On the Construction of "Optimal" Phylogenetic Trees

Geert De Soete *

Department of Psychology, University of Ghent
Henri Dunantlaan 2, B-9000 Ghent, Belgium

Z. Naturforsch. **38 c**, 156–158 (1983);
received September 29, 1982

Phylogenetic Tree, Classification, Molecular Evolution

An iterative algorithm for constructing the optimal phylogenetic tree from a given set of dissimilarity data is described. The procedure is applied for illustrative purposes an a data set compiled by Fitch and Margoliash.

A well-known method for studying molecular evolution starts off with calculating a measure of dissimilarity between certain proteins or nucleic acids of different species. These dissimilarity data are subsequently used to construct a phylogenetic tree (*cf. e.g.*, [1], [2]). A phylogenetic tree is a weighted tree in which the terminal nodes represent the species. The evolutionary distance or patristic difference between species i and j, denoted d_{ij}, is defined as the sum of the weights assigned to the links lying on the path that connects nodes i and j.

Let $\Delta = (\delta_{ij})$ be an n by n matrix containing the pairwise dissimilarities between n species, such that

$$\delta_{ii} = 0$$
$$\delta_{ij} > 0 \quad (i \neq j)$$
$$\delta_{ij} = \delta_{ji},$$

for all $i, j = 1, n$. It has been shown by several authors [3–5] that a necessary and sufficient condition for the existence of a (unique) phylogenetic tree is the so-called additive inequality

$$\delta_{ij} + \delta_{kl} \leqq \max(\delta_{ik} + \delta_{jl}, \delta_{il} + \delta_{jk}) \qquad (1)$$

for all i, j, k, and l. Whenever Δ satisfies (1), the tree can be constructed by straightforward means. In fact, every proof of the sufficiency of (1) entails a method for constructing the tree. Remark that condition (1) is equivalent to saying that the two largest of $(\delta_{ij} + \delta_{kl})$, $(\delta_{ik} + \delta_{jl})$, and $(\delta_{il} + \delta_{jk})$ should be equal for all i, j, k, and l.

In practice, however, an observed data set Δ never perfectly satisfies (1). In this case, one wants to find a phylogenetic tree whose patristic differences D are close to Δ. Fitch and Margoliash [1] suggested to select the tree which has the smallest percent "standard deviation" (PSD), which is defined as

$$\mathrm{PSD} = 100 \times \left[\frac{1}{N-1} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(\frac{\delta_{ij} - d_{ij}}{\delta_{ij}} \right)^2 \right]^{1/2},$$

where $N = n(n-1)/2$. The tree having the smallest PSD is called by Fitch and Margoliash the "optimal" phylogenetic tree. In this paper a method is described for constructing the optimal phylogenetic tree from a given Δ. As far as I know, this is the first direct algorithm for finding the optimal tree presented in the literature.

Since a set of distances D which perfectly satisfies the additive inequality, uniquely determines a phylogenetic tree, the problem of constructing the optimal phylogenetic tree reduces to finding the set of distances D which both satisfies the additive inequality *and* has the smallest PSD. This is equivalent to solving the following constrained nonlinear optimization problem

$$\text{minimize } L(D) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(\frac{\delta_{ij} - d_{ij}}{\delta_{ij}} \right)^2 \qquad (2)$$

subject to

$$d_{ij} + d_{kl} \leqq \max(d_{ik} + d_{jl}, d_{il} + d_{jk})$$
$$\text{for all } i, j, k, l.$$

Problem (2) can be solved by means of a sequential unconstrained minimization technique (*cf.* [6]). This technique consists of sequentially minimizing the augmented function

$$F(D, r) = L(D) + r P(D)$$

for an increasing sequence of values of r. The first component of $F(D, r)$ is monotonically related to the PSD, while the second one, $P(D)$, is a penalty function which measures how badly D satisfies the additive inequality. $P(D)$ is defined as

$$P(D) = \sum_{\Omega} (d_{ik} + d_{jl} - d_{il} - d_{jk})^2$$

---

* "Aspirant" of the Belgian "Nationaal Fonds voor Wetenschappelijk Onderzoek". I am indebted to Professor J. Hoste for providing computer facilities at the Institute of Nuclear Sciences at Ghent.

Reprint requests to G. De Soete.

0341-0382/83/0100-0156 $ 01.30/0

where Ω is the set of ordered quadruples (i, j, k, l) for which the additive inequality is violated, e.g.,

$$\Omega = \{(i,j,k,l) \mid d_{ij} + d_{kl} \leqq \min(d_{ik} + d_{jl}, d_{il} + d_{jk})\}.$$

Using the superscript q as iteration index, the iterative procedure for solving (2) can be schematically presented as follows:

(i) Initialize: $q=1$

$r^{(1)} = L(D^{(0)})/P(D^{(0)})$.

(ii) Minimize $F(D, r^{(q)})$ to obtain $D^{(q)}$.

(iii) Test for convergence: if $\sum_{i>j}\sum (d_{ij}^{(q)} - d_{ij}^{(q-1)})^2$ is less than some small constant stop, otherwise continue.

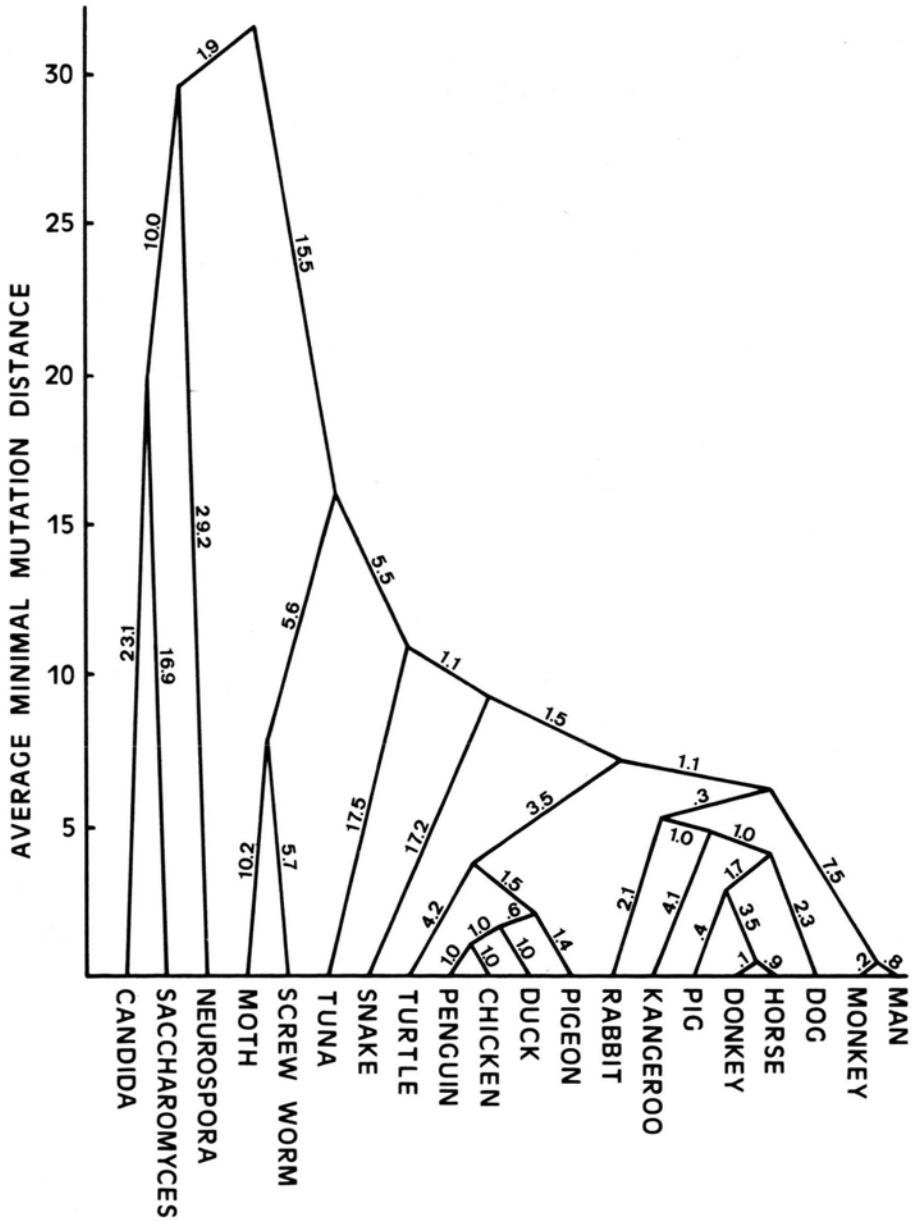

Fig. 1. Optimal phylogenetic tree for the Fitch and Margoliash data.

(iv) Update r: $r^{(q+1)} = 10 \times r^{(q)}$.
Increment q by one and go back to step (ii).

A computer program implementing this algorithm has been written in ANS FORTRAN and can be obtained from the author upon request. As initial estimate of D, $D^{(0)}$, the program uses

$$d_{ij}^{(0)} = \delta_{ij} + \varepsilon_{ij}$$

where ε_{ij} is a normal deviate. In step (ii) an unconstrained nonlinear minimization problem must be solved. This can be done in a variety of ways. Satisfactory results have been obtained with a conjugate gradient procedure incorporating automatic restarts due to Powell [7]. On the whole, the algorithm turns out to be quite efficient on a high-speed computer. Convergence usually occurs after cycling five or six times through the different steps described above.

The algorithm was applied for comparative purposes on a data set compiled by Fitch and Margoliash [1]. The data consist of mutation distances between twenty species and were obtained by counting the minimum number of mutations required to interrelate pairs of cytochromes *c*. Fitch and Margoliash tried out forty different phylogenetic trees. The lowest PSD obtained was 8.7. Figure 1 presents the phylogenetic tree constructed by the procedure outlined above. The PSD for this tree is 7.5. Following Fitch and Margoliash, the ordinate of an internal node is set equal to the average of the sums of all mutations in the lines of descent from the node. The reader is invited to compare this tree with the one presented in Figure 2 of [1].

[1] W. M. Fitch and E. Margoliash, Science **155,** 279 (1967).
[2] M. Goodman, J. Barnabas, G. Matsuda, and G. W. Moore, Nature **233,** 604 (1971).
[3] P. Buneman, J. Comb. Theory **17 B,** 48 (1974).
[4] A. J. Dobson, J. Appl. Prob. **11,** 32 (1974).
[5] A. N. Patrinos and S. L. Hakimi, Quart. Appl. Math. **30,** 255 (1972).
[6] A. V. Fiacco and G. P. McCormick, Non-linear programming: sequential unconstrained minimization techniques, Wiley, New York 1968.
[7] M. J. D. Powell, Math. Progr. **12,** 241 (1977).

Verantwortlich für den Inhalt: A. KLEMM
Satz und Druck: Konrad Triltsch, Würzburg