Datamonkey: rapid detection of selective pressure on individual sites of codon alignments

Sergei L. Kosakovsky Pond* and Simon D. W. Frost
Antiviral Research Center, University of California, San Diego, CA, 92103, USA

ABSTRACT
Summary: Datamonkey is a web interface to a suite of cutting edge maximum likelihood-based tools for identification of sites subject to positive or negative selection. The methods range from very fast data exploration to the some of the most complex models available in public domain software, and are implemented to run in parallel on a cluster of computers.
Availability: http://www.datamonkey.org. In the future, we plan to expand the collection of available analytic tools, and provide a package for installation on other systems.
Contact: spond@ucsd.edu

1 INTRODUCTION
The detection and quantification of evolutionary pressures that have contributed to genetic variation has been an active area of recent research (Yang, 2002) and has become an accepted part of a statistical toolbox used in sequence analysis. There are several popular statistical methods for the identification of rapidly evolving and unusually conserved sites in regions of protein coding sequences, which rely upon estimating site-specific synonymous (dS) and non-synonymous (dN) substitution rate parameters, and performing statistical tests to determine whether dS ≠ dN. Two widely used methods that make use of the evolutionary history of sampled sequences are the likelihood-based approach described by Nielsen and Yang (1998) and implemented in the popular PAML package, developed by Yang (1997), and a parsimony-based counting method of Suzuki and Gojobori (1999) implemented in the ADAPTSITE program, written by Suzuki et al. (2001). Kosakovsky Pond and Frost (2005b) proposed an integrative approach, combining the strengths of both of the above approaches and offering several new algorithms as well. Datamonkey is a web-based gateway to the suite of these algorithms, executed by HyPhy, a molecular evolution analysis platform (Kosakovsky Pond et al., 2004), running analyses in parallel on a cluster of computers, with a streamlined and easy-to-use interface.

2 METHODS
Datamonkey implements three complementary methods for detecting sites under selection. All theoretical and technical aspects of the methods and performance comparison can be found in Kosakovsky Pond and Frost (2005b).

*To whom correspondence should be addressed.

Single likelihood ancestor counting (SLAC) is a heavily modified and improved derivative of the Suzuki–Gojobori counting approach. SLAC can process an alignment with 100 sequences and 400 codons in about a minute, using likelihood-based branch lengths, nucleotide and codon substitution parameters and ancestral sequence reconstructions. SLAC has good power to detect non-neutral evolution in large (>50 sequences) alignments.

Fixed effects likelihood (FEL) is a new likelihood-based and statistically rigorous method to fit an independent dN and dS to every site in the context of codon substitution models and test whether dN ≠ dS. This method has been parallelized to run quickly on an MPI cluster and tends to be less conservative than SLAC on datasets of intermediate size (20–50 sequences).

Random effects likelihood (REL) is an improved variant of the Nielsen–Yang approach, which uses flexible but not overly parameter-rich rate distributions (Kosakovsky Pond and Frost, 2005a) and allows both dS and dN to vary across sites independently. Kosakovsky Pond and Muse (2005) suggest that accounting for nucleotide substitution biases and synonymous site-to-site variation helps reduce Type I errors. This method has been parallelized to run on an MPI cluster, and while it is the most powerful of the three methods, REL is somewhat susceptible to Type I errors, especially for small datasets, where parameter estimates are likely to have large associated errors.

3 IMPLEMENTATION
The interface has been constructed using open source, public domain components such as Apache Web server (http://www.apache.org) with custom Perl CGI and HyPhy batch language scripts which perform pre-processing of uploaded alignments and post-processing of analysis results (Fig. 1) and HyPhy scripts for executing the analyses. HyPhy runs complex analyses in parallel on clusters of computers which support the MPICH (http://www-unix.mcs.anl.gov/mpi/mpich/) implementation of the message passing interface (MPI) protocol. Presently, the analyses are hosted on a Linux cluster of eight dual processor Athlon MP 1.4 GHz nodes. The implementation is completely self-contained and allows users, among other things, the following:

(1) Upload an alignment in one of the several standard data formats, such as NEXUS, PHYLIP, MEGA or FASTA. The alignment is checked for validity, including the presence of stop codons.

(2) Run a locally hosted BLAST (Altschul et al., 1997) search on the sequences to classify the organisms.

(3) Perform phylogenetic reconstruction using an efficient implementation of the neighbor joining method (Saitou and Nei, 1987) and render high-quality PDF phylograms.

(4) Invoke a model selection procedure proposed in Kosakovsky Pond and Frost (2005a) to quickly decide which evolutionary...
model is appropriate for their alignments; this procedure is unique, as it explores 203 time reversible models, rather than a limited subset of ‘named’ models.

(5) Detect which sites in the alignment evolve adaptively and those which are functionally constrained. Our methods are orders of magnitude faster than popular existing methods, running essentially interactively, while offering a more statistically robust framework for estimating confidence in inferred results Kosakovsky Pond and Frost (2005b). The user can then run: SLAC (up to 150 sequences), FEL (up to 50 sequences) or REL (up to 25 sequences) to locate sites undergoing adaptive or purifying evolution. All methods provide progress updates and intermediate results.

(6) For SLAC analyses, the user can view inferred mutations for each site, optionally, map them onto a phylogeny.

(7) Generate charts to visualize the distribution of selective pressure, and other quantities, along sequences. This feature utilizes an open source plotting package, GNUPLOT, available at http://www.gnuplot.info.

(8) If all three methods are run on a given dataset, a comparative analysis integrating the three methods can be performed.

Analysis results can be downloaded or accessed on our server for up to 96 hours.

4 DISCUSSION
We believe that the availability of fast and statistically sound methods is critical to enabling sophisticated large scale analyses of sequence evolution. Datamonkey is linked to a cluster of computers so that analyses which would take a long time to run on a desktop computer can be run quickly. The use of a website allows the tools to be kept up-to-date centrally, without the need for the researcher to install and maintain the considerable hardware resources required to run computationally intensive analyses. In the future, we intend to distribute a complete package of components needed to install and configure Datamonkey on a POSIX compliant web server, with or without an SSH interface to MPI computer clusters.

ACKNOWLEDGEMENTS
This research was supported by the National Institutes of Health (AI47745, AI43638 and AI57167), the University of California Universitywide AIDS Research Program (grant number IS02-SD-701) and by a University of California, San Diego Center for AIDS Research/NIAID Developmental Award to S.D.W.F. (AI36214).
REFERENCES


