Estimation of relative effectiveness of phylogenetic programs by machine learning

Mikhail Krivozubov[1], Florian Göbels[3], and Sergei Spirin[2]

[1]Faculty of Bioengineering and Bioinformatics and [2]Belozersky Institute of Physico-Chemical Biology, Moscow State University; [3]Technische Universität München

E-mail: sas@belozersky.msu.ru

Introduction. There are many methods for reconstruction of phylogeny from multiple protein sequence alignments. These methods often give different results, so there is a problem of choosing the appropriate method for a given data set. It is also important to estimate if the data is good enough to produce a valuable reconstruction with any method.

In this work we tried to create expert systems that predict: 1) if the data is good for phylogeny reconstruction by Fitch – Margolish (FM) method; 2) which method would work better: FM or UPGMA. The choice of methods is motivated as follows. Fitch – Margoliash method (the program *fitch* of PHYLIP package) is proved to be one of the best phylogeny reconstruction methods. UPGMA (option of the program *neighbor* of PHYLIP package) is the best method using molecular clock assumption. For a given alignment, FM is three times more likely to perform better than UPGMA [1].

Methods. ***Test sets: alignments and etalons.*** We used two species sets, first containing 27 Fungi species and second containing 25 Metazoa species. For each species set, we selected all protein domain families from the Pfam database [2] which include sequences for each species in the set. From each family, we extracted all possible orthologous series. The etalon tree for each species set is obtained as a consensus of trees reconstructed with all the series (for each species set, the consensus was proved to be independent on the applied phylogeny reconstruction program). For each species set, a set of 5000 alignments was created. Each alignment was obtained as follows: randomly choose an orthologous series and 15 species and align the 15 sequences with *muscle 3.6*. Distance matrices were calculated with the program *protdist* of PHYLIP package. Finally we reconstruct two phylogenetic trees with each alignment, using *fitch* (FM method) and *neighbor* (UPGMA method). The quality of the reconstructions was estimated by comparison with the etalon (i.e., the species tree) using three different measures of tree similarity.

From the entire set of alignments, two subsets were extracted, called “good” and “bad”. The “good” subset consists of those alignments, for which the similarity of FM tree with the etalon is higher than a threshold in each (of the three) similarity measure. The “bad” subset consists of the alignments, for

which the similarity of FM tree with the etalon is lower than a threshold in each similarity measure. As the thresholds, the medians of the measures on the Fungi set were used. For Fungi, we obtained 2016 “good” and 1904 “bad” alignments, for Metazoa, 415 “good” and 3877 “bad” alignments.

Also another two subsets were extracted: “FM better” and “UPGMA better”. Each subset consists of alignments for which the corresponding program produces a tree that is closer to etalon according to all three similarity measures. Only alignments that give a good enough result with at least one of the methods are included into the subsets. For Fungi, there are 1418 “FM better” and 565 “UPGMA better” alignments; for Metazoa, 1312 “FM better” and 400 “UPGMA better” alignments.

Machine learning. For machine learning, we used a number of features derived from alignments (such as alignment length), distance matrices (such as deviations from additivity and ultrametricity), and trees themselves (the sum of branch lengths). The most informative features were values derived by comparison of a tree with a distance matrix or an alignment, for instance the objective function of the FM method. We used a random forest classifier with 1000 trees to classify alignments into “good”/“bad” and into “FM better”/“UPGMA better”. We evaluated the classifier on both the Metazoan and the Fungi set as follows: i) each classifier with 10-fold cross-validation on each set; ii) using a holdout set analysis, which means that we trained the classifier on the Fungi data set and evaluated it on the Metazoan data set.

Results. Training on Fungi set and evaluating on Metazoa set, 94% alignments classified as “bad” are really “bad”, while only 30% alignments classified as “good” are really “good” (a random choice would give 90% and 10%, respectively). For 10-fold cross-validation on Fungi, the values are 69% and 68% for the prediction, comparing with 51% and 49% for a random choice. Training on Fungi and evaluating on Metazoa, 78% alignments classified as “FM better” are really “FM better” and only 41% classified as “UPGMA better” are really “UPGMA better” (a random choice would give 77% and 23%, respectively, and 10-fold cross-validation on Metazoa gives 78% and 54%, respectively).

The work is supported by a joint grant of Russian Foundation of Basic Research (grant no. 12-04-91334) and German Research Foundation (grant IRTG 1563/1).

1. M.S.Krivozubov and S.A.Spirin (2010) Comparison of protein phylogeny reconstruction methods using natural protein sequences. *Moscow University Biological Sciences Bulletin*, **65** (4):139–141.

2. M. Punta et al. (2012) The Pfam protein families database. *Nucl. Acids Res.* **40** (D1):D290–D301.