

Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies

Korbinian Strimmer and Arndt von Haeseler

Zoologisches Institut, Universität München

A versatile method, quartet puzzling, is introduced to reconstruct the topology (branching pattern) of a phylogenetic tree based on DNA or amino acid sequence data. This method applies maximum-likelihood tree reconstruction to all possible quartets that can be formed from n sequences. The quartet trees serve as starting points to reconstruct a set of optimal n -taxon trees. The majority rule consensus of these trees defines the quartet puzzling tree and shows groupings that are well supported. Computer simulations show that the performance of quartet puzzling to reconstruct the true tree is always equal to or better than that of neighbor joining. For some cases with high transition/transversion bias quartet puzzling outperforms neighbor joining by a factor of 10. The application of quartet puzzling to mitochondrial RNA and tRNA^{Val} sequences from amniotes demonstrates the power of the approach. A PHYLIP-compatible ANSI C program, PUZZLE, for analyzing nucleotide or amino acid sequence data is available.

Introduction

In recent years the maximum-likelihood method for reconstructing phylogenetic relationships (Felsenstein 1981) has become more popular due to the arrival of powerful computers. The main advantage of a maximum-likelihood approach is the application of a well-defined model of sequence evolution to a given data set (Felsenstein 1988). Although the application of the maximum-likelihood method to biological data is now widespread, its computational complexity prevents computation for a large number of sequences. Generally, only slow programs for analyzing nucleotide or amino acid sequences are available (Felsenstein 1993; Yang 1995), although it is possible to speed up calculations by parallelizing the algorithm or using approximative techniques (Adachi and Hasegawa 1994; Olsen et al. 1994). Still, large trees can only be analyzed on massively parallel systems or by constraining the tree topology.

The principal goal of a maximum-likelihood analysis is the determination of a tree and corresponding branch lengths that have the greatest likelihood of generating the data. This task can be split into two parts: determining a tree topology and subsequently assigning branch lengths to the topology to obtain a maximum-likelihood estimate. Because the number of possible tree topologies grows exponentially with the number of sequences, all tree reconstruction methods that optimize an objective function have to rely on heuristic searches to find the best topology. Moreover, the optimization of branch lengths for a given topology is a tedious procedure for maximum-likelihood-based tree reconstruction methods and consumes a lot of computing time (Olsen et al. 1994). While maximum-likelihood procedures are generally slow for the general case of n sequences, the determination of the maximum-likelihood tree based on DNA or amino acid sequences poses no problem for four sequences. On the other hand, methods abound that

try to reconstruct a tree topology considering only the branching pattern of the $\binom{n}{4}$ different quartet trees that can be constructed from n sequences (Sattath and Tversky 1977; Fitch 1981; Bandelt and Dress 1986; Dress, von Haeseler, and Krüger 1986). It has been shown (Schöniger and von Haeseler 1993) that these distance-based methods exhibit performance similar to neighbor joining (Saitou and Nei 1987) while generally being much slower.

In this paper we describe a new method, quartet puzzling, for reconstructing phylogenetic relationships. This method reconstructs the maximum-likelihood tree for each of the $\binom{n}{4}$ possible quartets. In a so-called puzzling step the resulting quartet trees are then combined to an overall tree. During the puzzling step sequences are added sequentially in random order to an already-existing subtree. The position of a new sequence is determined by a voting procedure, considering all quartets. Finally, an intermediate tree relating n sequences is obtained. In general, there is no n -taxon tree that fits all the $\binom{n}{4}$ different quartet trees. Therefore, the puzzling step is repeated several times, thereby elucidating the landscape of possible optimal trees. The quartet puzzling tree is obtained as a majority-rule consensus (Margush and McMorris 1981) of all trees that result from multiple runs of the puzzling step. Depending on the phylogenetic information contained in the data, this tree may be binary or multifurcating. In addition to the tree topology the quartet puzzling tree also shows reliability values for each internal branch. In the next section of this paper the accuracy of the method is analyzed. As an illustration, quartet puzzling is applied to evaluate the phylogenetic relationship among the amniotes (Hedges 1994).

The Quartet Puzzling Algorithm

Quartet puzzling essentially is a three-step procedure, first reconstructing all possible quartet maximum-likelihood trees (maximum-likelihood step), then repeatedly combining the quartet trees to an overall tree (puzzling step), and finally computing the majority rule consensus of all intermediate trees giving the quartet puzzling tree (consensus step).

Key words: bootstrapping, consensus tree, maximum likelihood, tree reconstruction, reliability of internal branches, quartet puzzling, quartet trees.

Address for correspondence and reprints: Arndt von Haeseler, Zoologisches Institut, Universität München, Luisenstraße 14, D-80333 München, Germany. E-mail: arndt@zi.biologie.uni-muenchen.de.

Mol. Biol. Evol. 13(7):964–969, 1996

© 1996 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

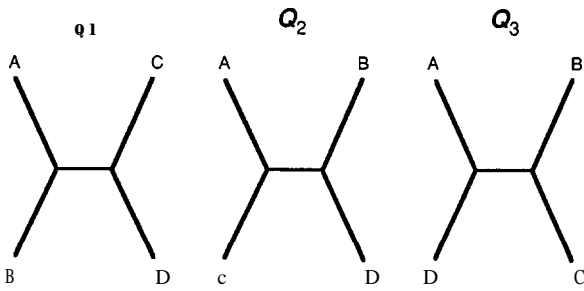


FIG. 1.—The three possible topologies for a four-taxon tree

The first step in the quartet puzzling analysis is the reconstruction of the branching pattern of all possible (4) quartets with maximum likelihood. For each quartet (A, B, C, D) three topologies $Q_1, Q_2,$ and Q_3 (fig. 1) exist with corresponding maximum-likelihood values $m_1, m_2,$ and m_3 . All topologies Q_i with $m_i = \max\{m_1, m_2, m_3\}$ are optimal topologies in the maximum-likelihood sense and are stored for the puzzling step. If there is more than one best topology, the branching pattern of the quartet (A, B, C, D) is not uniquely defined. In this case we choose randomly between the available optimal topologies every time we look up the branching pattern of (A, B, C, D). Thus, maximum-likelihood tree reconstruction induces a neighbor relation \parallel_{ml} between any four taxa A, B, C, and D (Bandelt and Dress 1986). The neighbor relation $AB\parallel_{ml} CD$ implies that taxa A and B and taxa C and D are neighbors with respect to each other. Note that in the corresponding tree Q, (fig. 1) the paths connecting the taxa A and B and the taxa C and D are disjoint.

Next, in the puzzling step, we aim to combine the quartet trees to an overall n-taxon tree. Generally the neighbor relation \parallel_{ml} on the set of all n taxa is not tree-like (Bandelt and Dress 1986), therefore it is necessary to apply approximation methods to obtain an overall tree topology (Sattath and Tversky 1977; Fitch 1981; Bandelt and Dress 1986; Dress, von Haeseler, and Kruger 1986). We suggest the following simple algorithm. First, the input order of the n taxa is randomized; let us assume that the order is A, B, C, D, E, The maximum-likelihood tree of the quartet (A, B, C, D) is now used as a seed for the overall n-taxon tree. Then taxon E is added to the subtree according to the following voting procedure: The neighbor relation \parallel_{ml} induces for every quartet (i, j, k, E) a clustering i, j versus k, E, say. It is obvious that taxon E should not be placed on a branch that lies on the path connecting i and j in the subtree. The edges where E should not be placed in the subtree are marked for every quartet (i, j, k, E). Thus, every branch in the subtree is assigned a score. If all different quartets containing taxon E and three taxa of the subtree are evaluated, species E is inserted at that branch in the tree that shows the lowest score. If the minimal score is attained for more than one edge, the sequence is inserted randomly at one of the equally good branches. Figure 2 illustrates the procedure for five taxa. The addition of a single taxon is repeated until an overall tree of n taxa is obtained. The randomized sequential insertion of se-

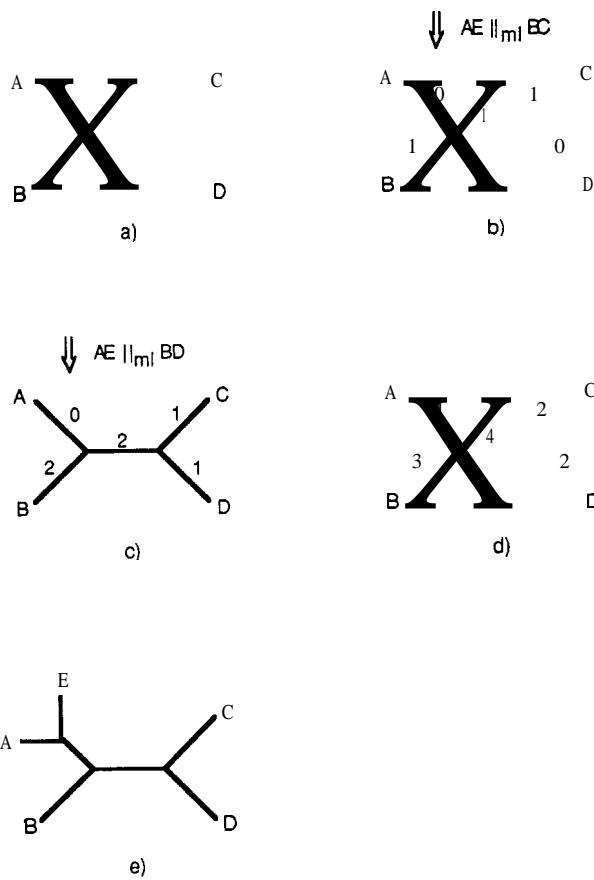


FIG. 2.—Addition of sequence E to the already-existing four-taxon tree (a). The neighbor relations are given by $AE\parallel_{m_1} BC, AE\parallel_{m_1} BD, AC\parallel_{m_1} DE,$ and $BD\parallel_{m_1} CE$. The relation $AE\parallel_{m_1} BC$ implies that the branches connecting B and C each get a score of one (b). (c) The score of the branches if $AE\parallel_{m_1} BD$ is evaluated. If all four quartets are analyzed, the branch leading to taxon A shows the lowest score (d). Hence, E is inserted at this branch (e).

quences may not always lead to the same tree topology for different runs of the puzzling step. Therefore, step two is repeated as often as possible, thereby elucidating the landscape of all possible optimal trees. Generally, the more taxa involved the more runs of the puzzling step are advised.

In the third step of the quartet puzzling algorithm a majority rule consensus (Margush and McMorris 1981) is computed from the intermediate trees resulting from the puzzling steps. We call this consensus tree the quartet puzzling tree. Depending on the phylogenetic information contained in the data the quartet puzzling tree is either completely resolved or shows multifurcations. In addition to the tree topology the quartet puzzling tree also provides information about the number of times a particular grouping occurred in the intermediate trees. If the resolution of phylogenetic relationships between a subset of sequences is unclear, the consensus tree will indicate it by displaying small reliability values for the corresponding internal branches. The repeated randomization of the input order of the taxa and subsequent computation of an intermediate tree results in a collection of locally optimal trees that are generated indepen-

dently of each other. In contrast, a collection of trees derived by procedures like branch swapping from one starting tree produces nonindependent trees (Penny et al. 1995). Thus, given the independence, the consensus tree gives a summary of all groups that occur in the majority of the intermediate trees.

The reliability values, i.e., the number of times the group is reconstructed during the puzzling steps, allow a simple interpretation of the phylogenetic information present in the data. Every intermediate tree represents a solution from the set of optimal trees. If we were able to compute all optimal trees, then all clusters that appear in more than 50% of the optimal trees fit into an overall tree (Margush and McMorris 1981). This not-necessarily bifurcating tree represents the total phylogenetic information. However, due to the limited number of puzzling steps, only a subset of all optimal trees are found. Therefore, it is advisable to trust only reliability values that are well above 50%. Note that the suggested reliability measure should not be confused with the usual bootstrap values. Whereas reliability values are an intrinsic result of the quartet-puzzling algorithm, bootstrapping is an external procedure that can be applied to any tree-building method. Quite remarkably however, it seems that both measures are highly correlated (unpublished data).

Quartet puzzling therefore is a simple method to get a phylogenetic tree and simultaneously an impression of how well the data are suited for a phylogenetic reconstruction.

Efficiency of Quartet Puzzling

It is easy to prove that quartet puzzling reconstructs the underlying tree if the neighbor relation $\|_{mt}$ is treelike (Bandelt and Dress 1986). However, real data hardly ever are treelike. To study the efficiency of our approach we employed a computer simulation, in which we compared the reconstructed trees with the model trees. We compared the efficiency of quartet puzzling with the performance of neighbor joining and maximum likelihood. The simulation settings are analogous to that employed in Schöniger and von Haeseler (1993).

Maximum likelihood was used as implemented in the PHYLIP DNAML version 3.5 program (Felsenstein 1993), quartet puzzling as implemented in version 2.3 of the PUZZLE program. Details about PUZZLE are given in the appendix. The results for the performance of neighbor joining are adopted from (Schöniger and von Haeseler 1993). Model trees T_1 and T_2 are displayed in figure 3. For each of the two model trees a variety of substitution rates a and b have been assumed. Sequences were evolved according to the Jukes-Cantor model (Jukes and Cantor 1969) and to Kimura's two-parameter model (Kimura 1980). The ratio t of the number of expected transitions to the number of expected transversions is $t=1/2$ in the Jukes-Cantor and $t=4$ in the Kimura case. Simulations were carried out with sequences of lengths 500 and 1,000. For each combination of parameters, 1,000 simulations were carried out. For DNAML, however, only 100 simulations were possible

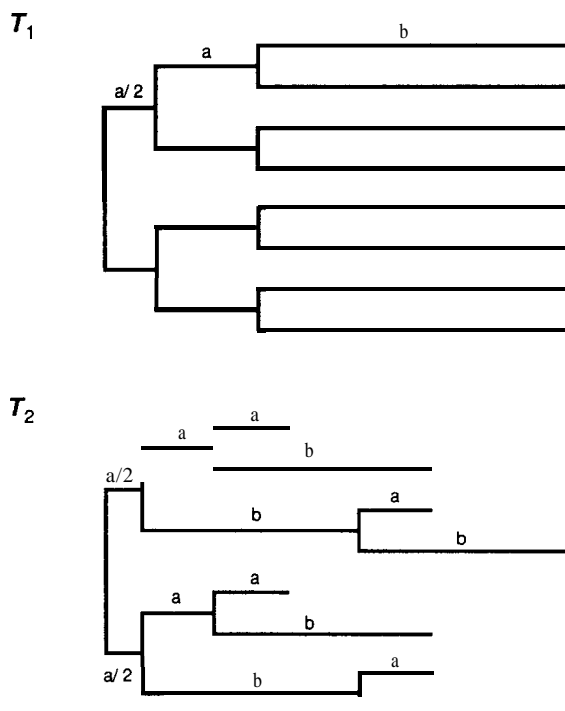


FIG. 3.—Model trees T_1 and T_2 with expected substitution rates a and b . In T_1 a molecular clock is assumed, whereas T_2 describes a situation of extreme rate heterogeneity in different branches of the tree.

due to the large computational costs involved. All programs were run with their default options set; only the transition/transversion parameter was set both in DNAML and PUZZLE equal to 1/2 or 4, according to the mode of assumed sequence evolution. Quartet puzzling was performed with 1,000 puzzling steps and with the approximation option for the maximum-likelihood computation invoked. Neighbor-joining results are displayed using Kimura corrected distances (Kimura 1980).

From tables 1 and 2 it is obvious that maximum likelihood generally outperforms both neighbor joining and quartet puzzling. Unfortunately, the computational costs of DNAML are prohibitively high when the number of taxa is large.

The performance of neighbor joining and quartet puzzling is different depending on the choice of parameters. As expected, an increase in sequence length leads to a better performance of each method. If sequences evolved according to a Jukes-Cantor model, both methods show a more or less identical efficiency. Quartet puzzling is slightly superior if the clock assumption is violated and if substitution rates are high. If sequences evolved under a Kimura model evolution with a transition/transversion ratio of $t=4$, the quartet puzzling method outperforms neighbor joining, irrespective of whether the tree follows a molecular clock (T_1) or not (T_2). For high substitution rates the efficiency of quartet puzzling is up to 10 times better than that of neighbor joining.

The Phylogeny of Amniotes

We have reanalyzed the concatenated sequences of amniote mitochondrial 12S rRNA, 16S rRNA, and t -

Table 1
Percentage of Correctly Reconstructed Trees Assuming Clocklike Evolution According to Tree T_1

SEQUENCE EVOLUTION		JUKES-CANTOR ($r = \frac{1}{2}$)			KIMURA ($r = 4$)		
l	ab	NJ	QP	ML	NJ	QP	ML
500 .	0.01/0.07	70.5	71.5	87	56.6	57.8	70
	0.02/0.19	52.0	54.4	63	23.1	42.5	48
	0.03/0.42	8.2	11.3	9	1.4	14.2	15
1,000	0.01/0.07	94.7	93.8	96	87.3	87.0	93
	0.02/0.19	86.8	86.0	85	59.3	75.3	85
	0.03/0.42	38.3	36.6	34	10.8	35.6	38

NOTE.—Estimates of efficiencies are based on 1,000 simulations (neighbor joining [NJ], quartet puzzling [QP]) and 100 simulations (maximum likelihood [ML]). Sequence length is denoted by l , branch lengths by a and b , and the expected transition-transversion ratio by r .

RNA^{val} genes (Hedges 1994). The data set comprises 15 species, among them six placental mammals, one bird, four reptiles, one frog and three lungfish sequences. More specifically, the species involved are *Neoceratodus forsteri* (lungfish, Australia), *Lepidosiren paradoxa* (lungfish, South America), *Protopterus* sp. (lungfish, Africa), *Xenopus laevis* (frog), *Trachemys scripta* (turtle), *Sphenodon punctatus* (sphenodontid), *Sceloporus undulatus* (lizard), *Alligator mississippiensis* (crocodilian), *Gallus gallus* (bird), *Homo sapiens* (human), *Phoca vitulina* (seal), *Bos taurus* (cow), *Balaenoptera physalus* (whale), *Mus musculus* (mouse), and *Rattus norvegicus* (rat). In addition the corresponding sequences from *Didelphis virginiana* (opossum) (Janke et al. 1994) and *Ornithorhynchus anatinus* (platypus) (Janke et al. 1996) were included in the analysis. The 17 sequences were aligned using CLUSTAL W (Thompson, Higgins, and Gibson 1994) resulting in an alignment of length 2,903. After removing ambiguous alignment positions, 2,439 sites remained for further analysis (data set available on request). In the PUZZLE program the model of sequence evolution by Hasegawa, Kishino, and Yano (1985) with the transition/transversion parameter set to $r = 1$ was selected and the approximation option for the maximum-likelihood computation was invoked.

A total of 2,380 four-species maximum-likelihood trees were reconstructed in the first step of the quartet puzzling algorithm. Among all quartet trees there were only 31 (1.3%) bad quartets. A quartet is called bad quartet if the following inequality is violated:

$$m_1 - m_2 > m_2 - m_3, \tag{1}$$

where $m_1 > m_2 > m_3$. Thus, for a bad quartet the best and the second best maximum-likelihood values are very close together, indicating that at least two different branching patterns are likely. A high percentage of bad quartets usually indicates that the data set is not very well suited for a phylogenetic analysis. If more than 10%–15% bad quartets are present the quartet puzzling tree is in general not completely resolved. In the cases investigated the bad quartets were distributed uniformly over all taxa (data not shown). In those situations the number of bad quartets can be viewed as a measure of the “background noise” obscuring the phylogenetic signal present in the data. For the amniote data the number of bad quartets is very small and hence the sequences are suitable for a phylogenetic analysis.

The resulting phylogeny, after performing 1,000 puzzling steps, is shown in figure 4. Note that only about 5 min CPU time on a standard personal computer (Macintosh 6100/66) were necessary to compute this quartet puzzling tree. The tree topology coincides more or less with the already-published tree (Hedges 1994). As expected from the small number of bad quartets, the support for the branches in the tree is very high. Our analysis supports the view that crocodilians are the closest living relatives of birds. In 100% of trees underlying the quartet puzzling tree the bird–crocodilian clade is found, indicating the clear separation from the remaining sequences in the tree. This high support from our analysis is matched by the high bootstrap support (Hedges 1994). Incidentally, the alternative clade placental mammals–bird was never detected in any of the 1,000 intermediate

Table 2
Percentage of Correctly Reconstructed Trees for Nonclocklike Evolution Assuming Tree T_2

SEQUENCE EVOLUTION		JUKES-CANTOR ($r = \frac{1}{2}$)			KIMURA ($r = 4$)		
l	ab	NJ	QP	ML	NJ	QP	ML
500	0.01/0.07	79.7	83.6	91	71.7	74.2	94
	0.02/0.19	64.8	75.3	93	38.6	65.8	92
	0.03/0.42	18.1	33.3	72	3.5	36.5	73
1,000	0.01/0.07	96.1	96.7	99	91.8	94.8	98
	0.02/0.19	91.3	93.5	99	67.6	88.4	99
	0.03/0.42	37.9	59.2	92	7.7	61.7	96

NOTE.—Abbreviations are explained in Table 1.

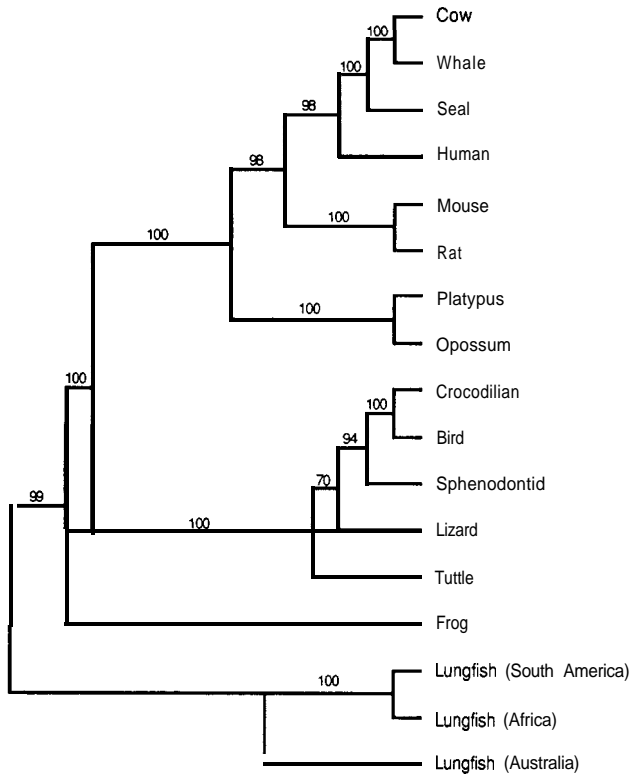


FIG. 4.—Quartet puzzling tree based on 1,000 puzzling steps. The reliability value of each internal branch indicates in percent how often the corresponding cluster was found among the 1,000 intermediate trees. Note that branch lengths are arbitrary; only the branching pattern is important. The lungfishes are used as outgroup to root the tree of the amniotes.

trees. Furthermore, we find that bird, crocodilian, and sphenodontid form a monophyletic group with a reliability value of 94%. The phylogenetic relationship among this group, lizard, and turtle is less clear because of a relatively low reliability of 70% for the corresponding internal branch. Contrary to the tree published in Hedges (1994), our branching pattern suggests that within the radiation of placental mammals the rodents branch off first and the humans are a sister group of the Ferungulata, a result in perfect agreement with other studies (Janke et al. 1994). Our results also support the sister group relationship of marsupials and monotremes (Janke et al. 1996). Thus, quartet puzzling analysis confirms the close relationship of birds and crocodilians and proposes a branching pattern of placental mammals which coincides with other analyses (Janke et al. 1994, 1996).

Discussion

We have presented the quartet puzzling method to reconstruct tree topologies from sequence data. This method computes the maximum-likelihood tree for all possible quartets. An intermediate *n*-taxon tree is computed in the so-called puzzling step. The repeated application of the puzzling step allows an assignment of reliability values to the groupings in the final quartet puzzling tree, a consensus tree built from all intermediate trees. If groups are found only occasionally in dif-

ferent runs of the puzzling step, they will obtain a low reliability value. In those situations it is more realistic to assume a multifurcation rather than a bifurcation.

Moreover, we have shown that quartet puzzling either shows performance comparable to or better than neighbor joining. If sequences evolved according to the Jukes-Cantor model and obeyed a molecular clock, both methods have a similar efficiency. In situations where neighbor joining performs badly, quartet puzzling has the advantage of not falling into the traps provided by the complex landscape of the tree space. The repeated application of the puzzling step prevents the method from getting trapped in local optima. This “trap avoiding” property stems from the various averaging procedures that are present in quartet puzzling. Finally, our analysis of the amniote sequence data shows that quartet puzzling performs very well on real as well as on simulated data sets.

Acknowledgments

Support from the DFG is greatly acknowledged. We also want to thank Jun Adachi and Masami Hasegawa for distributing MOLPHY as free software and generously providing the source code of their program. We thank the European Bioinformatics Institute for kindly distributing the PUZZLE program. Finally, we would like to thank Oliva Handt for carefully reading this manuscript and Svante Pääbo and all the members of his group for providing a stimulating environment.

APPENDIX

A computer program, PUZZLE, for analyzing nucleotide and amino acid sequence data with the quartet puzzling method is available. PUZZLE is menu-driven and PHYLIP-compatible. It is written in ANSI C and has been tested on all popular platforms (MacOS, MS-DOS, UNIX, VMS). Current versions can be retrieved over the Internet from the server of the European Bioinformatics Institute (<ftp://ftp.ebi.ac.uk/pub/software>). Parts of PUZZLE are taken from the free software MOLPHY by Jun Adachi and Masami Hasegawa (Adachi and Hasegawa 1994).

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1994. MOLPHY: programs for molecular phylogenetics. Version 2.2. Institute of Statistical Mathematics, Tokyo.

BANDELT, H.-J., and A. DRESS. 1986. Reconstructing the shape of a tree from observed dissimilarity data. *Adv. Appl. Math.* **7**:309–343.

DRESS, A., A. VON HAESLER, and M. KRUGER. 1986. Reconstructing phylogenetic trees using variants of the four-point condition. *Studien zur Klassifikation* **17**:299–305.

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

———. 1988. Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* **22**:521–565.

———. 1993. PHYLIP: phylogenetic inference package. Version 3.5c. Department of Genetics, University of Washington, Seattle.

- FITCH, W. M. 1981. A non-sequential method for constructing trees and hierarchical classifications. *J. Mol. Evol.* 18:30-37.
- HASEGAWA, M., H. KISHINO, and K. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22: 160-174.
- HEDGES, S. B. 1994. Molecular evidence for the origin of birds. *Proc. Natl. Acad. Sci. USA* 91:2621-2624.
- JANKE, A., G. FELDMAIER-FUCHS, N. GEMMELL, A. VON HAESELER, and S. PÄÄBO. 1996. The complete mitochondrial genome of the platypus (*Ornithorhynchus anatinus*) and the evolution of mammals. *J. Mol. Evol.* 42:153-159.
- JANKE, A., G. FELDMAIER-FUCHS, W. K. THOMAS, A. VON HAESELER, and S. PÄÄBO. 1994. The marsupial mitochondrial genome and the evolution of placental mammals. *Genetics* 137:243-256.
- JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21-132 in H. N. MUNRO, ed. *Mammalian protein metabolism*. Academic Press, New York.
- KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16: 111-120.
- MARGUSH, T., and F. R. MCMORRIS. 1981. Consensus n-trees. *Bull. Math. Biol.* 43:239-244.
- OLSEN, G. J., H. NATSUDA, R. HAGSTROM, and R. OVERBEEK. 1994. FastDNAML: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *CABIOS* 10:41-48.
- PENNY, D., M. STEEL, P. J. WADDELL, and M. D. HENDY. 1995. Improved analyses of human mtDNA sequences support a recent African origin of *Homo sapiens*. *Mol. Biol. Evol.* 12: 863-882.
- SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406-425.
- SATTATH, S., and A. TVERSKY. 1977. Additive similarity trees. *Psychometrika* 42:319-345.
- SCHONIGER, M., and A. VON HAESELER. 1993. A simple method to improve the reliability of tree reconstructions. *Mol. Biol. Evol.* 10:471-483.
- THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-4680.
- YANG, Z. 1995. *Phylogenetic Analysis by maximum likelihood (PAML)*. Version 1.1. Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, Pa.

PAUL M. SHARP, reviewing editor

Accepted April 25, 1996