Journal of Theoretical Biology 385 (2015) 20–30

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Progressive alignment of genomic signals by multiple dynamic time warping

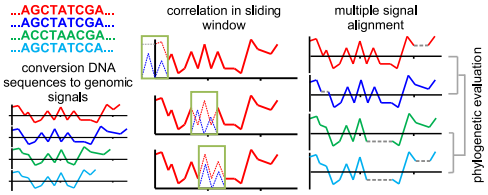
Helena Skutkova [a,*], Martin Vitek [a,b], Karel Sedlar [a], Ivo Provaznik [a,b]

[a] Department of Biomedical Engineering, Brno University of Technology, Technicka 12, 616 00 Brno, Czech Republic
[b] International Clinical Research Center – Center of Biomedical Engineering, St. Anne's University Hospital Brno, Pekarska 53, 656 91 Brno, Czech Republic

HIGHLIGHTS

- We propose a new method for adapting the lengths of multiple genomic signals.
- The multiple signal alignment combines clustering and dynamic time warping.
- We proposed a correlation based modification of dynamic time warping.
- The correlation in sliding window evaluates the local homology.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 17 November 2014
Received in revised form
21 July 2015
Accepted 3 August 2015
Available online 20 August 2015

Keywords:
Genomic signal processing
Multiple alignment
Correlation
Phylogenetic tree
Similarity distance

ABSTRACT

This paper presents the utilization of progressive alignment principle for positional adjustment of a set of genomic signals with different lengths. The new method of multiple alignment of signals based on dynamic time warping is tested for the purpose of evaluating the similarity of different length genes in phylogenetic studies. Two sets of phylogenetic markers were used to demonstrate the effectiveness of the evaluation of intraspecies and interspecies genetic variability. The part of the proposed method is modification of pairwise alignment of two signals by dynamic time warping with using correlation in a sliding window. The correlation based dynamic time warping allows more accurate alignment dependent on local homologies in sequences without the need of scoring matrix or evolutionary models, because mutual similarities of residues are included in the numerical code of signals.

1. Introduction

The growing volume of genomic data in public databases highlights the importance of discovering of more effective approaches to their processing. The utilization of signal processing tools for genomic data analyses forms a new sub-discipline of bioinformatics called genomic signal processing (Anastassiou, 2001). Through numerous special events and scientific meetings (e.g. Wang et al., 2004; Kung et al., 2010; Dougherty et al., 2005; Braga-Neto et al., 2010) it was determined that the genomic signal processing is not only an alternative approach but also plays a full-fledged role in analysis of genetic information. The main advantages of genomic signal processing consist of ability to detect and describe characteristic properties of genetic information which are invisible to the naked eye (Tao et al., 2013; Florquin et al., 2005; Song and Yan, 2012). The character based methods (4 characters for nucleotides A, C, G, T) typically use the analysis of only those point mutations which are visible. The computational demand of character based methods is another limitation in their application especially on a large amount of data. Here, we focus only on phylogenetic problems in bioinformatics. The growing number of whole genome records in databases is ideal for comprehensive

* Corresponding author. Tel.: +420 541 146 659.
E-mail addresses: skutkova@feec.vutbr.cz (H. Skutkova), vitek@feec.vutbr.cz (M. Vitek), sedlar@feec.vutbr.cz (K. Sedlar), provaznik@feec.vutbr.cz (I. Provaznik).

http://dx.doi.org/10.1016/j.jtbi.2015.08.007

phylogenetic study leading to building the tree of life (Delsuc et al., 2005). The complexity of the character-based methods such as maximum likelihood method makes their utilization impossible for this purpose (Chor and Tuller, 2006). On the other hand, the distance based methods using pairwise distances for classification are computationally feasible, but the previous indispensable step consisting of multiple sequence alignment (MSA) again complicates the calculation (Montanola et al., 2013; Wang and Jiang, 1994).

The phylogenetic analysis by genomic signal processing tools has two possible approaches. The first approach represented by "alignment-free" methods usually evaluates the pairwise distances based on difference of characteristic attributes (Vinga and Almeida, 2003). The characteristic attributes can be derived directly from the character sequences e.g. frequencies of words of a specific length as dinucleotides or triplets, commonly known as k-word or k-mer methods (Otu and Sayood, 2003; Deng et al., 2011; Kolekar et al., 2012; Yu, 2013; Wen et al., 2014). Alternatively, the alignment-free characteristic features can be extracted from signal or numerical representation of DNA by signal processing tools e.g. Fourier power spectrum or wavelet transform coefficients (Machado et al., 2011; Yin et al., 2014; Hoang et al., 2015). Into the last category of alignment-free methods can be included Chaos Game Representation (CGR) techniques using chaos theory for representation of genetic code (Deschavanne et al., 1999; Almeida et al., 2001). Generally, the alignment-free methods allow reduction of long character sequences of DNA to short representative numerical vectors (or matrices) for similarity evaluation without need of alignment. This transformation is almost always degenerative; the original genetic information of DNA sequences is lost. The result similarity measure is only global; it does not evaluate the local homology and in some cases (e.g. CGR methods) a global measure of similarity is difficult to interpret.

That is accomplished via the second approach represented by genomic signal classification methods. We start from the fact that the representation of DNA sequence by genomic signal is taxonomically specific (Cristea, 2003; Yao et al., 2008). The genomic signal represents sequence profile of changes of characteristic property depending on sequence position. The local homology in two or more sequences appears as a similar trend in their signals. The comparison of genomic signals based on local similarities requires the positional adjustment of similar segments. The similarity of signals with the same length (without the indels type of mutations) can be evaluated even without the alignment (Cristea and Tuduce, 2011; Cristea and IEEE, 2012). The majority of gene and especially genome sequences contains a large number of indels. A suitable tool for signal length alignment (sequence alignment) is required. The classification of genomic signals using the dynamic time warping (DTW) pairwise alignment was introduced in Skutkova et al. (2013). The mentioned approach for DNA classification is based on multiple alignment of more than two sequences. This paper presents a multiple DTW (mDTW) substitute for MSA in signal form. The mDTW was designed for phylogenetic classification of genomic signals with different lengths. The applicability of mDTW for phylogenetic study and its correspondence with standard phylogenetic methods will be subject to testing. The mDTW itself does not provide better results or computational speed up, but redundancy of genetic information observed in genomic signal offers the possibility of signal decimation and thereby operational complexity reducing (Skutkova et al., 2013; Sedlar et al., 2014).

2. Methodology

2.1. Signal specification

Although the main purpose of our new alignment technique is obvious from the title, the multiple DTW is also suitable for many other different applications besides genomic signal processing. Generally, the input data is set of 1D signals with different lengths requiring alignment. The same length allows the evaluation of mutual similarities e.g. phylogenetic analysis of genomic signals. The choice of cumulated phase (Cristea, 2002) as genomic signal representation used in further testing was implemented for its taxonomy specific features (Cristea, 2003; Skutkova et al., 2013). The main requirement for input signals is equidistant sampling of x-axis. The wide range of available numerical representations of DNA sequences is thus limited to graphical representation enabling the projection from multidimensional space (2D, 3D) to 1D, where the x-axis represents the position in DNA sequences and y-axis is some characteristic parameter changing in dependence on the sequence position. The alternative method for cumulated phase could be 1D projection of DNA "walk" curve (Berger et al., 2004), or nucleotide density curves (Maderankova and Provaznik, 2011). Outside the scope of genomic signals, most of waveforms (a curve showing the shape of a wave at a given time) fulfill requirements on the input data for the multiple DTW e.g. ECG, EEG, speech signal, etc. The time axis corresponds with the positional information in genomic signal. The signal parameters as length, size or amplitude influence the choice of parameters of multiple DTW as will be described below.

2.2. Multiple signal alignment

Our proposed algorithm for multiple genomic signal alignment performs a similar role as multiple sequence alignment algorithms for biological sequence of symbols representing nucleotides or amino acids. For this reason, we were inspired by them. Like many present commercial software, the multiple DTW alignment of signals is based on a progressive alignment algorithm using hierarchical clustering (Feng and Doolittle, 1987). The algorithm function is shown by flowchart in Fig. 1. Prior a systematic description of the all blocks that compose multiple alignment process, the principle of pairwise alignment shall be clarified. The common dynamic time warping algorithm allowing length adaptation of two signals (Skutkova et al., 2013; Sakoe and Chiba, 1978) was modified for our purposes. The proposed modification involves evaluation of local similarities by correlation instead of distance calculation.

2.2.1. Correlation based dynamic time warping – cDTW

The dynamic time warping is an algorithm of dynamic programming used to partial time shape matching of two signals. The basic principle consists of three main steps: calculation of table of local differences between each pair of signals samples; conversion of the local differences table to table of accumulated distances; searching for the best path through the table to obtain the optimal signal matching (Sakoe and Chiba, 1978). In our case, it is necessary to make two changes of this convention. The first one is formal, because the time series does not occur at genomic signals. The conversion consists of positional information coding in x-axis, instead of time variable.

The second modification justifies labeling "cDTW" in the subchapter title and in the fifth block in Fig. 1. The calculation of local differences between samples from two signals is replaced by determining the correlation coefficients in a sliding window. This idea combines dynamic programing with k-tuple sequence alignment approach on which are based some alternative heuristic alignment methods as FASTA (Pearson, 1998) or BLAST (Altschul et al., 1997). These heuristic sequence alignment approaches are used for a rapid scanning of sequences to localize sequence patterns (words) with length k. These k-tuples are scored by a scoring matrix, the words with small scores are eliminated and remaining words are locally aligned. The similarity evaluation by

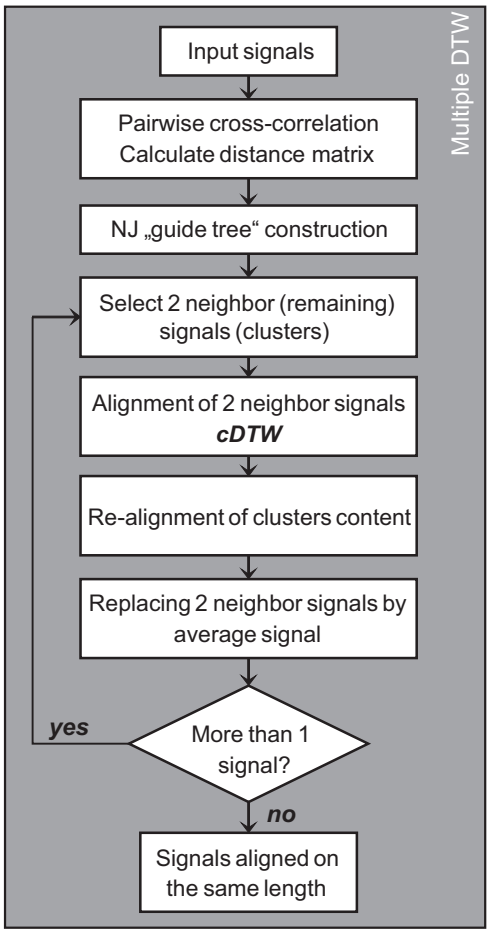


Fig. 1. Flowchart of multiple DTW.

correlation in a sliding window is similar to scoring of k-tuples but subsequent utilization of local correlation coefficient to create the distance matrix for dynamic programing is different.

The principle of this step is demonstrated in Fig. 2. The segment of one signal with length w is sequentially shifted along the second signal. The shift of window w along the second signal is marked by symbol j. After the window reaches the end of the second signal, the selected window in the first signal is shifted by one sample and the whole process is repeated. The shift of window in the first signal is indexed by i. The Pearson correlation coefficient (corrcoef) $c(i, j)$ is determined in each step for segments with the same length w from both signals addressed in table by i, j shift.

The Pearson corrcoef is within the range $\langle -1;\ +1\rangle$, where the value 1 corresponds with maximum similarity. The conventional evaluation of similarity in matrix of local differences or accumulated distances assumes that the cost of similarity between two same signals is equal to zero. The normalization of corrcoef values $c(i,j)$ to $c_N(i,j)$ is used for maintaining the principle of accumulated distances determination. The normalization is explained in Fig. 2B. The method of calculation of values in the table of accumulated distances $D(i,j)$ was chosen with respect to the requirement for an extension of both signals with the same weight

$$D(i,j) = \min\begin{cases} D(i-1,j)+c_N(i,j) \\ D(i-1,j-1)+c_N(i,j) \\ D(i,j-1)+c_N(i,j) \end{cases} \quad (1)$$

Fig. 2C shows the different states of window moving along the signal. The variant C1 explains the treating of boundary conditions. The signal S_2 is extended by half-window length on both sides through repetition of the first or the last value of signal S_2. The variant C1 shows the initial position of the window, where the shift of window j is equal to zero. The shift of window in Fig. 2C2 represents an ordinary case, where corrcoef between two segments from two signals with different trends and different positions $i \neq j$ is within the range $c \in \langle -1;\ +1)$.

The case where $c=1$ is shown in Fig. 2C3, the positions i, j of selected segments are still not equal, but trends are the same. In common DTW method, the local similarities are evaluated as a difference of values on the y-axis (e.g. amplitudes of signals). So, the different size of both signals on y-axis causes failure to detect similarities in the trend. The similar trend of genomic signals represents the similarities in genetic code and the different size on y-axis can be caused by different lengths of both signals. The influence of different length on the size of signals can be removed by detrendisation (Skutkova et al., 2013), but this step can cause loss of genetic information. However, our modified cDTW method evaluates the local similarity by Pearson correlation coefficient which is not affected by different offset of both signals.

The result of local similarities of signal segments in sliding window depends on choice of numerical map for conversion of DNA sequences to signal form and used metrics. The cumulated phase contains position specific information and also Pearson correlation coefficient depends on the order of elements in the vector. The conservation of positional information in chosen signal representation or similarity metrics is essential for the function of the algorithm.

2.2.2. *Progressive alignment utilization*

The pairwise alignment of two signals by DTW requires finding the optimal path in 2D matrix of accumulated distances. The each additional signal causes an increase of matrix dimension. It is hard to imagine finding an optimal path in more than 3D space (for three signals). In addition, it is very computationally demanding – the order of growth is approximately $O(n^m)$ for realization of m-way alignment of length n. The heuristic approach, such as progressive alignment, allows to reduce the order of growth of computational complexity to $O(mn^2)$ (Wang and Jiang, 1994). This is accomplished by systematic repetition of hierarchical aligning of two signals according to the guide tree.

The guide tree is the result of cluster analysis allowing sequential adjustment of signal pair from most similar to more distant. The Neighbor-Joining (NJ) method (Saitou and Nei, 1987) was chosen for construction of the guide tree by the example of CLUSTAL series algorithms (Chenna et al., 2003; Larkin et al., 2007). The adequacy of the choice of NJ guide tree is justified by the evolutionary correctness against the common cluster analysis methods (e.g. UPGMA). The respecting of least squares principle provides a minimum number of length adaptations for length integrity achievement (Bryant et al., 2007; Mihaescu et al., 2009).

The construction of the guide tree (3rd block in Fig. 1) depends on the previous calculation of the distance matrix (2nd block in Fig. 1). The signal cross-correlation (xcorr) was chosen for the similarity degree evaluation between two waveforms with different lengths. The alternative approach requiring pairwise alignment of all pairs of signals with subsequent similarity determining (e.g. by Euclidean distance metric) was computationally too demanding. A similar alignment solution in a symbolic sequence could be performed by a gapless pairwise local alignment (Karlin and Altschul, 1993). Whereas the calculation of similarity between signals for a guide tree construction is only approximate, the massive decimation of signal representations of long sequences (e.g. whole genomes) can be advantageous before the calculation of cross-correlations. The output of the cross-correlation is a

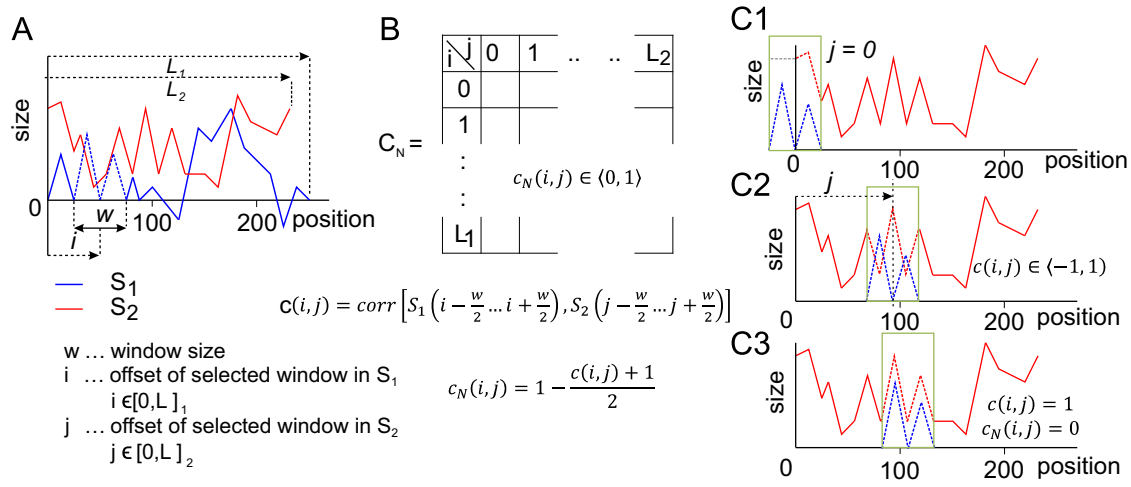


Fig. 2. The principle of cDTW determination: (A) Two signals S_1 and S_2 with different lengths L_1 and L_2 respectively. The selected segment from S_1 specified by length w and shift i is marked with a dotted line. (B) Table of normalization correlation coefficient c_N and equations for their calculation. (C) The principle of shifting the window with selected segment from signal S_1 along the signal S_2: (1) The boundary conditions solution; (2) Symbol j labels a general shift of window along signal S_2. Two selected segments marked by dotted lines are correlated. (3) The correlation between corresponding segments of genomic signals.

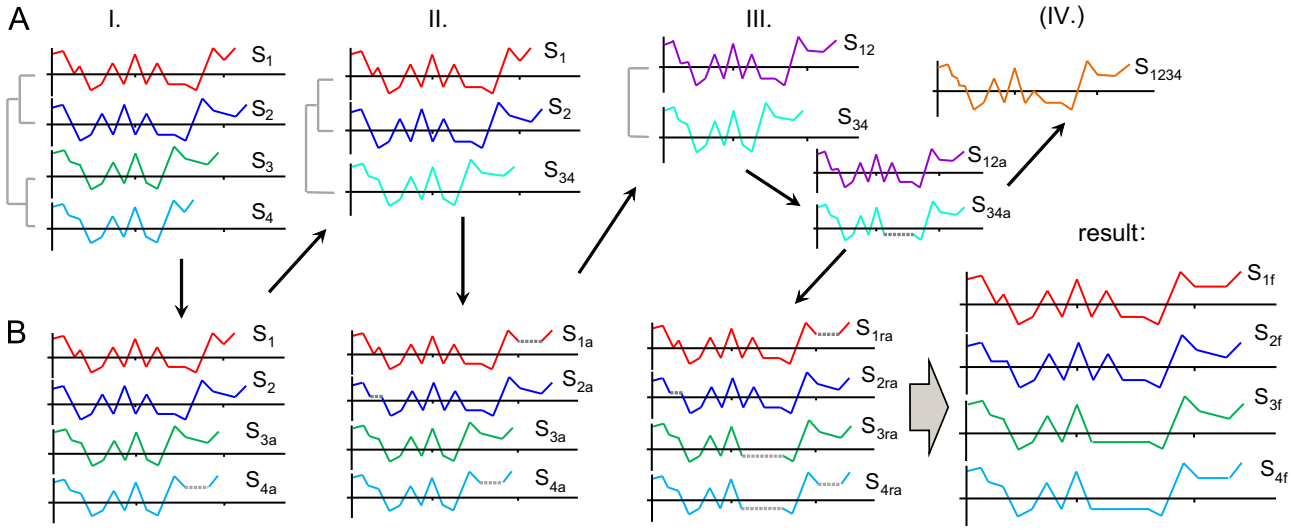


Fig. 3. The principle of mDTW determination: (A) A progressive alignment procedure using hierarchical clustering. (B) The development of signals length in each step of progressive alignment. I–IV represent the steps in progressive alignment. Symbols: a – aligned signals, ra – re-aligned signals, f – the final form of the signals.

cross-correlation signal (vector) R_{12} with length $2N-1$, where N is a length of the longest signal. The peak of xcorr signal corresponds to the position of the best match of mutually shifted signals. This maximum xcorr value was used as a similarity measure of two signals in distance table. The xcorr signal was normalized in order to eliminate the influence of different size (amplitude) of signals. The normalized value of the peak of xcorr signal r_{12NORM} is hereafter referred as the xcorrcoef. The normalization of xcorr signal was realized by dividing by the normalization coefficient n_{12}, so the auto-correlations at zero lag are identically 1.0.

$$R_{12}(m) = \begin{cases} \sum_{i=0}^{N-m-1} S_1(i+m)\cdot S_2(i) & m \geq 0 \\ R_{21}(-m) & m < 0 \end{cases},$$

$$n_{12} = \sqrt{\sum_{i=0}^{N} |S_1(i)| \cdot \sum_{j=0}^{M} |S_2(j)|},$$

$$r_{12NORM} = \frac{1}{n_{12}} \max(R_{12}), \qquad (2)$$

The S_1 and S_2 are vectors of values representing both signals with length N and M respectively and m is mutual shift of both signals. Before placing the values to the distance matrix for construction NJ guide tree, it is necessary to flip the range of values of xcorrcoef, so the maximum similarity of two signals is equal 0, as in the case of corrcoef for cDTW.

Fig. 3 demonstrates how the individual steps of progressive alignment adjust lengths of individual signals. At first, we have four signals (S_1–S_4) with different lengths. The distance matrix and NJ guide tree were determined (Fig. 3A – I step). According to the gray marked guide tree, the first selected neighbors (4th block in Fig. 1) are signals S_3 and S_4. The stage Fig. 3B – I shows the neighbor signals S_3 and S_4 aligned to the same length by pairwise cDTW and hereafter referred as S_{3a} and S_{4a} (a=aligned). The other two signals are left unchanged in the first step. The distance matrix and the guide tree are reduced by one, the average signal S_{34} is determined from signals S_{3a} and S_{4a} (Fig. 3A – II step). The average signal is equivalent to the sequence consensus, but two signal values are more feasible to average than two symbols (e.g. nucleotide symbols A, C, G and T). The utilization of average signal

does not cause distortion there, because the average signal of two previous aligned signals serves only as an auxiliary vector in clustering process. The average genetic information from them is not transmitted to next results.

The second nearest neighbors S_1 and S_2 are aligned in second step and the average signal S_{12} is determined just as in the first step. Now, we have two pairs of signals with the same length within the couple, but mutually different (Fig. 3B – II step). This fact respects the last cluster of guide tree containing two average signals S_{12} and S_{34} (Fig. 3A – III step). After next application of cDTW we obtain aligned signals S_{12a} and S_{34a}. The final alignment of all four input signals is achieved through realignment process (ra) according to optimal path from cDTW of S_{12a} and S_{34a} alignment. In this particular case, the signal insertion was placed only in S_{34a}, consequently the same insertion was placed on the same position in the signals S_3 and S_4.

Generally, the cluster content re-alignment block in Fig. 1 again performs alignment of signals previously aligned separately. The signal at lower levels in the cluster will be aligned multiple times, on the other hand, the signals represented by a separated external branch in the guide tree only once. The described four blocks (4th–7th blocks in Fig. 1) realize the cycle of progressive alignment terminated by 8th decision block. This corresponds to Fig. 3A – IV step, the last average signal S_{1234} does not require length adjustment to some other signal and the cycle is thus terminated. The result four signals (S_{1f}–S_{4f}) have the same length longer than any original, but this is not implicit. The final length of all signals must be equal or greater than maximal length of original signals.

3. Results and discussion

The results from testing of proposed multiple DTW algorithm are problematically displayable and comparable with MSA. The genomic signal alignment is suitable especially for long DNA sequences, but on such a large scale the alignment of symbolic form of DNA sequences becomes unreadable. The interpretation of results of proposed method will be presented on the particular application. The phylogenetic analysis is one of the most affected methods by wrong multiple sequence alignment. The classification of mutual similarity of properly aligned genomic signals will be compared with the results of phylogenetic analysis of symbolic sequences multiple alignment.

3.1. Test data sets

The standard phylogenetic markers 18S rRNA, 16S rRNA were selected for this purpose (Hillis and Dixon, 1991; Field et al., 1988). The already compiled data sets from regular phylogenetic studies were used for proper evaluation of proposed method. The first set consisting of 40 tetrapod 18S ribosomal RNA gene represents problematic alignment even for MSA in symbolic form (Xia et al., 2003). The second set was chosen for verifying the ability of the method to separate sequences of the same species. The 42 sequences of 16S rRNA genes from 6 primate species (family *Hominidae)* were tested for this purpose (Noda et al., 2001). The particular sequences names and their identifiers will be shown in result figures of phylogenetic trees.

The third dataset used in this paper was compiled to demonstrate the effectiveness of the proposed method for analysis of whole genome similarity. Fourteen whole genomes of bacteria from *Firmicutes* phylum were selected with regard to the possibility of evaluating taxonomy at class, family and species level. Their list and taxonomy classification is shown in Table 1.

The process of phylogenetic analysis of DNA sequences in the form of genomic signals is shown in Fig. 4. There is apparent division of block diagram in two major parts: genomic signal processing and genomic signal classification.

Table 1
List of used bacterial genomes.

NCBI ID	Species			Length [bp]
	Class	Order	Genus	
NC_020272	***Bacillus amyloliquefaciens* IT-45**			3 928 857
	Bacilli	*Bacillales*	*Bacillus*	
NC_019896	***Bacillus subtilis* str. BSP1**			4 043 754
	Bacilli	*Bacillales*	*Bacillus*	
NC_020244	***Bacillus subtilis* XF-1**			4 061 186
	Bacilli	*Bacillales*	*Bacillus*	
NC_015687	***Clostridium acetobutylicum* DSM 1731**			3 942 462
	Clostridia	*Clostridiales*	*Clostridium*	
NC_017174	***Clostridium difficile* M120**			4 308 325
	Clostridia	*Clostridiales*	*Peptoclostridium*	
NC_017175	***Clostridium difficile* M68**			4 047 729
	Clostridia	*Clostridiales*	*Peptoclostridium*	
NC_007622	***Staphylococcus aureus* RF122**			2 742 531
	Bacilli	*Bacillales*	*Staphylococcus*	
NC_007623	***Staphylococcus aureus* VC40**			2 692 570
	Bacilli	*Bacillales*	*Staphylococcus*	
NC_007168	***Staphylococcus haemolyticus* JCSC1435**			2 685 015
	Bacilli	*Bacillales*	*Staphylococcus*	
NC_008533	***Streptococcus pneumoniae* D39**			2 046 115
	Bacilli	*Lactobacillales*	*Streptococcus*	
NC_017591	***Streptococcus pneumoniae* INV104**			2 142 122
	Bacilli	*Lactobacillales*	*Streptococcus*	
NC_017592	***Streptococcus pneumoniae* OXC141**			2 036 867
	Bacilli	*Lactobacillales*	*Streptococcus*	
NC_003098	***Streptococcus pneumoniae R6***			2 038 615
	Bacilli	*Lactobacillales*	*Streptococcus*	
NC_008533	***Thermoanaerobacterium thermosaccharolyticum* DSM 571**			2 785 752
	Clostridia	*Thermoanaerobacterales*	*Thermoanaerobacterium*	

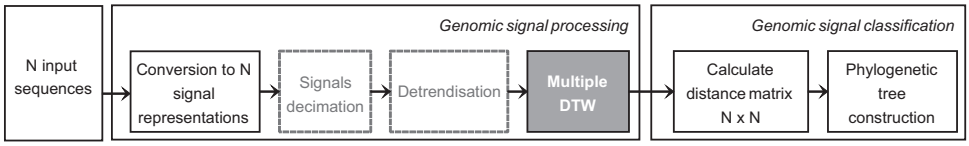


Fig. 4. The block diagram of phylogenetic analysis of genomic signals.

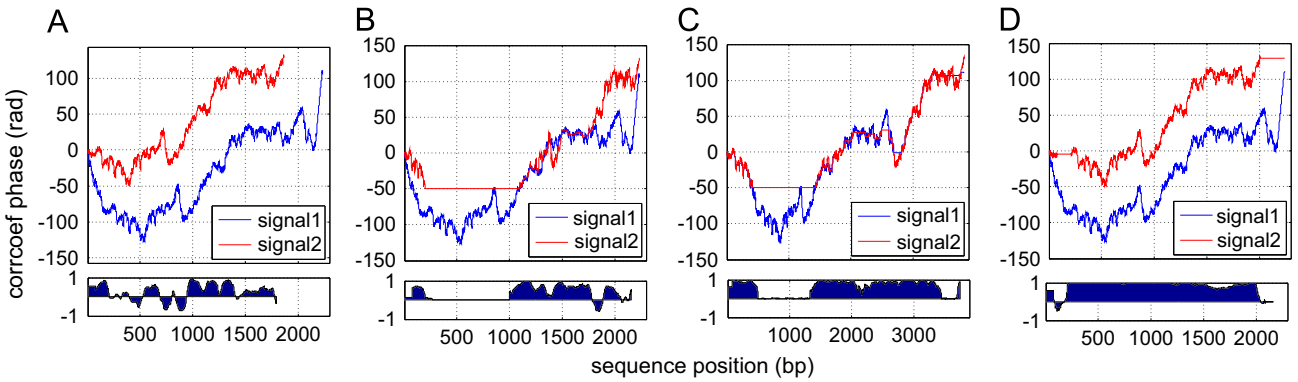


Fig. 5. The comparison of genomic signals alignment results by different DTW modifications: (A) Two original genomic signals of 18S rRNA genes from human (signal 1, longer) and rat (signal 2, shorter). (B) The genomic signals aligned by classic DTW method, the shorter signal is adapted to the longer one based on differences between the local values. (C) The genomic signals aligned by classic DTW method, both signals are mutually adapted based on differences between the local values. (D) The genomic signals aligned by cDTW method, both signals are mutually adapted based on correlation coefficients evaluated in a sliding window. The bottom figures – the similarity of genomic signals evaluated by correlation coefficients.

3.2. Genomic signal processing

The first super-block of genomic signal processing tools includes also the block of multiple DTW. The other three blocks form the pre-processing part differing according to the particular application. The choice of cumulated phase for conversion from symbolic representation of DNA to genomic signal was explained in Section 2.1. The numerical map for conversion to complex space is defined by complex coordinates: A $[1,j]$; C $[-1,-j]$; G $[-1,j]$; T $[1,-j]$. The different type of complex mapping provides the same results of signal alignment, but taxonomy specific trend of chosen map makes it advantageous for the subsequent phylogenetic analysis (Cristea, 2003). The trend of cumulated phase is evaluated as cumulative sum of phase component of complex coordinate of symbol in DNA sequence.

The dashed line marked blocks are optional. Their application is preferable especially for very long sequences as whole genome or chromosome. The signal decimation block, as its name suggests, realizes reduction of input signal size by downsampling. This step causes the loss of information at high frequencies, but frequency analysis of genomic signals proved that the major part of signal information is found on lower frequencies. Despite the loss of genetic information, the remaining information in genomic signals after downsampling allows to classify species based on whole genome sequences, which is hardly possible in normal situations.

Genomic signal representation is characterized by taxonomic features in a large scale (whole genomes) but also in partial segments (genes, CDS). However, the large scale trend complicates the recognition of local similarities, which is necessary for the signal alignment by standard DTW. The problem of large scale detrending can be easily solved, but distinguishing the large and the small scale trend is not trivial. The large scale detrendisation can cause loss of information about the local similarities in some cases. The cDTW utilization in mDTW allows aligning of signals without detrendisation, but still the very different trends of two signals limit the range of values for selection of window length w. The detrendisation allows us to use a smaller window with a smaller overlap, which reduces the computational demands. The utilizations of detrendisation and decimation blocks were already discussed in Skutkova et al. (2013) and Sedlar et al. (2014), so there are mentioned only marginally. The genes selected for testing and comparison with the usual phylogenetic method do not require their application.

3.2.1. The DTW modifications comparison

The DTW methods align local segments of signals based on local distances between values of signals (phase on y-axis for cumulated phase variant) on each position. The alignment of genomic signals requires adjustment of local similarities in trends of signals, because the similar trends represent the homology in DNA sequence. The insertion/deletion mutation type causes different size of local trends due to calculation process of cumulated phase, though the shape remains the same. The cDTW modification successfully solves this problem. Fig. 5 shows an alignment of two genomic signals by different modifications of DTW. The original signals are presented in Fig. 5A. The longer of both signals (18S rRNA human gene, K03432) contains in a comparison with the shorter one (18S rRNA rat gene, K01593) two considerable insertions at the beginning and at the end of the sequence. Although the middle parts of the signals have a practically the same shape, the different size of their values is evident. The application of classic DTW algorithm for length adaptation of shorter signal to longer (Fig. 5B) one failed in this case. Even common variant of DTW is adapting the length of both signals mutually (Fig. 5C), but this approach is still based on local differences between values which cannot estimate the insertion at the start and at the end of human signal. The best estimate of both insertions was performed by cDTW modification (Fig. 5C). The values of adjusted signals in Fig. 5B and C are more consistent but DTW causes distortion by adjustment of signals positions that do not correspond to each other.

3.2.2. The mDTW and multiple sequence alignment

The result of mDTW applied to 40 genomic signals of 18S rRNA is shown in Fig. 6B. The multiple sequence alignment of the same set of sequences realized by Clustal W2 algorithm (online tool with default settings available on website http://www.ebi.ac.uk/Tools/msa/clus

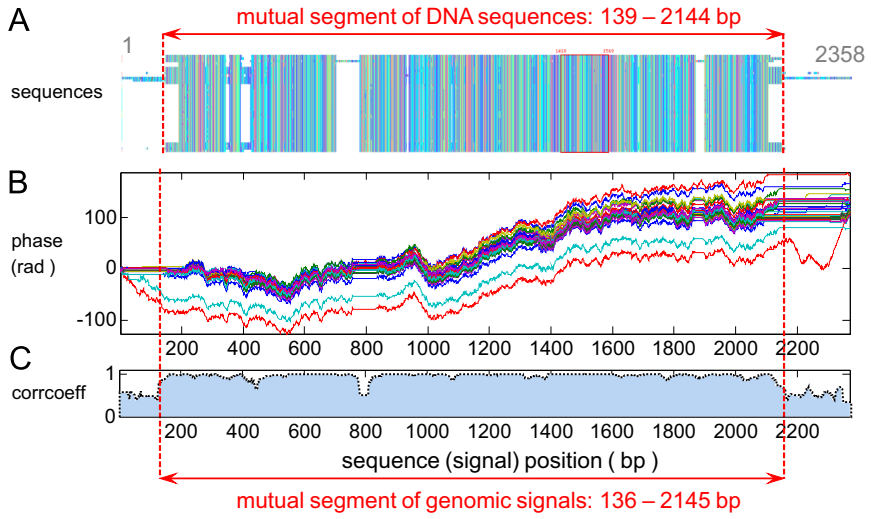


Fig. 6. The comparison of multiple alignment of genomic signals and symbolic DNA sequences of 40 tetrapod genes of 18S rRNA: (A) The illustration of the result of multiple sequence alignment realized by CLUSTAL W2 algorithm. (B) The result of multiple alignment of genomic signals realized by mDTW. (C) The mutual similarity of aligned signals evaluated by average correlation coefficients.

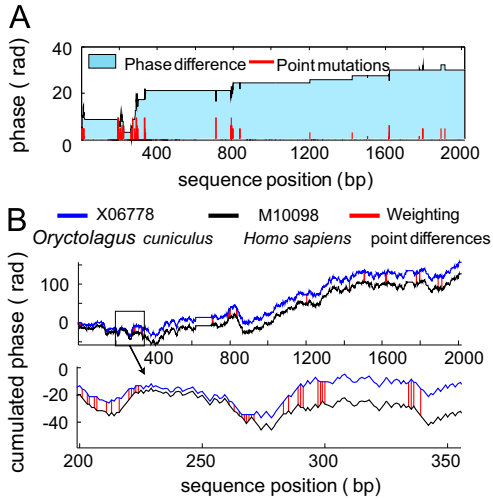


Fig. 7. Point mutation detection between two aligned genomic signals: (A) Phase difference is determined between two genomic signals represented by cumulated phase. Point mutations are absolute value of derivation of phase difference signal. (B) The column height of point mutations is weighted by local difference between signals bottom – the detail of segment between 200–350 bp.

talw2/) was evaluated for comparison (Fig. 6A). The indels segments represented by scoring gaps (—) in sequence form are represented by linear segments without a change of value in signal form. The advantage of using cumulated phase representation is that this signal form does not contain segments without phase changes. The aligned signals can be still re-converted to symbolic representation, where the plateaus are substitutable by gaps. The segment containing the greatest consensus in signal form on positions 136–2145 bp in aligned signals corresponds with mutual segment of DNA sequences selected on positions 139–2144 bp. The segment in symbolic representation of multiple alignment was selected by sequence consensus. The corresponding segment in signal alignment was found by similarity measure in Fig. 6C, evaluated as the average corrcoef in a floating window.

3.3. Genomic signals classification

The BIONJ method (Gascuel, 1997) for evaluation of phylogenetic relationship based on genomic signal similarity was chosen. The character based methods as maximum parsimony or maximum likelihood included in above mentioned phylogenetic studies are not applicable on signal data. The BIONJ is sufficiently precise and fast alternative of distance based methods, but phylogeny accuracy of result depend on similarity measure used as evolutionary distance in phylogenetic tree. The similarity analysis of genomic signals requires designing appropriate measure. The standard bootstrapping statistic test was used for evaluation of phylogenetic tree. Both results of phylogenetic analysis in Figs. 8 and 9 have marked bootstrap support values evaluated by 100 pseudo-replications.

3.3.1. The similarity measure of two aligned signals

The many distance metrics was described for the evaluation of similarity of two signals in signal processing (e.g. Euclidean distance, Cosine distance, City block distance). Most of them can be used for signals similarity classification, but the result scale of phylogenetic tree does not correspond to evolutionary measure. The bottom images in Fig. 5 or in Fig. 6C show a similarity degree between signals evaluated based on position. However, the pairwise similarity between two aligned signals must be evaluated by a single number. The similarity metric as proportional distance for genomic signals is necessary to additionally define. The calculation is based on the sum of the differences of two signals; in the case of cumulated phase representation it is a phase difference. Fig. 7A shows the dependence of phase difference between two signals on the sequence position. The gradual increase of the value of phase difference corresponds with the point mutations occurrence. The position of point mutations is evaluated by using derivation function of phase difference signal. The number of point mutations relative to the length of signals represents the proportional distance as it is usual for describing of symbolic sequences of DNA.

The proportional distance can be weighted by size of local differences between signals in point mutations positions as shown in Fig. 7B. The proportional distances determined as pairwise distances of all pairs of genomic signals can be composed to distance matrix (6th block in Fig. 4) and utilized for phylogenetic analyses. This modification is not appropriate for some evolutionary processes, but it allows enhancing the resolution of species within clusters where it is necessary.

3.3.2. Tetrapod phylogeny based on 18S rRNA

The similarity of aligned signals from Fig. 6B was evaluated by proposed pairwise proportional distance metric without weighting. The result phylogenetic tree is shown in Fig. 8. The basic division is to

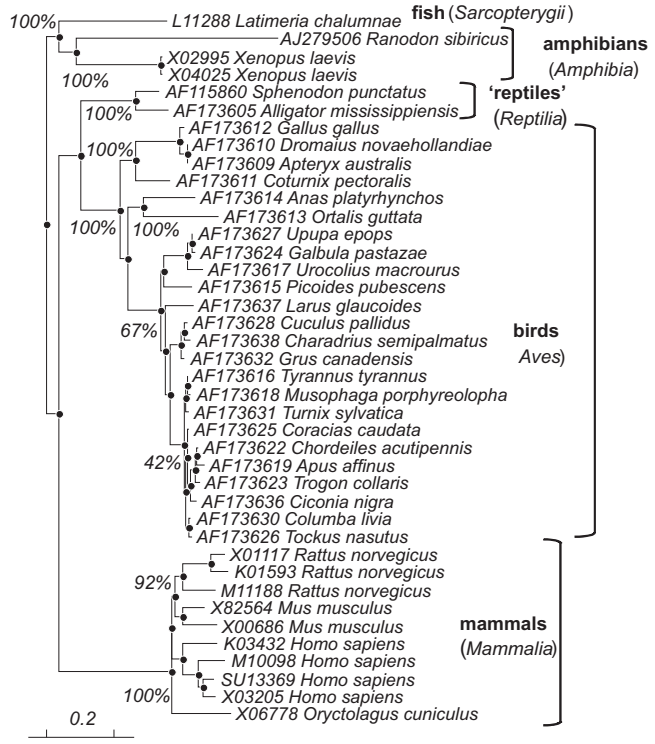


Fig. 8. The tetrapod phylogeny evaluated from the differences between aligned genomic signal representations of 18S rRNA genes.

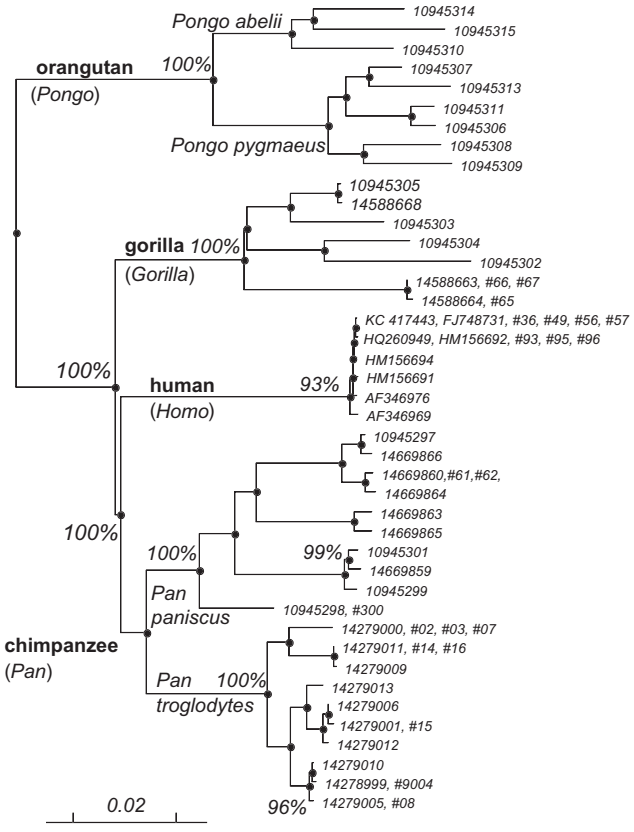


Fig. 9. The hominoid phylogeny evaluated from the differences between aligned genomic signal representations of 16S rRNA genes.

mammals, birds, reptiles, amphibians and one representative fish – *Latimeria chalumnae*. The inaccuracy in assignment of Latimeria close to amphibians cluster is caused by midpoint rooted method ensuring the tree integrity. The complication in alignment of 18S rRNA sequences lies in the different length of mammalian and avian – reptilian genes. The previous phylogenetic studies preferred the cutting off indels segments at the beginning and the end of mammalian sequences in multi-alignment process. Then, the phylogenetic analysis is due to the similarity of nucleotide frequencies between mammalian and avian sequences (GC content) susceptible to assignment of the birds closer to mammals than to reptiles (Huelsenbeck et al., 1996; Bernardi, 1993). So, the alignment-free methods for organisms classification based on characteristic nucleotide frequencies (Otu and Sayood, 2003; Yu, 2013; Kolekar et al., 2012) may be affected the same. The phylogenetic analysis of signals aligned based on mDTW correctly classifies the reptiles and the birds to each other before the mammals, because the similarity depends on the similar trend not the similar frequencies. The two reptiles are incorrectly assigned to one cluster, despite the current vertebrate taxonomy nomenclature (Meyer and Zardoya, 2003; Federhen, 2012) recognizing *Alligator mississippiensis* as a relative closer to *Aves*. This division is caused by insufficient genetic variability in data sets of 18S rRNA for their separation to individual clusters and is consistent with the Xia et al. (2003). The NJ based methods often group together one of the three rat sequences with the mouse sequences (Bruno et al., 2000), but the proportional distances of genomic signals allow the correct classification. The two records of rat sequences (K01593, X01117) evaluated as the closest in the tree are unified in NCBI database at present. It is interesting that the record X82564 of mouse gene from reference phylogenetic study is described as 45S pre rRNA gene in public database and has length 22 118 bp. The segment from this gene corresponding with other mouse 18S rRNA genes was selected by BLAST and used for our testing. The selected fragment had the specific trend in genomic signal common to 18S rRNA genes and was correctly assigned to the cluster with other rodents.

The internal division of avian cluster does not correspond to the conventional taxonomy of birds. The six avian species (AF173612 *Gallus gallus*, AF173614 *Anas platyrhynchos*, AF173610 *Dromaius novaehollandiae*, AF173609 *Apteryx australis*, AF173611 *Coturnix pectoralis*, AF173613 *Ortalis guttata*) are significantly separated from other 18 species as in Xia et al. (2003). The remaining 18 bird species are almost indistinguishable by conventional phylogenetic methods; their genetic variability consists of maximum 8 point mutations on 1.8 kbp. The 18S rRNA is not suitable for estimation of *aves* interspecies relationships. Fig. 8 shows that the proportional distance between avian species estimated from the genomic signal representation of 18S rRNA genes allows increase of their resolution within cluster. However, even there the increase of the differences between the species is not caused by contained genetic information, but due to the choice of computational window. The further adjustment of parameters as weight, window length or window shift would result to adding variability to classification process which does not correspond with the real genetic variability.

3.3.3. *The ability to distinguish the same species*

The 16S rRNA mitochondrion gene is known as a suitable phylogenetic marker able to distinguish even very close species (Galtier et al., 2009). The advantage of 16S rRNA as phylogenetic marker is presence of genetic variability in indels mutations not only in substitutions. The range of length of hominoid's genes in our study is between 1537–1715 bp. The difference of lengths almost 200 bp for such close species is appreciable and their phylogenetic analysis requires responsive tool for length adjustment. Fig. 9 shows the phylogeny reconstruction based on aligned genomic signal representations of 16S rRNA genes. The relationship of 6 great apes species was evaluated correctly (Prado-Martinez et al., 2013; Noda et al., 2001) with 100% bootstrap support. The weighted proportional distance of genomic signals was used for differentiation of several different

records from each species. The scale of weighted distance is influenced by size of cumulative phase, so result values can reach enormous size (e.g. millions of radians for whole genome). The normalization as in the case of corrcoef (2) was used for approximation on conventional values. The bootstrap support decreases slightly within clusters of some species. This fact is noticeable especially for the "youngest" species human and chimpanzees. The diversity of primary structure of these records within species is smaller than for the others. We deliberately added, in comparison with the original study, several other records for human 16S rRNA genes. It is obvious, that these records have practically the same primary structures and weighting of the proportional distances correctly did not add any extra information about genetic diversity.

3.3.4. The mDTW parameters settings

As in the case of multiple sequence alignment the accuracy of the result of mDTW depends on the appropriately selected alignment parameters. Advantage against to MSA is that the mDTW does not require the scoring matrix for weighting of substitution cost. Since, in numerical representation each sequence symbol has a specific numerical value characteristic according to its chemical properties, the substitution score is given by difference of values which represent two substituted symbols. Analogy to gaps penalties in MSA are weights of tree possible directions in calculation of table of accumulated distances (3).

$$D(i,j) = \min\begin{cases} D(i-1,j)+v_w c_N(i,j) \\ D(i-1,j-1)+d_w c_N(i,j) \\ D(i,j-1)+h_w c_N(i,j) \end{cases}, \quad (3)$$

The choice of weighting factors h_w, d_w, v_v ratio (horizontal, diagonal and vertical directions respectively) determines the preference between increasing or maintaining of the sequence length. The weights for horizontal and vertical directions should be consistent if not preferred by sequences otherwise. The diagonal weight is set to a value lower than the horizontal (vertical) weight if the original sequences contain mostly substitution type of mutation against to indels, then we prefer the maintaining of the sequence lengths.

The most important parameters are the length and the shift of the window in cDTW calculation. The generally recommended length of the window is values between 1/100 and 1/200 length of the longest sequence. The classification into order (for 18S rRNA) or interspecies variability evaluation corresponds with the theoretical expectations in this range of window lengths. The classification of 16S rRNA was slightly different only in clustering of very similar human records. The choice of length for construction of phylogenetic tree from 18S rRNA gene was complicated by poorly defined reference of avian genes clustering. Fig. 10 shows the dependence of classification variability based on window size. The classification differences were evaluated as number of different nodes against to topology in Fig. 8. based on the Robinson–Foulds metric (Robinson and Foulds, 1981). In the range of window length between 8–30 bp occurred changes in clustering of problematic 18 species of avian during the classification. The classification of other species to order was stable in this range. The range between 8–30 bp can be suitable for classification of organisms to order, the higher values allow reliable classification only to the class. The value of the window size must be an even because divisibility of the two is required in boundary conditions.

3.3.5. The whole genome alignment for estimation of bacterial phylogeny

The step for the shift of the window along the signal should be always equal 1 for evaluation of point mutations. Therefore, the previous two examples do not allow analysis of the influence of larger step to genomic signal alignment and subsequent classification. The comparison of whole genomes or chromosomes does not require such

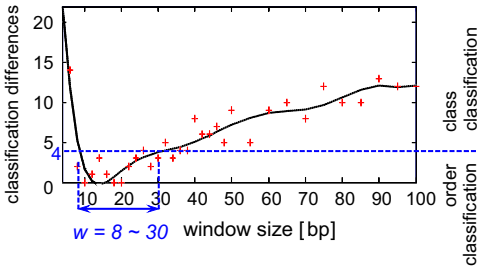


Fig. 10. The dependence of classification differences on the change of window size for cDTW performance – evaluated for 40 sequences of 18S rRNA vertebrate genes.

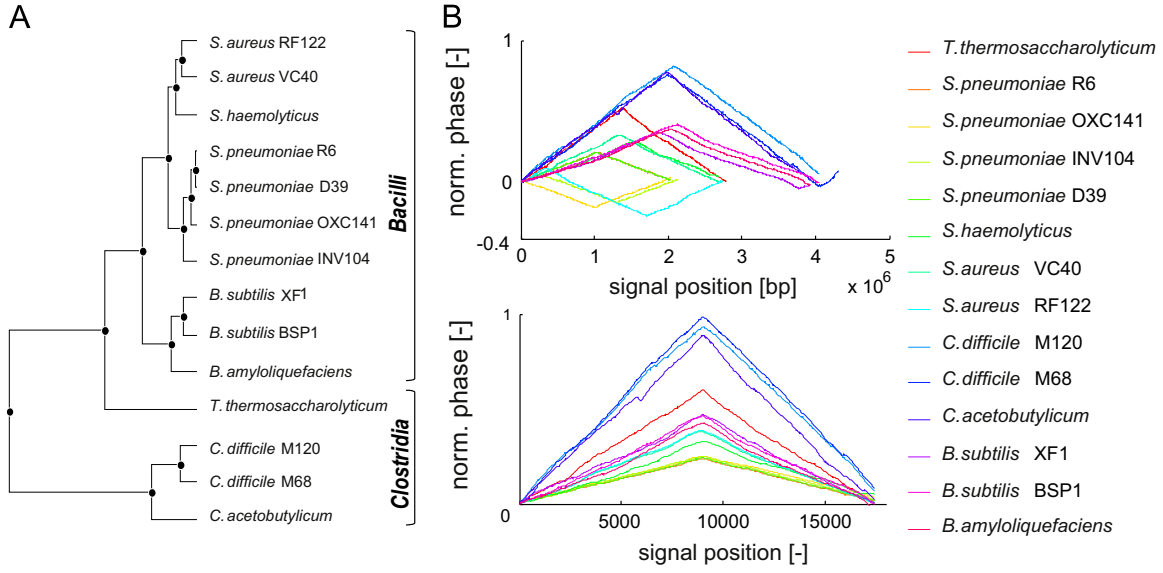


Fig. 11. Phylogeny classification of bacterial whole genome signals. (A) Phylogenetic tree for 14 representatives of *firmicutes* phylum reconstructed from multiple aligned 2^8 decimated genomic signals. (B) Cumulated phases of bacterial whole genome sequences (upper). Cumulated phases after re-assembly with oriC finder, decimation by 8 level of decomposition and multiple alignment to the same length.

sensitivity to point mutations; so it allows the higher step size up to the size of the window w. The set of whole bacterial genomes described in Section 3.1 was used for this purpose. The upper chart in Fig. 11B shows their cumulated phase representations. The whole bacterial genome records often have misidentified oriC, which causes piecewise refracted shape of the curves of genomic signals. The reassembly according to the correctly detected position of oriC (Gao and Zhang, 2008) is required before decimation and alignment. The original size of bacterial genome in millions of bps does not allow multiple alignment in signal or symbol representation of DNA sequences, but massively decimated genomic signals preserve enough information for phylogenomic classification (Sedlar et al., 2014). The signal downsampling done by dyadic discrete wavelet transform was used for decimation of genomic signal in range of 2^8–2^{17}. The smaller level of decomposition than 8 produces signals which are still too computationally demanding (on desktop PC). Downsampling of the shortest genomes with the factor of 2^{18} retains only one sample from original signal which is not enough for classification. (Sedlar et al., 2015) The choice of window size depends on length of aligned signals. The longer signals obtained with smaller level of decomposition require a larger window size. For the particular level of decomposition was chosen the size of the window as the smallest possible value providing the taxonomically most correct result according NCBI taxonomy (Federhen, 2012; Wolf et al., 2004), selected values are in Table 2. The reference phylogenetic tree in Fig. 11A was determined based on Euclidean distances of the 2^8 decimated signals multiple aligned with the size of correlation window $w=28$ and the step of window shift equal to 1. It took about 4 h of computational time. The result of alignment is shown in the lower chart of Fig. 11B. The classification of almost all species to clusters is correct according to NCBI taxonomy, only *T. thermosaccharolyticum* was assigned closer to *Bacilli* class than to *Clostridia*. The cause of this condition is probably different character of selected organisms (oxygen or temperature preferences).

The robustness of initial topology was analyzed in dependence on growing step size in four stages: 1, 2, $w/2$ and w. Result misclassification determined by RF distance against to the reference topology and processing time necessary for multiple alignment is in Table 2. The tree topology remained unchanged up to level of decomposition 12 with the window size equal to 1. The step increased to 2 samples retains topology too, but the processing time decreased about one third. One level of decomposition lower decimation allows preserving topology up to $w/2$ step size for more than half processing time savings. In addition to these marginal values, the optimum setting seems to be 11th level of decomposition with the step equal to 2, which provides sufficient accuracy for bacterial phylogenomic study with the result in minutes.

Table 2
The influence of the window step size on classification correctness and processing time.

Level of decomposition	w	RF distance for step size				Processing time [s] for step size			
		1	2	$w/2$	w	1	2	$w/2$	w
8	28	0	0	0	0	14,435.39	12,783.76	8314.04	5186.11
9	20	0	0	0	0	3409.42	3149.05	1872.02	1162.50
10	12	0	**0**	**0**	2	483.75	204.78	159.26	126.52
11	10	0	**0**	**0**	4	82.12	**66.34**	**38.47**	37.06
12	8	**0**	**0**	4	6	**19.44**	**12.56**	7.28	6.95
13	8	2	2	4	6	7.20	4.97	3.83	3.12
14	6	4	6	6	10	4.00	3.14	3.02	0.76
15	6	4	8	6	10	3.05	1.56	0.86	0.24
16	6	8	12	14	12	1.87	0.49	0.49	0.48
17	6	10	12	14	16	1.75	0.49	0.47	0.37

4. Conclusion

The multiple sequence alignment algorithms are often solved problem in bioinformatics nowadays. The wrong sequence position adjustment in multi-alignment process influences the outcome of majority part of subsequent sequence analysis e.g. motif finding, finding of homology segment, protein structures prediction and last but not the least also the phylogenetic analysis. The genomic signal processing proved to be a suitable just for the phylogenetic analysis, because the genomic signal representations of biological sequences have taxonomy specific trend. However, since there was not any equivalent tool for MSA of genomic signals, this article presents such a tool and tests its suitability for phylogenetic classification of genomic signals. The new method called multiple DTW abbreviated as mDTW has proved to be adequate for alignment of signal representations of two data sets of standard phylogenetic markers 16S rRNA and 18S rRNA genes. Multiple DTW, such like MSA, allows setting of several parameters based on sequence type: weights between elongation and maintaining sequence lengths, window length and shift for sensitivity to point or larger mutations. The advantage against to MSA is absence of scoring matrix choice, because substitution score is contained in numerical code of signals due to the choice of numerical map respecting similarities between characters. The resulting alignment of mDTW is also more robust due to correlation based partial pairwise alignments by cDTW. The principle of cDTW is alignment based on local homologies not only on point differences. Due to the fact, that in the genomic signal representation by cumulated phase the value of particular sample depends on all previous sample values, this idea is close to the construction of position-specific scoring matrix (PSSM). Of course, the greatest contribution of this approach for comparing biological sequences in signal forms lies in possibilities of massive signal decimation. The computational time for mDTW execution is approximately by an order higher than for MSA, due to the fact that cDTW analyses similarity of local segments and not only one position. However, decimation of signals in order of gene size by downsampling factor just 10 reduced the computational time by more than two orders. Moreover, ten-fold decimation still preserves more than 99.5% of original signal information, so the differences of distances in the phylogenetic tree are under 1% against to phylogenetic analysis without decimation (Skutkova et al., 2013). The redundancy of signal information in genomic signals grows linearly with length of signals which allows even thousand-fold downsampling for whole chromosomes (Sedlar et al., 2014). The computational complexity of progressive alignment of signals as well as sequences grows quadratically with signals (sequences) length. The computational time for progressive alignment of multiple signals quadratically decreases with their length decimation. Such a huge advantage is unreachable for character based MSA. In future, the increasing number of whole genome records pushes the phylogenetics into background of phylogenomics. The mDTW of decimated whole genome signals will be a valuable connecting link for phylogenomic studies.

Acknowledgment

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

References

Almeida, J.S., Carrico, J.A., Maretzek, A., Noble, P.A., Fletcher, M., 2001. Analysis of genomic sequences by chaos game representation. Bioinformatics 17, 429–437. http://dx.doi.org/10.1093/bioinformatics/17.5.429.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Anastassiou, D., 2001. Genomic signal processing. IEEE Signal Process. Mag. 18, 8–20. http://dx.doi.org/10.1109/79.939833.

Berger, J.A., Mitra, S.K., Carli, M., Neri, A., 2004. Visualization and analysis of DNA sequences using DNA walks. J. Frankl. Inst. 341, 37–53. http://dx.doi.org/10.1016/j.jfranklin.2003.12.002.

Bernardi, G., 1993. The vertebrate genome: isochores and evolution. Mol. Biol. Evol. 10, 186–204.

Braga-Neto, U., Kuang, R., Lahdesmaki, H., Vikalo, H., Yoon, B.-J., 2010. Genomic signal processing. Eurasip J. Adv. Signal Process. , http://dx.doi.org/10.1155/2010/137263.

Bruno, W.J., Socci, N.D., Halpern, A.L., 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol. Biol. Evol. 17, 189–197.

Bryant, D., Moulton, V., Spillner, A., 2007. Consistency of the neighbor-net algorithm. Algorithms Mol. Biol.2, 8. http://dx.doi.org/10.1186/1748-7188-2-8.

Cristea, P.D., 2002. Conversion of nucleotides sequences into genomic signals. J. Cell. Mol. Med. 6, 279–303. http://dx.doi.org/10.1111/j.1582-4934.2002.tb00196.x.

Cristea, P.D., 2003. Large scale features in DNA genomic signals. Signal Process. 83, 871–888. http://dx.doi.org/10.1016/S0165-1684(02)00477-2.

Cristea, P.D., Tuduce, R., 2011. Comparative analysis of mitochondrial DNA by using nucleotide genomic signals. In: Mamalis, A.G., et al., (Eds.), Applied Electromagnetic Engineering for Magnetic, Superconducting and Nanomaterials, vol. 670. pp. 507–516.

Cristea, P.D., IEEE2012. Building phylogenetic trees by using gene nucleotide genomic signals. In: Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 5549–5553.

Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6, 361–375. http://dx.doi.org/10.1038/nrg1603.

Deng, M., Yu, C., Liang, Q., He, R.L., Yau, S.S., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS One 6, e17293. http://dx.doi.org/10.1371/journal.pone.0017293.

Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., Fertil, B., 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol. 16, 1391–1399.

Dougherty, E.R., Datta, A., Sima, C., 2005. Research issues in genomic signal processing. IEEE Signal Process. Mag. 22, 46–68. http://dx.doi.org/10.1109/msp.2005.1550189.

Federhen, S., 2012. The NCBI taxonomy database. Nucleic Acids Res. 40, D136–D143. http://dx.doi.org/10.1093/nar/gkr1178.

Feng, D.-F., Doolittle, R., 1987. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. J. Mol. Evol. 25, 351–360. http://dx.doi.org/10.1007/BF02603120.

Field, K.G., Olsen, G.J., Lane, D.J., Giovannoni, S.J., Ghiselin, M.T., Raff, E.C., Pace, N.R., Raff, R.A., 1988. Molecular phylogeny of the animal kingdom. Science 239, 748–753.

Florquin, K., Saeys, Y., Degroeve, S., Rouze, P., Van de Peer, Y., 2005. Large-scale structural analysis of the core promoter in mammalian and plant genomes. Nucleic Acids Res. 33, 4255–4264. http://dx.doi.org/10.1093/nar/gki737.

Galtier, N., Nabholz, B., Glemin, S., Hurst, G.D., 2009. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Mol. Ecol. 18, 4541–4550. http://dx.doi.org/10.1111/j.1365-294X.2009.04380.x.

Gao, F., Zhang, C.T., 2008. Ori-Finder: a web-based system for finding oriCs in unannotated bacterial genomes. BMC Bioinform. 9, 79. http://dx.doi.org/10.1186/1471-2105-9-79.

Gascuel, O., 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14, 685–695. http://dx.doi.org/10.1093/oxfordjournals.molbev.a025808.

Hillis, D.M., Dixon, M.T., 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. Q. Rev. Biol. 66, 411–453.

Hoang, T., Yin, C., Zheng, H., Yu, C., Lucy, He, R., Yau, S.S., 2015. A new method to cluster DNA sequences using Fourier power spectrum. J. Theor. Biol. 372, 135–145. http://dx.doi.org/10.1016/j.jtbi.2015.02.026.

Huelsenbeck, J.P., Bull, J.J., Cunningham, C.W., 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. 11, 152–158.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson, J.D., 2003. Multiple sequence alignment with the clustal series of programs. Nucleic Acids Res. 31, 3497–3500. http://dx.doi.org/10.1093/nar/gkg500.

Chor, B., Tuller, T., 2006. Finding a maximum likelihood tree is hard. J. ACM 53, 722–744. http://dx.doi.org/10.1145/1183907.1183909.

Karlin, S., Altschul, S.F., 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. Proc. Natl. Acad. Sci. USA 90, 5873–5877.

Kolekar, P., Kale, M., Kulkarni-Kale, U., 2012. Alignment-free distance measure based on return time distribution for sequence analysis: applications to clustering, molecular phylogeny and subtyping. Mol. Phylogenetics Evol. 65, 510–522. http://dx.doi.org/10.1016/j.ympev.2012.07.003.

Kung, S.Y., Luo, Y., Mak, M.-W., 2010. Feature selection for genomic signal processing: unsupervised, supervised, and self-supervised scenarios. J. Signal Process. Syst. Signal Image Video Technol. 61, 3–20. http://dx.doi.org/10.1007/s11265-008-0273-8.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and clustal X version 2.0. Bioinformatics 23, 2947–2948. http://dx.doi.org/10.1093/bioinformatics/btm404.

Maderankova, D., Provaznik, I., 2011. Motive representation in nucleotide densities of bird's mitochondrial gene COX1. In: Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies. ACM, Barcelona, Spain, pp. 1–5.

Machado, J.A.T., Costa, A.C., Quelhas, M.D., 2011. Wavelet analysis of human DNA. Genomics 98, 155–163. http://dx.doi.org/10.1016/j.ygeno.2011.05.010.

Meyer, A., Zardoya, R., 2003. Recent advances in the (molecular) phylogeny of vertebrates. Annu. Rev. Ecol., Evol., Syst. 34, 311–338. http://dx.doi.org/10.1146/annurev.ecolsys.34.011802.132351.

Mihaescu, R., Levy, D., Pachter, L., 2009. Why neighbor-joining works. Algorithmica 54, 1–24. http://dx.doi.org/10.1007/s00453-007-9116-4.

Montanola, A., Roig, C., Guirado, F., Hernandez, P., Notredame, C., 2013. Performance analysis of computational approaches to solve multiple sequence alignment. J. Supercomput. 64, 69–78. http://dx.doi.org/10.1007/s11227-012-0751-4.

Noda, R., Kim, C.G., Takenaka, O., Ferrell, R.E., Tanoue, T., Hayasaka, I., Ueda, S., Ishida, T., Saitou, N., 2001. Mitochondrial 16S rRNA sequence diversity of hominoids. J. Hered. 92, 490–496. http://dx.doi.org/10.1093/jhered/92.6.490.

Otu, H.H., Sayood, K., 2003. A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19, 2122–2130. http://dx.doi.org/10.1093/bioinformatics/btg295.

Pearson, W.R., 1998. Empirical statistical estimates for sequence similarity searches. J. Mol. Biol. 276, 71–84. http://dx.doi.org/10.1006/jmbi.1997.1525.

Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A.E., Malig, M., Hernandez-Rodriguez, J., Hernando-Herraez, I., Prufer, K., Pybus, M., Johnstone, L., Lachmann, M., Alkan, C., Twigg, D., Petit, N., Baker, C., Hormozdiari, F., Fernandez-Callejo, M., Dabad, M., Wilson, M.L., Stevison, L., Camprubi, C., Carvalho, T., Ruiz-Herrera, A., Vives, L., Mele, M., Abello, T., Kondova, I., Bontrop, R.E., Pusey, A., Lankester, F., Kiyang, J.A., Bergl, R.A., Lonsdorf, E., Myers, S., Ventura, M., Gagneux, P., Comas, D., Siegismund, H., Blanc, J., Agueda-Calpena, L., Gut, M., Fulton, L., Tishkoff, S.A., Mullikin, J.C., Wilson, R.K., Gut, I.G., Gonder, M.K., Ryder, O.A., Hahn, B.H., Navarro, A., Akey, J.M., Bertranpetit, J., Reich, D., Mailund, T., Schierup, M.H., Hvilsom, C., Andres, A.M., Wall, J.D., Bustamante, C.D., Hammer, M.F., Eichler, E.E., Marques-Bonet, T., 2013. Great ape genetic diversity and population history. Nature 499, 471–475. http://dx.doi.org/10.1038/nature12228.

Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147. http://dx.doi.org/10.1016/0025-5564(81)90043-2.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Sakoe, H., Chiba, S., 1978. Dynamic-programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Process. 26, 43–49. http://dx.doi.org/10.1109/tassp.1978.1163055.

Sedlar, K., Skutkova, H., Vitek, M., Provaznik, I., 2014. Prokaryotic DNA signal downsampling for fast whole genome comparison. In: Piętka, E., et al. (Eds.), Information Technologies in Biomedicine, Volume 3, vol. 283. Springer International Publishing, pp. 373–383.

Sedlar, K., Skutkova, H., Vitek, M., Provaznik, I., 2015. Set of rules for genomic signal downsampling. Comput. Biol. Med. , http://dx.doi.org/10.1016/j.compbiomed.2015.05.022 (In press).

Skutkova, H., Vitek, M., Babula, P., Kizek, R., Provaznik, I., 2013. Classification of genomic signals using dynamic time warping. BMC Bioinform. 14, S1. http://dx.doi.org/10.1186/1471-2105-14-S10-S1.

Song, N.Y., Yan, H., 2012. Selection and mapping of DNA structural features for short gene recognition. Int. J. Data Min. Bioinform. 6, 675–691. http://dx.doi.org/10.1504/ijdmb.2012.050250.

Tao, M., Soliman, A.T., Mei-Ling, S., Yimin, Y., Shu-Ching, C., Iyengar, S.S., Yordy, J.S., Iyengar, P., 2013. Wavelet analysis in current cancer genome research: a survey. IEEE/ACM Trans. Comput. Biol. Bioinform. 10, 1442–14359. http://dx.doi.org/10.1109/TCBB.2013.134.

Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison-a review. Bioinformatics 19, 513–523. http://dx.doi.org/10.1093/bioinformatics/btg005.

Wang, L., Jiang, T., 1994. On the complexity of multiple sequence alignment. J. Comput. Biol. 1, 337–348. http://dx.doi.org/10.1089/cmb.1994.1.337.

Wang, X.D., Dougherty, E.R., Chen, Y.D., Peterson, C.O., 2004. Genomic signal processing – editorial. Eurasip J. Appl. Signal Process. 2004, 3–4. http://dx.doi.org/10.1155/s1110865704002756.

Wen, J., Chan, R.H., Yau, S.C., He, R.L., Yau, S.S., 2014. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. Gene 546, 25–34. http://dx.doi.org/10.1016/j.gene.2014.05.043.

Wolf, M., Muller, T., Dandekar, T., Pollack, J.D., 2004. Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. Int. J. Syst. Evol. Microbiol. 54, 871–875. http://dx.doi.org/10.1099/ijs.0.02868-0.

Xia, X., Xie, Z., Kjer, K.M., 2003. 18S ribosomal RNA and tetrapod phylogeny. Syst. Biol. 52, 283–295. http://dx.doi.org/10.1080/10635150390196948.

Yao, Y.H., Dai, Q., Nan, X.Y., He, P.A., Nie, Z.M., Zhou, S.P., Zhang, Y.Z., 2008. Analysis of similarity/dissimilarity of DNA sequences based on a class of 2D graphical representation. J. Comput. Chem. 29, 1632–1639. http://dx.doi.org/10.1002/jcc.20922.

Yin, C., Chen, Y., Yau, S.S.T., 2014. A measure of DNA sequence similarity by Fourier transform with applications on hierarchical clustering. J. Theor. Biol. 359, 18–28. http://dx.doi.org/10.1016/j.jtbi.2014.05.043.

Yu, H.-J., 2013. Segmented K-mer and its application on similarity analysis of mitochondrial genome sequences. Gene 518, 419–424. http://dx.doi.org/10.1016/j.gene.2012.12.079.