BMC Evolutionary Biology

SOFTWARE

Open Access

PhyDesign: an online application for profiling phylogenetic informativeness

Francesc López-Giráldez* and Jeffrey P Townsend*

Abstract

Background: The rapid increase in number of sequenced genomes for species across of the tree of life is revealing a diverse suite of orthologous genes that could potentially be employed to inform molecular phylogenetic studies that encompass broader taxonomic sampling. Optimal usage of this diversity of loci requires user-friendly tools to facilitate widespread cost-effective locus prioritization for phylogenetic sampling. The Townsend (2007) phylogenetic informativeness provides a unique empirical metric for guiding marker selection. However, no software or automated methodology to evaluate sequence alignments and estimate the phylogenetic informativeness metric has been available.

Results: Here, we present PhyDesign, a platform-independent online application that implements the Townsend (2007) phylogenetic informativeness analysis, providing a quantitative prediction of the utility of loci to solve specific phylogenetic questions. An easy-to-use interface facilitates uploading of alignments and ultrametric trees to calculate and depict profiles of informativeness over specified time ranges, and provides rankings of locus prioritization for epochs of interest.

Conclusions: By providing these profiles, PhyDesign facilitates locus prioritization increasing the efficiency of sequencing for phylogenetic purposes compared to traditional studies with more laborious and low capacity screening methods, as well as increasing the accuracy of phylogenetic studies. Together with a manual and sample files, the application is freely accessible at http://phydesign.townsend.yale.edu.

Background

Due to advances in sequencing technologies and decreasing costs, the number of genomes sequenced for species across the tree of life is increasing dramatically. Tools and databases for selecting single-copy orthologous loci [1,2] and designing successful primers for them [3,4] are available. However, orthology assessments for multiple genomes can provide thousands of candidate loci to sequence, and yet only a few of those have been commonly used as markers for phylogenetic studies [5]. To address the challenge of locus selection for sequencing in designing a phylogenetic study, Townsend [6] proposed a metric that provides a quantitative prediction of phylogenetic signal across historical times. Based on estimates of rates across sites, the phylogenetic informativeness metric facilitates prioritization of loci even when the taxa of interest have never been sequenced for a given locus. To estimate phylogenetic informativeness, prior data on the molecular evolutionary pattern of a locus is required. This prior information may be derived from three potential sources: 1) preliminary data on the candidate loci from a well-studied subset of the taxa of interest; 2) data on the candidate loci from a well-studied sister clade; or 3) comparative genomic data from sequenced genomes within and/or outside the clade of interest. Using this information for a number of classic and recent data sets for different time scales, the method successfully recapitulates the qualitative utility of loci [6-10]. However, no software or automated methodology to apply the phylogenetic informativeness metric has been available.

To facilitate wider application of phylogenetic informativeness analysis, we have developed the online application PhyDesign. PhyDesign features:

- Online accessibility, platform independence, and immediate access to software updates.

* Correspondence: francesc.lopez@gmail.com; jeffrey.townsend@yale.edu
Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut, USA

- User-friendly graphical interface with rich process feedback.
- Multiple locus analyses of amino acid and DNA alignments and multiple epoch integration.
- High quality graphical outputs in SVG format easily edited for final publication.
- Spreadsheet outputs including site rate estimates and phylogenetic informativeness values in rank-order of priority for the epochs of interest, as well as tables to reproduce the profiles for further processing.

PhyDesign is freely accessible online at http://phydesign.townsend.yale.edu. A manual and sample files to be used with the application, as well as a FAQ section, can be also found at the site. In addition, the source code and a Perl module to calculate the profiles are available for download.

Implementation

To estimate phylogenetic informativeness profiles [6], the PhyDesign application consists of 3 components:

1. A form to upload information and choose an application to calculate the evolutionary rates for each alignment site.
2. A table listing the evolutionary rate results.
3. A graphical interface to plot the phylogenetic profiles and calculate integration values.

On the client-side, these components of PhyDesign are coded in JavaScript/AJAX using jQuery libraries, facilitating dynamic communication between the server and the user without page reloading. Online input undergoes validation to detect possible errors before engaging the next step in the process. To communicate validation success or failure and its causes, unobtrusive temporary message boxes are featured. In addition, the web site uses modal windows to provide optional advanced settings. On the server-side, PhyDesign is implemented as a collection of Perl scripts and modules, integrated with third-party software used to calculate the evolutionary rates for each site in the alignment.

To run, PhyDesign requires at minimum an internet browser equipped with JavaScript, and SVG for the graphical output. The latest versions of major modern browsers have native support for JavaScript and for rendering SVG markup directly. Internet Explorer 8 and older versions require a plug-in to render SVG content. SVG format, as vector graphics format, is resolution-independent and easy to edit for final publication with common vector graphic editors such as Adobe Illustrator (commercially available) and Inkscape (open source). The online application has been tested as rendered by Internet Explorer, Firefox, Chrome and Safari. Although graphical representations of phylogenetic informativeness profiles are typically best kept to lower numbers of loci for purposes of display, PhyDesign was able to profile and produce quantitative results for 2000 loci of a length of 1000 bp each. For a higher volume of data, a Perl module to calculate the profiles is available for download.

Input form

To obtain the site rate distributions for each locus, two data entries are needed: (1) an alignment of loci of interest, pruned to contain a set of taxa for which the tree topology is fairly well known, and (2) an ultrametric tree for those taxa. The ultrametric tree can be a chronogram (an ultrametric tree with branch lengths proportional to time). Alternatively, branch lengths can be in unspecified molecular evolutionary units. PhyDesign reads amino acid and DNA alignments in NEXUS, FASTA, and Phylip formats, and trees in NEXUS and Newick formats. Formats are carefully checked via custom and BioPerl modules [11]. Data partitioning is permitted via NEXUS format. Once files are uploaded and checked, PhyDesign provides an interface for selecting and parameterizing any one of several third-party programs that estimate the site rates - HyPhy [12], DNArates [Olsen, unpublished] and Rate4site [13]. DNArates and Rate4site are called by the application "as is"; a *de novo* HyPhy script was implemented to estimate rates under the full diversity of time-reversible models. Due to the ability to specify of the evolutionary model and its parameters, HyPhy is the recommended and default option for DNA sequences. For amino acid sequences, Rate4site is available. To facilitate extensive analyses of large datasets, a link to the rates results is sent by email. A site rate file with the proper formatting will be generated after analysis of the alignments and the corresponding ultrametric tree. Supplied alignments are deleted from the server immediately after the site rates are calculated, and the site rate results are stored in the server for 24 hours. During this time, results can only be accessed by the link submitted. It is convenient, then, to save the site rate file for future use. Alternatively, if the rate distribution for each locus is known, PhyDesign also accepts direct supply of these rates to obtain the phylogenetic informativeness profiles.

Results and Discussion

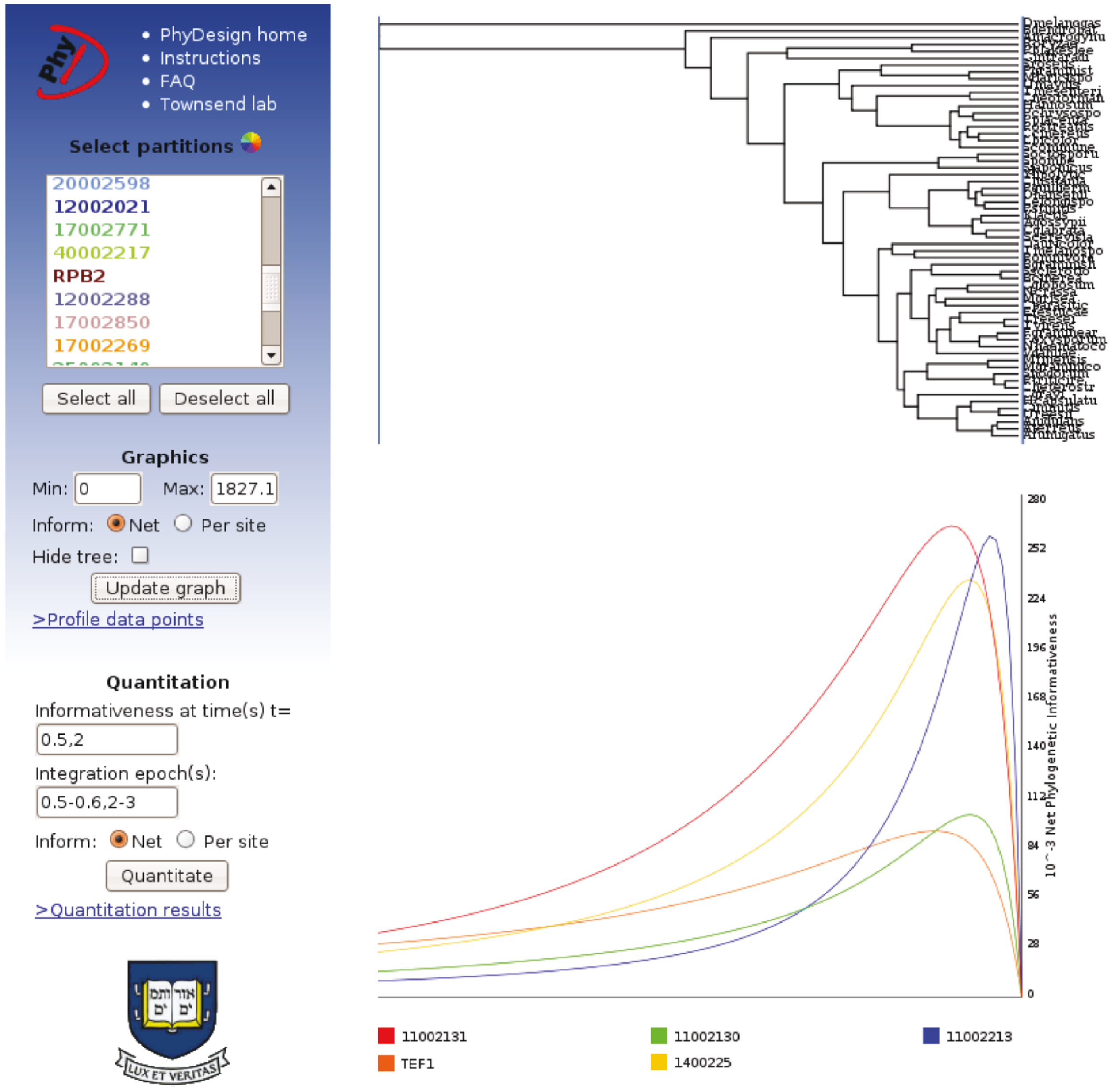
Output of rate vectors

The first result provided features a table with basic information about each locus analyzed, including the program used to analyze the rates, locus length, the number of sites for which a substitution rate could be calculated, and the number of faulty sites for which this calculation was not possible. Two files are offered as downloads: (1) a compressed file containing individual rate files for each locus, and (2) a single file containing rate site vectors for all loci. The latter can be downloaded for future uploading in the rate vectors form, eliminating the need to repeat rate calculations. Partitions to be profiled and the colors for representing them may then be specified interactively.

Profiles and area calculations

The second result provided is a graphical representation of phylogenetic informativeness and calculations of quantitative measures of informativeness. Two graphs are shown in the main section: the ultrametric tree and, aligned to it, the phylogenetic informativeness profiles (Figure 1).

These graphs can be readily downloaded in manipulable SVG format as displayed. In addition, a downloadable spreadsheet file with the profile data points is provided in the left panel, so that profiles can be replotted or reanalyzed with diverse software. The left panel facilitates further customization of the plots,



Min: 0 Max: 1827.1

Inform: ● Net ○ Per site

Hide tree: □

Informativeness at time(s) t= 0.5,2

Integration epoch(s): 0.5-0.6,2-3

Inform: ● Net ○ Per site

Figure 1 Screenshot of PhyDesign. Phylogenetic informativeness profiles are shown for five loci, aligned to a corresponding chronogram. The side bar is the interface through which the user can select loci and colors, adjust the range of time values, integrate over different epochs, and download the results.

including selection of loci to display, colors of profiles, adjustment of the range of time values, integration over different epochs, and downloading of the results. Integrating phylogenetic informativeness over specific epochs provides a metric for ranking loci. Integration values will be largest for the loci that have the highest probability of substitution in the given epoch that will not be obscured by subsequent evolution. Note, however, that phylogenetic informativeness plots display predicted signal and do not account for phylogenetic noise (homoplasy) caused by convergence or parallelism in divergent lineages. Thus, quantitative results should be thoughtfully considered in light of homoplasy that is likely to arise, significantly diminishing utility during epochs deeper than the peak of informativeness for a given profile. All informativeness values can be calculated on a net or per site basis. While the net phylogenetic informativeness quantifies signal as a whole, it is more subject to phylogenetic noise [14,15] than is phylogenetic informativeness per site, which maximizes both informativeness and cost-effectiveness, and more effectively minimizes noise. It also of conceptual interest to characterize the phylogenetic informativeness per site to compare relative power of genes without the confounding influence of gene length. A combination of shorter genes with a sequencing effort equal to that of a longer gene can lead to better results.

Conclusions

By providing these profiles, PhyDesign facilitates locus prioritization, increasing the efficiency of sequencing for phylogenetic purposes compared to traditional studies with more laborious and low capacity screening methods, as well as increasing the accuracy of phylogenetic studies. Future website implementations will include the latest theoretical advances developed in our research group, expanding the current phylogenetic informativeness methodology to quantify the effects of parallelism and convergence, as well as quantifying the utility of taxon addition [16].

Availability and requirements

Project name: PhyDesign

Project home page: http://phydesign.townsend.yale.edu

Operating system(s): Platform independent

Programming language: JavaScript and Perl

Other requirements: Internet browser with JavaScript and SVG for the graphical outputs (supported by all major browsers).

License: The tool is available online free of charge, and code is available under a Creative Commons Attribution-ShareAlike 3.0 Unported License.

Any restrictions to use by non-academics: None

Acknowledgements

We are grateful to S.L.K. Pond for helping with the HyPhy batch file design and E. Roberts for helping with the JavaScript programming. We also thank two anonymous reviewers for their thoughtful and useful comments.

Authors' contributions

FL-G and JPT conceived and designed the implementation. FL-G developed the online interface and server Perl scripts. FL-G wrote the manuscript. All authors read, edited, and approved the final manuscript.

Received: 3 February 2011 Accepted: 31 May 2011
Published: 31 May 2011

References

1. Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak MK, Douzery EJ: **OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics.** *BMC Evol Biol* 2007, **7**:241.
2. Marthey S, Aguileta G, Rodolphe F, Gendrault A, Giraud T, Fournier E, Lopez-Villavicencio M, Gautier A, Lebrun MH, Chiapello H: **FUNYBASE: a FUNgal phYlogenomic dataBASE.** *BMC Bioinformatics* 2008, **9**:456.
3. Boutros R, Stokes N, Bekaert M, Teeling EC: **UniPrime2: a web service providing easier Universal Primer design.** *Nucleic Acids Res* 2009, , **37 Web Server:** W209-13.
4. Contreras-Moreira B, Sachman-Ruiz B, Figueroa-Palacios I, Vinuesa P: **primers4clades: a web server that uses phylogenetic trees to design lineage-specific PCR primers for metagenomic and diversity studies.** *Nucleic Acids Res* 2009, **37 Web Server:** W95-W100.
5. Cummings MP, Meyer A: **Magic bullets and golden rules: data sampling in molecular phylogenetics.** *Zoology* 2005, **108(4)**:329-336.
6. Townsend JP: **Profiling phylogenetic informativeness.** *Syst Biol* 2007, **56(2)**:222-231.
7. Mahon BC, Neigel JE: **Utility of arginine kinase for resolution of phylogenetic relationships among Brachyuran genera and families.** *Mol Phylogenet Evol* 2008, **48(2)**:718-727.
8. Townsend JP, López-Giráldez F, Friedman R: **The phylogenetic informativeness of nucleotide and amino acid sequences for reconstructing the vertebrate tree.** *J Mol Evol* 2008, **67(5)**:437-447.
9. Schoch CL, Sung GH, López-Giráldez F, Townsend JP, Miadlikowska J, Hofstetter V, Robbertse B, Matheny PB, Kauff F, Wang Z, Gueidan C, Andrie RM, Trippe K, Ciufetti LM, Wynns A, Fraker E, Hodkinson BP, Bonito G, Groenewald JZ, Arzanlou M, de Hoog G, Crous PW, Hewitt D, Pfister DH, Peterson K, Gryzenhout M, Wingfield MJ, Aptroot A, Suh SO, *et al*: **The Ascomycota Tree of Life: A Phylum-wide Phylogeny Clarifies the Origin and Evolution of Fundamental Reproductive and Ecological Traits.** *Syst Biol* 2009, **58(2)**:224-239.
10. Tekle YI, Grant JR, Kovner AM, Townsend JP, Katz LA: **Identification of new molecular markers for assembling the eukaryotic tree of life.** *Mol Phylogenet Evol* 2010, **55(3)**:1177-1182.
11. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12(10)**:1611-1618.
12. Pond SLK, Frost SDW, Muse SV: **HyPhy: hypothesis testing using phylogenies.** *Bioinformatics* 2005, **21(5)**:676-679.
13. Mayrose I, Graur D, Ben-Tal N, Pupko T: **Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior.** *Mol Biol Evol* 2004, **21(9)**:1781-1791.
14. Klopfstein S, Kropf C, Quicke DLJ: **An evaluation of phylogenetic informativeness profiles and the molecular phylogeny of diplazontinae (Hymenoptera, Ichneumonidae).** *Syst Biol* 2010, **59(2)**:226-241.
15. Townsend JP, Leuenberger C: **Point of View: Taxon sampling and the optimal rates of evolution for phylogenetic inference.** *Syst Biol* 2011, **60(3)**:358-365.
16. Townsend JP, Lopez-Giraldez F: **Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships.** *Syst Biol* 2010, **59(4)**:446-457.

doi:10.1186/1471-2148-11-152
Cite this article as: López-Giráldez and Townsend: **PhyDesign: an online application for profiling phylogenetic informativeness.** *BMC Evolutionary Biology* 2011 **11**:152.