BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 19 2010, pages 2455–2457
doi:10.1093/bioinformatics/btq429

Sequence analysis

Advance Access publication July 29, 2010

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, CA, USA, [2]British Columbia Centre for Excellence in HIV/AIDS, Vancouver, British Columbia, Canada, [3]Department of Veterinary Medicine, University of Cambridge, Cambridge, UK and [4]Department of Medicine, Antiviral Research Center, University of California, San Diego, CA, USA

Associate Editor: David Posada

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 recently developed algorithmic methods for recombination detection, evolutionary fingerprinting of genes, codon model selection, co-evolution between sites, identification of sites, which rapidly escape host-immune pressure and HIV-1 subtype assignment. 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, and provide guidelines for their use in evolutionary biology.

Availability and documentation: http://www.datamonkey.org

Contact: spond@ucsd.edu

Received and revised on June 15, 2010; accepted on July 20, 2010

1 INTRODUCTION

Recent developments of high-throughput sequencing technologies have accelerated the rate at which genomic data are accumulating by orders of magnitude. Concurrent commoditization of cheap parallel computer systems (clusters, GPUs and 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 using three different phylogenetic methods (Kosakovsky Pond and Frost, 2005a; Kosakovsky Pond and Frost, 2005c) 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 completely redesigned the user interface, upgraded our cluster to one with 356 CPU cores, implemented 12 new analytical modules and a plethora of result processing and visualization features. Improvements to core algorithms in the HyPhy package (Kosakovsky Pond *et al.*, 2005), have resulted in significant (up to 10×) speedups and allowed us to increase the sizes of alignments that can be submitted.

*To whom correspondence should be addressed.

2 METHODS

2.1 Natural selection

2.1.1 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 and Frost, 2005a; Kosakovsky Pond and Frost, 2005c) to finding the sites in a multiple sequence alignment, which may have been affected by purifying or diversifying selection. These and nearly all other methods have been upgraded to correct for the confounding effect of recombination using the partitioning approach, whereby the alignment is partitioned (computationally, e.g. using **GARD**) into non-recombinant fragments, and each one of those is endowed with a separate phylogeny (Scheffler *et al.*, 2006). Result processing allows users to visualize and report the distribution of inferred substitutions on a site-by-site basis. The new **PARRIS** module furnishes a likelihood ratio test (LRT) for non-neutral evolution that is analogous to the original test of Nielsen and Yang (1998), but corrects for the confounding effect of recombination and permits synonymous substitution rates to vary from site to site.

2.1.2 'Population level' selection using **iFEL** When one is interested in 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.* (2006a), the **iFEL** (internal branches FEL) method is appropriate.

2.1.3 Lineage specific selection using **GABranch** This component executes a genetic algorithm (GA) search for lineages that are subject to differing mean selective pressures (Kosakovsky Pond and Frost, 2005b). Instead of addressing 'where in the gene has selection acted?' question that the previous tools are designed for, this analysis answers 'when in the past has selection acted?' question, assuming that selection acts uniformly across sites.

2.1.4 Directional evolution of protein sequences using **DEPS** In Kosakovsky Pond *et al.* (2008), we proposed a model-based test for directional evolution in protein sequences, capable of identifying such frequency changes, or, more generally, deviations from the 'background' substitution patterns that favor substitutions towards a particular residue. Given an amino-acid alignment and a rooted phylogenetic tree, **DEPS** reports whether or not there is evidence that a proportion of sites are evolving towards each of the 20 amino-acid residues. For those 'target' residues that pass this test, **DEPS** carries out an empirical Bayes analysis to pinpoint which sites may be directionally evolving towards a given residue, along with a heuristic interpretation of the type of selection that could have caused the inferred pattern of substitutions.

2.1.5 'Toggling' selection using **TOGGLE** The best example of toggling selection can be found in HIV-1 sequences, which can acquire mutations in

one host, e.g. in response to immune selection or drug therapy, and revert these mutations following subsequent transmissions to hosts that are not on treatment or do not raise the selecting immune response. Using the approach of Delport *et al.* (2008), **TOGGLE** searches a subset of sites identified by the user (e.g. based on inferred substitution patterns or association with immune targets) for evidence of elevated rates of substitution away from and back to the unknown wildtype residue. At every analyzed site, all possible wildtype residues are examined (several different wildtype residues can be consistent with an evolutionary history of a site), and those which return an (corrected) LRT $P < 0.05$ are reported. Visualization tools are available to assist in interpreting the patterns of substitutions at a site, and the evolutionary pathways between residues.

2.2 Recombination detection: *GARD* and *SCUEAL*

GAs for recombination detection (**GARD**) are a highly sensitive and accurate approach for screening alignments for evidence of phylogenetically incongruent segments (Kosakovsky Pond *et al.*, 2006b). Since the original release (Kosakovsky Pond *et al.*, 2006c), the **GARD** module in Datamonkey has been significantly upgraded, e.g. to automatically perform Kishino–Hasegawa tests for topological incongruence and compute Robinson–Foulds distances between conflicting topologies. This step helps tease apart the two most common causes of phylogenetic incongruence: recombination and heterotachy. A specialized refinement of **GARD** can be used for detecting recombination in a single sequence by screening against a reference alignment with a precomputed phylogeny (Kosakovsky Pond *et al.*, 2009). This type of analysis is most commonly used to infer the recombination or reassortment history of HIV-1 or Influenza A virus strains, and forms the basis of genetically delineated viral subtypes. The **SCUEAL** module currently implements HIV-1 subtyping based on the most frequently sequenced *pol* gene and is capable of processing several hundred sequences per hour.

2.3 Model selection

2.3.1 Protein model selection We have implemented a simple model selection procedure to rank 14 empirical amino acid substitution models (this list is regularly updated) using *AIC*, AIC_c and *BIC*, similar to the ideas of ProtTest (Abascal *et al.*, 2005). For each model, a version with published stationary frequencies and another (+F) with frequencies tabulated from the alignment under consideration are evaluated.

2.3.2 Codon model selection using **CMS** The problem of properly modeling mechanistic (synonymous versus non-synonymous) and empirical (the dependence of substitution rates on the amino acids encoded by the source and target codons) components of codon-based evolution is computationally challenging, as there are combinatorially many possible codon models. In Delport *et al.* (2010), we have described a statistical approach to partition all pairwise substitution rates into groups, akin to how, for example the HKY85 (Hasegawa *et al.*, 1985) model partitions nucleotide substitutions into transitions and transversions, and to search for well-fitting models of this type using a computationally feasible and accurate GA. The **CMS** analysis reports the number and membership of non-synonymous rate classes. Using multi-model based inference, **CMS** generates substitution rate profiles for each residue pair, determines the confidence with which each pair is allocated to a rate class and computes correlations between substitution rates and physico-chemical properties. We are currently developing a database with thousands of gene- and organism-specific codon evolutionary models to assist the users in selecting an appropriate evolutionary model for their alignments.

2.4 Evolutionary fingerprinting using *EVOBLAST*

The **EVOBLAST** module provides an implementation of the gene evolutionary fingerprinting procedure described in Kosakovsky Pond *et al.* (2010). It first fits a flexible generate bivariate distribution of synonymous and non-synonymous substitution rates to a coding sequence alignment (the appropriate number of rate classes determined automatically). An approximate posterior sample of the inferred rates is obtained and converted to an evolutionary fingerprint. Site-by-site inference of positive selection using this posterior sample is analogous to the Bayes Empirical Bayes (Yang *et al.*, 2005) approach that attempts to account for the errors in estimated model parameters. However, the primary purpose of **EVOBLAST** is to enable the comparison of inferred evolutionary properties between genes using Evolutionary Selection Distance, as described in the methodology paper. We are currently developing the functionality to compare users' alignments against a database of annotated (e.g. taxonomically and functionally) fingerprints and permit the users to add their own alignments to the database. In this fashion, it may be possible to create a large database of evolutionary properties of many genes sampled from different taxonomic levels to power quantitative comparisons of non-homologous sequence data.

2.5 Ancestral state reconstruction (ASR)

The **ASR** module accepts a partitioned alignment, provided, e.g. by **GARD**. Three different likelihood-based methods are used to recover ancestral sequences. First, the joint likelihood method finds the assignment of ancestral characters to maximize the likelihood over all such assignments (Pupko *et al.*, 2000). Second, for each site and ancestral sequence, the marginal likelihood method computes posterior weights for each ancestral character by marginalizing over all other ancestral characters (Yang *et al.*, 1995). Third, 100 samples are drawn from the joint posterior distribution of ancestral characters (Nielsen, 2002). Three ancestral sequences (one for each method) present in the strict consensus tree of all ancestral segments are returned, together with a report highlighting agreement and discrepancies between the methods.

2.6 Co-evolution between sites: *Spidermonkey*

The Spidermonkey module (Poon *et al.*, 2008) uses Bayesian network techniques and is geared towards identifying networks of interacting sites in an alignment, based upon the assumption that co-evolving sites will tend to acquire mutations along the same set of branches. Repeated inference with ancestral states sampled from the posterior distribution is useful to evaluate robustness.

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, either 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.

Funding: Joint Division of Mathematical Sciences/National Institute of General Medical Sciences Mathematical Biology Initiative through Grant NSF-0714991; National Institutes of Health (AI43638, AI47745 and AI57167); the University of California University wide AIDS Research Program (grant number IS02-SD-701); University of California, San Diego Center for AIDS Research/NIAID Developmental Award (AI36214 to S.D.W.F., S.L.K.P. and W.D.); Royal Society Wolfson Research Merit Award (in part to S.D.W.F.); Canadian Institutes of Health Research (CIHR) Fellowships Award in HIV/AIDS Research (200802HFE) (to A.F.Y.P.); The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: none declared.

REFERENCES

Abascal,F. *et al.* (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, **21**, 2104–2105.

Delport,W. *et al.* (2008) Frequent toggling between alternative amino acids is driven by selection in HIV-1. *PLoS Pathog.*, **4**, e1000242.

Delport,W. *et al*. (2010) CodonTest: modeling amino-acid substitution preferences in coding sequences. *PLoS Compt. Biol.*, **6**, e1000885.

Hasegawa,M. *et al.* (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Mol. Biol. Evol.*, **21**, 160–174.

Kosakovsky Pond,S.L. and Frost,S.D.W. (2005a) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, **21**, 2531–2533.

Kosakovsky Pond,S.L. and Frost,S.D.W. (2005b) A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol. Biol. Evol.*, **22**, 478–485.

Kosakovsky Pond,S.L. and Frost,S.D.W. (2005c) Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.*, **22**, 1208–1222.

Kosakovsky Pond,S.L. *et al.* (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, **21**, 676–679.

Kosakovsky Pond,S.L. *et al.* (2006a) Adaptation to different human populations by HIV-1 revealed by codon-based analyses. *PLoS Comp. Biol.*, **2**, e62.

Kosakovsky Pond,S.L. *et al.* (2006b) Automated phylogenetic detection of recombination using a genetic algorithm. *Mol. Biol. Evol.*, **23**, 1891–1901.

Kosakovsky Pond,S.L. *et al.* (2006c) GARD: a genetic algorithm for recombination detection. *Bioinformatics*, **22**, 3096–3098.

Kosakovsky Pond,S.L. *et al.* (2008) A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza a virus. *Mol. Biol. Evol.*, **25**, 1809–1824.

Kosakovsky Pond,S.L. *et al.* (2009) An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. *PLoS Comput. Biol.*, **5**, e1000581.

Kosakovsky Pond,S.L. *et al.* (2010) Evolutionary fingerprinting of genes. *Mol. Biol. Evol.*, **27**, 520–536.

Nielsen,R. (2002) Mapping mutations on phylogenies. *Syst. Biol.*, **51**, 729–739.

Nielsen,R. and Yang,Z.H. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, **148**, 929–936.

Poon,A.F.Y. *et al.* (2008) Spidermonkey: rapid detection of co-evolving sites using bayesian graphical models. *Bioinformatics*, **24**, 1949–1950.

Pupko,T. *et al.* (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol.*, **17**, 890–896.

Scheffler,K. *et al.* (2006) Robust inference of positive selection from recombining coding sequences. *Bioinformatics*, **22**, 2493–2499.

Yang,Z. *et al.* (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, **141**, 1641–1650.

Yang,Z. *et al.* (2005) Bayes Empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.*, **22**, 1107–1118.