinguish some bacterial taxa (e.g., the phylum *Deinococcus-Thermus* from the phyla *Thermotogae* and *Tenericutes* or the phylum *Tenericutes* from the phylum *Actinobacteria*). Some bacterial phyla, such as *Proteobacteria* and *Spirochaetes*, vary in their GC content; therefore, it is impossible to distinguish these from the other phyla (Fig. 3).

In whole genome bacterial spectrograms, a strong periodicity at 3 bp appears (Fig. 4). This periodicity reflects the coding regions in the DNA because synonymous codons can differ in their frequency of occurrence among the different genes within an organism. This arrangement creates a codon bias [28]. The nucleotide composition of the 3 bp line in the spectrogram can be determined approximately from its color; the resolution power on 3 bp

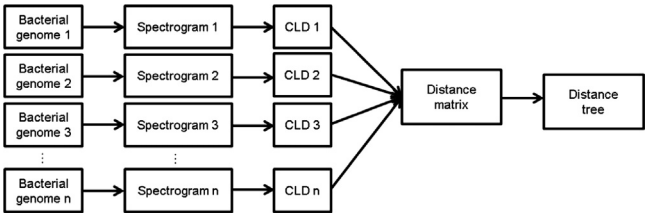

Fig. 1. Scheme of the method for determination of the relationship presented in [22]. The spectrogram is computed from the whole genome bacterial sequence, and the spectrograms are compared by the CLD to obtain the distance matrix. The distance tree, which represents the relationships of the bacteria, is built from the distance matrix.

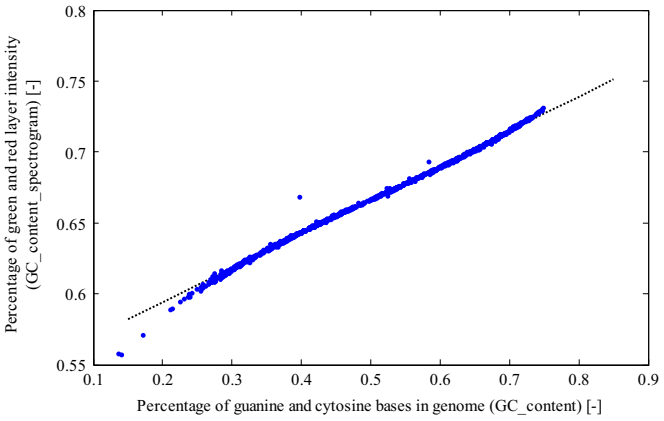

Fig. 2. Dependence of the percentage of the red and green layer of the color spectrogram on the GC content. The *x*-axis represents the real GC content in the genome (counted according to [30]), the *y*-axis represents the GC content as counted from the spectrogram – the percentage of the red and green layer from all three layers (red, green and blue) (assuming that the occurrence of guanine and cytosine is represented in the red and green layers in the color spectrogram). Each blue point in the scatter represents one bacterial genome. The black dotted line represents a linear fitting.

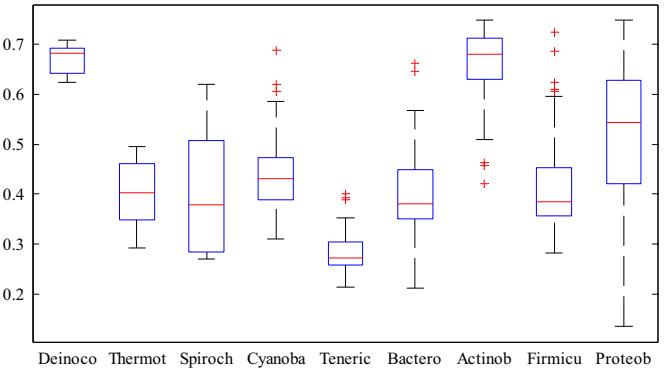

Fig. 3. Real GC content in bacterial phyla (counted according to [30]). For each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and whiskers extend to the most extreme data points that are not considered to be outliers. Outliers are plotted as red plus signs. Abbreviations of bacterial phyla: *Deinoco=Deinococcus-Thermus, Thermot=Thermotogae, Spiroch=Spirochaetes, Cyanoba=Cyanobacteria, Teneric=Tenericutes, Bactero=Bacteroidetes, Actinob=Actinobacteria, Firmicu=Firmicutes, and Proteob=Proteobacteria.*

depends on the window length. The color of the strip depends on the nucleotide composition and not on the quantity of the nucleotide; if the occurrence of thymine, cytosine and adenine is represented in the red, green and blue layer, respectively, the codons AAT and ATT appear in the spectrogram as the violet strip. Thus, there are seven different colors for 64 codons in the spectrogram. The white color is for the codons ACT, ATC, TAC, TCA, CTA, and CAT; the blue color is for the codons AAG, GAA, AGA, AGG, GAG, and GGA; the red color is for GGT, GTG, TGG, GTT, TGT, and TTG; the green color is for CGG, GCG, GGC, CCG, CGC, and GCC; the yellow color is for CTT, TCT, TTC, CCT, CTC, and TCC; the magenta color is for codons AAT, ATT and AGT; and the cyan-blue color is for AAC, ACC, and ACG. Our observation shows that the coloring of the line at 3 bp tends to be similar for organisms that belong to the same order. Codon usage differs among the orders; for example, the codon usage is different between the orders *Pseudomonadales* and *Enterobacteriales*, which belong to class *Gammaproteobacteria*, phylum *Proteobacteria*.

Strand asymmetry is associated with the single-origin mode of genome replication in bacterial genomes. The origin and terminus of the replication separate the genome into two regions that differ in the nucleotide composition. In the leading strand, guanine and thymine are preferred over cytosine and adenine, and vice versa in the lagging strand [29]. In a spectrogram, strand asymmetry is visible as a color intensity change. Spectrograms of some bacterial phyla show a strong color intensity change, e.g., *Firmicutes*, and some phyla do not, e.g., *Tenericutes* (Fig. 5).

2.2. Spectrogram computation

To compute the DNA spectrogram according to [31], a DNA sequence is divided into smaller overlapping segments. Each segment is first converted to four numerical vectors $u_A(n)$, $u_T(n)$, $u_C(n)$ and $u_G(n)$ using binary representation [23]. The vectors indicate the presence or absence of four nucleotides, A, T, C and G, respectively, at the nth position. Then, the spectra of the four binary vectors are computed by the discrete Fourier transform (DFT):

$$U_x(k) = \sum_{n=0}^{N-1} u_x(n) e^{-j(2\pi nk/N)}, \tag{1}$$

where N is the length of the binary vector $u_x(n)$, $n=0, 1, \ldots, N$ and k is the frequency. To obtain the color spectrogram, four spectra are reduced to three.

We investigated the impact of two chosen spectrum reducing methods:

$$\begin{aligned} x_R &= U_T[n] + \frac{1}{3} U_G[n], \\ x_G &= U_C[n] + \frac{1}{3} U_G[n], \\ x_B &= U_A[n] + \frac{1}{3} U_G[n]. \end{aligned} \tag{2}$$

$$\begin{aligned} x_R &= U_G[n] + \frac{1}{3} U_T[n], \\ x_G &= U_C[n] + \frac{1}{3} U_T[n], \\ x_B &= U_A[n] + \frac{1}{3} U_T[n]. \end{aligned} \tag{3}$$

where U_A, U_T, U_C, U_G are the spectra of the binary vectors u_A, u_T, u_C, u_G, respectively, and x_R, x_G, x_B represent three layers in the color spectrogram. The first spectrum reducing method in Eq. (2) maps thymine, cytosine and adenine occurrences into red, green and blue layers of the color spectrogram, respectively; guanine occurrences are registered in all three layers (Fig. 6A). The second spectrum reducing method in Eq. (3) maps the occurrences of guanine, cytosine and adenine to the red, green and blue layer, respectively, and the occurrences of thymine are in all three layers (Fig. 6B).

By computing the spectrum in short windows, sliding down a sequence, the spectrogram is constructed by depicting a single spectrum from one window as one column of a spectrogram

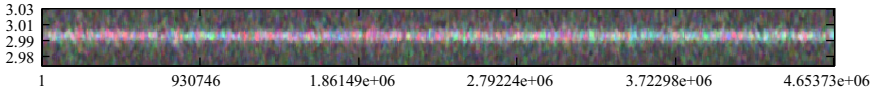

Fig. 4. Example of the 3 bp line in the spectrogram of *Yersinia pestis* (NC_003143). The occurrence of thymine, cytosine and adenine is represented in the red, green and blue layer, respectively, and the occurrence of guanine is represented in all three layers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

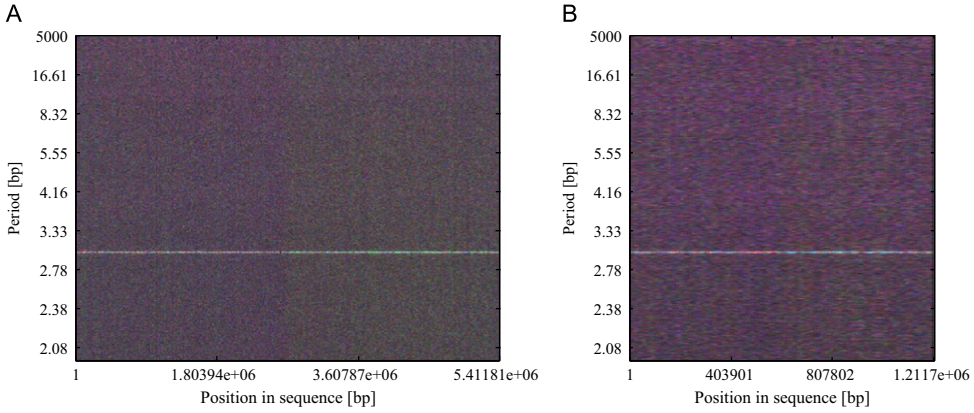

Fig. 5. Strand asymmetry in whole genome bacterial spectrograms. The occurrence of thymine, cytosine and adenine is represented in the red, green and blue layer, respectively, and the occurrence of guanine is represented in all three layers. (A) Spectrogram of *Bacillus cereus* (NC_004722, phylum Firmicutes) shows a strong color intensity change from violet to green, (B) spectrogram of *Mycoplasma mycoides* (NC_005364, phylum Tenericutes) does not show a color intensity change. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

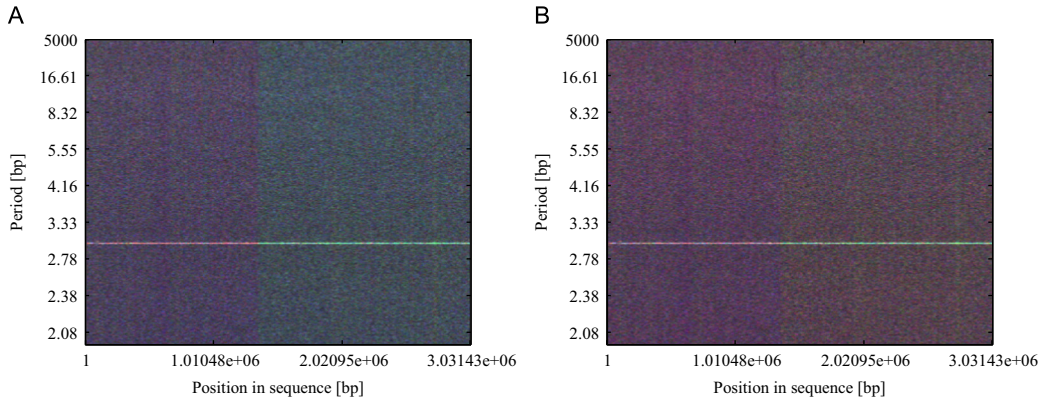

Fig. 6. Spectrogram of *Clostridium perfringens* (NC_003366). (A) Spectrogram constructed using the first type of spectrum reducing method: the occurrence of thymine, cytosine and adenine is mapped to the red, green and blue layer, respectively, and the occurrences of guanine are represented in all three layers (2). (B) Spectrogram constructed using the second type of spectrum reducing method: the occurrence of guanine, cytosine and adenine is mapped to the red, green and blue layer, respectively, and occurrences of thymine are represented in all three layers (3). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

image. Depicting many spectra from sliding windows, a spectrogram is obtained. Before rendering the spectrogram, each layer of the spectrogram was normalized because the DFT and reduction into three spectra do not guarantee values between 0 and 1. Normalization in two steps was performed according to [32]: all of the pixel values were divided three times by the global mean value of the pixels in all of the three layers. Then, the color pixel values that were greater than one were divided by the local maximum $\max(x_R, x_G, x_B)$.

To capture large patterns in the genome, the length of the sliding window was set to 5000 bp.

2.3. Spectrogram comparison by CLD

The relations between bacteria can be seen from the colors in the spectrograms; for closely related bacteria, the spectrograms showed similar color properties, which is in contrast to distant bacterial spectrograms, which are visually different. The image content could be represented by descriptors, and according to the MPEG-7 standard, the descriptors that are related to the color are: the dominant color descriptor, scalable color descriptor, color structure descriptor and CLD [33]. The spatial distribution of the colors reflects the properties of the genome. To capture this information, the CLD was used.

The image descriptor extraction process consists of four stages: dividing the image into smaller regions, representative color detection, DCT transformation and zigzag-scanning of the coefficients [33].

The spectrogram image is first split into n uniform rows and n uniform columns. We investigated the impact of the number of rows and columns n on the success of the classification. The average color is counted from each region, and the down-sampled image of n rows and n columns is obtained (Fig. 7). Before down-sampling, the number of rows in each spectrogram equals 2500 (which is half of the window length), and the number of columns depends on the length of the genomic sequence from which the spectrogram was computed. After down-sampling, the image dimension remains the same ($n \times n$).

According to the MPEG-7 standard, CLD is counted not in *RGB* color space but in YC_bC_r color space. A representative color image can be transformed to the YC_bC_r color space as follows:

$$Y = 0.299R + 0.587G + 0.114B$$
$$C_b = -0.169R - 0.331G + 0.5B$$
$$C_r = 0.5R - 0.419G - 0.081B \quad (4)$$

We investigated the impact of the color model (*RGB* and YC_bC_r) on the classification success.

Each of three matrixes (Y, C_b, C_r) is transformed by the discrete cosine transform (DCT) using the formula:

$$B_{pq} = \alpha_p \alpha_q \sum_{x=0}^{n-1} \sum_{y=0}^{n-1} Y_{xy} \cos \frac{\pi(2x+1)p}{2n} \cos \frac{\pi(2y+1)q}{2n}, \quad (5)$$

where n is the number of rows and columns of the input matrix Y, and p and q=0,1,…$n-1$. If p=0, then $\alpha_p = 1/\sqrt{M}$; otherwise, $\alpha_p = \sqrt{2/M}$; and if q=0, then $\alpha_q = 1/N$; otherwise, $\alpha_q = \sqrt{2/N}$.

Three matrixes of DCT coefficients are obtained, followed by zig-zag scanning. The purpose of the zig-zag scanning is to produce a DCT matrix coefficients vector. In this vector, low frequency coefficients are grouped at the beginning of the vector followed by higher frequency coefficients. The set of DCT coefficients is called the CLD. The distance between two descriptors can be computed by the Euclidean distance of DCT coefficients, as follows:

$$D = \sqrt{\sum_i (EY_i - EY_i')^2} + \sqrt{\sum_i (ECb_i - Cb_i')^2} + \sqrt{\sum_i (ECr_i - ECr_i')^2}, \quad (6)$$

where EY, ECb and ECr are three vectors of the first CLD, EY', ECb' and ECr' are three vectors of the second CLD, and i is number of considered elements in CLD. The distance D is one element of the distance matrix. In [34], 12 coefficients were considered (6 coefficients of Y, 3 for C_b and 3 for C_r). These coefficients capture low frequency information in the image. To distinguish the spectrograms, high frequency coefficients must be considered. We investigated the impact of the number of elements in CLD i on the success of the classification. If the distance matrix is calculated, then it is straightforward to construct a distance tree. In the literature, two building tree methods are often used in whole genome phylogeny: NJ (Neighbor-joining) and UPGMA (Unweighted Pair Group Method with Arithmetic mean).

2.4. Evaluation of a binary classification

We analyzed the classification of bacteria for each pair of studied phyla. The quality of the classification was measured by

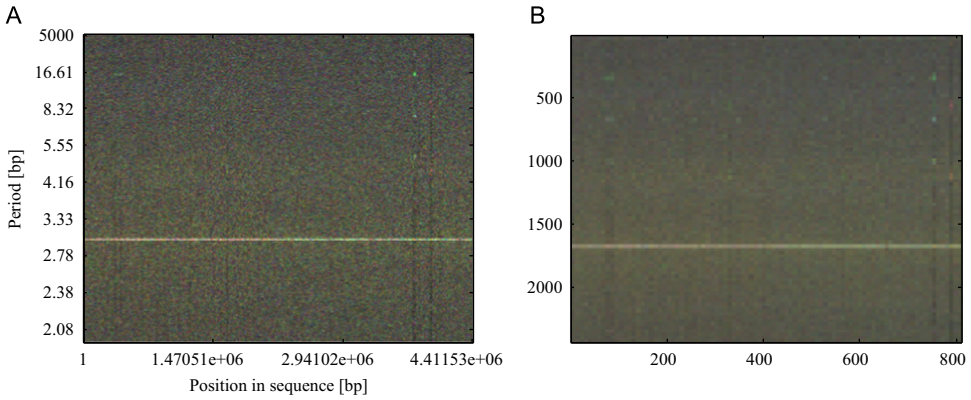

Fig. 7. Down-sampling of the spectrogram of *Mycobacterium tuberculosis* (NC_000962), (A) spectrogram before down-sampling, (B) spectrogram after down-sampling, where the number of rows and columns after down-sampling is $n=90$. After down-sampling, all of the spectrograms have the same dimension ($n \times n$). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Matthew's correlation coefficient (MCC) [35]. This coefficient is in essence a correlation coefficient between the observed and predicted binary classifications. MCC is calculated using the formula

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \quad (7)$$

where *TP* is the number of correctly classified bacteria from the first phylum, *TN* is the number of correctly classified bacteria from the second phylum, *FN* is the number of incorrectly classified bacteria from the first phylum, and *FP* is the number of incorrectly classified bacteria from the second phylum. MCC returns a value between -1 and $+1$, where $+1$ indicates a perfect prediction, 0 indicates a random prediction and -1 indicates total disagreement between the prediction and observation.

In Fig. 8A, the distance tree for *Actinobacteria* and *Firmicutes* is shown. The distance tree was constructed by the following parameters: the second type of spectra reduction method, the RGB color model, UPGMA as the building tree method, the number of regions $n=88$ and the number of DCT coefficients $i=1335$. The name of each node is created from the name of the phylum and the National Center for Biotechnology Information (NCBI) access number. Organisms from the phylum *Actinobacteria* are classified together in the upper branch, while organisms from the phylum *Firmicutes* are classified together in the lower branch. Eight *Actinobacteria* occur in the group *Firmicutes* incorrectly. In such a distance tree, $TP=15$, $TN=52$, $FP=0$, and $FN=8$. The success of sorting the organisms into two groups is measured by the probability that an organism was included in the correct branch. For example, Fig. 8A shows a distance tree in which the probability that an organism is from phylum *Firmicutes* is placed in the correct branch is $52/(52+0)=1$. On the other hand, the probability that an organism from the phylum *Actinobacteria* is placed in the correct branch is $15/(15+8)=0.652$. The quality of the classification expressed by MCC is 0.752.

For an optimal parameter selection, the quality of classification of bacteria into the correct phylum was studied. The impact of five parameters on the classification results was studied. We tested two types of spectrum reducing methods in Eqs. (2) and (3), and another three parameters were connected to the spectrogram comparison: the number of rows and columns n in the color representative image, the color model of the representative color image (*RGB* and YC_bC_r) and the number of considered elements in CLD i. The final tested parameter is the choice of building tree method; UPGMA and NJ methods were considered. Our aim was to choose conditions in which our method classifies the bacteria into the phyla correctly.

The bacteria chosen for analysis come from the NCBI list of reference bacteria, which was downloaded from the NCBI database in March 2013. For species that had sequenced genomes for more than one subspecies, we maintained one representative subspecies, which was chosen randomly. Then, four phyla with more than 10 organisms were chosen. This approach resulted in a data set of 166 prokaryotes—*Firmicutes* (52 organisms), *Actinobacteria* (23 organisms), *Tenericutes* (12 organisms) and *Proteobacteria* (79 organisms). The length of the whole genome sequence was between 742 431 bp and 10 467 782 bp. The total sequence length was 543 018 526 bp. The classification results produced by our method were compared to three methods used in bacterial phylogeny: the CVTree method, OGTree method and moment vector (MV) method. The capability of the classification was studied at the order level, also. A new dataset that was composed of 156 bacteria from phylum *Proteobacteria*, class *Gammaproteobacteria* was created.

3. Results and discussion

3.1. Parameter selection

In Fig. 9, the heat maps of MCC for different parameters are shown. Each element in a heat map represents the average MCC computed from 6 pairs of studied bacterial phyla: *Firmicutes* and *Actinobacteria*, *Firmicutes* and *Tenericutes*, *Firmicutes* and *Proteobacteria*, *Actinobacteria* and *Tenericutes*, *Actinobacteria* and *Proteobacteria*, and *Tenericutes* and *Proteobacteria*. In each heat map, the y-axis represents the number of rows and columns n in the color representative image, and the x-axis represents the number of elements in CLD i. The number of rows and columns n was set to be from 2 to 100 with a step of 2. The number of considered elements in CLD i was set to be from 5 to the maximum length of the CLD for a specific value of n; the result is $n \times n$. The step of parameter i was set to 5. The title of each heat map describes the color model of the representative color image and the building tree method. Four heat maps on the left (Fig. 9A–D) show the average MCC when the first method for spectra reduction in Eq. (2) was used. Four heat maps on the right (Fig. 9E–H) show the average MCC when the second method for spectra reduction in Eq. (3) was used.

The results show that for values n that are higher than 70, it is suitable to choose i to be lower than 4000. For i that is higher than 4000, the quality of the classification is low, and there are areas

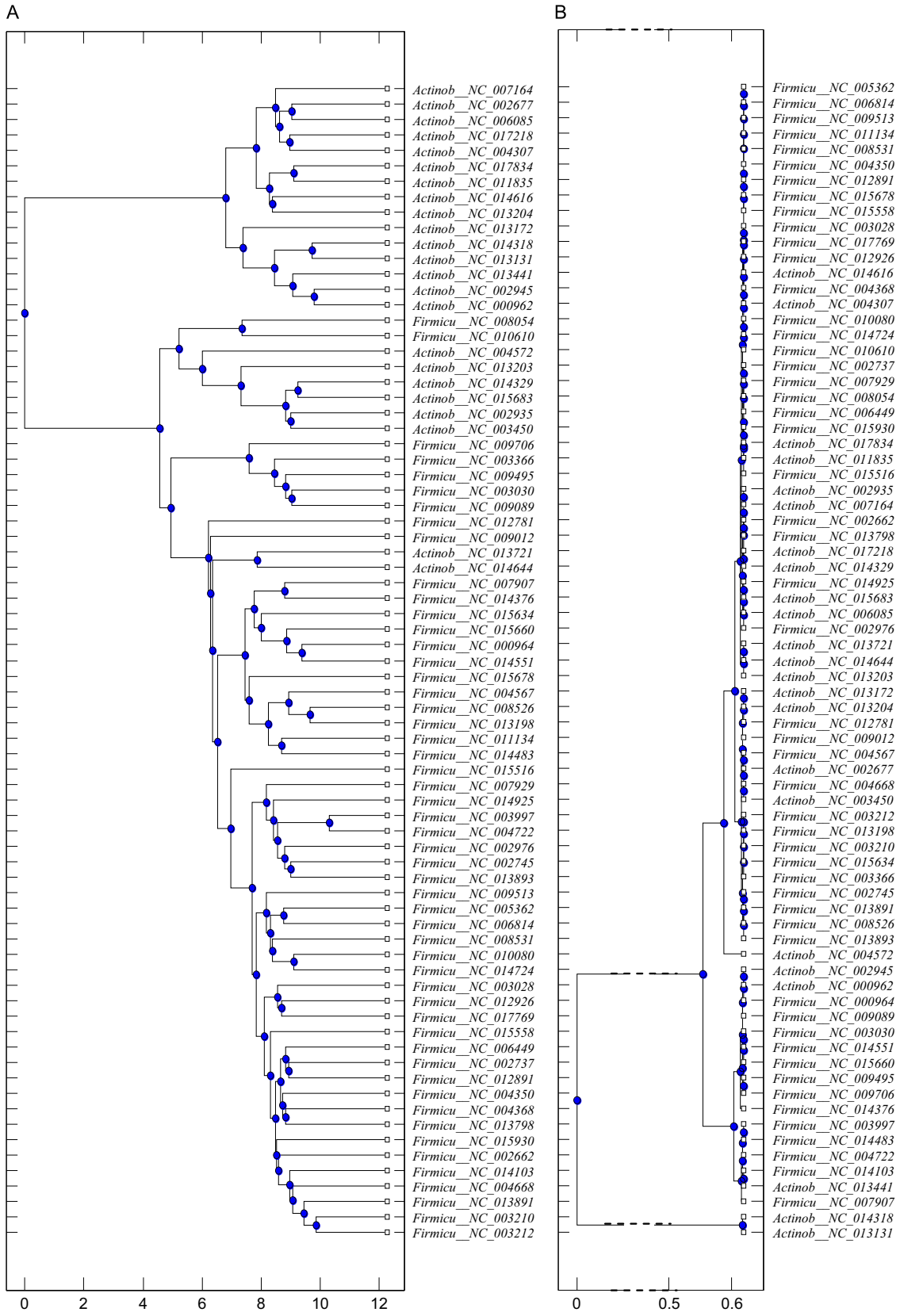

Fig. 8. Distance tree for Firmicutes and Actinobacteria. (A) Distance tree constructed by the SpectCMP method (parameters used: second type of spectra reduction method, RGB color model, building tree method UPGMA, number of regions $n=88$, number of DCT coefficients $i=1335$), (B) distance tree constructed by the MV method; two Actinobacteria are placed in separated branches, which leads to a worse MCC.

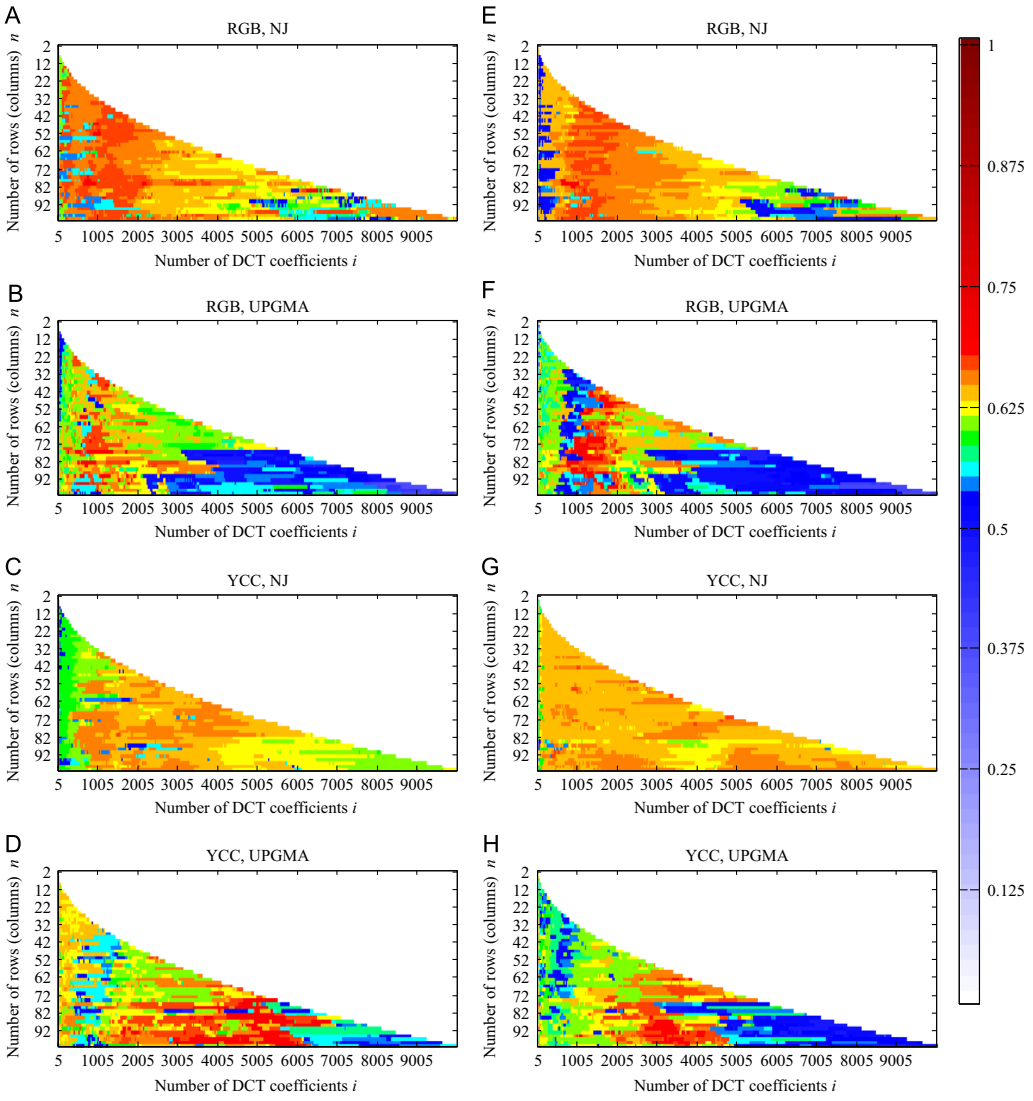

Fig. 9. Heat maps of the MCC. Each element in the heat map represents the average MCC computed from pairs of studied bacterial phyla. The *y*-axis represents the number of rows and columns *n* in the color representative image, and *n* was set from 2 to 100 with a step of 2. The *x*-axis represents the number of elements in CLD *i*, and *i* was set from 5 to n^*n with a step of 5. The title of each heat map describes the color model of the representative color image and the building tree method. Four heat maps on the left (A–D) show the average MCC when the first method for spectra reduction (2) was used. The four heat maps on the right (E–H) show the average MCC when the second method for spectra reduction (3) was used. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

with low MCC in the right parts of the heat maps (roughly 0.5—blue color in the heat map).

Focusing on the building tree method, when NJ is used as the building tree method (Fig. 9 A, C, E, G), a similar average MCC appears for almost each *n* and *i*. The color in the heat maps is more balanced than in the classification, where UPGMA is used (Fig. 9 B, D, F, H). The most balanced classification results occur in the case of the second type of spectrum reducing method, with YC_bC_r as the color model and NJ as the building-tree method (Fig. 9G).

Overall, the highest value of the average MCC 0.725 occurs in the second type of spectrum reducing method, the RGB color model and UPGMA (Fig. 9F) for $n=88$ and $i=1335$. Those parameters were chosen as default parameters in our method, and they were used in comparison with other methods as well. The proposed method with optimal parameter settings is called the SpectCMP method.

The distance matrix produced by the SpectCMP method for 166 bacteria can be visualized by classical multidimensional scaling [36,37]. Each studied organism is represented by a point in multi-dimensional space, where inter-point distances approximate elements in the distance matrix. By estimating the eigenvalues of the classical scaling configuration matrix, a sufficient number of dimensions is 2. Fig. 10 shows the studied organisms in the 2-dimensional plot. Each point in a distance map corresponds to one organism. The phylum *Proteobacteria* is divided into classes because it is the most diverse group of bacteria. In the distance map, the bacteria from the phylum *Tenericutes* and *Actinobacteria* are almost linearly separable. *Tenericutes* are only slightly overlapped with *Proteobacteria*. The distance map shows that *Firmicutes* are separated from *Alphaproteobacteria* and *Betaproteobacteria*.

3.2. Comparison of classification with other methods at the phyla level

Correct bacteria classification at the phylum level by the proposed method, called the SpectCMP method (Spectrogram CoMParison method), was compared for three methods used in bacterial phylogeny: CVTree method, OGTree method and MV method. Each method is based on different principles.

The CVTree method [17], which was developed for amino acid sequences, is capable of testing DNA sequences as well. The length of the counted overlapping strings was set to 12. NJ as the building

tree method was used to calculate a distance tree after the recommendations of the authors.

The OGTree method derives the distance among the genomes from the number of orthologous overlapping gene pairs [16]. The method considers the order of the overlapping genes, also. An online implementation of the OGTree method was used to classify the chosen bacteria [38]. For our analyses, default settings for the web user interface were used. The authors of the OGTree method recommend the UPGMA method for the reconstruction of phylogenies based on OG pairs. Therefore, the UPGMA method was used as the building tree method.

The MV method works with a graphical representation of DNA sequences [20]. After obtaining the DNA graphical curve, the moment vector is counted. The moment vector method contains one parameter—the number of considered components in the moment vector used for the comparative analysis of genomes. In our dataset, the number of components was set to 40 because the topology of the tree converges when the number of components is greater than 40. According to the authors, the distance matrix was computed by the Euclidean distance between the moment vectors, and UPGMA was used as the building tree method.

For each pair of studied phyla, the MCC and the probability of classifying the organisms to the correct phylum were analyzed by the methods described above. The parameters in the SpectCMP method were set according to the results in the previous chapter: the model of the representative color image was RGB_r, the number of rows and columns in the color representative image were $n=88$, the number of considered elements in CLD $i=1335$, the building tree method was UPGMA and the second type of spectrum reducing method was used in Eq. (3). The results are listed in Table 1.

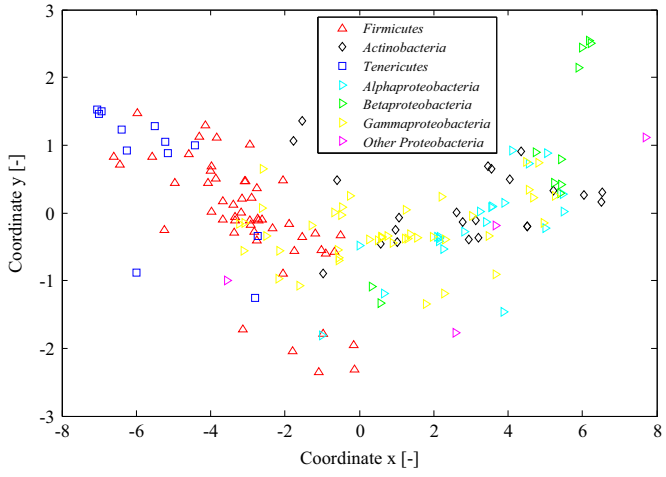

Fig. 10. Visualization of the distance matrix (SpectCMP method) by classical multidimensional scaling. Each point in the scatter diagram represents one studied organism. Inter-point distances approximate elements in the distance matrix. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The MV method and OGTree method give the worst results among all of the methods because some of the organisms were evaluated as outgroups, and thus, they were placed into separated branches (Fig. 8B). This approach leads to a worse result of classification because the first branch contains organisms that were evaluated as outgroups and the second branch the remaining organisms. All of the MCCs achieved by the OGTree and MV method are lower than 0.5. These methods are probably suitable for the classification of closely related organisms.

The CVTree method gives the highest average MCC; this method cannot separate *Proteobacteria* from *Actinobacteria*, which is similar to our method. The results from this method are similar to or much better than from our method. However, the CVTree method gives worse results in the separation of *Tenericutes* from *Actinobacteria* and *Tenericutes* from *Firmicutes* in comparison to our method. The advantage of the SpectCMP method versus the CVTree method is in the input data preparation; SpectCMP method does not require retrieving protein coding regions from the genome. The SpectCMP method works with coding as well as with noncoding regions.

3.3. Classification at the order level by the SpectCMP method

For practical use in medicine and microbiology, classification at lower taxonomical ranks is required. To prove the discrimination power of the SpectCMP method, the classification of bacteria from phylum *Proteobacteria*, class *Gammaproteobacteria* was performed at the order level. A total of 156 bacteria were chosen: 32 organisms from the order *Alteromonadales*, 61 organisms from the order *Enterobacteriales*, 10 organisms from the order *Chromatiales*, 16 organisms from the order *Pasteurellales*, 15 organisms from the order *Pseudomonadales*, 12 organisms from the order *Thiotrichales* and 10 organisms from the order *Xanthomonadales*. DNA sequences were downloaded from NCBI on May 2014.

The results of the classification are listed in Table 2. MCC for most of the pairs of orders is 1 or close to 1. In four cases, MCC is lower than 0.5, which indicates very similar DNA content for those pairs of orders (e. g., a close relationship is shown for *Chromatiales* and *Xanthomonadales* in [39]). However, classification at the order level is more successful than classification at the phylum level, and an average MCC is 0.760.

4. Conclusions

Spectrograms revealed the properties of genomes, which can be used in interspecies studies. In our work, we optimized the method parameters for the determination of relationships of

Table 1
Results of binary classification by the CVTree method, OGTree method, MV method and SpectCMP method. P1—probability that an organism from Phylum 1 is placed in the correct branch, P2—probability that an organism from Phylum 2 is placed in the correct branch, MCC—Matthew's correlation coefficient. Abbreviations of bacterial phyla: *Actinob.=Actinobacteria, Firmic.=Firmicutes, Proteob.=Proteobacteria, and Teneric.=Tenericutes.*

		CVTree method [16]			**OGTree method** [15]			**MV method** [17]			**Our method-SpectCMP method**		
Phylum 1	**Phylum 2**	**P1**	**P2**	**MCC**	**P1**	**P2**	**MCC**	**P1**	**P2**	**MCC**	**P1**	**P2**	**MCC**
Firmic.	*Actinob.*	1	0.870	0.907	0.942	0.591	0.175	1	0.087	0.249	1	0.652	0.752
Teneric.	*Actinob.*	1	0.826	0.787	0.917	0.348	0.287	1	0.087	0.178	0.833	1	0.876
Teneric.	*Firmic.*	1	0.846	0.713	0.417	0.962	0.473	1	0.250	0.243	0.833	1	0.896
Proteob.	*Firmic.*	0.949	1	0.939	1	0.013	0.071	1	0.013	0.071	0.760	0.962	0.707
Proteob.	*Actinob.*	0.456	0.826	0.241	1	0.013	0.054	0.987	0.087	0.184	0.608	0.652	0.218
Proteob.	*Teneric.*	1	0.917	0.951	1	0.013	0.041	1	0.013	0.041	1	0.833	0.902
Average MCC				0.756			0.184			0.161			0.725

Table 2
Results of binary classification at the order level by SpectCMP. P1—probability that an organism from Phylum 1 is placed in the correct branch, P2—probability that an organism from Phylum 2 is placed in the correct branch, MCC—Matthew's correlation coefficient. Abbreviations of bacterial orders: *Alter.* = *Alteromonadales, Chrom.*= *Chromatiales, Enter.*=*Enterobacteriales, Paste.*=*Pasteurellales, Pseud.*=*Pseudomonadales, Thiot.*=*Thiotrichales, Xanth.*=*Xanthomonadales.*

Order 1	Order 2	P1	P2	MCC	Order 1	Order 2	P1	P2	MCC
Chrom.	*Enter.*	1	1	1	*Pseud.*	*Paste.*	1	1	1
Chrom.	*Pseud.*	0.300	1	0.420	*Pseud.*	*Alter.*	1	0.750	0.775
Chrom.	*Paste.*	1	1	1	*Pseud.*	*Thiot.*	1	1	1
Chrom.	*Alter.*	1	1	1	*Pseud.*	*Xanth.*	0.600	0.900	0.524
Chrom.	*Thiot.*	1	1	1	*Paste.*	*Alter.*	1	0.250	0.378
Chrom.	*Xanth.*	0.300	1	0.420	*Paste.*	*Thiot.*	1	0.500	0.577
Enter.	*Pseud.*	0.180	1	0.315	*Paste.*	*Xanth.*	1	1	1
Enter.	*Paste.*	0.820	1	0.833	*Alter.*	*Thiot.*	1	0.500	0.577
Enter.	*Alter.*	0.820	0.750	0.571	*Alter.*	*Xanth.*	1	1	1
Enter.	*Thiot.*	1	0.500	0.577	*Thiot.*	*Xanth.*	1	1	1
Enter.	*Xanth.*	1	1	1					
Average MCC	0.760								

bacteria based on spectrogram comparisons. The quality of classification is affected most by the building tree method. The spectrum reducing method and color model type affected the results slightly. The optimal settings of the SpectCMP method are the following: the RGB model of the representative color image, division of the representative color image into 88 rows and columns, 1335 considered DCT coefficients in CLD, UPGMA as the building tree method, mapping guanine, cytosine and adenine to the red, green and blue layer, respectively, and mapping thymine to all three layers.

The proposed SpectCMP method is an alignment-free method that is suitable for comparing closely related bacteria as well as for comparing distantly related bacteria that do not share a common set of genes. The method works with whole genome sequences; it does not require retrieving protein coding regions from the genome before data analysis. It was shown that the SpectCMP method is suitable for the classification of bacteria at different taxonomic levels as well.

Conflict of Interest

None

Acknowledgments

This work has been supported by the grant project GACR P102/11/1068 and the European Regional Development Fund—Project FNUSA-ICRC (No. CZ.1.05/1.1.00/02.0123).

References

[1] S.T. Priest, S.T. Williams, Computer-assisted identification, Handbook of New Bacterial Systematics, 362-381, Academic Press, London, 1993.
[3] B. Budowle, J.A. Beaudry, N.G. Barnaby, A.M. Giusti, J.D. Bannan, P. Keim, Role of law enforcement response and microbial forensics in investigation of bioterrorism,, Croatian Med. J. 48 (2007) 437–449.
[4] B.K. Choi, C. Wyss, U.B. Gobel, Phylogenetic analysis of pathogen-related oral spirochetes,, J. Clin. Microbiol. 34 (1996) 1922–1925.
[5] J.E. Claridge, Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases, Clin. Microbiol. Rev. 17 (2004) 840–862. http://dx.doi.org/10.1128/CMR.17.4.840-862.2004.
[6] C.A. Petti, Detection and identification of microorganisms by gene amplification and sequencing,, Med. Microbiol. 44 (2007) 1108–1114.
[7] C.C. Thompson, F.L. Thompson, K. Vandemeulebroecke, B. Hoste, P. Dawyndt, J. Swings, Use of recA as an alternative phylogenetic marker in the family Vibrionaceae,, Int. J. Syst. Evol. Microbiol. 54 (2004) 919–924.
[8] B. Snel, P. Bork, M. Huynen, Genome phylogeny based on gene content,, Nat. Genet. 21 (1999) 108–110.
[9] Y. Guo, Z. Wen, X. Rong, Y. Huang, A multilocus phylogeny of the Streptomyces griseus16S rRNA gene clade: use of multilocus sequence analysis for streptomycete systematics,, Int. J. Syst. Evol. Microbiol. 58 (2008) 149–159. http://dx.doi.org/10.1099/ijs.0.65224-0.
[10] G. Devulder, M. Perouse de Montclos, J.P. Flandrois, A multigene approach to phylogenetic analysis using the genus Mycobacteriumas a model,, Int. J. Syst. Evol. Microbiol. 55 (2005) 293–302.
[11] C.U. Köser, M.T. Holden, M.J. Ellington, E.J. Cartwright, N.M. Brown, Rapid whole-genome sequencing for investigation of a Neonatal MRSA Outbreak,, New Engl. J. Med. (2012) 2267–2275. http://dx.doi.org/10.1056/NEJMoa1109910.
[12] S.T. Fitz-Gibbon, C.H. House, Whole genome-based phylogenetic analysis of free-living microorganisms,, Nucl. Acids Res. 27 (1999) 4218–4222.
[13] C.H. House, S.T. Fitz-Gibbon, Using homolog groups to create a whole-genomic tree of free-living organisms: an update,, J. Mol. Evol. 54 (2002) 539–547.
[14] J. Lin, M. Gerstein, Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels,, Genome Res. 10 (2000) 808–818.
[15] B.M. Moret, T. Warnow, Department, advances in phylogeny reconstruction from gene order and content data,, Methods Enzymol. (2005) 673–700.
[16] L.-W. Jiang, K.-L. Lin, C.L. Lu, OGtree: a tool for creating genome trees of prokaryotes based on overlapping genes,, Nucl. Acid Res. 36 (2008) 475–480. http://dx.doi.org/10.1093/nar/gkn240.
[17] J. Qi, B. Wang, B.-I. Hao, Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach,, J. Mol. Evol. 58 (2004) 1–11.
[18] G.E. Sims, S.-H. Kim, Whole-genome phylogeny of *Escherichia coli*/*Shigella* group by feature frequency profiles (FFPs),, Proc. Natl. Acad. Sci. U.S.A. 108 (2011) 8329–8334. http://dx.doi.org/10.1073/pnas.1105168108.
[19] P. Kolekar, M. Kale, U. Kulkarni-Kale, Alignment-free distance measure based on return time distribution for sequence analysis: applications to clustering, molecular phylogeny and subtyping,, Mol. Phylogenet. Evol. 65 (2012) 510–522. http://dx.doi.org/10.1016/j.ympev.2012.07.003.
[20] C. Yu, Q. Liang, C. Yin, R.L. He, S.S.-T. Yau, A novel construction of genome space with biological geometry,, DNA Res. 17 (2010) 155–168. http://dx.doi.org/10.1093/dnares/dsq008.
[21] K. Hatje, M. Kollmar, A phylogenetic analysis of the Brassicales clade based on an alignment-free sequence comparison method,, Front. Plant Sci. (2012) 1–12. http://dx.doi.org/10.3389/fpls.2012.00192.
[22] V. Kubicova, I. Provaznik, Relationship of bacteria using comparison of whole genome sequences in frequency domain,, Adv. Intell. Syst. Comput. 283 (2014) 397–408.
[23] D. Anastassiou, Frequency-domain analysis of biomolecular sequences,, Bioinformatics 16 (2000) 1073–1081. http://dx.doi.org/10.1093/bioinformatics/16.12.1073.
[24] H. Herzel, O. Weiss, E.N. Trifonov, 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding,, Bioinformatics 15 (1999) 187–193.
[25] L. Du, H. Zhou, H. Yan, OMWSA: detection of DNA repeats using moving window spectral analysis,, Bioinformatics 23 (2007) 631–633.
[26] D. Sharma, B. Issac, Spectral repeat finder (SRF): identification of repetitive sequences using Fourier transformation,, Bioinformatics 20 (2004) 1405–1412. http://dx.doi.org/10.1093/bioinformatics/bth103.
[27] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, R. Ramaswamy, Prediction of probable genes by Fourier analysis of genomic sequences,, Bioinformatics 13 (1997) 263–270. http://dx.doi.org/10.1093/bioinformatics/13.3.263.
[28] J.C. Shepherd, Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code,, J. Mol. Evol. 17 (1981) 94–102.
[29] J. Lobry, Asymmetric substitution patterns in the two DNA strands of bacteria,, Mol. Biol. Evol. 13 (1996) 660–665.
[30] M.T. Madigan, J.M. Martinko, Brock Biology of Microorganisms, Pearson-Prentice Hall, San Francisco, 2005.

[31] N. Dimitrova, Y.H. Cheung M. Zhang, Analysis and visualization of DNA spectrograms: open possibilities for the genome research, in Proceedings of the 14th Annual ACM International Conference on Multimedia, New York, 2006.
[32] N. Dimitrova, P. Manor Y.H. Cheung, Methods and Systems for Identification of DNA Patterns Through Spectral Analysis. United States Patent US 2009/0129647 A1, 2009.
[33] B.S. Manjunath, P. Salembier, T. Sikora, Introduction to MPEG-7: Multimedia Content Description Interface, Wiley, 2002.
[34] E. Kasutani A. Yamada, The MPEG-7 Color Layout Descriptor: A Com-pact Image Feature Description for High-speed Image/Video Segment Retrieval, in Image Processing, 2001, 2001. doi:10.1109/ICIP.2001.959135.
[35] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme„ Biochim. Biophys. Acta (BBA)—Protein Struct. 405 (1975) 442–451.
[36] G.A. Seber, Multivariate Observations, John Wiley ans Sons, Inc., New Jersey, 1984.
[37] L.A. Khayal, I. Provaznik, E. Tkacz, Differential analysis of neurodegenerative aging-related mitochondrial genes of long-lived naked mole-rat„ Int. J. Biosci. Biochem. Bioinf. 3 (2013) 75–79. http://dx.doi.org/10.7763/IJBBB.2013.V3.168.
[38] L.-W. Jiang, K.-L. Lin a.C.L. Lu, „OGtree: A Tool for Creating Genome Trees of Prokaryotes Based on Overlapping Genes, Department of Computer Science, National Tsing Hua University, Taiwan, [Online]. Available: ⟨http://genome.cs.nthu.edu.tw/OGtree/⟩.
[39] B. Gao, R. Mohan, R. Gupta, Phylogenomics and protein signatures elucidating the evolutionary relationships among the Gammaproteobacteria„ Int. J. Syst. Evol. Microbiol. 59 (2009) 234–247. http://dx.doi.org/10.1099/ijs.0.002741-0.