On the Construction of a Phylogenetic Tree

José Tohá, María Angélica Soto, and María Pieber

Departamento de Física, Laboratario de Biofísica, Facultad de Ciencias Físicas y Matemáticas,
Universidad de Chile, Casilla 5487, Santiago, Chile

Z. Naturforsch. **34 c**, 478 – 480 (1979) ; received January 19, 1979

Phylogenetic Dendrogram, Molecular Evolution

An algorithm for phylogenetic trees' construction is analyzed.

During the past years, a great deal of information on protein and nucleic acid sequence data has been obtained.

These data have been very useful to throw some light on the ancestral evolution of the species.

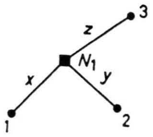
As it is well known, phylogenetic trees have been constructed considering that mutations have occurred at random and periodically.

In order to determine a node in a dendrogram, the two closest species are selected. This node is now calculated as an average of the values obtained from successive comparisons of these two species with each one of the rest of the species [1].

In each comparison a system of equations of three dependent variables is solved; thereby obtaining a unique solution.

For instance for Node 1

$$\begin{aligned}\text{Distance } 1 &\to 2 = x + y\\ \text{Distance } 1 &\to 3 = x + z\\ \text{Distance } 2 &\to 3 = y + z\,.\end{aligned}$$



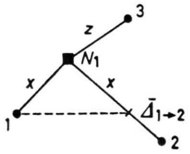
However, the observed differences between two species do not represent necessarily the real evolutionary mutation distance. As can be seen from reference [2] species that appear very different in respect to others, appear very close to each other. In an effort to search for the most probable nearest neighbour and to avoid errors in the construction of a phylogenetic tree, a simple algorithm is here described.

Replacing in system 1 in Fig. 1, the variable y by $x + \bar{\Delta} 1 \to 2$, the following is obtained:

$$\begin{aligned}\text{Distance } 1 &\to 2 = 2\,x + \bar{\Delta}_{1\to 2}\\ \text{Distance } 1 &\to 3 = x + z\\ \text{Distance } 2 &\to 3 = x + \bar{\Delta}_{1\to 2} + z\end{aligned}$$

where $\bar{\Delta}_{1\to 2}$ is the mean value of the differences that exist between elements one and two with respect to node 1. This mean value has been evaluated considering all distances of 1 and 2 in respect to the other elements of the system.

$$\vec{\Delta}_{1\to 2} = \quad (\text{Distance } z \to 3) - (\text{Distance } 1 \to 3)\,.$$



As can be deduced from this last equation, if the magnitude of the first member of the equation is different from $\bar{\Delta}_{1\to 2}$, an incoherence is evident.

For instance if

$$\bar{\Delta}\, 1 \to 2 < (\text{Distance } 2 \to 3) - (\text{Distance } 1 \to 3)$$

this may imply that

1) Distance 2 → 3 is greater than the real value
or 2) Distance 1 → 3 is smaller than the real value.

Simultaneously if

$\bar{\Delta}\, 2 \to 3$ is $>$ (Distance 1 → 3) – (Distance 1 → 2)
this may imply that

Distance 1 → 3 is smaller than the real value
or Distance 1 → 2 is greater than the real value.

Since

$$\bar{\Delta}\, 1 - 3 = (\text{Distance } 2 \to 3) - (\text{Distance } 1 \to 2)$$

then distance 1 → 3 may be smaller than its real value, and should be corrected in order to establish the first two closest values, which allow to start the dendrogram.

* This work was partially supported by O. E. A.

Reprint requests to Dr. J. Tohá.

0341-0382 / 79 / 0500-0478 $ 01.00/0

Table. Dissimilarity values for each indicated binary comparison of 16S rRNA *.

Organism	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0.000												
2	0.340												
3	0.400	0.400											
4	0.500	0.520	0.510										
5	0.470	0.510	0.490	0.400									
6	0.480	0.510	0.490	0.460	0.400								
7	0.750	0.730	0.750	0.740	0.770	0.750							
8	0.740	0.720	0.740	0.720	0.730	0.710	0.410						
9	0.800	0.760	0.790	0.770	0.770	0.780	0.490	0.480					
10	0.710	0.740	0.760	0.760	0.740	0.750	0.670	0.590	0.660				
11	0.920	0.920	0.890	0.910	0.910	0.900	0.950	0.940	0.930	0.900			
12	0.900	0.900	0.860	0.890	0.890	0.880	0.920	0.900	0.900	0.920	0.730		
13	0.900	0.900	0.900	0.900	0.900	0.890	0.920	0.910	0.920	0.890	0.760	0.740	0.000

* Taken from reference [2].

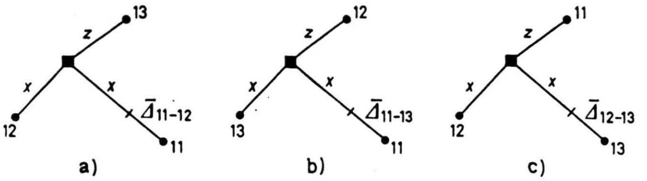
Organism

1	*M. arbophilicum*	6	*M. thermoautotrophicum*	10	*Methanoscarcina barkeri*
2	*M. ruminantium* PS	7	Cariaco isolate J. R-1	11	*Enteric-vibro sp.*
3	*M. ruminantium* M-1	8	Black Sea isolate J. R-1	12	*Bacillus sp.*
4	*M. formicicum*	9	*Methanospirillum hungatii*	13	Blue green sp.
5	M. Sp. M. o. H.				

From Table using dissimilarity instead of similarity values (1-Sab) it can be observed that microorganism Nr. 11, 12 and 13 appear more distant with respect to the other species and the closest to the initiation point of the tree [2].

According to data of this table the microorganisms Nr. 11 and 12 are the more related of this group, being its difference in the base composition of 16 ribosomal RNA of 73%. Nevertheless Nr. 11 and 12 are more different than 12 and 13 respect to the rest of the species.

Using the above described algorithm to obtain the first Node from this group [11 – 13] we have three possibilities:


13
z
x
x
12
$\bar{\Delta}_{11\text{-}12}$
11
a)
12
z
x
x
13
$\bar{\Delta}_{11\text{-}13}$
11
b)
11
z
x
x
12
$\bar{\Delta}_{12\text{-}13}$
13
c)

In the first case (a):

$$\bar{\Delta}_{11-12} = 0.022$$
$$\text{Distance } 11 \rightarrow 12 = 0.730 = 2x + 0.022$$
$$\text{Distance } 11 \rightarrow 13 = 0.760 = x + 0.022 + z$$
$$\text{Distance } 12 \rightarrow 13 = 0.740 = x + z$$
$$\text{Distance } (11 \rightarrow 13) - (12 - 13) = 0.760 - 0.740 = 0.020 \approx \bar{\Delta}_{11-12} = 0.022\,.$$

Case (b):

$$\bar{\Delta}_{11\rightarrow 13} = 0.012$$
$$\text{Distance } 11 \rightarrow 13 = 0.76 = 2x + 0.012$$
$$\text{Distance } 11 \rightarrow 12 = 0.73 = x + 0.012 + z$$
$$\text{Distance } 12 \rightarrow 13 = 0.74 = x + z$$
$$\text{Distance } (11 \rightarrow 12) - (12 \rightarrow 13) = -0.010 < \bar{\Delta}_{11\rightarrow 13} = 0.012\,.$$

In the case (c):

$$\bar{\Delta}_{13-12} = 0.009$$
$$\text{Distance } 12 \rightarrow 13 = 0.74 = 2x + 0.009$$
$$\text{Distance } 13 \rightarrow 11 = 0.76 = x + 0.009 + z$$
$$\text{Distance } 11 \rightarrow 12 = 0.73 = x + z$$
$$\text{Distance } (13 \rightarrow 11) - (11 \rightarrow 12) = 0.030 > \bar{\Delta}_{13\rightarrow 12} = 0.009\,.$$

According to case a), the observed distances 11 → 13 and 12 → 13 agree with the expected values.

In case b) on the contrary, distance 11 → 12 may be smaller, or distance 12 → 13 larger than the real value.

Similarly in case c), distance 13 → 11 could be larger, or distance 11 → 12 smaller than the real value.

However, analyzing the 3 cases, the distances 11 → 13 and 12 → 13 of case a) correspond to the expected value, indicating that distance 11 → 12 is diminished ≈ 0.020, a value representing $\bar{\Delta}_{11\to 12}$ allowing to improve the differences:

$$(11 \to 12) - (12 \to 13)$$

and $(13 \to 11) - (11 \to 12)$.

In conclusion, the most probable value for the distance $11 \to 12$ would be 0.750 instead of 0.730, implying that microorganisms 12 and 13 would be placed at the bottom of the tree.

[1] W. M. Fitch and E. Margoliash, Science **155**, 279 (1967).

[2] G. E. Fox, L. J. Magrum, W. E. Balch, R. S. Wolfe, and C. R. Woese, Proc. Nat. Acad. Sci. USA **74**, 4537 (1977).