Relative evolutionary rate inference in HyPhy with LEISR

Stephanie J. Spielman[1,✉] **and Sergei L. Kosakovsky Pond**[1]

[1]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA 19122

We introduce LEISR (*L*ikehood *E*stimation of *I*ndividual *S*ite *R*ates, pronounced "laser"), a convenient implementation in HyPhy to infer relative evolutionary rates from protein and nucleotide data. LEISR is based on the popular Rate4Site (1) approach for inferring relative site-wise evolutionary rates from protein data. We extend the original method by allowing for rate inference on both protein and nucleotide data, and we further overcome data-size limitations in Rate4Site. LEISR is available in HyPhy starting with version 2.3.7.

relative evolutionary rate | phylogenetics | molecular evolution
Correspondence: ***stephanie.spielman@temple.edu***

Introduction

Evolutionary rate inference is a fundamental analysis in computational molecular evolution (2). A widely-used tool for inferring evolutionary rates from phylogenetic protein data is Rate4Site, which exists both as a server and a command-line tool (1). Although this method has proven extremely useful over the years, garnering nearly 500 citations, Rate4Site has several limitations: i) it cannot accept more than ~ 200 sequences because of numerical underflow issues; ii) its optimizer often fails to converge if data are sufficiently complex even with relatively few $(25-100)$ sequences; and iii) it accepts primarily protein data only and has limited nucleotide utility. As the number of available genomic sequences continues to rapidly expand, tools to analyze large data sets of any genomic type (protein and nucleotide), are needed.

To this end, we introduce a generalized implementation of Rate4Site in the leading molecular evolution inference platform HyPhy (version $\geq 2.3.7$), which we term "LEISR" (**L**likehood **E**stimation of **I**ndividual **S**ite **Rates**). LEISR can be used to infer relative evolutionary rates from either nucleotide or protein data, thereby providing a flexible and fast platform for rate inference on non-coding data[1]. LEISR has been successfully tested with alignments containing up to 10,000 sequences, 2–3 orders of magnitude beyond what Rate4Site can manage. We note that LEISR is additionally MPI-enabled to support rapid inference from datasets with many sites, by distributing optimization tasks to multiple compute nodes.

Approach

LEISR proceeds in two steps. It first optimizes branch lengths under a user-specified substitution model (Table 1), and it then for each site, it infers a scaler parameter, r_s, that is used to uniformly scale all the branch lengths of the tree at the site. r_s can therefore be interpreted as the evolutionary rate at a specific site relative to the alignment-wide mean rate. As this algorithm proceeds, HyPhy will write markdown-formatted status-indicators to the console, including the inferred site-wise maximum-likelihood rate estimates with the 95% confidence interval (CI) obtained via profile likelihood. All final output is written to a JSON-formatted file, named as the input data file with the suffix <.LEISR.JSON>. Site-wise rates are stored in the top-level JSON field MLE, whose CONTENT field contains a row for each site's inferred rates. Individual values in each row correspond to information given in the HEADERS key. We note that a general description of HyPhy output JSON contents is available from http://www.hyphy.org.

Data type	Models
Nucleotide	GTR(6), HKY85(7), JC69(8)
Protein, Generalist	LG(9), WAG(10), JTT(11), JC69(8)
Protein, Specialist	mtMam(12), cpREV (13), HIVB/Wm (14), AB (15)

Table 1. Nucleotide and Protein models, both generalist and specialist, available for use in LEISR, as of HyPhy version 2.3.7. We note that users can apply other models with the use of HBL, the HyPhy batch language.

Users are free to transform these rates in a manner that suits their given analyses. For example, Rate4Site computes a standard score for each site, and other applications have called for normalizing each rate by the average (or median) gene-wide rate (16, 17). This latter scheme re-scales the average gene rate as 1, lending a more intuitive interpretation to each site's rate, i.e. a rate of 2 indicates that a site evolves twice as quickly as does the gene on average, and a rate of 0.5 indicates that a site evolves half as quickly as does the gene on average. As rate distributions are generally overdispersed and zero-inflated, it may be advisable to normalize by the median rather than the mean.

Rate4Site offers two statistical frameworks, maximum-likelihood (ML) (1) and empirical Bayes (18), for rate inference, where the latter approach requires a discrete gamma distribution of rates be used during branch length optimization. LEISR, on the other hand, only employs maximum-likelihood, and rate variation (if chosen) during branch length optimization can be modeled either with a discrete gamma distribution or the general discrete distribution (GDD) (19). Although we provide the option to employ rate variation for branch length optimization, we encourage users to opt for no

[1]We note that HyPhy contains several robust methods (including FEL (3), SLAC (3), and FUBAR (4)) for inferring site-wise evolutionary rates from codon data, and see Spielman et al. (5) for recommendations specifically on codon-level rate inference

rate variation. Indeed, the desired behavior for this method is for *only* the relative site-wise rates to contain information about site-wise evolutionary rate heterogeneity. If branch length optimization considers rate variation, then this information will be "split" between these two parameters (branch lengths and site rates).

Results

We confirmed that **LEISR** yields comparable inferences to **Rate4Site** using simulations. For each of three random phylogenies with 25, 50, and 100 taxa each, we simulated 10 replicate alignments, each with 100 sites, under the WAG model of protein evolution (10). Our simulations modeled rate heterogeneity among sites with a discrete gamma distribution with 20 categories and a shape parameter of 0.4. We note that each replicate number used the same model parameterizations for all three trees (i.e. replicate 1 employed the same model for 25, 50, and 100 taxa). Simulations were conducted using the Python simulation library **pyvolve** (20).

We then inferred relative evolutionary rates in **LEISR** in two modes: turning off rate heterogeneity during branch length optimization ("LEISR"), and specifying a four-category discrete gamma distribution during during branch length optimization ("LEISR+G"). We again inferred rates in two modes in Rate4Site: without rate heterogeneity during branch length optimization ("R4S") and with a four-category discrete gamma distribution ("R4S+G"). We conducted all **Rate4Site** inferences using their maximum-likelihood algorithm (in contrast to their empirical Bayes approach). Note that the default number of rate categories for this step in **Rate4Site** is 16, but **Rate4Site** failed with errors for all 100-taxa simulations. We therefore used four rate categories, to both achieve a fair comparison with **LEISR** and to ensure that **Rate4Site** could complete inferences. Finally, for each alignment inference, we normalized rate estimates by dividing all rates by the mean site rate estimate, as described earlier.

In Figure 1, we show r^2 values for Pearson's linear correlation between **LEISR** and **Rate4Site** inferences, computed across all simulations. Figure 2 shows, for a single representative simulation replicate of 100 taxa, the relationship between inferred site rates across different methods and/or parameterization. Overall, these results demonstrate a nearly complete agreement between **LEISR** and **Rate4Site**, with rate inferences showing the closest agreement when the same option for branch length optimization was specified (i.e. turned off or with a discrete gamma distribution). The r^2 values further increase as the number of taxa increases, although even with 25 taxa the agreement is remarkably high. This is expected because the precision of inference for individual site rates will increase for larger samples (more taxa). We therefore find that **LEISR** provides a robust and reliable platform that can be used in the place of **Rate4Site** when dataset size and/or complexity preclude **Rate4Site** use.

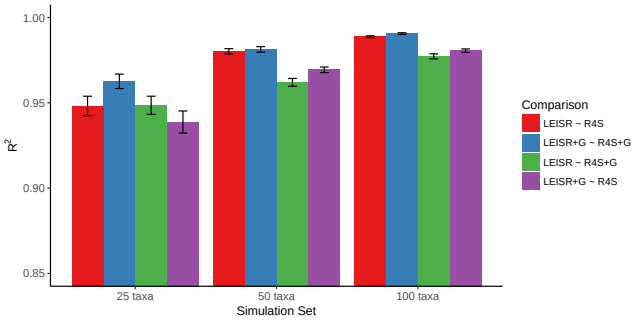

Fig. 1. Mean r^2 values (across 10 replicates) between inferred evolutionary rates across platforms and simulations. Bars represent the standard error of the mean. Note that the y-axis of this figure begins at $r^2 = 0.85$. All code to generate simulations and reproduce figures is available from https://github.com/sjspielman/leisr_validation.

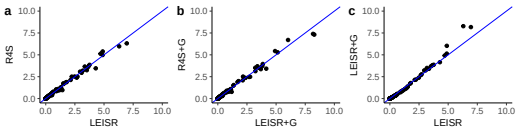

Fig. 2. Inferred evolutionary rates for a single simulation replicate with 100 taxa. The line shown in each panel is $y = x$. All code to generate simulations and reproduce figures is available from https://github.com/sjspielman/leisr_validation.

Conclusions

LEISR is accessible as the third option in the **HyPhy** analysis menu ("Relative evolutionary rate inference"), which calls the **HyPhy** batchfile **LEISR.bf**. Finally, we encourage users who cite **LEISR** to additionally cite **Rate4Site**, which provides the intellectual basis and historical precedent for our implementation.

ACKNOWLEDGEMENTS

This work was supported in part by grants R01 GM093939 (NIH/NIGMS), R01 AI134384 (NIH/NIAID), and U01 GM110749 (NIH/NIGMS).

Bibliography

1. T. Pupko, R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18:S71–S77, 2002.
2. J. Echave, S. J. Spielman, and C. O. Wilke. Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.*, 17:109–121, 2016.
3. S. L. Kosakovsky Pond and S. D. W. Frost. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.*, 22:1208–1222, 2005.
4. B. Murrell, S. Moola, A. Mabona, T. Weighill, D. Scheward, S. L. Kosakovsky Pond, and K. Scheffler. FUBAR: A Fast, Unconstrained Bayesian AppRoximation for inferring selection. *Mol. Biol. Evol.*, 30:1196–1205, 2013.
5. S. J. Spielman, S. Wan, and C. O. Wilke. A comparison of one-rate and two-rate inference frameworks for site-specific dn/ds estimation. *Genetics*, 204:499–511, 2016.
6. S. Tavare. Lines of descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.*, 26:119–164, 1984.
7. M. Hasegawa, H. Kishino, and T Yano. Dating the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22:160–174, 1985.
8. T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian protein metabolism*. Academic Press, New York, 1969.
9. S. Q. Le and O. Gascuel. An improved general amino acid replacement matrix. *Mol. Biol. Evol.*, 25:1307–1320, 2008.
10. S. Whelan and N. Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. *Mol. Biol. Evol.*, 18:691–699, 2001.
11. D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *CABIOS*, 8:275–282, 1992.

12. N. Yang, R Nielsen, and M. Hasegawa. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.*, 15:1600–1611, 1998.
13. J. Adachi, P. J. Waddell, W. Martin, and Hasegawa M. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast dna. *J. Mol. Evol.*, 50: 348–358, 2000.
14. D. C. Nickle, L. Heath, M. A. Jensen, Peter B. Gilbert, J. I. Mullins, and S. L. Kosakovsky Pond. Hiv-specific probabilistic models of protein evolution. *PLOS ONE*, 2 (6):e503, 2007.
15. A. Mirsky, L. Kazandjian, and M. Anisimova. Antibody-specific model of amino acid substitution for immunological inferences from alignments of antibody sequences. *Mol. Biol. Evol.*, 32:806–819, 2015.
16. B. R. Jack, A. G. Meyer, J. Echave, and C. O. Wilke. Functional sites induce long-range evolutionary constraints in enzymes. *PLOS Biology*, 14(5):1–023, 2016.
17. D.K. Sydykova, B.R. Jack, S.J. Spielman, and C.O. Wilke. Measuring evolutionary rates of proteins in a structural context. *F1000Research*, 6(185), 2017. doi: 10.12688/f1000research.12874.1.
18. I. Mayrose, D. Graur, Ben-Tal N., and T. Pupko. Comparison of site-specific rate-inference methods for protein sequences: Empirical bayesian methods are superior. *Mol. Biol. Evol.*, 1821.
19. S. Kosakovsky Pond and S. V. Muse. Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.*, 22:2375–2385, 2005.
20. S. J. Spielman and C. O. Wilke. Pyvolve: A flexible python module for simulating sequences along phylogenies. *PLOS ONE*, 10:e0139047, August 2015.