Split-Inducing Indels in Phylogenomic Analysis

Alexander Donath[a] *and Peter F. Stadler*[a−f]
[a]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center of Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig,
[b]Max-Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
[c]Fraunhofer Institut für Zelltherapie und Immunologie – IZI, Perlickstraße 1, D-04103 Leipzig, Germany
[d]Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria
[e]Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark
[f]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

Most approaches in molecular phylogenetics treat gaps im multiple sequence alignments as missing data or even completely exclude alignment columns with gaps. Here we show that gap patterns in large-scale, genome-wide alignments are themselves phylogenetically informative when properly filtered to reduce noise introduced by the alignment method. To this end, we introduce here split-inducing in/dels (splids) that define an approximate bipartition of the taxon set. We show both in simulated data and in case studies on real-life data that splids can be efficiently extracted from phylogenomic data sets. They provide a surprisingly clear phylogenetic signal and lead to quite accurate phylogenetic trees.

Introduction

Gaps in multiple sequence alignments are usually seen as a nuissance in molecular phylogenetics. In most studies, gaps are treated as missing data or alignment columns with gaps are even removed completely. Indeed, stochastic models of sequence evolution that deal explicitly with gaps have been investigated only recently (Rivas, 2005; Lèbre and Michel, 2010). Detailed evaluations show an overall improvement of phylogenetic reconstructions when gaps are modelled explicitly (Redelings and Suchard, 2007; Rivas and Eddy, 2008; Dwivedi and Gadagkar, 2009) but still report a negative effect of an increasing density of gap characters in multiple sequence alignments (Dwivedi and Gadagkar, 2009).

Between these few recent rigorous approaches to include gaps and the dismissal of gaps as missing data, indels have been incorporated in several ways into sequence-based phylogenetic analyses. The simplest one is to code gaps as 5th character state. Other authors have suggested the replacement of the gapped regions by a binary matrix that codes presence and/or absence of the respective indel (Simmons and Ochoterena, 2000). That is, the binary matrix is added to the "ungapped" sequence data and employed in tree inference. An extension of this simple indel coding (SIC) maximizes the amount of phylogenetic information in a parsimonious way by incorporating all indels (Müller, 2006).

Gaps in alignments are, of course, not features identifyable from the individual sequences. Instead, they appear as derived patterns inferred from sequence comparison only. Nevertheless, they convey a surprising amount of phylogenetic information. Shared multi-residue deletions, for instance, have been used to support hypothesis derived from molecular data in recent single gene analyses, see e.g. (Teeling et al., 2005). Multi-residue gaps in DNA as well as protein sequences have been reported as useful indicators of monophyletic groups (Lloyd and Calder, 1991). Single-residue gaps, on the other hand, occur more frequently than multi-residue gaps and show a higher amount of homoplasy, e.g. (Belinky et al., 2010). The same authors suggest that single-residue gaps should not be removed a priori from a data set based on a large taxon sampling, since they can still contain a phylogenetic signal.

The few studies of the phylogenetic information content of gap patterns were mostly conducted on limited sets of protein data. With the advent of high-throughput sequencing (nearly) complete genomes are becoming available at an increasing pace, from which large-scale genome-wide alignments can be constructed. Phylogenomics capitalizes on these developments and provides a wide diversity of phylogenetic information (Boussau and Daubin, 2010). We utilize these developments here to address the value of gap patterns from a phylogenomic perspective. The use of genome-wide datasets allow us to focus on the sub-class of indels only that define a "reasonably obvious" binary split among the taxa. As gaps are not part of sequence itself but the result of an alignment algorithm, however, we need to systematically investigate the impact of the alignment method on the phylogenetic information of the gap patterns.

Materials and Methods

Data sets

SIMULATED DATA. To test the performance of the method on multiple sequence alignments with indel formation according to a robust tree, we created a number of different artificial data sets, using `INDELible` V1.03 (Fletcher and Yang, 2009). The guide tree and background base frequencies were taken from the phastCons17way phastCons tree model file (Siepel et al., 2005) obtained from UCSC[1] and rescaled to have a maximum root-to-tip distance of 2. Indel rates and indel-size distributions are in most cases estimated based on pairwise alignments (e.g. human-mouse, primates, rodents (Lunter, 2007; Britten et al., 2003; Ogurtsov et al., 2004; Gu and Li, 1995)) but differ quite considerably. For example, estimates for the ratio of substitution rates to indel rates between mouse and human are ranging from 8 (Lunter, 2007) to 14 (Britten et al., 2003; Ogurtsov et al., 2004). It seems to be a good approximation to apply an indel rate in vertebrates at least

Key words: in/del, splits, molecular phylogeny, phylogenomics
E-mail: {alex,studla}@bioinf.uni-leipzig.de

Preprint.

[1] http://hgdownload.cse.ucsc.edu

as large as between human and mouse, however. Estimates suggest that the frequency of deletions is somewhat higher than the insertion frequency (Gu and Li, 1995; Zhang and Gerstein, 2003; Arndt and Hwa, 2004), with a ratio of deletion rate λ_d to insertion rate λ_i ranging from 1.3 to 4. We therefore created three different data sets using the F81 model (Felsenstein, 1981), two indel-size distributions and three different indel-rates, each consisting of 100 alignments with a length of 100.000 nt. The first two datasets use a simple geometric distribution with similar insertion and deletion rates ($\lambda_i = 0.03106$ and $\lambda_d = 0.04037$) but different probability values ($q_1 = 0.7$ and $q_2 = 0.55$, respectively). The third data set follows a Lavalette Distribution ($a = 1.5$, $M = 120$) with $\lambda_i = 0.02899$ and $\lambda_d = 0.03768$.

ENCODE DATA. In order to address the problem how the method behaves under real-life data and different alignment lengths we created two data sets from the ENCODE (Birney et al., 2007) project data, based on the December 2007 Multi-Species Sequence Analysis sequence freeze. The ingroup taxa include four apes (human, chimpanzee, Sumatran orangutan, and Northern white-cheeked gibbon), four Old World monkeys (Eastern Black-and-white colobus, vervet monkey, baboon, and rhesus macaque), four New World monkeys (dusky titi, owl monkey, marmoset, and squirrel monkey), two prosimians (small-eared galago and mouse lemur), tree shrew, four rodents (mouse, rat, guinea pig, and thirteen-lined ground squirrel), rabbit, cow, horse, two carnivores (dog and cat), three bats (greater horseshoe bat, little brown bat, and large flying fox), two insectivores (middle-African hedgehog and European shrew), armadillo, African elephant, tenrec, rock hyrax, platypus, and the South American short-tailed opossum. Chicken was used as outgroup to root the trees. The first data set contains only those alignments in which all 36 organisms were included. Alignments of only two ENCODE regions fulfilled this criteria: ENm001 (498 alignments) and ENm013 (67 alignments). To investigate how the method behaves under a considerable amount of missing data, as it is usually the case for genome wide alignments, a second data set was created, based on all ENCODE alignment regions with at least three species.

ENTEROBACTERIACEAE. The genomes of Bacteria are by far smaller than metazoan genomes. This makes creating whole-genome alignments computationally more feasible. On the other hand, their genomes frequently undergo recombination and they are known to import DNA from other organisms into their genomes. The imported DNA can then replace a homologous sequence. This often constitutes a problem for sequence based phylogenetic analysis. We have selected a subset of the Gram-negative Enterobacteriaceae family (Proteobacteria; Gammaproteobacteria; Enterobacteriales). Enterobacteriaceae normally inhabit the intestines of animals but are also found in plants and water. Many members are pathogens with a considerable clinically importance. We selected all Enterobacteria listed at [2] for which complete genomes were available at NCBI. Due to algorithmically limitations (see Section) we reduced the final data set to 54 species and removed all *E. coli strains* except K12 MG1655. *Buchnera aphidicola* was used as outgroup to root the trees. A complete taxon list is given in Suppl. Table 1.

[2] http://www2.unil.ch/comparativegenometrics/phylo.html

Alignments

GENOME-WIDE ALIGNMENTS. We used `TBA/MULTIZ` (Blanchette et al., 2004) for both the ENCODE and the enterobacteria data sets. This toolkit has been widely used for whole-genome alignments in large-scale comparative genomics studies (Birney et al., 2007; Bauer and Bailey, 2008). `TBA/MULTIZ` needs a guide tree that describes the relationship of the species to be aligned. In case of the ENCODE data set this tree is largely based on taxonomic information. The guide tree for the Enterobacteria data set was created on a set of orthologous proteins for an earlier analysis (kindly provided by Sven Findeiß, unpublished results): In brief, a set of orthologous proteins was derived using `ProteinOrtho` (Lechner et al., 2011) and pruned to contain only proteins present in all input taxa; these were aligned with `ClustalW` (Larkin et al., 2007); a neighbor-joining tree was calculated, using SplitsTree (Huson and Bryant, 2006). Since `TBA/MULTIZ` was not able to align the complete set 75 enterobacterial genomes (due the limitations in the internal data handling, M. Hou, pers. communication), we sub-selected 54 taxa for inclusion in our alignment.

A genome-wide alignment is the result of an extensive similarity search between at least two species. Due to evolutionary changes in genome organization, such as inversions and duplications, two genomes are virtually never completely co-linear, resulting in a decomposition of alignments into syntenic blocks. Practical procedure such as `TBA/MULTIZ` also use other features, such as large insertions, missing data in individual species, or low complexity regions, as additional breakpoints, so that relative small aligment blocks are produced. Not all of these blocks contain sequence from all taxa, both due to missing data in the sequence assemblies and because highly diverged regions of some taxa can not be reliably recognized as homologs.

RE-ALIGNMENT WITHOUT GUIDE TREES. The use of a guide tree for the genome alignments could conceivably create a bias in indel positioning. We therefore checked whether such a bias really exists and how other commonly used alignment programs perform. To this end we considered individual alignment blocks produced by `TBA/MULTIZ` and removed the gaps again. The genome-wide alignments thus are used only as a convenient means of extracting homologous regions.

A similar procedure was applied to the 'true' alignments of the simulated dataset which were at first separated in blocks with an average size of 140 nt and in the following treated as described below.

The gap-free sequences of each block were re-aligned with a variety of commonly used programs and algorithms: `ClustalW` (version 2.0.12) (Larkin et al., 2007), `Mafft` (v6.833b) (Katoh et al., 2005), `Muscle` (v3.7) (Edgar, 2004), `T-Coffee` (Version 8.97) (Notredame et al., 2000), `Prank` v.100802 (Löytynoja and Goldman,

2010), and Dialign-TX (version 1.0.2) (Subramanian et al., 2008). Dialign-TX differs from all other methods as it creates alignments from local pairwise sequence similarities without the use of explicit gap penalties. Approximately 2% of the ENCODE regions contain coding exons while the majority covers 'non-coding' sequences, such as introns, UTRs, and intergenic regions. It has been pointed out that, while performing fairly good on these sequences, TBA/MULTIZ's results on regions containing non-coding RNAs is not optimal (Wang et al., 2007). We therefore additionally selected ProbConsRNA (version 1.1) (Do et al., 2005), which is an experimental version of PROBCONS with parameters estimated from BRAliBASE II via unsupervised training (Gardner et al., 2005). All tools were used with default values except for Mafft which also offers two substantially different modes: L-INS-i (local optimal alignment) and G-INS-i (global optimal alignment).

Following realignment, gaps introduced at the 5' and 3' ends of the sequence blocks were interpreted as artifacts and hence coded as missing data, see also (Simmons and Ochoterena, 2000). As individual alignment blocks typically contain sequence data for only a subset of the input taxa, such missing taxa were also explicitly coded as missing data. The alignment blocks with two or more taxa and containing at least one gap character were then concatenated using a perl script (available with the Supplementary Material). Note that by construction the delimiting columns of each alignment block do not contain gap characters; concatenation therefore does not affect the gap patterns.

Splids

Gaps often appear in rather disordered clusters in automatically generated alignments. The encode of characters from gap patterns is not entirely trivial as soon as indels rather than individual gap characters are to be assessed. In the first step, the problem is reduced to *indel loci* consisting of connected indels. More precisely, an indel is a contiguous stretch of (overlapping) gap characters. Two indels overlap if there is an alignment column that is common to both of them, see Fig. 1.

In the second step the individual indel loci are examined in detail. First we identify all distinct intervals of gaps (circled numbers in Fig. 1). We call an indel a *splid* (*spl*it-inducing *in*d*el*) if it defines an approximate bipartition of the taxon set according to the following rules:

1. Only indels found in at least two sequences, having the same 5' and 3' end, and with a given minimum size are considered. By default, all indels of length at least two are considered. Thus indels (1), (2), (3), (5), (7), (8), (12), and (13) in Figure 1 are removed.

2. A splid does not overlap another indel that satiesfies the first condition. Thus indels (9) and (10) are excluded.

Splids are coded as binary characters marking their absence/presence pattern in the respective taxon. Missing sequence data in the alignment column of a splid was coded as "missing data". We optionally filter out splids that overlap an indel of length 1 occurring in at least two taxa (such as indel (13)). In this "strict mode" indel (11) is removed, while it is retained in "fuzzy mode". These alternative treatments of single-position gaps is motivated by the observation that they occur more randomly than multi-residue gaps, while still containing some phylogenetic information (Belinky et al., 2010).

The algorithm for the conversion of alignments to a binary character matrix is implemented in the C++ program gappy. The tool reads multiple sequence alignments in FASTA format. The user can select a minimum and maximum indel length for determining splids. By default, the output is a FASTA file, containing the binary coded splid absence/presence information, and some summary statistics. Output is also available in PHYLIP and NEXUS format.

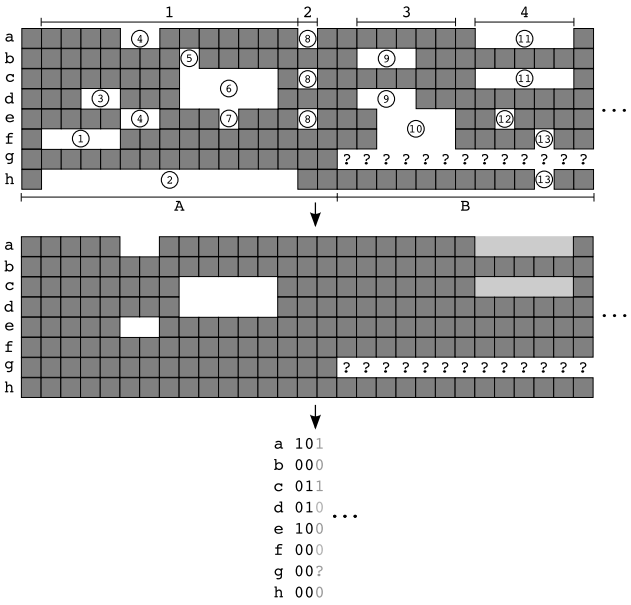

FIG. 1.—Non-trivial example of the determination of splids with size of at least 2 from a concatenated alignment set (A and B). Alignment A contains sequence data for all taxa, whereas B misses taxon *g*. At first, all indel loci are determined (1, 2, 3, and 4). Second, the loci are searched for indels constituting splids. From locus 1 indels (4) and (6) fulfill this criterion [(1) and (3), respectively, do not fulfill the splid criterion]. Indel (8) is too small. Locus 3 contains a set of conflicting splids [(9) and (10)]. Thus they are not included. If indel (11) is included in the final set of splids depends on the applied algorithm. In *strict* mode it is not included, due to the single-residue indel (13). In *fuzzy* mode, it is also included and taxon *g* is marked as missing data ("?") in the binary absence/presence coding.

Phylogenetic reconstruction and Analysis

TREE RECONSTRUCTION. Phylogenetic trees were calculated with the hybrid version of RAxML v7.2.8 (Stamatakis, 2006), using rapid bootstrapping with 100 random additions under the Gamma-model for binary characters (Stamatakis et al., 2008; Pattengale et al., 2010). Bootstrap support values were drawn on the best-scoring tree.

TREE COMPARISON. Many different distance measures are available to compare phylogenetic trees. The most sensitive one is the unweighted Robinson-Foulds distance (Robinson and Foulds, 1981), defined as the number d_{RF} of splits contained in exactly one of the two trees. The scaled version $d'_{RF} = d_{RF}/(n-3)$ takes values between 0 and 1, where n is the number of taxa. Its major drawback is that it does not emphasize local similarity, so that trees differing by the placement of a single taxon may have large RF distances (Penny et al., 1982). As an alternative we therefore employ the quartet distance (Estabrook et al., 1985), defined as the number of quartets that are subtrees of one but not the other input tree. The normalized quartet distance, $d'_Q = d_Q/\binom{n}{4}$, serves as a convenient distance measure between large phylogenetic trees. We use here `Phylonet` (version 2.3) (Than et al., 2008) and `QDist` (version 2.0.2) (Mailund and Pedersen, 2004) to compare the obtained trees with the underlying guide trees.

Results

Simulated Alignments

In order to test the quality of the phylogenetic signal provided by splids we first used simulated sequence data generated by `INDELible` along a known reference tree. Alignments were computed using nine different methods, see Supplementary material. Overall, 3000 trees were calculated from these alignments and the simulated `INDELible` reference alignments. On these artificial data set we observe nearly correct trees derived from splids (see Suppl. Figure 1). On these benign data, the choice of the alignment methods has little effect on the quality of the estimated phylogenies. No RF distances between reconstructed phylogeny and reference tree larger than 2 was observed. Indeed 82.57% of the trees were identical to the reference, and another 16.67% showed an RF distance of 1. Quartet distances draw a similar picture but allow a better differentiation of the respective methods. The majority of all trees (97.5%) from all methods have a $d'_Q \leqslant$ 0.122%, however. `ClustalW` performed worst, although the distance of the tree most dissimilar to the guide tree is only 1.68%. The best performance was observed for `Mafft-linsi`. Details can be found in the supplement.

In contrast to real data, however, these simulated test cases are rather homogeneous. Although rate heterogeneity among sites has long been accepted to be a biological more realistic assumption (see e.g. (Yang, 1996)). We therefore investigated two real-life examples in detail.

ENCODE Genomes

SMALL DATA SET. Depending on the alignment method, the concatenated re-alignments of the ENCODE data differed in length and hence in the total number of gaps that they contain. For the small ENCODE data set, `ClustalW` produced the shortest and `Dialign-TX` the longest alignment. It is no surprise that the number of splids grows with the number of alignment sites, Table 1. For the three `Mafft` algorithms, however, the number of splids decreases with alignment length. In particular, `Mafft-default` and `Mafft-linsi` seem to introduce more single-residue gaps or conflicting splits than `Mafft-ginsi`. Figure 2 shows the distribution of splid lengths for the different alignment algorithms. The alignment methods fall into two broad groups. While `Dialign-TX`, `T-Coffee`, `Prank` and `ProbConsRNA` yield a similar distribution to `TBA/MULTIZ`, we obtain only half as many splids $\leqslant$ 5nt from `Muscle`, `ClustalW`, and all three `Mafft` algorithms. There is, however, no systematic dependence on design features of the alignment methods such as global versus local alignments or progressive versus consistency based methods. While the splid-based phylogenies are nearly perfect on simulated data, we observe larger deviations that depends at least in part on the alignment methods when applying our approach to real-life data. On the other hand, in real data sets we do not have an absolute ground truth to compare to. Thus we discuss in following both the quality of the reconstructed phylogenies and the position of interesting taxa in some detail.

The monophyly of Afrotheria and the positioning of tenrec basal to elephant and rock hyrax is always recovered (Stanhope et al., 1998; Arnason et al., 2008), except by `Mafft-default`, which places tenrec basal to armadillo. The position of the placental root is still, at least

Table 1 Comparison of the number of sites of the final alignments and derived splids with length $\geqslant$ 2nt for the ENCODE data set containing only alignments with sequence information for all taxa.

Tool	No. of sites	No. of splids
`ClustalW`	79,006	793
`Dialign-TX`	96,990	2,163
`Mafft`	84,105	1,021
`Mafft-linsi`	83,578	1,245
`Mafft-ginsi`	83,123	1,279
`Muscle`	84,577	1,378
`ProbConsRNA`	86,277	1,927
`Prank`	96,622	2,047
`T-Coffee`	84,835	1,831
`TBA/MULTIZ`	90,726	2,032

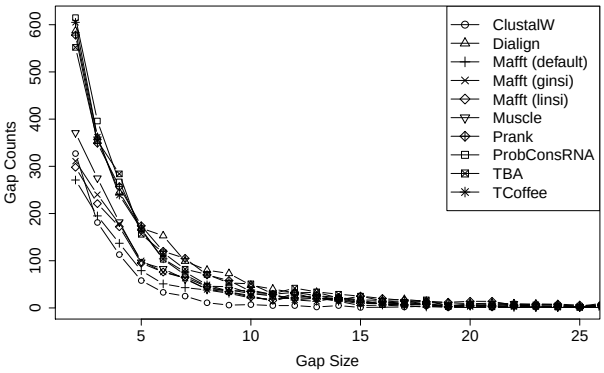

FIG. 2.—Number of splids $\geqslant$2nt for the different alignment methods for the ENCODE data set containing sequence information for all taxa.

to some extent, a matter of debate (Murphy et al., 2004; Springer et al., 2004; Murphy et al., 2007; Nikolaev et al., 2007). However, `Mafft-default` as well as the majority of all alignment programs correctly positions Afrotheria outside of Boreoeutheria (Prasad et al., 2008). Only splid data obtained from the `Muscle`, `ProbConsRNA`, and `T-Coffee` alignments places Afrotheria as sister group to Laurasiatheria (`ProbConsRNA` and `T-Coffee`) or inside Euarchontoglires (`Muscle`). Not even the original `TBA/MULTIZ` alignments contain enough supporting splids to position them outside of Boreoeutheria, however.

Three hypotheses concerning the positioning of Xenarthra are discussed in the literature: (1) basal-Afrotheria ((Boreoeutheria, Xenarthra); Exafroplacentalia), e.g. (Murphy et al., 2004; Nikolaev et al., 2007), (2) basal-Xenarthra ((Boreoeutheria, Afrotheria); Epitheria), e.g. (Kriegs et al., 2006), and (3) basal-Boreoeutheria ((Afrotheria, Xenarthra); Atlantogenata), e.g. (Wildman et al., 2007). Splid data is clearly in favor of the basal-Xenarthra hypothesis. `Prank` positions armadillo basal to Afrotheria, whereas `ProbConsRNA` and `T-Coffee` place it basal to Laurasiatheria and therefore inside Boreoeutheria. No tree supports the ENCODE guide tree which follows the basal-Afrotheria hypothesis.

Laurasiatheria are in most cases found to be monophyletic, with the exception of `Prank`, which places Insectivora basal to the remaining Boreoeutheria, and `ProbConsRNA` and `T-Coffee`, where Afrotheria are incorrectly positioned. Monophyly is also preserved for its major orders Insectivora (Eulipotyphla), Chiroptera (except `ProbConsRNA` and `T-Coffee`), and Carnivora. There is no clear result from splid data about the correct relationship within Laurasiatheria, which resembles the conclusions obtained elsewhere (Springer et al., 2004; Arnason et al., 2008; Prasad et al., 2008), although it seems to be consensus that Insectivora (Eulipotyphla) form the basal clade (Springer et al., 2004). All alignment methods support this view, except `Mafft-ginsi` and `Prank`. The evolutionary history of bats has long been a subject of discussion, with conflicting hypothesis depending on whether morphological or molecular data was employed. Earlier studies either traditionally suggested the monophyly of the suborders Megachiroptera (megabats) and Microchiroptera (microbats), e.g. (Simmons and Geisler, 1998), while other studies placed megabats together with the rhinolophoid microbats (Yinpterochiroptera), with the remaining microbats forming the suborder Yangochiroptera, e.g. (Hutcheon et al., 1998; Teeling et al., 2002). Splid data derived from most of the alignment methods support the novel view of chiropteran phylogeny and place *Rhinolophus ferrumequinum* together with *Pteropus vampyrus*, while *Myotis lucifugus* (Vespertilionidae) is found basal to them. Only `ProbConsRNA` follows the traditional view of a monophyly of megabats and microbats and is therefore similar to the results of the `TBA/MULTIZ` alignments which also clearly resembles the ENCODE guide tree.

The monophyly of Euarchontoglires (Euarchonta and Glires) could not be recovered from splid data obtained from `Muscle`, `T-Coffee`, and `ProbConsRNA`, because of the wrongly positioned Afrotheria, which also leads to a split of rodents (and therefore Glires) for `Muscle` and `T-Coffee` data. However, all other alignment methods clearly support the monophletic superorder Euarchontoglires.

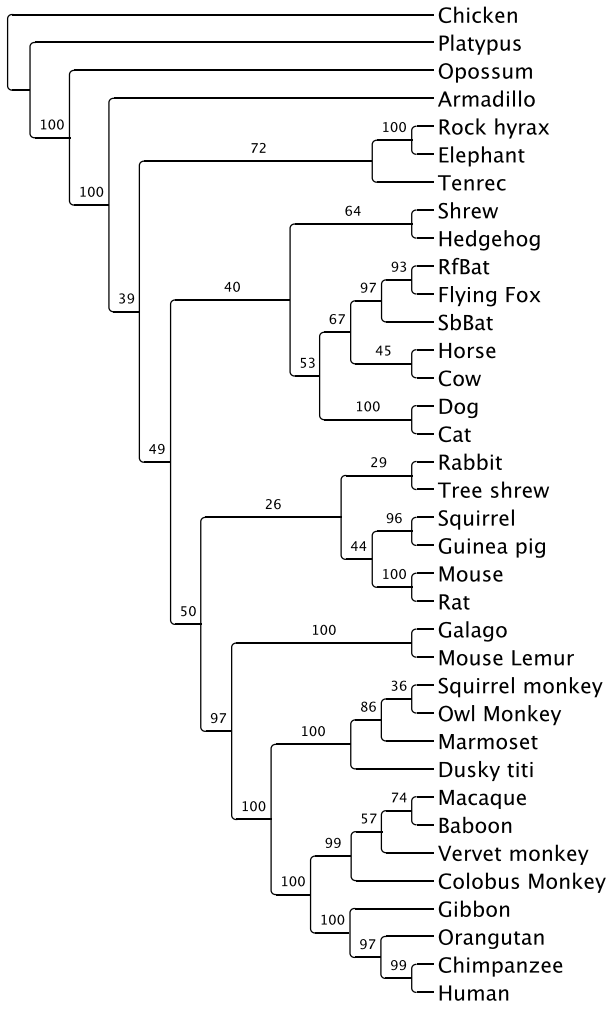

FIG. 3.—Cladogram with bootstrap values obtained from 100 rapid bootstrap inferences by `RAxML` (BINGAMMA) on the small ENCODE data set. The alignment was created with `Mafft-linsi` and splids $\geqslant$ 2nt were coded.

Among all groups analyzed, Glires are the most problematic one. While in all cases close taxa (e.g. Muridea) are found to be in a common subtree monophyly was only recovered from `ProbConsRNA` and `Prank` data. However, our results also reflect the unclear position of tree shrew. While some authors place them basal to Glires, others consider them to be basal to Primata. Even though our data set does not allow a clear conclusion, they are often found within or close to Glires. Only `Dialign-TX` positions them basal to Primata, although with rabbit as sister taxon. Splid data derived from `ProbConsRNA` and `T-Coffee` places them basal to (Afrotheria, Laurasiatheria) and (Afrotheria, Boreotheria), respectively.

Table 2 Overview of how the different monophyletic groups are recovered by the applied alignment tools. Trees were calculated on the small ENCODE data set, using RAxML (splids $\geqslant$ 2nt, BINGAMMA model). For single taxa and groups with more than two members the position in the tree is provided. For each tree the symmetric difference (Robinson-Foulds distance), the quartets distance, and the normalized quartets distance to the ENCODE guide tree is given. Carn. = Carnivora. "x" = correctly recovered, "-" = false positioning. See text for details.

	ClustalW	Dialign-TX	Mafft	Mafft-ginsi	Mafft-linsi	Muscle	Prank	ProbConsRNA	T-Coffee	TBA/MULTIZ
Afrotheria	x	x	-	x	x	x	x	x	x	x
	sister group Boreoeutheria	sister group to Boreoeutheria	sister group to Boreoeutheria	sister group to Boreoeutheria	sister group to Boreoeutheria	sister group to Euarchontoglires	sister group to Boreoeutheria	sister group to Laurasiatheria	sister group to Laurasiatheria	sister group to Euarchontoglires
((elephant, rock hyrax), tenrec)	x	x	-	x	x	x	x	x	x	x
Xenarthra	basal to Epitheria	basal to Epitheria	(basal to Epitheria)	basal to Epitheria	basal to Epitheria	basal to Epitheria	basal to Afrotheria	basal to Laurasiatheria	basal to Laurasiatheria	basal to Epitheria
Boreoeutheria	x	x	x	x	x	-	x	-	-	-
Laurasiatheria	x	x	x	x	x	x	-	x	x	x
Insectivora	x	x	x	x	x	x	x	x	x	x
Chiroptera	x	x	x	x	x	x	x	x	x	x
((rfbat, flying fox), sbbat)	x	x	x	x	x	x	x	-	x	-
Carnivora	x	x	x	x	x	x	x	x	x	x
				sister group to (Hystricognathi, Sciurognathi)						
horse	(bats, horse)	(Carn., horse)	(bats, horse)	((bats, cow), horse)	(cow, horse)	((bats, cow), horse)	((bats, cow), horse)	(((bats, cow), Carn.), horse)	(cow, horse)	(Carn., horse)
cow	((bats, horse), cow)	(((Carn., horse), bats), cow)	(((bat, horse), Carn.), cow)	(bats, cow)	(cow, horse)	(bats, cow)	(bats, cow)	(bats, cow)	(cow, horse)	(((Carn., horse), bats), cow)
Euarchontoglires	x	x	x	x	x	-	x	-	-	x
Glires	x	-	-	-	-	-	x	x	-	x
Rodentia	-	-	-	x	x	-	x	-	-	x
Muroidea	x	x	x	x	x	x	x	x	x	x
Rabbit	sister taxon to Muroidea	sister taxon to tree shrew; basal to Primata	basal to Euarchontoglires	sister taxon to tree shrew; basal to Rodentia	sister taxon to tree shrew; basal to Rodentia	basal to Euarchonta	basal to Rodentia	in Rodentia; basal to (Hystricognathi, Sciurognathi)	basal to Primata	basal to Rodentia
Primata	x	x	x	x	x	x	x	-	x	x
Strepsirrhini	x	x	x	x	x	x	x	x	x	x
Platyrrhini	x	x	x	x	x	x	x	x	x	x
(((squirrel m, marmoset), owl m), dusky titi)	x	-	-	-	-	-	-	-	x	-
Catarrhini	x	x	x	x	x	x	x	x	x	x
Cercopithecidae	x	x	x	x	x	x	x	x	x	x
(((baboon, macaque), vervet), colobus)	-	-	-	x	x	x	x	-	x	-
Hominoidea	x	x	x	x	x	x	x	x	x	x
(((chimp, human), orangutan), gibbon)	x	x	x	x	x	x	-	-	x	-
tree shrew	in Glires; basal to (Hystricognathi, Sciurognathi)	sister taxon to rabbit; basal to Primata	in Rodentia; basal to (Hystricognathi, Sciurognathi);	sister taxon to rabbit; basal to Rodentia	sister taxon to rabbit; basal to Rodentia	basal to (Hystricognathi, Sciurognathi)	basal to Euarchontoglires	basal to Afrotheria and Laurasiatheria	basal to Afrotheria and Boreoeutheria	basal to Glires
Robinson-Foulds distance	9	9	11	9	8	11	9	13	10	7
Quartets distance (at most 58,905)	2,314	2,980	3,052	2,664	2,043	7,024	4,028	9,714	9,458	3,932
Normalized Quartets distance	0.0393	0.0506	0.0518	0.0452	0.0347	0.1192	0.0684	0.1649	0.1606	0.0668

Table 3 Comparison of results for the large ENCODE data set. Splids $\geqslant$ 2nt were coded and trees were calculated with RAxML using the Gamma model for binary data.

	Mafft-linsi	ProbConsRNA	TBA/MULTIZ
# sites	35,505,276	36,464,967	37,689,662
# splids	529,153	946,184	919,908
d_{RF}	7	8	4
d_Q	4,936	4,746	862
d'_Q	0.0838	0.0806	0.0146

Table 4 Characteristics of the Enterobacteria data set. Splids $\geqslant$ 2nt were coded and trees were calculated using the Gamma model for binary data implemented in RAxML.

	Mafft-linsi	ProbConsRNA	TBA/MULTIZ
# sites	4,131,192	4,449,170	4,547,794
# splids	19,448	47,757	51,077

Almost all methods support the monophyly of Primates, as well as a monophyly of the respective sub- and parvorders. Only ProbConsRNA positions Strepsirrhini as sister group to Glires and (Laurasiatheria, Afrotheria).

As a quantitatively evaluation of the mammalian tree we consider their RF and quartet distances to the ENCODE reference tree, which – although not undisputed – well reflects the state of the art in mammalian phylogeny. The ProbConsRNA tree is most different from the reference with respect to both the RF and the quartet distance. Trees computed with T-Coffee and Muscle are only slightly better. However, when comparing the values of the two metrics for the other methods it becomes apparent that their results are quite different and show no clear correlation. While the RF distances of the Mafft-default and Muscle are similar, the quartet distance Mafft-default is smaller by about a factor of two. Overall, the Mafft-linsi algorithm clearly performed best, having the second lowest symmetric RF distance and a quartet distance of only 3.5% if compared to the ENCODE reference tree (Fig. 3). Surprisingly, trees based on splids from ClustalW, Dialign-TX, and all three Mafft algorithms outperformed the guide tree based TBA/MULTIZ alignments. The Probabilistic Alignment Kit Prank (Löytynoja and Goldman, 2010) has been advertised to produce gap placements that are phylogenetic more consistent compared to other alignment algorithms. In line with another recent study (Dessimoz and Gil, 2010), we were unable to confirm this claim for our data sets, however. We note, finally, that misplaced taxa in all trees generally had low bootstrap support.

LARGE DATA SET. Because of the computational resources required from the phylogenetic reconstruction we selected two methods for comparison on the large ENCODE data set: Mafft-linsi was chosen because it performed best on the small set. In order to check whether the increase in the size of the dataset improves the performance we also included ProbConsRNA, the method with the poorest performance on the small data set. In addition, we included the splid set derived from the original TBA/MULTIZ alignment, Table 3. An overall improvement was observed for both ProbConsRNA and Mafft-linsi. Two problematic nodes were observed in set, however. In the Mafft-linsi tree Afrotheria are now found as sister clade to Euarchontoglires even though they were placed in a more plaubsible position in the small ENCODE set. Also, Mafft-linsi places hedgehog and shrew outside of Afrotheria and Boreoeutheria and does not even recognize them as sister taxa. On the other hand, ProbConsRNA still positions Afrotheria as sister clade to Laurasiatheria and favors now an implausible position of tree shrew basal to Afrotheria and Boreoeutheria. However, the split of Euarchontoglires and (Afrotheria, Boreoeutheria) shows very low bootstrap support of 45. Of all species included in the large ENCODE data set, tree shrew has by far the smallest sequence coverage (approx. 10% of human), which likely contributes to its unstable position.

Unexpectedly, ProbConsRNA and Mafft-linsi yield similar results in terms of tree distances. For Mafft-linsi, the Robinson-Foulds distance dropped only slightly (7 vs. 8) while ProbConsRNA showed a much better result (8 vs. 13). On the other hand, a comparison based on local similarity showed a different picture. The quartet distance of the Mafft-linsi tree increased, with its normalized value more then twice as large as for the small ENCODE data set (0.0838 vs. 0.0347). The ProbConsRNA alignment-based tree, however, showed a normalized quartet distance of less than a half when compared to the result on the small ENCODE data set (0.0806 vs. 0.1649). Local as well as global similarity of the TBA/MULTIZ alignment-based tree to the ENCODE guide tree increased when compared to the performance on the small data set.

Enterobacteria

The length of the concatenated alignments of the Enterobacteria does not depend strongly on the alignment method, see Table 4. Nevertheless, the number of obtained splids is very different. More than twice as many splids could be found in ProbConsRNA and TBA/MULTIZ alignments, respectively, than in the Mafft-linsi alignments. Unfortunately, the rapid bootstrapping algorithm in RAxML contained a bug and was not able to calculate bootstrap support for the ProbConsRNA and TBA/MULTIZ splids. For the splids obtained from the Mafft-linsi alignments clear and distinct clusters could be retrieved for all taxa which largely resemble the phylogenetic relationships previously reported (Kuhnert et al., 2009), see Suppl. Figures 2 and 3. Members of the *Salmonella* genus form a clear cluster with subspecies *S. arizonae* (subsp. IIIa) well separated from the *S. enterica* (subsp. I) lineage. Within the *S. enterica enterica* subspecies similar serovars cluster together. *Yersinia*

pestis is seen as a clone of *Yersinia pseudotuberculosis* as they cannot be genetically distinguished (Achtman et al., 1999, 2004). However, high bootstrap values support the subtrees in which the two different strains are found to be monophyletic. Also *Y. enterocolitica* is positioned basal and therefore clearly distinguishable from the two species. In contrast to *Yersinia*, *Escherichia* and *Shigella* are found in one common subtree and cannot be separated from each other. This resembles the results found in several earlier studies (Karaolis et al., 1994; Stevenson et al., 1994; Fukushima et al., 2002) and further supports the view that *Shigella* strains are in fact clones of *E. coli*. *Enterobacter sp.* is clearly separated from *Cronobacter sakazakii*. This is also reported by Baldwin et al. (2009) and reflects the recent reclassification of the *Enterobacter sakazakii* species within the novel genus *Cronobacter* (Iversen et al., 2008).

Discussion

Indels are not features of individual sequences. Instead they are inferred by comparative analysis and, in practice, appear as gaps in multiple sequence alignments. In some alignment methods they are explicitly modelled and contribute to the score e.g. by means of affine gap costs. In other approaches they are given only implicitly. It is not unexpected, therefore, that the number and position of gaps depends quite strongly on the alignment algorithm. Nevertheless, gap positions can be phylogenetically informative.

We have focused here on a subclass of indels, namely those which can be found in more than one taxon and therefore define a split in the sequence set. Our definition and inference of such split-inducing indels (splids) is based on two basic principles that are largely accepted in the literature. First, indels at the same position, i.e. sharing the same end points in two sequences, are likely homologous. Second, independent single-residue insertions and deletions tend to occur more frequently than multi-residue indels. Hence they are expected to contribute a more noisy signal and hence are disregarded in our analysis. In addition, we employ technical conditions that help to reduce the noise introduced by single mis-aligned sequences, which are quite frequently observed in genome-wide alignments.

We have tested the information content of splids on three simulated and three real-life data sets and analyzed the capability of splids introduced by nine different alignment programs for phylogenetic inference by Maximum Likelihood (ML). For artificial data sets, which are generated from a known underlying phylogeny, we find that splid-based ML reconstruction leads to nearly perfect trees. On the real-life data sets we observe larger discrepancies between different alignment methods.

The splid-based phylogenies clearly recovered most of the undisputed monophyletic groups both in the mammalian and the bacterial data sets. Although there are clear differences in the alignment methods, the approach is surprisingly robust across a wide variety of alignment techniques. While expected a large influence of the guide tree on the reconstructed phylogeny. In particular for the indel-based approach, we observed that this effect is small when only splids are considered. Overall, alignment methods that put more emphasis on modelling indels, in particular those that employ an affine gap cost model, perform superior to alignment algorithm that consider indels only implicitly. For very large data sets, furthermore, we observe a decreasing influence of the alignment algorithm.

As with all other phylogenetic approaches, taxon sampling has a major influence on branch positions in very divergent taxonomic orders. This can been seen for example in the Laurasiatheria, where a small set of closer related taxa (e.g. bats or Carnivora) are embedded in a larger set of species more distantly related to the respective smaller subgroup. While the Chiroptera are always monophyletic and the correct phylogenetic position within their subtree can be recovered, their position within Laurasiatheria can not be unambiguously determined.

Increasing sequence length, and therefore splid information, does not lead to improved trees in all examples. This effect is likely related to the observation that alignments computed for large datasets have relatively large error rates, and maximum likelihood phylogenies computed on these alignments also have high error rates (Liu et al., 2010). In the case of low but roughly equal sequence amount for all taxa, the choice of the alignment algorithm, seems to have a higher effect within lower taxonomic orders, while groups resembling higher taxonomic orders are relatively stable and mostly correct positioned.

Supplementary material

Supplemental data, in particular the source code for `gappy` can be found at
http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/11-001

Acknowledgements

The authors want to thank Yvo Wezel and David Langenberger for fruitful discussions on the subject of character loss, Petra Pregel and Jens Steuck for making work so much easier, and the Center for Information Services and High Performance Computing (ZIH) of the TU Dresden for allowing us to use their resources[3]. This work was funded by the Deutsche Forschungsgemeinschaft under the auspices of SPP-1174 *Deep Metazoan Phylogeny* (project STA 850/2).

Literature Cited

Achtman, M., G. Morelli, P. Zhu, et al. 2004. Microevolution and history of the plague bacillus, Yersinia pestis. Proc. Natl. Acad. Sci. U.S.A. **101**:17837–17842.

Achtman, M., K. Zurth, G. Morelli, G. Torrea, A. Guiyoule, and E. Carniel. 1999. Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. Proc. Natl. Acad. Sci. U.S.A. **96**:14043–14048.

Arnason, U., J. A. Adegoke, A. Gullberg, E. H. Harley, A. Janke, and M. Kullberg. 2008. Mitogenomic relation-

[3] http://tu-dresden.de/zih/

ships of placental mammals and molecular estimates of their divergences. Gene **421**:37–51.

Arndt, P. F., and T. Hwa. 2004. Regional and time-resolved mutation patterns of the human genome. Bioinformatics **20**:1482–1485.

Baldwin, A., M. Loughlin, J. Caubilla-Barron, E. Kucerova, G. Manning, C. Dowson, and S. Forsythe. 2009. Multilocus sequence typing of Cronobacter sakazakii and Cronobacter malonaticus reveals stable clonal structures with clinical significance which do not correlate with biotypes. BMC Microbiol. **9**:223.

Bauer, D. C., and T. L. Bailey. 2008. Studying the functional conservation of cis-regulatory modules and their transcriptional output. BMC Bioinformatics **9**:220.

Belinky, F., O. Cohen, and D. Huchon. 2010. Large-scale parsimony analysis of metazoan indels in protein-coding genes. Mol. Biol. Evol. **27**:441–451.

Birney, E., J. A. Stamatoyannopoulos, A. Dutta, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature **447**:799–816.

Blanchette, M., W. J. Kent, C. Riemer, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. **14**:708–715.

Boussau, B., and V. Daubin. 2010. Genomes as documents of evolutionary history. Trends Ecol. Evol. **25**:224–232.

Britten, R. J., L. Rowen, J. Williams, and R. A. Cameron. 2003. Majority of divergence between closely related DNA samples is due to indels. Proc. Natl. Acad. Sci. U.S.A. **100**:4661–4665.

Dessimoz, C., and M. Gil. 2010. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. **11**:R37.

Do, C. B., M. S. Mahabhashyam, M. Brudno, and S. Batzoglou. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res. **15**:330–340.

Dwivedi, B., and S. R. Gadagkar. 2009. Phylogenetic inference under varying proportions of indel-induced alignment gaps. BMC Evol. Biol. **9**:211.

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

Estabrook, G. F., F. R. McMorris, and C. A. Meacham. 1985. Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units. Systematic Biology **34**:193–200.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

Fletcher, W., and Z. Yang. 2009. INDELible: a flexible simulator of biological sequence evolution. Mol. Biol. Evol. **26**:1879–1888.

Fukushima, M., K. Kakinuma, and R. Kawaguchi. 2002. Phylogenetic analysis of Salmonella, Shigella, and Escherichia coli strains on the basis of the gyrB gene sequence. J. Clin. Microbiol. **40**:2779–2785.

Gardner, P. P., A. Wilm, and S. Washietl. 2005. A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res. **33**:2433–2439.

Gu, X., and W. H. Li. 1995. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. J. Mol. Evol. **40**:464–473.

Huson, D. H., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. **23**:254–267.

Hutcheon, J. M., J. A. Kirsch, and J. D. Pettigrew. 1998. Base-compositional biases and the bat problem. III. The questions of microchiropteran monophyly. Philos. Trans. R. Soc. Lond., B, Biol. Sci. **353**:607–617.

Iversen, C., N. Mullane, B. McCardell, B. D. Tall, A. Lehner, S. Fanning, R. Stephan, and H. Joosten. 2008. Cronobacter gen. nov., a new genus to accommodate the biogroups of Enterobacter sakazakii, and proposal of Cronobacter sakazakii gen. nov., comb. nov., Cronobacter malonaticus sp. nov., Cronobacter turicensis sp. nov., Cronobacter muytjensii sp. nov., Cronobacter dublinensis sp. nov., Cronobacter genomospecies 1, and of three subspecies, Cronobacter dublinensis subsp. dublinensis subsp. nov., Cronobacter dublinensis subsp. lausannensis subsp. nov. and Cronobacter dublinensis subsp. lactaridi subsp. nov. Int. J. Syst. Evol. Microbiol. **58**:1442–1447.

Karaolis, D. K., R. Lan, and P. R. Reeves. 1994. Sequence variation in Shigella sonnei (Sonnei), a pathogenic clone of Escherichia coli, over four continents and 41 years. J. Clin. Microbiol. **32**:796–802.

Katoh, K., K. Kuma, H. Toh, and T. Miyata. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. **33**:511–518.

Kriegs, J. O., G. Churakov, M. Kiefmann, U. Jordan, J. Brosius, and J. Schmitz. 2006. Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol. **4**:e91.

Kuhnert, P., B. M. Korczak, R. Stephan, H. Joosten, and C. Iversen. 2009. Phylogeny and prediction of genetic similarity of Cronobacter and related taxa by multilocus sequence analysis (MLSA). Int. J. Food Microbiol. **136**:152–158.

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

Lèbre, S., and C. J. Michel. 2010. A stochastic evolution model for residue Insertion??Deletion Independent from Substitution. Comp. Biol. Chem. **34**:259–267.

Lechner, M., S. Findeiß, L. Steiner, M. Marz, P. F. Stadler, and S. J. Prohaska. 2011. `Proteinortho:` Detection of (Co-)Orthologs in Large-Scale Analysis. BMC Bioinformatics Submitted.

Liu, K., C. R. Linder, and T. Warnow. 2010. Multiple sequence alignment: a major challenge to large-scale phylogenetics. PLoS Curr **2**:RRN1198.

Lloyd, D. G., and V. L. Calder. 1991. Multi-residue gaps, a class of molecular characters with exceptional reliability for phylogenetic analyses. J. Evol. Biol. **4**:9–21.

Löytynoja, A., and N. Goldman. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinformatics **11**:579.

Lunter, G. 2007. Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. Bioinformatics **23**:i289–296.

Mailund, T., and C. N. Pedersen. 2004. QDist–quartet distance between evolutionary trees. Bioinformatics **20**:1636–1637.

Müller, K. 2006. Incorporating information from length-mutational events into phylogenetic analysis. Mol. Phylogenet. Evol. **38**:667–676.

Murphy, W. J., P. A. Pevzner, and S. J. O'Brien. 2004. Mammalian phylogenomics comes of age. Trends Genet. **20**:631–639.

Murphy, W. J., T. H. Pringle, T. A. Crider, M. S. Springer, and W. Miller. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. Genome Res. **17**:413–421.

Nikolaev, S., J. I. Montoya-Burgos, E. H. Margulies, J. Rougemont, B. Nyffeler, and S. E. Antonarakis. 2007. Early history of mammals is elucidated with the ENCODE multiple species sequencing data. PLoS Genet. **3**:e2.

Notredame, C., D. G. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. **302**:205–217.

Ogurtsov, A. Y., S. Sunyaev, and A. S. Kondrashov. 2004. Indel-based evolutionary distance and mouse-human divergence. Genome Res. **14**:1610–1616.

Pattengale, N. D., M. Alipour, O. R. Bininda-Emonds, B. M. Moret, and A. Stamatakis. 2010. How many bootstrap replicates are necessary? J. Comput. Biol. **17**:337–354.

Penny, D., L. R. Foulds, and M. D. Hendy. 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. Nature **297**:197–200.

Prasad, A. B., M. W. Allard, and E. D. Green. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol. Biol. Evol. **25**:1795–1808.

Redelings, B. D., and M. A. Suchard. 2007. Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. BMC Evol. Biol. **7**:40.

Rivas, E. 2005. Evolutionary models for insertions and deletions in a probabilistic modeling framework. BMC Bioinformatics **6**:63.

Rivas, E., and S. R. Eddy. 2008. Probabilistic Phylogenetic Inference with Insertions and Deletions. PLoS Comput Biol **4**:e1000172.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Mathematical Biosciences **53**:131 – 147.

Siepel, A., G. Bejerano, J. S. Pedersen, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. **15**:1034–1050.

Simmons, M. P., and H. Ochoterena. 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol. **49**:369–381.

Simmons, N. B., and J. Geisler. 1998. Phylogenetic relationships of *Icaronycteris*, *Archeonycteris*, *Hassianycteris*, and *Palaeochiropteryx* to extant bat lineages, with comments on the evolution of echolocation and foraging strategies in Microchiroptera. Bull. Am. Mus. Nat. Hist. **235**:1–182.

Springer, M. S., M. J. Stanhope, O. Madsen, and W. W. de Jong. 2004. Molecules consolidate the placental mammal tree. Trends Ecol. Evol. (Amst.) **19**:430–438.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**:2688–2690.

Stamatakis, A., P. Hoover, and J. Rougemont. 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst. Biol. **57**:758–771.

Stanhope, M. J., V. G. Waddell, O. Madsen, W. de Jong, S. B. Hedges, G. C. Cleven, D. Kao, and M. S. Springer. 1998. Molecular evidence for multiple origins of Insectivora and for a new order of endemic African insectivore mammals. Proc. Natl. Acad. Sci. U.S.A. **95**:9967–9972.

Stevenson, G., B. Neal, D. Liu, M. Hobbs, N. H. Packer, M. Batley, J. W. Redmond, L. Lindquist, and P. Reeves. 1994. Structure of the O antigen of Escherichia coli K-12 and the sequence of its rfb gene cluster. J. Bacteriol. **176**:4144–4156.

Subramanian, A. R., M. Kaufmann, and B. Morgenstern. 2008. DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms Mol Biol **3**:6.

Teeling, E. C., O. Madsen, R. A. Van den Bussche, W. W. de Jong, M. J. Stanhope, and M. S. Springer. 2002. Microbat paraphyly and the convergent evolution of a key innovation in Old World rhinolophoid microbats. Proc. Natl. Acad. Sci. U.S.A. **99**:1431–1436.

Teeling, E. C., M. S. Springer, O. Madsen, P. Bates, S. J. O'brien, and W. J. Murphy. 2005. A molecular phylogeny for bats illuminates biogeography and the fossil record. Science **307**:580–584.

Than, C., D. Ruths, and L. Nakhleh. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics **9**:322.

Wang, A. X., W. L. Ruzzo, and M. Tompa. 2007. How accurately is ncRNA aligned within whole-genome multiple alignments? BMC Bioinformatics **8**:417.

Wildman, D. E., M. Uddin, J. C. Opazo, G. Liu, V. Lefort, S. Guindon, O. Gascuel, L. I. Grossman, R. Romero, and M. Goodman. 2007. Genomics, biogeography, and the diversification of placental mammals. Proc. Natl. Acad. Sci. U.S.A. **104**:14395–14400.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. (Amst.) **11**:367–372.

Zhang, Z., and M. Gerstein. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res. **31**:5338–5348.