ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Maximum likelihood phylogenetic inference: An empirical comparison on a multi-locus dataset

Markus Göker[1], Alexandros Stamatakis[2]

[1] Organismic Botany/Mycology, Auf der Morgenstelle 1, Tübingen, University of Tübingen, Germany

[2] Laboratory for Computational Biology and Bioinformatics (LCBB), School of Computer & Communication Sciences,

Swiss Federal Institute of Technology, Station 14, CH-1015 Lausanne, Switzerland

Introduction

Maximum likelihood (ML) is currently considered to be one of the most reliable and robust optimality criteria for phylogenetic inference from nucleotide or amino acid sequence data. However, computing the ML function is demanding, particularly if large numbers of taxa and/or large numbers of characters are involved. Efficient implementations of tree search under the likelihood criterion are therefore required.

Based on a re-analysis of a recently published (Göker et al. 2006) DNA dataset comprising 72 representative taxa of plant-parasitic downy mildews (DM) and relatives (Oomycetes, Stramenopiles), five coding as well as non-coding loci and a total of 3921 characters, we compared phylogenetic trees computed with different ML programs. Using partition-specific substitution models and topological constraint options recently implemented in the latest version of RAxML (Stamatakis 2006), it turned out that the best trees obtained with this software are topologically different from the best trees from other programs (Guindon and Gascuel 2006; Jobb et al. 2004) and have higher likelihood values. Best RAxML trees are topologically rather uniform. Thus, results are independent of the ML models used.

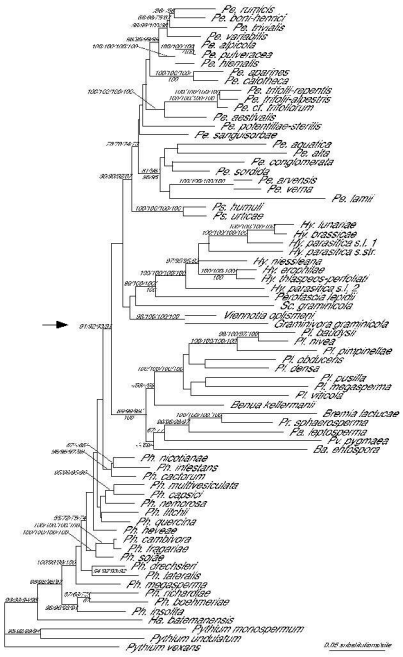

Figure 1. Maximum Likelihood tree from Göker et al. (2006) obtained with PhyML Version 2.4.4 under a GTR model of site substitution including estimation of GAMMA-distributed rate heterogeneity and a proportion of invariant sites. Numbers on branches are bootstrap values as obtained with PhyML and Treefinder (Version of October 2005) under GTR+GAMMA and three different modes of partitioning. Note consistently high support for a monophyletic group representing the downy mildews (DM; Peronosporales, Chromista) (arrow).

Surprisingly, strong statistical support for a systematically important arrangement (the monophyly of the downy mildews) disagreeing with the best ML trees inferred with RAxML has been observed in previous analyses of the same dataset in both distance (neighbour-joining) and ML bootstrap analyses. We therefore investigated whether this could be due to an artifact of the underlying tree search algorithms. A confirmation of this hypothesis could be of relevance for practitioners in phylogenetics since it implies that biases induced by starting trees and/or branch swapping algorithms could strongly affect maximum likelihood bootstrap values even in case of small to moderately-sized datasets.

Methods

We hypothesized that a too low number of starting trees followed by insufficient topological rearrangement may give misleading results even in moderately-sized empirical datasets. In the concrete case, this may be due to neighbour-joining starting trees being biased towards a certain topology since both TREEFINDER and PhyML use BIONJ to obtain staring trees for branch swapping and Neighbour-Joining bootstrapping resulted in considerable support for DM monophyly (Göker et al. 2006).

To empirically assess our hypothesis, RAxML was also executed with suboptimal starting trees and a less thorough branch swapping mechanism.

Search radius	Starting Tree		
	BIONJ	Parsimony	Random
2	0.680	0.608	[not determined]
10	0.252	0.486	0.377
15	0.253	0.465	0.343
estimated	0.275 (5-10)	0.490 (5-15)	0.369 (10-25)

Table 1. Proportion of bootstrap trees in agreement with DM monophyly as obtained under different RAxML tree search settings. Support is > 50% only if an extremely low SPR search radius of 2 is used. Numbers in parentheses denote the actual search radius used for bootstrapping in case initial rearrangement setting was automatically determined by RAxML (the default mode).

Search radius	Starting Tree		
	BIONJ	Parsimony	Random
2	-49828.395294	-49826.933050	-49828.497762
	-	-49833.059358	-49833.623398
10	-49827.801948	-49826.933035	-49826.932808
	-	-49833.059693	-49846.165360
15	[not determined]	-49826.932721	[not determined]
	[not determined]	-49846.436814	[not determined]
estimated	-49829.003776	-49826.933166	-49826.932897
	-	-49846.165618	-49846.845107

Table 2. Best ML values as obtained under different RAxML tree search settings. Upper values are best ML values obtained with trees indicating DM polyphyly, the values below are best ML values from trees in agreement with DM monophyly. Globally best trees ranging from -lnL=49826.932721 to -lnL=-49832.103404 all displayed DM polyphyletic and subdivided into two groups as in Fig. 2. PhyML Version 2.4.4 under GTR+GAMMA and default settings otherwise resulted in a best ML value of -49838.761763.

Results and Conclusions

It turns out that if if the size of the topological search space of RAxML is reduced, both best trees of a lower ML as well as higher bootstrap values for downy mildew monophyly are obtained. We thus conclude that the high bootstrap support values for DM monophyly as observed in Göker et al. (2006) represent an artifact. It turns out that a too small search radius in branch swapping may strongly affect bootstrap values even in moderately-sized datasets.

It is obvious that the most important steps in the evolution of downy mildews and their closest relatives are related to their parasitism on plant hosts (Göker et al. 2006), a conclusion which is not affected by the results presented here. Rather, contrary to our earlier findings, it must be regarded as unclear at present whether downy mildews arose once or twice from *Phytophthora*-like ancestors (Fig. 2).

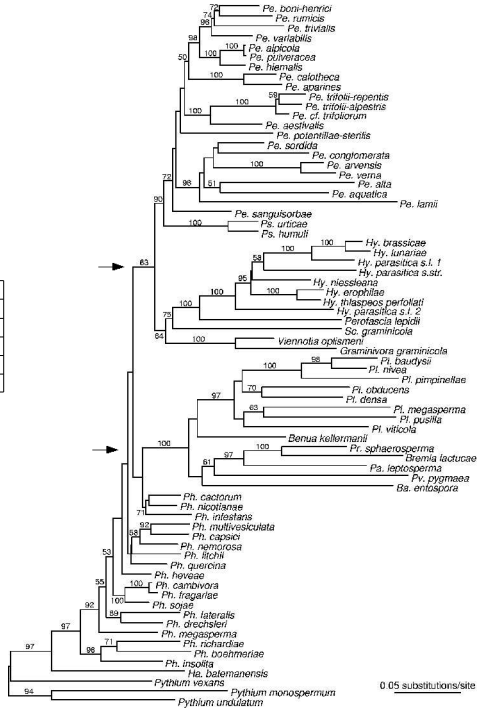

Figure 1. Globally best ML tree found with RAxML version 2.1.3 under GTR+GAMMA and default search values. The DM appear as polyphyletic and are subdivided into two groups (arrows). Bootstrap support for this arrangement is low, but the arrangement shown here *or* DM monophyly is present in almost 100% of the bootstrap trees (Table 1).

References:

- [1] Göker, M., Voglmayr, H., Riethmüller, A., Oberwinkler, F., 2006. How do obligate parasites evolve? A multi-gene phylogenetic analysis of downy mildews. Fungal Genetics and Biology, in press.
- [2] Stamatakis, A., 2006. RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. To be published in Bioinformatics. RAxML is available at http://icwww.epfl.ch/~stamatak/
- [3] Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696-704.
- [4] Jobb, G., Haeseler, A. von, Strimmer, K., 2004. TREEFINDER, A powerful graphical analysis environment for molecular phylogenetics. BMC Evol. Biol. 4, 18.