REFINEMENT OF PHYLOGENETIC SIGNAL IN MULTIPLE SEQUENCE ALIGNMENT: RESULTS OF SIMULATION STUDY

Rusin L.Y.*[*]*, Lyubetsky V.A.
Institute for Information Transmission Problems, RAS, Moscow, Russia
[*] Corresponding author: e-mail: roussine@yandex.ru

SUMMARY

Motivation: Disparate substitution rates within the different regions of homologous sequences and mutational saturation are well known to cause misalignment of sequences and to hamper accurate tree reconstruction. Therefore, there is a need in tools detecting and filtering out informational noise from the multiple alignment of sequence data; the tools will help to increase accuracy and resolution of phylogenetic analyses.

Results: We propose such a tool and tested its ability to improve the quality phylogenetic trees both on the biological COG data, and on the artificial data, where the ideal tree was known *a priory*. The key operation of the filtering is a removal of noisy columns. It was shown that the tool permits to reconstruct a tree closer to the "true" tree than is the tree reconstructed with data without removal. Procedure can be applied as a tool to pre-process multiple alignments and enhance phylogenetic inference.

INTRODUCTION

A common problem with large scale phylogenetic analyses is quality of primary sequence data. Many genomic applications require comparison of multiple phylogenies estimated from different families of orthologous genes in order to infer evolutionary events on a genomic scale. Prediction strength of this type of analysis in many respects will therefore depend upon reliability of individual reconstructions. Disparate substitution rates across regions of homologous sequences and mutational saturation are well known to result in elevated levels of homoplasy in the data and to overshadow available phylogenetic signal. The authors developed a procedure to detect and filter out informational noise from multiple alignment of protein sequence data, thus allowing one to considerably increase accuracy and resolution of phylogenetic analysis. In this work the procedure performance is studied with computer simulations.

MATERIALS AND METHODS

Phylogenetic software used at successive steps of the procedure was described in (Lyubetsky *et al.*, 2005) and includes originally developed programs that implement algorithms of computing the objective scoring function and constrained generation of random trees.

Simulations were conducted with the *evolver* program from PAML package (Yang, 1997). We have generated 1000 datasets, each consists of 40 amino acid sequences of length 300; the maximum-likelihood model parameters and branch lengths were obtained

from analysis of COG data, the same parameters were used in our previous studies (Lyubetsky *et al*., 2005).

Algorithm of the procedure was described in detail in (Lyubetsky *et al*., 2005). In essence, it uses a scoring function to rank columns of the alignment according to the consistency of the column's content with a *list of reliable clades* and gradually removes the least consistent ones until signal is refined to provide for better resolution of the tree. The list of reliable clades is basically the list of splits occurring in 70 % majority-rule consensus topology constructed after bootstrapping the intact alignment (i.e. before column removal). On each step of removing columns the g_1 statistic is estimated on current alignment (Hillis, Huelsenbeck, 1992) with the original algorithm of generating random trees strictly *compatible* with the list of reliable clades (Lyubetsky *et al*., 2005) and is used to determine the step, at which the procedure halts. The obtained alignment is considered optimal for tree reconstruction (definitive phylogenetic analysis).

SIMULATION RESULTS

Simulation studies were aimed at proving two statements: (1) removal of noisy columns permits to reconstruct a tree closer to the known tree than is the tree reconstructed with data without removal; (2) g_1 statistic estimated with the original algorithm of constrained random tree generation can be used to identify the refinement step, at which the procedure should be stopped.

The lists of reliable clades were quite different among the generated datasets, probably, due to unequal fraction of hypervariable sites retained in each of 500 replicates after bootstrapping the data. The datasets that produced well resolved consensus trees after bootstrapping were assumed to contain low amount of hypervariable sites and enough informative sites to produce a robust tree. Therefore, we used datasets (512 out of 1000), which produced consensus trees sufficiently unresolved to generate 100,000 constrained random topologies on their basis as test datasets to refine the signal.

Batch refinement of *in silico* generated datasets was continued for 10 steps. At each step, current alignment was analyzed to produce a phylogenetic tree and a g_1 score. Trees from successive steps were computed likelihoods against the *intact data* and compared using standard tests of phylogenies (approximately unbiased test, Kishino-Hasegawa test, Shimodaira-Hasegawa test).

In 100 % cases removing noisy columns permitted to reconstruct the tree, which is closer to the known tree used to simulate the data than is the tree obtained without refinement. In 91 % cases the tree with highest likelihood ("best" tree) was reconstructed at the step of the procedure, where the alignment produced the minimal (optimal) g_1 score, and in 53 % cases the difference in likelihood between the found "best" tree and the tree inferred with intact data was statistically significant. In 9 % cases the g_1 score continued to decrease beyond the step, at which the "best" tree is found, which might suggest that, although the signal related to poorly resolved branches of 70 %-consensus can be refined further, the columns needed to correctly reconstruct shallow parts of the whole tree (containing recent evolutionary events and, therefore, described by more variable regions) are already removed. Further studies will be conducted to develop measures of clade-specific noise removal. In the meantime, the described procedure can be used to refine alignments and improve phylogenetic inference with the advice to compare likelihoods of trees before and after column removal.

ACKNOWLEDGEMENTS

The authors are greatly indebted to O. Zverkov for valuable help.
The work was partly supported by grants RFBR Nos 05-04-49705 and ISTC 2766.

REFERENCES

Hillis D.M., Huelsenbeck J.P. (1992) Signal, noise, and reliability in molecular phylogenetic analyses. *J. of Heredity*, **83**, 189–195.

Lyubetsky V.A., Gorbunov K.Y., V'yugin V.V., Rusin, L.Y. (2005) Removing noise in multiple protein alignment. *Information processes*, **5**(5), 380–391.

Yang Z. (1997) PAML: a program for phylogenetic analysis by maximum likelihood version. *CABIOS*, **13**, 555–556.