Benchmarks

DIVEIN: a web server to analyze phylogenies, sequence divergence, diversity, and informative sites

Wenjie Deng[1], Brandon S. Maust[1], David C. Nickle[1,*], Gerald H. Learn[1,**], Yi Liu[1], Laura Heath[1], Sergei L Kosakovsky Pond[2], and James I. Mullins[1]
[1]Department of Microbiology, University of Washington School of Medicine, Seattle, WA and [2]Department of Medicine, University of California, San Diego, CA, USA

BioTechniques 48:405-408 (May 2010) doi 10.2144/000113370
Keywords: phylogeny; divergence; diversity; informative sites; center of tree; maximum likelihood

*D.C.N.'s current address is Rosetta Inpharmatics LLC, Seattle, WA 98109, USA.
**G.H.L.'s current address is University of Alabama at Birmingham, Birmingham, AL 35294, USA.

DIVEIN is a web interface that performs automated phylogenetic and other analyses of nucleotide and amino acid sequences. Starting with a set of aligned sequences, DIVEIN estimates evolutionary parameters and phylogenetic trees while allowing the user to choose from a variety of evolutionary models; it then reconstructs the consensus (CON), most recent common ancestor (MRCA), and center of tree (COT) sequences. DIVEIN also provides tools for further analyses, including condensing sequence alignments to show only informative sites or private mutations; computing phylogenetic or pairwise divergence from any user-specified sequence (CON, MRCA, COT, or existing sequence from the alignment); computing and outputting all genetic distances in column format; calculating summary statistics of diversity and divergence from pairwise distances; and graphically representing the inferred tree and plots of divergence, diversity, and distance distribution histograms. DIVEIN is available at http://indra.mullins.microbiol.washington.edu/DIVEIN.

Fast and accurate estimation of phylogenies and determination of genetic and phylogenetic divergence and diversity of molecular sequences are essential components of biological research. For a set of sequences, a typical phylogenetic analysis involves several steps, including multiple sequence alignment, phylogenetic reconstruction, visualization of the inferred tree, and calculation of evolutionary measures. A large number of phylogenetic analysis resources have been developed, as cataloged by Joseph Felsenstein (http://evolution.genetics.washington.edu/phylip/software.html), including web servers that provide an easy route to address specific evolutionary questions. For example, PhyML Online (1) performs maximum likelihood (ML) phylogenetic estimation under a wide range of evolutionary models. Phylemon (2) provides experts with a suite of online programs and a Java interface to build a phylogeny pipeline. Dereeper et al. recently made available Phylogeny.fr (3), which boasts an easy-to-use interface designed for the non-specialist combined with up-to-date programs that are frequently reserved for experts.

These tools provide excellent interfaces to phylogenetic reconstruction; however, there is an increasing demand by researchers for a tool that performs not only typical phylogenetic reconstructions, which most existing web servers do capably, but also enables downstream processing and interpretation. For example, calculating divergence and diversity measurements and genetic distance distributions from the phylogenetic output are usually very time-consuming processes that require caution if conducted manually to ensure that calculations are carried out correctly and that data has not been altered in the transfer among the several necessary software packages. Furthermore, reducing an alignment to only its phylogenetically informative sites—a position at which there are at least two different character states and each of those states occurs in at least two of the sequences—has proven to be a useful approach in recombination analysis (4–6) and visualizing extended alignments. Calculation of central sequences and comparison of a set of sequences to a consensus (CON), most recent common ancestor (MRCA), or center of tree (COT; an ancestral state that minimizes the phylogenetic distance from the specified sequences) (7–9) have been used in a variety of studies of sequence evolution, structure, function, and rational vaccine design.

The need for a unified web interface to integrate useful tools and perform automated phylogenetic and other genetic analyses (including summaries and visualization of the resulting data) led us to develop DIVEIN, which has four major components: (*i*) a pipeline to automatically guide a set of aligned sequences through phylogenetic tree estimation under a variety of evolutionary models, and visualization of the inferred tree; (*ii*) an interface to reconstruct MRCA/COT/CON sequences and reconstruct and visualize trees re-rooted by MRCA and COT sequences; (*iii*) calculation of genetic distance distributions, pairwise diversity and divergence from the MRCA/COT/CON; and (*iv*) an interface to detect, visualize, and numerically summarize phylogenetically informative sites as well as private mutations (found only in a single sequence) in an alignment.

DIVEIN runs on an Apache web server. The web interfaces are implemented via Perl CGI and JavaScript. Data manipulation and presentation employ standard Perl and BioPerl (10) modules. Maximum likelihood phylogenetic reconstructions use PhyML v3.0 (11), which applies a hill-climbing algorithm that adjusts tree topology and branch lengths simultaneously. The inferred tree can be viewed and edited through the included Archaeopteryx v0.955 β Java applet (www.phylosoft.org/archaeopteryx). The MRCA and COT sequences are reconstructed using a

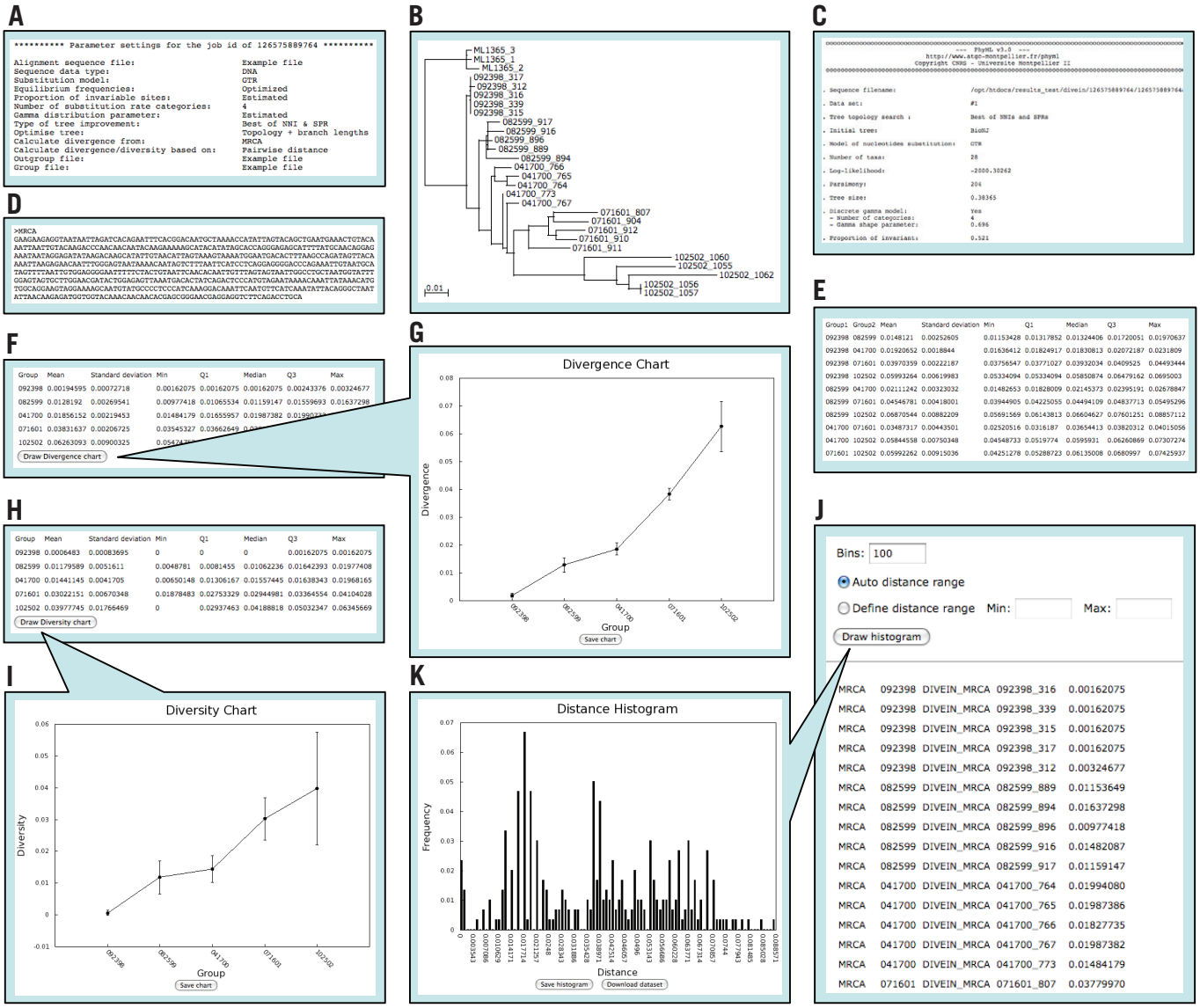

Figure 1. Screen shots of phylogeny/divergence/diversity output in DIVEIN. (A) Parameter settings for running the program. (B) Phylogenetic tree viewed through the Archaeopteryx Java applet. (C) Estimated evolutionary parameters. (D) Reconstructed MRCA sequence. (E) Summarized distances between groups. (F) Summarized divergence from MRCA for each group. (G) Divergence plot generated by clicking the Draw divergence chart button. (H) Summarized diversity for each group. (I) Diversity plot generated by clicking the Draw diversity chart button. (J) Distance matrix. (K) Distance distribution histogram generated by clicking the Draw histogram button.

joint maximum likelihood procedure (12) via HyPhy v2.0 (13), a scriptable software package for performing a wealth of evolutionary sequence analyses. Distance distribution histograms and divergence and diversity plots are generated using the open source Gnuplot graphing package (www.gnuplot.info). DIVEIN is hosted on a Linux computer with two quad-core Intel Xeon 2.5 GHz processors (8 cores) and 8 GB RAM. It is configured to run up to eight user-submitted projects simultaneously, with additional projects queued for later execution. Bootstrap replicates are limited to 100 because of computational resource limitations.

Given a collection of sequences, the divergence is derived by calculating the mean distance of all sequences from a reference or founder sequence and the diversity is given as the mean distance between all sequences (14). Using $d(i,j)$ to denote either the path length between nodes i and j in the reconstructed phylogenetic tree or a genetic distance between sequences i and j, we measure diver-

gence and diversity for a collection of N sequences as follows:

$$D_{divergence} = \frac{1}{N}\sum_{i=1}^{N} d(i, founder/reference)$$

[Eq. 1]

$$D_{diversity} = \frac{1}{N(N-1)}\sum_{\substack{i,j \\ i \neq j}}^{N} d(i,j)$$

[Eq. 2]

DIVEIN accepts aligned nucleotide or amino acid sequences in NEXUS, PHYLIP, or FASTA format. For phylogenetic analyses, users can perform ML estimation alone or include divergence/diversity analyses. They can calculate divergence from MRCA, COT, and/or CON, or any sequence in the alignment [MRCA calculations require a file listing sequence name(s) that belong to the outgroup]. Users can optionally provide a file that assigns input sequences to multiple groups and calculate divergence and diversity for each of those groups. If a group file is not provided, DIVEIN will assign all sequences to a single group, excluding the defined outgroup sequences. For COT analysis, users may upload a tree to reconstruct its COT. If the tree is not provided, DIVEIN will estimate one using either the general time reversible (GTR) (15) substitution model (for nucleotides) or LG, an improved general amino acid replacement matrix (16).

We have also included an informative sites module in DIVEIN that is useful for condensing sequence data to allow users to quickly identify sites that are changing within an alignment and more easily obtain an overview of complex and large data sets. To detect phylogenetically informative sites (those found in more than one sequence, and thus contributing to branch ordering), users can include a reference sequence at the top of the alignment, or DIVEIN will calculate the consensus of the alignment as the reference. Example data sets are provided to familiarize users with the correct input formats and expected output results. DIVEIN also provides the functionality to retrieve finished results via a previously assigned project ID.

When an analysis is finished, a randomly generated URL known only to the user initiating the analysis is sent to the user by email in order to view and download results, which are accessible on the server for 2 days. Users can locally view and edit phylogenetic trees and dynamically generate and download graphs of distance distribution histograms and divergence and diversity (if applicable). Sample screen shots of DIVEIN output (phylogeny/divergence/diversity) are shown in Figure 1. Using an example alignment of 28 DNA sequences with 624 sites (available on the DIVEIN web site), it takes <30 s to finish the entire analysis process. For the analysis of phylogenetically informative sites, the states at each informative site are displayed as an alignment and in a table.

In conclusion, DIVEIN performs fast, accurate, and automated phylogenetic analyses, including (*i*) informative sites detection, (*ii*) ML tree estimation under a variety of evolutionary models, (*iii*) MRCA, COT, and CON recon-

struction, (*iv*) distance distribution calculation, and (*v*) distance- and phylogenetic-based divergence and diversity measurements, along with resulting data summarization and visualization. Future versions will add the option to select the best-fit evolutionary model via ModelTest (17) and ProtTest (18) to reconstruct the phylogeny. Furthermore, we will incorporate other widely used phylogenetic analysis programs [e.g., MrBayes (19)] into DIVEIN to allow users easy access to other state-of-the-art molecular evolution analysis programs.

Acknowledgments

We thank John E. Mittler for discussions. This work was supported by grants from the US Public Health Services (grant nos. AI047734 and AI057005), including support to the Computational Biology Core of the University of Washington Center for AIDS Research (grant no. AI27757).

Competing interests

The authors declare no competing interests.

References

1. **Guindon, S., F. Lethiec, P. Duroux, and O. Gascuel.** 2005. PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. *33*:W557-W559.
2. **Tárraga, J., I. Medina, L. Arbiza, J. Huerta-Cepas, T. Gabaldón, J. Dopazo, and H. Dopazo.** 2007. Phylemon: a suite of web tools for molecular evolution, phylogenetics and phylogenomics. Nucleic Acids Res. *35*:W38-W42.
3. **Dereeper, A., V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J.F. Dufayard, S. Guindon, et al.** 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. *36*:W465-W469.
4. **Eshleman, S.H., M.J. Gonzales, G. Becker-Pergola, S.C. Cunningham, L.A. Guay, J.B. Jackson, and R.W. Shafer.** 2002. Identification of Ugandan HIV type 1 variants with unique patterns of recombination in pol involving subtypes A and D. AIDS Res. Hum. Retroviruses *18*:507-511.
5. **Gottlieb, G.S., L. Heath, D.C. Nickle, K.G. Wong, S.E. Leach, B. Jacobs, S. Gezahegne, A.B. van 't Wout, et al.** 2008. HIV-1 variation before seroconversion in men who have sex with men: analysis of acute/early HIV infection in the multicenter AIDS cohort study. J. Infect. Dis. *197*:1011-1015.
6. **Campbell, M.S., G.S. Gottlieb, S.E. Hawes, D.C. Nickle, K.G. Wong, W. Deng, T.M. Lampinen, N.B. Kiviat, and J.I. Mullins.** 2009. HIV-1 superinfection in the antiretroviral therapy era: are seroconcordant sexual partners at risk? PLoS One *4*:e5690.
7. **Nickle, D.C., M.A. Jensen, G.S. Gottlieb, D. Shriner, G.H. Learn, A.G. Rodrigo, and J.I. Mullins.** 2003. Consensus and ancestral state HIV vaccines. Science *299*:1515-1518.
8. **Nickle, D.C., M. Rolland, M.A. Jensen, S.L. Pond, W. Deng, M. Seligman, D. Heckerman, J.I. Mullins, and N. Jojic.** 2007. Coping with viral diversity in HIV vaccine design. PLOS Comput. Biol. *3*:e75.
9. **Rolland, M., M.A. Jensen, D.C. Nickle, J. Yan, G.H. Learn, L. Heath, D. Weiner, and J.I. Mullins.** 2007. Reconstruction and function of ancestral center-of-tree human immunodeficiency virus type 1 proteins. J. Virol. *81*:8507-8514.
10. **Stajich, J.E., D. Block, K. Boulez, S.E. Brenner, S.A. Chervitz, C. Dagdigian, G. Fuellen, J.G. Gilbert, et al.** 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. *12*:1611-1618.
11. **Guindon, S. and O. Gascuel.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. *52*:696-704.
12. **Pupko, T., I. Pe'er, R. Shamir, and D. Graur.** 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol. Biol. Evol. *17*:890-896.
13. **Pond, S.L., S.D. Frost, and S.V. Muse.** 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics *21*:676-679.
14. **Shankarappa, R., J.B. Margolick, S.J. Gange, A.G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C.R. Rinaldo, et al.** 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J. Virol. *73*:10489-10502.
15. **Tavare, S.** 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Am. Math. Soc. *17*:57-86.
16. **Le, S.Q. and O. Gascuel.** 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. *25*:1307-1320.
17. **Posada, D. and K.A. Crandall.** 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics *14*:817-818.
18. **Abascal, F., R. Zardoya, and D. Posada.** 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics *21*:2104-2105.
19. **Huelsenbeck, J.P. and F. Ronquist.** 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics *17*:754-755.

Received 5 June 2009; accepted 16 March 2010.

Address correspondence to James I. Mullins, Department of Microbiology, University of Washington School of Medicine, Seattle, WA 98195, USA. e-mail: jmullins@u.washington.edu