HAL
archives-ouvertes.fr

Phylogeny.fr: robust phylogenetic analysis for the non-specialist

A. Dereeper, Vincent Guignon, G. Blanc, S. Audic, S. Buffet, Jean-François Dufayard, Stéphane Guindon, Vincent Lefort, M. Lescot, J.-M. Claverie, et al.

▶ **To cite this version:**

A. Dereeper, Vincent Guignon, G. Blanc, S. Audic, S. Buffet, et al.. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Research, Oxford University Press, 2008, 36 (Web Server), pp.W465-W469. <http://www.phylogeny.fr/>. <10.1093/nar/gkn180>. <lirmm-00324099>

HAL Id: lirmm-00324099
https://hal-lirmm.ccsd.cnrs.fr/lirmm-00324099

Submitted on 24 Sep 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Phylogeny.fr: Robust Phylogenetic Analysis for the Non-Specialist

A. Dereeper[1,*], V. Guignon[2,*], G. Blanc[1], S. Audic[1], S. Buffet[1], F. Chevenet[3], J.-F. Dufayard[2], S. Guindon[2], V. Lefort[2], M. Lescot[1], J.-M. Claverie[1,] and O. Gascuel[2,**]**

[1]Information Génomique et Structurale (IGS), CNRS-UPR2589, IBSM, Marseille, France

[2]Méthodes et Algorithmes pour la Bioinformatique (MAB), LIRMM, CNRS – Univ. Montpellier II, France

[3]Génétique et Evolution des Maladies Infectieuses (GEMI), CNRS – IRD, Montpellier, France

* Joint first authors.

** Joint corresponding authors: Jean-Michel.Claverie@igs.cnrs-mrs.fr and Olivier.Gascuel@lirmm.fr

ABSTRACT

Phylogenetic analyses are central to many research areas in biology and typically involve the identification of homologous sequences, their multiple alignment, the phylogenetic reconstruction and the graphical representation of the inferred tree. The Phylogeny.fr platform transparently chains programs to automatically perform these tasks. It is primarily designed for biologists with no experience in phylogeny, but can also meet the needs of specialists; the first ones will find up-to-date tools chained in a phylogeny pipeline to analyze their data in a simple and robust way, while the specialists will be able to easily build and run sophisticated analyses. Phylogeny.fr offers three main modes. The "One Click" mode targets non-specialists and provides a ready-to-use pipeline chaining programs with recognized accuracy and speed: MUSCLE for multiple alignment, PhyML for tree building, and TreeDyn for tree rendering. All parameters are set up to suite most studies, and users only have to provide their input sequences to obtain a ready-to-print tree. The "Advanced" mode uses the same pipeline but allows the parameters of each program to be customized by users. The "A la Carte" mode offers more flexibility and sophistication, as users can build their own pipeline by selecting and setting up the required steps from a large choice of tools to suite their specific needs. Prior to phylogenetic analysis, users can also collect neighbors of a query sequence by running BLAST on general or specialized databases. A guide tree then helps to select neighbor sequences to be used as input for the phylogeny pipeline. Phylogeny.fr is available at: http://www.phylogeny.fr/**.**

INTRODUCTION

Reconstructing the evolutionary history of molecular sequences through phylogenetic analysis is at the heart of many biological research areas such as comparative genomics, functional prediction, detection of lateral gene transfer, or the identification of new micro-organisms. Starting from a sequence of interest, a typical phylogenetic analysis goes through successive steps that include the identification of homologous sequences, multiple alignment, phylogenetic reconstruction and graphical representation of the inferred tree. This process requires substantial computational resources depending on the number and length of the sequences, and on the methods being run.

A huge variety of models, approaches and computer programs are currently available, as can be seen from Joe Felsenstein's phylogeny software inventory[1]. The task of deciding which method to use, installing the corresponding programs, and running them on a local computer, is beyond the reach of most occasional users. Yet, phylogenetic trees have become a compulsory illustration (and referee's request) in most sequence-related studies. As a consequence, user-friendly but ancient programs are still widely used, though much improved methods now exist and are only used by the specialists, typically involved in molecular evolution or systematics.

In this context, Phylogeny.fr has been designed to provide a ready-to-use platform that transparently chains alignment and phylogeny programs in a comprehensive and flexible manner. Although phylogenetic specialists will be able to find up-to-date tools and run sophisticated analyses based on their favorite approaches and their own parameter settings, the primary philosophy of Phylogeny.fr is to assist biologists with no experience in phylogeny in analyzing their data in a simple and robust way, using methods corresponding to well accepted standards. Maximum-likelihood (ML) tree construction is the default option to infer phylogenies, which is commonly recognized (1) as the most accurate approach (along with Bayesian) in molecular phylogenetics.

Phylogeny.fr offers "one-stop-shopping" among a variety of leading methods for multiple sequence alignment, phylogenetic reconstruction and graphical representation of trees, and chains these methods into a pipeline that can be executed in three modes. The "One Click" mode is designed for biologists with no experience in bioinformatics; given a set of unaligned sequences, a predefined pipeline using MUSCLE (2),

[1] 1: http://evolution.genetics.washington.edu/phylip/software.html

Gblocks (3), PhyML (4) and TreeDyn (5) outputs the corresponding phylogenetic tree in a ready-to-print format. The “Advanced” mode allows the settings of each “One Click” tool to be customized by users. The “A la Carte” mode offers flexible choices in the pipeline steps, the tools and their settings to suite the more specific needs of experts.

Several other websites propose related services. PhyloBuilder (6) and PhyloBlast (7) are dedicated to proteins and gather homologs of a query sequence to build a phylogenetic tree using a distance or a parsimony method. POWER (8) infers a phylogenetic tree from a sequence set using a pipeline involving ClustalW (9) and PHYLIP (10) programs. Tarraga et al. recently proposed Phylemon (11) which provides experts with a suite of on-line programs and a Java interface to build a phylogeny pipeline. The main specificity of Phylogeny.fr is the combination of an interface designed for the non-specialists with up-to-date programs that are often reserved to experts. Moreover, Phylogeny.fr is able to analyze both DNA and protein sequences.

OVERVIEW

The Phylogeny.fr platform proposes three major components:

(i) a pipeline to reconstruct a phylogenetic tree from a set of sequences through an automated process that successively performs multiple sequence alignment, alignment refinement, phylogenetic reconstruction and, finally, a graphical representation of the resulting tree;

(ii) a fast parallel Blast (12) module to search for similar sequences of a query sequence. Sequence selection is facilitated by a “quick-and-dirty” guide tree based on Blast results, and by an estimate of final alignment length;

(iii) a suite of stand-alone phylogenetic programs.

All these components are offered in a user-friendly tabulated menu facilitating navigation, together with examples to familiarize users with the correct input and expected results. All programs used in the pipeline and their parameters and references are displayed in an “Analysis overview” that provides the main information to be indicated in a publication. In addition, the web interface enables users to visualize, manipulate, and locally edit their results through applet viewers such as Jalview (13) for alignment, or ATV (14) for phylogenetic tree.

PHYLOGENY PIPELINE

At the core of the system lies the pipeline allowing a phylogenetic tree to be built at once from a set of sequences. This pipeline consists of a succession of Perl modules wrapping different external software programs, which can be executed through the three modes described below. This process requires many additional programs that are not detailed here: (1) to make the junction between the main programs (format compatibility), (2) to output the results in several formats (e.g. PNG or PDF), or (3) to rearrange and modify the tree (e.g. Retree from PHYLIP is used to root the tree with the midpoint method or using an outgroup).

"One Click" Mode

The "One Click" mode targets users who do not wish to deal with program and parameter selection. By default, the pipeline is already set up to run and connect well recognized programs: MUSCLE for multiple alignment, Gblocks for automatic alignment curation, PhyML for tree building and TreeDyn for tree drawing.

Several studies showed that these programs are both fast and accurate. MUSCLE was assessed using several alignment databases, including the BAliBASE benchmark (15), on which it achieved the highest ranking of any method at the time of publication. PhyML was shown to be at least as accurate as other existing phylogeny programs using simulated data, while being one order of magnitude faster. Both programs have continuously been improved. These two programs are widely used (~650 and ~1000 citations in Web of Science, for MUSCLE and PhyML, respectively). It follows that they are robust, stable and bug free, and that the best way to use them has been reported by their numerous users. Notably, the default options and parameter values have been selected with care; they are used in the "One Click" mode and should be suited for most studies.

PhyML is run with the aLRT statistical test (16) of branch support. This test is based on an approximation of the standard Likelihood Ratio Test, and is much faster to compute than the usual bootstrap procedure. Both methods output the same trees; branch supports may differ but are generally highly correlated. Altogether, this mode makes routine phylogenetic analyses simple and fast. Users only need to load their set of sequences and submit their request. A few minutes later, the system will display a ready-to-print phylogenetic tree in a publication quality image. One has still the possibility to disable or not the

alignment refinement performed by the Gblocks program, which eliminates poorly aligned positions and divergent regions. It is also possible to specify an outgroup (midpoint rooting is the default option).

“Advanced” Mode

In the “Advanced” mode, the pipeline is the same as the one used in the “One Click” mode but users can now freely edit the settings of each program. The pipeline is flexible enough to enable users to select which steps to perform. In this manner, input data can be a set of non-aligned sequences, a sequence alignment or a tree in Newick format. Various file formats are supported as shown in Table 1. Furthermore, the system offers the possibility to inspect the results of each step before launching the next program, so that expert users can detect problems and properly adjust parameters accordingly.

“A la Carte” Mode

This mode presents the same interface as the “Advanced” mode (e.g., step-by-step execution, parameter settings) but offers the possibility of running, testing and comparing the efficiency of a larger panel of methods and programs. It proposes a user-friendly customizable pipeline on which users can select the steps to perform, and the program to use for each of them.

For the alignment step, this mode proposes T-Coffee (17) which ranks high in accuracy but is slower than MUSCLE, and its derivative 3DCoffee (18) that offers the unique possibility of incorporating protein structure information to improve alignments.

For phylogeny reconstruction, a choice is offered among the following approaches: maximum likelihood method using PhyML, parsimony using NONA (for DNA, 19) or Protpars (for proteins, 10), or distance-based methods using BioNJ (20) or Neighbor (NJ, 21) from PHYLIP (10). With BioNJ and NJ, evolutionary distances are calculated using Protdist (10) for proteins and Fastdist (22) for DNA, while bootstrap analysis is achieved by combining Seqboot and Consense (10). Distance methods are very fast and should be preferred for large scale analyses, or when performing bootstrap studies. Parsimony and ML methods are roughly as fast in practice (at least when using the programs selected here); they are able to analyze relatively large datasets, but become quite slow with bootstrap. ML is commonly reported as the most accurate approach, but parsimony does not require any substitution model to be selected, which appears preferable to some users. In practice, it is always a good idea to run several programs with various options

and verify the stability of the results. Phylogeny.fr greatly facilitates this conservative approach as users can easily modify the choices in the pipeline and run the server again.

The “A la Carte” mode makes it easy to integrate and compare both old reference programs - such as the basic tree viewers Drawgram and Drawtree, from PHYLIP - and new ones which are not yet well diffused in the scientific community – such as TreeDyn that offers sophisticated tools and options to draw trees. For example, this mode still provides ClustalW which is still the most popular alignment tool to date even though it has been shown to be less accurate than modern alignment programs (23).

Tree visualization and drawing

Following phylogenetic tree reconstruction by any of these three modes, an image of the tree is produced to facilitate its interpretation. Phylogeny.fr allows the tree image to be drawn and modified using a variety of TreeDyn options. These include: change of the tree shape (rectangular or radial), font style, color of the text and branches, display of branch supports, edition of taxon names, and root selection. The tree image can then be obtained in PNG and PDF formats and eventually incorporated in any artwork for publication.

SEQUENCE SEARCHING

Prior to a phylogenetic analysis, users can also build their initial dataset by running Blastall against several public sequence databases (e.g. 16S Database, GenBank NT, Swissprot). This search is performed in parallel on a 25-node Linux cluster, and allows for a quick exploration of the neighbors of a query sequence. A rough multiple alignment of the query and hit sequences is generated on the fly by piling the individual Blast pairwise alignments. This alignment serves to construct and display a NJ tree using uncorrected p-distances, so that sequences can be selected and incorporated into the dataset by simply clicking on the tree image. To facilitate this process, sequence names are colored according to the taxonomic group of the species they belong to. Furthermore, to help in the choice of sequences maximizing the number of gap-free sites available for phylogenetic reconstruction, an estimation of the length of the final multiple alignment is dynamically computed each time a sequence is selected. Once the selection step is completed, users can run the phylogeny pipeline or retrieve the selected sequences in FASTA format.

SERVER FEATURES AND LIMITATIONS

The platform currently runs on a dedicated server (PowerEdge 2850-Xeon 2.8GHz/2x2MB Dual Core), except for the Blast module which is parallelized on a 25-CPU cluster. Input limitations depend on the selected program and its computational speed, as shown in Table 1. MUSCLE and ClustalW are limited to 200 sequences, while T-Coffee and 3D-Coffee limitations are <50 sequences and <2000 sites. Distance-based phylogeny programs (i.e. NJ and BioNJ) have no limitation, while all other phylogeny programs are limited to: (sequence size) x (number of taxa)^2 < 10,000,000 (e.g. 100 sequences with length 1000). The number of bootstrap replicates – a highly sensitive parameter in terms of computational burden – is limited to 100 with all phylogeny programs, except NJ and BioNJ which are allowed for 500 replicates. These limitations should be relaxed in the near future, as the platform should be implemented on more powerful servers and multiple sites.

CONCLUSION AND FUTURE DEVELOPMENTS

To our knowledge, Phylogeny.fr is the first web server designed for both non-specialists and experts, which provides a complete automated phylogenetic analysis - from FASTA file to tree image - thanks to a pipeline of well trusted tools. Phylogeny.fr has been advertised in 2007 on a few diffusion lists. It currently processes about 2500 submissions per month from about 800 different users, who beta-tested the initial platform and now seem to be satisfied. Current waiting times with the “One Click” mode (without bootstrap) are ~2 and ~6 minutes for alignments of 50 DNA sequences with 1463 sites and of 40 proteins with 430 sites, respectively. When a bootstrap analysis is performed (“Advanced” and “A la carte” modes) waiting times are much longer (~100 times with 100 replicates) and the results are sent by email.

The modular architecture of the Phylogeny.fr pipeline will facilitate the addition of new features and new programs according to the evolution of the field. Phylogeny.fr is not committed to a particular set of programs. The methods constituting the backbone of the “One Click” mode today will be replaced by newer ones, if considered better by a large consensus in the field. The dethroned methods will then become part of the “A la carte” mode, to ensure compatibility with previous studies.

Among the further developments to come shortly, we will add the possibility of selecting the best substitution model to reconstruct the phylogeny, using MODELTEST (DNA, 24) or ProtTest (proteins, 25). Additional tools facilitating the selection of sequences after a BLAST search will be integrated, and the final

output tree will become clickable to help recovering information on the sequences, e.g. functional annotations from the Gene Ontology. Moreover, we plan to provide password-protected accounts to enable users to store and load their results, and to refine their previous analyses; researchers will thus be able to create and share on-line projects with their collaborators.

ACKNOWLEDGEMENTS

Sincere thanks to the authors of the many tools used by Phylogeny.fr. This project was supported by fundings from the “Réseau National des Génopoles” (RNG).

CONTRIBUTIONS

IGS and MAB-LIRMM teams equally contributed to this work. AD, VG: conceived and implemented the server, wrote the manuscript; GB, SA, SB, FC, JFD, SG, VL, ML: helped in the design of the server and/or implemented specific tools; JMC: initiated the project, managed the alignment side, wrote the manuscript; OG: managed the phylogeny side, wrote the manuscript.

REFERENCES

1. Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol. 1994 May;11(3):459-68.
2. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004,32:1792-1797.
3. Castresana J. Selection of conserved blocks for multiple alignments for their use in phylogenetic alignments. Mol Biol Evol. 2000,17(4):540-552.
4. Guindon S and Gascuel O. A simple, fast and accurate algorithm to estimate larges phylogenies by maximum likelihood. Syst Biol. 2003,52:696-704.
5. Chevenet F, Brun C, Banuls AL, Jacq B and Christen R. TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinformatics. 2006,7:439.
6. Glanville JG, Kirshner D, Krishnamurthy N, Sjölander K. Berkeley Phylogenomics Group web servers: resources for structural phylogenomic analysis. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W27-32.
7. Brinkman FS, Wan I, Hancock RE, Rose AM, Jones SJ. PhyloBLAST: facilitating phylogenetic analysis of BLAST results. Bioinformatics. 2001 Apr;17(4):385-7.

8. Lin CY, Lin FK, Lin CH, Lai LW, Hsu HJ, Chen SH, Hsiung CA. POWER: PhylOgenetic WEb Repeater--an integrated and user-optimized framework for biomolecular phylogenetic analysis. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W553-6.
9. Thompson JD, Higgins DG and Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nuc Acid Res. 1994,22:4673-4680.
10. Felsenstein J. PHYLIP (PHYLogeny Inference Package) version 3.6a2, distributed by the author, department of genetics, university of Washington, Seattle, 1993.
11. Tarraga J, Medina I, Arbiza L, Huerta-Cepas J, Gabaldon T, Dopazo J and Dopazo H. Phylemon: a suite of web tools for molecular evolution, phylogenetics and phylogenomics. Nuc Acid Res. 2007,35:38-42.
12. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. Basic local alignment search tool. J Mol Biol.1990,215
13. Clamp M, Cuff J, Searle SM and Barton GJ. The Jalview Java alignment editor. Bioinformatics. 2004,20:426-427.
14. Zmasek CM and Eddy SR. ATV: display and manipulation of annotated phylogenetic trees. Bioinformatics. 2001,17
15. Thompson JD, Plewniak F, Poch O. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics. 1999 Jan;15(1):87-8.
16. Anisimova M and Gascuel O. Approximate likelihood ratio test for branchs: A fast, accurate and powerful alternative. Syst Biol. 2006,55:539-552.
17. Notredame C, Higgins DG and Heringa J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000,302:205-217.
18. O'Sullivan O, Suhre K, Abergel C and Notredame C. 3DCoffee: combining protein sequences and structures whithin multiple sequence alignements. J Mol Biol. 2004,340:385-395.
19. Goloboff P. NONA (NO NAME) Version 2 1999, Published by the author, Tucumán, Argentina.
20. Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 1997,14:685-695.
21. Saitou N, and Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987,4(4):406-425.
22. Elias I and Lagergren J. Fast computation of distance estimators. BMC Bioinformatics. 2007,13:8-89.
23. Wallace IM, Blackshields G, Higgins DG. Multiple sequence alignments. Curr Opin Struct Biol. 2005 Jun;15(3):261-6.
24. Posada D, Crandall KA. MODELTEST: testing the model of DNA substitution. Bioinformatics. 1998;14(9):817-8.

25. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 2005 May 1;21(9):2104-5.

Program	Function	Input	Output	Speed	Current limitations	Use
Blastall 2.2.17	Sequence searching	Raw, FASTA	FASTA	Fast	None	Advanced and A la Carte
MUSCLE 3.6	Multiple alignment	FASTA, EMBL/Uniprot, GenBank	FASTA, Clustal, PHYLIP	Fast	<200 sequences	All modes Large dataset
T-Coffee 4.97	Multiple alignment	FASTA, EMBL/Uniprot, GenBank	FASTA, Clustal, PHYLIP, others	Very slow	<50 sequences <2000 sites	A la Carte Small dataset
3DCoffee 4.97	Multiple alignment using structural information	FASTA, EMBL/Uniprot, GenBank	FASTA, Clustal, PHYLIP, others	Very slow	<50 sequences <2000 sites	A la Carte Small dataset
ClustalW 1.83	Multiple alignment	FASTA, EMBL/Uniprot, GenBank	FASTA, Clustal, PHYLIP	Fast	<200 sequences	A la Carte Large dataset
Gblocks 0.91b	Alignment refinement	FASTA	FASTA, Clustal, PHYLIP	Fast	None	All modes Large dataset
PhyML 3.0	Phylogeny using maximum likelihood	FASTA, Clustal, PHYLIP, EMBL, PAUP*/Nexus	Newick	Fast to Slow	(sequence size) × (number of taxa)^2 < 10,000,000	All modes Medium to large dataset
NONA 2.0	DNA phylogeny using parsimony	FASTA, Clustal, PHYLIP, EMBL, PAUP*/Nexus	Newick	Fast to slow	(sequence size) × (number of taxa)^2 < 10,000,000	A la Carte Medium to large datasets
Protpars 3.66	Protein phylogeny using parsimony	FASTA, Clustal, PHYLIP, EMBL, PAUP*/Nexus	Newick	Very slow	(sequence size) × (number of taxa)^2 < 10,000,000	A la Carte Small to medium dataset
Dnadist Protdist BioNJ NJ PHYLYP 3.66	Phylogeny using distances	FASTA, Clustal, PHYLIP, EMBL, PAUP*/Nexus	Newick	Fast	None	A la Carte Large dataset
Bootstrap with PhyML, NONA, Protpars	Estimations of clade supports	FASTA, Clustal, PHYLIP, EMBL, PAUP*/Nexus	Newick	Very slow	<100 replicates	Advanced and A la Carte Small to medium datasets
Bootstrap with distance methods	Estimations of clade supports	FASTA, Clustal, PHYLIP, EMBL, PAUP*/Nexus	Newick	Slow	<500 replicates	A la Carte Medium to large datasets
TreeDyn 198	Tree rendering	Newick	PNG, PDF, TGF	Fast	None	All modes
Drawgram 3.66	Various tree shapes rendering	Newick	PNG, PDF	Fast	None	A la Carte
Drawtree 3.66	Unrooted tree rendering	Newick	PNG, PDF	Fast	None	A la Carte

Table 1: Main tools available on Phylogeny.fr.

Note : With fast programs Phylogeny.fr displays the results within a few minutes or less. Slow programs require about one hour or so, but their results are still displayed on-line, while the results of very slow program are sent by email.