

II. On the Construction of a Phylogenetic Tree

J. Tohá, M. A. Soto, and M. Pieber

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

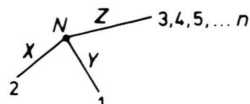
Z. Naturforsch. **34 c**, 1269–1271 (1979); received September 3, 1979

Phylogenetic Dendrograms, Molecular Evolution

For the construction of a phylogenetic tree an algorithm is described. This, allows the correction of the original data and the proper selection, in each step of the process, of the nearest neighbours of a common ancestor.

Assuming that differences at the molecular level can be correlated with the macroscopic dissimilarities, phylogenetic trees are built by utilizing sequences data of nucleic acids or homologous proteins of different species. However it is important to point out that the observed molecular changes do not represent necessarily the evolutionary trajectory in the differentiation process of the species. This is because: 1) repeated punctual mutations at the same molecular position are oversight; 2) the degenerate nature of genetic code will produce changes at the base level not relevant for the aminoacid translation; 3) technical analytic deficiencies and others. On the other hand the present absence of the possible common ancestors of neighbour species; and the non homogeneous distribution of the mutation during the evolutionary time space or along the compared molecules, add obstacles to the construction of a reliable dendrogram.

To build up a phylogenetic tree following the traditional methods [1], it is necessary in each step to know unambiguously which are the pair of nearest neighbours joined through a common ancestor; this is then properly located, averaging the values obtained through successive comparisons of the two nearest neighbours with each one of the rest of the elements (species) considered. In every comparison, a system of three independent equations is established permitting the solution of the three unknown distances there involved, as shown:

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


In a previous work, we described a method which allows the correction of the observed phylogenetic distances (table data), the corrected values are then used in the construction of the dendrogram. This algorithm is only valid if the three considered species are indeed nearly related.

In this communication we describe a modification, which generalizes the above mentioned algorithm, allowing to correct and select properly in each trial, the real two nearest neighbours of the data table, independent of the distances that mediate between species or nodes.

In this algorithm, the average value of the difference between the distances of the two nearest species and the rest of the elements, is determined from the table ($\bar{\Delta}$). Then, this difference is compared with that obtained from these two nearest elements (1 and 2) and the third involved (3) (see Figure).

For instance:

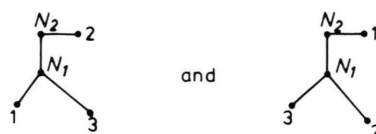
$$\text{Distance } 2 - 3 = x + z$$

$$\text{Distance } 1 - 3 = x + \Delta_{1-2} + z$$

$$(1 - 3) - (2 - 3) = \Delta_{1-2}$$

Δ_{1-2} is compared with $\bar{\Delta}_{1-2}$. If both values coincide, then the distances (2 - 3) and (1 - 3) are consistent with the rest of the table.

If this comparison is repeated for the following arrangements (cyclic permutation of 1, 2 and 3)



we can evaluate the degree of coincidence between the difference $(1 - 2) - (2 - 3)$ and the average value $\bar{\Delta}_{1-3}$ and between the difference $(1 - 3) - (1 - 2)$ and the average value $\bar{\Delta}_{2-3}$.

Reprint requests to Dr. J. C. Tohá.
0341-0382/79/1200-1269 \$ 01.00/0



Dieses Werk wurde im Jahr 2013 vom Verlag Zeitschrift für Naturforschung in Zusammenarbeit mit der Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V. digitalisiert und unter folgender Lizenz veröffentlicht: Creative Commons Namensnennung-Keine Bearbeitung 3.0 Deutschland Lizenz.

Zum 01.01.2015 ist eine Anpassung der Lizenzbedingungen (Entfall der Creative Commons Lizenzbedingung „Keine Bearbeitung“) beabsichtigt, um eine Nachnutzung auch im Rahmen zukünftiger wissenschaftlicher Nutzungsformen zu ermöglichen.

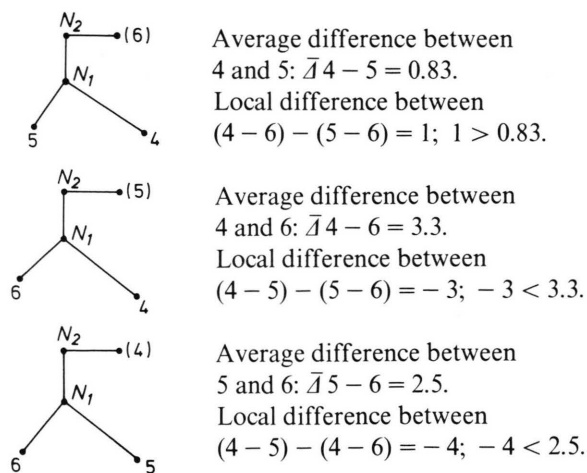
This work has been digitalized and published in 2013 by Verlag Zeitschrift für Naturforschung in cooperation with the Max Planck Society for the Advancement of Science under a Creative Commons Attribution-NoDerivs 3.0 Germany License.

On 01.01.2015 it is planned to change the License Conditions (the removal of the Creative Commons License condition “no derivative works”). This is to allow reuse in the area of future scientific usage.

The most convenient equation is selected according to: 1) The correction of the values from table has to be positive (because the values from Table underestimate in general the real number of accumulated mutations). 2) The correlated value has to be the smallest permitted. 3) After correction, in the arrangement selected, the distance that mediates between the two elements joined by the first ancestor has to be the smallest one.

Only one of the above mentioned arrangements fulfil these 3 conditions, in that, the corrections of discordant distances are performed and then the common node (ancestor) is located between the two real nearest elements following the traditional algorithm [1].

Example: The construction of the phylogenetic tree of cytochrome C, from a group of mammals (Table, lower half) is given, as an example of the application of the method. In this group, from Table I, the nearest neighbour elements are apparently species 4 and 5 and the third nearest neighbour is the species 6. Then, at the initiation of the dendrogram we compare the following arrangements, where the distances 4 – 5, 4 – 6 and 5 – 6 are respectively: 1, 5 and 4.



In the comparison, the local difference of distances that approaches better to the average \bar{A} is that of the first arrangement. There, after correction of the distance 5 – 6 to the value 4.2, all the 3 conditions are fulfilled and in consequence 4 and 5 are defined as the nearest neighbours. After correction, the legs of the first node N_1 are calculated following the traditional algorithm. Then, 4 and 5 are eliminated and instead of them, the distances of N_1 with the rest of

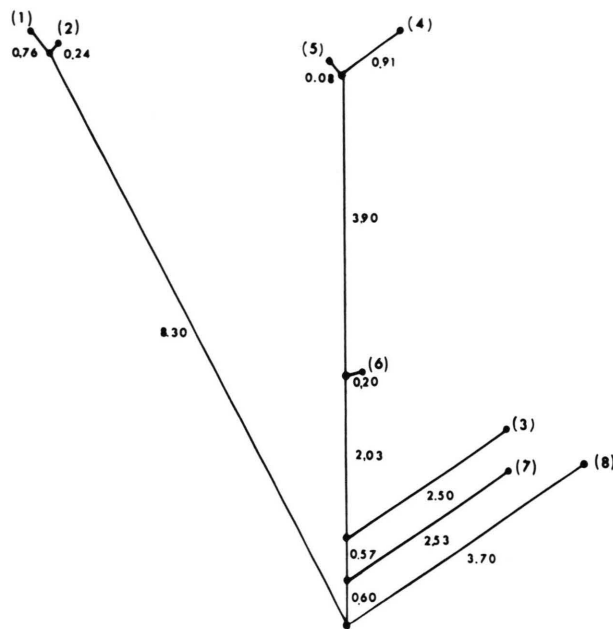


Fig. 1. Phylogeny of cytochrome C from a group of mammals. The dendrogram was constructed after correction of the original mutational Table, following the above described algorithm.

Table I. Mutational distance of cytochrome C from a group of mammals [1]. Upper right half of the table: original data. Lower left half of the table: reconstructed distances found by summing the leg lengths in the dendrogram (Fig. 1).

Species	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Man	(1)	0	1	13	17	16	13	12
Monkey (<i>Macacus</i> <i>Mulatta</i>)	(2)	1.0	0	12	16	15	12	11
Dog	(3)	12.7	12.2	0	10	8	4	6
Horse	(4)	17.1	16.6	9.3	0	1	5	11
Donkey	(5)	16.2	15.7	8.5	1.0	0	4	10
Pig	(6)	12.5	11.9	4.7	5.0	4.2	0	6
Rabbit	(7)	12.2	11.7	5.6	9.9	9.1	5.3	0
Kangaroo (<i>Canopus</i> <i>canguru</i>)	(8)	12.8	12.2	7.4	11.7	10.9	7.1	6.8

the elements are incorporated to the Table. After that, with a new set of apparent three nearest neighbours the possible arrangements are analyzed, following again the same above described process and so on.

The dendrogram thus obtained is shown in Fig. 1 and the respective table reconstructed from the distances mediating between the species, in the phylogenetic tree, appears in the lower half of the Table. All the distances that we found are real and positive values and the standard deviation in percentage obtained by comparison of data from the reconstructed

Table and the original one is 5.6, figure that shows the proximity of the data from the original and reconstructed Table.

Acknowledgement

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

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

[2] J. Tohá, M. A. Soto, and M. Pieber, *Z. Naturforsch.* **34 c**, 478 (1979).