Bioinformatics, 35(13), 2019, 2320–2322
doi: 10.1093/bioinformatics/bty966
Advance Access Publication Date: 27 November 2018
Applications Note

OXFORD

Phylogenetics

PhastWeb: a web interface for evolutionary conservation scoring of multiple sequence alignments using phastCons and phyloP

Ritika Ramani, Katie Krumholz, Yi-Fei Huang and Adam Siepel*

Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY 11724, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on May 23, 2018; revised on November 1, 2018; editorial decision on November 23, 2018; accepted on November 26, 2018

Abstract

Summary: The Phylogenetic Analysis with Space/Time models (PHAST) package is a widely used software package for comparative genomics that has been freely available for download since 2002. Here, we introduce a web interface (phastWeb) that makes it possible to use two of the most popular programs in PHAST, phastCons and phyloP, without downloading and installing the PHAST software. This interface allows users to upload a sequence alignment and either upload a corresponding phylogeny or have one estimated from the alignment. After processing, users can visualize alignments and conservation scores as genome browser tracks and download estimated tree models and raw scores for further analysis. Altogether, this resource makes key features of the PHAST package conveniently available to a broad audience.

Availability and implementation: PhastWeb is freely available on the web at http://compgen.cshl.edu/phastweb/. The website provides instructions as well as examples.

Contact: asiepel@cshl.edu

1 Introduction

In recent years, there have been enormous investments in complete genome sequencing of species that fall relatively close to one another on the tree of life, allowing for comparative genomic analyses on unprecedented scales. The PHylogenetic Analysis with Space/Time models (PHAST) software package has emerged as a popular and widely used toolkit for analyzing such comparative genomic data. PHAST is best known as the engine behind the Conservation tracks in the University of California, Santa Cruz (UCSC) Genome Browser, but it additionally includes several programs for phylogenetic modeling and functional element identification, as well as utilities for manipulating alignments, trees and genomic annotations.

Since 2002, PHAST has been available as a collection of command-line programs and supporting software libraries that users must download and install to apply to their own sequence data. However, traffic on the PHAST mailing list indicates that many users are exclusively interested in producing conservation scores or predicted conserved elements using the phastCons or phyloP programs (Siepel *et al.*, 2005), an application that requires running only a handful of existing programs. Installation of an entire client-side software package for this relatively straightforward application can strike users as excessive and inconvenient. In addition, users often wish to visualize their conservation scores and predicted conserved elements together with their multiple sequence alignment in a Genome Browser display, but the client-side PHAST package does not support such visualization (Hubisz *et al.*, 2011).

Here, we introduce an easy-to-use web interface to PHAST, called phastWeb, to facilitate conservation-scoring using PHAST and visualization using the UCSC Genome Browser. Users of phastWeb are able to circumvent the non-trivial process of installing the PHAST software, running several command-line tools, converting output formats and uploading data for visualization. Instead, all of the necessary steps are launched via a self-explanatory user interface and executed on our servers (Fig. 1A). Visualization is accomplished using the UCSC Genome Browser's 'track hub' mechanism (Raney *et al.*, 2014), as shown in Figure 1B. Users of phastWeb can either estimate phylogenetic trees, branch lengths and substitution models from their own datasets or accept pre-estimated models. Key intermediate data files

(such as phylogenetic models and 'wig' files of conservation scores) are made available for download.

2 Materials and methods

2.1 Getting started

The only required input for phastWeb is a sequence alignment file (in MAF, FASTA or PHYLIP format, with a maximum file size of 40 MB). In addition, the user may optionally provide a known phylogeny in Newick format or a pre-estimated neutral model (*.mod) file, if one is available from a previous analysis. All computations are accomplished on the server side, using phastCons, phyloP and other programs from PHAST, together with the phastWeb scripts. Our server has the capacity to handle various jobs depending on both the number of species and the length of the genomic segment represented in the input alignment. For example, an alignment of 20 species spanning 115 kbp of genomic sequence has an estimated processing time of about 10 min, as does an alignment with 75 species but a length of only 2 kbp (Fig. 1C). In contrast, an alignment with 40 species and a length of 3 Mbp has an estimated processing time of around 17 h. If a job is estimated to require more than 24 h, it will be rejected.

2.2 Estimating the neutral model [if not provided]

If the user has not provided a *.mod file, a neutral model must be estimated from the alignment. This step requires a tree topology defining the phylogenetic relationships of the aligned sequences (Pollard *et al.*, 2010). The user can choose to upload a known tree topology in Newick format, such as a published tree or one that has been estimated separately, or to have the topology estimated from the alignment using the neighbor joining method. If necessary, tree estimation is accomplished using the *neighbor* program from PHYLIP (Felsenstein, 2005). Once a tree is obtained, a neutral substitution model is estimated from the data using the phyloFit program in PHAST. The user has the option to upload a file defining the locations of sites likely to be free from the influence of natural selection (such as 4-fold degenerate sites in coding regions, ancestral repeats or intergenic regions). If this option is not selected, the model is estimated from all sites in the alignment under the assumption that most sites are not under selection (as is typically true for

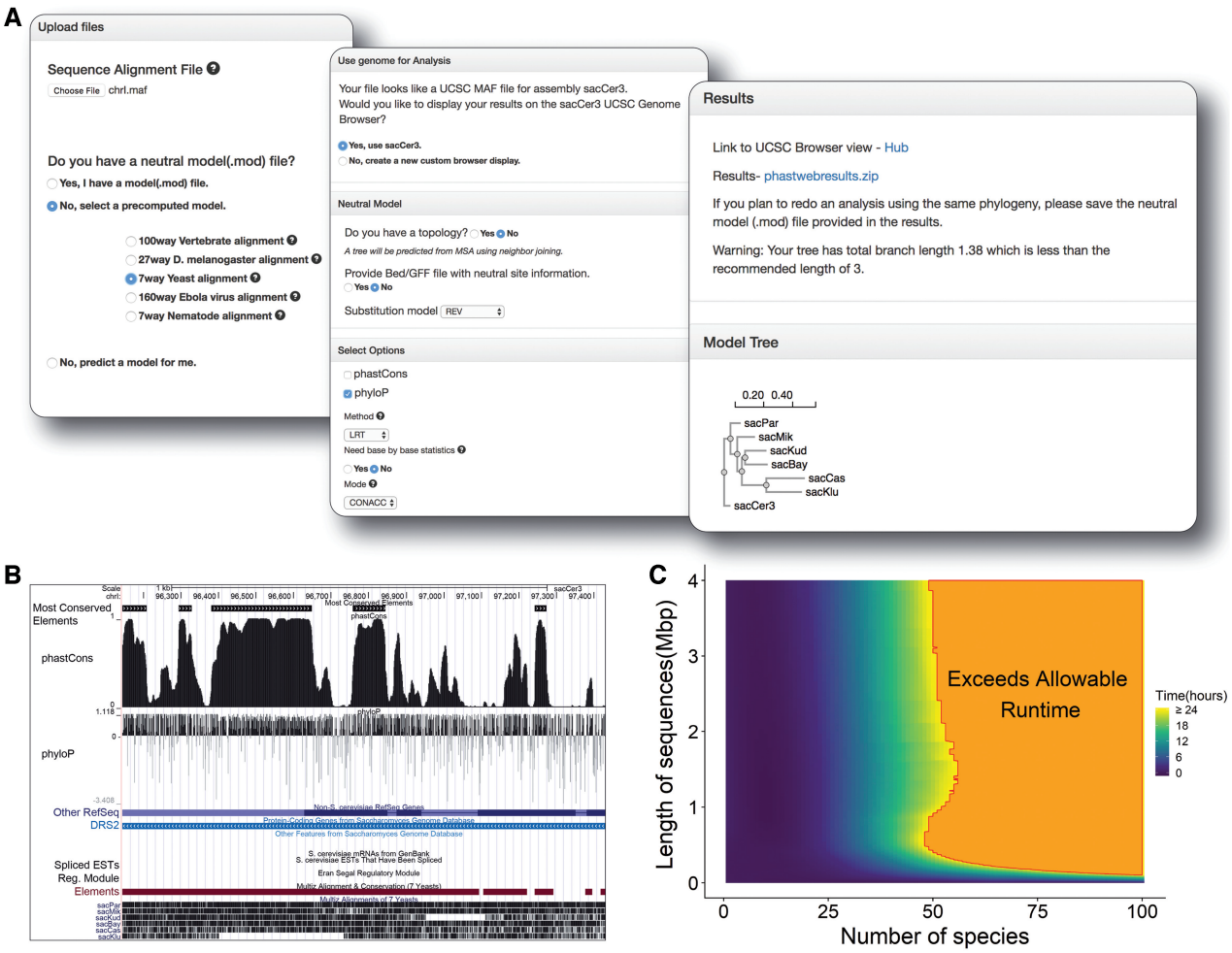

Fig. 1. Web Interface for phastWeb. (**A**) The interface prompts users to select options for running the phastCons and phyloP programs independently or together, after a.mod file is either provided by the user or selected from the available pre-computed neutral models. A neutral model can also be estimated from the alignment. The results page presents the zip file of phastWeb result with tree topology and a link to the UCSC Genome Browser's track hub. (**B**) The UCSC Genome Browser is used to display the generated conservation scores and conserved elements together with the reference genome. (**C**) Estimated run times based on the size of the alignment and the number of species

large genomes). The user also can select from one of several nucleotide substitution models implemented in PHAST.

2.3 Running phastCons and phyloP

Once a neutral model is obtained, the user can proceed with conservation scoring. The interface prompts the user to run the phastCons and phyloP programs independently or together. Sensible default parameters are provided for both programs but the user is free to customize them as desired, with guidelines provided in the online instructions.

2.4 Output

When a job is submitted, phastWeb provides an estimate of the required run time based on the size of the alignment and the number of species (see above). Once the results are available, the user receives an email with a link to a results page presented in three main parts, including: (i) a link to the UCSC Genome Browser's track hub displaying the generated conservation scores together with the reference genome and alignment; (ii) a zip file containing the phastCons and/or phyloP results (*.wig files), the tree topology (if estimated by *neighbor*), the neutral phylogenetic model estimated by phyloFit and the bigWig files for UCSC Genome Browser and (iii) an image (in scalable vector graphics [svg] format) displaying the neutral phylogeny used for the analysis. Users are encouraged to download all files generated by PHAST as these files are automatically deleted from our server after 7 days.

3 Conclusions

PhastWeb is an easy-to-use web-based interface to PHAST designed for users who wish to produce conservation scores or predict conserved elements from their own multiple sequence alignments, without downloading and installing the PHAST package. The system allows visualization of predictions in the UCSC Genome Browser and estimation of phylogenies and neutral models as needed. In addition to providing the basic functionality needed by many users, phastWeb can serve as an entry-point to a more elaborate conservation analysis using the full PHAST toolkit. If the phastWeb interface proves sufficiently useful, we may extend it to include other programs in PHAST, such as phyloFit and phastBias.

Acknowledgement

The authors thank Noah Dukler for preparing Figure 1 and designing the logo for the phastWeb website.

Funding

This work was supported by the National Institutes of Health [R01-HG008161 to A.S., R35-GM127070 to A.S.]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

Conflict of Interest: none declared.

References

Felsenstein,J. (2005) *PHYLIP (Phylogeny Inference Package) Distributed by the Author*, Version 3. Department of Genome Sciences, University of Washington, Seattle.

Hubisz,M.J. *et al.* (2011) PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.*, **12**, 41–51.

Pollard,K.S. *et al.* (2010) Detection of non-neutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.

Raney,B.J. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.