Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology
Wayne Delport 1, Art F. Y. Poon 2, Simon D. W. Frost 3 and Sergei L. Kosakovsky Pond 4

1 Department of Pathology, Antiviral Research Center, University of California, San Diego, USA
2 British Columbia Centre for Excellence in HIV/AIDS, Vancouver, British Columbia, Canada
3 Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom
4 Department of Medicine, Antiviral Research Center, University of California, San Diego, USA

ABSTRACT
Summary: Datamonkey is a popular web-based suite of phylogenetic analysis tools for use in evolutionary biology. Since the original release in 2005, we have expanded the analysis options to include several recently developed algorithmic methods. The traditional selection tools have also been augmented to include recent developments in the field. Here we summarize the analyses options currently available on Datamonkey.
Availability and documentation: http://www.datamonkey.org
Contact: spond@ucsd.edu

1 INTRODUCTION
Recent developments of high-throughput sequencing technologies have accelerated the rate at which genomic data are accumulating. Concurrent commoditization of cheap parallel computer systems (clusters, GPUs, multi-core systems) and rapid development of algorithmic, statistical and bioinformatics techniques have made it possible to analyze these genomic data with models of increased biological realism. To make such models, developed by ourselves and other groups, immediately useful to the life sciences community, we deployed a public web service to screen alignments of homologous sequences for signatures of natural selection (Kosakovsky Pond & Frost, 2005b; Kosakovsky Pond & Frost, 2005) on a 40-processor cluster in 2005. The server proved; Kosakovsky Pond & Frost, 2005 b
alignments of homologous sequences for signatures of natural
by ourselves and other groups, immediately useful to the life
Concurrent commoditization of cheap parallel computer systems
and rapid development of algorithmic, statistical and bioinformatics
tools have also been augmented to include recent developments
in the field. Here we summarize the analyses options currently
available on Datamonkey.
Availability and documentation: http://www.datamonkey.org
Contact: spond@ucsd.edu

1 INTRODUCTION
Recent developments of high-throughput sequencing technologies have accelerated the rate at which genomic data are accumulating. Concurrent commoditization of cheap parallel computer systems (clusters, GPUs, multi-core systems) and rapid development of algorithmic, statistical and bioinformatics techniques have made it possible to analyze these genomic data with models of increased biological realism. To make such models, developed by ourselves and other groups, immediately useful to the life sciences community, we deployed a public web service to screen alignments of homologous sequences for signatures of natural selection (Kosakovsky Pond & Frost, 2005b; Kosakovsky Pond & Frost, 2005) on a 40-processor cluster in 2005. The server proved to be popular, processing over 100,000 submitted jobs, many of which would require days or weeks of desktop CPU time. Since the original release, we have redesigned the user interface, upgraded our cluster to 356 CPU cores, implemented 12 new analytical modules and new result processing and visualization features. Improvements to HyPhy (Kosakovsky Pond et al., 2005) core algorithms have resulted in significant (up to 10×) speedups and allowed us to increase the size limits of alignments.

2 METHODS
Natural selection
Diversifying and purifying selection acting on sites. Datamonkey was originally designed to provide a front end to an implementation of three approaches (SLAC, FEL and REL; (Kosakovsky Pond & Frost, 2005b; Kosakovsky Pond & Frost, 2005)) to identifying purifying or diversifying selection. These and nearly all other methods now correct for the confounding effect of recombination through partitioning the alignment (e.g. using GARD) into non-recombinant fragments, each with a separate phylogeny (Scheffler et al., 2006). The new PARRIS test for non-neutral evolution is analogous to the original test of Nielsen & Yang (1998), but corrects for the confounding effect of recombination.
“Population level” selection. IFEL identifies selective pressures that are restricted to interior branches of the tree, e.g. as described in the context of population-level HIV-1 evolution in Kosakovsky Pond et al. (2006).
Lineage specific selection. GABranch executes a genetic algorithm (GA) search for lineages that are subject to differing mean selective pressures (Kosakovsky Pond & Frost, 2005a), assuming selection acts uniformly across sites.
Directional evolution of protein sequences. DEPS (Kosakovsky Pond et al., 2008) reports whether or not there is evidence that some sites are evolving towards each of the 20 amino-acid residues. Empirical Bayes analysis is used to identify directionally evolving sites.
“Toggling” selection. TOGGLE (Delport et al., 2008) searches a user-specified subset of sites for evidence of elevated rates of substitution away from and back to a wildtype residue, typically observed in host-pathogen systems where pathogens frequently escape host recognition, yet with a fitness cost.

Recombination Detection.
Recombination detection now includes tests for topological incongruence to separate recombination from heterotachy. GARD-type methods can also detect recombination in a single sequence by screening against a reference alignment (Kosakovsky Pond et al., 2009). SCUEAL currently implements this approach for HIV-1 subtyping based on the pol gene.

Model selection
Protein model selection. Given an alignment, we rank 14 empirical amino-acid substitution models using AIC, AICc and BIC, similar to the logic of ProtTest (Abascal et al., 2005).
Codon model selection. CMS is a statistical approach to partition all pairwise non-synonymous substitution rates into groups (Delport et al., view), and to search for well-fitting models using a computationally feasible and accurate GA.
Evolutionary Fingerprinting.

**EVOBLAST** provides an implementation of the gene evolutionary fingerprinting procedure described in Kosakovsky Pond et al. (2010). The primary purpose of **EVOBLAST** is to enable the comparison of inferred evolutionary properties between non-homologous genes using Evolutionary Selection Distance.

**Ancestral state reconstruction: ASR.**

Three different likelihood-based methods are used to recover and assess the robustness of ancestral sequences: joint likelihood (Pupko et al., 2000), marginal likelihood (Yang et al., 1995) and sampling of the joint posterior distribution of ancestral characters (Nielsen, 2002).

**Co-evolution between sites**

**Spidermonkey** (Poon et al., 2008) uses Bayesian networks and is geared towards identifying interacting sites in an alignment, based upon the assumption that co-evolving sites will tend to acquire mutations along the same set of branches.

3 IMPLEMENTATION

Datamonkey is implemented as a collection of Perl, HyPhy batch language and R scripts, with GnuPlot, GraphViz and GhostScript used for visualization. Data upload, CGI processing, SLAC analyses and result visualization is handled by a dedicated Mac OS X server, while all the other analyses are executed on a 356-core Linux Beowulf (SCYLD) cluster, as serial or MPI jobs. There are method-group FIFO queues to schedule submissions. Communication between the two systems is performed via SSH tunneling.

4 DISCUSSION

The ever-accelerating pace of methodological research and development places a premium on resources that avail computational and evolutionary biologists and bioinformaticians of fast, maintained and documented modern tools with a consistent and easy-to-use interface. As evidenced by the popularity of the original Datamonkey server, our approach of providing a web-based front end for running computationally intensive statistical sequence analysis tools on a large computer cluster continues to be well-received by the community and we fully intend to develop and extend the functionality of the service as new procedures and analyses are introduced.

ACKNOWLEDGEMENTS

**Funding:** This research was supported by the Joint Division of Mathematical Sciences/National Institute of General Medical Sciences Mathematical Biology Initiative through Grant NSF-0714991, the National Institutes of Health (AI43638, AI47745, and AI57167), the University of California University wide AIDS Research Program (grant number IS02-SD-701), a University of California, San Diego Center for AIDS Research/NIAID Developmental Award to SDWF, SLKP & WD (AI36214). SDWF is supported in part by a Royal Society Wolfson Research Merit Award. AFYP is funded by a Canadian Institutes of Health Research (CIHR) Fellowships Award in HIV/AIDS Research (200802HFE).

**Conflict of Interest:** none declared

REFERENCES


