PHYML Online: A Web Server for Fast Maximum Likelihood-Based Phylogenetic Inference

Stéphane Guindon, F. Le Thiec, Patrice Duroux, Olivier Gascuel

HAL Id: lirmm-00105317

https://hal-lirmm.ccsd.cnrs.fr/lirmm-00105317

Submitted on 11 Oct 2006

PHYML Online — a web server for fast maximum likelihood-based phylogenetic inference

Stéphane Guindon[a], Franck Lethiec[b], Patrice Duroux[b] and Olivier Gascuel[b] [1]

[a] Bioinformatics Institute & Allan Wilson Centre, University of Auckland, Private Bag 92019, Auckland, New Zealand, [b] Projet Méthodes et Algorithmes pour la Bioinformatique, LIRMM-CNRS, 161 Rue Ada, 34392-Montpellier Cedex 5, France.

[1]to whom correspondence should be addressed.
Olivier Gascuel.
LIRMM-CNRS, 161 Rue Ada.
34392-Montpellier Cedex 5.
FRANCE.
gascuel@lirmm.fr

Abstract

PHYML Online is a web interface to PHYML, a software that implements a fast and accurate heuristic for estimating maximum likelihood phylogenies from DNA and protein sequences. This tool provides the user with a number of options, e.g. nonparametric bootstrap and estimation of various evolutionary parameters, in order to perform comprehensive phylogenetic analyses on large data sets in reasonable computing time. The server and its documentation are available from `http://atgc.lirmm.fr/phyml`.

Introduction

The ever-increasing size of homologous sequence data sets and complexity of substitution models stimulate the development of better methods for building phylogenetic trees. Likelihood-based approaches (including Bayesian) provided arguably the most successful advances in this area in the last decade. Unfortunately, these methods are hampered with computational difficulties. Different strategies have then been used to tackle this problem, mostly based on stochastic approaches. Markov chain Monte Carlo methods are probably the most valuable tools in this context as they provide computationally tractable solutions to Bayesian estimation of phylogenies (1; 2).

Stochastic approaches have also been used to address optimisation issues in the maximum likelihood framework. Hence, simulated annealing (3) and genetic algorithms (4; 5) were proposed to estimate maximum likelihood phylogenies from large data sets. However, the hill climbing principle is usually considered faster than stochastic optimisation and sufficient for numerous combinatorial optimisation problems (6). Recently, Guindon and Gascuel (2003) described a fast and simple heuristic based on this principle, for building maximum likelihood phylogenies. Several simulation studies (7; 8) demonstrated that the tree topologies estimated with this approach are as accurate as those inferred using the best tree building methods cur-

2

rently available. These studies also showed that this new method is considerably faster than the other likelihood-based approaches. Using this heuristic, the analysis of large data sets is now achieved in reasonable computing time on any standard personal computer; e.g., only 12 min were required to analyse a data set consisting of 500 *rbc*L sequences with 1,428 base pairs from plant plastids.

This paper introduces PHYML Online, a web interface to the PHYML (PHYlogenetic inferences using Maximum Likelihood) software that implements the heuristic described by Guindon and Gascuel (2003). PHYML Online provides a number of useful options (e.g. nonparametric bootstrap), and proposes quite recent models of sequence evolution (e.g., WAG (9) and DCMut (10)). We first give an overview of the algorithm and present the web server thereafter.

Algorithm

The core of the heuristic is based on a well-known tree-swapping operation, namely "nearest neighbour interchange" (NNI), which defines three possible topological configurations around each internal branch (see (11)). For each of these configurations, the length of the internal branch that maximises the likelihood is estimated using numerical optimisation. The difference of likelihood obtained under the best alternative topological configuration and the current one defines a score. A score with positive value indicates that the best alternative topological configuration yields an improvement of likelihood. A score with negative value indicates that the current topological configuration cannot be improved at this stage and only the length of the internal branch is adjusted. Each internal branch is examined in this manner and ranked according to its score. The optimal length of external branches is also computed. These calculations are performed independently for every branch and define a set of (topological or numerical) modifications, each of which corresponding to an improvement of the current tree regarding the likelihood function.

The standard approach would only apply one of these modifications, typically that corresponding to the internal branch with best score. Here, a large proportion of all modifications computed previously is performed instead. This proportion is adjusted so as to increase the likelihood at each step, ensuring convergence of the algorithm. This way, the current tree is improved at each step, both in terms of topology and branch length, and only a few steps (usually a few dozen or less) are necessary to reach an optimum of the likelihood function. This explains the speed of this algorithm whose time complexity is $\mathcal{O}(pns)$, where p represents the number of refinement steps that have been performed and n is the number of sequences of length s.

PHYML Online

PHYML Online is a web interface to the PHYML algorithm (Figure 1). By default, the input data consists of a single text file containing one or more alignments of DNA or protein sequences in PHYLIP (12) interleaved or sequential format. Examples of sequence data sets in PHYLIP format are given in the 'User's guide' section of the web site.

Setting the parameters of a phylogenetic analysis through the interface is straightforward. The first step is the selection of the substitution model of interest. Alignments of homologous DNA and amino-acid sequences can be examined under a wide range of models (JC69, K80, F81, F84, HKY85, TN93 and GTR for nucleotides, and Dayhoff, JTT, mtREV, WAG and DCMut for amino acids). Variability of substitution rates across sites and invariable sites can also be taken into account. The parameters that model the intensity of the variation of rates across sites and the proportion of invariables sites can be fixed by the user or estimated by maximum likelihood. Note that the parameters of the substitution model can be estimated under a fixed tree topology or not. The fixed topology option is useful when describing the evolutionary process is more important than estimating the history of sequences.

An option is available to assess the reliability of internal branches using nonparametric

bootstrap (13) which is possible to achieve for even large data sets thanks to the speed of PHYML optimisation algorithm. The number of bootstrap replicates is fixed by the user. The bootstrap values are displayed on the maximum likelihood phylogeny estimated from the original data set. Trees estimated from each bootstrap replicate, as well as the corresponding substitution parameters, can also be saved in separate files for further analysis (e.g., computation of confidence intervals for the substitution parameters or estimation of a consensus bootstrap tree, as performed by PHYLIP's CONSENSE for instance).

Several data sets can be analysed in a single run. This option is especially useful in multiple gene studies. Multiple trees can be also be used as input and further optimised by the algorithm described above. This might prevent the tree searching heuristic to be trapped in local maxima. When combined with the fixed tree option, the multiple input trees approach also facilitates the comparison of the fit of different phylogenies estimated from a single data set. The 'User's guide' section gives details on the format of multiple sequence and tree files.

Sequences (and starting tree(s) if provided) are uploaded on our server, a 16-processor IBM computer running Linux 2.6.8-1.521custom SMP, and a maximum likelihood analysis is performed using the PHYML algorithm. Results are then sent to the user by electronic mail. The first file presents a summary of the options selected by the user, maximum likelihood estimates of the parameters of the substitution model that were adjusted, and the log likelihood of the model given the data. The second file shows the maximum likelihood phylogeny(ies) in NEWICK format. Trees can be viewed through an applet available on the PHYML Online server. This applet runs the program ATV (14) that provides numerous options to display and a manipulate large phylogenetic trees.

Availability

The PHYML Online server is located at "Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier" : `http://atgc.lirmm.fr/phyml`

PHYML can also be downloaded for local installation at `http://atgc.lirmm.fr/phyml/binaries.html`. The PHYML software has been implemented in C ANSI and is available under GNU general public licence. Sources are available upon request. Binaries, example data sets, sources and documentation are distributed free of charge for academic purpose only.

Acknowledgements

References

1. Rannala, B. and Yang, Z. (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.,* **43**, 304–311.

2. Huelsenbeck, J. P. and Ronquist, F. (2001) MrBayes: Bayesian inference of phylogeny. *Bioinformatics,* **17**, 754–755.

3. Salter, L. and Pearl, D. (2001) Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. *Syst. Biol.,* **50**, 7–17.

4. Lewis, P. (1998) A genetic algorithm for maximum likelihood phylogeny inference using nucleotide sequence data. *Mol. Biol. Evol.,* **15**, 277–283.

5. Lemmon, A. and Milinkovitch, M. (2002) The metapopulation genetic algorithm: an efficient solution for the problem of large phylogeny estimation. *Proc. Natl. Acad. Sci. USA.,* **99**, 10516–10521.

6. Aarts, E. and Lenstra, J. K. (1997) Local search in combinatorial optimization, Wiley, Chichester.

7. Guindon, S. and Gascuel, O. (2003) A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.,* **52**, 696–704.

8. Vinh, L. S. and vonHaeseler, A. (2004) IQPNNI: Moving fast through tree space and stopping in time. *Mol. Biol. Evol.,* **21**, 1565–1571.

9. Whelan, S. and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.,* **18**, 691–699.

10. Kosiol, C. and Goldman, N. (2004) Different versions of the dayhoff rate matrix. *Mol. Biol. Evol.,* **In press**.

11. Swofford, D., Olsen, G., Waddel, P., and Hillis, D. (1996) Phylogenetic inference. In Hillis, D., Moritz, C., and Mable, B., (eds.), *Molecular Systematics*, chapter 11 Sinauer Sunderland, MA.

12. Felsenstein, J. (1993) PHYLIP (PHYLogeny Inference Package) version 3.6a2, Distributed by the author, Department of Genetics, University of Washington, Seattle.

13. Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution,* **39**, 783–791.

14. Zmasek, C. and Eddy, S. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics,* **17**, 383–384.
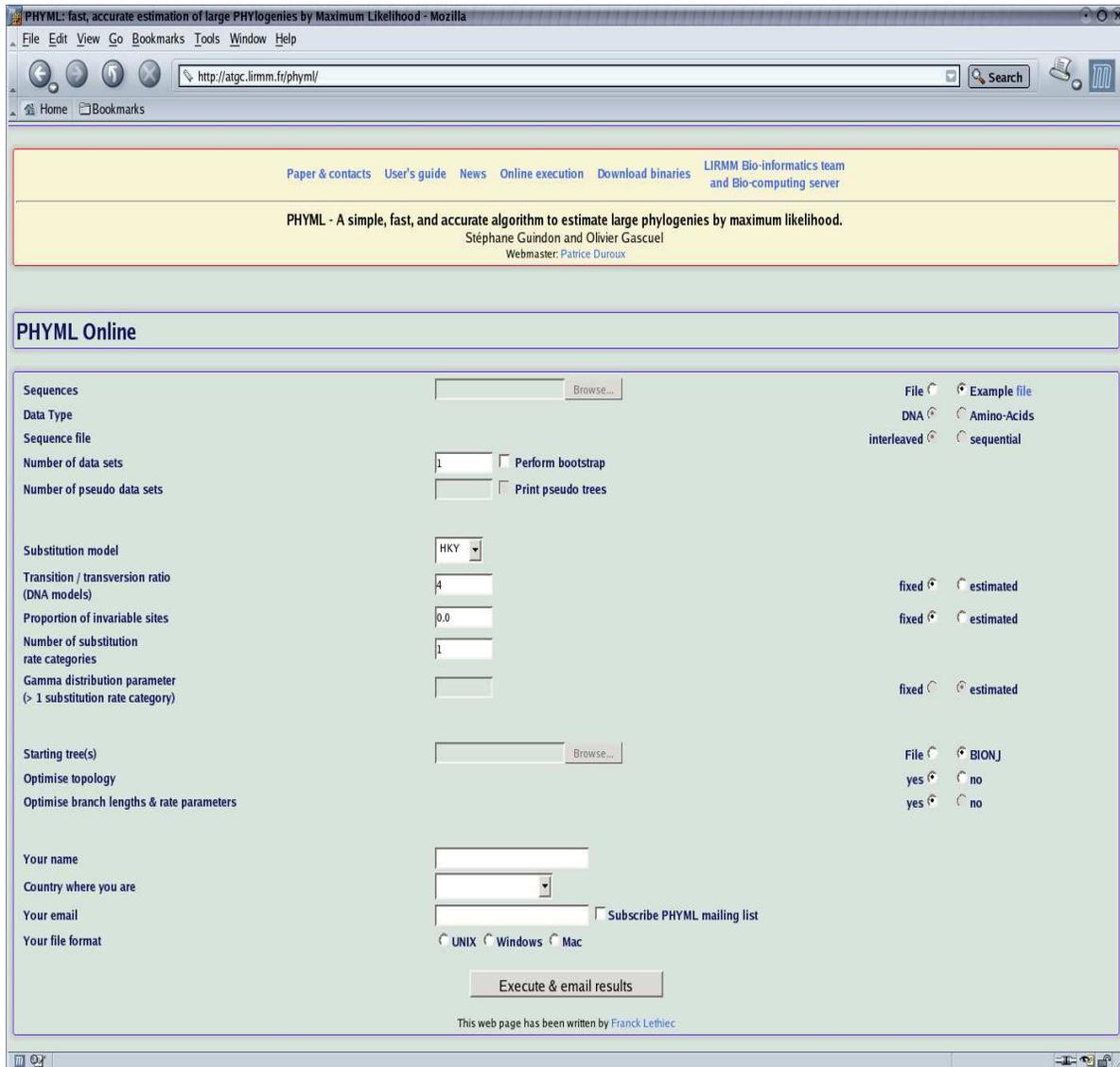
Figure 1. The PHYML Online interface.