PHYLOGENETIC TREE OF tRNAs USING A SIMPLE ALGORITHM

M. ANGELICA SOTO and JOSE TOHA

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

(Received 6 November, 1989)

Abstract. A simple method for phylogenetic tree construction is described. In this method each node is calculated considering the distance between the elements and the difference between these elements and an average element, allowing the selection of the most probable node.

Two examples of tRNA phylogenies (*E. coli* set and Phe family) are analyzed, giving both reliable trees. Data from these dendrograms give support to the idea of an early cloverleaf arising.

1. Introduction

An algorithm for the construction of phylogenetic trees is described.

As examples, the phylogenies of two sets of tRNAs have been constructed because of the great interest that represents this structure in the origin of the genetic code. On this problem, many advances have been reported about, for instance, origin and evolution of tRNA molecule, the primordial molecules involved in translation processes, the probable co-evolution with the amino acid set and so on (e.g. Cedergren *et al.*, 1972; Fitch and Upper, 1987; Nicoghosian *et al.*, 1987; Staves *et al.*, 1987; Eigen *et al.*, 1989). On the search of a more reliable phylogeny based on the greater number of recently published tRNA sequences, we have studied the *E. coli* tRNAs and the family of phenylalanine tRNAs (Sprinzl *et al.*, 1985).

In the method here described it is calculated an average tRNA molecule which is used as reference in the evaluation of the nodes. Moreover these nodes are accepted at each step only if there is agreement with the original table of distances, which allows the rapid and correct construction of the dendrogram.

As shown further, in the *E. coli* set the average tRNA resembles the origin of the tree, corresponding to a possible common ancestor.

In a complementary study, we have also analyzed the Phe tRNA family, considering in this case different species. In this example, the average tRNA can be considered as a possible origin only if mitochondrial and chloroplast systems are not included.

2. Method

This method uses two essential points for the construction of the dendrogram, which differentiates it from other described algorithms. One is the use of an averasge molecule representing the set of data of the table. The other implies that each node be fixed between the two nearest elements only if there is no contradiction with the other elements of the table.

Origins of Life and Evolution of the Biosphere **20**: 161–166, 1990.

CONSTRUCTION OF THE DENDROGRAM

– From the set of data, an average molecule is constructed. In the case of tRNAs, the average is constructed selecting for each position the most representative nucleotide of the set. This average molecule, which is a reference molecule reduces the eventual contradictions of the set of data, favoring a more representative calculation.

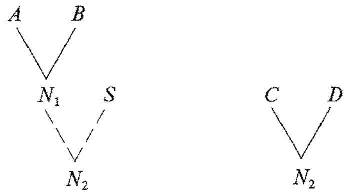
– The nearest species A and B, according to the distance matrix, are identified; let $x + y$ be the distance between A and B through the node N_1.

– The distances between A and the average and B and the average are evaluated in terms of the difference of their nucleotide sequences. Let $x - y$ be the difference between these two distances.

– From this pair of equations: $x{+}y$ and $x{-}y$ the values of x, y and the position of the node are determined.

– In the following step two options are possible:

(i) Consider the following smallest element of the table corresponding to the distance between species C and D to determine the node N_2.



(ii) The smallest distance may correspond to the value $N - S$ (S = any species of the table). However, if the distance S–A or S–B through the node N_2 is greater than the distance C–D, this option is rejected and the node N_2 is located between the species C and D. In this way we continue up to the end of the tree, trying at each step to remain as close as possible to the original table of distances, which in this method is the valid reference.

The examples considered in this communication correspond to:

– The tRNAs of 19 amino acids (Pro is not available) of *E. coli* 220 species (Sprinzl *et al.*, 1985).

– The Phe tRNAs of the following species: *Neurospora crassa*, *Lupinus luteus*, Barley, Wheat, Human placenta, Mouse liver, *Bombyx mori, Drosophila melanogaster*, Yeast, *Euglena gracilis*; Choroplasts: *Euglena gracilis, Spinacia oleracea*, Mitochondrias: Rat Morris Hepatoma, *Saccharomyces cerevisae*, Yeast (Sprinzl *et al.*, 1985).

To construct the distance matrix, the alignement of sequences used by Sprinzl *et al.* (1985) was employed, considering the deletions and insertions.

With the method above described were constructed the dendrograms corresponding to the table of distances of the whole *E. coli* tRNA sequences and those of the anticodon and aminoacyl sections separately.

In the case of the Phe set, the phylogenetic trees of the whole set was constructed and also the independent trees of the chloroplasts and mitrochondrias as well as of the remainder species.

In all cases the dendrograms were compared with the original table of distances evaluating their respective dispersions.

After the construction of the tree, the substitution of the tRNA located at the origin of the tree by the average is also statistically evaluated. The average approaches the origin of the tree if the rate of changes of the elements is comparable.

3. Results

E. coli tRNAs

Figure 1 displays the tree corresponding to the set of *E. coli* tRNAs, which agrees well with the table of distances. (Standard deviation = 10.9%).

From this tree it follows that:

- Chemically similar amino acids become near neighbours.

- Two patterns can be distinguished. That corresponding to the tRNAs of Leu, Ser and Tyr which display the greatest distances to the other tRNAs. These tRNAs evolved with the development of a big extra arm in their structure. In the other group, with more homogeneity in distances, can be individualized:

- Those tRNAs having the smallest distances between them, chemical similarity and similar codons:

Gly-Trp (GGG-UGG), Ile-Val (AUC-GUC), Thr-Ala (ACC-GCC), Asp-Asn-His (GAC-AAC-CAC), Cys-Phe (UGC-UUC).

- On the other hand, the tRNAs of Gln-Glu appear separate by small distances between their nodes.

In this tree, the average tRNA molecule approaches the origin of the group (excluding

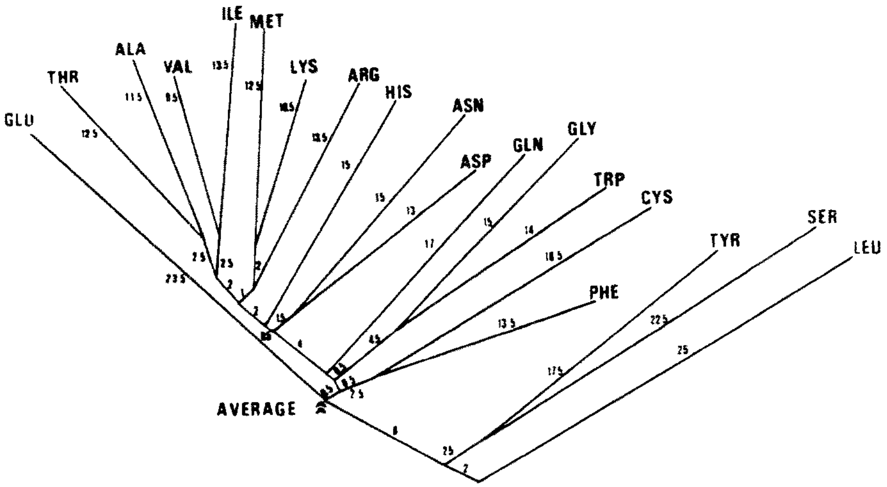


Fig. 1. Phylogenetic tree of *E. coli* tRNAs. (Dispersion : 10.9%).

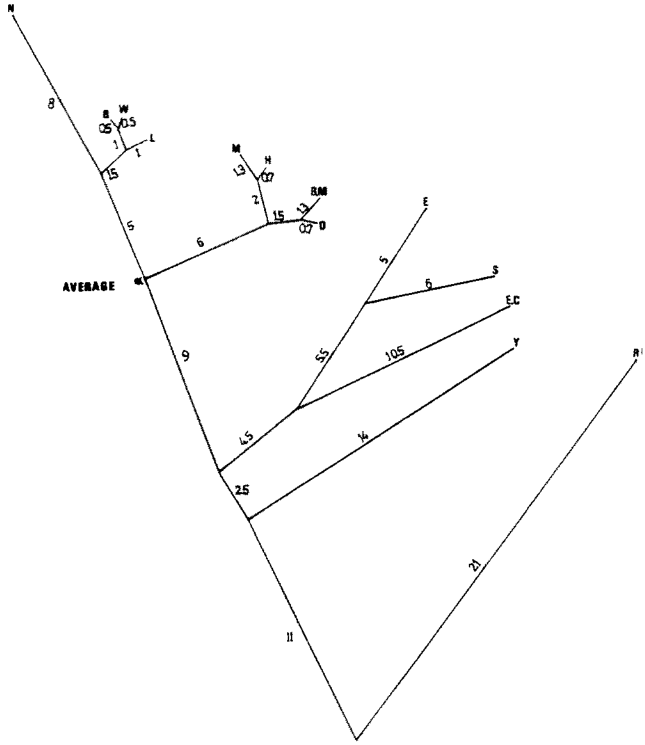


Fig. 2. Phylogenetic tree of Phe tRNAs of the following species: N = Neurospora *crassa*, L = *Lupinus luteus*, B = Barley, W = Wheat, M = Mouse liver, H = Human placenta, B M = *Bombyx mori*, D = *Drosophila melanogaster*, E C = *Escherichia coli*. Chloroplasts: E = *Euglena gracilis*, S = *Spinacia oleracea*. Mitochondrias: R = Rat Morris Hepatoma, Y = Yeast. Standard deviation of the tree: 6.6%.

Leu, Ser and Tyr) with a standard deviation of 28.7%.

Moreover, dendrograms corresponding to the anticodon and aminoacyl sections of the same set of tRNAs were constructed, displaying 16 and 17% of dispersion respectively. These greater dispersions respect to the tree of the whole tRNA molecule could probably be explained by the less homogeneity of these tables of distances, since it is not likely an independent evolution of every section.

Phe tRNAs

The phylogenetic tree corresponding to the Phe species (Figure 2) gave a standard deviation of 6.6% respect to the table of distances. In this tree, the average tRNA is located in the node separating the mitochondrial and chloroplast systems from the other species. This is not surprising because mitochondria use a simplified translation system with a reduced number of tRNAs (Jukes, 1983) and the same feature has been detected in chloroplasts (Ozeki *et al.*, 1987). The codes for mitochondria and chloroplasts have been derived from a eubacterial code (Jukes *et al.*, 1987).

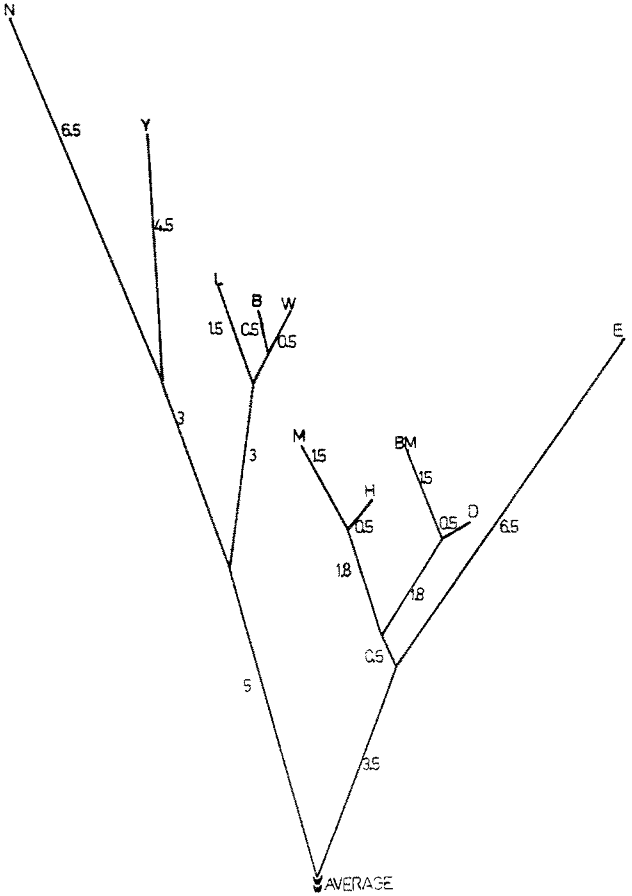


Fig. 3. Phylogenetic tree of Phe tRNAs of the following species: N = *Neurospora crassa*, Y = Yeast, L = *Lupinus luteus*, B = Barley, W = Wheat, M = Mouse liver, H = Human placenta, B M = *Bombyx mori*, D = *Drosophila melanogaster*, E = *Euglena gracilis*. Dispersion of the tree: 13.6%.

For these reasons we have constructed the separate dendrograms of chloroplasts and mitochondria and of the remainder species. The dendrogram of chloroplasts and mitochondria displays a 2.4% of dispersion respect to the experimental data and that of the remainder species a 13.6% of dispersion. This dendrogram is shown in Figure 3. The average tRNA of this set can be located at the origin of the tree, differing from this in a 19.6%.

4. Discussion

The algorithm here described is a rapid and efficient method to construct phylogenetic trees. In this method the average molecule, representing all the set of molecules under study can be used in the calculation of each node as a mean distance from the two vertiles involved in the node to the other vertices. This average can be located near the origin of the tree. In this case, it can be considered as a common origin. On the contrary, when the set of data is not homogeneous, there is not correspondance with the origin. Moreover, the nodes are determined only if there is a good agreement between the distance of the vertices joined by the node and the table of distances, being rejected those nodes which do not fulfill this requirement. This permits a reliable calculation at each step of the dendrogram construction.

In the two examples here analyzed, the *E. coli* and the Phe tRNAs, both average tRNAs are molecules satisfying the structural constraints of a tRNA molecule. In the *E. coli* set the average approaches the origin of the tree. As could be expected, the same situation is evidenced in the Phe dendrogram when mitochondria and chloroplast tRNAs are not considered (Figure 3), giving support to the idea of considering this structure as a primordial one.

In Figure 1, the *E. coli* tree is a bundle like (Eigen *et al.*, 1989) meaning a parallel evolution of every tRNA, which suggests an early origin of the cloverleaf structure as well as a probably early evolution of the tRNA aminoacyl synthetases. On the other hand, separate anticodon and aminoacyl trees have a less agreement with experimental data (16 and 17% respectively) indicating again the possible early formation of the tRNA molecule and its subsequent differentiation. Nevertheless, this configuration could be preceded by a simpler one involving probably the anticodon and aminoacyl arms since these regions present the greater rate of changes (mean value: 75 and 55% of changes respectively). Probably this precursor molecule evolved in accordance with the restrictions imposed by the extant tRNA pattern, to determine an efficient translation mechanism.

References

Cedergren, R. J., Cordeau, J. R. and Robillard, P.: 1972, *J. Theor. Biol.* **37**, 209–220.

Eigen, M., Lindemann, B. F., Tietze, M., Winkler-Oswatitsch, R., Dress, A. and Von Haeseler, A.: 1989, *Science* **244**, 673–679.

Fitch, W. M. and Upper, K. 1987, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 759–767.

Jukes, T. H.: 1983, *Adv. Space Res.* **3** (9), 107–111.

Jukes, T. H., Osawa, S., Muto, A. and Lehman, N.: 1987, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 769–776.

Nicoghosian, K., Bigras, M., Sankoff, D. and Cedergren, R.: 1987, *J. Mol. Evol.* **26**, 341–346.

Ozeki, H., Ohyama, K., Inokuchi, H., Fukusawa, H., Kohchi, T., Sano, T., Nakahigashi, K. and Umesono, K.: 1987, *Cold Spring Harbor Symp. Quant. Biool.* **52**, 791–804.

Sprinzl, M., Moll, J., Meissner, F. and Hartmann, T.: 1985, *Nucleic Acids Res.* **13**, r1–r49.

Staves, M. P., Bloch, D. P. and Lacey, J. C. Jr.: 1987, *Z. Naturforsch.* **42c**, 129–133.