HyPhy - Title TBD

This manuscript (permalink) was automatically generated from rdvelazquez/hyphy_release_manuscript@97d773a on May 17, 2019.

Authors

- **Sergei L KosaKovsky Pond**
  · spond · sergeilkp
  Institute for Genomics and Evolutionary Medicine, Temple University

- **Ryan D Velazquez**
  · rdvelazquez
  Institute for Genomics and Evolutionary Medicine, Temple University

Abstract

Here we announce the latest release of the HyPhy software package. HyPhy is designed for the analysis of genetic sequences using stochastic evolutionary models with common application to understanding the pressures exerted by natural selection. HyPhy is widely used, actively maintained, open source and freely available on a multitude of platforms. The codebase is available at https://github.com/veg/hyphy. Documentation, tutorials and downloads are available at https://hyphy.org.

Introduction

HyPhy (Hypothesis testing using Phylogenies) is an open source software package for comparative sequence analysis using stochastic evolutionary models. Since its initial release in 2005 [1] HyPhy has become an integral tool for the bioinformatics community with over 10,000 registered users, over 2,000 peer-reviewed citations and approximately 1,000 HyPhy jobs processed each week on the datamonkey web server [2,3,4]. Extensions and improvements to the HyPhy package have been ongoing since its inception, with active feedback between users and developers producing new features tailored to the specific needs of the research community. Here we announce the release of the newest version of HyPhy (version 2.4) and document how the software has been (1) packaged for easy use in a variety of settings (2) optimized for larger datasets (3) redesigned to follow modern bioinformatics best practices and (4) extended to include a broader set of standard analyses.

Packaged for Easy Use in a Variety of Settings

The users of HyPhy vary greatly in their technical proficiency, from biologists unfamiliar with the command line to bioinformaticians who want to incorporate HyPhy into their own software. To meet the needs of this divers user set, HyPhy has been packaged and distributed for use in multiple different forms as outlined in table 1.

Usage	Required Skills	Easily Extensible	Mac	Windows	Linux	Tutorial
Custom HBL Scripts	HBL	Yes	Yes		Yes	http://hyphy.org/about/#example-hbl-script
Command Line Invocation	Command line	Yes	Yes		Yes	To Be Written

Usage	Required Skills	Easily Extensible	Mac	Windows	Linux	Tutorial
Command Line Prompt	Command line	Yes	Yes		Yes	http://hyphy.org/tutorials/current-release-tutorial/
PhyPhy	Python	Yes	Yes		Yes	https://github.com/sjspielman/phyphy
DataMonkey	None		Web	Web	Web	http://datamonkey.org/help
HyPhy-GUI	None		Yes		Yes	http://hyphy.org/tutorials/current-release-tutorial_gui/
Galaxy	None	Yes	Web	Web	Web	https://galaxyproject.org/support/
MEGA	None		Yes	Yes	Yes	https://www.megasoftware.net/docs

Optimized for Larger Datasets

HyPhy has been optimized to analyze datasets with thousands of sequences and tens of thousands of sites. This optimization was accomplished by (1) integrating recent algorithmic advances from the fields of machine learning and natural language processing [5] into the core c++ implementation (2) incorporating fast Bayesian methods [6] (3) developing parallel implementations for commute intensive processes and (4) providing high performance computing infrastructure[2,3,4] free of charge to the global community, helping to democratize access to scientific computing resources [7].

Redesigned to Follow Modern Bioinformatics Best Practices

The design of the original HyPhy implementation emphasized convenient writing and execution of single HyPhy Batch Language (HBL) scripts. This decision manifested in the easy to use command line prompt which guided users interactively through

selecting the desired analysis and choosing the specific analysis parameters. Also, the HBL itself was designed to allow sophisticated models to be specified and fit with concise scripts, facilitated by extensive use of a global namespace. As the common use of HyPhy has shifted over the last decade from one-off analyses toward larger studies and use within pipelines, the demands on HyPhy have also shifted. HyPhy has therefore been redesigned to address the requirements of these changing use cases.

Usage as a typical command line tool (i.e. an executable name followed by key word arguments) has been added alongside the interactive command line prompt. This change, along with the ability to use relative paths to files, has made using HyPhy in pipelines and batch analyses seamless. In addition to being easier to use, the package is also now easier to install. HyPhy is now installable with bioconda [8] which has become the defacto package manager for scientific software. Users no longer need to concern themselves with dependencies, environments and build processes but can simply `conda install hyphy`.

The inner workings of HyPhy have also been improved, providing a more reliable package that is easier to extend. To this end, namespacing has been introduced. Before, all variables were automatically declared at a global scope. Namespaces helped facilitate the refactoring and standardization of the template batch files for easier comprehension and reuse. Additionally, extensive automated testing has been implemented. The automated testing is executed as a part of a continuous integration (CI) pipeline which which both expedites development and helps ensure healthy software is delivered. The automated testing includes: unit tests for over 90% of HBL functions, method tests on all the core analyses, and likelihood testing [9] which compares the likelihood values calculated by HyPhy with values calculated by other popular maximum-likelihood software packages and informs the developers if discrepancies are identified.

Extended to Include a Broader Set of Standard Analyses

Although the HyPhy package provides for limitless customization via writing HBL scripts, users can run many common analyses without needing to concern themselves with the HBL at all. The HyPhy package comes with pre-written HBL scripts for easily performing some of the most commonly used analyses. These analyses can be executed in the various places HyPhy is available and include:

- **aBSREL** [10,11] (adaptive Branch-Site Random Effects Likelihood) - Detect positive/diversifying selection at individual branches
- **BGM** [12] (Bayesian Graphical Models) - Test for co-evolving sites
- **BUSTED** [13] (Branch-Site Unrestricted Statistical Test for Episodic Diversification) - Test gene-wide selection at pre-defined lineages

- **FADE** (FUBAR Approach to Directional Evolution) - Evaluate if sites are subject to directional selection (not yet published)
- **FEL** [14] (Fixed Effects Likelihood) - Inferring non-synoymous and synonymous substitution rates on a per-site basis for smaller datasets using maximum likelihood
- **FUBAR** [6](Fast, Unconstrained Bayesian AppRoximation) - Infer non-synonymous and synonymous substitution rates on a per-site basis for larger datasets using Bayesian methods
- **GARD** [15] (Genetic Algorithm for Recombination Detection) - Screen multiple sequence alignment for recombination
- **MEME** [16] (Mixed Effects Model of Evolution) - Test the hypothesis that individual sites have been subject to episodic positive/diversifying selection
- **RELAX** [17] - Evaluate whether the strength of natural selection has been relaxed or intensified along a specific set of branches
- **SLAC** [14] (SingleLikelihood Ancestor Counting) - Infer non-synonymous and synonymous substitution rates on a per-site basis

References

1. **HyPhy: hypothesis testing using phylogenies**
S. L. K. Pond, S. D. W. Frost, S. V. Muse
Bioinformatics (2004-10-27) https://doi.org/bbcz2p
DOI: 10.1093/bioinformatics/bti079 · PMID: 15509596

2. **Datamonkey: rapid detection of selective pressure on individual sites of codon alignments**
S. L. K. Pond, S. D. W. Frost
Bioinformatics (2005-02-15) https://doi.org/c2bz9d
DOI: 10.1093/bioinformatics/bti320 · PMID: 15713735

3. **Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology**
W. Delport, A. F. Y. Poon, S. D. W. Frost, S. L. Kosakovsky Pond
Bioinformatics (2010-07-29) https://doi.org/bz8k5p
DOI: 10.1093/bioinformatics/btq429 · PMID: 20671151 · PMCID: PMC2944195

4. **Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes**
Steven Weaver, Stephen D Shank, Stephanie J Spielman, Michael Li, Spencer V Muse, Sergei L Kosakovsky Pond
Molecular Biology and Evolution (2018-01-02) https://doi.org/gdbs4n
DOI: 10.1093/molbev/msx335 · PMID: 29301006 · PMCID: PMC5850112

5. **Latent Dirichlet Allocation**
David Blei
Journal of Machine Learning Research (2003) http://www.jmlr.org/papers/v3/blei03a.html

6. **FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection**
B. Murrell, S. Moola, A. Mabona, T. Weighill, D. Sheward, S. L. Kosakovsky Pond, K. Scheffler
Molecular Biology and Evolution (2013-02-18) https://doi.org/gfxxzx
DOI: 10.1093/molbev/mst030 · PMID: 23420840 · PMCID: PMC3670733

7. **Cloud computing: A democratizing force?**
Nabil Sultan
International Journal of Information Management (2013-10) https://doi.org/gfxxzv
DOI: 10.1016/j.ijinfomgt.2013.05.010

8. **Bioconda: sustainable and comprehensive software distribution for the life sciences**

Björn GrüningRyan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris, Johannes Köster
Nature Methods (2018-07) https://doi.org/gd2xzp
DOI: 10.1038/s41592-018-0046-7 · PMID: 29967506

9. **testiphy / testiphy***GitLab* https://gitlab.com/testiphy/testiphy

10. **Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection**
Martin D. Smith, Joel O. Wertheim, Steven Weaver, Ben Murrell, Konrad Scheffler, Sergei L. Kosakovsky Pond
Molecular Biology and Evolution (2015-02-19) https://doi.org/f7dh98
DOI: 10.1093/molbev/msv022 · PMID: 25697341 · PMCID: PMC4408413

11. **A Random Effects Branch-Site Model for Detecting Episodic Diversifying Selection**
Sergei L. Kosakovsky Pond, Ben Murrell, Mathieu Fourment, Simon D.W. Frost, Wayne Delport, Konrad Scheffler
Molecular Biology and Evolution (2011-06-13) https://doi.org/frpb4w
DOI: 10.1093/molbev/msr125 · PMID: 21670087 · PMCID: PMC3247808

12. **Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models**
A. F. Y. Poon, F. I. Lewis, S. D. W. Frost, S. L. Kosakovsky Pond
Bioinformatics (2008-06-18) https://doi.org/d9mgkz
DOI: 10.1093/bioinformatics/btn313 · PMID: 18562270 · PMCID: PMC2732215

13. **Gene-Wide Identification of Episodic Selection**
Ben Murrell, Steven Weaver, Martin D. Smith, Joel O. Wertheim, Sasha Murrell, Anthony Aylward, Kemal Eren, Tristan Pollner, Darren P. Martin, Davey M. Smith, … Sergei L. Kosakovsky Pond
Molecular Biology and Evolution (2015-02-19) https://doi.org/f7djbj
DOI: 10.1093/molbev/msv035 · PMID: 25701167 · PMCID: PMC4408417

14. **Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection**
Sergei L. Kosakovsky Pond, Simon D. W. Frost
Molecular Biology and Evolution (2005-02-09) https://doi.org/bx9bg2
DOI: 10.1093/molbev/msi105 · PMID: 15703242

15. **Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm**
Sergei L. Kosakovsky Pond, David Posada, Michael B. Gravenor, Christopher H. Woelk, Simon D. W. Frost

Molecular Biology and Evolution (2006-07-03) https://doi.org/fvnfwn
DOI: 10.1093/molbev/msl051 · PMID: 16818476

16. **Detecting Individual Sites Subject to Episodic Diversifying Selection**
Ben Murrell, Joel O. Wertheim, Sasha Moola, Thomas Weighill, Konrad Scheffler, Sergei L. Kosakovsky Pond
PLoS Genetics (2012-07-12) https://doi.org/f34pgn
DOI: 10.1371/journal.pgen.1002764 · PMID: 22807683 · PMCID: PMC3395634

17. **RELAX: Detecting Relaxed Selection in a Phylogenetic Framework**
Joel O. Wertheim, Ben Murrell, Martin D. Smith, Sergei L. Kosakovsky Pond, Konrad Scheffler
Molecular Biology and Evolution (2014-12-23) https://doi.org/f64sp5
DOI: 10.1093/molbev/msu400 · PMID: 25540451 · PMCID: PMC4327161