# Analysis of distance matrices and construction of phylogenic trees

Pavel Perevedentsev[1], Mikhail Roytberg[2], Sergei Spirin[3]

[1]Higher School of Economics, Russian Federation
[2]Institute of Mathematical Problems of Biology RAS, Russian Federation, mroytberg@lpm.org.ru
[3]Moscow State University, Russian Federation

Quality of phylogenetic tree critically depends on the quality of distant matrix used by an algorithm constructing the tree [1]. Is it possible to reveal features of distance matrices allowing to choose the best one from the list of candidates or to improve the given matrix in way allowing improving the quality of resulting phylogenetic tree? As to our knowledge the question was not considered so far.

The approach we used is based on the well-known "4-leaves" feature of tree distances. Let A, B, C, D be arbitrary four leaves of a weighted unrooted tree and distance d(X, Y) between two leaves is the sum of weights of edges connecting X and Y. Consider three possible partitions of the 4-tuple {A, B, C, D} into two pairs and three corresponding sums d(A,B)+ d(C, D); d(A, C)+ d(B, D); d(A, D)+ d(B, C). Let SP0, SP1, SP2 be the sums ordered by increasing. The "4-leaves" equality claims that SP1=SP2 > SP0.

This allows one to formulate criteria of "tree-likelihood" of the distances between 4 leaves A, B, C, D given a distant matrix. Let SP0, SP1 and SP2 be the above sums. We have considered following quality factors: Q1 = SP2-SP1; Q2 = (SP2-SP1)/SP2; Q3 = (SP2-SP1)/)(SP1+SP2)/2); R1 = ((SP1+SP2)/2) -SP0; R2 = ((SP1+SP2)/2 -SP0)/SP0; R3= ((SP1+SP2)/2 -SP0)/((SP1+SP2+SP0)/3). The Q-factors reflect the needed coincidence between SP1 and SP2; the R-factors reflect the distance between nearest common predecessors of "paired" leaves.

To check predictive abilities of the factors we have performed computer experiments based on the set of bacterial genes, the set contains 103 families of aligned orthologous genes from 30 bacteria; the "true" phylogenetic tree for the bacteria was determined from the integral information about all genes. For each of 103 gene families we have constructed 3 distance matrices based on given multiple alignments: B-matrix (based on BLOSUM62 weight matrix), M-matrix (based on identical weight matrix) and J-matrix (based on Jones-Taylor-Thornton distance). Then starting from each of the 309 matrices we have constructed phylogenetic trees using neighbour-joining (NJ) algorithm. For each of the matrix and each 4-tuple of bacteria we have computed the following values: (1) values SP0, SP1, SP2 and the corresponding partition of the 4-tuple into pairs; (2) values Q1, Q2, Q3, R1, R2, R3; (3) partition of the 4-tuple into pairs corresponding to the “true” tree. We say that a 4-tuple is “good” if its partition based on the B-matrix of distances coincides with one based on the “true” tree. Otherwise the 4-tuple is called “bad”.

Our experiments show that factor R1 is most adequate to distinguish good and bad 4-tuples, its value is in correspondence with quality of a phylogenetic tree based on the matrix. This is demonstrated by Table 1. The table shows the data for 3 genes giving the worst, the best and medium quality of NJ-tree (the quality of an algorithmic tree is characterized by the number of common edges between it and the “true” tree, see column NJ in the Table 1). At the next step of our work we plan to design a novel algorithm constructing phylogenetic tree using preliminary analysis of distant matrices.

| NJ | NBad | MaxRBad | Good>0.3 | Good>0.1 | Bad>0.1 | %%Bad>0.1 | |
|---|---|---|---|---|---|---|---|
| PF00542 | 12 | 8476 | 28.00% | 1847 | 11333 | 1337 | 10.55% |
| PF01653 | 18 | 3595 | 19.00% | 4033 | 15326 | 171 | 1.10% |
| PF01765 | 24 | 1984 | 14.00% | 4230 | 16988 | 36 | 0.21% |

Table 1. NJ – see the text; NBad – number of bad 4-tuples; MaxRBad – maximal value of R1 factor for bad 4-tuples; Good>0.3 - number of good 4-tuples with R-factor > 0.3 (no bad 4-tuples has such value of the factor); Good>0.1 and Bad>0.1 – analogous data for good and bad 4-tuples with the cut-off 0.1; %%Bad>0.1 – percent of bad 4-tuples among all 4-tuples with R-factor > 0.1

1. The Performance of Neighbor-Joining Methods of Phylogenetic Reconstruction. Algorithmica 25:251–278.