Phyllogenic tree creation

Phylogenetic tree

A phylogenetic tree is a graphical representation of the kinship relationships between different taxonomic units that can be assumed to have a common ancestor. Family relationships are assessed here on the basis of morphological or genetic similarity. Instead of taxonomic units, individual biological species or even individual genes can appear in some trees.

The term tree is taken from graph theory, where it denotes an undirected connected acyclic graph. Vertices that are connected by edges to two or more other vertices are called *internal vertices*. The remaining vertices that are connected to only one other vertex are called *leaves*.

In the case of phylogenetic trees, each vertex represents a certain taxonomic unit, and an edge between two vertices indicates the relationship between the taxonomic units that these vertices represent. Depending on the type of phylogenetic tree, the length of an edge can indicate the time of evolution or the degree of similarity between the taxonomic units involved.

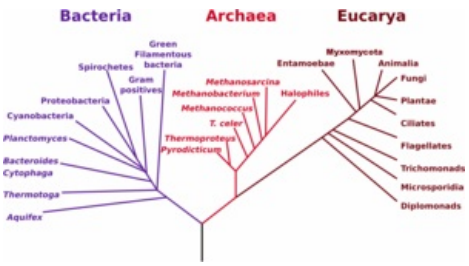
Types of phylogenetic trees

Unrooted phylogenetic tree

This type of tree depicts the relationships between taxonomic units without specifying their common ancestor.

Rooted phylogenetic tree

A rooted phylogenetic tree is a tree in which one of the internal vertices has been designated as the *root*. The edges of the tree thus acquired a natural orientation in the direction from the root to the leaves. The root represents the common ancestor of all taxonomic units represented by the tree. Each internal vertex represents the nearest ancestor of its descendants. However, internal vertices usually represent hypothetical taxonomic units that cannot currently be observed. In contrast, the leaves of a tree represent real taxonomic units.



Phylogenetic tree of life

It is possible to construct an unrooted tree from a rooted tree at any time simply by unmarking the root, the opposite procedure is only possible with additional information about the course of evolution.

Comparison of taxonomic units

The creation of phylogenetic trees is based on data on the similarity between individual taxonomic units. There are many ways to define this similarity. Recently, knowledge from the field of molecular biology has been widely used. We start from the base sequences in the genomes of individual biological species, or information about the relevant amino acid and protein products can also be used. Based on these data, it is possible to determine genetic distances between individual pairs of taxonomic units. The exact calculation of this distance first requires a suitable alignment of the compared DNA sequences - the so-called aligning. This is a computationally very difficult task (belongs to the class of NP-complete tasks), therefore a number of heuristic methods are used in practice, which are able to find at least a suboptimal solution in an acceptable time. For aligned sequences, it is possible to determine the distance, for example, based on the percentage of different bases between the sequences. More sophisticated methods attempt to estimate the number of mutations that are required to move from one sequence to another.

In addition to molecular biological data, morphological properties of the investigated taxonomic units can also be used. The calculation of distances in this case depends on the observed signs and the importance assigned to individual signs.

Methods of constructing phylogenetic trees

Distance methods

These methods are based on the distance matrix, which indicates the mutual distances between all pairs of taxonomic units for which we construct a phylogenetic tree. The genetic distance is used as the distance in this case.

UPGMA (Unweighted Pair Group Method with Arithmetic mean)

UPGMA, simply called Clustering Analysis, is the simplest algorithmic method of phylogenetic tree construction. The procedure is as follows:

1. Find the smallest value in the distance matrix (corresponds to the pair of taxonomic units that are closest to each other).
2. Merge relevant taxonomic units into one group and calculate the distance of this new group to all other taxonomic units. The distance of the taxonomic unit T to this new group S is calculated as the arithmetic mean of the distances between the unit T and all members of the group S. The group S can further be considered as a hypothetical taxonomic unit.
3. If we have more than one taxonomic unit, repeat the procedure from step 1.

If we graphically represent the clustering process during the described algorithm, we obtain the desired phylogenetic tree. The hypothetical taxonomic unit that arose last is its root.

The method of least squares

In this case, we construct all possible phylogenetic trees and evaluate which one is the best. We can perform the evaluation according to the following regulation:

$$Q = \sum_{i=1}^{N} \sum_{j=1}^{N} (D_{i,j} - d_{i,j})^2,$$

where $d_{i,j}$ is the distance between vertices i and j in the evaluated phylogenetic tree and $d_{i,j}$ is the distance between the corresponding taxonomic units in the distance matrix.

This procedure requires the construction and evaluation of all possible phylogenetic trees, which is similar to aligning an NP-complete problem.

The method of minimal evolution

The procedure is the same as for the least squares method, but we compare individual trees according to the sum of the lengths of all branches.

Neighbor-joining

At the beginning, one star tree is created, where there is one internal vertex, and all solved taxonomic units are represented by leaves. This tree is gradually decomposed by clustering the nearest taxonomic units so that the total length of the tree is reduced as much as possible at each step.

Maximum parsimony

The maximum parsimony method tries to find a phylogenetic tree that requires the smallest possible number of evolutionary events that would have to occur if this tree corresponded to the course of evolution. In some cases, different weights are assigned to individual evolutionary events when evaluating trees, for example, if it is known that some nucleotides or amino acids mutate more easily or worse than others.

In the basic variant, this method again requires the construction of all possible phylogenetic trees and their subsequent evaluation. To make the search of the tree space more efficient, you can use, for example, the Branch and bound method, which is able to limit the search to only "promising" trees.

Maximum likelihood method

This is based on statistical methods and a posteriori probability. We are trying to estimate the probability that the statistical hypothesis represented by a particular phylogenetic tree is true for the data we have. For hypothesis H and data D, this probability can be calculated as follows:

$$P(H|D) = P(H).\frac{P(D|H)}{P(D)},$$

where P(D|H) is the probability that we observe the actual data D, given that the hypothesis H is true.

The method requires a substitution model, on the basis of which we determine the probability of individual evolutionary changes (mutations). A tree that needs more of these changes to explain the available phylogenetic data will have less credibility than a tree that gets by with fewer changes. In addition, we also note the length of the individual branches.

Links

Related articles

- Phylogenic systems

- Evolution

External links

- Phylogenic tree (English Wikipedia)

Resources

- Vladimír Hampl: Lecture on molecular taxonomy (http://web.natur.cuni.cz/~vlada/moltax/)
- Computational phylogenetics (English Wikipedie)