Prokaryotic DNA Signal Downsampling for Fast Whole Genome Comparison

Karel Sedlar[1], Helena Skutkova[1], Martin Vitek[1,2], and Ivo Provaznik[1,2]

[1] Department of Biomedical Engineering, Brno University of Technology, Technicka 12, CZ 616 00 Brno, Czech Republic
sedlar@feec.vutbr.cz
[2] International Clinical Research Center - Center of Biomedical Engineering, St. Anne's University Hospital Brno, Pekarska 53, CZ 656 91 Brno, Czech Republic

Abstract. Classification of prokaryotes is mainly based on molecular data, since next-generation sequencing platforms provide fast and effective way to capture prokaryotes' characteristics. However, two different bacterial strains of the same genus can differ in the specific parts of their genomes due to copious amounts of repetitive and transposable parts. Thus, finding an ideal segment of genome for comparison is difficult. Conventional character-based methods rely on multiple sequence alignment, rendering them extremely computationally demanding. Only small parts of genomes can be compared in reasonable time. In this paper, we present a novel algorithm based on the conversion of the whole genome sequences to cumulative phase signals. Dyadic wavelet transform (DWT) is used for lossy compression of phase signals by eliminating redundant frequency bands. Signal classification is then performed as cluster analysis using Euclidean metrics where sequence alignment is replaced by dynamic time warping (DTW).

Keywords: prokaryotes, genomic signal, cumulated phase, compression, classification, dwt, dtw.

1 Introduction

The classification of organisms is one of the fundamental questions in biology. It is based mainly on molecular characters since DNA is the carrier of heredity [1]. However, new sequencing techniques allow cheap assembly of the entire genomes, particularly prokaryotic genomes formed by single circular chromosomes, since the classical methods of comparison are still unable to process whole chromosomes. This is caused by multiple sequence alignment that is computationally too demanding, even impossible for sequences of length of several Mbp. Only small parts of chromosomes, e.g. single genes, can be processed. Unfortunately, various genes are evolving at different rates, which may not reflect the evolutionary development rate of the whole organism. On the other hand, the conversion from character sequence to numeric signal brings the possibility of using digital

E. Pietka, J. Kawa, and W. Wieclawek (eds.), *Information Technologies in Biomedicine, Volume 3*, Advances in Intelligent Systems and Computing 283,
DOI: 10.1007/978-3-319-06593-9_33,

signal processing techniques for lossy compression. Despite lossy compression, we are able to preserve reasonable amounts of information and significantly reduce the computational demands, so that whole genomes can be compared in a very short time. Several digital signal processing techniques can be used for compression. Here, we present an approach using dyadic wavelet transform (DWT) [2] for its speed and effectiveness.

Other disadvantage of sequence alignment is the necessity of using scoring matrices. Comparison of several sequences based on various scoring matrices leads to a number of different results. This is caused by presumptions concerning the specific speed of evolution of an organism, which is unknown. Multiple sequence alignment can be replaced by dynamic time warping (DTW). It is an algorithm of dynamic programming used for signal alignment. Unlike the multiple sequence alignment, DTW is not dependent on substitution matrix and works with individual nucleotides changes. The previous utilization of DTW in DNA signals alignment can be found in [3].

2 Materials and Methods

A set of several bacterial whole genome sequences was used for comparing our approach with the classical character processing method performed on 16S rRNA, which are the most commonly used short barcode sequences for prokaryotes' identification [4]. Later study shows that using only short sequences can brings many imprecisions [5]. Sequences were obtained from GenBank database at NCBI (`http://www.ncbi.nlm.nih.gov/genbank/`). The characterization of sequences used for analysis is summarized in Table 1.

2.1 Sequence Conversion

A number of techniques for for converting DNA sequences to genomic signals have been published [6], though not all of them can be used for whole genome classification. The preservation of all biological properties is the essential condition during conversion. Thus, we chose cumulated phase signal representation [7]. In this representation each of the nucleotides A, C, G, T occurring in the DNA is reflected in the complex plane in manner such that appropriate complex numbers maintain information on the nucleotides' chemical similarities, see Figure 1(a). Every character along a sequence is replaced by its complex number during transformation: A [1,j]; C [-1,-j]; G [-1,j];T [1,-j].

By the definition, the complex number phase have values $(-\pi, +\pi\rangle$. Using trigonometric functions, we can easily calculate the phase of our four numbers:

$$\{\varphi_A, \varphi_C, \varphi_G, \varphi_T\} = \left\{\frac{\pi}{4}, -\frac{3\pi}{4}, \frac{3\pi}{4}, -\frac{\pi}{4}\right\}. \quad (1)$$

Table 1. The specifications of sequences from seven organisms

Organism	Chr accession	Chr length (bp)	16S length (bp)
Escherichia coli str. K-12	NC_000913.2	4639675	1403
Lactobacillus casei	NC_008526.1	2895264	1568
Lactobacillus crispatus	NC_014106.1	2043161	1552
Lactobacillus gasseri	NC_008530.1	1894360	1579
Salmonella bongori	NC_015761.1	4460105	1542
Salmonella enterica CT18	NC_003198.1	4809037	1542
Salmonella enterica LT12	NC_003197.1	4857432	1542
Thermococcus ga. EJ3	NC_012804.1	2045438	1539
Thermococcus sp. 4557	NC_015865.1	2011320	1496
Pyroccocus fu. COM1	CP_003685.1	1909827	1519
Bibersteinia trehalosi	NC_020515.1	2407846	1528
Proteus mirabilis HI4320	NC_010554.1	4063606	1542
Bordetella per. Tohama I	NC_002929.2	4086189	1992
Acidovorax ebreus TPSY	NC_011992.1	3796573	1971
Thauera sp. MZ1T	NC_011662.2	4496212	1985

The actual signal is gained using cumulating phase numbers (1) of appropriate nucleotides along the sequence or it can be computed directly from character sequence by:

$$\theta_{cum} = \frac{\pi}{4} \left[3 \left(n_G - n_C\right) + \left(n_A - n_T\right)\right], \tag{2}$$

where n_X is number of nucleotide X in the sequence, from the first to the current location.

The representation of the DNA sequence by cumulated phase keeps the positional information, which enables the mutual comparison of two sequences. Also it maintains the chemical and structural information about the original sequence [7, 8]. The main reason for choosing cumulated phase signals is their large scale feature [9]. The shape of prokaryotic whole genome cumulated phase signal is typical for each organism. Moreover, signals of related organisms are more alike than the signals of evolutionary farther ones. Although in a negligible number of cases, two different strains of genomes of the same genus can be more dissimilar due to horizontal transfer of genetic information, which is common in prokaryotes [10]. The downsampled cumulated phase signals of seven organisms are shown in Figure 1(b).

The shape of a signal is mainly formed by the ratio of purines and pyrimidines, especially by those with strong bonds. Almost linear purines-rich subsequences alternating linear pyrimidines-rich subsequences along the DNA are evident. Signals usually end with the phase close to zero because of the second Chargaff's rule [11]. Due to these features, signals are suitable for massive downsampling.

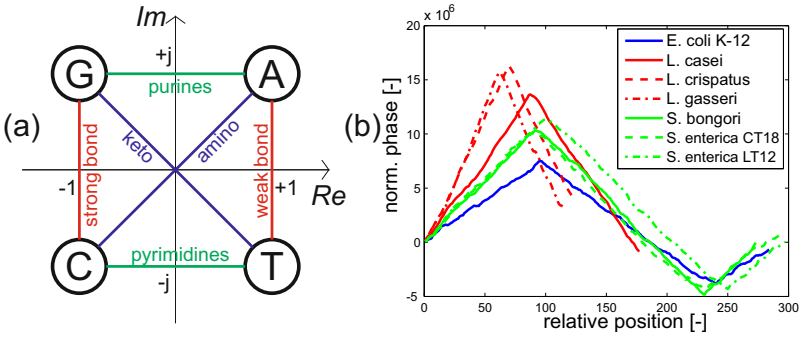

Fig. 1. (a) Complex representation of nucleotides, (b) Downsampled cumulated phase of DNA sequences of seven different organisms

2.2 Analysis of Signals

Genomic signals are discrete signals with progression along the DNA sequence; thus, they can be processed using any discrete transformation [12]:

$$\langle f(n), \psi(n) \rangle = \sum_{-\infty}^{+\infty} f(n)\psi(n), \tag{3}$$

where $f(n)$ represents sequence of signal samples and $\psi(n)$ belongs to the basis functions that determine the type of transformation.

Unlike other biological signals e.g. ECG, where sampling rate f_s is given by resolution of the sensing device, the sampling frequency of genomic signal is equal to the length of the DNA sequence. This makes it many times higher than f_s of other biological signals and massive downsampling is needed. Spectral analysis provided by discrete Fourier transform (DFT) can show possibilities of downsampling. To be able to perform DFT, the signal has to be periodic. The cumulated phase is defined at interval $\langle 1, N \rangle$, where N is number of nucleotides in the sequence, which could be taken as one period of signal on $(-\infty, +\infty)$. Consequently, the frequency axis can be divided into N equal units $\Omega = 2\pi/NT$ and DFT can assign to signal $f(n)$ new coefficients of discrete spectrum series $F(k)$ in the frequency domain, having the same length:

$$DFT\{f(n)\} = F(k) = \sum_{n=1}^{N} f(n)\ e^{-jk\Omega nT}. \tag{4}$$

The spectrum of *Escherichia Coli* in the Figure 2 had to be zoomed in due to the fact that the peripheral spectral lines are more than 10^{10} high making other lines unable to be observed. In the zoomed spectrum, only up to 10^5 other lines can be observed. These higher frequency components show changes to adjacent nucleotides and form only a noisy background of the genome. This information

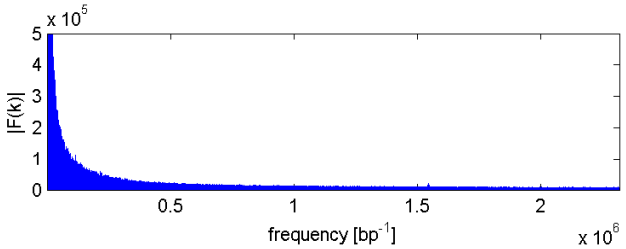

Fig. 2. Limited Fourier spectrum for *E. Coli*

is redundant, for comparison useless because these components are very similar to all genomes. On the contrary, low frequencies carry information about large scale features of signal, e.g. upward parts of signals formed by purines-rich subsequences or downward parts made of pyrimidines-rich subsequences. These components are species-specific. Thus, we are able to reduce a significant part of the spectrum by removing higher frequency components without compromising large scale information. Removing part of the spectrum allows us to downsample the signal, avoiding aliasing. Theoretically, simple lowpass filter could preprocess the signal for downsampling. Very long impulse response of the filter would be needed since the signal sampling frequency is equal to its length.

2.3 Signal Downsampling

We avoided the necessity of very long filter impulse response by using another transformation for discrete signals - discrete time wavelet transform ($DTWT$). For our purpose, the special case of wavelet transform - dyadic $DTWT$, was employed. This technique is characterized by utilizing parameters that are power of two. Using the relation between correlation and convolution, we can define dyadic wavelet transform for genomic signal as discrete convolution:

$$y_m(n) = \sum_{i=-\infty}^{+\infty} x(i)h_m(2^m n - i) = \sum_{i=-\infty}^{+\infty} h_m(i)x(2^m n - i), \tag{5}$$

as a signal decomposition by bank of discrete octave filters with impulse responses $h_m(n)$. Then the sampling frequency of signal $y_m(n)$ on output of m^{th} filter is 2^m times lower than the sampling rate f_s of the input signal $x(n)$.

There are two parameters that we had to set, the shape of the wavelet and the extent of the degree of decomposition. To reduce the organism comparison analysis time, we tested several simple wavelets. The best results were obtained using the basic Haar wavelet [13]. The shape of this wavelet is rectangular, thus computation of the transform is extremely fast. We found more complex wavelets as unsuitable because they can change the shape of the signal inappropriately

and the computation is more demanding. The Haar wavelet stands for two simple filters with impulse responses:

$$h_h(n) = \{-0.7071; 0.7071\} \tag{6}$$

$$h_d(n) = \{0.7071; 0.7071\}. \tag{7}$$

To determine the maximum possible downsampling factor, we used percentage root-mean-square difference (PRD) between the original signal and the downsampled signal by dyadic wavelet decomposition, that was resampled to the initial sampling rate:

$$PRD = \sqrt{\frac{\sum_{i=1}^{n}(x_0(i) - x_r(i))^2}{\sum_{i=1}^{n}(x_0(i) - \bar{x_0})^2}} \cdot 100\%, \tag{8}$$

where x_0 stands for original signal and x_r for resampled signal, both of length n.

PRD dependency of tested signals on the degree of decomposition with error bars along the curve is shown in Figure 3. Up to level 14 of the decomposition, the dependency shows a linear trend with reasonable values of percentage root-mean-square difference and its standard deviations. From level 15 of the decomposition, the dependency changes to a quadratic trend with high values of PRD and its standard deviations. As optimum, we selected decomposition degree of 14. Further analysis with higher level of decomposition leads to unsatisfactory results due to the loss of too much information. On the contrary, lower level of decomposition takes more computational time without any benefits.

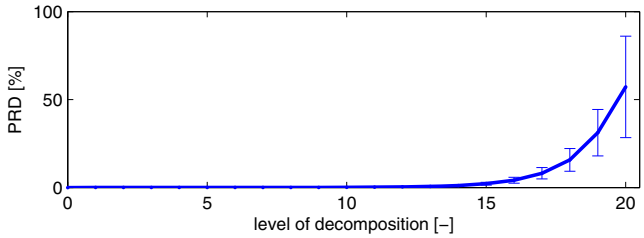

Fig. 3. Percentage root-mean-square difference as a function of degree of decomposition for 7 tested organisms

Of course the same degree of decomposition has to be used for all analyzed signals in order to maintain the ratio of lengths among the signals. Figure 1(b) shows the batch of our downsampled signals that were used for PRD analysis. The length of signals is only about 300 samples, unlike the original length of sequences in milions of bases.

2.4 Signal Alignment

Signals have to be aligned prior to conducting genome comparison. Since the lengths of various genomes can vary, multialignment of more signals would bring

about the incorporation of high number of gaps. Therefore, we decided to use pairwise alignment of every signal pair instead. This leads to maximum preservation of the genetic information for each comparison. We utilized dynamic time warping algorithm (DTW) [3, 14], which is similar to Needleman-Wunsch [15] or Smith-Waterman [16] algorithms for character sequence alignment. DTW is based on minimizing the distance between the pair of signals to be aligned.

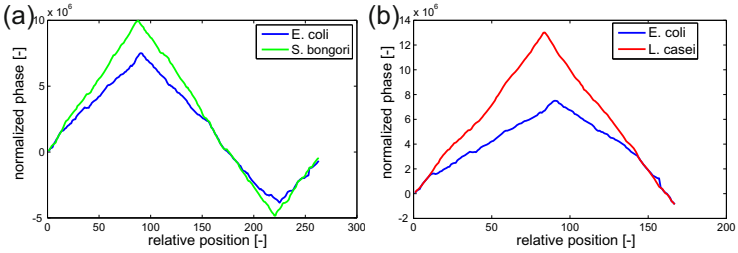

Fig. 4. (a) Alignment of *E. coli* and *S. bongori* signals of similar length, (b) Alignment of *E. coli* and *L. casei* signals of different lengths

When both signals are approximately of the same length, DTW is similar to global alignment as shown in Figure 4(a), where complete information from both signals were used. When the signals of different lengths are aligned, DTW is similar to local alignment where corresponding purines-rich and pyrimidines-rich subsequences are aligned and other parts of the longer signal are eliminated as shown in Figure 4(b). In both cases, maximum signal information is maintained.

2.5 Organism Comparison

The aligned pair of signals has the same length n. Their distance can be computed using the Euclidean metric:

$$d = \sqrt{\sum_{i=1}^{n} [x(i) - y(i)]^2}, \tag{9}$$

where $x(n)$ and $y(n)$ are aligned signals.

From the distances of each pair of signals we were able to construct the distance matrix and process it by cluster analysis. We used the unweighted pair group method with arithmetic mean ($UPGMA$) for achieving the best distinction of clusters [17]. The entire method was compared to the standard analysis based on multialignment of 16S rRNA sequences processed by the same clustering method. To prove that the parameters of the proposed method are correct and applicable in general, we added the rest of organisms from Table 1 to cluster analysis.

Table 2. Taxonomical classes of tested organisms

Class	Assigned number
Betaproteobacteria	1
Thermococci	2
Bacilli	3
Gammaproteobacteria	4

Selected organisms belong to four different taxonomical classes. Each class is assigned a number according to Table 2. These numbers are used for describing the organisms in cluster analysis shown in Figure 5.

Figure 5(a) representing the results of the proposed method. Four clusters correspond to real taxonomical classes. The results show that two different strains of

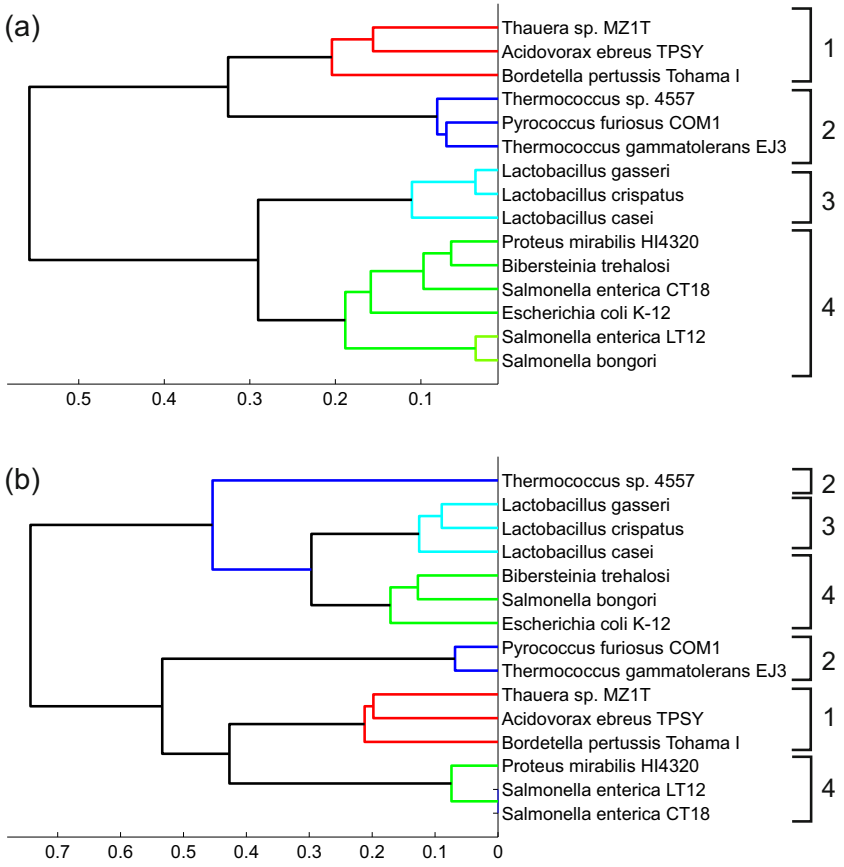

Fig. 5. (a) Cluster analysis based on whole genome signals, (b) Cluster analysis of 16S rRNA sequences

Salmonella enterica have quite diverse genomes, but they are still assigned to the *Gammaproteobacteria* cluster. The method is not dependent on the length of a genome but rather on information it contains. It assigns *Bibersteinia trehalosi* to the cluster of related organism despite its genome having half the length beside other *Gammaproteobacteria*. This is probably caused by two possible behavior of DTW, which can be similar to global or local alignment according to the specific situation. The result of the classical character processing method is shown in Figure 5(b). This approach cannot distinguish between the various strains of bacteria because 16S rRNA sequences are very commonly the same for different strains of the same genus. It splits *Gammaproteobacteria* into two very distant clusters. Also *Thermococci* cluster is split in an inappropriate way.

Whole genome comparison is significantly more robust. Redundant information that can be the source of bias was filtered out during signal downsampling. Only two signals are compared at a time, which leads to maximum utilization of the relevant information. Short 16S rRNA sequences should contain similar information across all genomes; however, multiple sequence alignment brings much imprecision to the analysis because of its high complexity.

3 Conclusion

In this paper, we present a novel method for classifying whole genome DNA sequences. Due to the use of sequences conversion to cumulated phase signals and massive downsampling, the method has low computational requirements in comparison to traditional methods. Thus, extremely long sequences with lengths of millions pair of bases, like whole genomes, can be processed. The method was tested on complete genome sequences records of prokaryotic organisms obtained from the GenBank database at NCBI.

Current bioinformatics does not provide an adequate technique for preprocessing whole genome signals; the proposed approach can be used as a new standard for this purpose. Although the genomic signal processing is a relatively new scientific field, it provides a high number of conversion techniques for obtaining genomic signal from character sequence. The cumulated phase signal representation was chosen for its specific properties suitable for downsampling. The low frequency band, formed by purines/pyrimidines ratio, carries the main information about the genome. The results of whole genome signals spectral analysis were used for designing appropriate downsampling technique, which allows downsampling of extremely large signals like whole prokaryotic genomes by more than ten thousand times. The dyadic wavelet transform was chosen for its ability to easily downsample signals with very high sampling rate. The level 14 of decomposition by DWT was set according to the percentage root-mean-square difference analysis of the selected signals; the percentage losses of original signals information do not exceed 1 percent.

Sequence multiple alignment, one of the most problematic issues in traditional DNA classification methods, was replaced by modification of dynamic time warping for genomic signals. The principal utilization of DTW does not

provide faster or less computationally demanding result, the calculation speed up is given only by the previous signal decimation step. However, the alignment of genomes in signal form using DTW offers another advantage; it is not necessary to choose specific parameters for sequence alignment like scoring matrix or gap penalization based on DNA biological properties. The genomic signal carries biological and chemical properties in a specific shape of signal, further biological adaptation of the alignment process is not necessary.

The results of the proposed method were compared to the traditional character processing technique based on multiple alignment of short parts of sequences represented by common phylogenetic marker 16S rRNA genes. Our approach reproduced the real taxonomical division with higher success than the traditional method and due to the independence on the genome length, it is now possible to conduct an extensive comparative analysis that would, otherwise, not be realizable by conventional techniques.

Acknowledgement. Supported by European Regional Development Fund - Project FNUSA-ICRC (No. CZ.1.05/1.1.00/02.0123) and by the grant project GACR P102/11/1068 NanoBioTECell.

References

1. Mayr, E., Bock, W.J.: Classifications and other ordering systems. Zool. Syst. Evol. Research 40, 169–194 (2002)
2. Cohen, A., Daubechies, I., Vial, P.: Wavelets on the Interval and Fast Wavelet Transforms. Applied and Computational Harmonic Analysis 1(1), 54–81 (1992)
3. Skutkova, H., Vitek, M., Babula, P., Kizek, R., Provaznik, I.: Classification of genomic signals using dynamic time warping. BMC Bioinformatics 14, S1 (2013)
4. Bittner, L., Halary, S., Payri, C., Cruaud, C., de Reviers, B., Lopez, P., Bapteste, E.: Some considerations for analyzing biodiversity using integrative metagenomics and gene networks. Biology Direct 5 (2010)
5. Chapple, D.G., Ritchie, P.A.: A Retrospective Approach to Testing the DNA Barcoding Method. PloS One 8(11) (2013)
6. Anastassiou, D.: Genomic Signal Processing. IEEE Signal Processing Magazine 18(4), 8–20 (2001)
7. Cristea, P.D.: Conversion of nucleotides sequences into genomic signals. Journal of Cellular and Molecular Medicine 6(2), 279–303 (2002)
8. Yau, S.S.T., Wang, J.S., Niknejad, A., Lu, C., Jin, N., Ho, Y.K.: DNA sequence representation without degeneracy. Nucleic Acids Research 31(12), 3078–3080 (2003)
9. Cristea, P.D.: Large scale features in DNA genomic signals. Signal Processing 83, 871–888 (2003)
10. Hao, W., Golding, G.B.: Patterns of Bacterial Gene Movement. Mol. Biol. Evol. 21(7), 1294–1307 (2004)
11. Sorimachi, K.: A Proposed Solution to the Historic Puzzle of Chargaff's Second Parity Rule. The Open Genomics Journal 2(1), 12–14 (2009)
12. Jan, J.: Digital signal filtering, analysis and restoration. Institution of Electrical Engineers (2000)
13. Daubechies, I.: Ten lectures on wavelets. CBMS-NSF conference series in applied mathematics. SIAM Ed (1992)

14. Berndt, D., Clifford, J.: Using dynamic time warping to find patterns in time series, New York, vol. 398, pp. 359–370 (1994)
15. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(3), 443–453 (1970)
16. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of Molecular Biology 147(1), 195–197 (1981)
17. Sokal, R., Michener, C.: A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin 38, 1409–1438 (1958)