Computational and Structural Biotechnology Journal 17 (2019) 118–126

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Rapid Bacterial Species Delineation Based on Parameters Derived From Genome Numerical Representations

Denisa Maderankova *, Robin Jugas, Karel Sedlar, Martin Vitek, Helena Skutkova

Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Technicka 12, 61600 Brno, Czech Republic

ARTICLE INFO

Article history:
Received 30 August 2018
Received in revised form 7 December 2018
Accepted 20 December 2018
Available online 9 January 2019

Keywords:
Bacterial genome
Species delineation
Comparative genomics
Numerical representation
Genomic signal processing

ABSTRACT

Species delineation based on bacterial genomes is an essential part of the research of prokaryotes. In silico genome-to-genome comparison methods are computationally demanding, but much less tedious and error prone than the wet-lab methods. In this paper, we present a novel method for the delineation of bacterial genomes based on genomic signal processing. The proposed method uses numerical representations of whole bacterial genomes, phase signal and cumulated phase signal, from which four parameters are derived for each genome. The parameters characterize a genome and their calculation is independent of the other genomes comprising a delineation dataset. The delineation itself is processed as a calculation of the parameters' average similarity. The method was statistically verified on 1826 bacterial genomes. A similarity threshold of 96% was set based on the receiver operating characteristic curve that featured sensitivity of 99.78% and specificity of 97.25%. Additionally, comparative analysis on another 33 bacterial genomes was conducted using standard delineation tools as these tools were not able to process the dataset of 1826 genomes using desktop computer. The proposed method achieved comparable or better delineation results in comparison with the standard tools. Besides the excellent delineation results, another great advantage of the method is its small computational demands, which enables the delineation of thousands of genomes on a desktop computer. The calculation of the parameters takes tens of minutes for thousands of genomes. Moreover, they can be calculated in advance by creating a database, meaning the delineation itself is then completed in a matter of seconds.

1. Introduction

One of the first tasks in the research of any newly studied organism lies in its correct taxonomic placement. While the taxonomy of higher eukaryotes is less complicated for relatively easily distinguishable species formed by a group of organisms that can interbreed [1], the majority of the Tree of Life is formed by microbial species. This domination lies in their abundance, with the estimation of the total number of microbial cells on Earth thought to be 10^{30}, and in their richness as this amount is formed mainly by 10^6–10^8 separate prokaryotic genospecies [2].

Unfortunately, compiling the taxonomy of asexual microbial organisms is not easy and requires a combination of genotypic, phenotypic, and chemotaxonomic information [3]. Due to the advances in biological molecular techniques, the genotypic traits play the main role in microbial species delineation. While some of the genotypic techniques utilize selected barcoding of parts of genomes, the others compare whole genomes. The first group mainly uses techniques for massive genotyping of pathogenic bacteria in epidemiologic studies and includes the utilization of electrophoresis, PCR, and amplicon sequencing [4–7]. The second group serves as a tool for the correct taxonomic placement of a new organism.

In the past, only a few techniques could offer genome-wide comparisons between organisms and DNA–DNA hybridization was considered the recommended standard for delineating species for a long time [8]. A massive reduction in sequencing costs brought a new, wide range of bioinformatics strategies for species delineation in silico by comparing their genome sequences. These included a wide range of techniques for calculating average nucleotide identity (ANI) [9,10]. The ANI of two genomes was first calculated using all shared orthologous protein coding genes [9] and the method was later improved by cutting one genome into 1020 bp fragments that are searched for in the second genome using the BLAST algorithm [11].

Another approach to calculating ANI searches for maximal unique matches (MUM) based on alignment via suffix trees [12,13]. These computationally derived similarities are closely related to former lab-derived hybridization values [14] and can be easily obtained from genome sequences. The taxonomic placement of every new genome should be verified using these approaches as many of the genomes in the databases are mislabeled [15]. For this purpose, there is a wide

* Corresponding author.
E-mail addresses: maderankova@vut.cz (D. Maderankova), jugas@feec.vutbr.cz (R. Jugas), sedlar@feec.vutbr.cz (K. Sedlar), vitek@feec.vutbr.cz (M. Vitek), skutkova@feec.vutbr.cz (H. Skutkova).

https://doi.org/10.1016/j.csbj.2018.12.006

range of online and standalone tools, e.g. JSpeciesWS [16], Orthologous Average Nucleotide Identity Tool (OAT) [17], Genome-to-Genome Distance Calculator (GGDC) [18], ANI tool by Kostas lab [14], ANItools web [19], dRep [20], Microbial Species Identifier (MiSI) [21], etc.

For the purpose of species delineation using ANI-based methods, a query genome is compared with many other genomes in a database. This task is computationally very demanding as the query genome has a very low similarity to most of the compared genomes that belong to different taxonomic groups. Most of the comparisons are unnecessary and massively increase the computational time.

In this paper, we present an alternative approach for species delineation utilizing four statistical parameters derived from the original genome using genomic signal processing. We used phase signals and cumulated phase signals [22], which are suitable for pairwise and multiple comparisons [23,24]. From these signals, we were able to derive four unique values characterizing individual genomes, which led to a massive reduction of data without affecting the results of the comparison for delineation purposes [25]. A calculation of the similarity of these parameters expresses the similarity between the genomes. Moreover, our method for species delineation significantly reduces the necessary computational time and the delineation accuracy is better or at least comparable with the accuracy of the other tools.

2. Methods

2.1. Statistical Parameters Representing Whole Genome

There are many types of numerical representations of nucleotide sequences [26–28]. Each numerical representation highlights the different characteristics of an original nucleotide sequence and is suitable for a different type of subsequent analysis. We chose to use the phase signal and the cumulated phase signal, which are very easy to calculate and represent the DNA sequence with a vector of numerals instead of symbols. The phase signal is a sequence of values corresponding to the phases of complex numbers assigned to each nucleotide. The assignment is not arbitrary and it is a projection of the nucleotide tetrahedron to the complex plane, where all nucleotide IUPAC symbols can be represented [22]. The numerical map is: A = +1 + j, C = −1 – j, G = −1 + j, T = +1 – j, R = +j, Y = −j, S = −1, W = +1, M = K = N = 0. The respective phases' values in radians are: A = $\pi/4$, C = $-3\pi/4$, G = $3\pi/4$, T = $-\pi/4$, R = $\pi/2$, Y = $-\pi/2$, S = π, W = 0 or 2π, M = K = N = 0.

The cumulated phase is a cumulative sum of all previous phase values, but it can also be calculated directly from the DNA sequence according to Eq. 1 [22]:

$$c_k = \frac{\pi}{4}\left[3\left(n_{G,k} - n_{C,k}\right) + \left(n_{A,k} - n_{T,k}\right)\right], \tag{1}$$

where $n_{A,k}$, $n_{C,k}$, $n_{G,k}$, and $n_{T,k}$ refer to several corresponding nucleotides from the beginning of the sequence to the position k. The phase signal and the cumulated phase signal have the same length L as the original sequence and a reverse transformation is possible. An advantage of the cumulated phase signal is its suitability for the visualization of the whole genome sequence. The visualization can reveal the global trend of the genome, which is not noticeable in the original symbolic sequence or phase signal [28]. For example, the majority of bacterial genomes have a characteristic arrow shape when the cumulated phase signal begins in the region of replication origin (oriC). This helps to predict the position of the oriC region in newly assembled genomes [29].

For a whole genome comparison, it is preferable that sequence records start in the same position and the oriC region is an obvious choice. Unfortunately, genomes starting in another position may still occur in the GenBank database. To eliminate the possible influence of different starts, a simple signal-based rearrangement of sequences was made in a similar manner as was used for the purpose of oriC localization [29].

We used a very simple and computationally undemanding three-step method that only required the identification of the global maximum and minimum. Firstly, a cumulated phase signal of the sequence was calculated. The genome record may begin in an arbitrary region and thus, the first value of the cumulated phase signal may be a false minimum. Secondly, the maximal value of the cumulated phase signal was found, and the signal was rearranged to begin at the position of the maximal value. This rearrangement means that a part of the signal from the beginning to the maximum was simply moved behind the last signal's value and the last signal's value was added to each value of the moved part to make an offset (see Fig. 1). The symbolic sequence was rearranged accordingly. The last step was a localization of the true minimal value, which was the global minimum of the rearranged signal. The symbolic sequence and the cumulated phase signal were again rearranged to begin at the position of the true minimum. The value of the true minimum was then subtracted from all the signals' values so the cumulated phase signal's first value was 0. Subsequently, the phase signal was calculated from the rearranged sequence.

Many diverse parameters representing global features of individual genomes can be calculated from these two signals. We tested common statistical and mathematical parameters, such as the standard deviation, the sum of differences of adjacent signal values, the length of signal, the area under the cumulated phase signal, different types of angles in the cumulated phase signal, distribution of phase signal values, etc. Based on cross-correlation analysis and analysis of the parameter's value distribution for genomes of the same species and different species, the number of final parameters was reduced. Four parameters were chosen as suitable representatives of genome variability. Together, they exhibited excellent discriminative power.

The first parameter was a sum of the differences of the adjacent phase signal values divided by the length of the signal according to:

$$Diff_p = \frac{1}{L}\sum_{k=1}^{L-1}\left|p_k - p_{k+1}\right|, \tag{2}$$

where L is the length of the signal/sequence and p_k is the phase value at position k. The difference depends on an order of nucleotides, e.g. its value for the sequence AAAGGG is 0.26 and for the sequence AGAGAG, which has the same nucleotide content but in a different order, it is 1.31.

The second parameter Tr_0 was the number of transitions between the positive and negative values of the phase signal and vice versa. The count Tr_0 corresponds to the sum of dinucleotides AC, AT, CA, CG, GC, GT, TA, and TG that occurred in the symbolic sequence. The third parameter Tr_{CG} was the number of all possible transitions between the phase signal's values corresponding to the nucleotides C and G. Tr_{CG} is an equivalent of the sum of dinucleotides CC, CG, GC, and GG. Both parameters were normalized by the signal's length L–1, which represented the total number of transitions between the two values in the signal. Although these parameters could be calculated directly from the original symbolic sequence, this would require separate calculations for each dinucleotide. The processing of the phase signal required only a few numerical operations (subtractions and sums) applied on the whole signal to obtain counts for all dinucleotides, e.g. sequence CAGG CAG has $Tr_0 = 3/6$ and $Tr_{CG} = 2/6$.

The fourth parameter was the average growth angle A_{cp} of the cumulated phase signal. The angle was calculated as an average value of the angles for N positions from the beginning of the signal to the maximal value of the signal:

$$A_{cp} = \frac{1}{N}\sum_{k=1}^{N} A_{cp}^{k}, \quad A_{cp}^{k} = \tan^{-1}\left(\frac{c_{ki}}{i}\right), \quad i = k\frac{i_{max}}{N}, \tag{3}$$

where c_{ki} is the cumulated phase value at position ki and i_{max} is the position of the maximal value of the cumulated phase signal. The number of positions was set to $N = 10$ as a trade-off between precision and

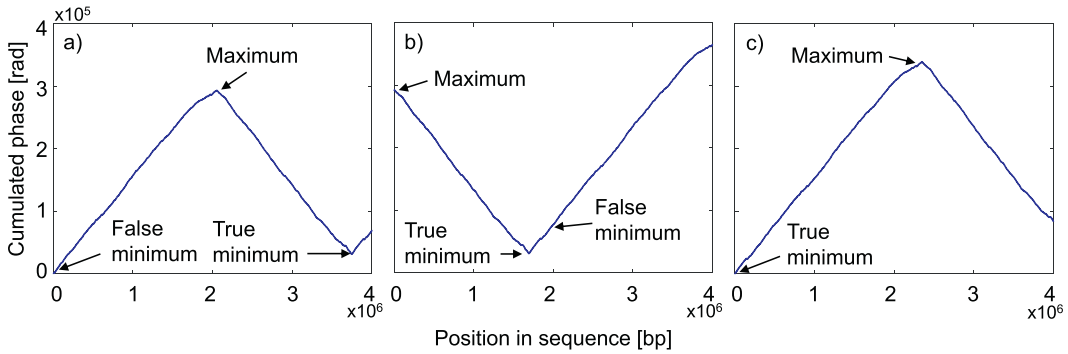

Fig. 1. Three-step rearrangement of the cumulated phase signal of the whole bacterial genome of *Bordetella parapertussis*: a) the genome does not begin in the oriC region and the cumulated phase has a false minimum; b) rearrangement according to the global maximum; c) rearrangement according to the true minimum.

computational demands. The cumulated phase signal can have an uneven rise with several local maxima and the angle can vary depending on the position. The average value smooths the local differences of the signal's growth.

We propose a rapid delineation method based on these parameters. The Whole Genome Parameter (WGP) method is based on comparing the parameters ($Diff_p$, Tr_0, Tr_{CG}, and A_{cp}) of individual genomes. This method does not require any alignment of the genomes or their annotation. Moreover, the parameters can be calculated in advance and saved for further analysis using different compositions of the dataset. Fig. 2 shows the diagram of the WGP method. When processing the delineation analysis of the given dataset, the parameters of all pairs of genomes were compared and the absolute values of the differences between the parameters were calculated. The differences of each parameter were then normalized according to their range in the whole dataset to obtain values from 0 to 1. The normalization was needed to obtain comparable ranges of differences of all parameters. The distance (difference) $\overline{d}^{A,B}$ of each pair of genomes *A* and *B* was an average value of the normalized distances calculated for the parameters. A percentage similarity of the genomes was then $s^{A,B} = 100 * (1 - \overline{d}^{A,B})$.

2.2. Standard Delineation Tools

Four freely available web-based and standalone tools that are standardly used for delineation were chosen for testing: JSpeciesWS [16,30], OAT [17], GGDC [18,31], and ANI/AAI-Matrix from Kostas lab [14]. These tools were chosen as they enable delineation analysis using standard desktop computer with Windows operating system which is the most common equipment.

The JSpeciesWS is an online service that provides pairwise comparison of complete or draft genomes by calculating the ANI values or the tetranucleotide signature frequency correlation coefficients (TETRA) [32]. The ANI values are calculated using the BLAST algorithm (ANIb) or the MUMmer alignment tool (ANIm). For both methods, a similarity threshold of 95–96% is suggested to sufficiently differentiate species. The TETRA is an alignment-free method based on characteristic

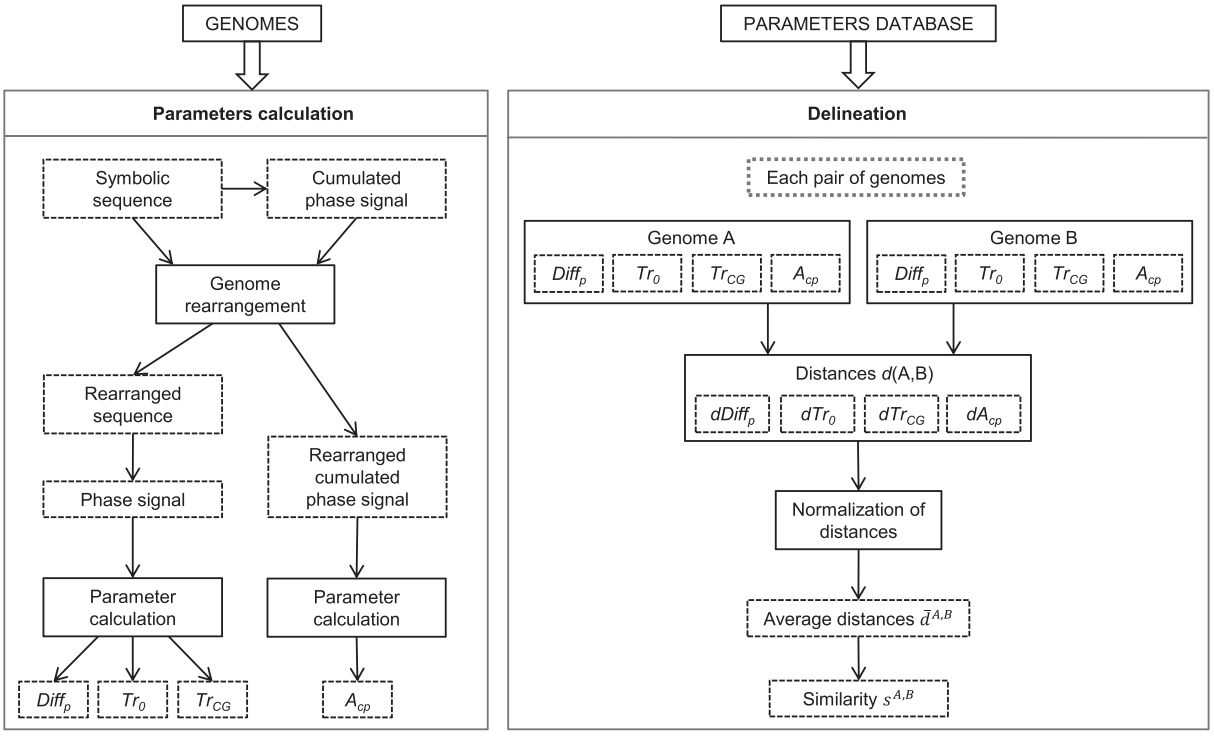

Fig. 2. Diagram of the WGP method. Left: independent parameter calculation for genomes; right: subsequent delineation of a dataset.

frequency occurrences of all 256 possible combinations of tetranucleotides. Closely related genomes have a similar distribution of the frequencies with a high correlation coefficient of >0.99, which corresponds to the ANI value of >96% [30].

The OAT (Orthologous Average Nucleotide Identity Tool) is standalone tool based on OrthoANI method [17] which is a reciprocal version of ANI calculation using BLAST. The reciprocal means that comparison of two genomes is the same for both pair combinations. Therefore, only half of the comparisons is needed to analyze whole dataset. The similarity threshold is 95–96%.

The Genome-to-Genome Distance Calculator (GGDC) is a web tool that uses statistical models of digital DNA-DNA hybridization and is based on the Genome Blast Distance Phylogeny program (GBDP). First, the BLAST algorithm is applied to find the high-scoring segment pairs (HSPs) between two compared genomes. Second, the GBDP uses three different formulas to calculate the distance between the genomes: sum of all HSPs' lengths/sum of genomes' lengths (ANI-f1), identities in HSPs/sum of all HSPs' lengths (ANI-f2), and identities in HSPs/sum of both genomes' lengths (ANI-f3). Then, the distance is converted to an analogous DNA-DNA hybridization (dDDH) value using a generalized linear model. The formula ANI-f2 is recommended and the threshold for the species delineation is set to 70%.

The ANI/AAI-Matrix is a web tool by Kostas lab. Beside the ANI values, the tool also calculates average amino acid identity (AAI), which is better for less related organisms with ANI <75% [33]. The ANI values are calculated using BLAST and the delineation threshold is 95%.

3. Results and Discussion

3.1. Statistical Validation of WGP Method

The WGP method was statistically validated on an extensive dataset of whole bacterial genomes that were downloaded from the GenBank database. The dataset comprises 1826 genomes in total. The composition of the dataset was designed to contain enough sequences of the same species and also different species. Within the dataset, the first 350 genomes belong to seven species: *Acinetobacter baumannii*, *Bacillus cereus*, *Bacillus subtilis*, *Escherichia coli*, *Klebsiella pneumoniae*, *Salmonella enterica*, and *Staphylococcus aureus*, where each species is equally represented by 50 genomes.

The remaining 1476 genomes were downloaded from the NCBI Reference Sequence Database. Only one representative genome of each species was included in the dataset, except for the seven species listed above. Within the 1476 genomes, the genomes with synonym species names were not eliminated. This means the dataset may contain more sequences of the same species, except for the first 350 genomes, however, the number is considered low enough that their influence on the analysis was negligible.

The parameters $Diff_p$, Tr_0, Tr_{CG}, and A_{cp} were calculated for all 1826 genomes. Fig. 3 shows the distribution of the parameters for the seven species represented by 50 genomes each. Any one parameter was not able to sufficiently discriminate all species alone, but each parameter was able to discriminate between different groups of species. For example, the strongest parameter Tr_{CG} discriminated all species quite well, but there were four genomes of *Bacillus subtilis* that overlapped with *E. coli*. Although the parameter A_{cp} seemed to be the worst at species

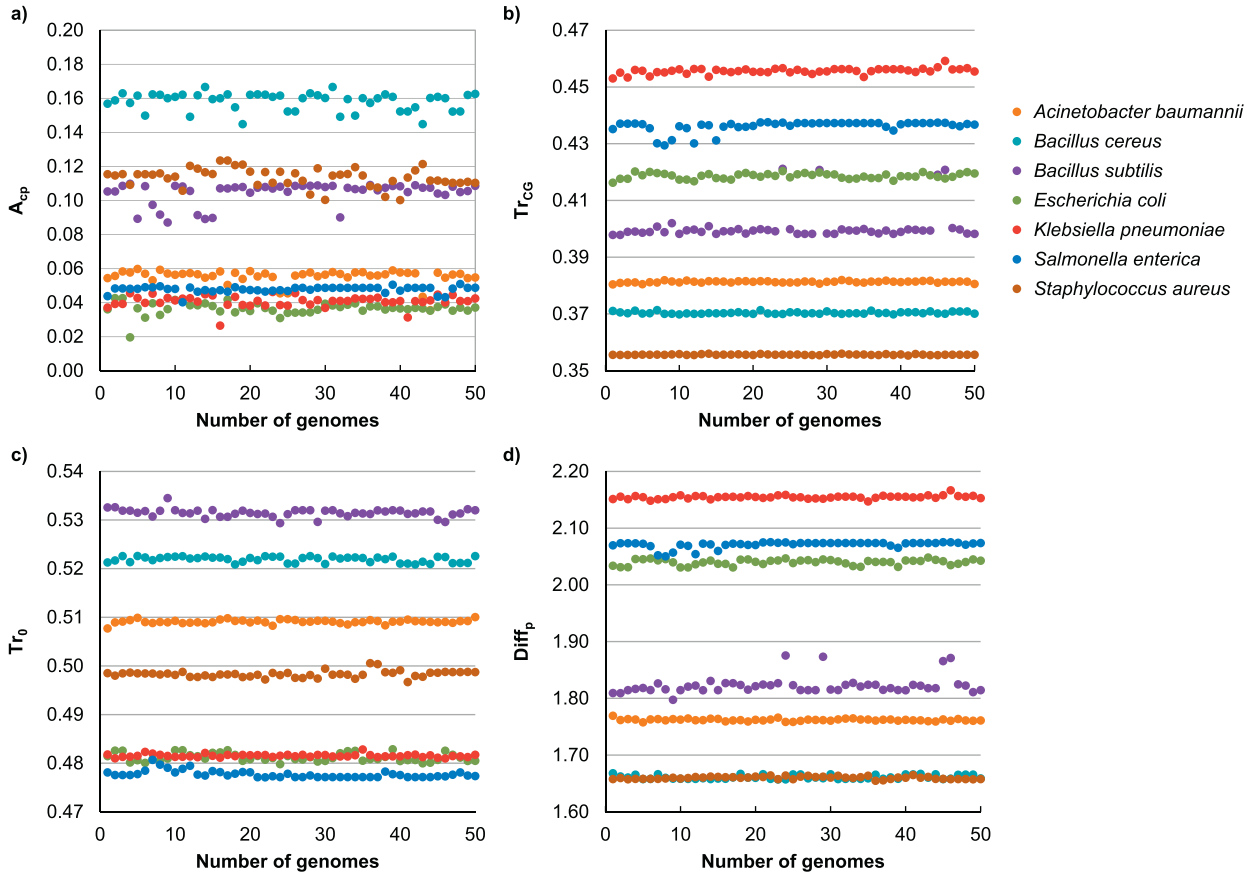

Fig. 3. Distribution of the parameters for 350 genomes: a) the average growth angle A_{cp} of the cumulated phase signal; b) Tr_{CG} – the number of transitions between the phase values of nucleotides C and G; c) Tr_0 – the number of transitions between the positive and the negative values of the phase signal; d) $Diff_p$ – the sum of differences of the adjacent phase values.

discrimination in comparison with the other parameters, a subsequent analysis showed its significance.

The normalized distances of the parameters were calculated for all pairs of genomes, which meant 1.66 million genome-to-genome comparisons. The WGP method is reciprocal which means a comparison of genomes *A* and *B* is the same as *B* and *A*. Therefore, only half of all possible genomes' pairs was calculated. The overall distance $\overline{d}_4$ of each genome pair was calculated as an average value of the normalized distances of the four parameters. To assess the contribution of each parameter to the overall distance, $\overline{d}_2$ was calculated as the average distance of parameters Tr_{CG} and $Diff_p$ and $\overline{d}_3$ was calculated as the average distance of parameters Tr_{CG}, $Diff_p$, and Tr_0. Genome similarities $\overline{s}_2$, $\overline{s}_3$, and $\overline{s}_4$ and similarities for each parameter ($sDiff_p$, sTr_0, sTr_{CG}, and sA_{cp}) were derived from the distances.

Sensitivity and specificity were calculated for a similarity threshold within the range 90% to 98% (see Additional file 1). The similarity threshold was used to divide the genome similarity values into four groups: true positives (TP) were similarity values above the threshold of a genome pair belonging to the same species, true negatives (TN) were similarity values below the threshold of a genome pair of two different species, false positives (FP) were similarity values above the threshold of a genome pair of two different species, and false negatives (FN) were similarity values below the threshold of a genome pair belonging to the same species. Receiver operating characteristic (ROC) curves were constructed where sensitivity was a function of 100 – specificity (see Fig. 4).

FN could occur only in the case of the 350 sequences of the seven species as the dataset did not contain two or more sequences for any other species. The validation showed a low level of FN. Thus, the sensitivity of the WGP method is very high for delineation based even on only one parameter. The sensitivity decreased with an increase in the similarity threshold level. For a threshold of 98%, the lowest sensitivity of 84.35% belonged to the parameter A_{cp}. The parameter with the best sensitivity of 99.94% for the same threshold was Tr_0. The delineation based on the average of all parameters had sensitivity within the range of 97.39% to 100%.

The specificity of each parameter was lower than the sensitivity and in contrast to the sensitivity; it rose with higher values of the similarity threshold. For a threshold of 90%, the specificity ranged from 43.77 to 77.86% and it rose to 85.42–99.69% for a threshold of 98%. The biggest benefit of the combination of parameters was the significant improvement in the specificity. Fig. 4 shows the ROC curves for the delineation based on the similarity of parameter Tr_{CG}, the average of Tr_{CG} and $Diff_p$, the average of Tr_{CG}, $Diff_p$, and Tr_0, and finally the average of all parameters that gave the best results for both sensitivity and specificity. Based on the ROC diagram, the delineation threshold was set to 96%.

When the described genome rearrangement according to the minimal value of the cumulated phase signal was omitted, the delineation results were negatively affected. The sensitivity of the delineation based on $\overline{s}_4$ decreased by approximately 3% and the specificity by 1% for the suggested threshold of 96% and the results were not better for the other thresholds (see Additional file 1).

3.2. Comparison of the WGP Method with the Standardly Used Tools

The extensive dataset used for the statistical validation of the WGP method could not be analyzed by the standardly used tools based on ANI calculation on a standard desktop computer as these methods are extremely computationally demanding and require computing clusters or grid. For comparison purposes, a small dataset containing only 33 bacterial genomes of nine different species was assembled. Each species was represented by at least two strains. Five species were Gram-positive: *Bacillus cereus*, *Bacillus licheniformis*, *Bacillus subtilis*, *Clostridium acetobutylicum*, and *Clostridium beijerinckii*. Four species were Gram-negative: *Escherichia coli*, *Klebsiella pneumoniae*, *Shigella flexneri*, and *Shigella sonnei*. It has been suggested that the genus *Shigella* should be considered as a subgenus of *E. coli* [34,35], however, we decided to count *Shigella* as a different species and test whether the whole genome-based delineation was able to distinguish these genomes from *E. coli*. Additional file 2 provides the names and GenBank accession numbers of all genomes in the dataset.

Fig. 5 shows the cumulated phase signals of the 33 bacterial genomes. The Gram-positive bacteria showed significant differences

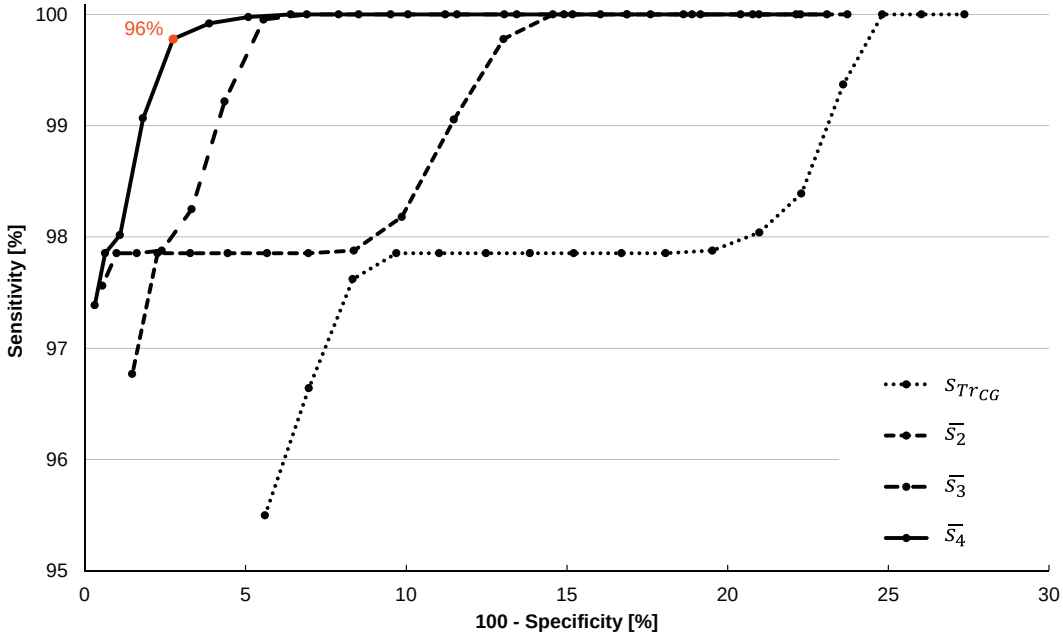

Fig. 4. The ROC curves for the delineation based on sTr_{CG}, $\overline{s}_2$ (the average similarity of parameters Tr_{CG} and $Diff_p$), $\overline{s}_3$ (the average similarity of parameters Tr_{CG}, $Diff_p$, and Tr_0), and $\overline{s}_4$ (the average similarity of all parameters) for the similarity threshold 90–98%, step 0.5%. The red value is the similarity threshold for which the best sensitivity and specificity results were obtained.

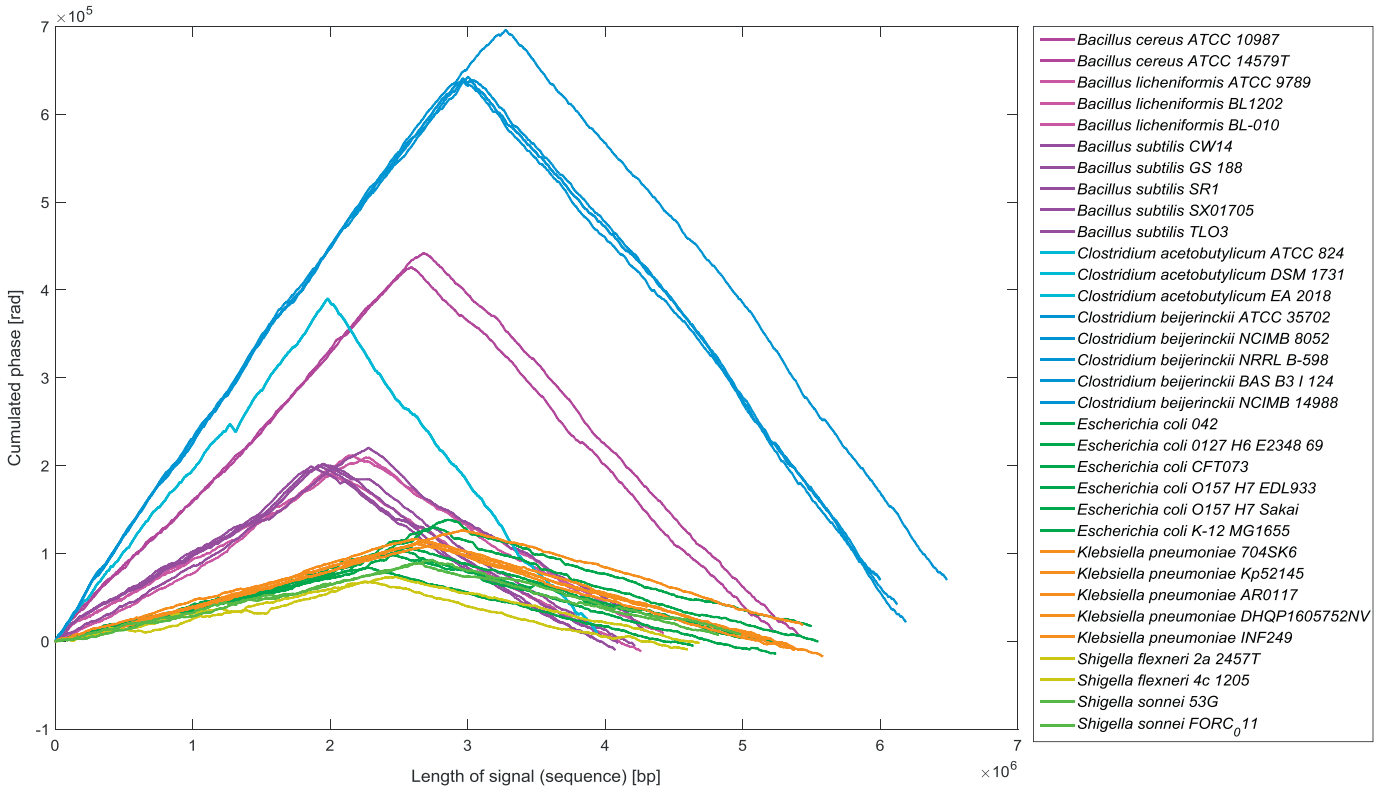

Fig. 5. The cumulated phase signals of 33 bacterial genomes of nine different species. Signals of different strains of one species have the same color.

between their signals, whereas the Gram-negative bacteria were not clearly visually separated according to the species. As can be seen, the strains of *Bacillus cereus* differ in length by 187.5 kbp and the other two *Bacillus* species have much shorter genomes. The signal of *B. licheniformis* ATCC 9789 is visually more similar to *B. subtilis* than to the other two *B. licheniformis* strains. Similarly, *B. subtilis* CW14 is closer to the *B. licheniformis* strains. The signals of the three *Clostridium acetobutylicum* strains are visually almost identical and the genomes' lengths are much shorter than the closely related genomes of *Clostridium beijerinckii*, which are the longest genomes in the dataset. One of them, *C. beijerinckii* NCIMB 14988, is from 298.5 kbp to 486.3 kbp longer than the others.

All ANI-based tools were used with their default or recommended settings. The delineation results of all the tested methods are visualized as heatmaps (see Additional file 3 and Fig. 6).

The online tool JSpeciesWS can only analyze 15 genomes in one run, therefore, the dataset of 33 genomes needed to be divided into several sub-datasets to achieve all the pairwise comparisons. The tool displays the estimated duration of the computation and for a pair of genomes, the estimated duration was 80 s using BLAST (ANIb), 20 s using MUMmer (ANIm), and 6 s using TETRA. The dataset of 33 genomes required 1056 genome-to-genome comparisons, which had an estimated duration of 23.47 h for ANIb, 5.87 h for ANIm, and 1.76 h for TETRA. For comparison, the computational time for the WGP method was 106 s for the parameter calculations of the 33 genomes and 0.02 s for the delineation itself using a standard desktop computer without parallelization.

The delineation based on the ANIb was unsuccessful for two strains of *Bacillus cereus* with an average similarity 91.05%, which was below the 96% threshold (see Fig. 5c). Likewise, the method had a problem with the *Bacillus subtilis* CW14, which had an average similarity of 92.68% with the other *B. subtilis* strains. An average interspecies similarity of the *Bacillus* species was 69.52%. All strains of the two *Shigella* species had a similarity with *E. coli* above the threshold and the average similarity of this group was 97.26%. All other genomes were correctly delineated. The results based on the ANIm algorithm resembled the ANIb-based results (the similarities differed in tenths of a percent, see Fig. 6d).

The TETRA method of the JSpeciesWS tool correctly delineated the problematic *Bacillus cereus* and *Bacillus subtilis* strains and the *Shigella* strains were again delineated together with the *E. coli* strains (see Fig. 6b).

The OAT in its graphical user interface version accepts up to 10 genomes. It enables to use different BLAST versions and use multiple processing cores on your computer. The 33 genomes dataset was divided into several sub-datasets. The average processing time of one genome pair was 122 s using default BLAST version and running on a desktop computer (Intel ® Core™ i5–3330 CPU @ 3.00GHz 3.20 GHz, 32 GB RAM, 64-bit operational system Windows 7 Professional). Similarly to JSpeciesWS, the delineation was unsuccessful for the two *Bacillus cereus* strains with similarity 91.61% and for *Bacillus subtilis* CW14 strain with average similarity 93.02% in comparison with other *B. subtilis* strains. The similarity values produced by OAT and JSpeciesWS varied by a maximum of several tenths of a percent.

The GGDC web tool analyzed the whole dataset in one run (the limit is 50 genomes). The estimated time for one genome pair was about 1 min, which was approximately 17.6 h for all comparisons of the 33 genomes. The results for the three formulas differed, but the variance was not significant for most of the genomes (see Fig. 6e, f, and g). The results differed significantly only for the two strains of *Bacillus cereus*. In this case, the ANI values of the three formulas were 69.4%, 44.9%, and 65.1%. The lowest ANI value was obtained for the recommended formula ANI-f2 and the value was far below the 70% threshold, whereas the other two values were only slightly under the threshold.

The ANI-f1 values were under the threshold for some *E. coli* strains and for most of the comparisons between the *E. coli* and *Shigella* strains. The comparisons between *Shigella flexneri* and *Shigella sonnei* were above the threshold. Beside *Bacillus cereus*, the ANI-f2 values were also significantly under the threshold for the *Bacillus subtilis* CW14 strain. All *E. coli* strains were above the threshold and the comparisons of *E. coli* with the *Shigella* strains were also above the threshold. The ANI-

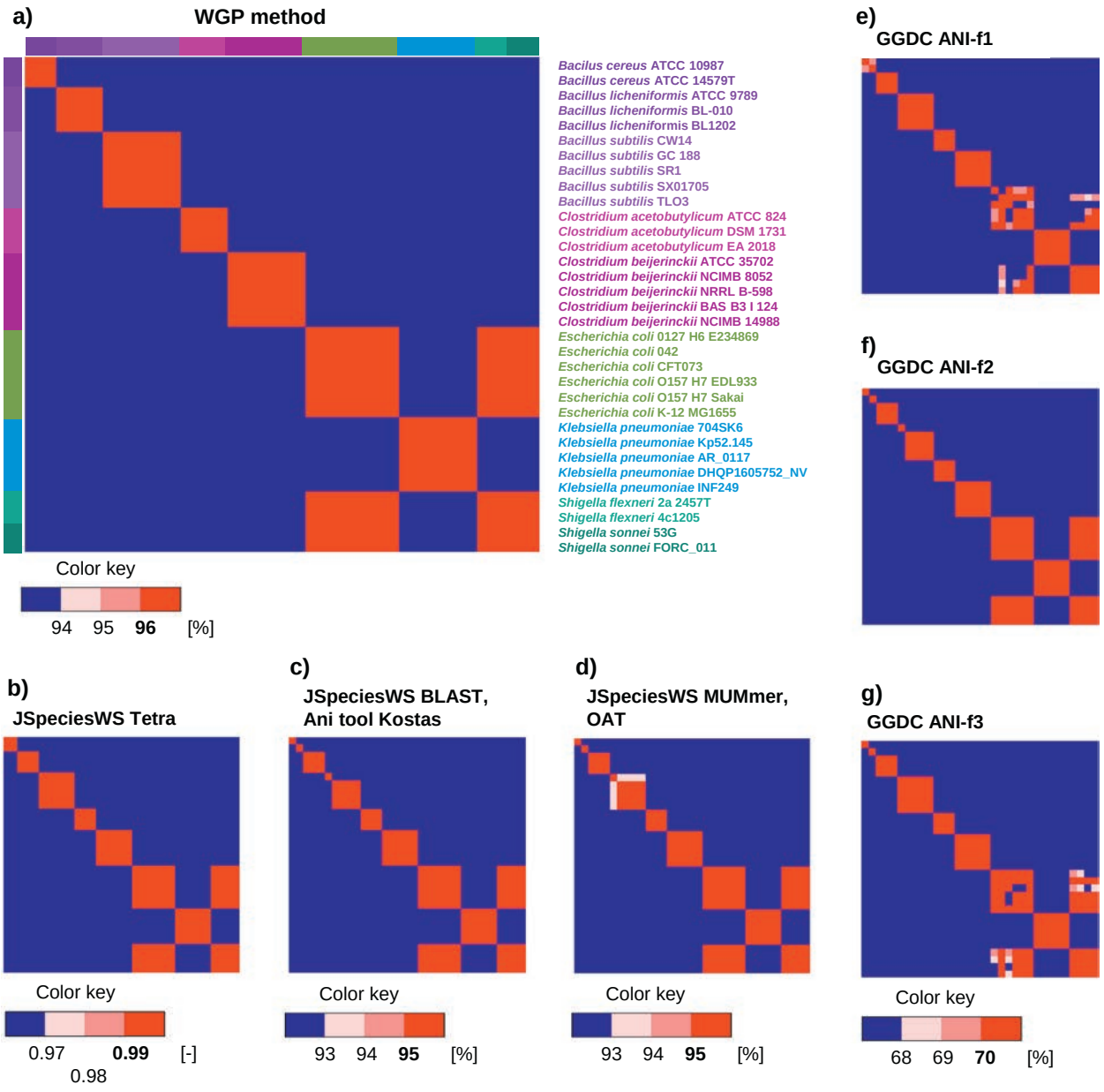

Fig. 6. Visualization of the delineation results: a) our proposed WGP method; b) online tool JSpeciesWS using method TETRA; c) online tool JSpeciesWS using BLAST and Ani tool by Kostas lab; d) online tool JSpeciesWS using MUMmer and OAT; e) tool GGDC using formula ANI-f1; f) GGDC tool using the recommended formula ANI-f2; and g) GGDC tool using formula ANI-f3. The methods that have the same heatmap for the given threshold share one subplot. On the diagonal, there is a self-comparison of each genome where self-similarity is 100%. Similarity values above the threshold (bold value in the color key) are highlighted in red and similarity values 2% below the threshold are blue.

f3 values were slightly under the 70% threshold for some comparisons between the *E. coli* strains and for the comparisons between *E. coli* and the *Shigella* strains.

Similar to JSpeciesWS, the ANI tool by Kostas lab is based on BLAST and is computationally demanding. The tool limits the analysis to 50 genomes. The delineation results (see Fig. 6c) were under the 95% threshold for the *Bacillus cereus* strains and *Bacillus subtilis* CW14. The *Shigella* strains were delineated together with the *E. coli* strains.

The whole set of 33 genomes was analyzed also by proposed WGP method. The phase signals and the cumulated phase signals were calculated. From the signals, the vector of the four WGP method parameters was calculated for each genome. Each parameter vector was compared with the parameter vectors of all other genomes and the average similarities $\overline{s_4}$ were calculated. According to the 96% delineation threshold, all genomes belonging to the same species had a similarity above the threshold (see Fig. 6a). The similarities between genomes of different species were under 94%, apart from the *Shigella* species. *Shigella flexneri* and *Shigella sonnei* were delineated to *E. coli* with an average similarity of 97.84%, whereas the average intraspecies similarity of the *E. coli* strains was 98.51%.

Although another Gram-negative bacteria, *Klebsiella pneumoniae* (average intraspecies similarity 98.80%), had a similar cumulated phase signal to the *E. coli* and *Shigella* species, the average interspecies similarity between *K. pneumonia* and the other Gram-negative bacteria was 85.67%. The highest interspecies similarity of 93.27% occurred between the strains of *Clostridium acetobutylicum* and *Clostridium beijerinckii*. Both species had intraspecies similarities above 99%, even the strain *C. beijerinckii* NCIMB 14988, which had the longest genome. The average similarity between *Bacillus* species was 79.78%. The two strains of *Bacillus cereus*, for which delineation was problematic using

Table 1
The elapsed and estimated processing times of the WGP method and the standardly used delineation tools.

Method/tool	Time [sec]		
	2 genomes	33 genomes	1826 genomes
WGP	4.5/0.006	71.6/0.013	3048/19.9
JSpeciesWS BLAST, ANI tool by Kostas	*80*	*84,480*	*266.6*10^6*
JSpeciesWS MUMmer	*20*	*21,120*	*66.6*10^6*
JSpeciesWS TETRA	*6*	*3168*	*9.9*10^6*
OAT	*122*	*64,416*	*203.3*10^6*
GGCD ANI-f1, ANI-f2, ANI-f3	*60*	*63,360*	*199.9*10^6*

Legend: The elapsed and estimated processing times are for one pair of genomes, the dataset of 33 genomes, and the dataset of 1826 genomes without parallelization of the task. For the WGP method, the first value is the parameter calculation time and the second value is the time taken for the comparison of the parameters. Estimated times are in italics.

the ANI-based methods, had a similarity of 98.90%. The strain *B. subtilis* CW14 had the lowest similarity with other strains of the same species; however, the average value was 96.76%, which was sufficiently above the threshold.

3.3. Computational Demands

Table 1 shows the computational time needed for the delineation of one pair of genomes, for the dataset of 33 genomes that was used for the comparative analysis, and for the dataset of 1826 genomes that was used for the WGP method's verification. The WGP method was implemented in Matlab 2015a without parallelization of the task and using a common desktop computer (Intel ® Core™ i5-3330 CPU @ 3.00GHz 3.20 GHz, 32 GB RAM, 64-bit operational system Windows 7 Professional). In Table 1, the WGP method has two values separated by a slash. The first value corresponds to the elapsed time for the parameter calculations and the second value corresponds to the elapsed time for the delineation itself, which means the comparisons of the parameters. For the other tools, the values are based on estimations provided by the tools for one pair of genomes and the dataset of the 33 genomes. The tested tools, using their stand-alone version on the same desktop computer or web version, were not able to process the dataset of 1826 genomes and the time estimations show that such analysis is unrealizable without using large computing clusters or grid. For example, the estimated time for the BLAST-based methods is >3000 days without parallelization of the task. The WGP method calculated the four parameters for the 1826 genomes in 50.8 min and the delineation took 19.9 s. With parallelization using four cores, the times for the WGP method were four times smaller.

The reciprocal methods (WGP, TETRA, and OAT using OrthoANI method) have advantage as they analyze each genome pair once; the similarity does not depend which genome serves as a query. Therefore, the methods compare $(N \times N - N)/2$ genome pairs where N is the number of genomes. Non-reciprocal methods compare $N \times N - N$ genome pairs. The number of comparisons was reflected in the processing time estimations.

4. Conclusions

The proposed WGP method uses four parameters to globally represent a genome. The parameters are calculated from the phase signal and the cumulated phase signal, which are numerical representations of the genome's DNA sequence. The parameters are the sum of the differences of adjacent phase signal values, the number of transitions between the positive and the negative values of the phase signal, the number of transitions between the phase values corresponding to the nucleotides C and G, and the average growth angle of the cumulated phase signal. The parameters' calculation is fast, straightforward, deterministic, and independent of the composition of a dataset. These parameters can be calculated in advance and stored in a database to be used anytime in different delineation analyses. The WGP method enables delineation of an extensive dataset on a standard desktop computer.

The method was verified on a dataset of 1826 RefSeq bacterial genomes. Any WGP method parameter alone was not sufficient to delineate the genomes with sufficiently high sensitivity and specificity. The best results were obtained for an average distance of all four parameters. Based on the verification and the ROC curve, the similarity threshold was set to 96%. For this threshold, the WGP method had a sensitivity of 99.78% and a specificity of 97.25%.

As the standardly used delineation tools were not able to process the dataset of 1826 genomes using desktop computer due to the computational demands and their restrictions on dataset size, the comparison of methods/tools was conducted using a much smaller dataset comprising only 33 bacterial genomes of nine species. The ANI-based methods had a problem with delineating some strains of *Bacilus cereus* and *Bacilus subtilis*. All the compared tools and the WGP method assigned the strains of *E. coli*, *Shigella flexneri*, and *Shigella sonnei* to one group. The analysis showed that these species are difficult to distinguish on a whole-genome level.

Besides the *E. coli* and *Shigella* group, the WGP method produced a perfect delineation faster than the ANI-based methods. Moreover, the ANI-based methods except OrthoANI tool analyze one pair of genomes twice because one genome serves as a "query" and the *se*cond as a "reference" and the results of the comparisons slightly differ. This double analysis increases the computational time. This is not an issue for the WGP method as the similarity between genomes *A* and *B* is the same as for *B* and *A*, therefore, only half the comparisons are needed. Furthermore, the BLAST and MUMmer algorithms are very complex and computationally demanding, making an analysis of hundreds or even thousands of genomes impossible on a desktop computer and require large computing clusters or grid.

One of the biggest problems in contemporary bacterial genome research is the huge amounts of data that need to be analyzed. The methods used to process these data need to be computationally effective, which is not the case for the available delineation tools. The proposed signal-based method can tackle this problem reliably in what has previously been an unattainable time, even without task parallelization. Additional parameters can be added to the existing ones if needed (by extending the vector of four parameters). For example, to obtain sufficient resolution within species or groups of closely related species.

The WGP method is based on four parameters that globally represent genomes and is a powerful tool for the processing of huge amounts of data. We consider this method to be a newly proven concept that has significant advantages for genomic signal processing. As the development of nanopore sequencing technology is expected in the near future, genomic signal processing methods may be of great importance. The nanopore technology produces a current signal that has to be converted into a symbolic sequence. This conversion can be skipped and the genome can be analyzed in its signal representation using the genomic signal processing methods. The WGP method derives the parameters from the phase representations of a genome and equivalent parameters can be derived directly from the nanopore produced signal.

We have shown that a delineation based on several parameters representing a whole genome is not only possible, but highly effective and precise. The aim of third-generation sequencers is to enable everybody everywhere to perform DNA sequencing that requires only low computational demands. Our delineation method reduces a whole genome from millions of symbols to only four significant values. This enables the comparison of extensive numbers of microorganisms even without online access to large databases.

The WGP method can be downloaded as Matlab source codes: https://www.ubmi.feec.vutbr.cz/en/publications/wgp-genome-delineation/. The present version of the software is not suitable for comparison of genomes assembled in multiple scaffolds.

Funding

This work was supported in part by the grant project of the Czech Science Foundation [GACR 17-01821S].

Conflict of Interest

The authors declare that they have no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2018.12.006.

References

[1] Zachos FE. Taxonomy: species splitting puts conservation at risk. Nature 2013;494(7435):35. https://doi.org/10.1038/494035c.
[2] Simon C, Daniel R. Metagenomic analyses: past and future trends. Appl Environ Microbiol 2011;77(4):1153–61. https://doi.org/10.1128/AEM.02345-10.
[3] Prakash O, Verma M, Sharma P, Kumar M, Kumari K, et al. Polyphasic approach of bacterial classification - an overview of recent advances. Indian J Microbiol 2007;47(2):98–108. https://doi.org/10.1007/s12088-007-0022-x.
[4] Maiden MC, Jansen Van Rensburg MJ, Bray JE, Earle SG, Ford SA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 2013;11(10):728–36. https://doi.org/10.1038/nrmicro3093.
[5] Powell W, Morgante M, Andre C, Hanafey M, Vogel J, et al. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol Breed 1996;2(3):225–38. https://doi.org/10.1007/BF00564200.
[6] Hardys H, Balick M, Schierwater B. Applications of random amplified polymorphic DNA (RAPD) in molecular ecology. Mol Ecol 1992;1(1):55–63.
[7] Brhelova E, Kocmanova I, Racil Z, Hanslianova M, Antonova M, et al. Validation of minim typing for fast and accurate discrimination of extended-spectrum, beta-lactamase-producing Klebsiella pneumoniae isolates in tertiary care hospital. Diagn Microbiol Infect Dis 2016;86(1):44–9. https://doi.org/10.1016/j.diagmicrobio.2016.03.010.
[8] Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, et al. Report of the Ad Hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Evol Microbiol 1987;37:463–4. https://doi.org/10.1099/00207713-37-4-463.
[9] Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A 2005;102(7):2567–72. https://doi.org/10.1073/pnas.0409727102.
[10] Auch AF, von Jan M, Klenk HP, Göker M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci 2010;2(1):117–34. https://doi.org/10.4056/sigs.531120.
[11] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389–402. https://doi.org/10.1093/nar/25.17.3389.
[12] Deloger M, El Karoui M, Petit MA. A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. J Bacteriol 2009;191(1):91–9. https://doi.org/10.1128/JB.01202-08.
[13] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. Versatile and open software for comparing large genomes. Genome Biol 2004;5:R12. https://doi.org/10.1186/gb-2004-5-2-r12.
[14] Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 2007;57(1):81–91. https://doi.org/10.1099/ijs.0.64483-0.
[15] Figueras MJ, Beaz-Hidalgo R, Hossain MJ, Liles MR. Taxonomic affiliation of new genomes should be verified using average nucleotide identity and multilocus phylogenetic analysis. Genome Announc 2014;2(6). https://doi.org/10.1128/genomeA.00927-14 (e00927-14).
[16] Richter M, Rosselló-Móra R, Glöckner FO, Peplies J. JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. Bioinformatics 2015;32(6):929–31. https://doi.org/10.1093/bioinformatics/btv681.
[17] Lee I, Kim YO, Park SC, Chun J. OrthoANI: an improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Microbiol 2016;66:1100–3. https://doi.org/10.1099/ijsem.0.000760.
[18] Auch AF, Klenk HP, Göker M. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. Stand Genomic Sci 2010;2(1):142–8. https://doi.org/10.4056/sigs.541628.
[19] Han N, Qiang Y, Zhang W. ANItools web: a web tool for fast genome comparison within multiple bacterial strains. Database (Oxford) 2016:baw084. https://doi.org/10.1093/database/baw084.
[20] Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J 2017;11:2864–8. https://doi.org/10.1038/ismej.2017.126.
[21] Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, et al. Microbial species delineation using whole genome sequences. Nucleic Acids Res 2015;43(14):6761–71. https://doi.org/10.1093/nar/gkv657.
[22] Cristea PD. Conversion of nucleotides sequences into genomic signals. J Cell Mol Med 2002;6(2):279–303.
[23] Skutkova H, Vitek M, Babula P, Kizek R, Provaznik I. Classification of genomic signals using dynamic time warping. BMC Bioinforma 2013;14(Suppl10):S1. https://doi.org/10.1186/1471-2105-14-S10-S1.
[24] Skutkova H, Vitek M, Sedlar K, Provaznik I. Progressive alignment of genomic signals by multiple dynamic time warping. J Theor Biol 2015;385:20–30. https://doi.org/10.1016/j.jtbi.2015.08.007.
[25] Sedlar K, Skutkova H, Vitek M, Provaznik I. Set of rules for genomic signal downsampling. Comput Biol Med 2016;69:308–14. https://doi.org/10.1016/j.compbiomed.2015.05.022.
[26] Mendizabal-Ruiz G, Román-Godínez I, Torres-Ramos S, Salido-Ruiz RA, Morales JA. On DNA numerical representations for genomic similarity computation. PLoS One 2017;12(3):e0173288. https://doi.org/10.1371/journal.pone.0173288.
[27] Bielińska-Wąż D. Graphical and numerical representations of DNA sequences: statistical aspects of similarity. J Math Chem 2011;49:2345. https://doi.org/10.1007/s10910-011-9890-8.
[28] Cristea PD. Representation and analysis of DNA sequences. In: Dougherty ER, Shmulevich I, Chen J, Wang ZJ, editors. Genome signal processing and statistics. New York: Hindawi Publising Corporation; 2005. p. 15–65.
[29] Maderankova D, Sedlar K, Vitek M, Skutkova H. The identification of replication origin in bacterial genomes by cumulated phase signal. 2017 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB); 2017. p. 1–5. https://doi.org/10.1109/CIBCB.2017.8058561.
[30] Richter M, Roselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 2009;106(45):19126–31. https://doi.org/10.1073/pnas.0906412106.
[31] Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinform 2013;14:60. https://doi.org/10.1186/1471-2105-14-60.
[32] Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO. Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol 2004;6(9):938–47. https://doi.org/10.1111/j.1462-2920.2004.00624.x.
[33] Rodriguez-R LM, Konstantinidis KT. Bypassing cultivation to identify bacterial species. Microbe 2014;9(3):111–8. https://doi.org/10.1128/microbe.9.111.1.
[34] Lan R, Reeves PR. Escherichia coli in disguise: molecular origins of Shigella. Microbes Infect 2002;4(11):1125–32. https://doi.org/10.1016/S1286-4579(02)01637-4.
[35] Pettengill EA, Pettengill JB, Binet R. Phylogenetic analyses of Shigella and enteroinvasive Escherichia coli for the identification of molecular epidemiological markers: whole-genome comparative analysis does not support distinct genera designation. Front Microbiol 2016;6(1573). https://doi.org/10.3389/fmicb.2015.01573.