

Bayesian Probabilities and Quartet Puzzling

Korbinian Strimmer,* Nick Goldman,† and Arndt von Haeseler*

*Zoologisches Institut, Universität München; and †Department of Genetics, University of Cambridge

Quartet puzzling (QP), a heuristic tree search procedure for maximum-likelihood trees, has recently been introduced (Strimmer and von Haeseler 1996). This method uses maximum-likelihood criteria for quartets of taxa which are then combined to form trees based on larger numbers of taxa. Thus, QP can be practically applied to data sets comprising a much greater number of taxa than can other search algorithms such as stepwise addition and subsequent branch swapping as implemented, e.g., in DNAML (Felsenstein 1993). However, its ability to reconstruct the true tree is less than that of DNAML (Strimmer and von Haeseler 1996). Here, we show that the assignment of penalties in the puzzling step of the QP algorithm is a special case of a more general Bayesian weighting scheme for quartet topologies. Application of this general framework leads to an improvement in the efficiency of QP at recovering the true tree as well as to better theoretical understanding of the method itself. On average, the accuracy of QP increases by 10% over all cases studied, without compromising speed or requiring more computer memory.

Consider the three different fully-bifurcating tree topologies Q_1 , Q_2 , and Q_3 for four taxa (fig. 1). Denote by m_1 , m_2 , and m_3 their corresponding maximum-likelihood (not log-likelihood) values. Note that $m_1 + m_2 + m_3 \ll 1$. Evaluation via Bayes' theorem of the three tree topologies given uniform prior information leads to posterior probabilities

$$p_i = \frac{m_i}{m_1 + m_2 + m_3} \quad (1)$$

for each quartet Q_i (Lindgren 1976, Kishino and Hasegawa 1989), with $p_1 + p_2 + p_3 = 1$. From an inferential point of view it is natural to use these Bayesian probabilities p_i as weights w_i for each quartet in the tree-building process.

If we investigate the puzzling step of the original QP algorithm more closely we see that for each quartet a penalty of 1 is assigned for the quartet topology that shows the highest maximum-likelihood value and a penalty of 0 is assigned for the other two topologies. Thus, an implicit weighting of quartet topologies occurs using weights $w_{\max} = 1$, $w_{\text{other}} = 0$. In fact, when the probabilities p_i are computed for real data sets, most of the quartets show posterior probabilities $p_{\max} \approx 1$, $p_{\text{other}} \approx 0$, justifying the ad hoc QP procedure. However, if sequences are short or very closely related and if therefore not all quartet trees can be confi-

dently resolved, the Bayesian posteriors p_i may deviate substantially from this simple picture. It is desirable to incorporate the additional information provided by the probabilities p_i into the QP algorithm to improve the tree reconstruction process in these cases.

The most straightforward implementation of this idea is to use for each quartet of taxa (A, B, C, D) the Bayesian probabilities p_i as weights w_i . For each of the three possible topologies Q_i , and not only for the one showing the highest maximum likelihood, we assign a penalty of w_i along the corresponding branches (Strimmer and von Haeseler 1996). We call this approach QP using continuous weights, in contrast to the originally proposed QP where implicitly three discrete weights are applied. However, this new procedure has drawbacks. First, three times more assignments, and floating-point instead of integer calculations, are necessary and, consequently, the tree reconstruction process is slowed down significantly. Second, much more memory space is needed to store all the different weights w_i . This problem prohibits the use of this method for larger numbers of taxa.

We have therefore investigated another natural extension of the original QP procedure for assigning penalties. It is well known that there are three different unrooted binary trees connecting four taxa. However, if we also consider the completely unresolved star tree and the three partially resolved quartet trees then we count in total seven different topologies (Eigen, Winkler-Oswatitsch, and Dress 1988). To each of the seven trees corresponds a set of discrete weights w_i (fig. 1), according to which bifurcating trees may be obtained by resolving the partially resolved networks. When quartets are evaluated in the maximum-likelihood step of the QP algorithm we choose among these seven permitted sets of weights by selecting that which minimizes the least-squares distance

$$d = \sum_{i=1}^3 (p_i - w_i)^2. \quad (2)$$

Thus, we approximate the Bayesian probabilities p_i by one of the seven sets of discrete weights w_i . This procedure has several advantages. Not only can all necessary storage now be done without demanding additional computer memory, but it also makes it possible to avoid floating-point calculations when QP penalties are distributed over branches while still applying fractional weights. If we have decided on one of the three completely resolved quartet trees Q_i (fig. 1) we assign a penalty of 1 along the corresponding branches as usual. However, if we have ended up in one of the four non-bifurcating quartets we assign a penalty of 1 for precisely one randomly chosen bifurcating quartet topology that is compatible with it. In this way we effectively apply weights w_i as an average over all puzzling steps performed, but still use only integer calculation. We call this approach QP using discrete weights.

Key words: accuracy, assigning penalties, Bayesian evaluation, posterior probabilities, quartet networks, quartet puzzling.

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. 14(2):210–211. 1997

© 1997 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

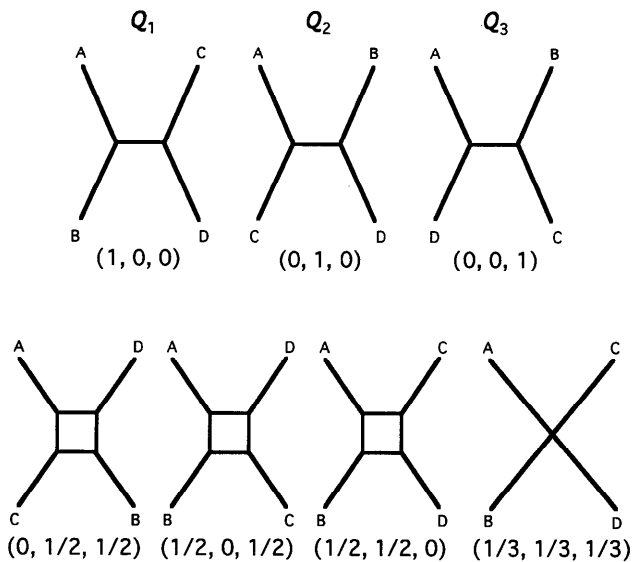


FIG. 1.—The seven possible binary and nonbinary trees for a quartet of taxa (A, B, C, D), and the corresponding discrete weights (w_1, w_2, w_3). The Q_i are the three completely resolved quartet trees.

To evaluate the accuracies of the two modifications of the QP algorithm, we have repeated the computer simulations used in the original paper (Strimmer and von Haeseler 1996). We adopt that setup and those conventions. For several evolutionary scenarios, the performance of QP at reconstructing the correct tree was studied. Our results comparing the original QP (A), QP using discrete weights (B), and QP using continuous weights (C) are shown in Table 1 for the clock-like eight-taxon tree T_1 , and in Table 2 for the non-clock-like eight-taxon tree T_2 . In all cases examined methods B and C show a significant increase in efficiency over method A. This is very pronounced for high substitution rates. However, the differences of algorithms B and C are only very small. Procedure C is slightly better than method B except for high substitution rates on tree T_1 , where B performs better. This is remarkable, as method B is computationally faster and needs less computer memory than method C. We think that this is because

Table 1
Percentage of Correctly Recovered Trees T_1 if the Original QP Algorithm (A), QP Using Discrete Weights (B), and QP Using Continuous Weights (C) Are Applied

SEQUENCE EVOLUTION	a/b	TREE T_1					
		JC ^a			Km ^b		
1		A	B	C	A	B	C
500 ..	0.01/0.07	71.5	79.6	81.2	57.8	69.8	74.7
	0.02/0.19	54.4	69.7	66.3	42.5	63.3	58.6
	0.03/0.42	11.3	28.5	20.5	14.2	33.4	21.4
1,000 ..	0.01/0.07	93.8	93.5	96.9	87.0	89.2	92.3
	0.02/0.19	86.0	91.5	93.5	75.3	85.0	86.1
	0.03/0.42	36.6	52.9	42.8	35.6	57.2	48.6

NOTE.—Terminology and setup follow Strimmer and von Haeseler (1996). Sequence length is denoted by l , branch lengths by a and b , and the expected transition-transversion ratio by T .

^a Jukes-Cantor ($T = 1/2$).

^b Kimura ($T = 4$).

Table 2
Percentage of Correctly Recovered Trees T_2 if the Original QP Algorithm (A), QP Using Discrete Weights (B), and QP Using Continuous Weights (C) are Applied

SEQUENCE EVOLUTION	a/b	TREE T_2					
		JC ^a			Km ^b		
1		A	B	C	A	B	C
500 . . .	0.01/0.07	83.6	86.0	87.3	74.2	80.5	81.7
	0.02/0.19	75.3	84.9	84.7	65.8	77.4	78.5
	0.03/0.42	33.3	47.2	47.4	36.5	52.1	54.1
1,000 . . .	0.01/0.07	96.7	97.3	97.6	94.8	95.4	95.8
	0.02/0.19	93.5	96.2	96.4	88.4	91.5	92.1
	0.03/0.42	59.2	69.6	70.0	61.7	73.6	76.6

NOTE.—Terminology and setup follow Strimmer and von Haeseler (1996). Abbreviations are explained in Table 1.

^a Jukes-Cantor ($T = 1/2$).

^b Kimura ($T = 4$).

B already accounts for the seven possible unrooted trees of a quartet of taxa which cannot be substantially improved by considering a continuous spectrum of weights.

The QP algorithm using discrete weights (B) has been incorporated in version 2.5 of the PUZZLE program (Strimmer and von Haeseler 1996). There, it replaces algorithm A, which was used in versions 2.4 and earlier. PUZZLE can be retrieved over the Internet from the server of the European Bioinformatics Institute (<ftp://ftp.ebi.ac.uk/pub/software>).

Acknowledgments

We thank Gunter Weiss for helpful comments on the manuscript. Support from the Deutsche Forschungsgemeinschaft (K.S. and A.v.H.) is greatly appreciated. N.G.'s research is supported by a Wellcome Trust Fellowship in Biodiversity Research. K.S. and A.v.H. thank the European Bioinformatics Institute for kindly distributing the PUZZLE program. K.S. would also like to thank the Department of Genetics, University of Cambridge, for its hospitality during the short visit when this work was initiated.

LITERATURE CITED

- EIGEN, M., R. WINKLER-OSWATITSCH, and A. DRESS. 1988. Statistical geometry in sequence space: a method of quantitative comparative sequence analysis. *Proc. Natl. Acad. Sci. USA* **85**:5913–5917.
- FELSENSTEIN, J. 1993. PHYLIP: phylogenetic inference package. Version 3.5c. Department of Genetics, University of Washington, Seattle.
- KISHINO, H., and M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from {DNA} sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* **29**:170–179.
- LINDGREN, B. W. 1976. *Statistical theory*. 3rd edition. Macmillan, New York.
- STRIMMER, K., and A. VON HAESELER. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**:964–969.

PAUL M. SHARP, reviewing editor

Accepted November 7, 1996