Available online at www.sciencedirect.com

SCIENCE DIRECT®

Gene 345 (2005) 81–90

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Information theory reveals large-scale synchronisation of statistical correlations in eukaryote genomes

Manuel Dehnert[a], Werner E. Helm[b], Marc-Thorsten Hütt[a,*]

[a]Bioinformatics Group, Department of Biology, Darmstadt University of Technology, D-64287 Darmstadt, Germany
[b]Mathematics and Science Faculty, University of Applied Sciences, D-64295 Darmstadt, Germany

Received 15 September 2004; received in revised form 18 October 2004; accepted 9 November 2004
Available online 4 January 2005
Received by M. Porto

Abstract

We study short-range correlations in DNA sequences with methods from information theory and statistics. We find a persisting degree of identity between the correlation patterns of different chromosomes of a species. Except for the case of human and chimpanzee inter-species differences in this correlation pattern allow robust species distinction: in a clustering tree based upon the correlation curves on the level of individual chromosomes distinct clusters for the individual species are found. This capacity of distinguishing species persists, even when the length of the underlying sequences is drastically reduced. In comparison to the standard tool for studying symbol correlations in DNA sequences, namely the mutual information function, we find that an autoregressive model for higher order Markov processes significantly improves species distinction due to an implicit subtraction of random background.

Keywords: Genome signature; Mutual information; Markov process; Clustering tree

1. Introduction

The correlation structure of a DNA sequence has been extensively studied within the last decade, particularly with methods from information theory (see, e.g., Herzel et al., 1994; Holste et al., 2000; Grosse et al., 2002; Berryman et al., 2004). In these investigations, the mutual information function has turned out to reveal important features of the information content of DNA sequences. The mutual information function shows a clear difference between coding and non-coding DNA (due to codon usage), a feature that has been found to be species independent (Grosse et al., 2000). The enormous explicatory power of information theory in general and the mutual information function in particular comes out clearest when properties of some information observable can be related to biological components. For the case of Alu repeats this has been achieved by Holste et al. (2003). In many cases the biological origin of such correlations is still far from being fully understood (see, e.g., Ouyang et al., 2004; Garte, 2004 for recent studies). The bulk of scientific efforts to apply information theory as well as various other statistical methods to studying DNA sequences, however, focused–and still focuses (Berryman et al., 2004)–on long-range correlations (Voss, 1992; Li and Kaneko, 1992; Peng et al., 1992; Karlin and Brendel, 1993; Buldyrev et al., 1995). Large-scale statistical properties of DNA sequences have recently received a lot of attention in the statistical physics community. There, long-range patterns have been analysed with methods from dynamical systems theory (Nicolay et al., 2004) and fractal geometry (Garte, 2004). From a completely different perspective, with the advent of fully sequenced genomes studies of genome-wide properties

Abbreviations: TCC, tree colour coding; MI, mutual information; DAR(p), discrete autoregressive process of the order p.

[*] Corresponding author. Tel.: +49 6151 163202; fax: +49 6151 164630.

E-mail address: huett@bio.tu-darmstadt.de (M.-T. Hütt).

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.11.026

emerged with a special emphasis on possible phylogenetic information that can be found on this large-scale level. These studies show, e.g., that phylogeny can be retrieved from the number of common genes in different unicellular species (Snel et al., 1999) or from integration sites of interspersed elements (retroposons) (Shimamura et al., 1997; Nikaido et al., 1999). In (Rokas et al., 2003) it has been shown that phylogenetic accuracy increases with taking into account ever more information on the basis of orthologous genes along a genome.

On a more quantitative level (and–more importantly–without looking at particular biological features or components of the sequence) attempts have been made to set up genome "signatures", from which distance matrices can be computed and clustering trees can be obtained. Karlin et al. showed that dinucleotide abundances reveal species information (Karlin and Ladunga, 1994; Karlin and Mrázek, 1997; Gentles and Karlin, 2001). Hao et al. studied n-word distributions for prokaryotes (Hao and Qi, 2003; Qi et al., 2004).

We have recently shown that phylogenetic reconstruction is also possible from genome-wide statistical correlations (Dehnert et al., submitted). The correlation strength of two nucleotides in a distance k is the key ingredient of this analysis. Here we will extend this investigation to a larger set of eukaryotic species and we will see that in this case the relation between the different correlation patterns goes far beyond a simple phylogenetic interpretation. Particularly the cases where species clustering based on the correlation pattern does not coincide with purely phylogenetic relations are expected to reveal the biological origin of the striking synchrony in correlation patterns we find with methods from information theory.

2. Material and methods

2.1. Data sets

Genomic sequences were downloaded from public databases. From the site ftp://ftp.ensembl.org/pub/ the sequences of the mosquito *Anopheles gambiae* (anopheles—22.2b), the zebrafish *Brachydanio rerio* (zebrafish—22.3b), the nematode *Caenorhabditis elegans* (celegans—22.116a), the fruit fly *Drosophila melanogaster* (fly—22.3a), the chicken *Gallus gallus* (chicken—22.1), the human *Homo sapiens* (human—22.34d), the mouse *Mus musculus* (mouse—22.32b), the chimpanzee *Pan troglodytes* (chimp—22.1), and the rat *Rattus norvegicus* (rat—22.3b) have been obtained. From the site http://www.ncbi.nlm.nih.gov the sequences of the yeast *Saccharomyces cerevisiae*, (NC_001133-NC_001148), the cress *Arabidopsis thaliana*, (NC_003070, NC_003071, NC_003074-NC_003076), the malaria parasite *Plasmodium falciparum*,(NC_000910, NC_000521, NC_004325-NC_004331, NC_004314-NC_004318) have been obtained. Data for *Encephalitozoon cuniculi*, (AL391737, AL590442-AL590451), *Ashbya gossypii*, (AE016814-AE016820) have been taken from the site http://www.ebi.ac.uk/genomes. Unidentified nucleotides have been discarded for this analysis. We checked that substituting unidentified nucleotides by random nucleotides (instead of omitting them from the sequence) has no significant influence on the correlation curves. Chromosome 16 from *G. gallus* is regarded as an outlier: more than 20% of this chromosome consist of unidentified nucleotides. It has been excluded from the analysis.

2.2. Mathematical background: mutual information function and DAR(p) process

We formulate two methods for quantifying the correlation between nucleotides in a DNA sequence. Let $p(i)$ denote the probability of finding the symbol i from the alphabet $\mathbf{A}$={A,G,C,T } in a given DNA sequence and let $p^{(k)}$ (i,j) be the probability of finding the symbols i and $j \in \mathbf{A}$ at a distance k in the sequence. The mutual information function is then defined as (see Ebeling et al., 1998 for the general background and Grosse et al., 2000 for applications to DNA sequences)

$$I(k) = \sum_{(i,j)\in\mathbf{A}^2} p^{(k)}(i,j)\log_2 \frac{p^{(k)}(i,j)}{p(i)p(j)}. \tag{1}$$

The function $I(k)$ quantifies the amount of information one obtains from the symbol i on a symbol j at a distance k within the sequence. It is, therefore, a measure of the strength of average correlation between the symbols i and j at a distance k. A more refined method for quantifying the average correlation of two nucleotides at a distance k is given by the parameters of a discrete autoregressive process of the order p, DAR(p). On an elementary level, a DAR(p) process is a model to generate symbol sequences with a Markov property of higher order (Dehnert et al., 2003). By estimating the parameters of a DAR(p) process from a given DNA sequence quantities approximating the correlation strength at a certain distance are obtained up to the Markov order p. How does the model work? Let X_n be the n-th symbol in a sequence generated by a DAR(p) process. Then X_n is given by the following recursion relation (Jacobs and Lewis, 1978, 1983; Dehnert et al., 2003)

$$X_n = V_n X_{n-A_n} + (1 - V_n)Y_n. \tag{2}$$

The first term in this relation is responsible for the Markov property mentioned above, while the second term introduces an amount of uncorrelated randomness into the sequence. The stochastic variable V_n can have two values namely V_n=1 and V_n=0 and, consequently, serves as a switch between the two terms on the right-hand side of Eq. (2). The value V_n=1 occurs with probability ρ while the other value V_n=0 occurs with the remaining probability

$1-\rho$. The quantity ρ constitutes the first parameter of the DAR(p) process. The remaining parameters are contained in the stochastic variable A_n which takes the values $A_n=1, 2, 3,\ldots, p$ with probabilities α_1 α_2 α_3 $\ldots,\alpha_p$, respectively. Lastly, the quantity Y_n denotes a random symbol chosen from the alphabet **A** according to a specified distribution. The random variables V_n, A_n, and Y_n are assumed to be independent. The sequence X_n itself has a p-th order Markov property (i.e. it has a finite memory of length p). By construction the parameters α_k represent the strength of correlation between two symbols in the sequence at a distance k. From Eq. (2) it is seen that the values of α_k jointly with regulate, how often a symbol X_n in the sequence is determined by a symbol X_{n-k}, which occurs at the k-th previous position. To some extent, the vector $\{\alpha_1\ \alpha_2\ \alpha_{3,\ldots,}\ \alpha_p\}$ determines the intrinsic memory of the sequence generated with this DAR(p) process. Instead of using the DAR(p) process as a model (i.e. a method for generating a sequence) it can also be fitted to a given DNA sequence. More precisely, the parameters of the DAR(p) process can be estimated from a given sequence using a Yule–Walker-type formalism. For details on this estimation process, see Dehnert et al. (2003). Compared to the mutual information function this second method of quantifying the strength of correlations at a certain distance is significantly superior, because any random background in the sequence is effectively eliminated from this correlation strength with the help of the parameter ρ.

Consequently, we have two different methods for quantifying short-range correlations of DNA sequences: the mutual information function, yielding a correlation vector $\{I(1),\ldots,I(p)\}$ and the DAR(p) process, where the correlation vector is given by the parameters $\{\alpha_1\ldots,\alpha_p\}$. In the following both measures of correlation are estimated from the sequence data. Throughout this investigation we have chosen $p=30$. For the p-dependence of our results, see (Dehnert et al. Bioinformatics, in preparation).

2.3. Distance matrix, clustering algorithms, and bootstrap replicates

The distance between two correlation curves can be measured by summing up the absolute differences in each component. For the DAR(p) correlation vector the distance is then given by

$$d_{a,b} = \| \vec{\alpha}^{(a)} - \vec{\alpha}^{(b)} \|_1 = \sum_{k=1}^{p} \left|\alpha_k^{(a)} - \alpha_k^{(a)}\right| \tag{3}$$

where $\vec{\alpha}^{(s)} = \left(\alpha_1^{(s)}, \cdots, \alpha_p^{(s)}\right)$ denotes the correlation curve of a chromosome s and $\|\cdot\|_1$ denotes the L_1 norm of the difference vector. By calculating all pairwise distances of correlation curves one obtains a distance matrix. Clustering trees have been generated from such distance matrices with the UPGMA algorithm using the software package PHYLIP (Felsenstein, 2004). Bootstrap replicates have been obtained by randomly deleting 20% of pairs of components entering the computation of $d_{a,b}$. The tree shown in Fig. 2 has been calculated as a 50% majority-rule (extended) consensus trees (CONSENS, PHYLIP package) and displayed using the software tool TREEVIEW (Page, 1996). A tree obtained by a neighbour-joining procedure (with *C. elegans* as an out-group) shows the same features of chromosome clustering, similarly high bootstrap values and the same quality of species distinction as the tree given in Fig. 2.

2.4. Tree colour coding

For comparing clustering trees we sort each tree in a universal way respecting its topology and, furthermore, we choose an efficient one-dimensional graphical representation of the sorted tree. Our representation of a clustering tree as a line of colours (tree colour coding, TCC) for assessing its internal order is based upon the following sorting algorithm: the branches on each level of the tree are sorted alphabetically, with each branch being identified by the alphanumerically lowest element (chromosome identifier) it contains. Then the order of chromosomes as they appear in the Newick representation of the sorted tree is taken and all chromosomes of a species are labeled with the same colour. This sorting algorithm slightly overestimates the overall order in the tree, as different branches containing chromosomes of the same species can become direct neighbours in the colour line, even if one of them also contains chromosomes of another species. For example, if branch 1 contained only elements of type a while its neighbour branch 2 contained both a's and b's, one or more of the a's in branch 2 will be sorted to lie adjacent to the a's of branch 1. Thus the a's would look exaggeratedly dense in the one-dimensional line of colours. As this effect is of the same order of magnitude for all trees we analysed, the TCC plot nevertheless provides a convenient tool for comparing the quality of chromosome sorting of different trees.

2.5. t-Value

To measure the difference of two species' correlation vectors at a certain index position k we use the absolute value of the t-statistic, i.e.

$$t_k(A,B) = \frac{\bar{\alpha}_k(A) - \bar{\alpha}_k(B)}{\sqrt{\frac{\sigma_k^2(A)}{n(A)} + \frac{\sigma_k^2(B)}{n(B)}}} \tag{4}$$

where, for a fixed index position k,$\bar{\alpha}_k(S)$ denotes the mean and $\sigma_k^2(S)$ the variance calculated over all $n(S)$ chromosomes of species S which are included in the analysis. In the following we use the absolute values of

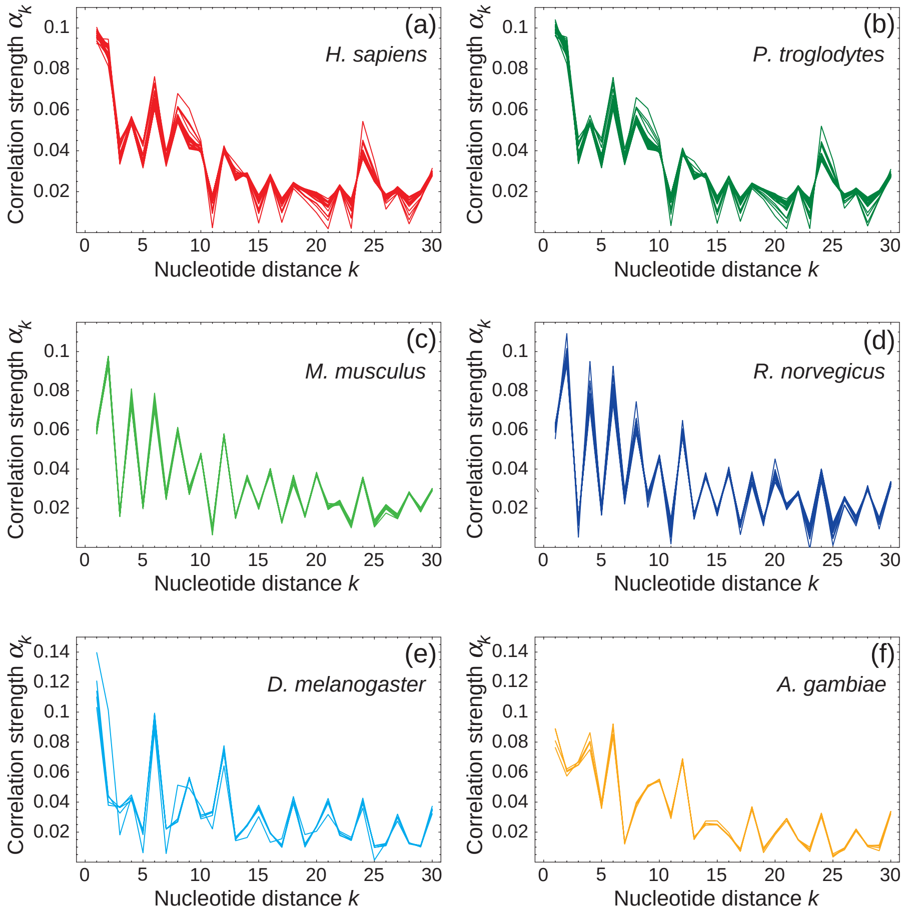

Fig. 1. Correlation curves for chromosomes of the following species (a) *H. sapiens* [22 curves], (b) *P. troglodytes* [23 curves], (c) *M. musculus* [19 curves], (d) *R. norvegicus* [20 curves], (e) *D. melanogaster* [5 curves], and (d) *A. gambiae* [4 curves]. In all cases the correlation curve is given by the parameter vector $\vec{\alpha}$ of a DAR(30) process.

$t_k(A, B)$, normalised to unity with respect to summation over k.

3. Results and discussion

3.1. Correlation curves

In the following the correlation strength of two nucleotides at a distance k is represented by the parameter vector $\vec{\alpha}$ of a DAR(p) process. Fig. 1 shows the correlation curves for distances $k=1,\ldots,30$ calculated for all chromosomes of six species. This figure already reveals the essence of our finding: all chromosomes of a single species follow essentially the same correlation curve. In fact, for the examples given in Fig. 1 even the degree of inter-species similarity shows a certain pattern. Systematic differences between two species' correlation curves seem to increase with the species' evolutionary distance. Later we will see the limit of this simple observation, but for the cases shown in Fig. 1 this phylogenetic aspect is rather striking.

In our previous investigation (Dehnert et al., submitted) we found that the correlation curves for sex chromosomes show a strong deviation from a species average correlation curve and, consequently, frequently lie at the edges of a species' cluster in the respective clustering tree. We therefore omitted them from the present analysis.

3.2. Clustering tree

In order to study this visual impression on a more quantitative level we compute pairwise distances of the correlation curves as described in Section 2.3. This yields a distance matrix on the correlation curves to which a clustering algorithm can be applied. The resulting clustering tree represents a very efficient method of condensing the information contained in the correlation curves into a single relational structure. In our previous work based upon only six species we found that essentially all chromosomes are automatically sorted into the appropriate species cluster (Dehnert et al., submitted). This remarkable feature of chromosome clustering corresponds to the high degree of intra-species synchrony of the correlation curves observed in Fig. 1. The clustering tree for 125 chromosomes of eight different species based upon the distance matrix obtained from the correlation curves is shown in Fig. 2. Note that no

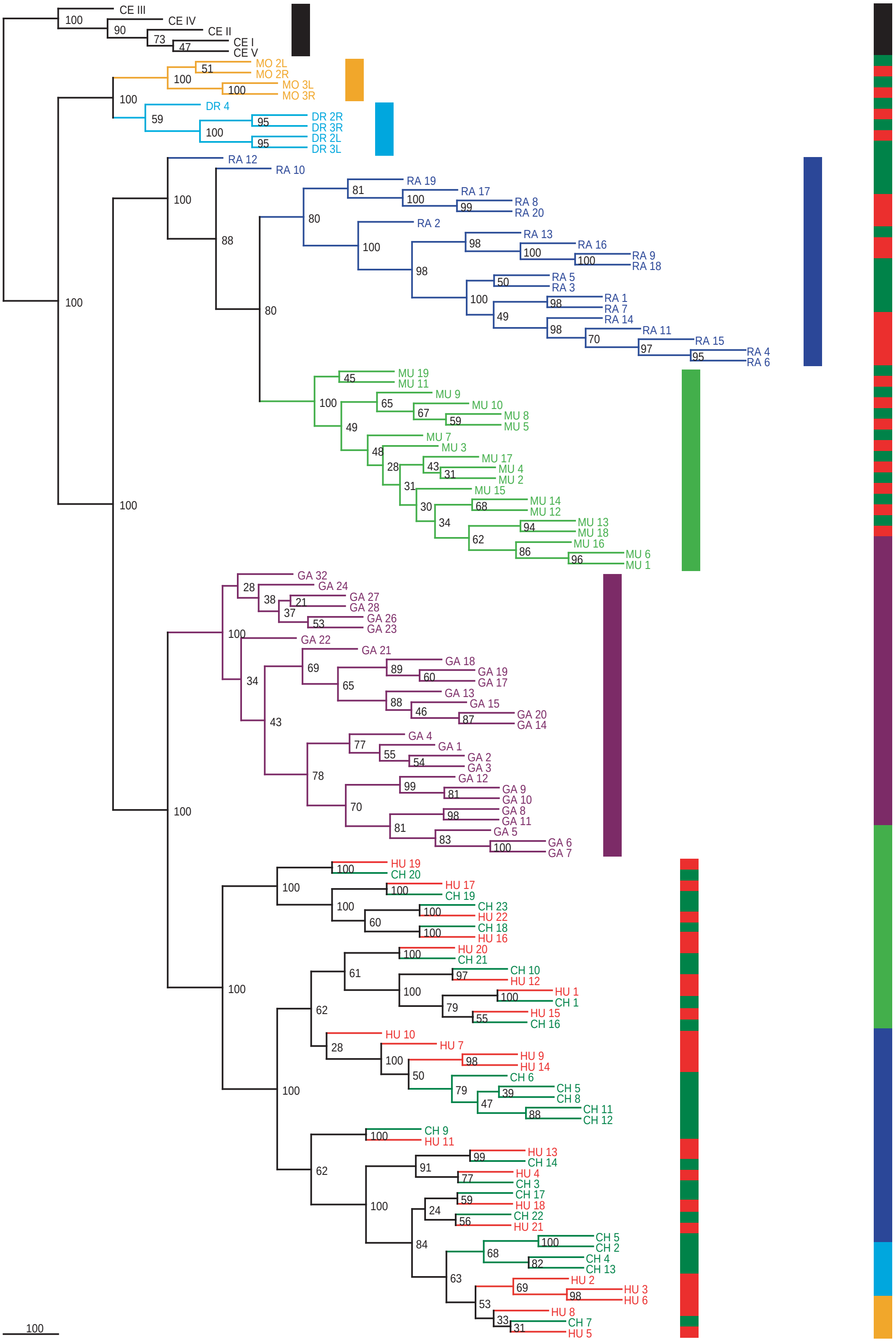

Fig. 2. Clustering tree for 125 chromosomes of eight eukaryotic species. The figure shows the consensus tree of 100 bootstrap replicates (for details see Methods). Numbers above branches indicate bootstrap values. By attributing a colour to each species, each cluster can formally be represented by a colour line fragment (shown on the right-hand side of each cluster). These colour line fragments are sorted via the TCC algorithm (see Methods) and lead to the colour line displayed on the right-hand side of the figure. The following species are included: *A. gambiae* (MO), *C. elegans* (CE), *D. melanogaster* (DR), *G. gallus* (GA), *H. sapiens* (HU), *M. musculus* (MU), *P. troglodytes* (CH), and *R. norvegicus* (RA). The number after the two letter abbreviation for the species indicates the number of the respective chromosome.

information other than the distance of correlation curves enters the clustering process. In particular, all chromosomes are treated as individual taxa. Both, the clustering of chromosomes pertaining to the same species and those aspects of the tree clearly in correspondence with evolutionary species differentiation emerge from the correlation curves alone. One observes that in almost all cases the chromosomes of a single species form a cluster of their own. The most obvious exception is the complete mixture of human and chimpanzee chromosomes. Clearly the correlation curves' capability of distinguishing between species stops at so small evolutionary distances. One important feature of Fig. 2 is that most of the human and chimpanzee chromosomes form pairs, which are mostly found to be robust with respect to changes in the range p of nucleotide distances and in sequence length (cf. also the discussion of Fig. 3). As far as we see this chromosome pairing cannot be immediately attributed to some quantitative feature of the sequence. The systematic comparison of biological features of chromosomes is, however, still at the beginning. Particularly, the human/chimpanzee chromosome mixing suggests studying certain pairs with methods of comparative genomics. The pairs HU 19/CH 20 and HU 17/CH 19 for example have very high bootstrap values in Fig. 2. Note that one of the chromosome pairs, HU 21 and CH 22, seems also be similar in a variety of biological aspects discussed in

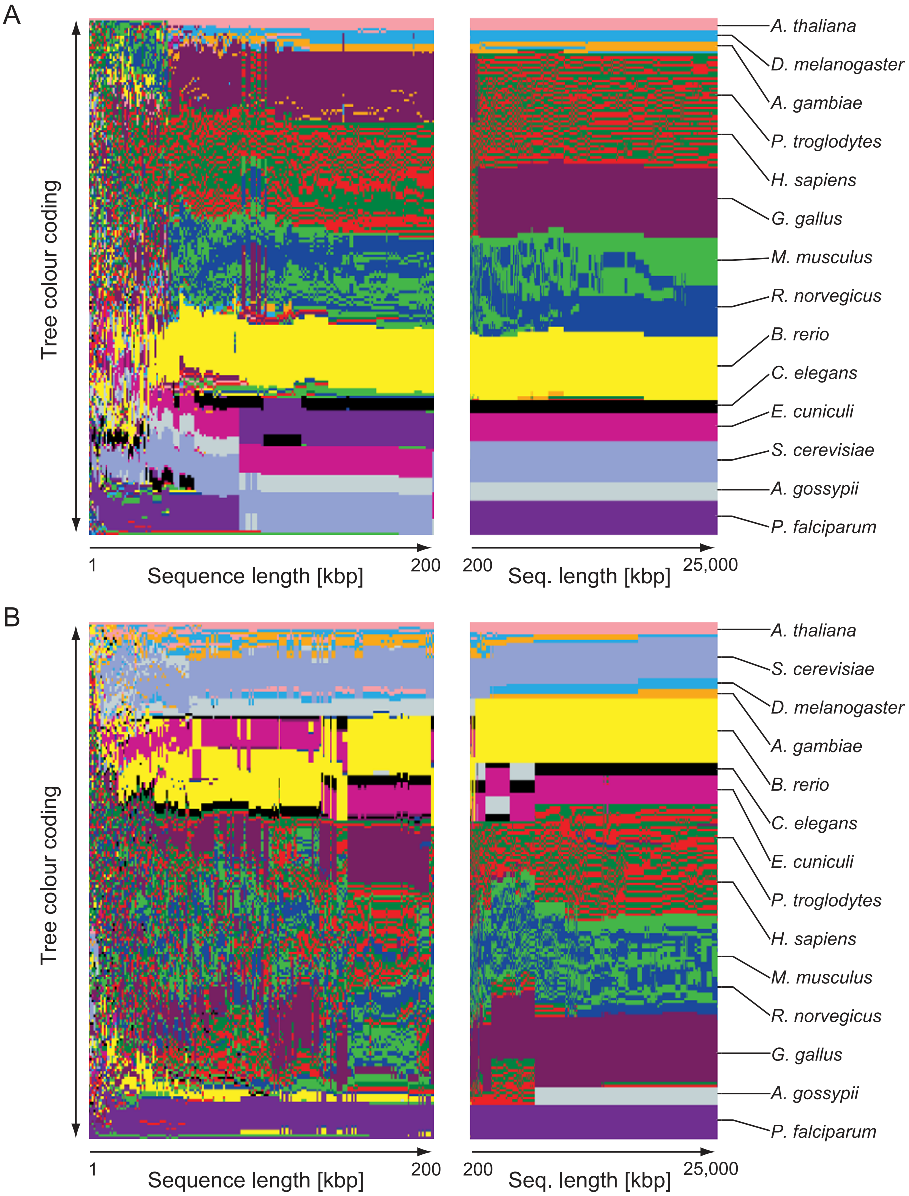

Fig. 3. Tree colour coding (TCC) plot for 14 eukaryote species for (A) the DAR(p) and (B) the $I(k)$ representations of the correlation vectors. The length of the underlying DNA sequences is varied. For each length a clustering tree is computed and then translated into a colour line with the TCC algorithm (cf. Fig. 2). Starting with the first 1 kbp of each of the 203 chromosomes of the 14 species the sequence lengths are simultaneously increased with step sizes 1 kbp (up to 200 kbp) and 10 kbp (up to 25 Mbp). In case of exceeding the length of a chromosome before reaching 25 Mbp the length is kept constant at the maximum possible length.

comparative genomics (Watanabe et al., 2004). Indeed, pairs of human and chimpanzee chromosomes often follow pairings obtained from gene homology content. We are currently studying this point more systematically. We expect such structural properties to be an important source of information on what biological features of a sequence determine a chromosome's position in the clustering tree and, consequently, a chromosome's correlation curve. In Fig. 2 we also find an example where the clustering tree based on correlation curve distances clearly deviates from a pure phylogenetic tree (a term strictly applicable only on the species level, not on the level of single chromosomes), as the position of chicken in spite of the high degree of clustering of its chromosomes is misplaced from a phylogenetic point of view. It should be pointed out that at this level, the clustering tree derived from the L_1-norm-based distances could be a too coarse representation of the correlation curves. In spite of their neighbouring positions the correlation curves of human and chicken display a vast amount of systematic differences over the whole range of nucleotide distances k (cf. Fig. 4, later in section 3.4). Nevertheless, these deviations from pure phylogeny may help reveal the biological properties of the chromosome sequences responsible for the high degree of synchrony between a species' correlation curves.

3.3. Dependence on sequence length

An important question is, how the chromosome clustering depends on the amount of underlying sequence information. This is interesting for example to understand if the correlation curves from Fig. 1, when computed, e.g., for a small sequence segment, could in principle serve as a means to identify the species this segment belongs to, as well as the most probable chromosome. In order to study the dependence of chromosome sorting on sequence length we developed a tool for monitoring the change of a clustering tree as a function of some parameter (see Methods). The idea of this tree colour coding (TCC) plot is to apply topologically allowed branch switches to bring a large set of trees (each tree belonging to a certain value of the parameter) as close to the same predefined (e.g. alphanumerical) order as possible and then colour-code the sequence of taxa (by assigning a single colour to all chromosomes of one species), resulting for each tree in a single colour line, which represents the chromosome sorting. In Fig. 2 the corresponding colour line is shown next to the clustering tree. On the level of the TCC plot the focus of attention is on chromosome clustering instead of on the detailed progression of the branching. It is, therefore, appropriate to include only species with more than, e.g., 4 chromosomes into the analysis, as then the order of some colour segment can be meaningfully evaluated.

At this level it is fruitful to compare the DAR(p) process and $I(k)$ as a measure for constructing clustering trees based on correlation in DNA sequences. Fig. 3 compares the two TCC plots. In both cases an increase in sequence length results in a significant increase of overall order of the colour lines representing the respective clustering trees. The chromosome clustering is seen by the colour homogeneity of each line of the TCC plot. In this direct comparison the superiority of the DAR(p) correlation vector compared to the correlation vector from the mutual information function is obvious. In particular the DAR(p) correlation vector at the highest sequence lengths included in the TCC plot is capable of fully separating the mouse (green) and rat (blue) chromosome clusters. For the mutual information correlation vector the highest sequence lengths shown in Fig. 3 are not sufficient to achieve this distinction even though an increase in clustering quality with increasing sequence length is observed. For the DAR(p) correlation vector it is seen that a remarkably small amount of data is necessary

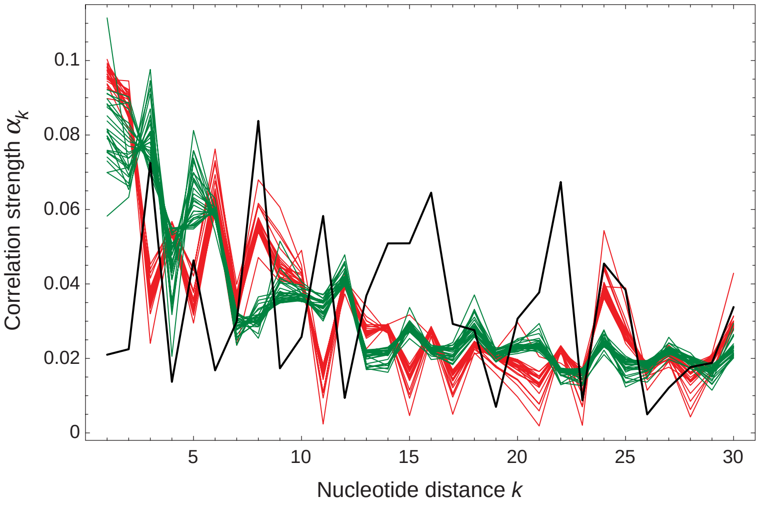

Fig. 4. Correlation curves [based upon the parameter vector $\vec{\alpha}$ of a DAR(30) process] for chromosomes of human (red, 22 curves, same as Fig. 1a) and chicken (dark green, 27 curves). The $|t|$-value curve (see Methods) is shown in black. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(approximately a few thousand bases) to establish already rather ordered clustering patterns. The single case where species distinction fails (in accordance with the clustering tree based on the full sequence information shown in Fig. 2) is the case of human and chimpanzee. It is, however, seen that with increasing sequence length stable horizontal stripes are marked out which are a clear indication of stable pairs of chromosomes being formed. Occasional jumps of whole blocks in the TCC plot are a side effect of the sorting algorithm, where a single outsider chromosome within a homogeneous cluster can define the alphanumeric label of this cluster and, consequently, induce a jump of this cluster, when the outsider chromosome, e.g., disappears from the cluster as sequence length is increased. For very short sequences (up to about 30 kbp) the estimation process of the correlation vector fails. This lack of convergence is clearly seen in the TCC plot, where the overall order of the colour line is lost spontaneously. This length scale can be viewed as the statistical limitation of our method: shorter sequences cannot be sorted on the basis of their correlation vector.

Note that the chromosome clustering in general (and particularly the distinction of mouse and rat) also depends on the range of nucleotide distances. For example, the sequence length, for which full distinction is achieved,

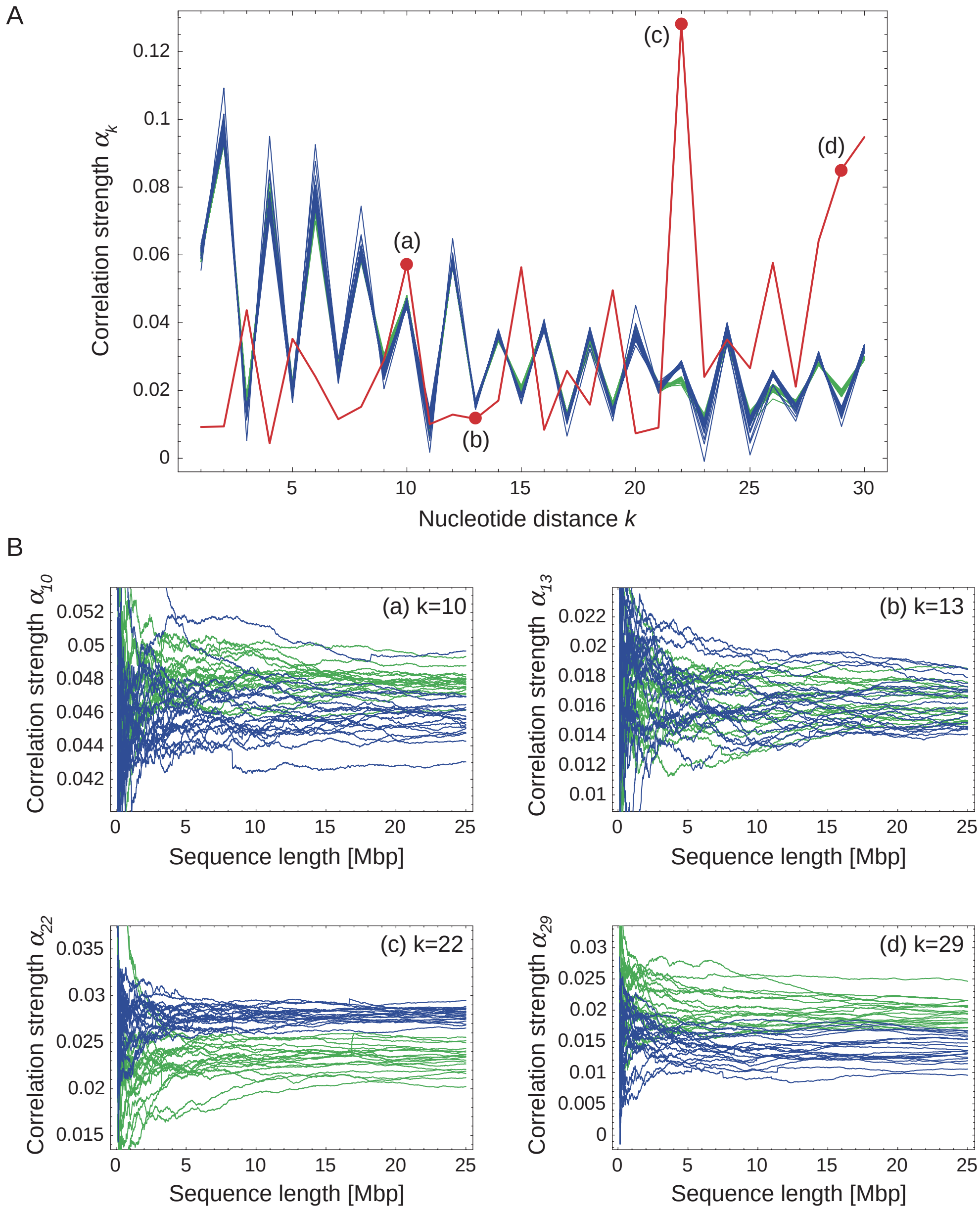

Fig. 5. (A) Correlation curves for chromosomes of mouse (green, 19 curves, same as Fig. 1c) and rat (blue, 20 curves, same as Fig. 1d), together with the |t|-value curve (red). A high (low) |t|-value indicates a large (small) contribution of the component αk to the distinction of the of mouse and rat, respectively. (B) Dependence of the correlation strength αk on sequence length for chromosomes of mouse and rat at distances (a) k=10, (b) k=13, (c) k=22, and (d) k=29. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

decreases substantially (i.e. the distinctive power of our method increases), when p is increased (Dehnert et al., in preparation).

3.4. A detailed look at the correlation curves

The correlation curves are the basis of our analysis. Already in Fig. 1 it is seen that the information, on which species distinction is based, is not uniformly distributedin the range of nucleotide distances k. One can therefore ask for which distances the correlation strength contributes the most to the distinction of two species. To measure the relative amount of species distinguishing information we use the absolute t-values, which are well-defined measure in statistics (see Methods). From Figs. 2 and 3 it is easy to identify interesting species pairs whose distinction can be studied using this elementary method. In Fig. 4 the distinction of human and chicken chromosomes is discussed. The striking aspect about Fig. 4 is that the proximity of the two species clusters in the tree shown in Fig. 2 does not necessarily correspond to a high degree of similarity in the corresponding correlation curves. It is seen in Fig. 4 that the two families of correlation curves display clear and systematic differences distributed over the whole range of nucleotide distances k. The $|t|$-value (black curve) as a function of distance k has pronounced peaks at those values of k where visual inspection of the two families of correlation curves would expect the largest amount of information for distinguishing the two species.

The interesting aspect about the second species pair namely mouse and rat is obvious from Fig. 3. Even for the DAR(p) correlation vector this pair of species needs the largest amount of sequence information to achieve a separation of the respective chromosomes. The two families of correlation curves (which are taken from Fig. 1b and c) almost coincide in Fig. 5A. Here the $|t|$-value curve (shown in red) provides a sensitive measure of the remaining systematic differences between the two families of correlation curves. In particular it is seen from the $|t|$-value curve that the information on which species distinction can be based is highly localised (cf. the peak at k=22) as compared to the other species pair discussed in Fig. 4. The $|t|$-value curve in Fig. 5A can serve as a rapid guideline for selecting particular values of k for which the dependence of mouse/rat distinction on sequence length can now be investigated on the level of individual entries in the correlation vector. This is shown in Fig. 5B. There the dependence of individual components α_k on the length of the underlying sequence is shown for the maximal $|t|$-value (c), two intermediate $|t|$-values (a and d), and a very low $|t|$-value (b). In all four cases a significant stabilisation with increasing sequence length is observed (most curves seem to approach constant values; fluctuations and large deviations from this value are reduced). It is also seen that the distinction of the two families of values is not possible in all four cases. Indeed, the $|t|$-value is a good indicator as to when separation can be achieved. Most notably in the case of the largest $|t|$-value the distinction between the two families persists down to very short sequences. From our studies of $|t|$-value curves so far pairs of species seem to fall into different categories: one type, where the distinctive information is highly localised in a few components of the correlation curve, and a second type, where this information is distributed over a wide range of values in k.

4. Conclusions

We have shown that the pattern of short-range statistical correlations between nucleotides is highly similar for the chromosomes of a species. This intra-species similarity is accompanied by systematic inter-species differences allowing species distinction with very high accuracy, unless evolutionary distance is small. Within the present paper we studied the systematics of both, the chromosome clustering based upon this synchrony in correlation patterns and the systematics of species distinction with this pattern.

Due to the wider scope of species this extension of our previous work shows clear and robust deviations from a purely phylogenetic interpretation, when the correlation patterns are translated into inter-species relations. Furthermore, with human and chimpanzee this selection of species contains one example, where species distinction seems no longer possible, even at full sequence length. We expect that these cases will point towards the specific biological properties shaping the correlation curves and will, ultimately, reveal the mechanism for the high degree of synchrony in the correlation patterns of a species' chromosomes. How can statistical studies of sequence properties help reveal the biological mechanisms responsible for the intra-species synchrony of the correlation curves? One key strategy is to identify (biologically or statistically) well-defined sequence components, eliminate them (e.g., by omitting them from the sequence or by substituting them with random segments) and then assess the impact of this operation on the correlation curves. For two such components, namely CpG islands and Alu repeats, we recently performed such an investigation (Dehnert, Helm, Hütt, in preparation).

The TCC plot allows us to study the dependence of chromosome clustering on the length of the underlying sequences. In addition, it enables us to efficiently compare our two methods for quantifying correlations (mutual information function and p-th order Markov process). Particularly for an analysis based upon the DAR(p) correlation vector some chromosome clustering is found already for rather short sequences (a few thousand nucleotides). It is, therefore, possible that refining the method of estimating correlations even further may lead to an efficient and fast tool for identifying large (more than a few thousand nucleotides) insertions into a chromosome sequence.

References

Berryman, M., Allison, A., Abbott, D., 2004. Mutual information for examining correlations in DNA. Fluctuation Noise Letters 4, 237–246.

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

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

Ebeling, W., Freund, J., Schweitzer, F., 1998. Komplexe Strukturen: Entropie und Information. Teubner, Stuttgart.

Felsenstein, J., 2004. PHYLIP (Phylogeny Inference Package) version 3.6 (alpha3) Distributed by the author. Department of Genome Sciences, University of Washington, SE.

Garte, S., 2004. Fractal properties of the human genome. J. Theor. Biol. 230, 251–260.

Gentles, A.J., Karlin, S., 2001. Genome-scale compositional comparisons in eukaryotes. Genome Res. 11, 540–546.

Grosse, I., Herzel, H., Buldyrev, S.V., Stanley, H.E., 2000. Species independence of mutual information in coding and noncoding DNA. Phys. Rev., E 61, 5624–5629.

Grosse, I., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J., Stanley, H., 2002. Analysis of symbolic sequences using the Jensen–Shannon divergence. Phys. Rev., E 65, 041905.

Hao, B., Qi, J., 2003. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. IEEE proceedings of the computational systems bioinformatics.

Herzel, H., Ebeling, W., Schmitt, A., 1994. Entropies of biosequences: the role of repeats. Phys. Rev., E 50, 5061–5071.

Holste, D., Grosse, I., Buldyrev, S., Stanley, H., Herzel, H., 2000. Optimization of coding potentials using positional dependence of nucleotide frequencies. J. Theor. Biol. 206, 525–537.

Holste, D., Grosse, I., Beirer, S., Schieg, P., Herzel, H., 2003. Repeats and correlations in human DNA sequences. Phys. Rev., E 67, 061913.

Jacobs, P., Lewis, P., 1978. Discrete time series generated by mixtures III: autoregressive processes (DAR(p)). Tech. Rep. NPS55-78-022, Naval Postgraduate School, Monterey, California.

Jacobs, P., Lewis, P., 1983. Stationary discrete autoregressive-moving average time series generated by mixtures. J. Time Ser. Anal. 4, 19–36.

Karlin, S., Brendel, V., 1993. Patchiness and correlations in DNA sequences. Science 259, 677–680.

Karlin, S., Ladunga, I., 1994. Comparisons of eukaryotic genomic sequences. PNAS 91, 12832–12836.

Karlin, S., Mrázek, J., 1997. Compositional differences within and between eukaryotic genomes. PNAS 94, 10227–10232.

Li, W., Kaneko, K., 1992. Long-range correlation and partial $1/f^{\alpha}$spectrum in a noncoding DNA sequence. Europhys. Lett. 17, 655–660.

Nicolay, S., Argoul, F., Touchon, M., d'Aubenton Carafa, Y., Thermes, C., Arneodo, A., 2004. Low frequency rhythms in human DNA sequences: a key to the organization of gene location and orientation? Phys. Rev. Lett. 93, 108101.

Nikaido, M., Rooney, A.P., Okada, N., 1999. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: hippopotamuses are the closest extant relatives of whales. PNAS 96, 10261–10266.

Ouyang, Z., Wang, C., She, Z.-S., 2004. Scaling and hierarchical structures in DNA sequences. Phys. Rev. Lett. 93, 078103.

Page, R.D.M., 1996. TREEVIEW: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. 12, 357–358.

Peng, C.-K., et al., 1992. Long-range correlations in nucleotide sequences. Nature 356, 168–170.

Qi, J., Wang, B., Hao, B., 2004. Whole genome prokaryote phylogeny without sequence alignment: a K-string composition approach. J. Mol. Evol. 58, 1–11.

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

Shimamura, M., et al., 1997. Molecular evidence from retroposons that whales form a clade within even-toed ungulates. Nature 388, 666–670.

Snel, B., Bork, P., Huynen, M.A., 1999. Genome phylogeny based on gene content. Nat. Genet. 21, 108–110.

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

Watanabe, H., et al., 2004. DNA sequence and comparative analysis of chimpanzee chromosome. Nature 429, 382–388.