Computers in Biology and Medicine 69 (2016) 308–314

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Set of rules for genomic signal downsampling

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

ARTICLE INFO

Article history:
Received 10 November 2014
Accepted 26 May 2015

Keywords:
Genomic signal
Cumulated phase
Downsampling
Compression
DWT
Sequence identification
Phylogeny

ABSTRACT

Comparison and classification of organisms based on molecular data is an important task of computational biology, since at least parts of DNA sequences for many organisms are available. Unfortunately, methods for comparison are computationally very demanding, suitable only for short sequences. In this paper, we focus on the redundancy of genetic information stored in DNA sequences. We proposed rules for downsampling of DNA signals of cumulated phase. According to the length of an original sequence, we are able to significantly reduce the amount of data with only slight loss of original information. Dyadic wavelet transform was chosen for fast downsampling with minimum influence on signal shape carrying the biological information. We proved the usability of such new short signals by measuring percentage deviation of pairs of original and downsampled signals while maintaining spectral power of signals. Minimal loss of biological information was proved by measuring the Robinson–Foulds distance between pairs of phylogenetic trees reconstructed from the original and downsampled signals. The preservation of inter-species and intra-species information makes these signals suitable for fast sequence identification as well as for more detailed phylogeny reconstruction.

1. Introduction

Current DNA, RNA and protein sequence comparison is done especially by character-processing techniques [1]. Multiple sequence alignments and the complexity of the character-based methods (e.g. Maximum Likelihood) limit their use to small dataset [2] only. They do not meet today's requirements for processing of large amount of data produced by increasingly cheaper sequencing technologies [3]. It is caused by character-based representation in which we are not able to extract important information for example for auxiliary alignment during multiple sequence alignment of very long sequences or to provide pilot alignment of reads during a de novo assembly process. A different way of how to treat the data is a bioinformatic subdiscipline called genomic signal processing [4]. Character-based representation (A, C, G, and T for nucleotides of DNA) can show only point differences between sequences and it is hard to be read by the human eye. The first signal representations of DNA sequences began to appear right with the onset of Sanger's sequencing to show sequence features in the larger scale than only point mutations and to visualize the information for a human eye [5]. Last decade showed that different genomic signals are usable not only for visualization but also for solving various bioinformatic tasks e.g. organism comparison, sequence alignment, gene prediction etc. [6,7] making genomic signal processing part of bioinformatics that is developing very rapidly. Just in the last few months, numerous new signal representations for DNA [8–11] as well as for protein sequences were described [12,13]. The latest results show that genomic signal processing is not only a full-fledged alternative for character based methods but it can even provide analyses that are unable to be done by character-processing techniques e.g. fast whole-genome comparison [14] or construction of guiding tree for multiple sequence alignment.

Genomic signal processing techniques for phylogeny reconstruction and sequence comparison can be divided into two groups. The first group of “alignment-free” methods differs substantially from the character based methods. Sequences are compared in pairs according to the difference in characteristic attributes e.g. frequencies of k-mers [15,16] or by using signal processing techniques e.g. Fourier transform (FT) for spectral analysis or dyadic wavelet transform (DWT) for revealing periodicities in DNA [17–19]. Although this approach allows fast long sequence comparison, it suffers from inability to evaluate local differences because the resulting similarity is always global. The second group of “alignment-dependent” methods can compare local differences. Algorithm for pairwise alignment using dynamic time warping (DTW) was described recently [20]. Although these techniques have the same complexity as the corresponding character-based methods, they have an advantage over them by

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

http://dx.doi.org/10.1016/j.compbiomed.2015.05.022

possibility of processing compressed data. Sampling rate of a character sequence is given by a number of the characters and it cannot be reduced, because dropping of character would change a nature of a sequence. Genomic signals can be significantly downsampled without a negative effect on the result of an alignment and a similarity measure, as proved in [14,20]. However, no rule for genomic signal downsampling was given. In this paper, we examined fast algorithm for signal decimation using dyadic wavelet transform (DWT). By measuring information loss according to the level of decimation for sequences of various lengths, we were able to set the rule for downsampling the signal depending on the length of the original sequence. The rule was also verified for loss of biological information using phylogenetic trees. We measured Robinson–Foulds distance of phylogenetic trees reconstructed from original data and downsampled signals. Thus, using computationally undemanding algorithm for downsampling, we are able to reduce the amount of data provided to signal alignment which is computationally demanding NP-hard (non-deterministic polynomial-time hard) problem [21].

2. Materials and methods

2.1. Test dataset

To examine the possibility of genomic signal downsampling we used a set of 420 sequences divided into 7 groups each of 60 sequences according to their lengths. The shortest sequences of cytochrome c oxidase I (COX1) genes of eukaryotes, commonly used as short barcode sequence for identification of organisms [22], were obtained from Boldsystems database (http://www.boldsystems.org/). Other sequences containing 16S rRNA and complete ACTA1 genes of eukaryotes, whole mitochondrial genomes of eukaryotes, whole genomes of viruses, whole bacterial plasmids and whole bacterial genomes were obtained from GenBank database at NCBI (http://www.ncbi.nlm.nih.gov/genbank/). First two groups contains only coding sequences, other groups contains both, coding and non-coding DNA. Unlike COX1 and 16S rRNA, eukaryotic ACTA1 genes consist of exons and introns. The summary of sequences is given in Table 1.

For the biological validation, we used another 2 smaller sets of whole mitochondrial genomes of eukaryotes and whole bacterial genomes obtained from GenBank database. Accession numbers of these sequences are mentioned in Fig. 5.

2.2. Genomic signal

Downsampling techniques are applicable to “alignment dependent” methods. However, not all of the signal representations are suitable for fast comparison equally. Various signals e.g. Z curve [23], DNA walks [24–26] or phase visualizations [27] differ in dimensionality, primary feature representation or species specificity. We used cumulated phase signal representation for several purposes [27], because its’ 1D nature makes it very suitable for alignment and easy to compute with. Also it preserves appropriate features in the large scale and after massive downsampling [14].

Sequence conversion is done by projection of nucleotides in the complex plane in the manner such that appropriate complex numbers maintain information on nucleotides' chemical similarities: A [1,j], C [−1,−j], G [−1,j], T [1,−j]. Using trigonometric functions, we are able to calculate the phase of these four numbers: $\{\phi_A, \phi_C, \phi_G, \phi_T\}=\{\pi/4,-3\pi/4,3\pi/4,-\pi/4\}$. The signal form of a sequence is done by cumulating of nucleotides' phase numbers along the sequence [28]. In case of RNA sequences, we treat U as T.

This 1D signal is similar to other biological 1D signals e.g. ECG, EEG signals, and can be processed by similar tools, e.g. FT, DWT of DTW. However, several differences can be found. Sampling rate f_s of cumulated phase is given by the length of the sequence not by a sensing device. Spectral analysis provided by discrete Fourier transform (DFT) can show possibilities of downsampling by revealing frequency bands carrying the main information. 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 $(-\infty,+\infty)$. Discrete spectrum $F(k)$ in the frequency domain has the same length as the signal:

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

2.3. Signal analysis

Sampling rate equal to the sequence length makes direct downsampling problematic for longer sequences, because it increases demands on the antialiasing filter, mainly in terms of length of impulse response. We proposed fast and simple solution based on dyadic wavelet transform (DWT) [14]. However DWT has been previously used for revealing periodicities in DNA [29], it has not been used for DNA signal downsampling to extract large scale feature of cumulated phase signal. Using the relation between correlation and convolution, we can define dyadic wavelet transform for genomic signal as discrete convolution:

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

which represents signal decomposition by a bank of discrete octave filters with impulse responses $h_m(n)$ [30]. 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 $x(n)$. Using the Haar wavelets standing for 2 filters, with short impulse responses $h_h(n)=\{-0.7071;\ 0.7071\}$ and $h_d(n)=\{0.7071;\ 0.7071\}$, makes downsampling very fast. Such a short impulse response also minimalizes delays that may affect the signal shape in an inappropriate way, e.g. rounding peaks, original shape deformation.

The preservation of signal information was measured as percentage root-mean-square difference (PRD) between the original and the downsampled signal that was again resampled to the initial sampling rate:

$$\mathrm{PRD} = \sqrt{\frac{\sum_{i=1}^{n}(x_0(i)-x_r(i))^2}{\sum_{i=1}^{n}(x_0(i)-\overline{x_0})^2}}\cdot 100\%, \tag{3}$$

Table 1
The specification of test sequences.

	1	2	3	4	5	6	7
Sequence	COX1	16S rRNA	ACTA1	Whole mtDNA	Whole genome	Whole plasmid	Whole genome
Taxa	Eukaryotes	Eukaryotes	Eukaryotes	Eukaryotes	Virus	Bacteria	Bacteria
Average length [bp]	652	1 441	2859	16,335	28,962	383,646	3,830,130
Standard deviation [bp]	2	300	1064	981	1620	141,706	1,708,995

where x_0 stands for original signal and x_r for signal that was resampled to the original sampling rate by inverse DWT with lost bands replaced by zero vectors, both of length n. This value better represents the biological information that signals carry than measuring loss in the power spectrum as proposed in [20], because it takes into account both, nucleotide changes as well as large scale feature of a signal. Moreover, the overall spectral energy can be preserved using normalizing constant $2^{-m/2}$ for the output of m^{th} filter in DWT.

3. Results

3.1. Spectral analysis

From the definition of the cumulated phase, the signal always begins by zero phase. In combination with its specific large scale feature, the mean value of the signal is non-zero. This means that Fourier spectrum (1) of the signal contains direct component (DC). However this component depends on the length and the size of the signal, it does not carry any information. DC is always affected by multiplicative effect caused by downsampling with DWT and by summing effect caused by signal alignment with dynamic time warping. This component can be eliminated by setting the mean value of the signal to zero, as shown in Fig. 1a and b. By eliminating DC, we are able to measure PRD depending on the level of decomposition of DWT while maintaining the power spectrum of the signal. Long genomic signals are suitable for downsampling because cumulated phase has a tendency to produce slow trend, thus it cumulates genetic information at a low frequency. The main information for short signals is also carried by low frequencies, as shown in Fig. 1c–e. On the other hand, their low frequency band is longer relative to the entire spectrum in comparison with long signals. Thus, the short signals are also suitable for downsampling, but with lower decimation factor.

3.2. Signal downsampling

Complexity of dyadic wavelet transform (2) due to the length of the sequence n is linear ($O(nm)$) making the algorithm fast. Maximum level of decomposition depends on the length of the original signal. When value of 2^m, where m is level of decomposition, reaches the length of the original signal, the downsampled signal is represented by only one sample. PRD dependency of tested signals on the degree of decomposition with error bars along the curves is shown in Fig. 2. The dependence of both PRD mean value as well as its standard deviation seems to be quadratic. For every group Pearson's correlation coefficient r between test data and predicted values by corresponding quadratic function satisfies the condition $r > 0.9$. For all levels of decomposition, at least 99,9% of spectral energy of the original signal was preserved. The preservation of the reasonable amount of information depends on the length of the original sequence. For every group of the test sequences, the threshold can be found.

3.3. Setting the rule

Since the PRD dependency shows quadratic character, we set maximum level of decomposition for the signals in manner to prevent mean value of PRD exceeds 1%. With every another decomposition the loss of information increases greatly. Maximum levels of decomposition for average PRD value below 1% are marked with a blue arrow in Fig. 2. Allowed levels of decomposition and average length of downsampled signals are shown in Table 2. For COX1 and 16S rRNA sequences, no downsampling is possible to maintain PRD below 1%.

For setting degree of decomposition from Table 2, PRD value for different signals can vary greatly. However most PRD values are located below 1%, extremes can be found as shown in Fig. 3. For the first two groups, no PRD statistics can be given, since no downsampling was possible. In other groups, maximum PRD values exceed the value of 1% since signals can have various lengths. On the other hand, also in these groups majority of PRD values lay below 1%.

A representative for each group of test dataset was taken to show comparison of original signal with its downsampled version. For each group, level of decomposition set in Table 2 was used. Because both, the length and the amplitude of the original and downsampled signals differ, the values were normalized to show the overlap of the signals (Fig. 4).

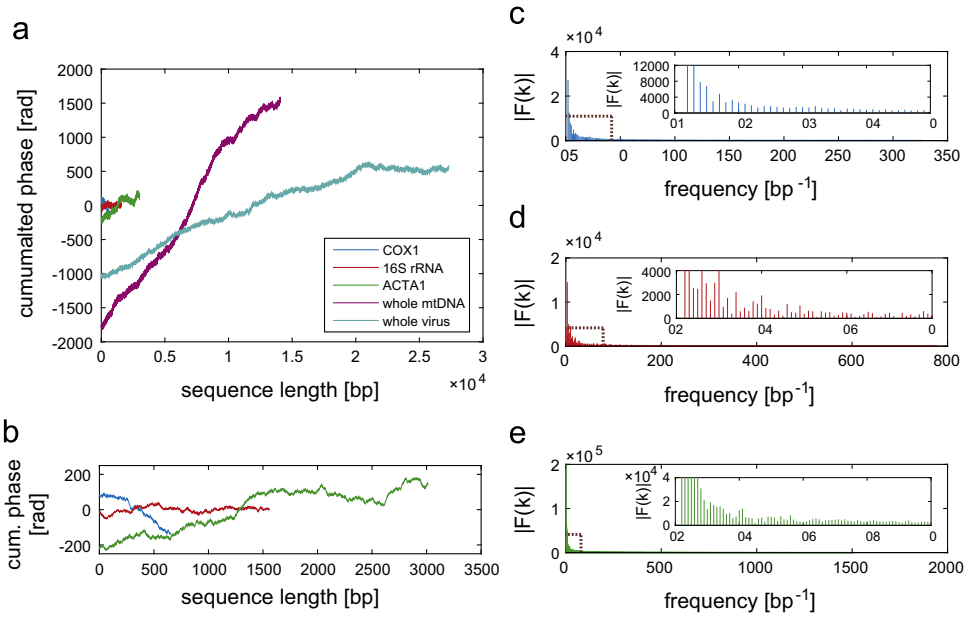

Fig. 1. Cumulated phase signals (a) without direct component picked from the dataset (b) detail of 3 shortest signals, Fourier spectrum up to $f_s/2$ and zoomed part of spectrum for (c) COX1 (d) 16S rRNA and (e) ACTA1 signals.

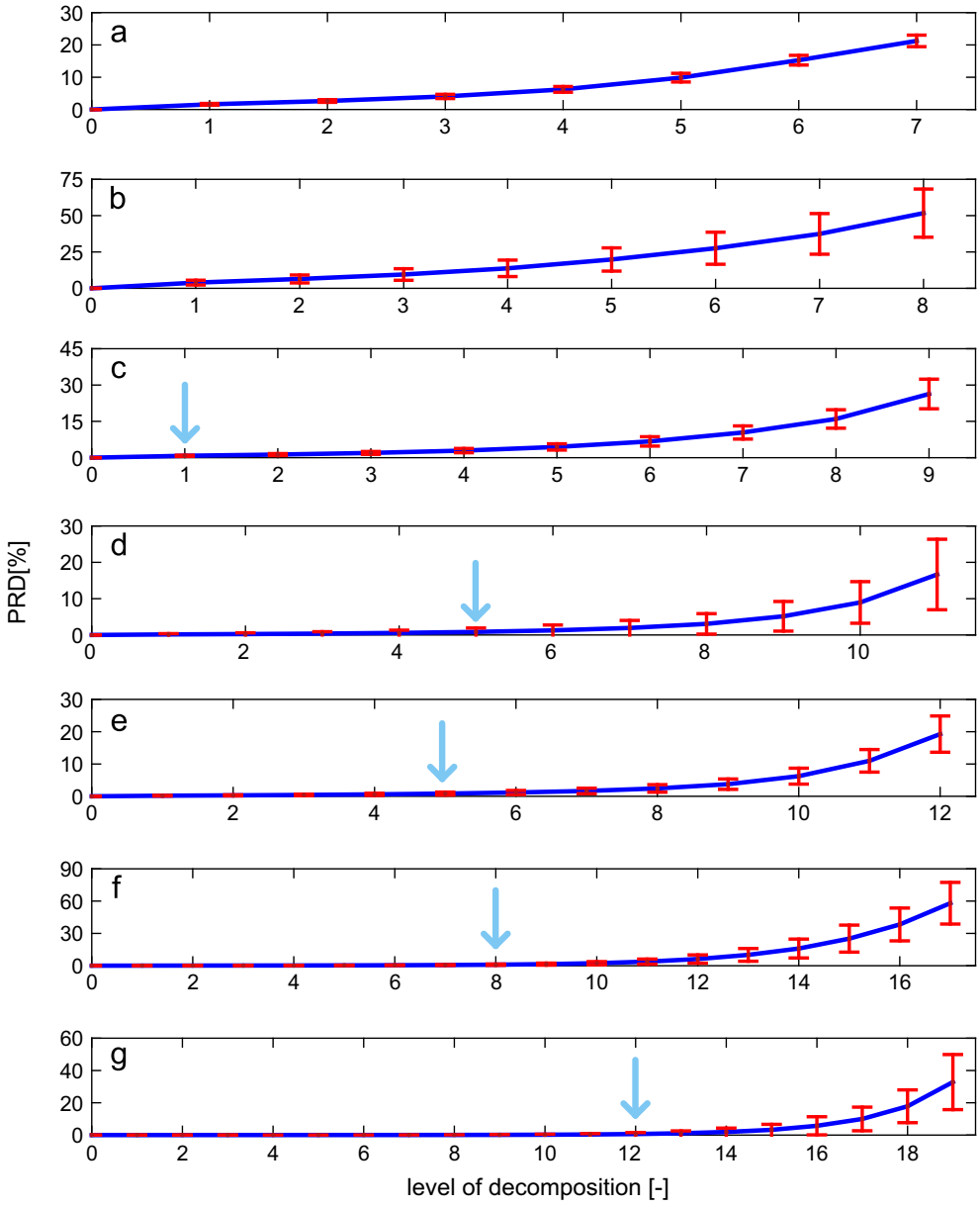

Fig. 2. Percentage root-mean-square differences (blue) with their standard deviations (red) as a function of degree of decomposition for (a) COX1 (b) 16S rRNA (c) ACTA1 (d) whole mtDNA (e) whole virus genome (f) whole plasmid genome (g) whole bacterial genome datasets. Blue arrows show maximum levels of decomposition for average PRD value below 1%. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Downsampling threshold set for individual groups of test sequences and average lengths of sequences after downsampling.

	1	2	3	4	5	6	7
Sequence	COX1	16S rRNA	ACTA1	Whole mtDNA	Whole viral genome	Whole bacterial plasmid	Whole bacterial genome
Level of decomposition	0	0	1	5	5	8	12
Average length after DWT [–]	652	1441	715	256	453	750	935

For biological validation of proposed loss of information, we used 2 smaller sets of previously unused sequences for phylogenetic analysis based on downsampled signals. The results of phylogeny for downsampled signals processed by method proposed in [14] are represented as cladograms shown in Fig. 5. Cladograms for group of mitochondrial sequences (c) are the same as cladogram provided by the ClustalW character based method [31] until 5th level of decomposition by DWT according to Robinson–Foulds distance [32] for comparing of phylogenetic trees, while cladograms of signals with PRD value above 1% show changes in one or more nodes. Sequences of whole bacterial genomes are too long to be processed by character-based methods due to its time and space complexity. Downsampled signal can be processed down to level 8 of decomposition. From level 8 to 13 of decomposition cladograms (d) are the same showing individual clusters of Bacilli (black), Betaproteobacteria (red), Gammaproteobacteria (green) and Thermococci (blue). With every other level of decomposition, thus PRD value higher than 1%, cladograms start to

differ in one or more nodes. For set level of decomposition according to PRD value, computational time (Intel Core i5, 3.2 GHz) is several times lower than for original signals as shown in Fig. 5a and b.

We found the relationship between the length of the original signal and decimation factor, while maintaining the 1% PRD, to be linear (see Fig. 6a) given by the equation:

$$decf = 0.0011 \cdot n, \tag{4}$$

where *decf* is decimation factor and n is the length of the original sequence. A linearity can be confirmed by sufficiently high Pearson's correlation coefficient ($r > 0.99$) between test data and predicted values. For our data and predicted function, we get $r = 0.9992$. Using DWT transform for decimation, simple $\log_2$ function can be used to get level of decomposition of DWT corresponding to the decimation factor. Decimation factor function and its deviations from the test dataset is shown in Fig. 6.

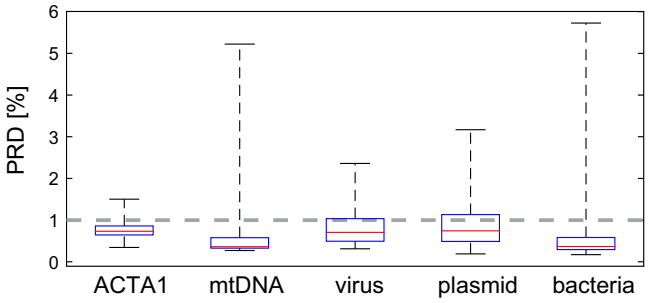

Fig. 3. Boxplot of percentage root-mean-square differences for degrees of decomposition set for each group of sequences according to Table 2.

Lengths of downsampled signals $n_{downsampled}$ can be computed as:

$$n_{downsampled} = \frac{n}{decf} = \frac{n}{0.0011n} = \frac{1}{0.0011} \cong 909, \tag{5}$$

where n is the length of the original signal and *decf* is the decimation factor. The equation shows that average length of downsampled signals depends only on selected PRD value. Thus, average length of downsampled signals, while maintaining the 1% PRD, is always 909 samples.

Due to the quadratic PRD dependency on the decimation factor, we consider 1% PRD as suitable threshold. We also validated this value with following phylogenetic analysis. On the other hand, accuracy requirements may vary among different applications. Examples of suitable decimation factor according to the sequence length and PRD value are shown in Table 3. For usability of DWT, values are rounded to the nearest integer powers of 2, so they represent level of decomposition by DWT. Cases for which downsampling is not available are marked as N/A.

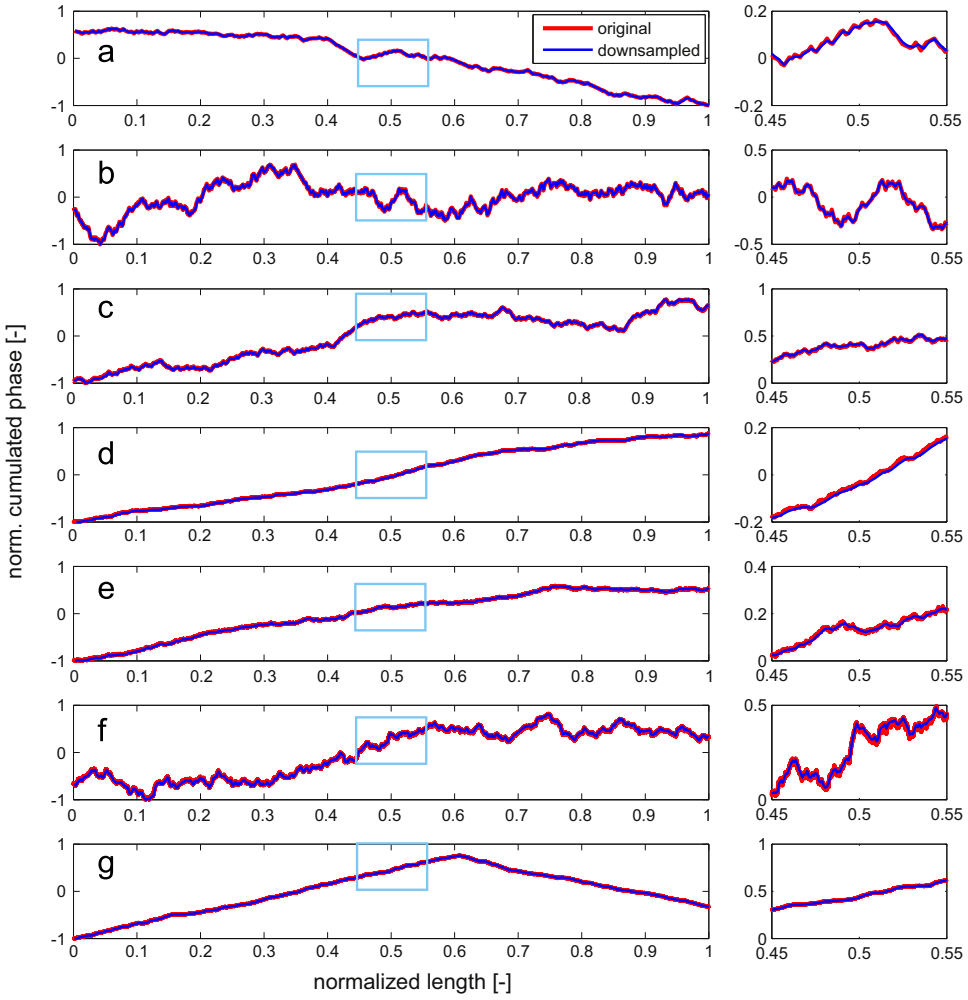

Fig. 4. Pairs of original and downsampled signals and their 10 × zoom randomly chosen from (a) COX1 (b) 16S rRNA (c) ACTA1 (d) whole mtDNA (e) whole virus genome (f) whole plasmid genome (g) whole bacterial genome datasets.

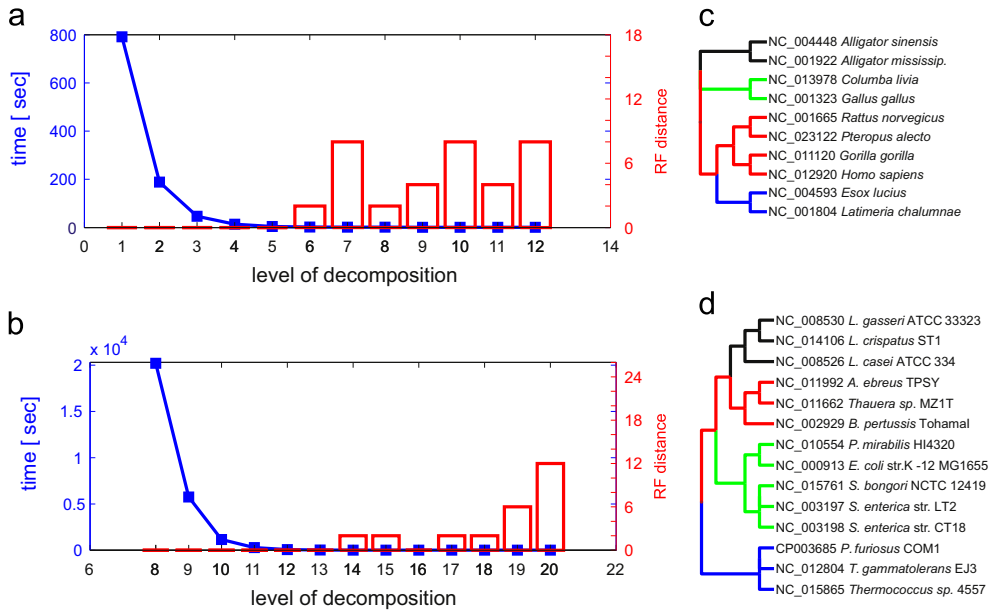

Fig. 5. Computational time and Robinson–Foulds distance for set of (a) whole mitochondrial genomes of eukaryotes and (b) whole bacterial genomes. Phylogenetic trees of downsampled signals for (c) whole mitochondrial genomes of eukaryotes and (d) whole bacterial genomes. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

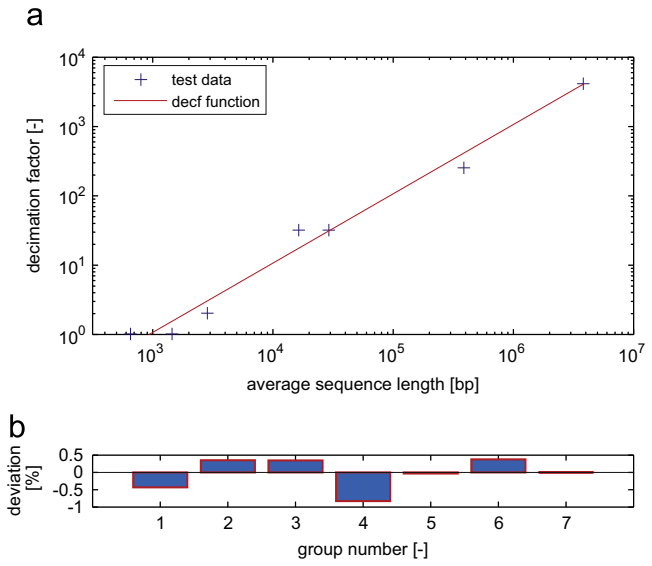

Fig. 6. Decimation factor and predicted decimation factor function (a) for the test data and (b) percentage deviation of decimation factor function from the values of test data.

Table 3
Decimation factors for sequences of various lengths set according to the proposed rule and required percentage deviation.

Sequence length [bp]	PRD [%]			
	0.01	0.1	1	10
100	N/A	N/A	N/A	2
1,000	N/A	N/A	N/A	5
10,000	N/A	N/A	3	8
100,000	N/A	3	7	12
1,000,000	1	6	10	15

4. Conclusion

Genomic signal processing is a bioinformatics sub-discipline that is undergoing rapid development at the moment. Lots of new genomic signal representations were described recently, as well as techniques for their processing. The greatest advantage of the genomic signal alignment-dependent methods over the standard character based methods is possibility of processing data treated by lossy compression. In the paper, we examined the redundancy of the genetic information carried by genomic signal. By processing large dataset of sequences of different lengths obtained from Boldsystems and GenBank databases, we were able to set the rule for maximum possible genomic signal downsampling ratio according to the length of an original DNA sequence. To validate the rule, we provided phylogenetic analysis on another sets of eukaryotic and prokaryotic sequences that were not used for setting the rule.

The bioinformatic comparison of a set of sequences is based on multiple sequence alignment which is NP-complete problem with exponential complexity $O(n^s)$, where n is the length of the alignment and s is the number of sequences in alignment. Because no polynomial solution can be found, the only way how to reduce computational time for the same set of sequences is to reduce the length of sequences. Using very fast DWT algorithm which complexity is $O(nm)$, where n is the length of the sequence and m is DWT degree of decomposition, reduce computational operations for alignment 2^{ms} times. With this reduction of genomic signals, it is now possible to conduct an extensive comparative analysis that would not be realizable by conventional techniques, e.g. for test group of bacterial genomes, the computational time is reduced approximately 10^{216} times. For analysis done using conventional techniques, the reduction is also significant. For test group of ACTA1 genes, the computational time is reduced approximately 10^{18} times. This reduction of computational demands is possible due to the large redundancy of genomic sequences, hence signals. However fast algorithm for genomic signals decimation was proposed recently, no general rule for downsampling of sequences of various lengths was given.

Here, we examined redundancy of genetic information stored in cumulated phase signal representation based on large dataset of real sequences of various lengths. Our results show, that the main information of the signals is carried by low frequency bands independently on sequence length and sequence nature. Thus, cumulated phase signals are suitable for downsampling in general. By measuring percentage change of downsampled signals across different domains of life (Archea, Bacteria, Virus, Eukaryotes), we set the rule for genomic signal downsampling ratio according to the length of an original DNA sequence. The rule was also validated using phylogenetic analysis of eukaryotic and prokaryotic sequences by measuring Robinson–Foulds distance between original and downsampled signals. Moreover, for given PRD value, the length of a downsampled signal is independent on the length of an original sequence. Due to the quadratic dependency of PRD and exponential dependency on computational time on the decimation factor, we consider 1% PRD value as suitable threshold. For such a level of maintaining information, any DNA sequence can be represented by downsampled cumulated phase signal with average length of 909 samples.

Conflict of interest statement

None declared.

Acknowledgments

This study is 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. Computational resources were provided by the MetaCentrum under the program LM2010005 and the CERIT-SC under the program Centre CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, Reg. no. CZ.1.05/3.2.00/08.0144.

References

[1] E. Mayr, W.J. Bock, Classifications and other ordering systems, J. Zool. Syst. Evol. Res. 40 (4) (2002) 169–194. http://dx.doi.org/10.1046/j.1439-0469.2002.00211.x.
[2] B. Chor, T. Tuller, Finding a maximum likelihood tree is hard, J. ACM 53 (5) (2006) 722–744. http://dx.doi.org/10.1145/1183907.1183909.
[3] W.W. Soon, M. Hariharan, M.P. Snyder, High-throughput sequencing for biology and medicine, Mol. Syst. Biol. 9 (1) (2013) 1–14. http://dx.doi.org/10.1038/msb.2012.61.
[4] D. Anastassiou, Genomic signal processing, IEEE Signal Process. Mag. 18 (4) (2001) 8–20. http://dx.doi.org/10.1109/79.939833.
[5] E. Hamori, J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, J. Biol. Chem. 258 (2) (1983) 1318–1327.
[6] E. Dougherty, Genomic signal processing [Life Sciences], IEEE Signal Process. Mag. 29 (3) (2012) 124–129. http://dx.doi.org/10.1109/MSP.2012.2185868.
[7] S.Y. Kung, Y. Luo, M-W. Mak, Feature selection for genomic signal processing: unsupervised, supervised, and self-supervised scenarios, J. Signal Process. Syst. 61 (1) (2010) 3–20. http://dx.doi.org/10.1007/s11265-008-0273-8.
[8] W. Hou, Q. Pan, M. He, A novel representation of DNA sequence based on CMI coding, Phys. A: Stat. Mech. Appl. 409 (2014) 87–96. http://dx.doi.org/10.1016/j.physa.2014.04.030.
[9] B-S. Jeong, A.T.M. Golam Bari, M. Rokeya Reaz, S. Jeon, C.-G. Lim, H-J Choi, Codon-based encoding for DNA sequence analysis, Methods 67 (3) (2014) 373–379. http://dx.doi.org/10.1016/j.ymeth.2014.01.016.
[10] X. Ding, C-C. Cao, X. Sun, Intrinsic correlation of oligonucleotides: a novel genomic signature for metagenome analysis, J. Theor. Biol. 353 (2014) 9–18. http://dx.doi.org/10.1016/j.jtbi.2014.02.039.
[11] W. Hou, Q. Pan, M. He, A novel 2D representation of genome sequence and its application, J. Comput. Theor. Nanosci. 11 (8) (2014) 1745–1749. http://dx.doi.org/10.1166/jctn.2014.3561.
[12] Y. Yao, S. Yan, J. Han, Q. Dai, P. He, A novel descriptor of protein sequences and its application, J. Theor. Biol. 347 (2014) 109–117. http://dx.doi.org/10.1016/j.jtbi.2014.01.001.
[13] T. Ma, Y. Liu, Q. Dai, Y. Yao, P. He, A graphical representation of protein based on a novel iterated function system, Phys. A: Stat. Mech. Appl. 403 (2014) 21–28. http://dx.doi.org/10.1016/j.physa.2014.01.067.
[14] K. Sedlar, H. Skutkova, M. Vitek, I. Provaznik, Prokaryotic DNA signal downsampling for fast whole genome comparison, Information Technologies in Biomedicine, Vol. 3, 2014373. http://dx.doi.org/10.1007/978-3-319-06593-9_33, Advances in Intelligent Systems and Computing.
[15] HJ. Yu, Segmented K-mer and its application on similarity analysis of mitochondrial genome sequences, Gene 518 (2) (2013) 419–424. http://dx.doi.org/10.1016/j.gene.2012.12.079.
[16] P. Kolekar, M. Kale, U. Kulkarni-Kale, Alignment-free distance measure based on return time distribution for sequence analysis: applications to clustering, molecular phylogeny and subtyping, Mol. Phylogenet. Evol. 65 (2) (2012) 510–522. http://dx.doi.org/10.1016/j.ympev.2012.07.003.
[17] C. Yin, Y. Chen, S. S-T. Yau, A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering, J. Theor. Biol. 359 (2014) 18–28. http://dx.doi.org/10.1016/j.jtbi.2014.05.043.
[18] V. Kubicova, I. Provaznik, Relationship of bacteria using comparison of whole genome sequences in frequency domain, Information Technologies in Biomedicine, Vol. 3, 2014397. http://dx.doi.org/10.1007/978-3-319-06593-9_35, Advances in Intelligent Systems and Computing.
[19] L. Pinello, G. Lo Bosco, G.-C. Yuan, Applications of alignment-free methods in epigenomics, Brief Bioinform. 15 (3) (2014) 419–430. http://dx.doi.org/10.1093/bib/bbt078.
[20] H. Skutkova, M. Vitek, P. Babula, R. Kizek, I. Provaznik, Classification of genomic signals using dynamic time warping, BMC Bioinform. 14 (Suppl 10) (2013) S1. http://dx.doi.org/10.1186/1471-2105-14-S10-S1.
[21] I. ELIAS, Settling the intractability of multiple alignment, J. Comput. Biol. 13 (7) (2006) 1323–1339. http://dx.doi.org/10.1089/cmb.2006.13.1323.
[22] V. Savolainen, R.S. Cowan, A.P. Vogler, G.K. Roderick, R. Lane, Towards writing the encyclopaedia of life: an introduction to DNA barcoding, Philos. Trans. R. Soc. B: Biol. Sci. 360 (1462) (2005) 1805–1811. http://dx.doi.org/10.1098/rstb.2005.1730.
[23] C.-T. Zhang, R. Zhang, H.-Y. Ou, The Z curve database: a graphic representation of genome sequences, Bioinformatics 19 (5) (2003) 593–599. http://dx.doi.org/10.1093/bioinformatics/btg041.
[24] M.A. Gates, Simpler DNA sequence representations, Nature 316 (6025) (1985) 219. http://dx.doi.org/10.1038/316219a0.
[25] S. S-T. Yau, DNA sequence representation without degeneracy, Nucleic Acids Res. 31 (12) (2003) 3078–3080. http://dx.doi.org/10.1093/nar/gkg432.
[26] J.A. Berger, S.K. Mitra, M. Carli, A. Neri, Visualization and analysis of DNA sequences using DNA walks, J. Frankl. Inst. 341 (1–2) (2004) 37–53. http://dx.doi.org/10.1016/j.jfranklin.2003.12.002.
[27] P.D. Cristea, Conversion of nucleotides sequences into genomic signals, J. Cell. Mol. Med. 6 (2) (2002) 279–303. http://dx.doi.org/10.1111/j.1582-4934.2002.tb00196.x.
[28] P.D. Cristea, Large scale features in DNA genomic signals, Signal Process. 83 (4) (2003) 871–888. http://dx.doi.org/10.1016/S0165-1684(02)00477-2.
[29] F. Cui, M.V. Sirotin, V.B. Zhurkin, Impact of Alu repeats on the evolution of human p53 binding sites, Biol. Direct 6 (1) (2011) 2. http://dx.doi.org/10.1186/1745-6150-6-2.
[30] J. Jan, Digital Signal Filtering, Analysis and Restoration, The Institution of Electrical Engineers, London, ISBN 08-529-6760-8407s.
[31] K.-B. LI, ClustalW-MPI: ClustalW analysis using distributed and parallel computing, Bioinformatics 19 (12) (2003) 1585–1586. http://dx.doi.org/10.1093/bioinformatics/btg192.
[32] D.F. Robinson, L.R. Foulds, Comparison of phylogenetic trees, Math. Biosci. 53 (1–2) (1981) 131–147. http://dx.doi.org/10.1016/0025-5564(81)90043-2.