Informational structure of two closely related eukaryotic genomes

Manuel Dehnert,[1] Werner E. Helm,[2] and Marc-Thorsten Hütt[1]
[1]*Computational Systems Biology, School of Engineering and Science, International University Bremen, Campus Ring 1, D-28759 Bremen, Germany*
[2]*Mathematics and Science Faculty, University of Applied Sciences, D-64295 Darmstadt, Germany*
(Received 23 January 2006; published 15 August 2006)

Attempts to identify a species on the basis of its DNA sequence on purely statistical grounds have been formulated for more than a decade. The most prominent of such genome signatures relies on neighborhood correlations (i.e., dinucleotide frequencies) and, consequently, attributes species identification to mechanisms operating on the dinucleotide level (e.g., neighbor-dependent mutations). For the examples of *Mus musculus* and *Rattus norvegicus* we analyze short- and intermediate-range statistical correlations in DNA sequences. These correlation profiles are computed for all chromosomes of the two species. We find that with increasing range of correlations the capacity to distinguish between the species on the basis of this correlation profile is getting better and requires ever shorter sequence segments for obtaining a full species separation. This finding suggests that distinctive traits within the sequence are situated beyond the level of few nucleotides. The large-scale statistical patterning of DNA sequences on which such genome signatures are based is thus substantially determined by mobile elements (e.g., transposons and retrotransposons). The study and interspecies comparison of such correlation profiles can, therefore, reveal features of retrotransposition, segmental duplications, and other processes of genome evolution.

DOI: 10.1103/PhysRevE.74.021913 PACS number(s): 87.10.+e, 87.14.Gg, 02.50.−r, 02.50.Ga

I. INTRODUCTION

The statistical properties of DNA sequences received a substantial amount of scientific attention in the last few years. In particular empirical distributions of various genomic elements have been studied [1–8]. At the same time, large effort went into the modeling of such distributions [9–12], as well as in the analysis of large-scale sequence properties with nonstandard tools [13–16].

From this theoretical perspective, the genome is a dynamically expanding object on an evolutionary scale. An important next step is to understand the effect that characteristics of genome evolution may have on statistical properties of a DNA sequence. Over the last decade short-range correlations in DNA sequences have proven quite informative in this respect. Starting from the early finding that coding and noncoding sequence segments possess mutual information functions with striking differences due to codon usage in the coding segments [17], an ever more detailed look at short-range correlations has allowed in specific incidences to relate correlation properties with biological function. Two important findings in this line of thought are the relation of 10-11 bp periodicities with DNA supercoiling [18] and the identification of the signature of Alu repeats as peaks in the correlation function [19]. Statistical properties of oscillations and fluctuations in DNA sequences are still a topic of intense research [15,20,21]. In addition to such features common to many species the correlation pattern as a whole also contains features that reflect species identity. Species comparison on the level of such global statistical properties so far mostly focused on differences in dinucleotide frequencies [22–24] and n-word distributions [25,26]. In all these cases investigations could be extended to provide rather robust algorithms for species distinction. The biological origin of species information being present on this large-scale statistical level is, however, far from being understood. While it is often argued that DNA repair mechanisms and certain species-dependent characteristics of tertiary structure of the DNA molecule may result in such systematic word count differences [27], the only candidate to account for such differences on a quantitative level currently seems to be neighbor-dependent mutations which are known to differ between species on purely biochemical grounds [28–31]. Particularly for the observed features of word distributions it is debated whether the key processes leading to these patterns are situated on the level of very few nucleotides (e.g., neighbor-dependent mutations [32], microsatellite expansion [33,34], or locally acting repair mechanisms [24]) or on a larger scale (e.g., longer repetivive elements, preferential insertion sites of mobile DNA, etc.). Here we apply a new method (introduced in [35]) for quantifying statistical correlations in DNA sequences in order to show that correlation-based species distinction *increases* with increasing distances. This strongly supports a view, where such correlations are determined by (and, consequently, reveal) properties of longer repetitive elements. In addition, the new method, which is based on a discrete autoregressive model, is compared to the mutual information function, which constitutes a standard approach for quantifying correlations in DNA sequences (see, e.g., [17,19]).

In two recent studies [36,37] we analyzed the average correlation of a symbol within a DNA sequence with another symbol at a distance k up to distances of a few tens of nucleotides. We find that these correlation profiles, when analyzed for a variety of eukaryotic species, display a high degree of intraspecies similarity and systematic interspecies differences. Intriguingly, these interspecies differences seem to increase with evolutionary distance, i.e., a cluster tree based upon distances of the correlation profiles sorts all chromosomes involved into fully separated species clusters within the tree and approximates the corresponding phylogenetic

tree of these species [36]. Species distinction on the basis of these correlation profiles fails at too small evolutionary distances. Human and chimpanzee chromosomes, for example, cannot be distinguished with this method. This analysis has been performed on the level of full chromosomal sequences. When passing to smaller segments, species identification deteriorates until (at sequence lengths of approximately 200 kbp) the corresponding cluster tree fails to provide species clusters [37]. The first pair of species, within this study, for which distinction breaks down with a decreasing sequence length, is the mouse (*Mus musculus*) and the rat (*Rattus norvegicus*). This pair, therefore, provides an ideal starting point to study the statistical information within a DNA sequence leading to this phenomenon of correlation-based species distinction. Within the present paper we, therefore, investigate mouse/rat distinction based on their respective correlation profiles. We analyze the informational structure of genomic sequences based on the correlation pattern as a function of nucleotide distances. By informational structure we understand the clustering structure resulting from similarities and differences of the (information-based) correlation patterns. In contrast to our previous investigations [36,37], where the aim was to look at the simultaneous distinction of a large number of species and to understand the systematics of this correlation observable by comparison with the species' phylogeny, we now focus on only two evolutionary rather close eukaryotic species, namely *Mus musculus* (mouse) and *Rattus norvegicus* (rat), and show how species identity encoded in such statistical sequence properties changes with the amount of sequence information and with the range of nucleotide distances considered. We argue that in contrast to the above-mentioned neighbor-oriented hypotheses the patterning of eukaryotic genomes by repetitive elements explains most of these statistical features. Our main finding supporting this view is that species distinction *increases* with an increased range of correlations taken into account. In order to further validate the involvement of repetitive DNA in these correlations we eliminated all annotated repeats from the sequence and observe that the species distinction on the basis of the new correlation curves disappears.

The basis of our correlation analysis is the parameter estimation process from [35]. A linearized version of this procedure is described in Appendix A. This estimation process gives access to the correlation parameters from a given DNA sequence. Differences between these parameters lead to a distance matrix on the level of chromosomes which in turn can be translated into a clustering tree. Studying the parameter dependence of the resulting clustering tree requires a tool for efficiently comparing two similar trees in the degree of clustering they provide. This is achieved with a sorting algorithm exploiting topologically allowed branch permutations. In Appendix B we summarize this algorithm, together with other methods, and show how it can be implemented by letting it act upon the Newick representation of such a clustering tree. Appendix C lists the download sites of the sequences. Section II describes our results, both on the level of the correlation curves (Sec. II A), and on the level of the clustering trees' parameter dependence (Sec. II B). The implications of our findings are discussed in Sec. III.

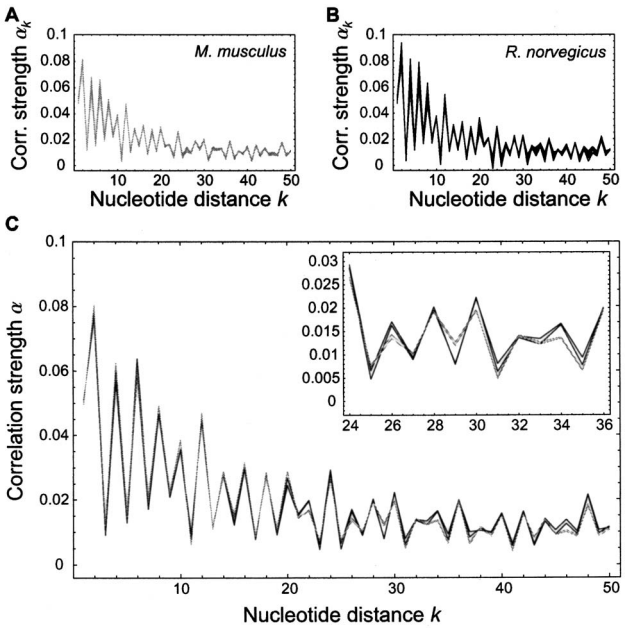

FIG. 1. Correlation curves for the Markov representation for $p=50$. (A) All chromosomes of *M. musculus* [19 curves], (B) All chromosomes of *R. norvegicus* [20 curves]. (C) Exemplary correlation curves for the two species overlaid in one diagram, namely chromosomes 1 and 2 of *M. musculus* (MU 1, MU 2) and *R. norvegicus* (RA 1, RA 2), respectively. Here and in the following, all information pertaining to *M. musculus* chromosomes is shown in gray, while its *R. norvegicus* counterpart is given in black. In all cases, sex chromosomes have been omitted from the analysis.

II. RESULTS

A. Correlation curves

In our previous studies, correlations were evaluated with a well-known tool from information theory, namely the mutual information function, and a new method based on higher-order Markov processes. We found that the Markov-based method has a superior performance in revealing species identity behind a sequence. All definitions for these two approaches are given in Appendix A. By taking the parameter vector $\vec{\alpha}$ as a measure of the correlation strength we have two different representations, namely the mutual information function (MI representation) and the parameter vector $\vec{\alpha}$ of the DAR(p) process (Markov representation). In the following we will mostly focus on the Markov representation due to its better performance in this analysis. In Sec. II B the two representations are directly compared in terms of the sequence length dependence of the observed informational structure. We have previously shown that the correlation profiles for the same species almost lie on top of each other and species, which are close in an evolutionary sense, seem to show similar (but distinct) correlation curves [36,37]. Figure 1 summarizes these findings for the case of *M. musculus* and *R norvegicus* chromosomes for $p=50$.

Already in this first analysis step, some evidence for an increasing species distinction with increasing distances is found. Figure 1(c) suggests that systematic differences between the two species' correlation curves increase with

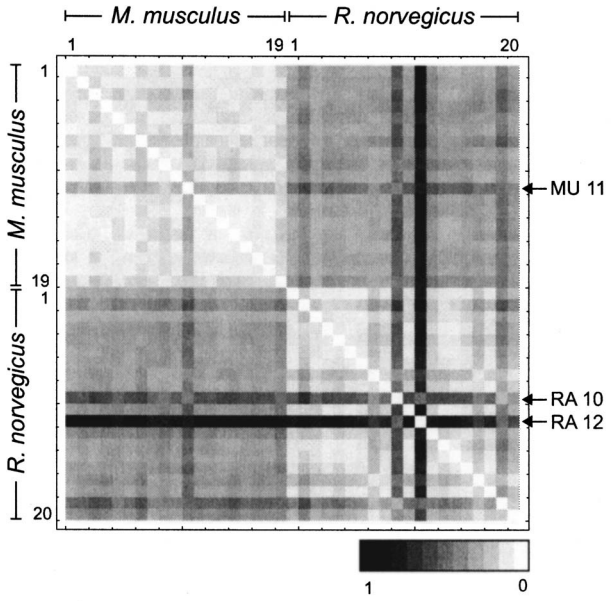

FIG. 2. Density plot of the distance matrix for all chromosomes of *M. musculus* and *R. norvegicus*. Pairwise distances of correlation curves have been calculated with the L_1 norm. The resulting normalized distance matrix has then been represented in grayscale coding according to the displayed lookup table. The three outliers (MU 11, RA 10, and RA 12) are indicated by arrows.

nucleotide distance k. We will discuss this point in more detail at the end of Sec. II B. The most striking aspect of Fig. 1 is that the correlation curves are extremely similar within a species. This phenomenon is common to all eukaryotic species we have so far investigated [36]. In this particular case, however, differences between the two species are extremely small due to the phylogenetic proximity of the two species. An important question, therefore, is if enough interspecies differences remain upon which to base species distinction. For the first two chromosomes of each species the corresponding correlation curves are overlaid in Fig. 1(c). It is seen that the two pairs of curves nearly coincide over almost the whole range of nucleotide distances. Nevertheless small systematic differences appear, e.g., at $k=22$, 26, and 29.

Figure 2 shows the distance matrix for all chromosomes from *M. musculus* and *R. norvegicus* in grayscale coding. This figure has two interesting features. First, a clear block-like structure is seen indicating that the blocks of interspecies entries in the distance matrix are systematically different from the two (diagonal) blocks of intraspecies distances. Thus adding more curves to the comparison in Fig. 1 greatly enhances the systematic differences instead of diminishing them. Second, three chromosomes can be made out in the distance matrix which do not fall into this general pattern, namely MU 11, RA 10, and RA 12. It is interesting to note that RA 12 is known to have an elevated recombination rate compared to other rat chromosomes [38]. Obviously their respective distances to all other chromosomes of the two species lie outside the range found for the other combinations. The clustering tree shown in Fig. 3, which represents the informational structure, confirms these visual impressions from the distance matrix. While the overall systematics

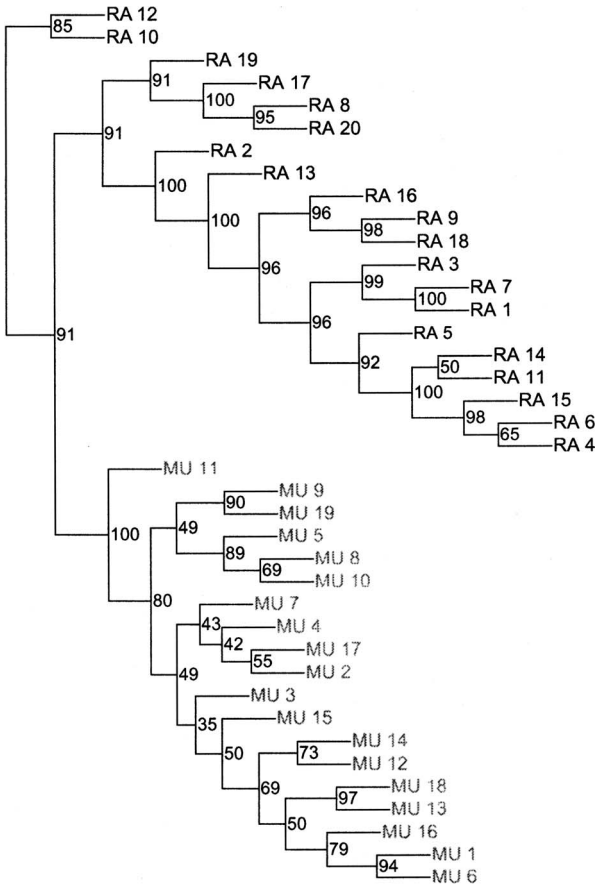

FIG. 3. Clustering tree for chromosomes of *M. musculus* (MU) and *R. norvegicus* (RA) based on the data from Fig. 1. Bootstrap values for 100 bootstrap replicates (see Appendix B) are shown. As before, the number after the two letter abbreviation for the species indicates the number of the respective chromosome.

contained in these correlation profiles, which allows a very high level of species distinction, is the focus of our study, it is nevertheless worthwhile to look at the biological features of these outlying chromosomes. Figure 4 displays two properties of all mouse and rat chromosomes, namely the *GC* content (i.e., the sum over *C* and *G* frequencies) and the

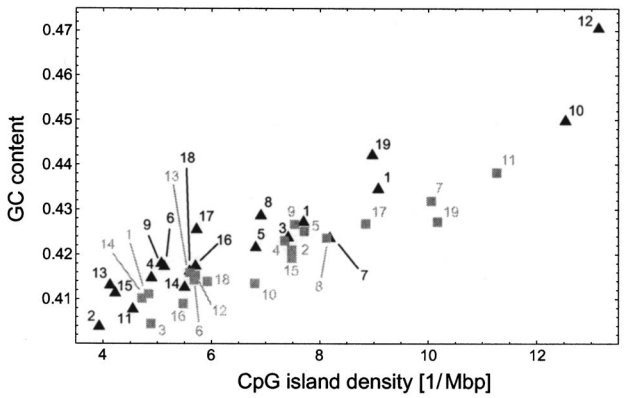

FIG. 4. *GC* content vs *CpG* island density for mouse (squares) and rat (triangles) chromosomes. Data have been obtained from the UCSC Genome Browser database (see Appendix C).

CpG island density (i.e., qualitatively speaking, the average number of regions with an elevated fraction of dinucleotides *CG* per Mbp; for a more detailed definition of *CpG* islands, see [39]). *CpG* islands are known to positively correlate with the location of regulatory regions [39]. The outliers in our clustering tree are immediately marked out as extreme cases in these properties as well. From this observation a scenario emerges, where extreme dinucleotide compositions occasionally override all other information in the correlation profiles, but where in all other cases the key features of the profiles are determined by more long-range properties than this dinucleotide level. In particular, the two quantities provided in Fig. 4 do not explain other features of the clustering tree than the three outlying chromosomes. The observed subclusters, as well as the general distinction between mouse and rat, are beyond this level.

Let us return to the overall properties of the clustering tree in Fig. 3. Apart from the outlying chromosomes MU 11, RA 12, and RA 10, two distinct clusters of *R. norvegicus* and *M. musculus* chromosomes are observed. Thus the correlation patterns of *M. musculus* and *R. norvegicus* for $p=50$ in the Markov representation reveal the species identity of the underlying sequences. Another interesting aspect of the informational structure displayed in Fig. 3 is the patterning of each species cluster into smaller subclusters. The high bootstrap values indicate that these subclusters contain systematic information on chromosome similarity and are not an artifact of random clustering of values in the distance matrix. The systematically lower bootstrap values in the mouse part of the tree (i.e., the lower significance of the mouse subtree structure) are consistent with the lower variance in the ensemble of mouse correlation curves observed in Fig. 1.

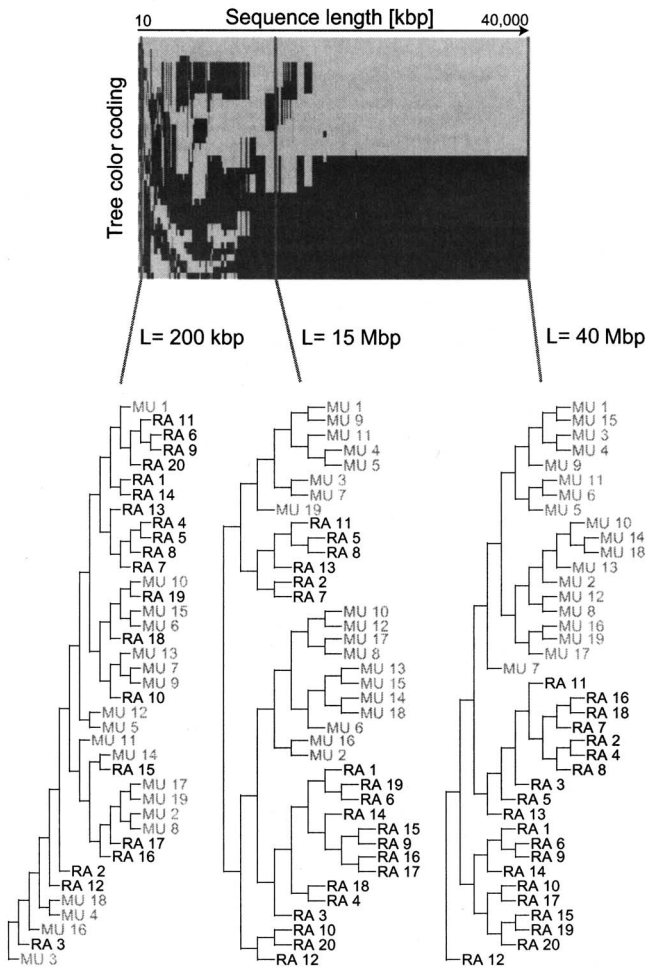

FIG. 5. Tree color coding plot for the Markov representation. The length of the underlying DNA sequences is simultaneously increased starting with the first 10 kbp of each chromosome up to 40 Mbp with a step size of 10 kbp. In the case of exceeding the length of a chromosome before reaching 40 Mbp the length is kept constant at the maximum possible length. For each length the clustering tree is computed and then translated into a TCC line with the help of the TCC algorithm. These lines form the TCC plot. For three different sequence lengths, namely $L=200$ kbp, $L=15$ Mbp, and $L=40$ Mbp the corresponding (sorted) clustering trees are shown.

B. Parameter dependences visualized by TCC

In Sec. II A the general procedure of our analysis has been lined out, where chromosomes are represented by their respective correlation curves from which pairwise distances of correlation patterns can be calculated. The resulting distance matrix can then be translated into a clustering tree. For the case of full chromosomes and high Markov order, namely $p=50$, this analysis led to the high species distinction seen in Fig. 3. In this section we will analyze how the species identity encoded in the correlation curves depends on the two key parameters of our analysis, namely the length of the sequences used to compute the correlation curves and the range of nucleotide distances taken into account. In the case of the Markov representation of the correlation pattern this range is determined by the Markov order p. It is clear that the capacity of distinguishing between species on the basis of the correlation curves will decrease when the amount of underlying sequence information is reduced. It is, however, not immediately clear how much sequence information is needed to reveal species identity. How long must a segment of, e.g., a *M. musculus* chromosome be in order to allow the resulting correlation curve to reflect the "mouseness" of the segment, i.e., to display the characteristic signature of *M. musculus* chromosomes seen in Fig. 1? We studied this length dependence of chromosome clustering by computing a large number of clustering trees for different lengths of the underlying sequences, then aligning these clustering trees with the help of the TCC algorithm described in Appendix B, and lastly displaying all the resulting grayscale lines as a TCC diagram. The result is shown in Fig. 5. In order to provide an idea to what extent homogeneity of a TCC line reflects the clustering and therefore the species distinction present in the corresponding clustering tree Fig. 5 also contains trees for three different values of the sequence length L. Obviously also on the level of the actual clustering tree an increased mixing of chromosomes of the two species is observed as sequence length decreases. The TCC plot thus turns out to be a reliable instrument for monitoring the change of species information as a function of the amount of underlying sequence informa-

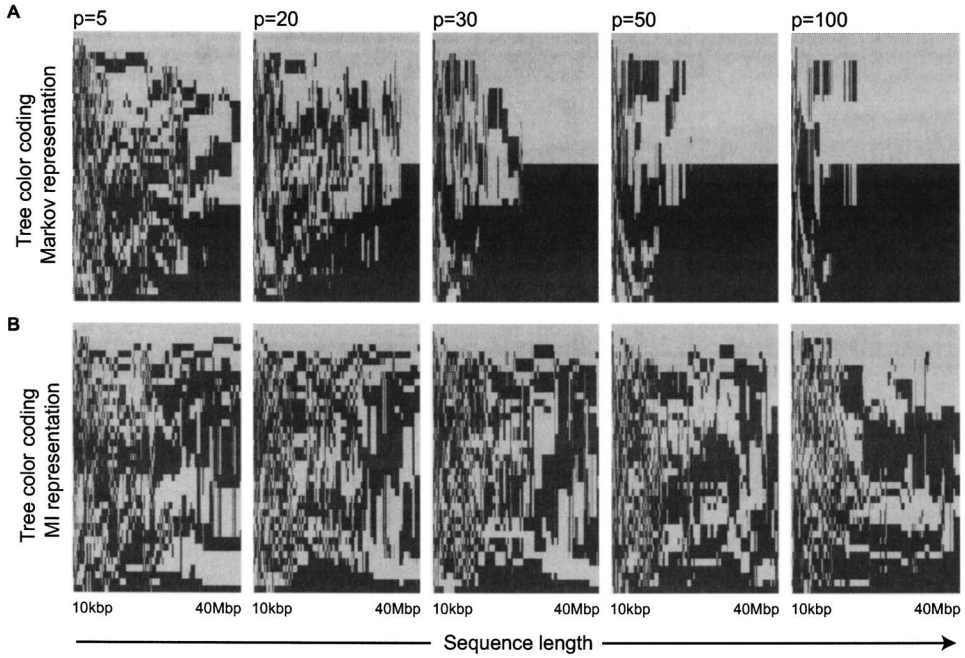

FIG. 6. Tree color coding plots for the two representations of the correlation pattern for different values of p. In this way, both the length of the underlying DNA sequences and the range p of correlations are varied. (A) Markov representation with correlation ranges $p=5$, 20, 30, 50, and 100. (B) MI representation with correlation ranges $p=5$, 20, 30, 50, and 100.

tion. When read from right to left, i.e., in the direction of decreasing sequence length, the TCC plot in Fig. 5 reveals that the capacity of distinguishing between *M. musculus* and *R. norvegicus* persists down to a certain sequence length. From that point on further down to shorter sequences species information gradually decreases. First, large blocks of chromosomes are misplaced, then these blocks become smaller and more numerous and, finally, no species information is observed beyond a random level. After having demonstrated that the TCC plot allows an efficient (although not perfect; cf. Appendix B) representation of species distinction as a function of some model parameter we can now turn to the main result of our investigation. Figure 6 shows the sequence length dependence of the TCC coding line for five different values of the nucleotide distance range p. Results are given both for the Markov representation and for the MI representation. Note that the TCC plot in the Markov representation for $p=50$ is the same as the TCC plot in Fig. 5. Figure 6 clearly demonstrates the dramatic differences between the Markov and the MI representations. In the MI representation (bottom row) an increase in p only leads to a very moderate increase in clustering quality at high sequence lengths. Increasing p for the Markov representation on the other hand (upper row) substantially enhances chromosome clustering and, therefore, the amount of species identity. In particular the range of sequence lengths where complete clustering is achieved becomes larger with ever higher p.

From Fig. 6 it is not clear what nucleotide distances k contribute most to this increase in species distinction, when passing from $p=50$ to 100. It is, therefore, interesting to look at the case $p=100$ on the level of correlation curves. This is shown in Fig. 7. The dashed curve is the $|t|$-value, which is a measure for the species distinction capacity residing in this component of the correlation vector [36]. One has

$$|t_k(A,B)| = \left| \frac{\bar{\alpha}_k(A) - \bar{\alpha}_k(B)}{\sqrt{\frac{\sigma_k^2(A)}{n(A)} + \frac{\sigma_k^2(B)}{n(B)}}} \right|, \tag{1}$$

where, for a fixed index position k, $\bar{\alpha}_k(S)$ denotes the mean and $\sigma_k^2(S)$ the variance of α_k calculated over all $n(S)$ chromosomes of species S which are included in the analysis. It is seen that particularly the region between $k=70$ and 90 contributes substantially to the species distinction. For many values of k in this range the two groups of curves display clearly visible systematic differences. Note that this $|t|$-value is a heuristic measure for the distribution of species distinction *within a given range of distances*. Changing p will alter the distribution of $|t|$-values.

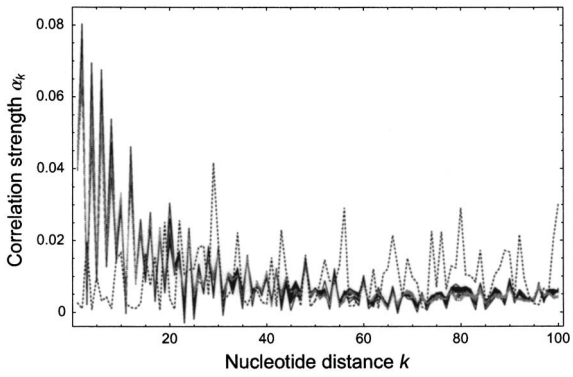

FIG. 7. Correlation curves in the Markov representation for $p=100$. As before, *M. musculus* curves are shown in gray, *R. norvegicus* curves are given in black. In addition, the (normalized) $|t|$-value is shown (dashed curve) as a measure of systematic differences between the two families of correlation curves in each component k.

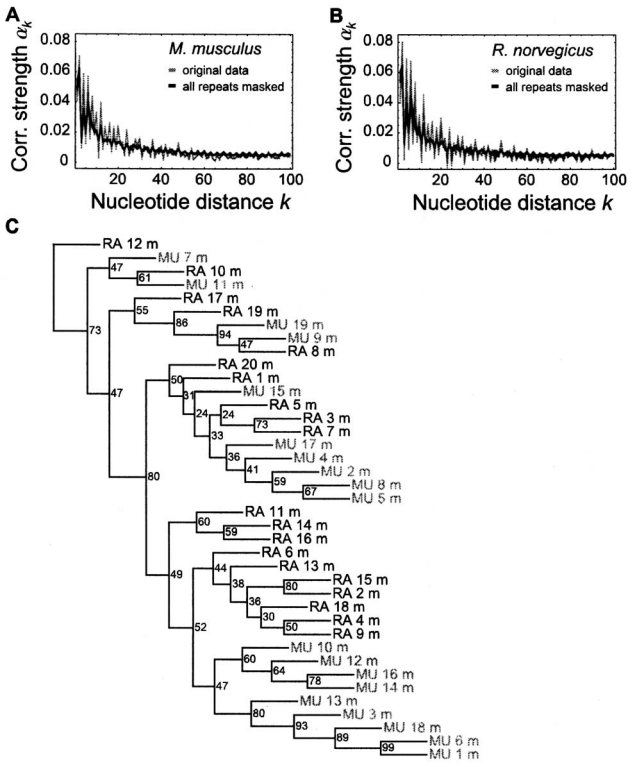

FIG. 8. Correlation curves for masked and original chromosomes of (A) *M. musculus* and (B) *R. norvegicus* for the Markov representation for $p=100$. Part (C) shows the corresponding clustering tree based on the masked sequences, where bootstrap values are obtained as described in Appendix B.

A first step in assessing the role of repeats in the correlation pattern is to eliminate all annotated repetitive DNA from the sequences and recalculate the correlation curves. In order to achieve this, we downloaded sequence data from the *Ensembl* servers (see Appendix C), in which repeats have been identified with the help of the *RepeatMasker* software [40]. In principle, two methods of preparing these repeat-free sequences are possible: (i) cutting out the repeats or (ii) substituting them by random sequence segments. Here we used the first method. We checked that both methods essentially lead to the same correlation curves. In particular, we checked that the change in the correlation curves does not result from the reduction in sequence length upon elimination of repeats. The corresponding curves, together with the previous (unaltered) correlation curves, and the clustering tree based upon the new correlation curves are given in Fig. 8. We observe that (a) the correlation curves are substantially modified by eliminating the repeats, (b) systematic species differences are strongly reduced (and, particularly, the clustering tree no longer allows species distinction), and (c) the new correlation curves still display a high degree of synchronization, suggesting that a residual systematic signal beyond repetitive DNA persists in the correlation patterns.

III. DISCUSSION

We studied the capacity to distinguish between species, which is contained within the statistical correlations in DNA sequences at short and intermediate-range distances. The focus of our study has been on the distinction of two eukaryotic species, which are closely related in evolution, namely mouse and rat. We observe that the distinction increases with an increasing symbol distance. This phenomenon, which is already tracable in the correlation profiles themselves, can be quantitatively studied by looking at the sequence lengths necessary for species distinction. The correlation profiles, which provide the basis of our subsequent clustering analysis, have been obtained with an approach of evaluating such symbol correlations in DNA sequences. In contrast to standard methods from information theory, this approach is capable of revealing the identity of the two closely related species on the basis of these correlation patterns.

A key result of our study is that both, in the Markov and in the MI representations, systematic differences in the correlation patterns of the two species increase with nucleotide distance. This effect is more pronounced for the Markov representation than for the MI representation. This finding strongly suggests that for these two species, the distinctive correlation features extend far beyond the dinucleotide level.

In a certain sense, this view may turn out to be more interesting (and far more productive) than the more local, neighbor-centered approach. Understanding these statistical features as signatures of repetitive elements provides a link to processes of genome evolution, such as retrotransposition and microsatellite growth, which pattern the genome on an evolutionary time scale. By linking the elementary process-oriented properties with genome-wide statistical patterns, this view of the correlation profile as a process signature clearly points towards a system biology treatment of genome evolution. The corresponding view of an evolving genome as a dynamical system governed by local rules and rate equations shaping the sequence has moved into the focus of scientific attention within the last years (see, e.g., [9,11,12,15]).

In order to further confirm this evidence, we intend to analyze the impact that a subsequent removal of classes of repetitive elements from the sequence has on the correlation profiles. Removing *all* annotated repeats from the sequence has a huge impact on the correlation profiles (see Fig. 8). However, the correlation curves remain highly synchronized. The differences between the original and the reduced correlation curves (as well as the remaining amount of synchronization) have to be understood in terms of processes of genome evolution. A good starting point could be local sequence properties like GC fluctuations and their relation to repetitive DNA (see, e.g., [21,41]). We believe that the systematic features of the correlation curves after eliminating the repeats can, to a certain extent, be related to structural properties of the (three-dimensionally arranged) DNA molecule. The longer-range goal of these studies, as pointed out in the Introduction, is to link features in the correlation profile with properties of the processes involved in distributing repetitive elements within the genome. The fact that in this intermediate-range regime of symbol distances the correlation profiles contain an ever increasing distinctiveness of these two species, supports this perspective.

APPENDIX A: CORRELATION MEASURES

A DNA sequence consists of nucleotides adenine (A), guanine (G), cytosine (C), and thymine (T). Let $p(i)$ denote

the probability of finding the symbol i of the alphabet $A=\{A,G,C,T\}$ in a given DNA sequence and let $p^k(i,j)$ be the probability of finding the symbols i and $j \in A$ at a distance k in a given sequence. The mutual information function is defined as (see, e.g., [17])

$$I(k) = \sum_{(i,j)\in A^2} p^{(k)}(i,j)\log_2 \frac{p^{(k)}(i,j)}{p(i)p(j)}, \tag{A1}$$

where $I(k)$ quantifies the amount of information one obtains from a symbol i about a symbol j in a distance k within the sequence. In this way it is a measure of the strength of average correlation between the symbols i and j at a distance k. The vector $\{I(1),\ldots,I(p)\}$ constitutes the pth order mutual information (MI) representation of a DNA sequences correlation pattern.

A more refined method of quantifying this average correlation is given by the parameters of a discrete autoregressive process of order p, DAR(p), as described in [35]. The process can be used to generate symbol sequences with higher order Markov properties. The DAR(p) process is defined by the following recursion relation [35,42,43]:

$$X_n = V_n X_{n-A_n} + (1 - V_n)Y_n, \tag{A2}$$

where X_n denotes the nth symbol in the sequence, which is determined by the memory part or by the purely random part of the recursion. V_n serves as a switch between the two parts. With probability ρ the stochastic variable V_n has the value 1 and with the remaining probability $1-\rho$ the quantity V_n takes on the value 0. In the case of $V_n=1$ a symbol from the history of the process (i.e., a previous symbol in the sequence) is chosen. A_n denotes the number of steps one goes back in the sequence for selecting this new symbol. A_n takes values from 1 to p with the respective probabilities $\alpha_1,\alpha_2,\ldots,\alpha_p$. In the case of $V_n=0$ a symbol is chosen at random from some marginal distribution $\vec{\pi}$. In this way, ρ quantifies the amount of randomness in the sequence. The result is a pth order Markov process. In addition to simulating sequences for certain parameter values, all the process' parameters can be estimated from a given sequence. We will use the parameter vector $\vec{\alpha}$ obtained by a Yule-Walker formalism as a measure for correlation strength in a distance k. This estimation process basically consists of two steps. First an empirical autocorrelation function in symbolic space is estimated for a given DNA sequence. The *ad hoc* estimator for these quantities is defined as follows [43]:

$$\hat{r}(k) = 1 - \sum_{a_i \in A} B_m(k,a_i)\frac{1}{1-\pi(a_i)} \tag{A3}$$

for $k=1,2,\ldots$ with

$$B_m(k,a_i) = \frac{1}{m-k}\sum_{a_j \neq a_i \in A}\sum_{l=1}^{m-k} \delta_{a_j}(x_l)\delta_{a_i}(x_{l+k}), \tag{A4}$$

where $\delta_y(x)=1$, if $x=y$, and 0 otherwise.

The second step leads from these quantities $\hat{r}(k)$ to the actual parameters of the DAR(p) process. In order to obtain this parameter vector $\vec{\alpha}$ one has to deal with a set of nonlinear Yule-Walker equations [35,43], which relate the (theoretical) r- and ϕ-parameters. For our purposes it is convenient to observe that this set of equations can be linearized by the substitution $\phi_k=\rho\alpha_k$. One obtains:

$$r(1) = \phi_1 r(0) + \phi_2 r(1) + \cdots + \phi_p r(p-1), \tag{A5}$$

$$r(2) = \phi_1 r(1) + \phi_2 r(0) + \cdots + \phi_p r(p-2), \tag{A6}$$

$$\vdots \qquad \vdots \qquad . \qquad \vdots \qquad \vdots$$

$$r(p) = \phi_1 r(p-1) + \phi_2 r(p-2) + \cdots + \phi_p r(0), \tag{A7}$$

where $r(0)=1$. Since $\sum_{k=1}^{p}\alpha_k=1$, the following normalization condition on the level of the quantities ϕ_k holds:

$$\sum_{k=1}^{p}\phi_k = \sum_{k=1}^{p}\alpha_k\rho = \rho\sum_{k=1}^{p}\alpha_k = \rho. \tag{A8}$$

Inserting $\hat{r}(1),\hat{r}(2),\ldots,\hat{r}(p)$ for $r(1),r(2),\ldots,r(p)$, the p equations can be solved for the p parameters with $\phi_k=\rho\alpha_k$ and $k=1,\ldots,p$. The vector $\vec{\alpha}$ resulting from this estimation process is the pth order Markov representation of the correlation pattern. The key advantage of this Markov representation is that the contribution of a random background is effectively absorbed by the parameter ρ, which does not enter our subsequent clustering analysis. We will see that, as a consequence, the minimal segment length necessary for a distinction of *M. musculus* and *R. norvegicus* chromosomes is substantially lower for the Markov representation than for the MI representation.

APPENDIX B: DISTANCE MATRIX, CLUSTER ALGORITHM, AND TREE COLOR CODING (TCC) DIAGRAMS

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 between the chromosomes a and b is then given by

$$d_{a,b} = \|\vec{\alpha}^{(a)} - \vec{\alpha}^{(b)}\|_1 = \sum_{k=1}^{p}|\alpha_k^{(a)} - \alpha_k^{(b)}|, \tag{B1}$$

where $\vec{\alpha}^{(s)}=(\alpha_1^{(s)},\ldots,\alpha_p^{(s)})$ 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 are obtained by a UPGMA algorithm based on the distance matrices using the software package PHYLIP [44]. Bootstrap replicates have been obtained by randomly deleting 20% of pairs of components entering the computation of $d_{a,b}$. The software component CONSENS in the PHYLIP package has been used to calculate a 50% majority-rule (extended) consensus tree. Dendograms have been displayed using the software tool TREEVIEW [45].

As pointed out in the previous section our method of quantifying the correlation pattern of a DNA sequence de-

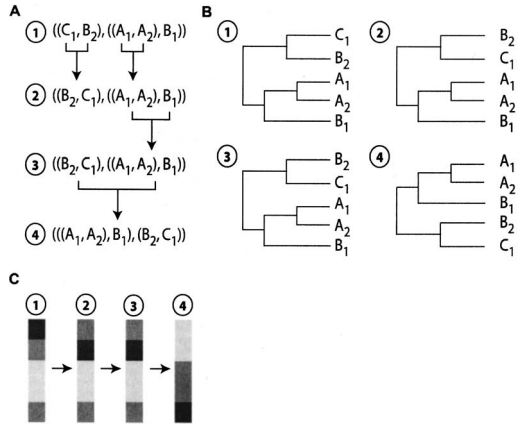

FIG. 9. Schematic view of the tree color coding (TCC) algorithm. (A) Operations upon the Newick representation are shown for a simple clustering tree of five taxa from three different species. Starting from the unsorted tree (1) the TCC algorithm yields a sorted tree (4) by iterative application of branch switches. (B) Same operations as in (A), but now on the actual dendogram corresponding to the tree from (A). (C) Visualization of the original tree (1), intermediate steps (2) and (3) and the final, sorted tree (4) as TCC lines.

pends on several parameters. The most important parameters are the range of correlation (i.e., the order p of the corresponding representation) and the sequence length. As described above, correlation curves are translated into a distance matrix, which in turn is converted into a clustering tree. When studying the parameter dependence of our result we are consequently confronted with the task of comparing a substantial number of different clustering trees. For our purposes the key observable on such a clustering tree is the quality of species distinction, i.e., how pronounced the formation of clusters appears in the tree. Comparing such trees requires a universal sorting of the branches. To this end we developed a sorting algorithm which translates such a clustering tree into a simple line of colors, where the number of color changes basically reflects the amount of clustering in the underlying tree and therefore the quality of species distinction. The algorithm acts upon the Newick representation of a clustering tree, where entries in a list represent taxa and matching brackets denote objects linked by branches. Our algorithm, which is illustrated in Fig. 9, first acts upon the innermost branches and performs an alphanumerical sorting of the corresponding taxa using only topologically allowed branch switches. In the next step one moves to the next higher order of branches and applies sorting there. Whenever one encounters nonelementary objects at the end of one branch (i.e., a subtree instead of a single taxon) the alphanumerically lowest object in the subtree serves as a label for the subtree itself. After passing through all hierarchical levels in the clustering tree all taxa are sorted as close to alphanumerical order as topology of the tree allows. Coloring all taxa according to their species affiliation leads to a color line whose homogeneity directly reflects the degree of clustering observed in the original tree and, furthermore, can be immediately compared with any other tree consisting of the same taxa due to the universal order of taxa approximated by the sorting algorithm. The tree color coding algorithm slightly overestimates the overall order in the tree, as different branches containing taxa of the same species can become direct neighbors in the color line, even if one of them also contains chromosomes of another species. Figure 9 provides three different views on our tree color coding algorithm, illustrated with a very simple tree consisting of five taxa from three different species. Part A illustrates the operations acting upon the Newick representation of this tree. It is clearly seen which constellations lead to a change of order of the elements inside the brackets. In part B the same operations are shown for the actual clustering tree where branch switches correspond to the change of elements in part A. Even though the color line stands at the end of the TCC algorithm it is nevertheless instructive to see how the sequence of colors evolves during the operations of the algorithm. This is depicted in part C. Note that the TCC algorithm systematically overestimates the degree of clustering found in the tree because taxa from the same species which lie in different clusters can end up in adjacent positions in the TCC line due to sorting. This is, for example, seen for taxa B1 and B2 in Fig. 9. It results from neglecting in the color line all topological information coming from higher-order branches, i.e., basically from mapping a tree to a one-dimensional line. It is, however, also clear from Fig. 9 that this systematic error is rather small when much more taxa than species (colors) are involved.

APPENDIX C: DATA SETS

Original and repeat-masked DNA sequences for *M. musculus* and *R. norvegicus* were downloaded from the site ftp://ftp.ensembl.org/pub/release-22/. Corresponding information on *CpG* islands have been obtained from the UCSC Genome Browser site http://genome.ucsc.edu/ for the UCSC mouse release *mm*4 and the UCSC rat release *rn*3. Sex chromosomes have been omitted. 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.

[1] J. Jurka, O. Kohany, A. Pavlicek, V. V. Kapitonov, and M. V. Jurka, Proc. Natl. Acad. Sci. U.S.A. **101**, 1268 (2004).
[2] S. Yang, A. F. Smit, S. Schwartz, F. Chiaromonte, K. M. Roskin, D. Haussler, W. Miller, and R. C. Hardison, Genome Res. **14**, 517 (2004).
[3] A. L. Price, E. Eskin, and P. A. Pevzner, Genome Res. **14**, 2245 (2004).
[4] J. M. Greally, Proc. Natl. Acad. Sci. U.S.A. **99**, 327 (2002).

[5] C. Rizzon, G. Marais, M. Gouy, and C. Biemont, Genome Res. **12**, 400 (2002).
[6] C. Chen, A. J. Gentles, J. Jurka, and S. Karlin, Proc. Natl. Acad. Sci. U.S.A. **99**, 2930 (2002).
[7] I. Ovchinnikov, A. B. Troxel, and G. D. Swergold, Genome Res. **11**, 2050 (2001).
[8] S. Temnykh, G. DeClerck, A. Lukashova, L. Lipovich, S. Cartinhour, and S. McCouch, Genome Res. **11**, 1441 (2001).
[9] P. W. Messer, P. F. Arndt, and M. Lässig, Phys. Rev. Lett. **94**, 138103 (2005).
[10] Y. Zhou and B. Mishra, Proc. Natl. Acad. Sci. U.S.A. **102**, 4051 (2005).
[11] Y.-H. Chen, S.-L. Nyeo, and C.-Y. Yeh, Phys. Rev. E **72**, 011908 (2005).
[12] L.-C. Hsieh, L. Luo, F. Ji, and H. C. Lee, Phys. Rev. Lett. **90**, 018101 (2003).
[13] H.-D. Chen, C.-H. Chang, L.-C. Hsieh, and H.-C. Lee, Phys. Rev. Lett. **94**, 178103 (2005).
[14] W. Peng, H. Levine, T. Hwa, and D. A. Kessler, Phys. Rev. E **69**, 051911 (2004).
[15] S. Nicolay, F. Argoul, M. Touchon, Y. d'Aubenton Carafa, C. Thermes, and A. Arneodo, Phys. Rev. Lett. **93**, 108101 (2004).
[16] Y. Xiao, Y. Huang, M. Li, R. Xu, and S. Xiao, Phys. Rev. E **68**, 061913 (2003).
[17] I. Grosse, H. Herzel, S. V. Buldyrev, and H. E. Stanley, Phys. Rev. E **61**, 5624 (2000).
[18] H. Herzel, O. Weiss, and E. N. Trifonov, Bioinformatics **15**, 187 (1999).
[19] D. Holste, I. Grosse, S. Beirer, P. Schieg, and H. Herzel, Phys. Rev. E **67**, 061913 (2003).
[20] W. Li and D. Holste, Comput. Biol. Chem. **28**, 393 (2004).
[21] W. Li and D. Holste, Fluct. Noise Lett. **4**, L453 (2004).
[22] S. Karlin and I. Ladunga, Proc. Natl. Acad. Sci. U.S.A. **91**, 12832 (1994).
[23] S. Karlin and J. Mrázek, Proc. Natl. Acad. Sci. U.S.A. **94**, 10227 (1997).
[24] A. J. Gentles and S. Karlin, Genome Res. **11**, 540 (2001).
[25] D. T. Pride, R. J. Meinersmann, T. M. Wassenaar, and M. J. Blaser, Genome Res. **13**, 145 (2003).
[26] J. Qi, B. Wang, and B. Hao, J. Mol. Evol. **58**, 1 (2004).
[27] S. Karlin, A. M. Campbell, and J. Mrázek, Annu. Rev. Genet. **32**, 185 (1998).
[28] C. Coulondre, J. Miller, P. Farabaugh, and W. Gilbert, Nature (London) **274**, 775 (1978).
[29] A. Razin and A. Riggs, Science **210**, 604 (1980).
[30] S. Hess, J. Blake, and R. Blake, J. Mol. Biol. **236**, 1022 (1994).
[31] P. Arndt, C. Burge, and T. Hwa, in *Proceedings of the 6th Annual International Conference on Computational Biology (RECOMB 2002)* (ACM Press, New York, 2002), pp. 32–38.
[32] P. F. Arndt and T. Hwa, Bioinformatics **21**, 2322 (2005).
[33] B. Borstnik and D. Pumpernik, Phys. Rev. E **71**, 031913 (2005).
[34] B. Borstnik and D. Pumpernik, Europhys. Lett. **65**, 290 (2004).
[35] M. Dehnert, W. E. Helm, and M.-T. Hütt, Physica A **327**, 535 (2003).
[36] M. Dehnert, W. E. Helm, and M.-T. Hütt, Gene **345**, 81 (2005).
[37] M. Dehnert, R. Plaumann, W. E. Helm, and M.-T. Hütt, J. Comput. Biol. **12**, 545 (2005).
[38] M. I. Jensen-Seaman, T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin, C.-F. Chen, M. A. Thomas, D. Haussler, and H. J. Jacob, Genome Res. **14**, 528 (2004).
[39] D. Takai and P. Jones, Proc. Natl. Acad. Sci. U.S.A. **99**, 3740 (2002).
[40] A. Smit, R. Hubley, and P. Green, *RepeatMasker Open-3.0. at* http://www.repeatmasker.org(1996–2004).
[41] W. Li and D. Holste, Phys. Rev. E **71**, 041910 (2005).
[42] P. Jacobs and P. Lewis, Technical Report NPS55-78-022, Naval Postgraduate School, Monterey, CA, 1978 (unpublished).
[43] P. Jacobs and P. Lewis, J. Time Ser. Anal. **4**, 19 (1983).
[44] J. Felsenstein, *PHYLIP* (*Phylogeny Inference Package*) *version 3.6* (*alpha3*), distributed by the author. Department of Genome Sciences, University of Washington, Seattle, 2004.
[45] R. D. M. Page, CABIOS, Comput. Appl. Biosci. **12**, 357 (1996).