Gene 435 (2009) 104–118

Contents lists available at ScienceDirect

Gene

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

Methods paper

Multivariate autoregressive model for a study of phylogenetic diversity

K.J. Blinowska *, B. Trzaskowski, M. Kaminski, R. Kus

Department of Biomedical Physics, Warsaw University, ul. Hoza 69, 00-681 Warsaw, Poland

ARTICLE INFO

Article history:
Received 24 September 2008
Received in revised form 17 December 2008
Accepted 5 January 2009
Available online 10 February 2009

Received by I. King Jordan

Keywords:
Multivariate autoregressive model
DNA sequences
Clustering trees
Phylogenies
Globin family
Interspecies distance

ABSTRACT

We present a computationally effective model to parameterize DNA sequences in a way describing comprehensively its auto and cross-correlation structure. The approach is based on four-channel Multivariate Autoregressive Model (MVAR). The model was applied to a study of genes from the globin family for 6 vertebrate species. First, the sequences were coded as four signals (corresponding to the nucleotides), which were fitted to a four-channel MVAR. From the correlation matrices the vectors of model coefficients were calculated as functions of the nucleotide distance. The between-chromosomes and inter-species differences were best distinguished in the cross-coefficients binding different nucleotide sequences. For clustering purposes different metrics were tested and then two clustering procedures (Nearest Neighbor and UPGMA) were applied. The clustering trees and consensus trees were constructed for exons, introns and whole genes. The results were in agreement with the known dependencies between the chromosomes of the globin family. The orthological genes for different species were grouped together. Inside these groups the phylogenetically close organisms were localized in proximity.

1. Introduction

The vast amount of genetic characterizations of various organisms, including the fundamental, huge database collected in the framework of the Human Genome Project, require application of new mathematical methods and tools in order to extract biologically significant features. This concerns in particular retrieval of phylogenetic information contained in the sequenced genomes.

A broad repertoire of statistical methods have been applied to DNA series including estimation of power spectra e.g. (Fukushima et al., 2002), Mutual Information function e.g. (Holste and Grosse 2003), wavelet transform e.g. (Arneodo et al., 1995; Audit et al., 2001), and fractal analysis e.g. (Voss 1992; Stanley et al., 1999). Short-range and long-range correlations were found in DNA sequences (Trifonov 1998). The short range 3 base pairs (bp) periodicities were connected with properties of coding sequences (Fukushima et al., 2002) and they were found to be species independent (Grosse et al., 2000).

More recently, one-channel Autoregressive Model (AR) (Chakravarthy et al., 2004) and Discrete Autoregressive Model (DAR) (Dehnert et al., 2003) was introduced to DNA analysis. Consecutively, DAR was also used to estimate the correlation patterns of chromosomes (Dehnert et al., 2005a) and the construction of the clustering tree for eukaryotic species. In phylogenetic information retrieval, statistical methods such as the study of correlations (Dehnert et al., 2005b) and dinucleotide relative abundances (Karlin and Mrazek, 1997) were used.

The first important step in statistical analysis of DNA is the codification of the sequences in a way amenable to digital signal processing and without loss of information. Different methods of transcodification have been used. Voss (1992) proposed converting DNA sequence into four binary sequences corresponding to four nucleotides. In the position where the given nucleotide occurred a value of one was inserted, other positions were filled with zeroes. Other methods which become popular are DNA walks, which are the cumulative sum of binary signals obtained with rules such as purine versus pyrimidine or strong versus weak bonds. Different rules for codification may be found in Bernaola-Galvan et al. (2002). Numerical mapping of nucleotides involved also representation of nucleotides in the complex space (Anastassiou, 2002) and derived from this approach, real number representation (Chakravarthy et al., 2004). However, this kind of representation is to a large degree arbitrary. The method of K-strings (Yu and Anh, 2004) involved assigning to the nucleotides numbers from 0–3 range.

In this paper we propose a method of codification of DNA sequences and their parameterization in a way that allows for efficient construction of phenograms. The sequence of each kind of four nucleotides is transformed into a continuous sampled signal and then all four signals are fitted into a multichannel autoregressive model (MVAR). We demonstrate that our model correctly reproduces known properties of DNA series such as the presence of 1/3 frequency, and the difference in statistical properties of exons and introns. Then we use measures derived from the model

* Corresponding author. Tel.: +48 22 55 32 144; fax: +48 22 55 32 320.
E-mail address: kjbli@fuw.edu.pl (K.J. Blinowska).

0378-1119/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2009.01.009

```
  A T G G C A C T T G T C A A A T A C A G C A C A G T T T T T
A 100000000010000000000000101010001000100000100010000000000000
C 000000001000100000000010000000000010000010001000000000000000
G 000010100000000000100000000000000000001000000000100000000000
T 001000000000001010001000000000100000000000000000001010101010
```

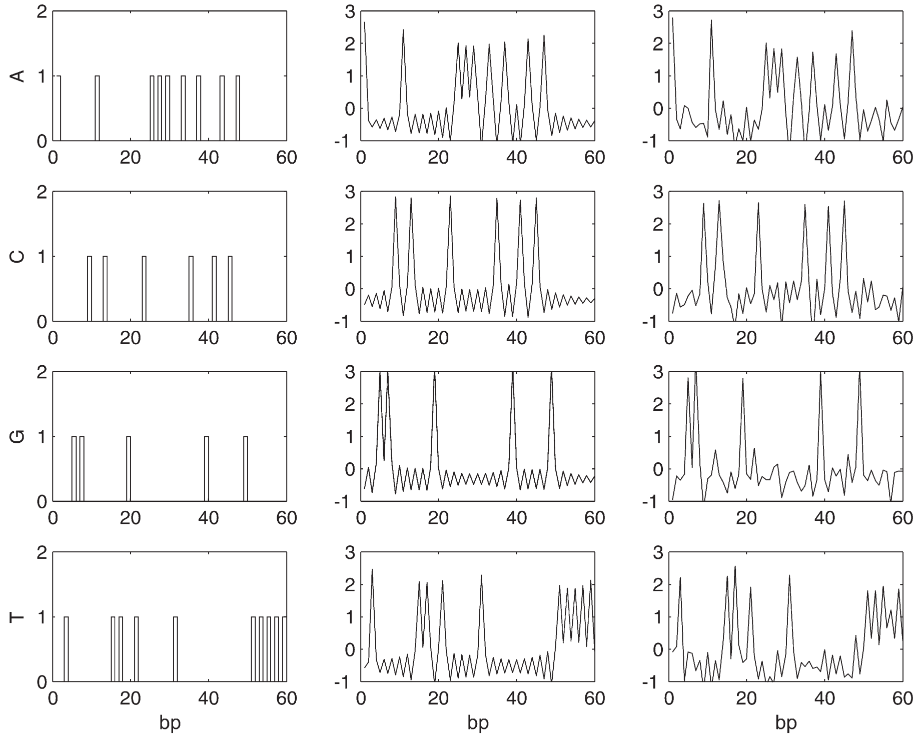

Fig. 1. Coding of DNA sequence. Upper part of picture—coding of the exemplary sequence into four numerical series (marked at left). Below, for 4 nucleotides from the left to the right: binary signal, signal filtered by low-pass Butterworth filter, signal after addition of 10% noise. On horizontal axis distance in base pairs (bp).

coefficients for construction of clustering trees for the globin family in a computationally efficient way.

2. Method

2.1. Data sets

DNA sequences were downloaded from public database EMBL[1]. They included: *Canis lupus* (CYGB, HBB, HBE1, MB, NGB), *Gallus gallus* (CYGB, HBA2, HBB, HBE1, HBG1, HBM, HBZ, MB, NGB), *Homo sapiens* (CYGB, HBA1, HBA2, HBB, HBD, HBE1, HBG1, HBG2, HBQ1, HBZ, MB, NGB), *Mus musculus* (CYGB, HBA1, HBB, HBZ, MB, NGB), *Pan troglodytes* (HBB, HBG1, HBG2, HBZ, MB, NGB) and *Rattus norvegicus* (CYGB, HBA1, HBA2, HBG1, MB, NGB). The details concerning used sequences may be found in Table 1 in the Appendix I.

Analyses were performed on: whole genes, exons obtained by sealing exon sequences of a given gene and introns obtained in the same way.

2.2. Codification of DNA sequences

The transformation of point processes into continuous signals allows for application of a broad range of methods developed for analysis of time series. To make DNA sequences suitable to this kind of approach we have used the method proposed for transformation of point processes into continuous signals (Kamiński et al., 2001; Kocsis and Kaminski, 2006). The method proved to be useful in estimation of relations between spike trains.

The procedure is as follows: each of the four channels was assigned to a specific nucleotide. The occurrence of a nucleotide at a given position was marked as 1 and the lack of it as 0 in the corresponding channel (similarly to the method of Voss, 1992). Then additional zeroes were inserted between each position. The insertion of additional zeroes provided (according to the Nyquist rule) a correct sampling frequency of $f_s = 2/\text{bp}$. The binary data were converted to continuous signals by application of low-pass order 1 Butterworth filter (Parks and Burrus, 1987), with 0.95 cut-off frequency. Filtering a signal is based on a multiplication of the spectral representation of the signal by a function called the transfer matrix of the filter. This operation changes the original spectrum of the signal leaving only the desired range of frequencies. In the time domain this operation is equivalent to convolution of the signal with a series of time domain filter coefficients. After that step the signal becomes a slowly changing, smooth waveform. Special care is needed to preserve position of the maxima in that signal, which correspond to nonzero values in the original point process. To make the filtered data better match the stochastic character of the autoregressive model, a small amount (10%) of white uncorrelated noise was added. The noise values were drawn from zero mean Gaussian distribution generated by the Matlab® routine randn with amplitude rescaled by a factor 0.1. Because the noise is not correlated in any way with the original signal its presence will not disturb the signal correlation structure from sample to sample, described by the model parameters $\mathbf{A}(i)$. The procedure of DNA sequences transformation is illustrated in Fig. 1.

[1] http://www.ebi.ac.uk/embl

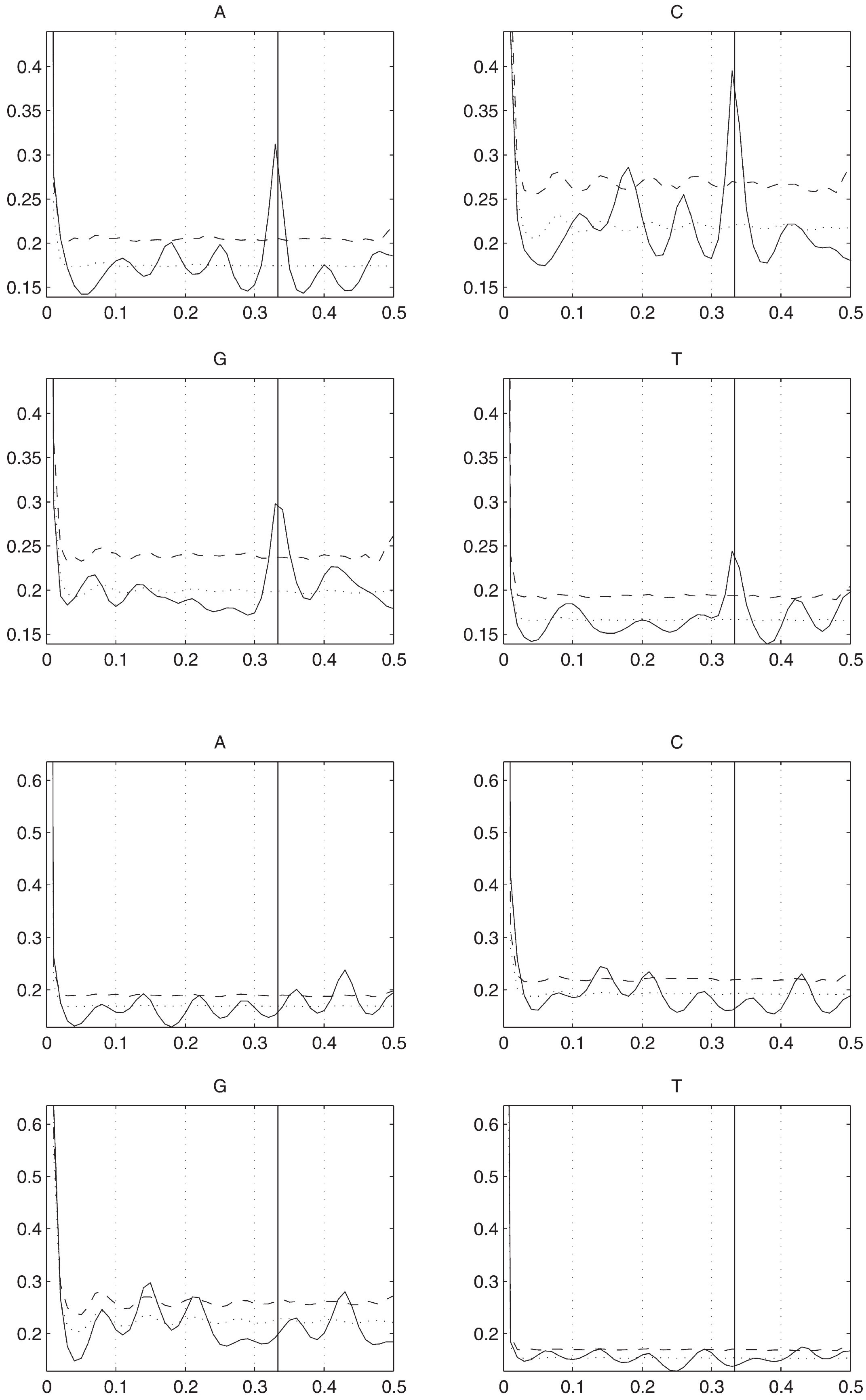

Fig. 2. Power spectrum of nucleotides of exon and intron sequences for human zetaglobin obtained from MVAR model—solid line; broken line—critical level for significance $\alpha = 0.05$; dotted line—average value for random distribution. Four upper pictures represent exon, four lower pictures intron spectra.

2.3. Multivariate autoregressive model

The method of codification proposed above allows for the use of repertoire of signal processing methods to the analysis of DNA sequences. Among them the multivariate autoregressive model (MVAR) proved to be an excellent tool for determination of interrelation between signals for multichannel processes e.g.: (Blinowska and Kaminski, 2006), (Kaminski and Liang, 2005).

The four channels representing nucleotide sequences were fitted simultaneously to the MVAR model. The MVAR model assumes that $\boldsymbol{X}(t)$—a sample of data at a position t—can be expressed as a sum of its j previous values weighted by model coefficients $\mathbf{A}$ plus a random component $\boldsymbol{E}(t)$:

$$\boldsymbol{X}(t) = \sum_{j=1}^{p} \mathbf{A}(j)\boldsymbol{X}(t-j) + \boldsymbol{E}(t). \quad (1)$$

The p is called the model order. For a k-channel process $\boldsymbol{X}(t)$ and $\boldsymbol{E}(t)$ are vectors of size k and the coefficients $\mathbf{A}$ are $k \times k$-sized matrices.

The calculation of the MVAR model coefficients is based on the estimation of the matrix of correlations between channels. Methods for finding model coefficients may be found in textbooks e.g.: (Marple, 1987; Lütkepohl, 1993); here we have used the Yule–Walker method. The outline of the method is presented in Appendix II.

The statistical criteria for determination of the model order (e.g. Akaike, 1974) did not show the distinct minimum, so the problem of optimal model order selection was approached by searching for a balance between the tendency to increase the accuracy of the fit by increasing the model order and the deterioration of statistical properties of the estimate for higher model order. The number of MVAR parameters, which is equal to k^2p, has to be several times higher than the number of data points, which is equal to kN (N is the number of data points in one channel). Considering the length of the shortest sequence in our data we have chosen model order 30. We have checked that the performance of the model is not very sensitive to the value of the model order. A change of model order in the range 30 ± 10 weakly influenced the shape of the spectra and the values of the coefficients.

2.4. Construction of phenograms

The construction of phenograms is based on the measures of the similarities between the points in the parameters space. To quantify the similarity a measure of the distance has to be introduced. After testing several measures of distance we have chosen the average correlation distance. The correlation distance d_{xy} was calculated according to the formula:

$$d_{xy} = 1 - \frac{C_{xy} + 1}{2} \quad (2)$$

where

$$C_{xy} = \frac{\sum_i x_i y_i}{\sqrt{\sum_j x_j^2 \sum_k y_k^2}} \quad (3)$$

and x, y are vectors of MVAR coefficients. Average correlation distance was obtained by summation of all sixteen d_{xy} values and division by their number.

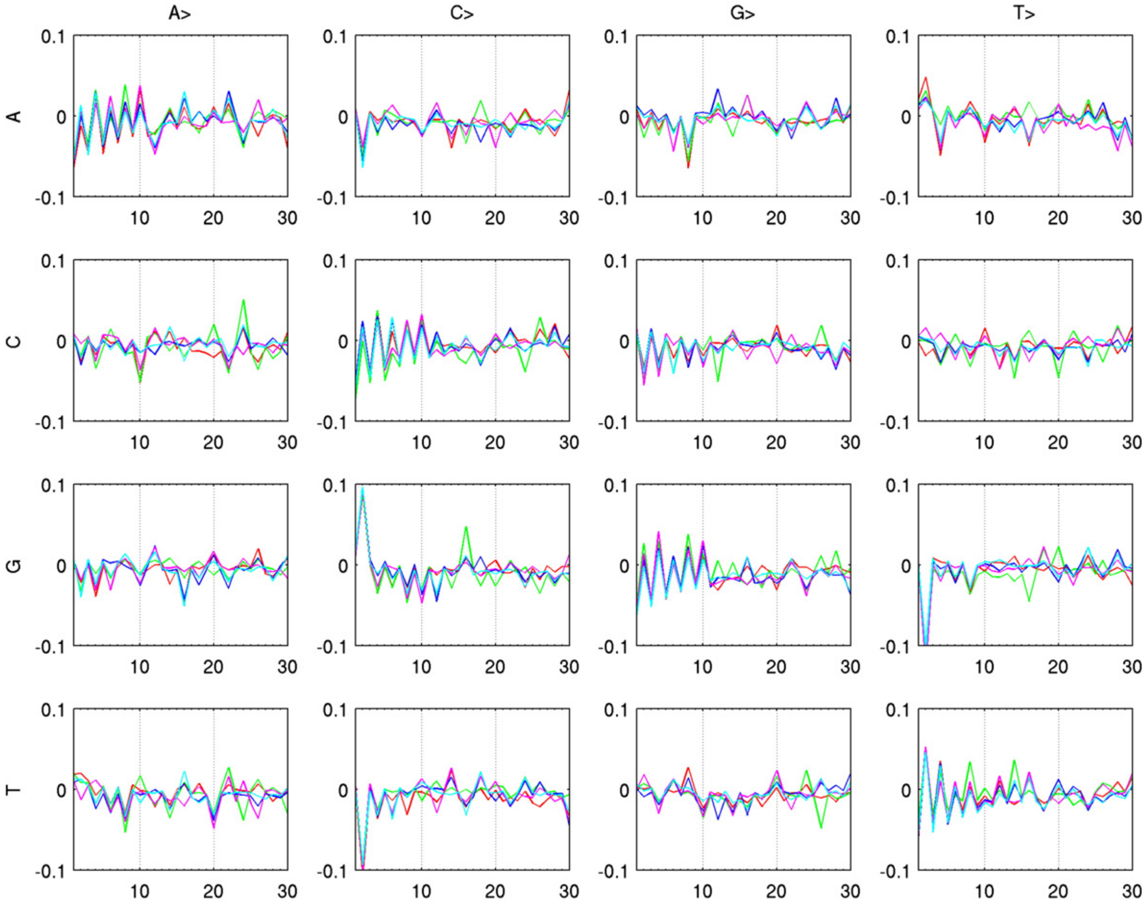

Fig. 3. MVAR coefficients $A_{ij}(m)$ as functions of nucleotide distance. For exon sequences of betaglobin for different species: human—blue, chimpanzee—cyan, dog—red, mouse—magenta, chicken—green). On the diagonal $A_{i=j}(m)$, off-diagonal $A_{i\neq j}(m)$.

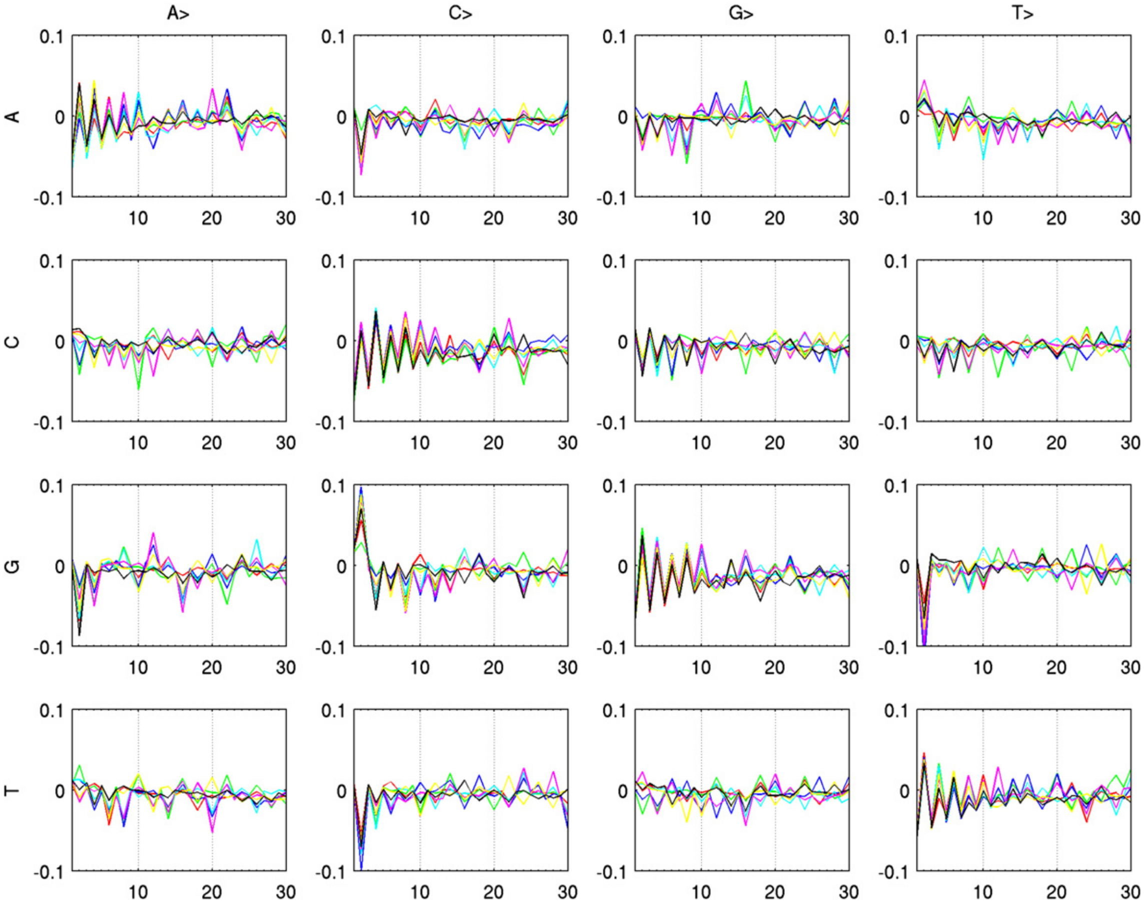

Fig. 4. MVAR coefficients for human exon sequences of 7 genes from globin family. The colors correspond to the following genes: CYGB—red, HBA1—green, HBB—blue, HBE1—cyan, HBG1—magenta, MB—yellow, NGB—black.

On the basis of the above defined measures the distance matrix was defined, which served for the construction of phenograms. Using the phylogenetic numerical methods two-dimensional unrooted bifurcational diagrams were constructed in the Euclidean space. The calculations were conducted by means of the PHYLIP package[2] for two algorithms: Neighbor-Joining Method (NJ) and Unweighted Pair Group Method with Arithmetic Mean (UPGMA).

2.5. Estimation of significance

The tests of significance of power spectra and correlations of model coefficients vectors obtained for different globins/species involved comparison with the random nucleotide sequences. For each analyzed data segment a random sequence of nucleotides was generated as permutations without repetitions, and for this sequence model coefficients were calculated. This procedure was repeated 1000 times and the distributions of function values were found. The significance levels of rejecting the hypothesis of no difference in respect to random sequences were calculated for correlations between MVAR coefficients for different species/globins. In the case of power spectra, from the estimated distributions percentiles were calculated on a significance level of 0.05.

In order to find the consensus trees the bootstrap method was applied. Bootstrap was realized by randomly drawing MVAR model coefficients (with repetitions), then the values of distances were calculated for the matrices of coefficients. Usually the jackknife method (random elimination of some components) is applied for obtaining consensus trees, however, the advantage of bootstrap is that the dimension of the samples pool is the same as in the original data (Efron, 1987).

3. Results

In the first step of our analysis we have considered the power spectra, which can be easily calculated from MVAR coefficients (the formula for calculation of spectra is given in Appendix II). It is known that for coding sequences (contrary to non-coding sequences) at 1/3 (bp) a spectral maximum occurs, which was found by Fourier analysis, e.g.: by Voss (1992), Buldyrev et al. (1995) and Fukushima et al. (2002). This phenomenon is connected with codon structure, which consists of three nucleotides (Buldyrev et al., 1995). The typical examples of exon and intron spectra are shown in Fig. 2. For the exons we have found in almost all cases statistically significant maxima corresponding to frequency 1/3, which was not the case for introns, where peaks of this frequency rarely appeared. This observation validates the application of MVAR model to the DNA sequences.

In the MVAR model the value of the signal at the given point is described by means of the preceding samples of the same channel and also by means of samples of all other channels multiplied by the model coefficients. By fitting a four channel MVAR model to the data, the matrix of coefficients of dimensions: 4×4×p is obtained:

$$M(m) = \begin{pmatrix} M_{AA}(m) & M_{AC}(m) & M_{AG}(m) & M_{AT}(m) \\ M_{CA}(m) & M_{CC}(m) & M_{CG}(m) & M_{CT}(m) \\ M_{GA}(m) & M_{GC}(m) & M_{GG}(m) & M_{GT}(m) \\ M_{TA}(m) & M_{TC}(m) & M_{TG}(m) & M_{TT}(m) \end{pmatrix}, m = 1, \dots, p. \quad (4)$$

[2] http://evolution.genetics.washington.edu/phylip.html

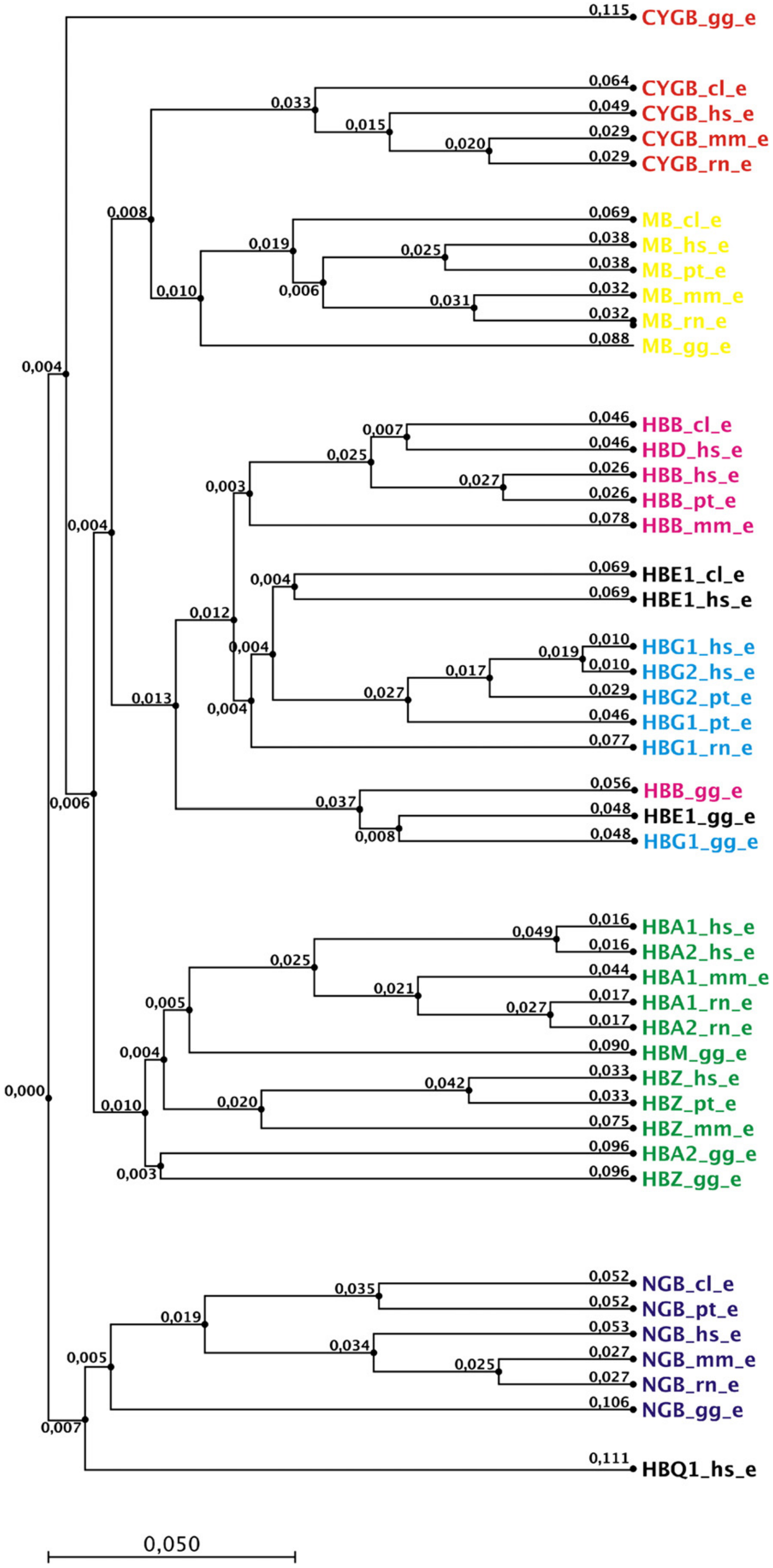

Fig. 5. Clustering tree for exon sequences of globin family genes. Numbers represent the distances between groups connected by the given knot. Measure of the distance—average correlation coefficient d_{xy}, clustering algorithm UPGMA. The abbreviations of species names: hs—*Homo sapiens*, pt—*Pan troglodytes*, cl—*Canis lupus*, rn—*Rattus norvegicus*, mm—*Mus musculus*, gg—*Gallus gallus*.

The matrix element from column j and row i describes the influence of channel j on the channel i as a function of index m, which gives the backward distance along the sequence measured in samples of 2/bp.

We have calculated the model coefficients for different globins and species. In Fig. 3 the coefficients obtained for the betaglobin exon sequence for different species are shown and in Fig. 4 the coefficients obtained for different human globins are illustrated. We can observe for the betaglobin gene that the MVAR coefficients are rather similar for different species (Fig. 3). The biggest differences can be observed for chicken. The differences in the coefficients values are bigger in the case of different genes in one species (Fig. 4).

In Fig. 3 as well as in Fig. 4 the differences between the coefficients are bigger for off-diagonal pictures. It means that most important for the distinction are the cross-dependences between four channels describing nucleotide sequences. Considering the shape of the functions it is easy to observe that the evolution of coefficients as a function of m bears the similarity, not the particular value at a given point. It suggests the choice of correlation as a measure of similarity between vectors of MVAR coefficients for different globins and species.

In order to compare the interspecies similarities of globins we have calculated the correlations between MVAR coefficients $A_{ij}(m)$. Magnitudes of the correlations between human and four other species for myoglobin sequences are shown Table 2 (Appendix I) together with the significance levels. The highest values were obtained for correlations between human and chimpanzee, as might have been expected. For all the studied globins the highest correlations between coefficients were observed for phylogenetically close species i.e. human and chimpanzee, mouse and rat.

From the correlations the distances were calculated. The average correlation distances between human and four species for myoglobin sequence are shown in Table 3 (Appendix I). The obtained distance between human and chimpanzee is the smallest one, the biggest are the distances between chicken and mammals.

These results were encouraging to attempt the construction of the clustering trees. Fig. 5 shows the phenogram for exon sequences obtained by means of the average correlation distance and the UPGMA algorithm. Inspecting Fig. 5 one can easily see the grouping of the orthologous genes belonging to different organisms, which might have been expected from the values of the correlations between vectors of model coefficients. In Fig. 5 one can distinguish groups: CYGB, MB, NGB with HBQ1, HBB–HBD, HBE–HBG, HBA–HBM–HBZ and the group of three genes of chicken: HBB, HBE1 and HBG1. Inside the groups the closest positions were occupied by the genes of related species namely: human and chimpanzee, mouse and rat.

In order to verify the obtained classification the consensus trees were calculated by means of the bootstrap method. The consensus tree obtained for exons by means of UPGMA algorithm is shown in Fig. 6. It differs from the tree obtained for original data, however the main tendencies are conserved, namely main groups of globins are clustered together. The exception was the cytoglobin family, which was split.

We have constructed also clustering trees and consensus trees for introns and for whole genes (Figs. 7, 8, 9, 10). In Fig. 7 the phenogram for intron sequences is shown. We can observe that even for these non-coding sequences main groups of globins are clustered together, however evolutionary close organisms rather than subgroups appear together. This is the case for the hemoglobin family. In the case of introns the consensus tree shown in Fig. 8 the clustering of similar globins is not as good as in previous figures. The hemoglobin family is split by a neuroglobin group. Cytoglobins do not form a distinct cluster.

The phenogram for whole genes (Fig. 9) bears resemblance to the phenogram for exons in respect of grouping families of genes. We can notice in the dendrograms of whole genes groups corresponding to the α-, β-, γ-globins, NGB, MB, CYGB, however, the ordering is not as good as it was in the case of exon sequences. The consensus tree for the whole genes is shown in Fig. 10. The structure of the tree is similar as in Fig. 9. There are tendencies of grouping together different globins for the same organisms. This is especially the case for chicken which was to a smaller degree also manifested in other dendrograms.

4. Discussion

Most statistical methods applied so far address long range properties of DNA sequences and they have a statistical limitation on the required length of sequences, however correlations on smaller scale (hundreds of base pairs) give information on the codon structure directly related to regulatory functions. Therefore, here we have concentrated on the short DNA sequences for demonstration of the classification power of the multichannel autoregressive approach. However, the method can be used as well for the sequences of thousands or tenths of thousands bp.

The application of the four-channel MVAR model allowed for establishing relations between the nucleotides occurrences as a function of nucleotide distances in terms of model coefficients. The evolution of coefficients $A_{ij}(m)$ linking different nucleotides ($i \neq j$) showed bigger variability depending on the kind of gene and also on the kind of species than autoregressive coefficients for a given nucleotide ($i=j$). This fact shows an advantage of the multivariate approach in statistical evaluation of DNA sequences, which gives an exhaustive and explicit description of the correlation structure of DNA sequences. In the proposed approach the relations between the occurrences of all four nucleotides are estimated in the framework of one MVAR model. It has been shown (Blinowska et al., 2004) that relations between the channels of a process may be found correctly only if all interdependent series are considered simultaneously in one multichannel model.

Comparison of our results with the phylogenetic tree of globins[3] shows very good concordance. Namely, according to the above reference, alpha and beta subunits of hemoglobin as well as myoglobin form the main branches of a phylogenetic tree and then smaller ramifications corresponding to the species are formed. This pattern is congruent with our findings. Cytoglobins, myoglobins, neuroglobins, α-globins and β-globins families are separated and form a main branches of the phylogenetic tree (Fig. 5). The subgroups inside the α- and β-globin families are ordered correctly as well, with only a few exceptions concerning very closely related globins. In case of consensus tree (Fig. 6) main branches enumerated above are still well distinguished, there are more exceptions in the case of subgroups, but the main structure is preserved. The ordering of main groups is also the case for introns, with the exception of cytoglobins (Fig. 7). Inside the main groups species rather than subgroups are clustered. This may be interesting information in tracing the evolution of particular species/globins. The 72 β-globin family genes from different species were used for phylogenetic reconstruction by Aguileta et al. (2004). They found several displacements supported by high bootstrap proportions. They conjecture that a potential source of conflict between the gene tree and the species tree could be gene conversion. This effect might be also the reason of some displacements observed by us.

The globin chromosomes are not among those which changed much during the evolution, so one may expect difficulties in distinguishing genes belonging to close species, however our method showed grouping of such species as a *Homo sapiens* and chimpanzee, rat and mouse (Tables 2 and 3). The distinction between the subgroups of genes such as HBG1–HBG2, HBA1–HBA2 was not pronounced, in this case rather the same species, not the same genes were grouped

[3] http://www.muhlenberg.edu/depts/biology/courses/bio152/BioinformaticsLab/betaglobininfo.html

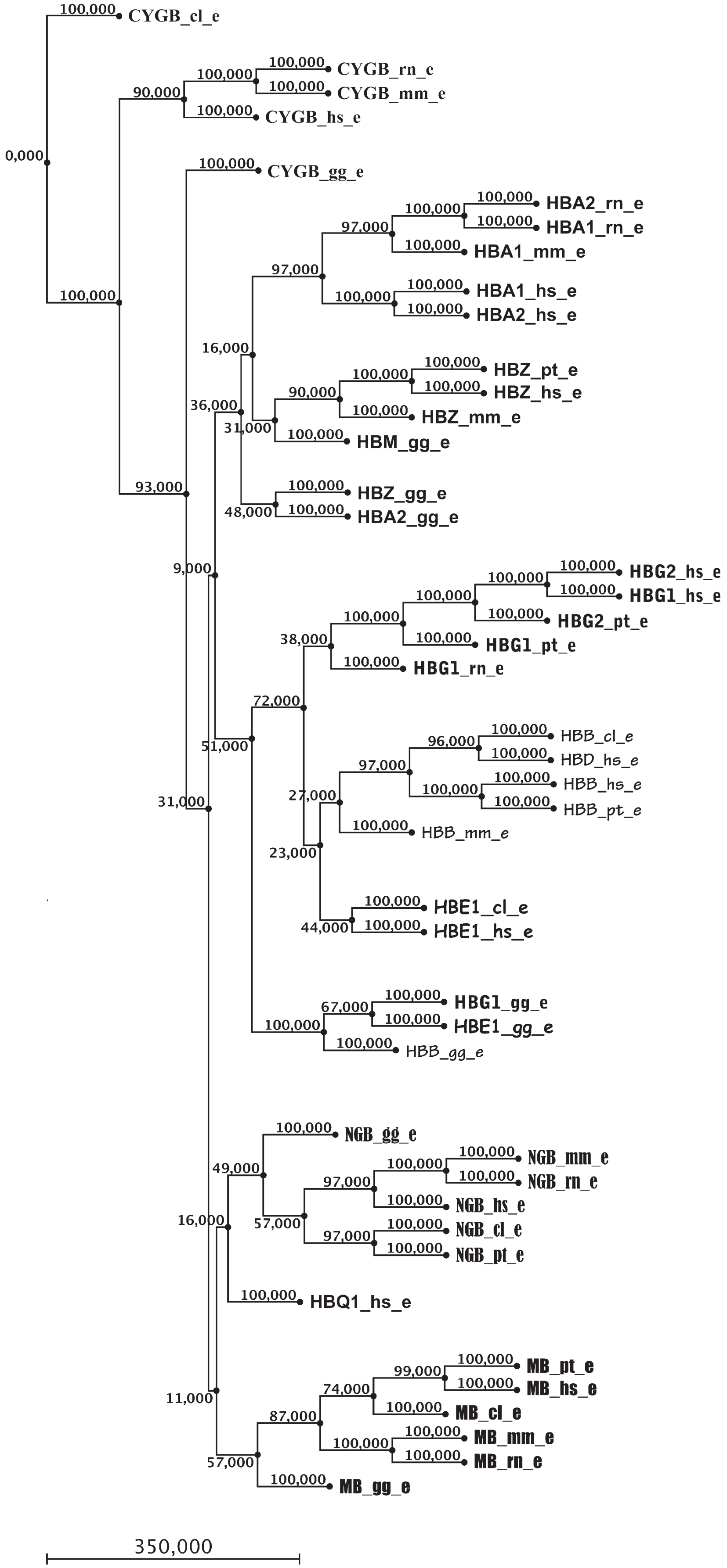

Fig. 6. Consensus tree for exon sequences of globin family genes for 100 bootstrap trials. Numbers represent the distances between groups connected by the given knot. Measure of the distance—average correlation coefficient d_{xy}, clustering algorithm UPGMA. The abbreviations of species names as in Fig. 5.

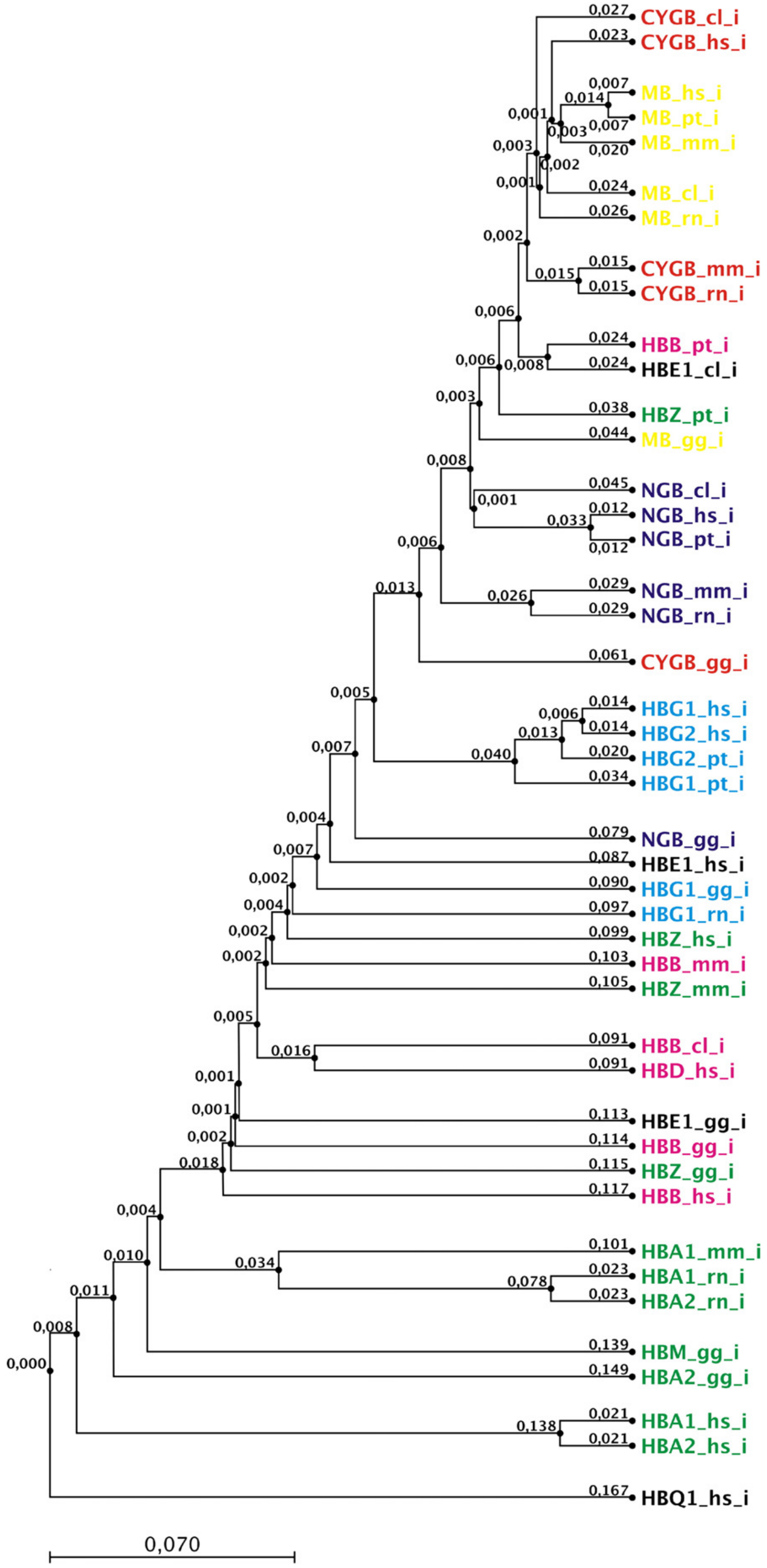

Fig. 7. Clustering tree for intron sequences of globin family genes. Numbers represent the distances between groups connected by the given knot. Measure of the distance—average correlation coefficient d_{xy}, clustering algorithm UPGMA. The abbreviations of species names as in Fig. 5.

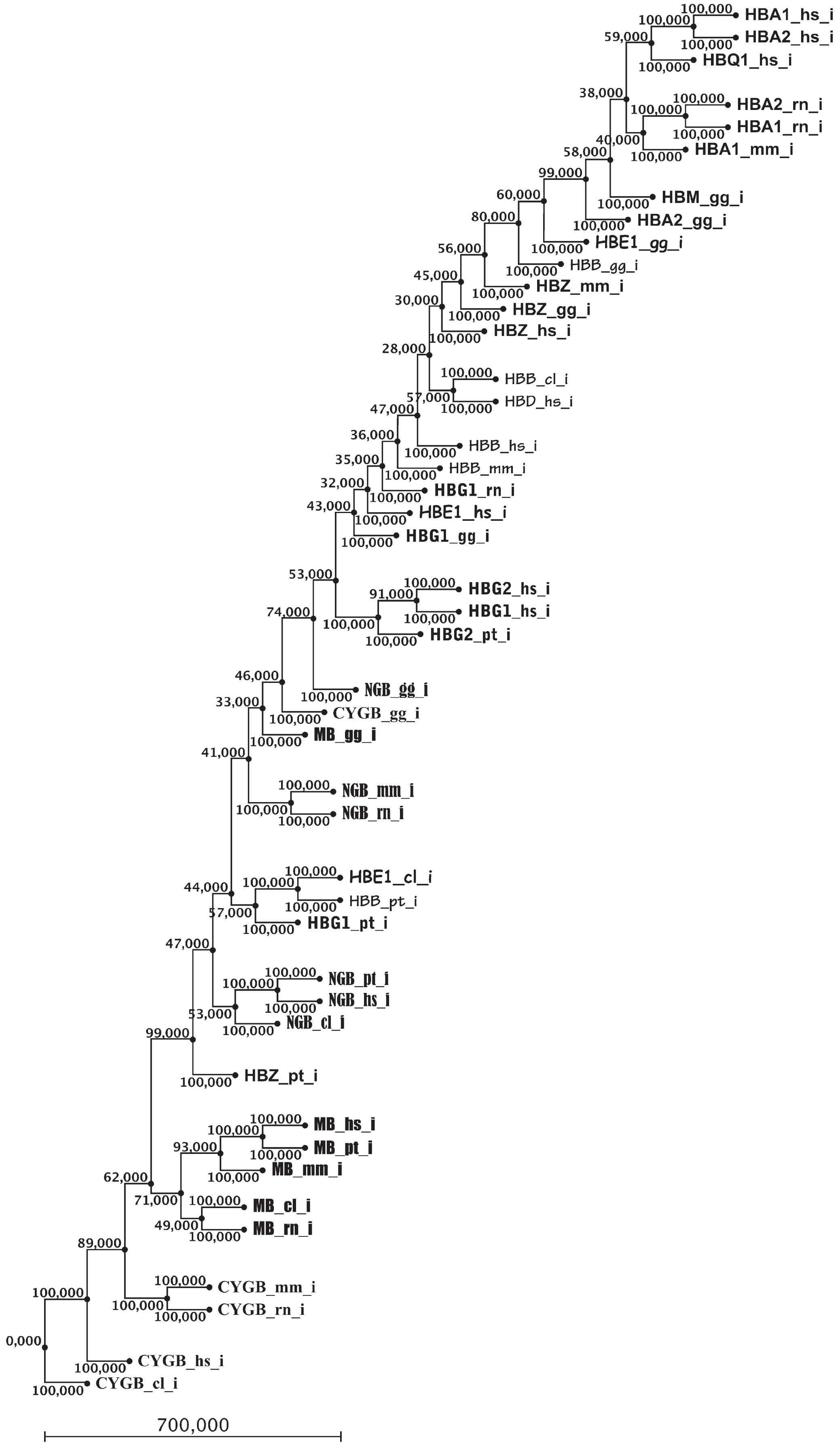

Fig. 8. Consensus tree for intron sequences of globin family genes. Numbers represent the distances between groups connected by the given knot. Measure of the distance—average correlation coefficient d_{xy}, clustering algorithm UPGMA. The abbreviations of species names as in Fig. 5.

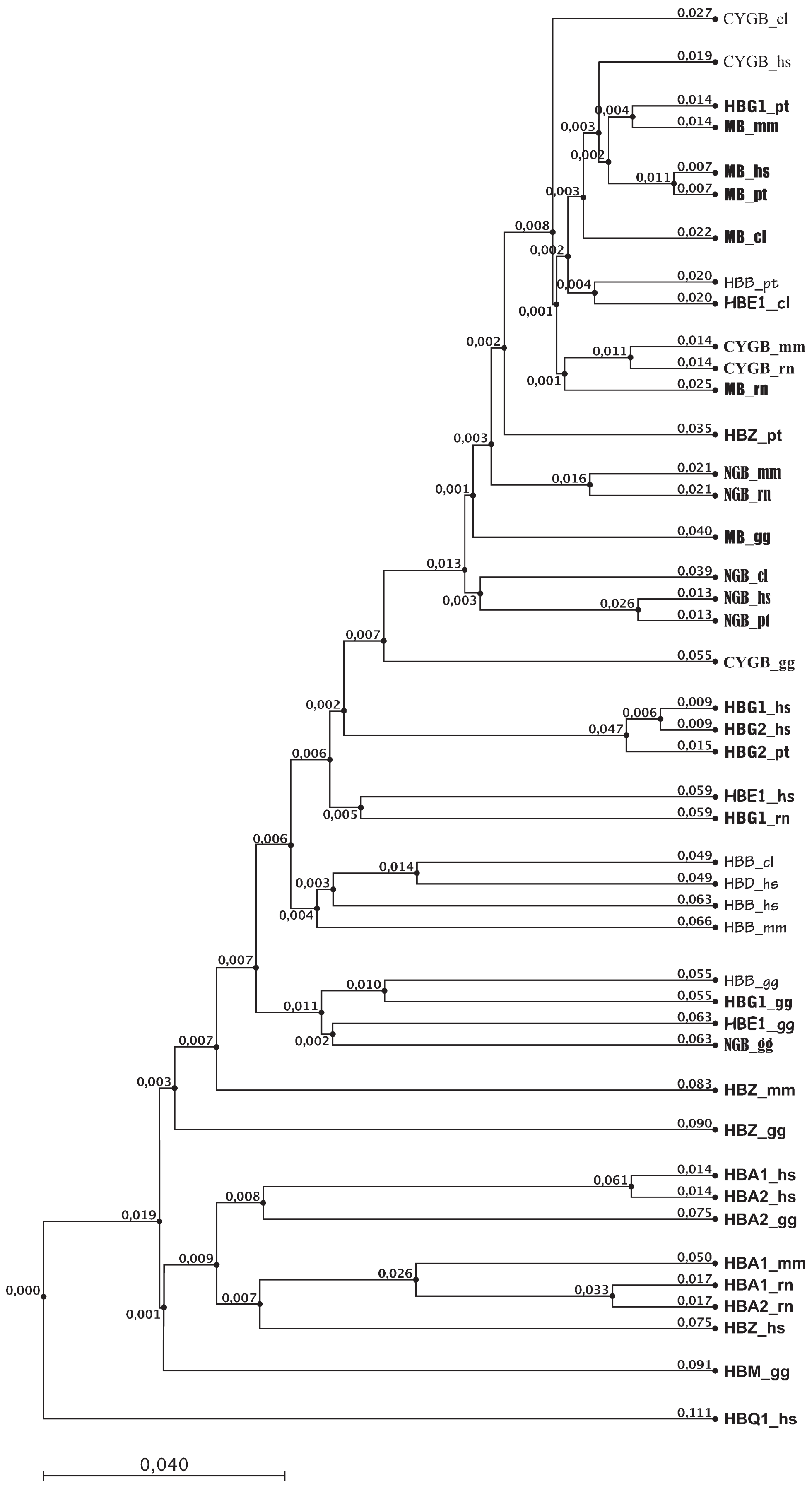

Fig. 9. Clustering tree for whole genes from globin family. Numbers represent the distances between groups connected by the given knot. Measure of the distance—average correlation coefficient d_{xy}, clustering algorithm UPGMA. The abbreviations of species names as in Fig. 5.

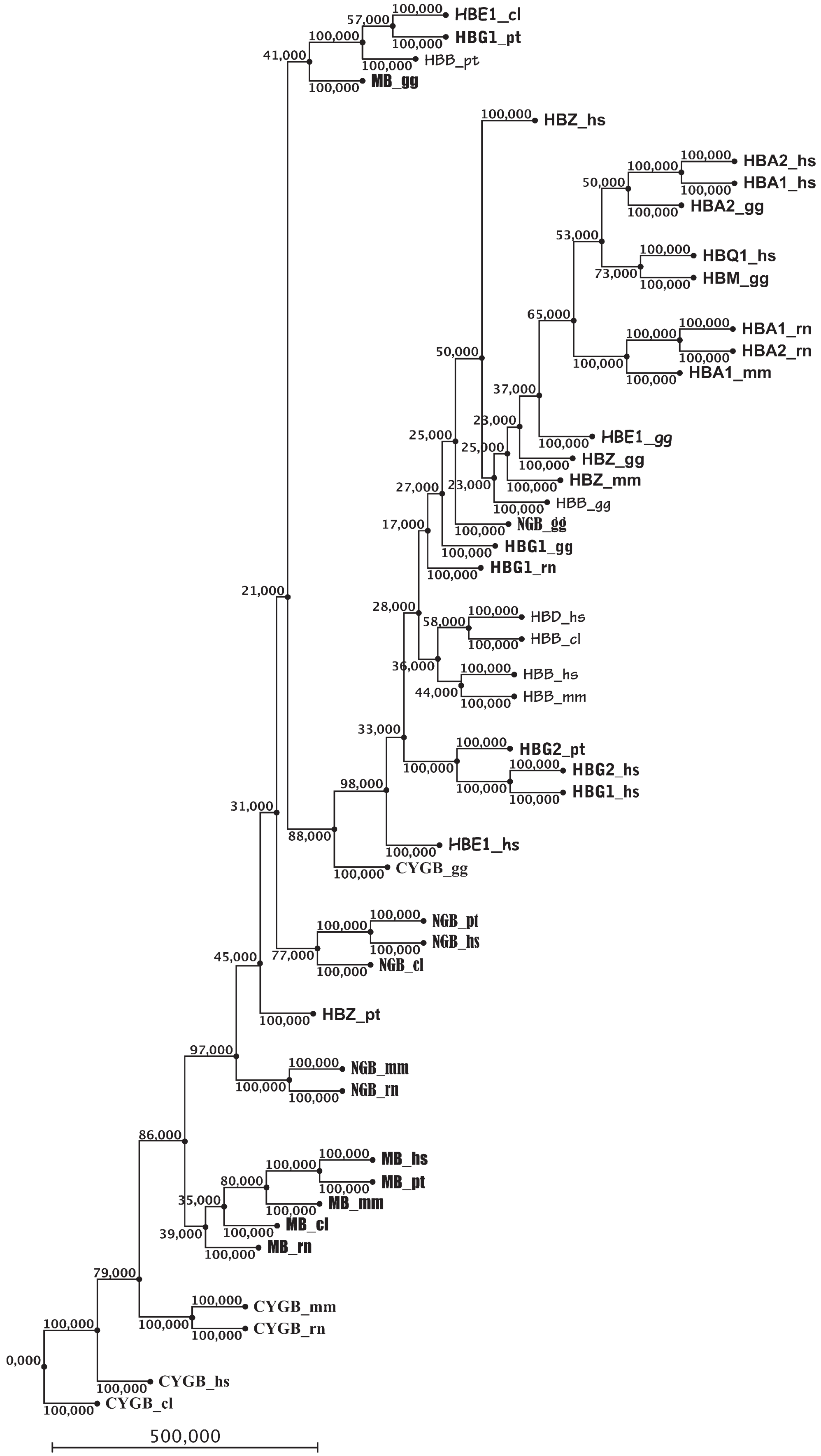

Fig. 10. Consensus tree for whole genes from globin family. Numbers represent the distances between groups connected by the given knot. Measure of the distance—average correlation coefficient d_{xy}, clustering algorithm UPGMA. The abbreviations of species names as in Fig. 5.

together. The analysis concerning subgroups of genes for different species may be helpful in finding phylogenies for evolutionary close organisms.

In Rokas et al. (2003) the incongruence in phylogenies obtained from single genes was pointed out. Our results for different genes give to a large degree coherent phylogenetic dependencies, although they concern highly related species, which in some statistical analyses are difficult to sort out: e.g. human and chimpanzee. In the phylogenetic analyses of whole genomes the interspecies differences, not the differences between genes, are dominant, however the study of exon, intron and gene sequences may bring a closer look at the particular evolutionary divergence.

Our results have shown that also from intron sequences the phylogenetic information may be retrieved, which may be helpful in analysis of evolutionary diversity. There is emerging evidence that interspecies differences may be found also in non-coding sequences e.g.: Willows-Munro et al. (2005). According to Luo et al. (1998) evolutionary dependences may be found in intron sequences, however they are expressed in another “genetic language”. The use of intron sequences for genomic and evolutionary analysis was discussed in Irmia and Roy (2008). It was pointed out that the appeal of intronic sequences for phylogenetics of recent divergences owes to their plausibly being both more rapidly and more neutrally evolving than protein coding. Relatively fast evolving intron sequences are especially suitable for resolving relationships between highly related animals, e.g. they have been applied successfully to explain evolutionary relationships among *Cetartiodactyla* (Matthee et al., 2001), *Leporidae* (Matthee et al., 2004) and also *Bovidae* (Matthee and Davis, 2001). The method proposed by us seems to be particularly useful to the study of closely related species, since it can operate on selected not very long intron sequences.

The proposed method is computationally very efficient. The computations were conducted on personal computer with Intel Core 2 Duo 2.0 GHz processor and 2048 MB RAM. The MVAR coefficients were calculated in real time. The construction of clustering trees encompassing 44 genes took 22 s. The generation of consensus trees based on 100 bootstraps (UPGMA method) took 250 s. Taking into account the amount of genetic information contained in data bases, waiting to be sorted, the speed of computations is not to be ignored.

Our approach has several advantages in comparison with the other methods applied so forth:

- The traditional methods of dendrogram construction based on multiple sequence alignment involve heavy computations and usually require powerful computers. The databases of the Gene Banks accumulate more and more material and there is an increasing need for fast methods to sort the stored information. The method proposed by us is very fast and can be run at the average PC. In comparison with the other statistical methods like DAR it is still much faster.
- Our method performs better than other statistical methods like DAR. Namely DAR model (Dehnert et al., 2005b) contrary to our model failed to distinguish human and chimpanzee genomes and chicken was placed close to human and chimpanzee, whereas in our model chicken was separated from mammals.
- Our approach can be applied to the long and to the short sequences as well. Other statistical methods require long sequences e.g. the DAR model required minimal length of a sequence of the order of 30 kbp (Dehnert et al., 2005b). Quite often, especially for the study of phylogenies of very close species the evolutionary information is encoded in short sequences (sometimes not only exons but also intron sequences give crucial information as it was pointed out above), also short sequences give information on the codon structure directly related to regulatory functions. Our method allows for comparison of many short sequences in a very efficient way.
- The power of the method comes from the fact that we include information of statistical relations *between* different nucleotides sequences; that part of multichannel data sets is not taken into account by other methods.

The method described in this article introduces a new approach to the study of phylogenic relations. It offers complete and explicit information on the auto- and cross correlation dependencies between nucleotide occurrences. Due to efficient parameterization of sequences the clustering trees can be obtained with a high level of confidence and in a computationally effective way. The method may become a useful tool in the study of phylogenetic diversity.

Acknowledgments

This work was supported by CoE BioExploratorium grants WKP_1/1.4.3/1/2004/44/44/115/2005 and WKP_1/1.4.3/2/2005/74/193/385/2006 and Polish Ministry of Science and Higher Education grant for Institute of Experimental Physics of Warsaw University.

Appendix I

Table 1
Sequence lengths of exons and introns used for construction of phenograms

Specie	Gene	Length of nucleotide sequences		
		Gene	Exonic	Intronic
Canis lupus	CYGB	9235	961	8274
	HBB	1529	679	850
	HBE1	9885	576	9309
	MB	10015	1137	8878
	NGB	4170	384	3786
Gallus gallus	CYGB	7644	540	7104
	HBA2	789	536	253
	HBB	1146	495	651
	HBE1	1191	540	651
	HBG1	1632	537	1095
	HBM	835	426	409
	HBZ	1434	429	1005
	MB	3816	902	2914
	NGB	2335	483	1852
Homo sapiens	CYGB	10343	1951	8392
	HBA1	842	576	266
	HBA2	834	575	259
	HBB	1606	626	980
	HBD	1800	774	1026
	HBE1	1794	816	978
	HBG1	1586	584	1002
	HBG2	1591	583	1008
	HBQ1	844	651	193
	HBZ	1651	589	1062
	MB	16591	1170	15421
	NGB	5822	1876	3946
Mus musculus	CYGB	8719	2331	6388
	HBA1	820	564	256
	HBB	1380	610	770
	HBZ	1512	596	916
	MB	35182	1056	34126
	NGB	5009	1616	3393
Pan troglodytes	HBB	15098	765	14333
	HBG1	35717	1156	6711
	HBG2	1775	764	1011
	HBZ	12797	703	12094
	MB	32841	797	32044
	NGB	4478	456	4022
Rattus norvegicus	CYGB	9767	2112	7655
	HBA1	856	556	300
	HBA2	844	544	300
	HBG1	1379	444	935
	MB	7232	947	6285
	NGB	5068	1554	3514

Table 2
Correlation coefficients between human and five other species for exonic sequences of myoglobin gene (in parentheses significance levels)

		A	C	G	T
A	*Pan troglodytes*	0.906 (0.062)	0.878 (<0.001)	0.853 (0.004)	0.942 (<0.001)
	Canis lupus	0.868 (0.399)	0.862 (0.004)	0.593 (0.595)	0.766 (0.026)
	Rattus norvegicus	0.870 (0.337)	0.783 (0.029)	0.563 (0.563)	0.719 (0.067)
	Mus musculus	0.903 (0.128)	0.798 (0.016)	0.544 (0.689)	0.798 (0.012)
	Gallus gallus	0.877 (0.292)	0.887 (<0.001)	0.659 (0.323)	0.850 (0.006)
C	*Pan troglodytes*	0.778 (0.019)	0.930 (0.017)	0.888 (<0.001)	0.706 (0.071)
	Canis lupus	0.843 (0.007)	0.911 (0.138)	0.767 (0.172)	0.770 (0.045)
	Rattus norvegicus	0.756 (0.067)	0.919 (0.066)	0.811 (0.039)	0.726 (0.078)
	Mus musculus	0.745 (0.065)	0.926 (0.041)	0.749 (0.150)	0.789 (0.029)
	Gallus gallus	0.455 (0.770)	0.933 (0.018)	0.624 (0.442)	0.572 (0.433)
G	*Pan troglodytes*	0.921 (0.001)	0.970 (<0.001)	0.934 (0.019)	0.903 (<0.001)
	Canis lupus	0.845 (0.017)	0.919 (<0.001)	0.950 (0.011)	0.802 (0.019)
	Rattus norvegicus	0.916 (<0.001)	0.932 (<0.001)	0.897 (0.223)	0.683 (0.212)
	Mus musculus	0.854 (0.008)	0.967 (<0.001)	0.932 (0.051)	0.682 (0.234)
	Gallus gallus	0.769 (0.065)	0.920 (<0.001)	0.890 (0.220)	0.613 (0.382)
T	*Pan troglodytes*	0.862 (0.001)	0.766 (0.024)	0.864 (0.002)	0.923 (0.017)
	Canis lupus	0.619 (0.230)	0.670 (0.191)	0.660 (0.278)	0.935 (0.014)
	Rattus norvegicus	0.710 (0.074)	0.653 (0.210)	0.504 (0.674)	0.761 (0.876)
	Mus musculus	0.719 (0.098)	0.758 (0.060)	0.546 (0.621)	0.833 (0.621)
	Gallus gallus	0.415 (0.804)	0.663 (0.192)	0.529 (0.647)	0.921 (0.033)

Table 3
The average correlation distances for myoglobin exons for six species (below diagonal) and statistical significances estimated by means of bootstrap (1000) repetitions (above diagonal)

	Homo sapiens	*Pan troglodytes*	*Canis lupus*	*Rattus norvegicus*	*Mus musculus*	*Gallus gallus*
Homo sapiens		<0.001	<0.001	<0.001	<0.001	<0.001
Pan troglodytes	0.073		<0.001	<0.001	<0.001	<0.001
Canis lupus	0.124	0.123		<0.001	<0.001	<0.001
Rattus norvegicus	0.150	0.121	0.156		<0.001	0.002
Mus musculus	0.138	0.114	0.147	0.062		0.001
Gallus gallus	0.171	0.167	0.165	0.204	0.175	

Appendix II

The autoregressive (AR) model is based on an assumption that a value of the series $\boldsymbol{X}$ at a point t can be expressed by its p previous values with certain coefficients $\mathbf{A}(i)$ plus a noise component $\mathbf{E}$ (see Eq. (1)). This type of a model is widely used e.g. in economics and especially in biomedical signal analysis. The AR model describes *stochastic* time series, randomly changing its values from point to point; the random component $\mathbf{E}$ in Eq. (1) represents the stochastic part of the modeled signal. When we observe a set of k signals simultaneously, such process is called multivariate or multichannel, $\boldsymbol{X}$ and $\boldsymbol{E}$ are vectors of size k and coefficients $\mathbf{A}$ are matrices of size $k \times k$.

There are many methods of finding AR model parameters for the given data. We used the Yule–Walker method, which will be briefly described below. First, we must calculate the cross-correlation matrix $\mathbf{R}$ of the signals for time lags s ranging from 0 to p:

$$\mathbf{R}(s) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}(i)\boldsymbol{X}^{\mathrm{T}}(i+s), \quad s = 0, \ldots, p. \tag{5}$$

Then we notice that if we multiply both sides of the model equation (Eq. (1)) by $\boldsymbol{X}^{\mathrm{T}}(i-s)$ and take expectations of both sides the result can be expressed using the correlation matrix $\mathbf{R}$:

$$\sum_{i=1}^{p} \mathbf{A}(i)\mathbf{R}(i{-}s) + \mathbf{0} = \mathbf{R}(-s) \tag{6}$$

where $\mathbf{0}$ represents zero matrix—result of expectations taken on product of signal and noise, which are uncorrelated. After transposition of both sides of Eq. (6) we get

$$\sum_{i=1}^{p} \mathbf{R}^{\mathrm{T}}(i{-}s)\mathbf{A}^{\mathrm{T}}(i) = \mathbf{R}^{\mathrm{T}}(-s). \tag{7}$$

Repeating this for $s = 1, \ldots, p$ we get a set of linear equations to solve where $\mathbf{A}$ are unknowns and $\mathbf{R}$ are calculated from the data:

$$\begin{pmatrix} \mathbf{R}^{\mathrm{T}}(0) & \mathbf{R}^{\mathrm{T}}(1) & \cdots & \mathbf{R}^{\mathrm{T}}(p{-}1) \\ \mathbf{R}^{\mathrm{T}}(-1) & \mathbf{R}^{\mathrm{T}}(0) & & \mathbf{R}^{\mathrm{T}}(p{-}2) \\ \vdots & & \ddots & \vdots \\ \mathbf{R}^{\mathrm{T}}(1{-}p) & \cdots & \cdots & \mathbf{R}^{\mathrm{T}}(0) \end{pmatrix} \begin{pmatrix} \mathbf{A}^{\mathrm{T}}(1) \\ \mathbf{A}^{\mathrm{T}}(2) \\ \vdots \\ \mathbf{A}^{\mathrm{T}}(p) \end{pmatrix} = \begin{pmatrix} \mathbf{R}^{\mathrm{T}}(-1) \\ \mathbf{R}^{\mathrm{T}}(-2) \\ \vdots \\ \mathbf{R}^{\mathrm{T}}(-p) \end{pmatrix}. \tag{8}$$

Solving this set we obtain the set of model coefficients $\mathbf{A}(i)$ describing properties of the original data.

Eq. (1) can be easily transformed to describe relations in the frequency domain. After rewriting Eq. (1) in the following form (with sign of $\mathbf{A}$ changed)

$$\boldsymbol{E}(t) = \sum_{j=0}^{p} \mathbf{A}(j)\boldsymbol{X}(t{-}j) \tag{9}$$

the application of Z transformation (which is analogous to Fourier transformation in this case) yields

$$\boldsymbol{E}(f) = \mathbf{A}(f)\boldsymbol{X}(f) \tag{10}$$

After multiplying both sides of Eq. (10) by A^{-1} we get

$$\begin{gathered} \boldsymbol{X}(f) = \mathbf{A}^{-1}(f)\boldsymbol{E}(f) = \mathbf{H}(f)\mathbf{E}(f) \\ \text{where}: \quad \mathbf{H}(f) = \left(\sum_{m=0}^{p} \mathbf{A}(m)\exp(-2\pi i m f \Delta t)\right)^{-1} \end{gathered} \tag{11}$$

From the form of that equation we see that the model can be considered as a linear filter with white noises $\boldsymbol{E}(f)$ on its input—right hand of the equation (flat dependence on frequency) and the signals $\boldsymbol{X}(f)$ on its output—left hand of the equation. The matrix of filter coefficients $\mathbf{H}(f)$ is called the transfer matrix of the system. It contains information about all relations between data channels in the given set.

From the matrix $\mathbf{H}(f)$ estimators such as spectra and coherences can be calculated. It easily follows that the spectral matrix is given by:

$$\mathbf{S}(f) = \boldsymbol{X}(f)\boldsymbol{X}^{*}(f) = \mathbf{H}(f)\boldsymbol{E}(f)\boldsymbol{E}^{*}(f)\mathbf{H}^{*}(f) = \mathbf{H}(f)\mathbf{V}\mathbf{H}^{*}(f) \tag{12}$$

(asterisk denotes a transpose and complex conjugate operation). The matrix $\mathbf{S}(f)$ contains auto-spectra of each channel on the diagonal and cross-spectra off the diagonal.

References

Aguileta, G., Bielawski, J.P., Yang, Z., 2004. Gene conversion and functional divergence in the β-globin gene family. J. Mol. Evol. 59, 177–189.

Akaike, H., 1974. A new look at statistical model identification. IEEE Trans. Automat. Contr. 19, 716–723.

Anastassiou, D., 2002. Genomic signal processing. IEEE Signal Process. Mag. 90, 1859–1867.

Arneodo, A., Bacry, E., Graves, P.V., Muzy, J.F., 1995. Characterizing long-range correlations in DNA sequences from wavelet analysis. Phys. Rev. Lett. 74 (16), 3293–3296.

Audit, B., Thermes, C., Vaillant, C., d'Aubenton-Carafa, Y., Muzy, J.F., Arneodo, A., 2001. Long-range correlations in genomic DNA: a signature of the nucleosomal structure. Phys. Rev. Lett. 86 (11), 2471–2474.

Bernaola-Galvan, P., Carpena, P., Román-Roldán, R., Oliver, J.L., 2002. Study of statistical correlation in DNA sequences. Gene 300, 105–115.

Blinowska, K.J., Kaminski, M., 2006. Multivariate signal analysis by parametric models. In: Schelter, B., Winterhandler, M., Timmer, J. (Eds.), Handbook of Time Series Analysis — Recent Theoretical Developments and Applications. Wiley-VCH, pp. 373–409.

Blinowska, K.J., Kus, R., Kaminski, M., 2004. Granger causality and information flow in multivariate processes. Phys. Rev. E 70, 050902.

Buldyrev, S.V., et al., 1995. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. Phys. Rev. E 51 (5), 5084–5091.

Chakravarthy, N., Spanias, A., Iasemidis, L.D., Tsakalis, K., 2004. Autoregressive modeling and feature analysis of DNA sequences. EURASIP J. Appl. Signal Process. 1, 13–28.

Dehnert, M., Helm, W.E., Hütt, M.-Th., 2003. A discrete autoregressive process as a model for short-range correlations in DNA sequences. Physica, A 327, 535–553.

Dehnert, M., Plaumann, R., Helm, W.E., Hütt, M.-T.H., 2005a. Genome phylogeny based on short-range correlations in DNA sequences. J. Comput. Biol. 12 (5), 545–553.

Dehnert, M., Helm, W.E., Hütt, M.-Th., 2005b. Information theory reveals large-scale synchronisation of statistical correlations in eukaryote genomes. Gene 345, 81–90.

Efron, B., 1987. The jackknife, the bootstrap and other resampling plans. Vol. 38 of CBMS-NSF Regional Conference Series in Applied Mathematics. S.I.A.M.

Fukushima, A., et al., 2002. Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis. Gene 300, 203–211.

Grosse, L., Hertzel, H., Buldyrev, S.V., Stanley, H.E., 2000. Species independence of mutual information in coding and non-coding DNA. Phys. Rev. E 61, 5624–5629.

Holste, D., Grosse, I., 2003. Repeats and correlations in human DNA sequences. Phys. Rev. E 67, 061913.

Irmia, M., Roy, S.W., 2008. Spliceosomal introns as tools for genomic and evolutionary analysis. Nucleic Acids Res. 36, 1703–1712.

Kaminski, M., Liang, H., 2005. Causal influence: advances in neurosignal analysis. Crit. Rev. Biomed. Eng. 33 (4), 347–430.

Kamiński, M., Ding, M., Truccolo, W., Bressler, S., 2001. Evaluating causal relations in neural systems: Granger causality, directed transfer function (DTF) and statistical assessment of significance. Biol. Cybern. 85, 145–157.

Karlin, S., Mrazek, J., 1997. Compositional differences within and between eukaryotic genomes. Proc. Natl. Acad. Sci. U. S. A. 94, 10227–10232.

Kocsis, B., Kaminski, M., 2006. Dynamic changes in the direction of the theta rhythmic drive between supramammillary nucleus and the septohippocampal system. Hippocampus 16 (6), 531–540.

Luo, L., Lee, W., Jia, L., Fengmin, J., Tsai, L., 1998. Statistical correlation of nucleotides in a DNA sequence. Phys. Rev. E 58 (1), 861–871.

Lütkepohl, H., 1993. Introduction to Multiple Time Series AnalysisSecond Edition. Springer-Verlag, Berlin-Heidelberg Germany.

Marple, S.L., 1987. Digital Spectral Analysis With Applications, Prentice-Hall Signal Processing Series. Simon & Schuster, New Jersey.

Matthee, C.A., Davis, S.K., 2001. Molecular insights into evolution of the family *Bovidae*: a nuclear DNA perspective. Mol. Biol. Evol. 18, 1220–1230.

Matthee, C.A., Burzlaff, J.D., Taylor, J.F., Davis, S.K., 2001. Mining the mammalian genome for artiodactyls systematics. Syst. Biol. 50, 367–390.

Matthee, C.A., Jansen van Vuuren, B., Bell, D., Robinson, T.J., 2004. A molecular supermatrix of the rabbits and hares (Leporiadae) allows for the identification of five intercontinental exchanges during the Miocene. Syst. Biol. 53, 443–447.

Parks, T.W., Burrus, C.S., 1987. Digital Filter Design. John Wiley & Sons, New York.

Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.

Stanley, H.E., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Simons, M., 1999. Scaling features of noncoding DNA. Physica, A 273, 1–18.

Trifonov, E.N., 1998. 3-, 10.5-, 200- and 400-base periodicities in genome sequences. Physica, A 249, 511–516.

Voss, R.F., 1992. Evolution of long-range fractal correlations and 1/*f* noise in DNA base sequences. Phys. Rev. Lett. 68, 3805–3808.

Willows-Munro, S., Robinson, T.J., Matthee, C.A., 2005. Utility of nuclear DNA intron markers at lower taxonomic levels: phylogenetic resolution among nine Tragelaphus spp. Mol. Phylogenet. Evol. 35, 624–636.

Yu, Z.-G., Anh, V., 2004. Phylogenetic tree of prokaryotes based on complete genomes using fractal and correlation analyses. Australian Computer Society Inc. In: Yi-Pimng, Phoebe Chen (Ed.), Conferences in Research and Practice in Information Technology, Vol. 29.