Genome Phylogeny Based on Short-Range Correlations in DNA Sequences

MANUEL DEHNERT,1 RAINER PLAUMANN,2 WERNER E. HELM,2 and MARC-TH. HÜTT1

ABSTRACT

The surprising fact that global statistical properties computed on a genomewide scale may reveal species information has first been observed in studies of dinucleotide frequencies. Here we will look at the same phenomenon with a totally different statistical approach. We show that patterns in the short-range statistical correlations in DNA sequences serve as evolutionary fingerprints of eukaryotes. All chromosomes of a species display the same characteristic pattern, markedly different from those of other species. The chromosomes of a species are sorted onto the same branch of a phylogenetic tree due to this correlation pattern. The average correlation between nucleotides at a distance $k$ is quantified in two independent ways: (i) by estimating it from a higher-order Markov process and (ii) by computing the mutual information function at a distance $k$. We show how the quality of phylogenetic reconstruction depends on the range of correlation strengths and on the length of the underlying sequence segment. This concept of the correlation pattern as a phylogenetic signature of eukaryote species combines two rather distant domains of research, namely phylogenetic analysis based on molecular observation and the study of the correlation structure of DNA sequences.

Key words: information theory, eukaryote genomes, species distinction.

1. INTRODUCTION

A significant progress in phylogenetic analysis in recent years was brought about by discussing specific events in the noncoding parts of the DNA sequence, namely, the spread of retroposons (or so-called interspersed elements) throughout the genome with time (Nikaido et al., 1999). Until now, however, few attempts have been made to address phylogenetic questions from a genome-wide perspective (Nikaido et al., 1999; Karlin and Ladunga, 1994; Karlin and Mrázek, 1997; Gentles and Karlin, 2001; Abe et al., 2003; Rokas et al., 2003). Karlin and Mrázek (1997) have analyzed the systematic differences in dinucleotide frequencies within and between species and obtained a biologically plausible phylogenetic tree for mitochondrial and nuclear genomes. Abe et al. (2003) found that a special class of neural networks (so-called self-organizing maps) can learn to distinguish species on the basis of oligonucleotide frequencies.

1Bioinformatics Group, Department of Biology, Darmstadt University of Technology, D-64287 Darmstadt, Germany.
2Mathematics and Science Faculty, University of Applied Sciences, D-64295 Darmstadt, Germany.

545
Snel et al. (1999) have studied the distribution of genes along a chromosome as an observable for unicellular species and obtained a high correlation with the 16S rRNA phylogenetic tree which constitutes an important standard of molecular phylogeny (Woese and Fox, 1977). Extracting phylogenetic properties from genomewide statistical observables has also been attempted recently for prokaryotes by Qi et al., who analyzed asymmetries in the distribution of “words” of length \(n\) (the so-called \(n\)-word distribution) (Qi et al., 2004; Hao and Qi, 2003). An alternative method for efficiently condensing genomewide information is given by the average correlation between two nucleotides with a distance \(k\) in the DNA sequence. A large amount of research has focused on long-range correlations in DNA up to distances of thousands of basepairs (Li and Kaneko, 1992; Peng et al., 1992; Voss, 1992; Karlin and Brendel, 1993; Buldyrev et al., 1995; Holste et al., 2003). Short-range correlations received far less attention. The two most important contributions to correlations at distances up to a few tens of basepairs are period-three oscillations of the correlation strength, which characterize coding regions, and oscillations with a period around 10 and 11 basepairs, which correspond to DNA bending properties and are a reflection of the double-helical structure (Trifonov and Sussman, 1980; Trifonov, 1998). Information theory has turned out to be a particularly convenient framework for investigating such properties. The period-three oscillations lead to pronounced differences of, e.g., the mutual information function at small distances for coding and noncoding DNA segments, respectively. It was observed that these differences are species independent (Grosse et al., 2000). Recently, two peaks of the mutual information function around a distance of a few hundred basepairs have been related to the internal structure of Alu repeats (Holste et al., 2003).

2. RESULTS

We show that, on the scale of chromosomes, the distance dependence of nucleotide correlations serves as a fingerprint of that particular eukaryote species and that phylogenetic information can be retrieved from this correlation pattern. A convenient representation of the correlation strength of two nucleotides at a distance \(k\) in a DNA sequence is given by the value of the mutual information (MI) function \(I(k)\) at this particular value of \(k\) (cf. Methods). A more refined form of quantifying correlations up to distances \(k_{\text{max}}\) of a few tens of nucleotides has recently been described (Dehnert et al., 2003). There, a Markov process of the order \(k_{\text{max}}\) is used to approximate the DNA sequence. The parameters \(\{\alpha_k, k = 1, \ldots, k_{\text{max}}\}\) of this Markov process (representing the “memory” of the sequence) then give a good quantitative estimate of the correlation strength at a distance \(k\). We focus on the results for the Markov representation. Results for the MI representation are found to be similar to the Markov representation, with the latter, however, yielding a slightly better phylogenetic reconstruction. Figure 1A–C shows the strength of correlation as a function of the nucleotide distance \(k\) in the Markov representation for three different species.

The diagram of each species contains the individual correlation curves for all chromosomes. It is seen that (i) different chromosomes follow the same qualitative curve and (ii) the three different species display strong qualitative differences in the correlation curves. When performing a principal components analysis on the basis of these curves, we find three separate clusters of points in a plane consisting of the first two principal components (Fig. 1D). This two-dimensional analysis quickly reaches its limits when more species are taken into account. One then has to pass to a pairwise comparison of the correlation curves. In the following, we use the sum of absolute values of the differences in each component of the correlation vector as a robust measure for the distance of two correlation curves. This leads to a distance matrix involving all chromosomes of the six species as taxa, which in turn gives rise to a phylogenetic tree. An important implicit test of our analysis is to regard the different chromosomes of each species as individual taxa. Figure 2 shows the resulting phylogenetic tree (a term strictly applicable only on the species level, not on the level of single chromosomes) of six species.

It is seen that these chromosomes form clusters according to their species identity. Except for the X chromosomes of mouse and rat and two additional chromosomes of the rat, we find a complete sorting of the chromosomes to the corresponding species cluster. We have, indeed, seen that in the correlation curves the X and Y chromosomes often display a slight but systematic deviation from the pattern followed by all other chromosomes of a particular species. Apart from a phylogenetically plausible relative position of the different species in the resulting tree, the degree of chromosome sorting is the most important observable for assessing the quality of this phylogenetic reconstruction.
SHORT-RANGE CORRELATIONS

FIG. 1. Examples of correlation curves for the Markov representation. The correlation vectors \( \{ \alpha_k, k = 1, \ldots, k_{\text{MAX}} \} \) obtained by the Markov representation are displayed for (A) all human chromosomes [24 curves], (B) all mouse chromosomes [20 curves] and (C) all Drosophila chromosomes [6 curves]. The first two principal components of these 30-dimensional correlation vectors are calculated and displayed in (D).

Next we evaluate the robustness of the phylogenetic tree with respect to internal parameters of our method. To some extent, robustness of a particular branch in the tree under such a parameter change also implies a high consensus of this branch. As a further support of this point, we apply conventional bootstrapping methods (Efron and Tibshirani, 1993) directly. This procedure leads to the bootstrap probabilities indicated at each node in Fig. 2. When performing the same analysis for a larger tree by adding the chromosomes of nine additional species, this chromosome sorting persists, with only 5 out of 124 chromosomes being sorted inaccurately (i.e., not into their respective species cluster; data not shown). In order to systematically study the change of overall order of a phylogenetic tree with internal parameters, we assign a specific color to all chromosomes of a species and then represent the tree as a sequence of colors (a tree color coding, TCC, plot). Due to the translation algorithm (described in Methods), this color representation shows whether all leaves (representing the chromosomes) of the same species can be brought into a consecutive order without altering the topology of the tree. All other branching information is discarded in this translation process. We find that even though possible errors on higher levels have still to be analyzed by visualizing the full phylogenetic tree, small alterations at the lowest branch level, as well as the overall degree of chromosome sorting can be efficiently monitored by following the line of colors as a function of some parameter.

From a practical point of view the most interesting parameter of our phylogenetic reconstruction method is the length of the DNA sequence entering the analysis. The phylogenetic tree shown in Fig. 2 has been obtained from the full sequence of each chromosome. The corresponding tree color coding line is shown as part of the inset in Fig. 2. In Fig. 3 we show with the same color coding the dependence of chromosome sorting on the length of the sequence.

While all our tests confirm that over a wide range of parameter constellations the same degree of ordering is observed, varying the sequence length provides the clearest insight into the mechanisms of our investigation. For this particular test, we kept the sequence lengths of all taxa identical and started from the beginning of each sequence. It is seen that even down to 100 kbp a certain order exists, although an appropriate chromosome sorting for the most closely related species in our analysis, namely mouse and rat, requires consideration of much longer sequences. This distinction between the chromosomes of mouse
FIG. 2. Clustering tree for six eukaryote species. The main part of the figure shows the 50% majority-rule (extended) consensus tree of 100 bootstrap replicates, where bootstrap replicates have been obtained as described in Methods. Numbers above branches indicate bootstrap values. For further reference on the original branch lengths, the inset shows branch lengths obtained from the UPGMA method. Branches are of the same order as in the consensus tree. In the inset, branch labels have been omitted for visual clarity. In the corresponding tree color coding line, it is seen how the TCC representation reflects the actual chromosome sorting of the tree. Due to the alphabetical sorting of the chromosomes in the TCC representation, RA (blue) and MU (green), as well as MO (orange) and DR (turquoise), exchange their position in the tree color coding line in relation to the clustering tree. The inaccurately sorted chromosome MU X remains clearly visible. The following species are included: *Anopheles gambiae* (African malaria mosquito (MO)); *Caenorhabditis elegans* (CE); *Drosophila melanogaster* (fruit fly (DR)); *Homo sapiens* (human (HU)); *Mus musculus* (mouse (MU)); *Rattus norvegicus* (rat (RA)). The number after the two-letter abbreviation for the species indicates the number of the respective chromosome.
FIG. 3. Tree color coding plot for 15 eukaryote species. The influence of the length of the underlying DNA sequence on the chromosome sorting of the tree is displayed. Starting with the first 1 kbp of each of the 124 chromosomes of 15 eukaryote species, the sequence lengths are simultaneously increased in steps of 1 kbp up to 25,000 kbp. In case of exceeding the length of a chromosome before reaching 25,000 kbp, the length is kept constant at the maximum possible length.
and rat, which is clearly the most demanding item in this task of phylogenetic reconstruction, also brings out significant differences between our two representations of the correlation strength. Figure 3 shows that a mixing (i.e., a lack of sorting) of chromosomes of these two species persists up to very high sequence lengths (about 10 Mbp), much longer than for the other species in our analysis. In the corresponding figure for the MI representation (data not shown), a comparable degree of order appears only at much higher sequence lengths (about 21 Mbp). An adequate way of comparing the two representations is to look at the quality of chromosome sorting in an even more condensed form given by a single number for each tree, namely, the tree ordering parameter (cf. Methods for the definition of that parameter) as a function of sequence length. This is done in Fig. 4A.

The persistent lack of distinction between mouse and rat chromosomes in the MI representation yields systematically lower values of the corresponding tree ordering parameter, which in this way illustrates each representation’s capacity to achieve this distinction. Another parameter of our method is the correlation range or the length $k_{\text{max}}$ of the correlation vectors displayed in Figs. 1A–C. Figure 4B gives a qualitative impression of this parameter dependence. Already for small $k_{\text{max}}$ (e.g., $k_{\text{max}} = 5$), we find a high degree of order in the resulting phylogenetic tree. Values of $k_{\text{max}}$ with $k_{\text{max}} > 12$ already lead to almost the same quality of phylogenetic reconstruction as in the case of $k_{\text{max}} = 30$, which we have discussed in detail above. In the range $15 < k_{\text{max}} < 30$, changes in the corresponding phylogenetic trees mostly occur at the level of outmost branches, i.e., as intraspecies switches, and are, therefore, not accounted for by the tree ordering parameter displayed in Fig. 4B. The corresponding tree color coding plot (shown as an inset in Fig. 4B) confirms the impression of a highly stable tree for $k_{\text{max}} > 15$.

3. DISCUSSION

In spite of the various tests of robustness of our phylogenetic tree, from our point of view, the most impressive display of methodological stability is given by the fact that the chromosomes of each species are clustered, even though they enter the analysis as individual taxa. In certain cases, the pronounced lengths of subbranches, as well as a stability of such patterns with respect to parameter changes and to different methods of quantifying correlations, imply that the relative positions of chromosomes within a single species as reconstructed by our method contains further biological information.

On the basis of our results, it is now possible to explore the quantitative influence of certain functionally or statistically well-defined elements of the sequence on these correlation strengths and, consequently, to assess their relevance for a phylogenetic reconstruction. Even though dinucleotide frequencies are observed to contain a significant amount of phylogenetic information, an attempt to link this statistical observation to specific functional components of the sequence requires taking into account correlations at larger distances. We expect this additional degree of freedom, the amount of phylogenetic information in each component of the correlation vector, to be an important item in identifying the functional components in the sequence...
and, subsequently, the biological mechanism responsible for the existence of phylogenetic information on the global, genomewide level. Repetitive elements in the noncoding parts of the genome, as well as the rates of their amplification, serve as markers (or characteristics) of evolutionary divergence. The resulting distributions, e.g., of Alus and SINEs, as well as of pseudogenes throughout a genome, are a source of biological, particularly phylogenetic, information (Batzer and Deininger, 2002; Hamdi et al., 1999). Empirical evidence for interspecies differences between these distributions is just beginning to emerge (Zhang et al., 2004). Linking these findings to phylogeny on a quantitative level most probably requires studying compositional information in the sequence beyond the dinucleotide level. With our analysis of the distance dependence of genomewide correlations, we provided a step in this direction. An intriguing remaining question is, which mechanism synchronizes the correlation pattern of chromosomes leading to this remarkable degree of similarity within the chromosomes of a species.

4. METHODS

Sequence data

DNA sequences have been obtained from public databases (e.g., www.ebi.ac.uk/genomes/eukaryota.html). All sequence data have been downloaded in October 2003. The percentages of unidentified nucleotides (usually marked with “N” in the downloaded sequences) vary between species and chromosomes. Unidentified nucleotides have been discarded for the analysis. By inserting the “N”s of randomly selected chromosomes into chromosomes of other species (i.e., by discarding the nucleotides at these positions), we checked explicitly that differences in numbers and distributions of unidentified nucleotides do not influence the clustering analysis.

Correlation vectors

The mutual information function $I(k)$ has been computed by using standard estimates of pair probabilities $p^{(k)}(i, j)$ of observing nucleotides $i$ and $j$ at a distance $k$. One then has $I(k) = \sum_{i,j \in S} p^{(k)}(i, j) \log_4 \left( \frac{p^{(k)}(i, j)}{p(i)p(j)} \right)$, where $S$ denotes the set of nucleotide states, $S = \{A, C, G, T\}$ and $p(i)$ is the estimated probability of observing the single nucleotide $i \in S$. For calculation of the Markov representation, we apply an estimation scheme (Jacobs and Lewis, 1978, 1983) which has been recently adapted to DNA sequences (Dehnert et al., 2003). This scheme involves numerically solving the Yule–Walker equations based upon a discrete autoregressive process of order $k_{max}$ with an empirical autocorrelation function as an estimator. The resulting vector $(\alpha_k, k = 1, \ldots, k_{max})$ constitutes the Markov representation (of order $k_{max}$). Due to the incapacity of such a process to represent negative correlations (Jacobs and Lewis, 1978), some of the components $\alpha_k$ can become negative and, therefore, it is no longer possible to interpret them as probabilities. This does not affect the phylogenetic reconstruction. In general terms, our approach separates the intrinsic correlation within a chromosome sequence from its general random background. The Markov representation contains only this intrinsic component. This focus constitutes the main reason for the superiority over the MI representation based on the mutual information function. Let $\vec{v}^{(a)} = (v_{1}^{(a)}, v_{2}^{(a)}, \ldots, v_{k_{max}}^{(a)})$ denote the correlation vector of taxon $a$. Depending on the representation, the components $v_{k}^{(a)}$ are given by $\alpha_k$ or by $I(k)$. The distance between taxa $a$ and $b$ is measured by $d_{ab} = \| \vec{v}^{(a)} - \vec{v}^{(b)} \|_1$, i.e., summing the absolute values of differences in each component ($L_1$ norm).

Phylogenetic analysis

Trees have been generated from distance matrices with the UPGMA algorithm using the software package PHYLIP (Felsenstein, 2004). The UPGMA algorithm and the corresponding dendograms are explained in detail in (Kaufman and Rousseeuw, 1990) where the combination of $L_1$ norm and UPGMA is recommended as a robust clustering technique. We checked that qualitatively the same tree as in Fig. 2 is obtained with the neighbor-joining algorithm (Saitou and Nei, 1987) using C. elegans as an outgroup. We use a schematic tree display in the main part of Fig. 2, and we use average distances between clusters as calculated by the UPGMA method in the inset. Bootstrap replicates have been obtained by randomly deleting 20% of pairs of components $(v_{i}^{(a)}, v_{i}^{(b)})$ entering the computation of $d_{ab}$. Using the software
component CONSENS in the PHYLIP package, a 50% majority-rule (extended) consensus tree has been obtained and displayed using the software tool TREEVIEW (Page, 1996).

Tree color coding

Our representation of such a tree as a line of colors (tree color 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 lexicographically lowest element (chromosome) 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 color. This sorting algorithm slightly overestimates the overall order in the tree, as different branches containing chromosomes of the same species can become direct neighbors in the color 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 neighbor 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 colors. As this effect is of the same order of magnitude for all trees we analyzed, the TCC plot nevertheless provides a convenient tool for comparing the quality of chromosome sorting of different trees. In certain cases, it is convenient to further condense such a TCC plot into a single curve. This can be achieved by counting the numbers of neighboring identical colors in each line of the TCC plot. This quantity, when properly normalized, gives an estimate of the tree order. Let $M_1$ be the maximum number of color changes in a TCC line (i.e., the number of chromosomes minus one), $M_2$ the minimum number of color changes (which corresponds to the number of species minus one), and $A$ the number of actual color changes in a specific TCC line. Then the tree ordering parameter $O(A)$ is defined by

$$O(A) = \frac{(M_1 - A)}{(M_1 - M_2)}.$$  

The quantity $O(A)$ always lies between 0 (maximal disorder) and 1 (maximal order).

REFERENCES


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


Address correspondence to:
Marc-Th. Hütt
Bioinformatics Group
Department of Biology
Darmstadt University of Technology
D-64287 Darmstadt, Germany

E-mail: huett@bio.tu-darmstadt.de