UPGMA is an arithmetic averaging unweighted group pair algorithm. It is used in numerical taxonomy to reflect the degree of similarity in the representation of groups. We assume that the rate of gene replacement is constant, with this algorithm, we can construct molecular system trees.

To facilitate analysis, we need to define a continuous sum-distance function first. We define that the distance between two taxa is proportional to the total length of the branches connecting the two taxa in the phylogenetic tree. In this case, the distance between taxon a and taxon b is + if the taxon and taxon are connected by two sides of length and which passing through an intermediate node .

The basic idea of UPGMA is to first select the pair of taxa (DNA sequence) with the smallest distance from the distance matrix, naming them and respectively, and then combine the two taxa to form a new taxon , which are regarded as their ancestors. Recalculate the distance from the new taxon u. Its calculation formula is:

The new taxon formed by each merge is actually a cluster, which is represented by an internal node. This node is the same distance from the node where **x** and **y** are located, and its value is equal to half of . The distance to other nodes is calculated according to the above formula. After each merge, we need to modify the distance matrix. The process is repeated until all taxa have been combined into one category.

The UPGMA process is as follows:

1. Initialization: make each taxonomic unit a class of its own. If there are taxonomic units, there are classes at the beginning. The size of each class is 1, and each class is represented by leaf nodes.

2. Execute the following cycle:

(1) Find two classes x and y with minimum distance in the distance matrix;

(2) A new cluster u was established as the merge of taxon x and taxon y.

(3) Create a new internal node in the tree, generate two new branches. Connecting and to , making be the parent node of and , and connect to each leaf of and . The length of the node is assigned as (The calculation and definition of GLD have been given in Model 1), so and are calculated;

(4) Calculate the distance between and other taxon according to formula:

Here, and respectively represent the number of classes in taxon and , and ij d is the distance between class in taxon and class in taxon .

(5) Update the distance matrix. Delete the row and column corresponding to x and y in the distance matrix, add new rows and columns for , and establish a new distance matrix;

Repeat the loop until there is only one class left

The UPGMA method often results in incorrect topological structures when the evolutionary rate of different lineages is very different or parallel evolution with homologous sequences is present. Moreover, when the state space of the evolutionary tree is large, the operability of the UPGMA method is extremely poor, so the use of this tree construction method is extremely limited. Trees obtained by the UPGMA method are usually rooted trees, as roots are easily obtained under the assumption of a constant rate of evolution. Trees constructed by other phylogenetic inference methods are usually rootless trees because it is difficult to identify the roots at different evolutionary rates.