Evolutionary Bioinformatics

SHORT REPORT

PhyLIS: A Simple GNU/Linux Distribution for Phylogenetics and Phyloinformatics

Robert C. Thomson

Center for Population Biology and Department of Evolution and Ecology, University of California, Davis, CA 95616 USA.
Email: rcthomson@ucdavis.edu

Abstract: PhyLIS is a free GNU/Linux distribution that is designed to provide a simple, standardized platform for phylogenetic and phyloinformatic analysis. The operating system incorporates most commonly used phylogenetic software, which has been pre-compiled and pre-configured, allowing for straightforward application of phylogenetic methods and development of phyloinformatic pipelines in a stable Linux environment. The software is distributed as a live CD and can be installed directly or run from the CD without making changes to the computer. PhyLIS is available for free at http://www.eve.ucdavis.edu/rcthomson/phylis/.

Keywords: linux, phylogenetics, phyloinformatics, operating system

Evolutionary Bioinformatics 2009:5 91–95

This article is available from http://www.la-press.com.

Introduction

Phylogenetic methods are playing a growing role in nearly all fields of biology. As researchers accumulate data, the size of phylogenetic analyses and the scope of inferences for which they are employed have both increased dramatically.[1] Increasingly, researchers employ a diverse array of methods that have been developed by a large and talented group of biologists and programmers. While the availability of these methods is an obvious boon for biology as a whole, the sheer number of software packages that are now regularly employed in phylogenetic research forces biologists and system administrators to spend growing amounts of time installing, configuring and maintaining software rather than focusing on research.

The availability of cheap multi-core processors exacerbates this issue, and now many even-moderately sized labs routinely build small clusters or groups of several phylogenetic workstations. Phyloinformatic research, in particular, depends largely on processing power and highly parallel analyses that are spread across many computers and processors. As phyloinformatic pipelines become more complex and sophisticated, careful standardization of operating systems across computers also becomes more complex. While this growing computational power gives researchers the ability to be more creative in attempting complex and time-consuming analyses, the process of compiling, installing, and configuring software for these machines becomes increasingly repetitive, error-prone, and time-consuming. This problem is, in principle, easily solved. What is needed is a simple, stable platform that is specifically geared toward performing phylogenetic analyses. Perhaps the simplest (from a user perspective) solution to this need is a lightweight Linux-based operating system, geared specifically toward phylogenetic and phyloinformatic research. Existing Linux distributions such as Bio-Linux[2] and SciBuntu[3] represent useful steps in this direction. However, these distributions are aimed at a more general usership, and thus do not incorporate many of the packages that are now standard tools for phylogenetic analysis. Moreover, because phylogenetic methods are currently experiencing rapid development, there is a need for a distribution specifically focusing on this area. PhyLIS aims to fill this need.

Implementation

PhyLIS v1.0 is a free GNU/Linux distribution based on the popular and user-friendly Ubuntu Linux distribution. The name PhyLIS is an acronym for Phylogenetic Linux for Informatics and Systematics. The distribution comes with most commonly used phylogenetic software pre-compiled, installed, and configured, which allows this software to be executed by simply typing the appropriate command (Table 1). PhyLIS also contains popular scripting languages (and appropriate phyloinformatic packages) including Perl (with BioPerl), Python (with BioPython), and R (with several packages). It implements parallel versions of several particularly processor-intensive programs using MPI (including BEST,[4] MrBayes,[5] and raxML).[6] PhyLIS aims to present a complete phylogenetic workbench for all steps of analysis from sequence data manipulation to alignment and tree search, including visualization (for alignments and trees), model selection, divergence time estimation, macroevolutionary analyses and tools for automation and batch analysis.

The distribution intentionally re-uses most of the system maintenance packages from Ubuntu, making the actual use of the operating system very similar to Ubuntu (and Debian, upon which Ubuntu is based). Overall, the non-phylogenetic aspects of PhyLIS (e.g. installation, updating software, file system structure) have been kept as close as possible to Ubuntu in order to preserve the large amount of development that the Ubuntu team has put into ensuring an easy-to-use operating system. Because of this, navigating, updating and using the operating system is largely intuitive for users that are already familiar with more widely used operating systems. The bundled phylogenetic tools (Table 1) are available via the command line interface, allowing for straightforward batch analyses and scripting. Software that employs a graphical user interface can be run using graphical launchers on the desktop, in addition to the command line.

Installation

PhyLIS is distributed as a live CD and can be used in two ways. First, it can be booted from the CD without making changes to the underlying operating system. This is useful, for example, for temporarily employing computers (which may not be configured for phylogenetics, or may be configured for other purposes)

Table 1. Bundled phylogenetic software packages and commands used to call them.

Software Package	Command
Ade4[7]	(R package, see documentation)
Ape[8]	(R package, see documentation)
Aptreeshape[9]	(R package, see documentation)
BEAST[10]	beast
BEAUTi[10]	beauti
BEST (including a parallel version)[4]	mbbest, mbbest_mpi
Bioperl[11]	(Perl package, see documentation)
Biopython[12]	(Python package, see documentation)
Blast2 package[13]	blast2, blastclust, blastall, megablast, etc.
Bucky[14]	bucky
Clustalw[15]	clustalw
Dialign2[16]	dialign2–2
FigTree[17]	figtree
Garli[18]	garli
Geiger[19]	(R package, see documentation)
Hmmer[20]	(See documentation)
jModelTest[21]	jmodeltest
LogAnalyser[17]	loganalyser
LogCombiner[17]	logcombiner
MAFFT[22]	mafft
Mesquite (graphical and headless versions)[23]	run_mesquite.sh, run_headless_mesquite.sh
MrBayes (including a parallel version)[5]	mb, mb_MPI
Muscle[24]	muscle
Ouch[25]	(R package, see documentation)
Paloverde[26]	paloverde
PHYLIP[27]	phylip
Phylocom[28]	phylocom
PhyML[29]	phyml
Physim[30]	(R package, see documentation)
Phyutility[31]	phyutility
POA[32]	poa
R	R
r8s[33]	r8s
RaxML (including a parallel version)[6]	raxmlHPC, raxmlHPC_MPI
Readseq[34]	readseq
Seaview[35]	seaview
Seq-gen[36]	seq-gen
T-coffee[37]	t_coffee
Tracer[17]	tracer
TreeAnnotator[17]	treeannotator
TreeLogAnalyser[17]	treeloganalyser

to run phylogenetic analyses when not in use for their primary purpose. At the completion of the analysis, the results can be transferred to permanent storage and the machine rebooted, restoring it to its previous configuration and operating system. The live CD mode is also useful for testing PhyLIS with little time commitment before deciding whether to install it.

Second, PhyLIS can be directly installed from the live CD using a simple graphical installer that allows for a new, complete installation (erasing the previous operating system), or a partitioned installation (allowing for dual boot systems). The distribution has been tested and can be installed on most 32- and 64-bit PCs (including Apple computers that use Intel processors).

Conclusions

PhyLIS aims to simplify the process of carrying out complex phylogenetic analyses and has utility both for individual researchers and for teaching environments. The operating system presents a large suite of tools in a stable platform and should be useful for system administrators performing many installations. However, it is also simple enough to use that individual researchers with little previous Linux experience can employ it effectively. PhyLIS is under active development and undergoes periodic updates every six months to incorporate new versions of software and minor bug fixes. Users are encouraged to request additional software or features that would enhance the utility of the operating system; these will be incorporated into future releases of PhyLIS. The latest release is freely available at http://www.eve.ucdavis.edu/rcthomson/phylis.

Acknowledgements

I thank Phil Spinks, Ian Wang, and two anonymous reviewers for comments on an earlier version of this report. I am also grateful to many early users of PhyLIS for suggestions and feedback. Development of PhyLIS has not been supported by any specific funding, though it grew out of projects funded under a National Science Foundation Doctoral Dissertation Improvement Grant [DEB-0710380], a Graduate Student Award from the Society of Systematic Biologists, and funding from the UC Davis Center for Population Biology.

Conflict of Interest

The author reports no conflicts of interest.

References

1. Hillis DM. The tree of life and the grand synthesis of biology. In: Cracraft J, Donoghue MJ, eds. *Assembling the tree of life*. New York: Oxford University Press; 2004:545–7.
2. NERC Environmental Bioinformatics Centre. Bio-Linux. Version 5.0. 2009; Available from http://nebc.nox.ac.uk/tools/bio-linux.
3. Anjar U. Scibuntu: Ubuntu Linux for scientists. Version 0.4-beta. 2006; Available from http://scibuntu.sourceforge.net/index.html.
4. Liu L, Pearl DK. Species trees from gene trees: Reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst Biol*. 2007;56:504–14.
5. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 2003;19:1572–4.
6. Stamatakis A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 2006;22:2688.
7. Chessel D, Dufour AB, Thioulouse J. The ADE4 package I: One-table methods. *R News*. 2004;4:5–10.
8. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*. 2004;20:289–90.
9. Bortolussi N, Durand E, Blum M, Francois O. apTreeshape: statistical analysis of phylogenetic tree shape. *Bioinformatics*. 2006;22:363–4.
10. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*. 2007;7:214–22.
11. Stajich J, Block D, Boulez K, et al. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*. 2002;12:1611–8.
12. Chapman B, Chang J. Biopython: Python tools for computational biology. *ACM SIGBIO Newsletter*. 2000;20:15–9.
13. Altschul S, Madden T, Schaffer A, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389–402.
14. Ane C, Larget B, Baum D, Smith S, Rokas A. Bayesian estimation of concordance among gene trees. *Mol Biol Evol*. 2007;24:412–26.
15. Chenna R, Sugawara H, Koike T, et al. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*. 2003;31:3497–500.
16. Morgenstern B. DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*. 1999;15:211–8.
17. Available from http://beast.bio.ed.ac.uk/.
18. Zwickl DJ. *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD Thesis*, University of Texas at Austin; 2006.
19. Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. GEIGER: investigating evolutionary radiations. *Bioinformatics*. 2008;24:129–31.
20. Eddy SR. Profile hidden Markov models. *Bioinformatics*. 1998;14:755–63.
21. Posada D. jModelTest: Phylogenetic model averaging. *Mol Biol Evol*. 2008;25:1253–6.
22. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002;30:3059–66.
23. Maddison W, Maddison D. Mesquite: a modular system for evolutionary analysis. Version 2.6. 2009; Available from http://mesquiteproject.org.
24. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792–7.
25. Butler MA, King AA. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *Am Nat*. 2004;164:683–95.
26. Sanderson MJ. Paloverde: an OpenGL 3D phylogeny browser. *Bioinformatics*. 2006;22:1004–6.
27. Felsenstein J. PHYLIP (phylogeny inference package). Version 3.68. 2009; Available from http://evolution.genetics.washington.edu/phylip.html.
28. Webb C, Ackerly D, Kembel S. Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics*. 2008;24:2098–100.
29. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 2003;52:696–704.
30. Weir J, Schluter D. PhySim: phylogenetic tree simulation package. Version 1.0. 2007; Available from http://cran.r-project.org/web/packages/PhySim/index.html.

31. Smith SA, Dunn CW. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics*. 2008;24:715–6.
32. Grasso C, Lee C. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. *Bioinformatics*. 2004;20:1546–56.
33. Sanderson MJ. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*. 2003 2003;19:301–2.
34. Gilbert D. Readseq. Version 2. 2001; Available from http://iubio.bio.indiana.edu/soft/molbio/readseq/java/.
35. Galtier N, Gouy M, Gautier C. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Bioinformatics*. 1996;12:543–8.
36. Rambaut A, Grass NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*. 1997;13:235–8.
37. Notredame C, Higgins DG, Heringa J. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*. 2000;302:205–17.