BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 7 2001
Pages 662–663

PAL: an object-oriented programming library for molecular evolution and phylogenetics

Alexei Drummond[1] *and Korbinian Strimmer*[2,*]

[1]*School of Biological Sciences, University of Auckland, 3A Symonds Street, Auckland, New Zealand and* [2]*Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK*

Received on January 20, 2001; accepted on March 15, 2001

ABSTRACT

Summary: Phylogenetic Analysis Library (PAL) is a collection of Java classes for use in molecular evolution and phylogenetics. PAL provides a modular environment for the rapid construction of both special-purpose and general analysis programs. PAL version 1.1 consists of 145 public classes or interfaces in 13 packages, including classes for models of character evolution, maximum-likelihood estimation, and the coalescent, with a total of more than 27 000 lines of code. The PAL project is set up as a collaborative project to facilitate contributions from other researchers.

Availability: The program is free and is available at http://www.pal-project.org. It requires Java 1.1 or later. PAL is licensed under the GNU General Public License.

Contact: a.drummond@auckland.ac.nz; korbinian.strimmer@zoo.ox.ac.uk

Supplementary information: An online description of the Application Programming Interface (API) of all public classes in PAL is available at http://www.pal-project.org/api/.

A rich methodology has developed for the study of molecular evolution and phylogenetics, and new techniques for analyzing sequence data are being proposed at an increasing pace. Implementation of these algorithms in a computer program requires infrastructure such as a representation of trees and alignments and methods for reading and writing of sequence data. Even though this infrastructure is operationally identical across different programs, the monolithic character of many current programs, code incompatibilities and lack of documentation renders it difficult to reuse existing code.

Phylogenetic Analysis Library (PAL) is an effort to compile a core collection of data structures and methods for molecular sequence analysis that are compatible with each other, well documented and can easily be reused. Thus PAL aims to shortcut the development of analysis programs and simultaneously provide a modular framework for testing of new methodology. PAL is entirely written in the Java programming language. Java enforces a clean object-oriented design and its simplicity and mature language features are advantageous for collaborative development. Moreover, freely-available native compilers such as the GNU Java compiler (available from http://gcc.gnu.org) allow compilation of Java into native machine code for maximum efficiency.

A list of the 13 Java packages available in PAL version 1.1 is shown in Table 1. There are in total 145 different classes and interfaces, e.g. for

- reading and writing sequence alignments, distance matrices, and trees,
- modeling substitution of nucleotide and amino acids (REV, TN, HKY, F84, F81, JC; Dayhoff, JTT, MTREV24, BLOSUM, VT, WAG, CPREV models), incorporating rate variation over sites (Liò and Goldman, 1998),
- maximum-likelihood estimation of pairwise distances and of branch lengths in a tree, both for unconstrained and various clock-like trees (Rambaut, 2000),
- simulating coalescence intervals and estimation of demographic parameters (Donnelly and Tavaré, 1995),
- performing likelihood ratio and chi-square tests and for comparison of phylogenetic hypotheses using the Kishino–Hasegawa and Shimodaira–Hasegawa tests (Goldman *et al.*, 2000),
- manipulating alignments and trees,
- simulating data sets along trees,
- optimizing uni- and multivariate functions by various methods including a Genetic Algorithm (GA), computing numerical derivatives, creating simulation-quality random numbers (Matsumoto and Nishimura, 1998), and sorting,
- creating formatted text input and output by classes extending the standard Java IO library,

*To whom correspondence should be addressed.

Table 1. Java Packages available in PAL (version 1.1)

Package	Classes	Description
pal.alignment	9	Data structures and utilities for sequence alignments
pal.coalescent	14	Modeling of population genetic processes using the coalescent
pal.datatype	6	Classes and tools for describing sequence data types
pal.distance	9	Data structures and methods to compute genetic distances
pal.eval	5	Classes for evaluating evolutionary trees and estimating parameters
pal.gui	7	GUI components for some special objects (e.g. trees)
pal.io	4	Classes to simplify Java text input/output
pal.math	19	Optimization methods, special functions, numerical derivatives, etc.
pal.misc	14	Classes that don't fit in any of the other packages
pal.statistics	11	Distributions, bootstrap estimators, tree tests, etc.
pal.substmodel	22	Classes describing models of substitution (e.g. REV, Dayhoff)
pal.tree	18	Data structures and tools for modifying and constructing trees
pal.util	7	Various utility classes (e.g. sorting)
	145	

- reconstructing neighbor-joining, UPGMA and sUPGMA trees (Drummond and Rodrigo, 2000), and estimating least-squares branch lengths on trees (weighted and unweighted LS),
- translating nucleotide to amino acid sequences, and
- accessing mathematical special functions (e.g. Γ, error, and Binomial function) and pdf, cdf, and quantile functions of statistical distributions (e.g. Γ, χ^2, exponential, Gaussian distribution).

For an exhaustive list, source code and documentation please visit the web site.

The general design of the library follows standard rules for object-oriented programming. Where possible, complex objects such as alignments and trees are represented by a simple interface class only, with additional classes for its implementation and for utility methods. Thus, any external implementations of PAL interfaces can still use the corresponding methods and infrastructure provided in PAL.

This facilitates cross-linking PAL with other packages. For example, plans exist to interface PAL with Mesquite, a modular environment for evolutionary analysis developed by Maddison and Maddison (2001). Already, PAL provides the computational engine of Vanilla, a set of command-line programs written in Java by K.S. (see PAL web page) and it is also used by the Java program PEBBLE, written by A.D., Greg Ewing and Matthew Goode (http://www.cebl.auckland.ac.nz). PAL has also been used as a test bed for new methodology, e.g. in Strimmer and Moulton (2000).

PAL is an ongoing effort, the package has been updated regularly since the start of the project in 1999. Scheduled extensions for upcoming releases of PAL are, for example, tree and model searches based on Markov Chain Monte Carlo (MCMC) and GA techniques. The project is set up as a collaborative effort, and contributions by other researchers to this repository are greatly welcome.

ACKNOWLEDGEMENTS

We thank Allen Rodrigo for discussion and for providing hardware through his NIH Grant GM59174. We also thank Matthew Goode for significant contributions. This work is supported by a New Zealand FRST Bright Futures Scholarship (A.D.) and Emmy Noether-Fellowship by the Deutsche Forschungsgemeinschaft (K.S.).

REFERENCES

Donnelly,P. and Tavaré,S. (1995) Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.*, **29**, 401–421.

Drummond,A. and Rodrigo,A.G. (2000) Reconstructing genealogies of serials samples under the assumption of a molecular clock using serial-sample UPGMA. *Mol. Biol. Evol.*, **17**, 1807–1815.

Goldman,N., Anderson,J.P. and Rodrigo,A.G. (2000) Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.*, **49**, 652–670.

Liò,P. and Goldman,N. (1998) Models of molecular evolution and phylogeny. *Genome Res.*, **8**, 1233–1244.

Maddison,W.P. and Maddison,D.R. (2001) Mesquite: a modular system for evolutionary analysis. http://mesquite.biosci.arizona.edu/mesquite/mesquite.html.

Matsumoto,M. and Nishimura,T. (1998) Mersenne Twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans. Modeling Comput. Simul.*, **8**, 3–30.

Rambaut,A. (2000) Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, **16**, 395–399.

Strimmer,K. and Moulton,V. (2000) Likelihood analysis of phylogenetic networks using directed graphical models. *Mol. Biol. Evol.*, **17**, 875–881.