JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 10, 2009

Pp. 1487–1500
DOI: 10.1089/cmb.2009.0106

Research Article

Estimating Mutation Distances from Unaligned Genomes

BERNHARD HAUBOLD,[1] PETER PFAFFELHUBER,[2] MIRJANA DOMAZET-LOŠO,[1,3] and THOMAS WIEHE[4]

ABSTRACT

Alignment-free distance measures are generally less accurate but more efficient than traditional alignment-based metrics. In the context of genome sequence analysis, the efficiency gain is often so substantial that it outweighs the loss in accuracy. However, a further disadvantage of alignment-free distances is that their relationship to evolutionary events such as substitutions is generally unknown. We have therefore derived an estimator of the number of substitutions per site between two unaligned DNA sequences, K_r. Simulations show that this estimator works well with "ideal" data. We compare K_r to two alternative alignment-free distances: a *k*-tuple distance and a measure of relative entropy based on average common substring length. All three measures are applied to 27 primate mitochondrial genomes, eight whole genomes of *Streptococcus agalactiae* strains, and 12 whole genomes of *Drosophila* species. In each case, the cluster diagrams based on K_r are equivalent to or significantly better than those based on the two alternative measures. This is due to the fact that in contrast to the alternative measures K_r is derived from an explicit model of evolution. The computation of K_r is efficiently implemented in the program kr, which can be downloaded freely from the internet.

Key words: alignment-free distance, number of substitutions, genome comparison, suffix tree, shortest unique substring.

1. INTRODUCTION

"THE AFFINITIES OF ALL THE BEINGS OF THE SAME CLASS have sometimes been represented by a great tree. I believe this simile largely speaks the truth." This quote from the *Origin of Species* indicates that Darwin was not the first to imagine descent as a branching tree when he drew the famous single figure for his book (Darwin, 1859). Indeed, the tree metaphor had been used by a number of prominent biologists before, including Lamarck (Archibald, 2008). Rather than priority, what is remarkable here is Darwin's clarification that a phylogeny represents "affinities between all the beings of the same class." The complement of "affinities" are distances.

Distances between organisms are traditionally defined with respect to homologous traits. Since the advent of molecular biology, the traits most widely used for phylogeny reconstruction have been protein

[1]Department of Evolutionary Genetics, Max-Planck-Institute for Evolutionary Biology, Plön, Germany.
[2]Faculty of Mathematics & Physics, University of Freiburg, Freiburg, Germany.
[3]Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia.
[4]Institute of Genetics, University of Cologne, Cologne, Germany.

and nucleotide residues. Homologous residues are identified by sequence alignment, and the corresponding algorithms have been developed and refined since the early 1970s (Waterman, 1995).

Today, alignment algorithms can be applied to complete genome sequences from diverse organisms yielding a wealth of evolutionary insights. However, more complex organisms tend to have larger genomes. Kauffman observed that organismal complexity as quantified by the number of cell types generally only grows as the square root of the amount of DNA per cell (Kauffman, 1993). This leads to the well-known large genomes of organisms with many cell types such as mammals.

Alignment of very long sequences is computationally challenging (Dewey and Pachter, 2006). Fortunately, an alignment is not necessary for computing distances between sequences, and alignment-free sequence comparison has been developed since the 1980s (Blaisdell, 1986). These methods fall in two classes: approaches based on fixed-length words and resolution-free methods (Vinga and Almeida, 2003). The latter class is rather heterogeneous, and includes information theoretical measures, such as relative entropy (Ulitsky et al., 2006).

In-between full alignment on the one hand an alignment-free methods on the other are metrics based on homology at a higher level than individual residues. These include distances computed from the number of inversions (Adam and Sankoff, 2008), gene content (Huson and Steel, 2004), or coverage by BLAST hits (Henz et al., 2005).

A good distance measure should be linear with evolutionary time. Since evolutionary time can usually not be observed directly, it is inferred from the number of certain kinds of events that have occurred since divergence. Such events might be genome rearrangements (Adam and Sankoff, 2008), insertions/deletions (Lunter et al., 2006), or nucleotide substitutions (Jukes and Cantor, 1969). The latter is most widely used because it can easily be computed from an alignment and is to a first approximation linear in time (Zuckerkandl and Pauling, 1965).

The advantage of alignment-free distance measures is their efficient computation compared to the construction of alignments to infer substitution rates. Their disadvantage is, however, that their growth is monotone, but not necessarily linear in time. This is the reason why these measures are usually only used to reconstruct the topology of phylogenies and not their branch lengths.

Our aim in this study is to derive an alignment-free distance measure that allows branch-length reconstruction in addition to topology reconstruction. Specifically, we propose a method to estimate the number of substitutions per site from unaligned genomes. We develop this distance measure on the background of our previous work on the repeat structure of DNA sequences using the concept of shortest unique substrings, which we call shustrings. These constitute the shortest unique prefixes of each suffix in a sequence (Haubold et al., 2005). Based on shustrings, we have previously defined a measure of genome repetitiveness, the index of repetitiveness, which is closely related to the information theoretical relative entropy (Haubold and Wiehe, 2006). A similar quantity has been used as a distance measure for DNA and protein sequences (Ulitsky et al., 2006). However, we will show that, without an explicit model of how a given distance measure changes over time, its utility is reduced.

In the following, we derive our repeat-based estimator of the number of substitutions per site, K_r. We explore by simulation the domain of K_r before applying it to three data sets: 27 primate mitochondrial genomes, the complete genomes of eight strains of *Streptococcus agalactiae* (Tettelin et al., 2005), and the complete genomes of 12 *Drosophila* species (*Drosophila* 12 Genomes Consortium, 2007). We compare the results based on K_r to two alternative distance metrics representing the two major classes of alignment-free distance measures: a *k*-tuple distance measure recently shown to be useful in the reconstruction of trees from highly divergent sequences (Yang and Zhang, 2008), and a measure of relative entropy (Cover and Thomas, 2006) based on average common substring lengths (Ulitsky et al., 2006). In each case, K_r gives biologically meaningful results that are either equivalent or substantially better than the next best alternative.

2. APPROACH AND DATA

2.1. Derivation of K_r

K_r measures the distance between pairs of double-stranded DNA sequences. Let Q and S be such a pair of sequences, which we call *query* and *subject*, respectively. For every suffix of Q, $Q[i..|Q|]$, we determine the shortest prefix that is absent from S and call these shortest prefixes *shortest unique substrings*, or *shustrings*

(Haubold et al., 2005, 2008; Haubold and Wiehe, 2006). We wish to know the expected shustring length as a function of the number of substitutions that separate Q and S since they diverged from their last common ancestor. We start deriving this relationship by establishing the probability density function of shustring lengths for unrelated query and subject before generalizing this to related pairs of sequences.

For $1 \le i \le |Q|$ and $1 \le i' \le |S|$, let

$$X_{i,i'} := \min\{k : Q[i..i+k-1] \neq S[i'..i'+k-1]\}$$

and

$$X_i^* := \max_{1 \le i' \le |S|} X_{i,i'}.$$

In other words, X_i^* refers to the length of the shustring starting at position i in Q. Moreover, we set $K_{i,x,\bullet}$ as the number of nucleotides $\bullet$ in the substring $Q[i..i+x-1]$, $\bullet = A,C,G,T$, and summarize

$$\underline{K}_{i,x} := (K_{i,x,A}, K_{i,x,C}, K_{i,x,G}, K_{i,x,T}).$$

Let the frequencies of the nucleotides found in Q and S be denoted by

$$\underline{p} := (p_A, p_C, p_G, p_T) \quad \text{and} \quad \underline{q} := (q_A, q_C, q_G, q_T), \tag{1}$$

respectively. Assuming that Q and S are strings arising by independent draws from $\underline{p}$ and $\underline{q}$, we find

$$\mathbb{P}(\{X_{i,i'} \le x\} | Q[i..i+x-1]) = 1 - \prod_{j=i}^{i+x-1} q_{Q[j]} = 1 - \underline{q}^{\underline{K}_{i,x}}. \tag{2}$$

Here, we have ignored edge effects for i close to $|Q|$ or i' close to $|S|$, which is equivalent to assuming that both sequences are long. By writing $\binom{x}{\underline{k}} := \binom{x}{k_A \ldots k_T}$, we can summarize the probability of observing a particular nucleotide composition in a shustring as

$$\mathbb{P}\{\underline{K}_{i,x} = \underline{k}\} = \binom{x}{\underline{k}} \mathbb{P}\{Q[i..i+x-1]\} = \binom{x}{\underline{k}} \underline{p}^{\underline{k}}.$$

Next we make the approximation that $X_{i,1}, \ldots, X_{i,|S|}$ are independent, i.e., shustrings do not overlap. While this is clearly not true, we shall see that this assumption is justified for long sequences. In that case we obtain, using conditional expectations,

$$\mathbb{P}\{X_i^* \le x\} = \mathbb{E}[\mathbb{P}(\{X_i^* \le x\} | \underline{K}_{i,x})] = \mathbb{E}\left[\left(1 - \underline{q}^{\underline{K}_{i,x}}\right)^{|S|}\right] = \sum_{\underline{k}} \binom{x}{\underline{k}} \underline{p}^{\underline{k}} \left(1 - \underline{q}^{\underline{k}}\right)^{|S|}. \tag{3}$$

Since we are dealing with double stranded DNA, we can make the simplifications $q_G = q_C$, $q_A = q_T$, $p_G = p_C$, $p_A = p_T$ and define the GC-content of Q as $2p = p_G + p_C$ and that of S as $2q = q_C + q_G$. Then the desired probability density function of shustring lengths for unrelated query/subject can be expressed as

$$\begin{aligned}
\mathbb{P}\{X_i^* \le x\} &= \sum_{\underline{k}} \binom{x}{\underline{k}} \underline{p}^{\underline{k}} \left(1 - \underline{q}^{\underline{k}}\right)^{|S|} \\
&= \sum_{\underline{k}} \binom{k_G + k_C}{k_G} \binom{k_A + k_T}{k_A} \binom{x}{k_G + k_C} p^{k_G + k_C} \left(\frac{1}{2} - p\right)^{k_A + k_T} \left(1 - q^{k_G + k_C} \left(\frac{1}{2} - q\right)^{k_A + k_T}\right)^{|S|} \\
&= \sum_{k=0}^{x} 2^x \binom{x}{k} p^k \left(\frac{1}{2} - p\right)^{x-k} \left(1 - q^k \left(\frac{1}{2} - q\right)^{x-k}\right)^{|S|}.
\end{aligned} \tag{4}$$

This generalizes the corresponding probability density function for a single sequence derived as equation (1) by Haubold et al. (2005).

Next we consider pairs of related sequences that evolve under a one-parameter mutation model where every nucleotide changes into every other nucleotide with equal rate. This mutation model is also known as

the Jukes-Cantor model (Jukes and Cantor, 1969). Assume for simplicity that query and subject have identical nucleotide composition, $\underline{p} = \underline{q}$. Let d' denote the number of uniformly distributed segregating sites that have accumulated since the divergence of query and subject. The probability that a particular query substring of length x, $Q[i..i+x-1]$ has an exact match $S[i..i+x-1]$ at the homologous position in the subject is $(1-d'/|S|)^x$. Note that $X_{i,i}$ is the length of the homologous exact matches. Since $X_{i,i} \ll |S|$, we can write

$$\mathbb{P}\{X_{i,i} > x\} = \left(1 - \frac{d'}{|S|}\right)^x \approx e^{-xd'/|S|}. \tag{5}$$

This means that $X_{i,i}$ is approximately exponentially distributed with parameter $d := d'/|S|$ and hence has an expectation of $1/d$. Note that in this case $X_{i,i}$ is independent of $Q[i..|Q|]$. Hence, with a similar calculation as for the unrelated case, we find that

$$\begin{aligned}\mathbb{P}\{X_i^* \le x\} &= \mathbb{P}\{X_{i,i} \le x\} \cdot \mathbb{E}\left[\mathbb{P}\left(\left\{\max_{i' \ne i} X_{i,i'} \le x\right\} | \underline{K}_{i,x}\right)\right] \\ &\approx (1 - e^{-xd}) \cdot \sum_{k=0}^{x} 2^x \binom{x}{k} p^k \left(\frac{1}{2} - p\right)^{x-k} \left(1 - p^k \left(\frac{1}{2} - p\right)^{x-k}\right)^{|S|},\end{aligned} \tag{6}$$

where the sum is again over all $\underline{k} = (k_A, k_C, k_G, k_T)$ with $k_A + \cdots + k_T = x$ and $2p$ is the GC-content of both sequences, as above. Note that, for known values of p and $|S|$, the distance d is the only parameter on the right hand side of (6). We can therefore express the expectation of the average shustring length,

$$\ell_{Q,S} := \frac{1}{|Q|} \sum_{i=1}^{|Q|} X_i^*, \quad \text{i.e.} \quad \mathbb{E}[\ell_{Q,S}] = \sum_{x=1}^{|S|} x(\mathbb{P}\{X_i^* \le x\} - \mathbb{P}\{X_i^* \le x-1\}), \tag{7}$$

as a function of d. Thus, equation (7) establishes the sought relationship between divergence, d, and the (expected) shustring length under this divergence. Using a moment-based approach, substituting the average shustring length for its expectation, we can now compute the divergence given an observed average shustring length. In a final step such a divergence value is converted to our new mutation distance using the Jukes-Cantor equation (Jukes and Cantor, 1969):

$$K_r = -\frac{3}{4} \ln\left(1 - \frac{4}{3} d\right).$$

2.2. *Asymmetric values of* K_r

The average shustring length varies, depending on which of a pair of sequences we label query and subject. This translates to asymmetric K_r values, which is of course unacceptable in a metric. There are two main sources of asymmetric average shustring lengths: (i) local homology and (ii) copy number variation in shared elements (Fig. 1). If homology is only local, shustring lengths from non-homologous regions indicated as dotted lines will deflate the average shustrings length (Fig. 1A). This problem is mitigated by using equation (4) to exclude shustring lengths from the analysis that would be observed by chance alone.

A greater number of a shared repeat element in the query than in the subject leads to an increase in average shustring length compared to the converse case (Fig. 1B). Segmental duplication is a single event

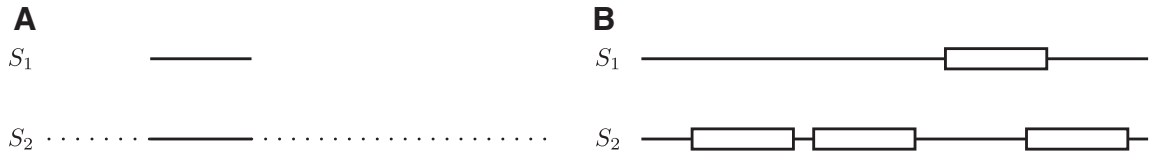

FIG. 1. Sources of asymmetry in the average shustring length, $\ell(Q, S)$. (**A**) Sequences S_1 and S_2 differ in length and as a result share only local homology (—), in which case $\ell(S_1, S_2) > \ell(S_2, S_1)$. (**B**) S_1 contains a lower copy number of a particular genetic element (box) than S_2, in which case $\ell(S_1, S_2) < \ell(S_2, S_1)$.

in evolution, while a potentially large number of mutations is necessary to reverse its effect on the average shustring length. We therefore decided to use the smaller of the two values leading to a greater divergence.

2.3. Implementation

At the heart of the computation of our distance measure lies the calculation of the average shustring length. This is done by indexing query and subject in a generalized suffix tree to look up individual shustring lengths. We used a suffix array (Manzini and Ferragina, 2002) to simulate suffix tree traversal (Abouelhoda et al., 2002). Details of the corresponding algorithm are given in Haubold et al. (2008). The resulting program, kr, can be accessed via a web interface at http://guanine.evolbio.mpg.de/kr/. The site allows submission of a file containing the sequences to be compared in any format and returns the K_r-based phylogeny. In addition, the C source code of kr is freely available from this web site under the GNU General Public License.

2.4. Alternative distance measures

We implemented two alternative alignment-free distance measures to compare K_r to: the k-tuple distance (Yang and Zhang, 2008) and the average common substring distance (Ulitsky et al., 2006). The k-tuple distance is defined as

$$d_{\mathrm{ktup}}(Q,S) = \sum_{i=1}^{4^k} (Q_i - S_i)^2,$$

where Q_i is the frequency of the i-th tuple of length k in Q, and S_i is the frequency of the i-th tuple of length k in S. Following Yang and Zhang (2008), we used tuples of length $k = 5$.

The average common substring length, as defined by Ulitsky et al. (2006), is identical to the average common shustring length defined above minus one. Given the average common substring length, $L(Q, S)$, the corresponding distance is defined as the average between the two versions of

$$d_{\mathrm{acs}}(Q,S) = \log(|S|)/L(Q,S) - 2\log(|Q|)/|Q|.$$

2.5. Cluster analysis

Square matrices of the three distances investigated—K_r, d_{acs}, and d_{ktup}—were subjected to cluster analysis using the bionj algorithm as implemented in the program bionj (Gascuel, 1997). The resulting trees were either midpoint rooted or outgroup rooted using retree, which is part of the PHYLIP package (Felsenstein, 2005). Trees were compared using the Symmetric Distance (Robinson and Foulds, 1981) and the Branch Score Distance (Kuhner and Felsenstein, 1994) as implemented in the PHYLIP program treedist (Felsenstein, 2005). Another PHYLIP program, drawgram, was used to draw the cluster diagrams.

Bootstrapping of phylogenies was carried out using the block bootstrap approach with a block size of 500 bp. Consensus trees were constructed from the bootstrapped trees using the PHYLIP program consense.

2.6. Data sets

Three data sets representing the range of input sizes typically encountered in phylogenomic studies were analyzed: 27 primate mitochondrial genomes (446.23 kb total), complete genomes of eight strains of the bacterial pathogen *Streptococcus agalactiae* (17.39 Mb), and complete genomes of 12 *Drosophila* species (2.03 Gb).

The primate mitochondrial genomes were downloaded from GenBank and analyzed without further editing (Table 1).

The *S. agalactiae* strains, which had previously been analyzed by Tettelin et al. (2005), were downloaded from GenBank (Table 2). All contig sequences of a single strain were concatenated before analysis. Multilocus sequence data for the sequence types corresponding to these strains was obtained from the database mlst.net (Aanensen and Spratt, 2005).

The *Drosophila* dozen genomes consisting of up to 14,547 contigs were downloaded from http://rana.lbl.gov/drosophila/caf1/all_caf1.tar.gz. Unsequenced regions marked by Ns were removed as they would inflate the average shustring length. In addition, the chromosome or contig sequences of each species were concatenated before analysis.

Table 1. Primate Mitochondrial Genomes Analyzed in This Study

No.	*Name*	*Genbank common name*	*Accession*
1	*Cebus albifrons*	White-fronted capuchin	NC_002763.1
2	*Chlorocebus aethiops*	African green monkey	NC_007009.1
3	*Chlorocebus pygerythrus*	Green monkey	NC_009747.1
4	*Chlorocebus sabaeus*	Green monkey	NC_008066.1
5	*Chlorocebus tantalus*	Green monkey	NC_009748.1
6	*Colobus guereza*	Guereza	NC_006901.1
7	*Cynocephalus variegatus*	Sunda flying lemur	NC_004031.1
8	*Gorilla gorilla*	Western Gorilla	NC_001645.1
9	*Homo sapiens*	Human	NC_001807.4
10	*Hylobates lar*	Common gibbon	NC_002082.1
11	*Lemur catta*	Ring-tailed lemur	NC_004025.1
12	*Macaca mulatta*	Rhesus monkey	NC_005943.1
13	*Macaca sylvanus*	Barbary ape	NC_002764.1
14	*Nasalis larvatus*	Proboscis monkey	NC_008216.1
15	*Nycticebus coucang*	Slow loris	NC_002765.1
16	*Pan paniscus*	Pygmy chimpanzee	NC_001644.1
17	*Pan troglodytes*	Chimpanzee	NC_001643.1
18	*Papio hamadryas*	Hamadryas baboon	NC_001992.1
19	*Pongo pygmaeus*	Bornean orangutan	NC_001646.1
20	*Pongo pygmaeus abelii*	Sumatran orangutan	NC_002083.1
21	*Presbytis melalophos*	Mitred leaf monkey	NC_008217.1
22	*Procolobus badius*	Western red colobus	NC_008219.1
23	*Pygathrix nemaeus*	Douc langur	NC_008220.1
24	*Pygathrix roxellana*	Golden snub-nosed monkey	NC_008218.1
25	*Semnopithecus entellus*	Hanuman langur	NC_008215.1
26	*Tarsius bancanus*	Horsfield's tarsier	NC_002811.1
27	*Trachypithecus obscurus*	Dusky leaf monkey	NC_006900.1

Table 2. *Streptococcus agalactiae* Genomes and the Corresponding Multilocus Sequence Types Analyzed in This Study

No.	*Strain*	*Accession*	*Sequence type*
1	18RS21	AAJO01000000	ST19
2	2603V/R	AAJP01000000	ST110
3	515	AAJQ01000000	ST23
4	NEM316	AAJR01000000	ST23
5	A909	AAJS01000000	ST7
6	CJB111	CP000114	ST1
7	COH1	AE009948	ST17
8	H36B	AL732656	ST6

3. RESULTS

3.1. Simulation study

We simulated pairs of sequences with a known number of substitutions per site, K. The pairwise distances d_{ktup}, d_{acs}, and K_{r} were then computed and compared to K. Figure 2A shows that, for $0 \leq K \leq 0.55$, the value of d_{ktup} remains close to zero throughout. In contrast, d_{acs} grows rapidly with low values of K and then levels off. K_{r}, finally, approximates moderate values of K very well. For greater evolutionary distances, the expectation of K_{r} continues to grow with K, but the error bars also grow substantially and the estimated values are not centered on the true values any more (Fig. 2B). This "phase transition" in the behavior of K_{r}

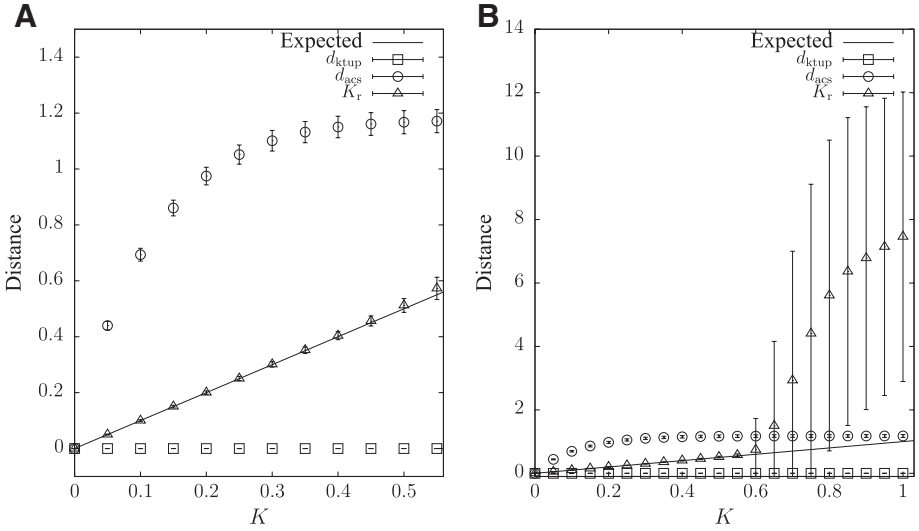

FIG. 2. Pairwise distances as a function of the number of substitutions per site, K. (**A**) Range of substitutions/site (K) values that are well approximated by K_{r}. (**B**) Range of K values with "phase transition" of K_{r}. Each symbol represents the mean $\pm$ standard deviation of 10^4 iterations with sequence pairs of length 100 kb each and GC content of 0.5.

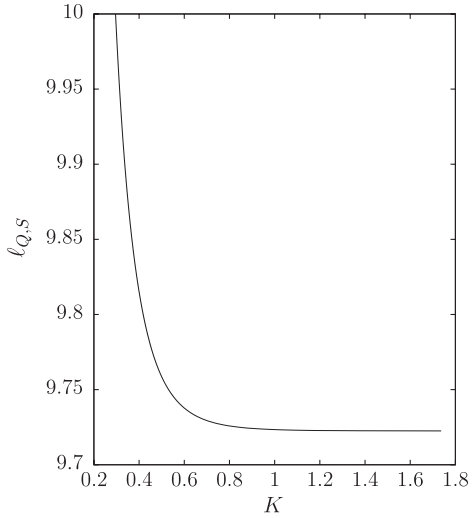

FIG. 3. Average shustring length ($\ell_{Q,S}$) as a function of the number of substitutions per site, K (sequence length, 100 kb; GC content, 0.5).

as a function of K is due to the fact that for large values of K the slope of the average shustring length rapidly approaches zero (Fig. 3).

3.2. Clustering primate mitochondrial genomes

Figure 4 compares the cluster diagrams of primate mitochondrial genomes based on d_{ktup} (**A**), d_{acs} (**B**), K_{r} (**C**), and alignment (**D**). The d_{ktup} tree is quite rough and contains numerous unresolved nodes. Much more resolution is obtained when using d_{acs}. For instance, the ape clade marked by an asterisk ($*$, Fig. 4B) has the well-known correct topology. In contrast, *Papio hamadyras* in the clade marked by a bullet ($\bullet$) ought to

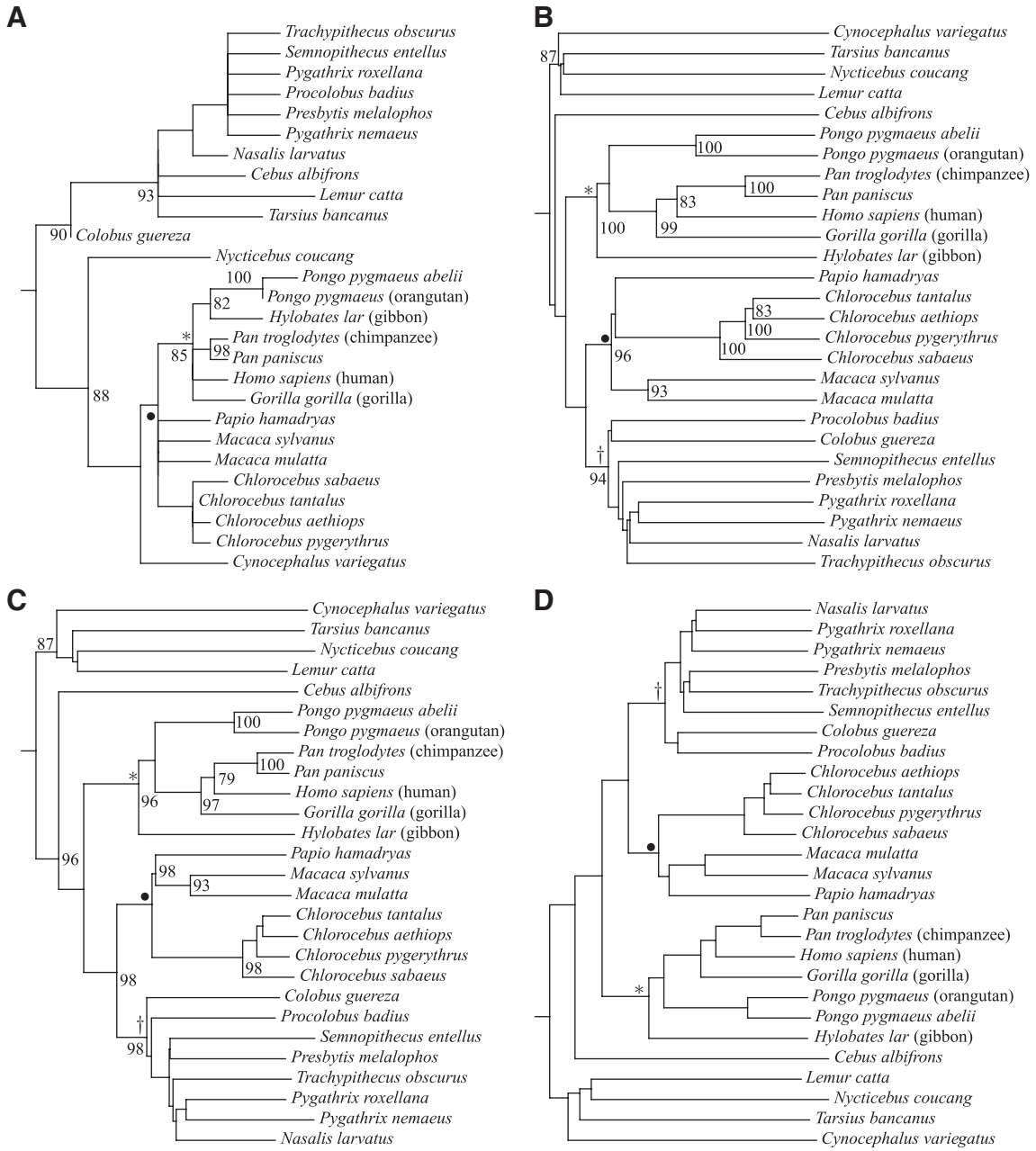

FIG. 4. Midpoint-rooted neighbor joining trees of 27 primate mitochondrial genomes. (**A**) Based on d_{ktup}. (**B**) Based on d_{acs}, bootstrap values $\geq$75% shown. (**C**) Based on K_{r}, bootstrap values $\geq$75% shown. (**D**) Alignment-based tree, bootstrap support for all nodes $\geq$95%. $*$, ape clade (Hominoidea); •, Cercopithecinae among the Old World monkeys (Cercopithecidae), †, Colobinae.

cluster with the macaques rather than with the green monkeys (*Chlorocebus*). K_{r} resolves both of these clades correctly (Fig. 4C). However, the Colobinae (†) remain clustered incorrectly by all three alignment-free distance measures.

Figure 4B,C also displays bootstrap values of 75% or higher. All nodes in the corresponding alignment-based tree had bootstrap values of 95% or higher (Fig. 4D). The lower self-consistency of our method compared to an alignment-based approach is due to the relatively small amount of sequence data provided by mitochondrial genomes ($\approx$16.5 kb). We will return to this issue when analyzing the roughly 100 times longer genomes of *S. agalactiae* below.

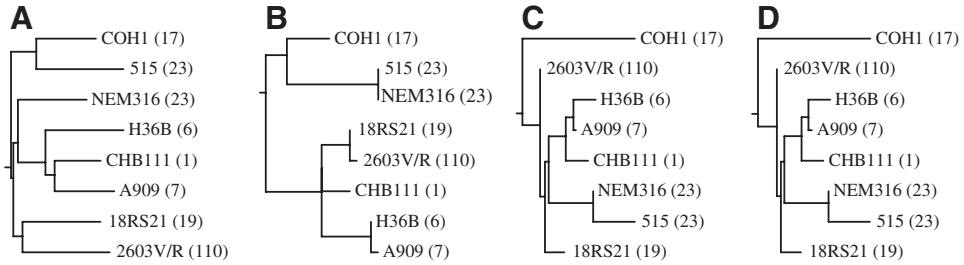

FIG. 5. Midpoint-rooted neighbor joining trees of eight strains of *Streptococcus agalactiae*; strain designations are followed by multilocus sequence types in brackets. (**A**) Gene content tree redrawn from Tettelin et al. (2005). (**B**) Tree based on aligned MLST data. (**C**) Tree based on d_{acs} applied to complete genomes. (**D**) Tree based on K_{r} applied to complete genomes.

Bootstrap support quantifies the consistency of different parts of the input data with respect to the applied tree reconstruction method (Felsenstein, 1985). It says little about the precision of this method. In order to quantify the precision of the cluster diagrams displayed in Figure 4, we computed the three Branch Score Distances (Kuhner and Felsenstein, 1994) between the alignment-free trees on the one hand (Fig. 4A–C) and the alignment-based tree on the other (Fig. 4D). The distances were ranked as follows

$$K_r(0.07) < d_{\text{acs}}(0.39) < d_{\text{ktup}}(1.54).$$

In other words, the alignment-based tree is most closely approximated by the K_{r} tree. If we measure the distance between the trees in terms of topology only (Symmetric Distance [Robinson and Foulds, 1981]), we get the same ranking of distances to the alignment tree:

$$K_r(8) < d_{\text{acs}}(10) < d_{\text{ktup}}(28).$$

3.3. *Clustering* Streptococcus agalactiae *genomes*

Streptococcus agalactiae causes sepsis in neonates and as a result is an intensely studied group of gram-positive bacteria. Tettelin and colleagues reconstructed the phylogeny of eight strains of *S. agalactiae* from the complete genome sequences using gene content as their distance metric (Tettelin et al., 2005). We reproduce the resulting tree in Figure 5A. The authors were surprised to find that this phylogeny did not cluster strains 515 and NEM316, even though they are members of the same multilocus sequence type. Figure 5B displays a tree based on the available multilocus sequence typing (MLST) data showing the identity of strains 515 and NEM316. This agreed with the trees based on d_{acs} and K_{r}, which both clustered NEM316 and 515 (Fig. 5C,D). Cluster analysis of *S. agalactiae* based on d_{ktup} failed, because all branches of the resulting tree had length zero.

The *S. agalactiae* trees in Figure 5C,D have identical topologies and very similar relative branch lengths. The topological distances of the gene content tree (Fig. 5A) and the $K_{\text{r}}/d_{\text{acs}}$ trees (Fig. 5C,D) on the one hand and the MLST tree on the other (Fig. 5B) are summarized as

$$K_r(6) = d_{\text{acs}}(6) < \text{gene content}(10).$$

The bootstrap support for all nodes on the K_{r} and the d_{acs} tree was 100%. Compare this to the lower bootstrap support for the primate trees in Figure 4B,C, which were based on roughly 1% of the genome length of *S. agalactiae*.

3.4. *Clustering* Drosophila *genomes*

The accepted phylogeny of the twelve sequenced *Drosophila* species is shown for reference in Figure 6A (*Drosophila* 12 Genomes Consortium, 2007). The tree based on the d_{ktup} metric when applied to the complete genomes differs markedly from this; for example, *D. sechellia* is not shown as the closest neighbor of *D. simulans*, as would be expected (Fig. 6B). In contrast, the tree based on d_{acs} does cluster these two species and is on the whole much closer to the accepted phylogeny. Some differences remain, though; for example, *D. ananassae* ought to be part of the melanogaster group. The cluster diagram based

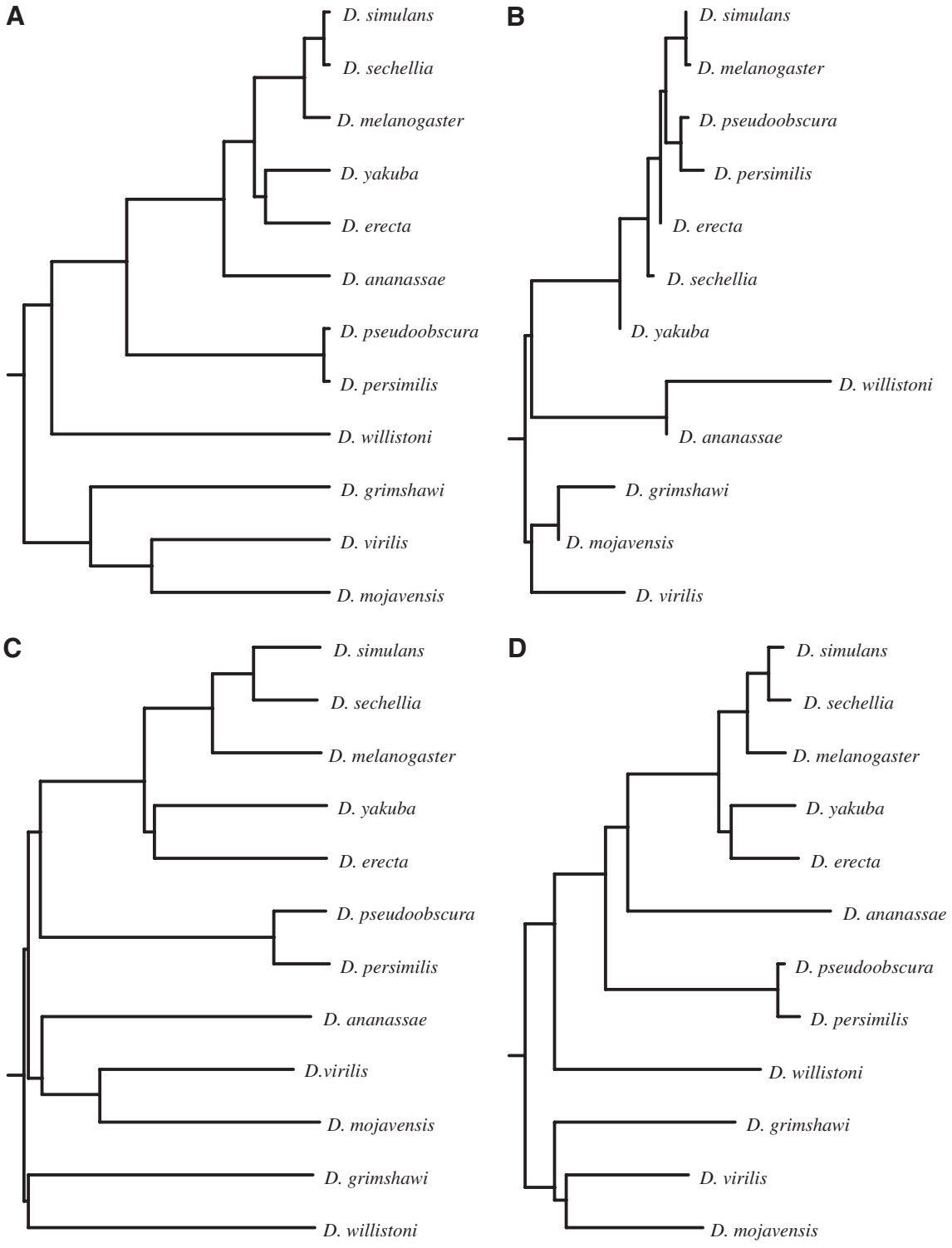

FIG. 6. Outgroup-rooted neighbor-joining trees of 12 *Drosophila* species computed from complete genome sequences. (**A**) Accepted phylogeny (*Drosophila* 12 Genomes Consortium, 2007). (**B**) Based on d_{ktup}. (**C**) Based on d_{acs}. (**D**) Based on K_{r}.

on K_r, finally, is topologically identical to the accepted phylogeny (Fig. 6D). The topological distances between the accepted phylogeny on the one hand (Fig. 6A) and the alignment-free reconstructions on the other (Fig. 6B–D) are ranked as

$$K_r(0) < d_{\text{acs}}(6) < d_{\text{ktup}}(14).$$

4. DISCUSSION

Phylogenies are the central metaphor of evolutionary biology (Archibald, 2008). Any particular phylogeny implies distances between the taxa investigated. Since the advent of molecular biology, taxa are increasingly represented by a sample of their genome sequence, and the computation of distances between molecular sequences has attracted a correspondingly large amount of research interest over the years (Yang, 2006).

Alignment-free distances have for some time now complemented the more widely used alignment-based measures, but it is the current availability of genome-scale sequence data that has sparked renewed interest in alignment-free distances (Vinga and Almeida, 2003). Any such measure needs to fulfil two necessary conditions for usefulness: biological relevance and efficient implementation.

Biological relevance is ensured by explicitly stating the connection between the metric and some known measure of evolutionary distance. We constructed K_r as an estimator of the expected number of substitutions per site, K, by deriving the probability density function of shustring lengths as a function of the number of mismatches per site, d. The resulting equation (7) can be used to compute the expected average shustring length. By equating an observed average shustring length with its expectation, we can solve numerically for d. The relationship between d and the number of substitutions, finally, is given by the oldest distance measure for DNA sequence data, the Jukes-Cantor equation.

The Jukes-Cantor equation has been refined in various ways since its inception, particularly by taking into account nucleotide-specific mutation rates and rate variation between sites (Yang, 2006). In principle, it might be possible to re-derive equation (7) for more complex mutation models. However, Figure 2 shows that at least for the simplest of evolutionary models K_r approaches the number of substitutions per site much better than the other two measures investigated.

The distance measures we compared K_r to are representative of two great classes of alignment-free distance measures: those based on word frequencies (d_{ktup}) and those based on the information-theoretical concept of relative entropy (d_{acs}). k-Tuple distances are easy to compute and are used in various contexts including phylogeny reconstruction in bacteria (Pride et al., 2003), guide tree reconstruction for multiple sequence alignment (Larkin et al., 2007), and phylogeny reconstruction for highly divergent sequences (Yang and Zhang, 2008). However, for the small to moderate evolutionary distances considered in this study, they turned out to be less useful (Figs. 4A and 6B).

The average common substring length is very similar to the average shustring length on which K_r is based. Both measures are therefore restricted to closely related DNA sequences as they saturate with evolutionary distance (Fig. 2B). The comparison between K_r and d_{acs} demonstrated the importance of expressing a quantity such as the average common substring length as a function of specific evolutionary events: for primate mitochondrial DNA and for *Drosophila* genomes K_r outperformed d_{acs} (Figs. 4 and 6). However, when applied to the genomes of *S. agalactiae*, both methods obtained the same topology (Fig. 5B,C). This is an instructive case, as the *S. agalactiae* strains were the most closely related of the taxa analyzed. In fact, they were so closely related that d_{ktup} failed to resolve any phylogeny at all. In contrast, both d_{acs} and K_r returned distances whose topology agreed better with the multilocus sequence data than a published analysis based on gene presence/absence (Tettelin et al., 2005).

Alignment-free distance measures generally trade speed for precision. This is the reason why, for instance, a k-tuple distance is used in clustalw for fast guide tree construction (Larkin et al., 2007). Guide tree construction in ''quicktree'' mode of clustalw is fast indeed, taking only 0.5 s on a 3-GHz Intel Xeon system to compute a tree of the 27 primate mitochondrial genomes. This is orders of magnitudes faster than the default guide tree construction mode based on all pairwise alignments. However, the quality of the resulting tree is approximately as low as that of our d_{ktup} tree shown in Figure 4A. Given that the quality of the K_r tree (Fig. 4C) is much better than that, K_r might be useful as a fast guide tree construction method. This brings us to the question, how fast is K_r compared to other distance measures?

Word-frequency measures have very fast run times that are linear in the combined lengths of the input sequences, $O(l)$. In addition, they only need to store the word frequency table, so their memory requirement is $O(4^k)$. In contrast, the run times of both d_{acs} and K_{r} are dominated by suffix array construction. For the algorithm underlying the suffix array implementation used by us (Manzini and Ferragina, 2002), this would be $O(c \times l \log l)$, where c depends on the time it takes to compare two suffixes. When sorting identical suffixes, this time can become large so that the worst case run time is $O(l^2 \log l)$, but in most applications it is much closer to $O(l \log l)$. In addition to computing a suffix array, K_{r} also requires the estimation of d from equation (7), which makes it somewhat slower than d_{acs}. Here are a few example run times: On a 3 GHz Intel Xeon system it took 0.15 s to compute d_{ktup} for the 27 primate mitochondrial genomes, 8.99 s to compute d_{acs}, and 10.83 s to compute K_{r}. In other words, our implementation of d_{ktup} is over 70 times faster than our implementation of K_{r}. While this may make our distance measure seem slow, it is still fast compared to the computation of the corresponding multiple sequence alignment which takes 8977 s using muscle (Edgar, 2004) and 8657 s using clustalw (Larkin et al., 2007). These programs are designed for protein sequence alignment. In contrast, MAVID is a program designed for aligning multiple long sequences (Bray and Pachter, 2004) and it analyzes the primate data in 16.94 s, which is close to the run time of our more approximate K_{r} measure.

As expected from the complexity formulas, the run time difference increased between d_{ktup} on the one hand and $d_{\text{acs}}/K_{\text{r}}$ on the other when applied to the *Drosophila* data. On our test system d_{ktup} was computed in 6 min and 43 s, while the computation of d_{acs} took roughly 403 times as long and the computation of K_{r} 481 times as long amounting to 2 d, 5 h, 45 min, and 46 s. We tried to align the *Drosophila* data using MAVID, but ran out of memory on a 64 GB computer. Moreover, given that the *Drosophila* data consists of unordered contigs, it is not clear how a global multiple sequence alignment as computed by MAVID might even sensibly be defined.

Notice that the difference in run time between d_{acs} and K_{r} is appreciable. This would make it desirable to approximate the sum in equation (6) by a probability density function. Still, it is remarkable that 2 days suffice to analyze the full genomes of the *Drosophila* dozen and return with the accepted branching order. In their current form the run times of the measures investigated can be summarized as

$$d_{\text{ktup}} \ll d_{\text{acs}} < K_r.$$

It is important to realize that the run time of K_{r} depends on details of how shustrings are looked up and how equation (7) is solved. These might be revised in the future, particularly since efficient algorithms for the construction of suffix arrays are a lively topic of current research (Puglisi et al., 2007). However, this does not materially affect the central contribution of our study, which is to construct an alignment-free estimator of the number of pairwise substitutions. This estimator is not only applicable to unaligned genomes but also to unordered contigs, as was the case with the *Drosophila* dozen.

The work presented here can be extended in at least two ways: by revising the calculation of K_{r}, and by estimating other statistics from the average shustring length. Figure 3 shows that the moment-based estimation of K developed in this paper fails for $K > 0.55$ because then the average shustring length is essentially constant. A maximum likelihood estimator might be more powerful and allow distance estimation for more divergent sequences. This would be important if K_{r} were to be applied in guide tree reconstruction.

As to estimating other statistics, population parameters such as the scaled neutral mutation rate (Hartl and Clark, 1997, p. 319), θ, are a primary target, since genetic diversity is so low that average shustring length would be a very sensitive indicator. We expect that there is going to be heightened interest in alignment-free estimation of such population genetic parameters in the wake of ongoing comparative sequencing projects such as the 1000 genomes projects for *Arabidopsis*, *Drosophila*, and human.

ACKNOWLEDGMENTS

We thank Angelika Börsch-Haubold and an anonymous reviewer for comments that improved this article. This work was supported by the German Federal Ministry of Education and Research (BMBF) through the Freiburg Initiative for Systems Biology (grant 0313921 to P.P.) and the German Research Foundation (grant SFB680 to T.W.).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Aanensen, D., and Spratt, B. 2005. The multilocus sequence typing network: mlst.net. *Nucleic Acids Res.* 33, W728–W733.

Abouelhoda, M.I., Kurtz, S., and Ohlebusch, E. 2002. The enhanced suffix array and its applications to genome analysis. In *Proc. 2nd Workshop Alg. Bioinform.*, 449–463. Lecture Notes in Computer Science 2452, Springer-Verlag.

Adam, Z., and Sankoff, D. 2008. The ABCs of MGR with DCJ. *Evol. Bioinform.* 4, 69–74.

Archibald, J.D. 2008. Edward Hitchcock's pre-Darwinian (1840) "tree of life." *J. History Biol.* DOI 10.1007/s10739-008-9163-y.

Blaisdell, B.E. 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Nat. Acad. Sci. USA* 83, 5155–5159.

Bray, N., and Pachter, L. 2004. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res.* 14, 693–699.

Cover, T.M., and Thomas, J.A. 2006. *Elements of Information Theory*, 2nd ed. Wiley, Hoboken, NJ.

Darwin, C. 1859. *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life* [1985 edition]. Penguin, London.

Dewey, C., and Pachter, L. 2006. Evolution at the nucleotide level: the problem of multiple whole-genome alignment. *Hum. Mol. Genet.* 15 (spec. no. 1), R51–R56.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791.

Felsenstein, J. 2005. PHYLIP (phylogeny interference package), version 3.6.

Gascuel, O. 1997. BIONJ: an improved version of the nj algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* 14, 685–695.

Hartl, D.L., and Clark, A.G. 1997. *Principles of Population Genetics*, 3rd ed. Sinauer, Sunderland, MA.

Haubold, B., Domazet-Lošo, M., and Wiehe, T. 2008. An alignment-free distance measure for closely related genomes. *Lect. Notes Bioinform.* 5267, 87–99.

Haubold, B., Pierstorff, N., Möller, F., et al. 2005. Genome comparison without alignment using shortest unique substrings. *BMC Bioinform.* 6, 123.

Haubold, B., and Wiehe, T. 2006. How repetitive are genomes? *BMC Bioinform.* 7, 541.

Henz, S., Huson, D., Auch, A., et al. 2005. Whole-genome prokaryotic phylogeny. *Bioinformatics* 21, 2329–2335.

Huson, D.H., and Steel, M. 2004. Phylogenetic trees based on gene content. *Bioinformatics* 20, 2044–2049.

Jukes, T.H., and Cantor, C.R. 1969. Evolution of protein molecules, 21–132. In Munro, H.N., ed., *Mammalian Protein Metabolism*, *Volume 3*, Academic Press, New York.

Kauffman, S.A. 1993. *The Origin of Order*. Oxford University Press, Oxford.

Kuhner, M.K., and Felsenstein, J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol., Evol.* 11, 459–468.

Larkin, M., Blackshields, G., Brown, N., et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948.

Lunter, G., Ponting, C.P., and Hein, J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLOS Comput. Biol.* 2, 2–12.

Manzini, G., and Ferragina, P. 2002. Engineering a lightweight suffix array construction algorithm. *Proc. 10th Annu. Eur. Symp. Algorithms* (*ESA '02*) 698–710.

Pride, D.T., Meinersmann, R.J., and Wassenaar, T.M. 2003. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.* 13, 145–158.

Puglisi, S.J., Smyth, W.F., and Turpin, A. H. 2007. A taxonomy of suffix array construction algorithms. *ACM Comput. Surv.* 39, 4.

Robinson, D.F., and Foulds, L.R. 1981. Comparison of phylogenetic trees. *Math. Biosci.* 53, 514–525.

Tettelin, H., Masignani, V., Cieslewicz, M., et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae:* implications for the microbial "pan-genome." *Proc. Natl. Acad. Sci. USA* 102, 13950–13955.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450, 203–218.

Ulitsky, I., Burstein, D., Tuller, T., et al. 2006. The average common substring approach to phylogenomic reconstruction. *J. Comput. Biol* 13, 336–350.

Vinga, S., and Almeida, J. 2003. Alignment-free sequence comparison—a review. *Bioinformatics* 19, 513–523.
Waterman, M.S. 1995. *Introduction to Computational Biology; Maps, Sequences and Genomes*. Chapman & Hall/CRC, London.
Yang, K., and Zhang, L. 2008. Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction. *Nucleic Acids Res*. 36, e33.
Yang, Z. 2006. *Computational Molecular Evolution*. Oxford University Press, New York.
Zuckerkandl, E., and Pauling, L. 1965. Molecules as documents of evolutionary history. *J. Theoret. Biol.* 8, 357–366.

Address correspondence to:
Dr. Bernhard Haubold
Max-Planck-Institute for Evolutionary Biology
Department of Evolutionary Genetics
Plön, Germany

E-mail: haubold@evolbio.mpg.de