An Alignment-Free Distance Measure for Closely Related Genomes

Bernhard Haubold[1], Mirjana Domazet-Lošo[1,2], and Thomas Wiehe[3]

[1] Max-Planck-Institute for Evolutionary Biology, Department of Evolutionary Genetics, Plön, Germany
[2] Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
[3] Institute of Genetics, Universität zu Köln, Cologne, Germany

Abstract. Phylogeny reconstruction on a genome scale remains computationally challenging even for closely related organisms. Here we propose an alignment-free pairwise distance measure, K_{r}, for genomes separated by less than approximately 0.5 mismatches/nucleotide. We have implemented the computation of K_{r} based on enhanced suffix arrays in the program `kr`, which is freely available from `guanine.evolbio.mpg.de/kr/`. The software is applied to genomes obtained from three sets of taxa: 27 primate mitochondria, eight *Staphylococcus agalactiae* strains, and 12 *Drosophila* species. Subsequent clustering of the K_{r} values always recovers phylogenies that are similar or identical to the accepted branching order.

1 Introduction

Gene phylogenies do not necessarily coincide with organism phylogenies. This well known observation leads to the idea of reconstructing phylogenies from all available genetic information, that is, from complete genomes. In fact, the study of whole genome phylogenies started as soon as suitable data became available [8]. In spite of much progress since then, the computational obstacles to such analyses are still considerable and a good part of bioinformatics is concerned with solving them [6].

To the uninitiated the reconstruction of genome phylogenies might appear to simply involve the scaling up of available techniques for reconstructing gene phylogenies: compute a multiple sequence alignment and estimate the genealogy from that. However, in the wake of the first genome projects it proved difficult if not impossible to scale existing gene-centered alignment software from input of a few kilo bases to several mega bases. This left two avenues to explore: development of more efficient alignment algorithms and development of alignment-free methods of distance computation.

In the years following publication of the first genomes of free-living organisms, phylogenomics—as the field concerned with reconstructing phylogenies from genomes became known—made great strides on both counts [14]. Alignment algorithms and alignment tools have received most attention as they are useful in many sequence comparison tasks [6]. In contrast, alignment-free sequence comparison has a more narrow applicability, the classical case being phylogeny reconstruction from pairwise distances [3]. The great advantage of this approach is that it obviates the computationally intensive alignment step. In fact, alignment-free distance measures may even be

C.E. Nelson and S. Vialette (Eds.): RECOMB-CG 2008, LNBI 5267, pp. 87–99, 2008.

used in the computation of multiple sequence alignments. For example, pairwise distances based on exact word (k-tuple) matches [27] underlie the fast mode of guide tree construction in the popular multiple sequence alignment program `clustalw` [18].

Two classes of methods for alignment-free sequence comparison can be distinguished: (i) methods based on word frequencies, the utility of which may depend on the word length chosen, and (ii) resolution-free methods, where no such parameter choice is necessary [26]. These methods have been applied to, for example, phylogeny reconstruction from γ-protobacterial genomes [5] and the analysis of regulatory sequences in metazoan genomes [17]. One disadvantage of alignment-free methods is that there is generally no model to map their results to evolutionary distances. Models describing the mutation probabilities of homologous nucleotides have been continuously refined since the pioneering work on this topic by Jukes and Cantor in the late 1960's [16,29]. However, a recent study indicates that k-tuple distances may be highly accurate when compared to conventional model-based distances [28].

We have developed a new alignment-free distance measure, which we call K_{r}. The central idea of our approach is that closely related sequences share longer exact matches than distantly related sequences. In the following we derive K_{r}, describe its implementation, and demonstrate its utility through simulation. We then apply it to three data sets of increasing size: 27 primate mitochondrial genomes, eight complete genomes of the bacterial pathogen *Streptococcus agalactiae*, which is a leading cause of bacterial sepsis in neonates [24], and the complete genomes of twelve species of *Drosophila* [25]. In each case cluster analysis of K_{r} values recovers a topology that is close or identical to the accepted phylogeny.

2 Approach and Data

2.1 Definition of K_{r}

Consider two sequences, $Q = \mathtt{TATAC}$ and $S = \mathtt{CTCTGG}$, which we call *query* and *subject*, respectively. For every suffix of Q, $Q[i..|Q|]$, we look up the shortest prefix, $Q[i..j]$, that is absent from S. This special prefix is called a *Shortest Absent Prefix* (SAP) and denoted by q_i. We start by examining the first suffix of our example query, which covers the entire sequence: $Q[1..|Q|] = \mathtt{TATAC}$. Its shortest prefix, $Q[1..1] = \mathtt{T}$, does occur in S and hence we extend it by one position to get $Q[1..2] = \mathtt{TA}$, which is absent from S yielding our first SAP, $q_1 = \mathtt{TA}$. Next we determine the shortest prefix of $Q[2..|Q|] = \mathtt{ATAC}$ that is absent from S and find $q_2 = \mathtt{A}$, and so on. Notice that there is no prefix of $Q[5..|Q|] = \mathtt{C}$ that is absent from S. In this case we define $q_i = Q[i..|Q|+1]$; in other words, we pretend that Q (and S) are terminated by a unique sentinel character (\$) to guarantee that q_i exists for all i. Finally we have the SAPs $q_1 = \mathtt{TA}, q_2 = \mathtt{A}, q_3 = \mathtt{TA}, q_4 = \mathtt{A}$, and $q_5 = \mathtt{C\$}$.

Our algorithm is based on the lengths of the SAPs, $|q_i|$. The key insight leading from these lengths to a distance measure is that if Q and S are closely related, they are characterized by many long exact repeats between Q and S. As a consequence, SAP lengths will tend to be greater than if Q and S are only distantly related.

To make this notion rigorous, we define the observed aggregate SAP length

$$A_{\mathrm{o}} = \sum_{i=1}^{|Q|} |q_i|$$

and its expectation, A_{e}, which can be computed either analytically [12] or through shuffling of S. Next, we take the logarithm of $A_{\mathrm{o}}/A_{\mathrm{e}}$ and normalize this quantity by the maximum value it can take to define the index of repetitiveness, I_{r}

$$I_{\mathrm{r}}(Q,S) = \frac{\ln(A_{\mathrm{o}}/A_{\mathrm{e}})}{\ln(\max(A_{\mathrm{o}})/A_{\mathrm{e}})}, \tag{1}$$

where

$$\max(A_{\mathrm{o}}) = \begin{cases} \binom{|Q|+2}{2} - 1 & \text{if } |Q| \le |S| \\ (|Q| - |S| + 1)(|S| + 1) + \binom{|S|+1}{2} - 1 & \text{otherwise.} \end{cases} \tag{2}$$

We therefore have

$$\sim 0 \le I_{\mathrm{r}} \le 1.$$

The ceiling of the I_{r} domain is exact—any pair of identical query and subject sequences are maximally repetitive and have $I_{\mathrm{r}} = 1$. In contrast, the floor is an expectation for reasonably long shuffled sequences of any GC content. The definition of I_{r} presented here extends an earlier version [13] by adding the query/subject distinction and the normalization.

We used simulations to explore the relationship between I_{r} and the number of pairwise mismatches per nucleotide, d. One thousand pairs of 10 kb long sequences with a fixed d were generated and Figure 1 displays d as a function of simulated $\ln(I_{\mathrm{r}})$ values. The shape of the bottom right hand part of the curve tells us that in pairs of similar sequences few mutations have a large effect on I_{r}. We found that the relationship between divergence and I_{r} could conveniently be modeled with the statistical software R [22] using two logistic functions, one covering $\ln(I_{\mathrm{r}}) > -2.78$ and the other covering the rest. Given these two functions, we define the number of pairwise differences based on the I_{r}, d_{r}:

$$d_{\mathrm{r}} = \begin{cases} \frac{0.1380}{1+e^{(-2.2016-\ln(I_{\mathrm{r}}))/-0.5307}} & \text{if } \ln(I_{\mathrm{r}}) > -2.78 \\ \frac{0.6381}{1+e^{(-5.5453-\ln(I_{\mathrm{r}}))/-1.7113}} & \text{otherwise.} \end{cases} \tag{3}$$

The dashed line in Figure 1 indicates that this model gives a useful approximation of the simulated values shown as dots.

Finally, the number of pairwise mismatches, d_{r}, was converted into our distance measure, K_{r}, using the formula by Jukes and Cantor [16]:

$$K_{\mathrm{r}} = -\frac{3}{4} \ln\left(1 - \frac{4}{3} d_{\mathrm{r}}\right). \tag{4}$$

2.2 Asymmetric Values of K_{r}

In general and depending on which sequence is designated query, the two resulting I_{r} values differ, that is, $I_{\mathrm{r}}(S_1, S_2) \neq I_{\mathrm{r}}(S_2, S_1)$. Direct application of equations (3) and (4)

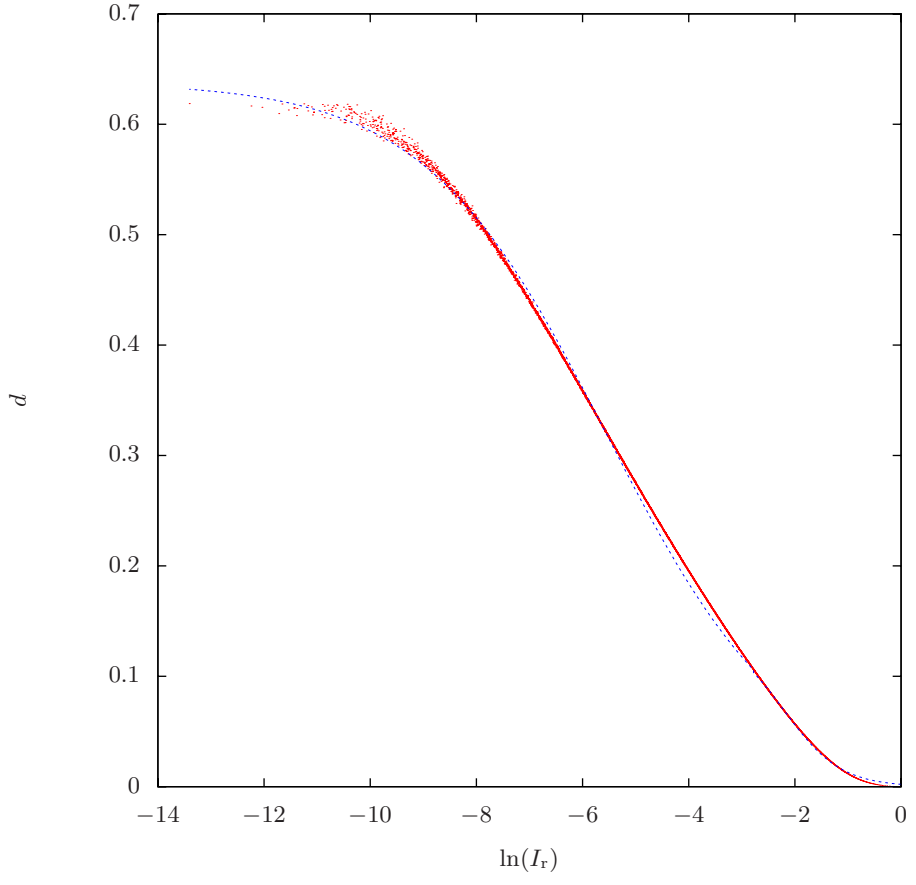

Fig. 1. Simulated (dots) and modeled (dashed) relationship between the number of pairwise differences per site, d, and the index of repetitiveness, I_{r}. Each dot represents an average of 1000 I_{r} values calculated from 1000 pairs of sequences characterized by a given value of d. The model relationship is stated in Equation (3).

would translate this inequality into asymmetric matrices of K_{r} values, which is unacceptable for a metric. In the case of two "ideal" sequences devoid of insertions/deletions and repetitive elements, the difference is due to stochastic placement of mutations along a DNA sequence. However, indels and shared repetitive elements may cause systematic differences between the two possible query/subject configurations.

Figure 2A shows an example in which S_1 has undergone large deletions and as a result is much shorter than S_2, i.e. S_2 is only locally homologous to S_1. In this case $I_{\mathrm{r}}(S_1, S_2) < I_{\mathrm{r}}(S_2, S_1)$. However, regions in S_2 that have no homologue in S_1 are characterized by SAPs that are only as long as expected by chance. We have therefore implemented a global and a local mode for K_{r} computation. In the global mode all SAPs are included in the analysis. In the local mode the user can set the fraction,

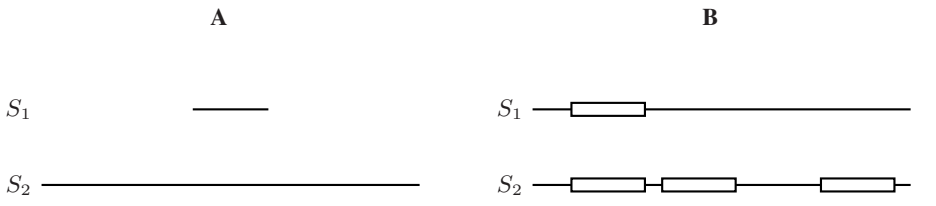

Fig. 2. Sources of asymmetric I_r values. **A**: S_2 is only locally homologous to S_1, in which case $I_\mathrm{r}(S_1, S_2) < I_\mathrm{r}(S_2, S_1)$; **B**: S_1 contains a lower copy number of a genetic element than S_2, in which case again $I_\mathrm{r}(S_1, S_2) < I_\mathrm{r}(S_2, S_1)$.

say 0.5, of SAP lengths compatible with randomness that are excluded from the analysis. All applications to real data presented in this paper were computed using the local mode.

Figure 2B illustrates variation in the copy number of a shared element: S_1 contains one copy of the element and S_2 three, which again leads to $I_\mathrm{r}(S_1, S_2) < I_\mathrm{r}(S_2, S_1)$. Since many mutations are necessary to reverse the effect of a single gene duplication on I_r, we always chose the lower of the two values for the computation of K_r.

2.3 Implementation

Conceptually, SAP lengths are determined in a single bottom-up traversal of a generalized sufix tree [11] containing the forward and reverse strands of the query and subject data sets. Each internal node in this tree, n, is classified as *isQuery* if the subtree rooted on it has leaves referring to positions in the query sequences, and as *isSbjct*, if the subtree rooted on it has leaves referring to positions in the subject sequences. Both properties propagate up the tree. If n *isQuery* and *isSbjct*, its child nodes, c_i, are searched for two relevant cases: First, c_i may be a leaf referring to a query position x. In that case the desired SAP length, $|q_x|$, is the string depth of c_i plus 1. Second, c_i may be an internal node with the property *isQuery* but not *isSbjct*. Then the leaves of the subtree rooted on c_i are looked up and the string depth of c_i plus 1 is the desired length of the SAPs referred to by these leaves.

We based the implementation of the suffix tree traversal on its more space-efficient sister data structure, the enhanced suffix array [2]. For this purpose we used the suffix array library by Manzini and Ferragina [19], as it is fast and space-efficient [21]. In its original form, the library was limited to the analysis of $2^{31} \approx 2 \times 10^9$ characters, which we have re-engineered to a limit of $2^{63} \approx 9 \times 10^{18}$ characters.

Our program for calculating the K_r is called `kr`. It takes as input a set of FASTA-formatted sequences and returns a distance matrix in PHYLIP [10] format. The program can be accessed via a simple web interface at

`http://guanine.evolbio.mpg.de/kr/`

The C source code of `kr` is also available from this web site under the GNU General Public License.

2.4 Phylogenetic Analysis

Phylogenies based on sequence alignments were computed using the neighbor joining algorithm [23] implemented in `clustalw` [18]. Phylogenies based on K_{r} values were computed using the neighbor joining algorithm implemented in the software package PHYLIP [10]. Phylogenetic trees were also drawn using PHYLIP.

It is highly desirable to attach confidence measures to individual nodes in a phylogeny. A popular method for achieving this is bootstrap analysis [7]. The central question in any bootstrap analysis is, what is the unit to be sampled with replacement (bootstrapped)? In traditional bootstrap analysis of phylogenies, columns of homologous nucleotides are sampled with replacement from the underlying multiple sequence alignment [9]. This cannot be applied in the context of an alignment-free distance measure such as K_{r}. Instead, we propose to sample random fragments of 500 bp length with replacement from the original sequences.

Table 1. Primate mitochondrial genomes analyzed in this study

#	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

#	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

2.5 Data Sets

Three sets of genomes were analyzed: 27 primate mitochondrial genomes (total of 446.23 kb), genomes of eight *S. agalactiae* strains (17.39 Mb), and the genomes of twelve *Drosophila* species (2.03 Gb).

The 27 primate mitochondrial genomes available from Genbank were downloaded and compared without any further editing (Table 1).

The eight *S. agalactiae* genomes previously analyzed by [24] were downloaded from Genbank and subjected to K_{r} computation without further editing (Table 2). Complete multilocus sequence data for the sequence types corresponding to the these genomes was obtained from `mlst.net` [1].

The 12 *Drosophila* genomes consisting of up to 14,547 contigs each were downloaded from

`http://rana.lbl.gov/drosophila/caf1/all_caf1.tar.gz`

Unsequenced regions in these genomes marked by `N` were removed before K_{r} analysis, as these generate suffixes with long matching prefixes that distort the K_{r}.

3 Results

3.1 Clustering of Simulated DNA Sequences

Figure 1 demonstrates that the model underlying the computation of K_{r} is reasonably exact for divergence $d \leq 0.5$, which roughly corresponds to $\ln(I_{\mathrm{r}}) \geq -8$, or $I_{\mathrm{r}} \geq 0.0003$. In order to explore the utility of K_{r} for sequence clustering, we simulated a set of 12 DNA sequences of 10 kb with a maximal d of 0.5, that is, by distributing 5000 segregating sites on a random topology generated using the coalescent simulation program `ms` [15]. The true phylogeny of these sequences is shown in Figure 3A. Neighbor joining analysis of the 66 Jukes-Cantor distances between the dozen simulated sequences yielded the phenogram shown in Figure 3B, which is topologically identical to the true phylogeny. The branch lengths of Figure 3A and B also look almost indistinguishable. However, they differ in numerical detail as illustrated for the edges

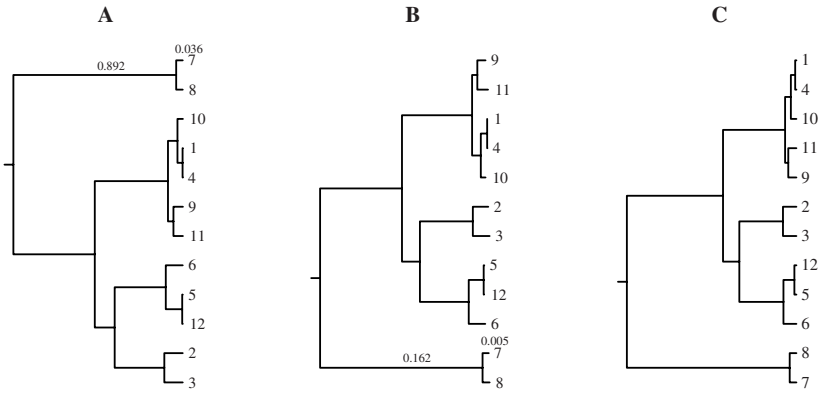

Fig. 3. Reconstructing the phylogeny of 12 simulated sequences. **A:** True phylogeny; **B:** phylogeny based on multiple sequence alignment by `clustalw` [18]; **C:** phylogeny based on K_{r}. The small numbers on the edges leading from taxon 7 to the root illustrate branch length differences between phylogenies **A** and **B**.

connecting taxa 7 and 8 to the root. Phylogeny reconstruction based on K_{r} returned the tree shown in Figure 3C. It is topologically identical to the cluster diagram based on standard pairwise distances (Figure 3B). Again, the branch lengths also look very similar but the diagram reveals small differences such as the distance between taxa 5 and 12, which is larger in the K_{r} phylogeny than in the other two. We shall see that the K_{r} measure has a tendency to overestimate terminal branch lengths.

Next we investigate the performance of K_{r} when applied to real sequences.

3.2 Clustering Primate Mitochondrial Genomes

Figure 4 displays two phylogenies of primate mitochondrial genomes, one based on K_{r} (A), the other on a multiple sequence alignment by `clustalw` (B). The two trees share important clades, particularly groups of closely related taxa. For example, the well-known great ape clade (asterisk in Figure 4) is resolved correctly using K_{r}. In contrast, within the Cercopithecinae (bullet in Figure 4) *Pio hamadryas* ought to cluster with the macaques (Figure 4B) rather than with the green monkeys (*Chlorocebus*, Figure 4A).

3.3 Clustering *Streptococcus agalactiae* Genomes

Tettelin and colleagues analyzed the complete genomes of eight *S. agalactiae* strains and reconstructed their phylogeny by comparing gene content [24]. Surprisingly, they obtained a phylogeny that did not cluster strains 515 and NEM316, even though these belong to the same multilocus sequence type (ST23; Table 2). In our K_{r} phylogeny of complete genomes these strains again appear as closest neighbors with 100% bootstrap support (Figure 5A). Overall the topology of this phylogeny is similar to a `clustalw` tree based on multilocus sequence data (Figure 5B). In contrast to the topology, the branch lengths derived from the two methods differ markedly, with the

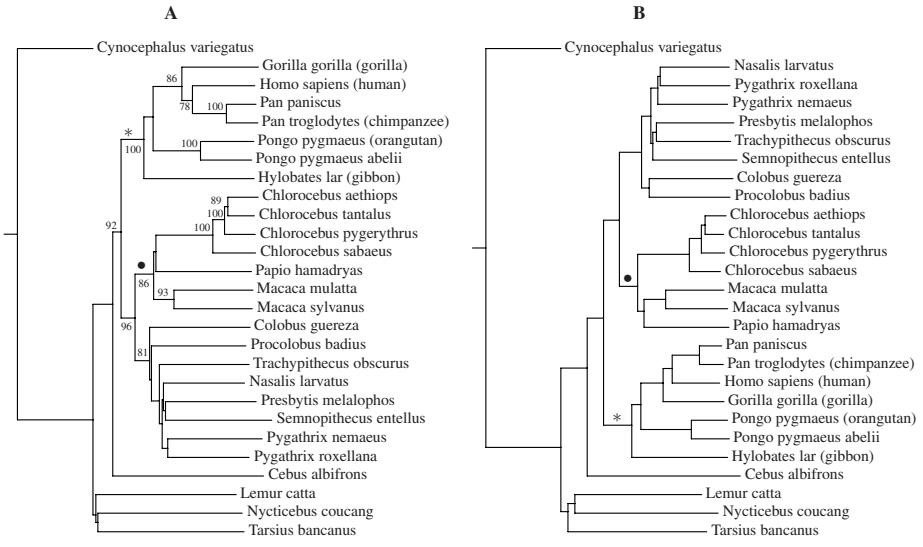

Fig. 4. Phylogeny of 27 primate mitochondrial genomes. The asterisk ($*$) marks the ape clade (Hominoidea), the bullet ($\bullet$) the Cercopithecinae among the old world monkeys (Cercopithecidae). **A**: Distance estimates based on K_{r}, bootstrap (100 replicates) greater than 75% are shown; **B**: distances based on multiple sequence alignment, all bootstrap values were greater than 95%.

external branches being much longer in the K_{r} tree. This is not simply a consequence of the K_{r} tree being computed from whole genomes and the `clustalw` tree from multilocus sequence data. When we subjected the same multilocus sequence data to K_{r}

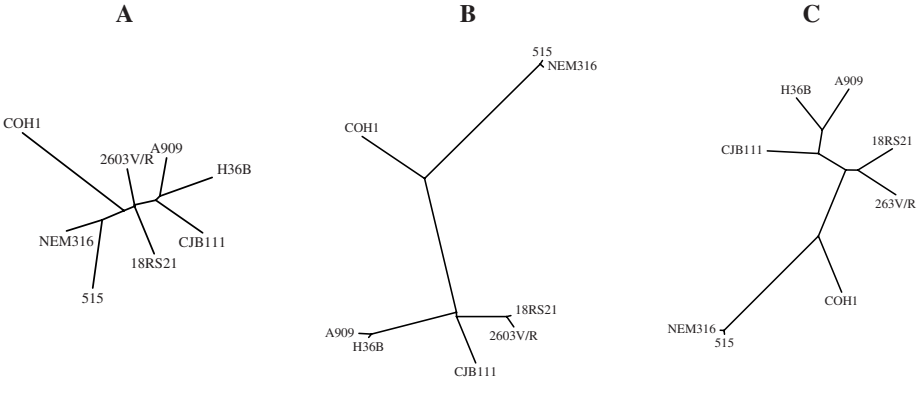

Fig. 5. Phylogenies of eight *Streptococcus agalactiae* strains. **A**: Based on K_{r} and whole genomes, all bootstrap values (100 replicates) were 100%; **B**: same set of organisms as **A**, but tree based on an alignment of multilocus sequence data using `clustalw`; **C**: same organisms and data as in **B**, but clustering based on K_{r}.

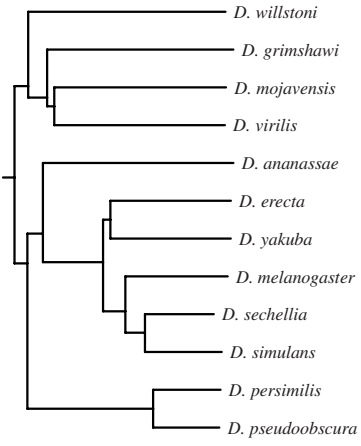

Fig. 6. Midpoint-rooted neighbor-joining tree of 12 *Drosophila* species based on K_{r} and complete genome sequences

analysis, we obtained the tree shown in Figure 5C. This is topologically similar to the alignment-based tree but has longer terminal branches.

3.4 Clustering *Drosophila* Genomes

Calculating the K_{r} values for the 12 *Drosophila* species investigated took four days and 18 hours of CPU time on a computer with 64 GB RAM. The resulting phylogeny in Figure 6 has the same topology as the tree computed as part of the *Drosophila* dozen project [25].

4 Discussion

In this study we calculate the phylogeny of 12 *Drosophila* species from their raw genome sequences using a new measure of sequence similarity, K_{r}. This is defined with ease of implementation and scalability in mind. For this reason the underlying idea is simple: if we compare a query to a closely related subject, for every suffix taken from the query one finds on average a suffix in the subject with a long common prefix. Specifically, we concentrated on the shortest prefixes of query suffixes that are absent from the subject. The entire computation of K_{r} is based on the lengths of these shortest absent prefixes, SAPs. There are three reasons for this: (i) SAPs are on average longer for closely related pairs of sequences than for divergent pairs; (ii) we have previously derived the distribution of SAP lengths expected by chance alone [12], which allows us to normalize the observed lengths by their expectation; and (iii) the exact matching strategy for distance computation we propose is very quick as it is based on enhanced suffix array traversal [2].

Technicalities aside, our approach is to transform exact match lengths to distances using the Jukes Cantor model [16]. This is the oldest and simplest model of nucleotide

evolution. It is clear that its application across species with strong intra-genomic variation in mutation rates as observed in *Drosophila* [20] violates the model assumption of rate uniformity across residues and positions. However, the very large amount of sequence information contained in the *Drosophila* genomes leads to the recovery of the correct clades from K_{r} in spite of the simplifications of the model.

The trade-off between speed and precision is well known in the field of sequence alignment. For example, `clustalw` has a slow, accurate and a fast, approximate mode ("quicktree") for guide tree computation. Like our K_{r} calculation, the fast mode of guide tree reconstruction is based on alignment-free pairwise sequence comparison. However, `kr` is both faster and more sensitive than the quicktree mode. For example, `kr` takes half as long as `clustalw` in quicktree mode to compute the guide tree for the 27 primate mitochondrial genomes. The difference in run time grows to 12-fold for a simulated sample of 27 sequences that are 100 kb long, that is 6 times longer than the primate mitochondrial genomes. In addition, K_{r} tends to resolve closely related sequences better than the quicktree mode (not shown).

The reason for this sensitivity to small differences in sequence similarity was apparent in the long terminal branches of the phylogeny based on multilocus sequence data (Figure 5C) compared to the alignment-based phylogeny (Figure 5B). This emphasis on recent mutations is already apparent in the simulated relationship between divergence, d, and I_{r} (Figure 1). The lower right corner of this graph indicates that the addition of few mutations to a pair of identical sequences has a strong effect on the I_{r} and hence on K_{r}. This suggests great sensitivity to differences among closely related sequences, leading to the long terminal branches observed in *S. agalactiae* (Figures 5A and C) and *Drosophila* (Figure 6). Sensitivity and speed of execution make `kr` a promising tool for the computation of guide trees that can be used as input to multiple sequence alignment programs such as `clustalw` or the more powerful `MAVID` [4].

The sensitivity of K_{r} restricts its application to closely related DNA sequences, which is an important limitation of our method. Figure 1 allows us to quantify the range of diversity values for which K_{r} computations might be attempted: For divergence values greater than 0.5 the relationship between d and I_{r} becomes increasingly noisy. Under the Jukes-Cantor model of sequence evolution [16] a d-value of 0.5 corresponds to 0.82 substitutions/site. Substitution rates in *Drosophila* genes vary between 11.0×10^{-9} and 27.1×10^{-9}/site/year [20]. If we take the average of these values (19.5×10^{-9}), we arrive at a maximum evolutionary distance of 43.4 million years for our method. This is approximately the divergence time of the *Drosophila* clade analyzed in Figure 6. Taxa with lower substitution rates could, of course, be analyzed to correspondingly greater evolutionary distances, but this rough calculation illustrates the caveat that K_{r} should only be applied to closely related genomes. Given this proviso, our distance measure gives biologically meaningful results on scales ranging from mitochondrial to metazoan nuclear genomes.

Acknowledgements

We thank Peter Pfaffelhuber, Angelika Börsch-Haubold, and an anonymous reviewer for comments that improved this manuscript.

References

1. Aanensen, D.M., Spratt, B.G.: The multilocus sequence typing network: mlst.net. Nucleic Acids Res. 33(Web Server issue) , W728–W733 (2005)
2. Abouelhoda, M.I., Kurtz, S., Ohlebusch, E.: The enhanced suffix array and its applications to genome analysis. In: Proceedings of the second workshop on algorithms in bioinformatics. Springer, Heidelberg (2002)
3. Blaisdell, B.E.: A measure of the similarity of sets of sequences not requiring sequence alignment. Proceedings of the National Academy of Sciences, USA 83, 5155–5159 (1986)
4. Bray, N., Pachter, L.: MAVID: Constrained ancestral alignment of multiple sequences. Genome Research 14, 693–699 (2004)
5. Chapus, C., Dufraigne, C., Edwards, S., Giron, A., Fertil, B., Deschavanne, P.: Exploration of phylogenetic data using a global sequence analysis method. BMC Evolutionary Biology 5, 63 (2005)
6. Dewey, C.N., Pachter, L.: Evolution at the nucleotide level: the problem of multiple whole-genome alignment. Hum. Mol. Genet. 15(Spec. No. 1), R51–R56 (2006)
7. Efron, B.: Bootstrap methods: another look at the Jackknife. The Annals of Statistics 7, 1–26 (1979)
8. Eisen, J.A.: Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Research 8, 163–167 (1998)
9. Felsenstein, J.: Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791 (1985)
10. Felsenstein, J.: PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle (2005)
11. Gusfield, D.: Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, Cambridge (1997)
12. Haubold, B., Pierstorff, N., Möller, F., Wiehe, T.: Genome comparison without alignment using shortest unique substrings. BMC Bioinformatics 6, 123 (2005)
13. Haubold, B., Wiehe, T.: How repetitive are genomes? BMC Bioinformatics 7, 541 (2006)
14. Hervé, P., Delsuc, F., Lartillot, N.: Phylogenomics. Annual Review of Ecology, Evolution, and Systematics 36, 541–562 (2005)
15. Hudson, R.R.: Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18, 337–338 (2002)
16. Jukes, T.H., Cantor, C.R.: Evolution of protein molecules. In: Munro, H.N. (ed.) Mammalian Protein Metabolism, vol. 3, pp. 21–132. Academic Press, New York (1969)
17. Kantorovitz, M.R., Robinson, G.E., Sinha, S.: A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics 23, i249–i255 (2007)
18. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G.: Clustal w and clustal x version 2.0. Bioinformatics 23(21), 2947–2948 (2007)
19. Manzini, G., Ferragina, P.: Engineering a lightweight suffix array construction algorithm. In: Möhring, R.H., Raman, R. (eds.) ESA 2002. LNCS, vol. 2461, pp. 698–710. Springer, Heidelberg (2002)
20. Moriyama, E.N., Gojobori, T.: Rates of synonymous substitution and base composition of nuclear genes in Drosophila. Genetics 130(4), 855–864 (1992)
21. Puglisi, S.J., Smyth, W.F., Turpin, A.H.: A taxonomy of suffix array construction algorithms. ACM Comput. Surv. 39, 4 (2007)
22. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2007) ISBN 3-900051-07-0

23. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylgenetic trees. Molecular Biology and Evolution 4, 406–425 (1987)
24. Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., Fraser, C.M.: Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. USA 102(39), 13950–13955 (2005)
25. Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450, 203–218 (2007)
26. Vinga, S., Almeida, J.: Alignment-free sequence comparison—a review. Bioinformatics 19, 513–523 (2003)
27. Wilbur, W.J., Lipman, D.J.: Rapid similarity searches of nucleic acid and protein data banks. Proceedings of the National Academy of Sciences, USA 80, 726–730 (1983)
28. Yang, K., Zhang, L.: Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction. Nucleic Acids Res. 36(5), e33 (2008)
29. Yang, Z.: Computational Molecular Evolution. Oxford University Press, Oxford (2006)