Algorithm for Construction of Phylogenetic Trees

José Tohá, M. A. Soto, and H. Chinga

Biofísica, Departamento de Física, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Casilla 487-3, Santiago, Chile

Z. Naturforsch. **44c**, 312–316 (1989); received August 12/November 2, 1988

Phylogenetic Tree Algorithm, Algorithm for Evolutionary Dendrograms

An algorithm is described for the construction of phylogenetic trees.
The algorithm is based on the progressive correction of data along the tree construction.
For the correction, the average value of the difference between each pair of neighbour elements to the rest of the table is considered.

An Algorithm for Construction of Phylogenetic Trees

Molecular phylogenetic trees are built up based on data of homologous sequences of nucleic acids or proteins of different species [1–16]. Unfortunately the observed changes in these sequences, not always represent faithfully the real distance or evolutionary span extended between the compared species, because: a) repeated point mutations at the same position of the DNA molecule are oversight, b) the degenerate nature of the genetic code, prevents the exact reckoning of changes produced at the third position of the codon and c) technical analytic deficiencies or others, contribute to the configuration of a very common type of table of distances displaying internal incoherence.

To overcome these difficulties we developed a simple method for the construction of phylogenetic trees, based on a progressive correction of the original table of data, provided that the distance table satisfies the ultrametric inequality (*i.e.* $d(x,y) \max\{d(x,z), d(y,z)\}$ for all x, y, z [17].The corrections performed at every step of the process consider the normalization of the data converging to a node or vertex, in accordance with the average distance of these points to the rest of the elements of the table [18, 19].

In this manuscript we communicate a general algorithm and the flow chart for its computational use, in such a way that in each trial the two nearest neighbours of the table are selected unambiguously. Moreover, we compare in the method two options: one where the values of the table are normalized before the beginning of the tree construction and the second one, more convenient as shown forward, where the correction to the table data is performed at each step along the dendrogram building. Also we analyze the degree of goodness of fit of the tree distances, after comparison with the distances taken from the original table. On the other hand, we compare the tree distances with the data of an ideal table constructed after a hierarchic arrangement of its elements.

Reprint requests to Dr. J. Tohá.

Verlag der Zeitschrift für Naturforschung, D-7400 Tübingen
0341–0382/89/0300–0312 $ 01.30/0

The Method

Firstly we select from the table the nearest pair of species displaying the shortest distance.

Then we calculate the average difference value of the distance of every element of the table and the two selected species.

As we need at least three elements and their distances to calculate the position of the common node joining the two selected species, we choose as third element that displaying a difference of its distances to the two nearest ones, equal or at least closest to the average difference value. When this difference is not equal to the average difference, a correction is performed in the distance values of the first and third element or in the second and third element, to obtain a similar value to the average (as shown forward). Moreover the same correction is accomplished in the distance of the first and second element to maintain the proportionality of the correction.

The correction introduced is usually positive because, as previously mentioned [18, 19], the data of the table are in general lower in magnitude than the real values or distances.

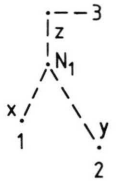
So, let as be 1 and 2 the nearest neighbours and 3 the selected reference.

Then:

$$\text{distance}\,(1\rightarrow 2) = \vec{x} + \vec{y} \qquad (1)$$

and

$$\text{distance}\,(2\rightarrow 3) - (1\rightarrow 3) = (\vec{y} + \vec{z}) - (\vec{x} + \vec{z}) = \vec{y} - \vec{x} \qquad (2)$$



The $\vec{y} - \vec{x}$ difference: $\blacktriangle(1,2)$, is then compared with the average difference $\bar{\blacktriangle}(1, 2) = \frac{1}{n-2} \sum_{i=3}^{n} (i-2) - (i-1)$ where, in the example, $\bar{\blacktriangle}(1, 2)$ is the value obtained after averaging all the differences between the distances of every element of the table to the elements 1 and 2. If one of the differences coincides with the $\bar{\blacktriangle}(1, 2)$ value, the third element involved in this difference is chosen for the resolution of the system of equations (1) and (2), and the $\vec{x}$ and $\vec{y}$ length determination $[\vec{x} = (1 \rightarrow N_1),\ \vec{y} = (2 \rightarrow N_1)]$; if not, the nearest difference to the average $\bar{\blacktriangle}$ is selected and the third element involved in this comparison is utilized for the determination of $\vec{x}$ and $\vec{y}$ values after the correction, in this case, of the $(1 \rightarrow 3)$ or the $(2 \rightarrow 3)$ distance value to attain a difference between both, equal to the average delta $\bar{\blacktriangle}(1, 2)$. If $\bar{\blacktriangle}(1, 2)$ is greater than $\blacktriangle(1, 2)$ then, the correction applies on the larger value $(2 \rightarrow 3)$ which has to be augmented. If, on the contrary, $\bar{\blacktriangle}(1, 2)$ is lower than $\blacktriangle(1, 2)$, the $(1 \rightarrow 3)$ distance value is increased. All that in accordance with the statement above mentioned on the generally positive character of the corrections. Obviously, to be consistent an equal positive correction is introduced at the $(1 \rightarrow 2)$ distance value. The $(1 \rightarrow 2)$ corrected value is considered in the calculation of the node position if, on spite of the correction of its distance, still is the shortest distance of the table. If not the following shortest distance of the table is considered instead of the mentioned $(1 \rightarrow 2)$ value and the procedure repeated.

After solving the equations (1) and (2) for $\vec{x}$ and $\vec{y}$ values, the position of the node is defined and then the distances from the node to the other points of the table are calculated. These values replace then the distances of 1 and 2 to the remainder elements. Then we continue looking for the following shortest distance in the table and we repeat the same routine up to the end of the data. In the very unusual situation of detecting multiple pairs of data with equal minimal pairwise distances, should be necessary to go ahead in more than one direction up to the point where the following successive results define the optimal order or sequence.

The Example

The example chosen as probe of the method corresponds to the evolutionary distance matrix for various archebacterial 16*S* rRNAs taken from the article of: Laurie Achenbach-Richter, Karl O. Stetter, and Carl R. Woese, Nature **327,** 348–349 (1987) [20]. The table of distances is the following:

Table I.

Species	Evolutionary distances								
	1	2	3	4	5	6	7	8	9
1. *Archeoglobus* VC-16	–								
2. *Tc. celer*	16.2	–							
3. *Mc. vannielli*	22.7	21.0	–						
4. *M. formicicum*	23.2	21.5	21.8	–					
5. *Ms. hungatei*	24.2	26.3	28.8	24.6	–				
6. *H. volcanii*	29.3	28.1	30.1	27.3	25.2	–			
7. *P. occultum*	22.2	18.4	26.2	25.6	31.4	33.0	–		
8. *S. solfataricus*	26.9	24.9	29.0	29.6	34.4	36.3	12.3	–	
9. *T. tenax*	24.7	20.8	29.9	29.0	33.5	34.1	12.1	17.4	–
10. *Tt. maritima*	38.3	36.1	45.0	44.1	46.8	48.4	36.2	42.1	38.5

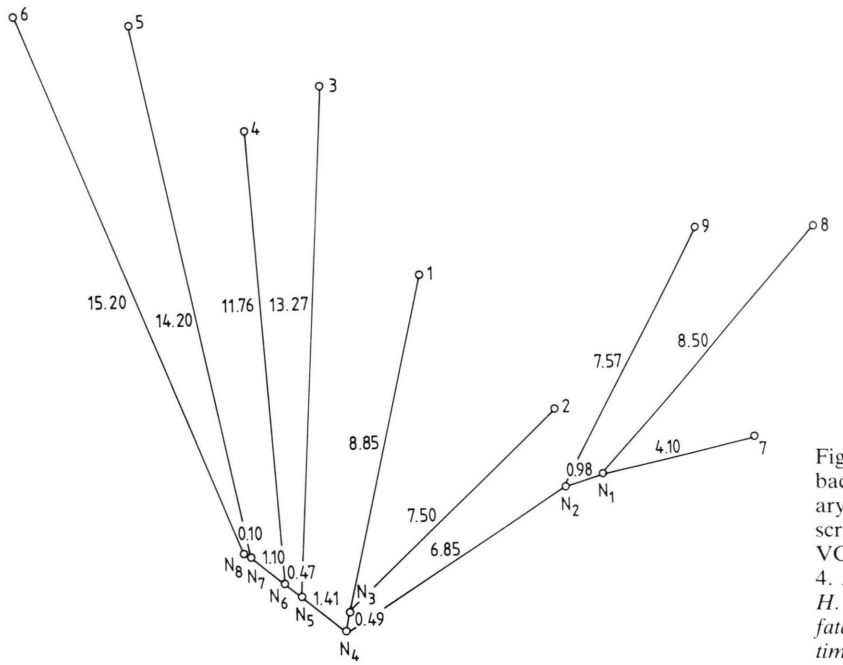


Fig. 1. Phylogenetic tree for archaebacteria derived from the evolutionary distances of Table I by the described algorithm. 1. *Archaeoglobus* VC-16; 2. *Tc. celer;* 3. *Mc. vannielli;* 4. *M. formicicum;* 5. *Ms. hungatei;* 6. *H. volcanni;* 7. *P. occultum;* 8. *S. solfataricus;* 9. *T. tenax* and 10. *Tt. maritima.*

As shown in the table the nearest neighbours are the species 7 and 9, but this entrance is rapidly discarded because of the magnitude of the necessary corrections performed to advance from this entrance, so the following smallest distance 12.3, joining the 7 and 8 elements is selected to start the tree construction. The average difference between every element of the table and the 7 and 8 species is $\bar{\blacktriangle} = 4.4$ and the particular difference that approaches more to this value is the $(1 \rightarrow 8)$ distance minus the $(1 \rightarrow 7)$ distance equal to: 4.7. Then as the difference between $4.7 - 4.4 = 0.3$ this amount is added to the $(1 \rightarrow 7)$ distance which is now of 22.5 instead of the former 22.2. The same value 0.3 is also added to the $(7 \rightarrow 8)$ distance which is now equal to 12.6, still the lower figure of the table. After that we proceed to calculate the values of the $\vec{x}$ and $\vec{y}$ segments defining the first node, using the equations:

(1) Where $\vec{x} + \vec{y} = 12.6$ and

(2) $\vec{y} - \vec{x} = 4.40$, solving for $\vec{x}$ and $\vec{y}$

$$\vec{x} = (7 \rightarrow N_1) = 4.1$$

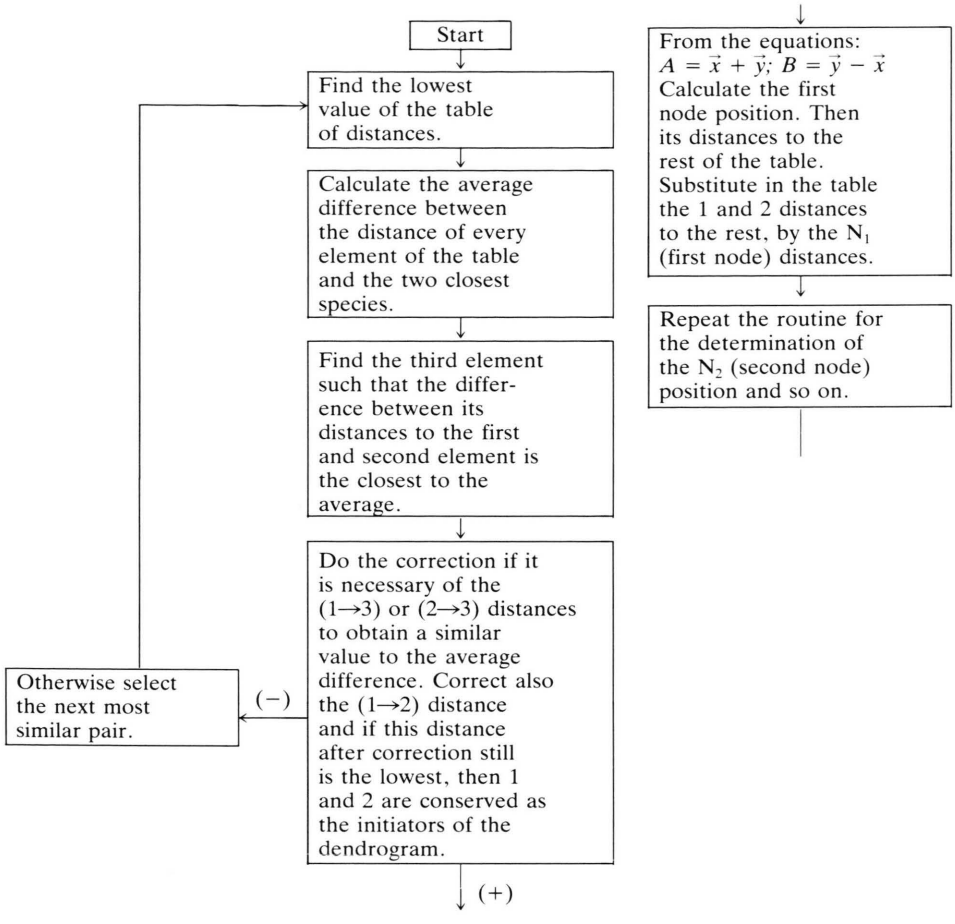
$$\vec{y} = (8 \rightarrow N_1) = 8.5.$$

After that we calculate all the distances mediating between N_1 (the first node) and the rest of elements of the table. These figure replace in the table the distances between the elements 7 and 8 to the rest. And the routine is repeated looking for the following shortest distance joining these two elements and the N_2 node, in this case between N_1 and element 9 (distance equal 8.45).

The dendrogram obtained following this procedure for the group of archaebacterial 16*S* rRNAs is similar to that communicated by Achenbach-Richter *et al.* [20], moreover the table of data reconstructed from the tree distances, corresponds fairly well with the data of the original table. So, the standard deviation, in percentage, of the original data and the reconstructed from the tree, is of 4.4%, which agrees with the deviation of the data obtained from the dendrogram displayed in the publication of Achenbach-Richter *et al.*, when compared with the original table. On the other hand, for a different comparison, we constructed using the data of the original table, a hierarchically ordered table and then we compare with that table the data: of the original table, the data of the Achenbach-Richter *et al.* dendrogram and the data of our calculated tree. The standard deviation in percentage of this comparison were respectively: 15.40%, 13.40% and 8.80%.

We analyzed also a variation of the algorithm before described, where at first, all the data of the table were corrected in accordance with the average difference distance of each couple of elements to the rest

Flow Chart of the Algorithm:



of the elements of the table. But, this variation did not give better results than the method before described of successive corrections.

Discussion

The algorithm above described facilitates the rapid finding at every step of the dendrogram of the appropriate couple of most nearby species and moreover permits the introduction of corrections to table data, sometimes disconnected.

The results or distances obtained in the dendrogram by the application of the algorithm agree fairly well with the original data, but more than that, they correlate better with data of a re-ordered table following a hierarchic distribution of figures which is probably a better parameter to appreciate the degree of certainty of the evolutionary trees found, than the original data. In the algorithm is relevant that at the first steps the average value of the difference or distances between the two selected neighbours and the rest of points, is calculated considering a high number of elements. That gives an appreciable degree of certainty to the eventual corrections performed. On the other hand, at the end of the procedure, the average difference is calculated with only few data available, but fortunately, as at every step, corrections have been introduced, then the universe of points disposable at the time, if scanty, is nevertheless enough reliable and accurate.

Acknowledgements

This work was partially supported by: Departamento Técnico de Investigación.

[1] W. Henning, Annu. Rev. Entomol. **10,** 97 (1965).
[2] L. Cavalli-Sforza and A. F. W. Edwards, Evolution **21,** 550 (1967).
[3] W. M. Fitch and E. Margoliash, Science **155,** 279 (1967).
[4] M. O. Dayhoff and R. V. Eck, Atlas of Protein Sequence and Structure, p. 7, National Biomedical Research Foundation, Silver Spring 1968.
[5] T. H. Jukes and C. H. Cantor, Mammalian Protein Metabolism (H. N. Munro, ed.), pp. 21–23, Academic Press, New York 1969.
[6] J. S. Farris, Amer. Nat. **106,** 645 (1972).
[7] G. W. Moore, M. Goodman, C. Callahan, R. Holmquist, and H. Moise, J. Mol. Biol. **105,** 15 (1976).
[8] E. M. Prager and A. C. Wilson, J. Mol. Evol. **11,** 129 (1978).
[9] L. C. Klotz, N. Komar, R. L. Blanken, and R. M. Mitchell, Proc. Natl. Acad. Sci. **76,** 4516 (1979).
[10] H. Holmquist and D. Pearl, J. Mol. Evol. **16,** 211 (1980).
[11] W. H. Li, Proc. Natl. Acad. Sci. **78,** 1085 (1981).
[12] H. J. Wagner, J. Theor. Biol. **91,** 621 (1981).
[13] F. Tajima and M. Nei, Mol. Biol. Evol. **1,** 269 (1984).
[14] D. Penny and M. D. Hendy, Syst. Zool. **34,** 75 (1985).
[15] D. F. Feng and R. F. Doolittle, J. Mol. Evol. **25,** 351 (1987).
[16] M. Nei, Molecular Evolutionary Genetics, p. 287, Columbia University Press, Columbia 1987.
[17] P. H. A. Sneath and R. R. Sokal, Numerical Taxonomy, p. 121, W. H. Freeman and Company, San Francisco 1973.
[18] J. Tohá, M. A. Soto, and M. Pieber, Z. Naturforsch. **34c,** 478 (1979).
[19] J. Tohá, M. A. Soto, and M. Pieber, Z. Naturforsch. **34c,** 1269 (1979).
[20] L. Achenbach-Richter, K. O. Stetter, and C. R. Woese, Nature **327,** 348 (1987).