RATE VARIATIONS, PHYLOGENETICS, AND PARTIAL ORDERS

Sonja J. Prohaska[1,2], *Guido Fritzsch*[3,4], *and Peter F. Stadler*[5,1,2,4,6]

[1]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe NM 87501, USA
[2]Department of Theoretical Chemistry, University of Vienna,
Währingerstraße 17, A-1090 Wien, Austria;
[3]Institute of Biology II: Zoologie, Molekulare Evolution und Systematik der Tiere,
University of Leipzig, Talstrasse 33, D-04103 Leipzig, Germany
[4]Interdisciplinary Center for Bioinformatics, and
[5]Bioinformatics Group, Department of Computer Science
University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
[6]RNomics Group, Fraunhofer Institut for Cell Therapy and Immunology (IZI),
Deutscher Platz 5e, D-04103 Leipzig, Germany
sonja@santafe.edu, gfritz@rz.uni-leipzig.de, studla@bioinf.uni-leipzig.de

ABSTRACT

The systematic assessment of rate variations across large datasets requires a systematic approach for summarizing results from individual tests. Often, this is performed by coarse-graining the phylogeny to consider rate variations at the level of sub-claded. In a phylo-geographic setting, however, one is often more interested in other partitions of the data, and in an exploratory mode a pre-specified subdivision of the data is often undesirable. We propose here to arrange rate variation data as the partially ordered set defined by the significant test results.

1. INTRODUCTION

Rate variations are an important source of information in evolutionary biology. Typically, one devises so-called relative-rate tests (RRTs) for statistically significant rate variations between two species [1, 2, 3, 4] or between subgroups of species [5, 6]. Group tests, however, require an initial hypothesis about which species to summarize. In particularly in an exploratory phase this is typically undesirable, since rate variations can be associated with many very different mechanisms, for clade-specific changes in mutation rates to differences in population structure.

In this contribution we therefore introduce an explorative approach to summarizing the results of many pairwise RRTs. The basic idea is to arrange the individual statistically significant pair-wise test results in a partially ordered set. Inspection of the Hasse diagram of this graph can then be used to identify systematic rate variations. In particular, this approach has the potential to highlight systematic rate variations even if they do not conform to a phylogenetic tree but correlate with other variables, such as migratory history.

2. RELATIVE RATE PO-SET

2.1. Po-Sets

Recall that a partially ordered set, *po-set* for short, is a set X together with a relation $\preceq$ satisfying
(P0) $x \preceq x$.
(P1) $x \preceq y$ and $y \preceq x$ implies $x = y$.
(P2) $x \preceq y$ and $y \preceq z$ implies $x \preceq z$.
A finite po-set $(X, \preceq)$ can be respresented as directed acyclic graph G (by drawing an arc $x \leftarrow y$ whenever $x \preceq y$ and $x \neq y$). The Hasse diagram of G is the subgraph H of G with the same vertex set X, and an arc $x \to y$ if $x \to y$ is an arc in G and there is no $z \neq x, y$ such that z lies on a directed path from x to y in G.

2.2. Substitution Rates

Let $\mathcal{X}$ be a set a taxa, which we represent here by their (aligned) nucleic acid or peptide sequences of length n. Furthermore, let $\mathfrak{T}$ be the underlying phylogenetic tree. Each interior vertex w of the tree can be specified as the *last common ancestor* $w = \mathrm{lca}(A, B)$ of two of the descents A and B of w so that the path connecting A and B runs through w.

The Hamming distance $d_{AB} = |\{i|A_i \neq B_i\}|$ counts the positions i in which the characters of the sequences differ. Now consider a triple (A, B, C) of sequences. The quantities

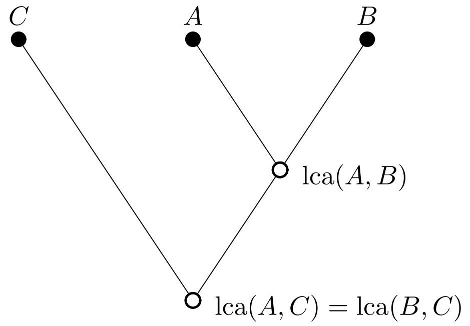
$$\begin{aligned} a_{ABC} &= |\{i|A_i = B_i = C_i\}|, \\ m_{AB|C} &= |\{i|A_i = B_i \neq C_i\}|, \\ m_{AC|B} &= |\{i|A_i = C_i \neq B_i\}|, \\ m_{BC|A} &= |\{i|B_i = C_i \neq A_i\}|, \\ w_{ABC} &= |\{i|A_i \neq B_i \neq C_i \neq A_i\}| \end{aligned} \tag{1}$$

distiguish five classes of alignment positions: (i) constant positions, (ii) positions in which all three sequence differ and (iii) three classes of positions in which two sequences are the same and the third one ins different.

The Hamming distance d_{AB} can be decomposed into three different components w.r.t. to a third sequence C. These correspond to the sequence position where C agrees with B (but not with A), the positions where C agrees with A (but not with B), and those where all three sequences differ:

$$d_{AB} = m_{BC|A} + m_{AC|B} + w_{ABC} \tag{2}$$

Now consider a subtree of $\mathfrak{T}$ consisting of three taxa A, B, C so that C is an outgroup to A and B:



(3)

Let us denote by a and b the lengths of branches between A, B and $\mathrm{lca}(A,B)$, respectively. We have

$$\begin{aligned} 2a = d_{AC} + d_{AB} - d_{BC} &= 2m_{BC|A} + w_{ABC} \\ 2b = d_{BC} + d_{AB} - d_{AC} &= 2m_{AC|B} + w_{ABC} \end{aligned} \tag{4}$$

and hence

$$a - b = m_{BC|A} - m_{AC|B} \,. \tag{5}$$

Note that $m_{BC|A}$ and $m_{AC|B}$ count independent sequence positions, while the Hamming distances are dependent via the common term w_{ABC}. Equ.(5) is the basis of Tajima's relative rate test [2], while the older Wu & Li test [3] uses the difference $d_{AC} - d_{BC}$. Alternatively, one might want to employ a suitable maximum likelihood test to assess the significance of branch length differences [1, 4].

We can estimate the relative rate of evolution along the branches a and b for those comparisons that are statistically signficant according to the relative rate test of choice. In the following, it will be more convenient to use the following logarithmic measure

$$\eta_{AB} = \begin{cases} \ln\frac{a}{b} & \text{if } a - b \text{ is statistically significant} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Next we show that for ideal data we do not have to fear contradictory results of relative rate tests involving different triples of taxa selected from the tree $\mathfrak{T}$. Recall that the distances d_{AB} of leafs A and B in a additive metric tree $\mathfrak{T}$ are defined as the sum of the lengths of the edges along the unique path that connects A and B in $\mathfrak{T}$.

More precise, we have the following

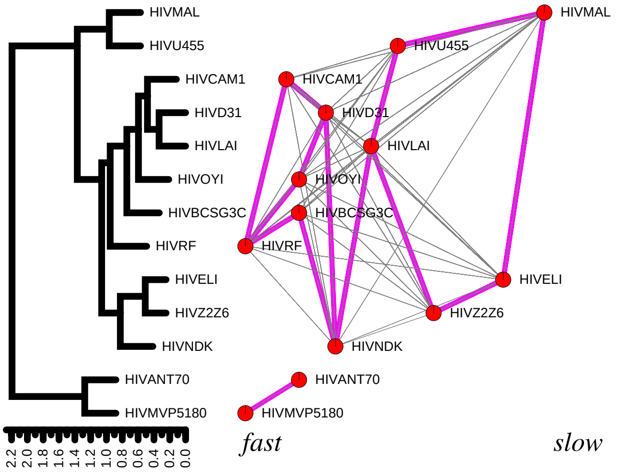
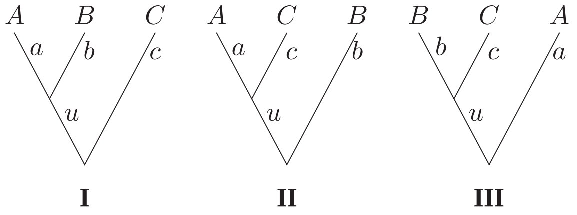


Figure 1. Example of a relative rate poset. Data are 5'UTRs of HIV-1. Thin lines in the r.h.s. panel indicate significant Tajima tests, the thick lines represent the associated Hasse diagram of the partially ordered set.

Theorem 1. *The directed graph associated with $\boldsymbol{\eta}$ is acyclic provided d is an additive tree metric on $\mathcal{X}$.*

Proof. First, we observe that $\boldsymbol{\eta}$ is antisymmetric by construction, $\eta_{AB} = -\eta_{BA}$. Thus there are no cycles of length 2. Next assume $\eta_{AB} > 0$ and $\eta_{BC} > 0$. We have to consider the following three cases



Translating the assumption in inequalities of branch lengths in each of the three cases yields:

(I) $a > b$ and $b + u > c$ implies $a + u > c$, i.e., $\eta_{AC} \geq 0$.

(II) $a + u > b$ and $b > c + u$ implies $a > c$, i.e., $\eta_{AC} \geq 0$.

(III) $a > b + u$ and $b > c$ implies $a > c + u$, i.e., $\eta_{AC} \geq 0$.

These three inequalities for η_{AC} assume that the underlying statistical test is "sane" in the sense that it never returns a significantly larger rate for the short branch. Thus $\eta_{AB} > 0$ and $\eta_{BC} > 0$ always implies $\eta_{AC} \geq 0$. Now consider a chain of taxa $\{A^j | 1 \leq j \leq m\}$ such that $\eta_{A^{j-1}A^j} > 0$ for $2 \leq j \leq m$. By repeated application of the this result we conclude $\eta_{A^k,A^l} \geq 0$ for any $l > k$, i.e., the $\{A^j\}$ cannot be part of a directed cycle. Since there is an edge from node i to node j iff $\eta_{i,j} > 0$, we conclude that the corresponding graph is a DAG, and hence the matrix $\boldsymbol{\eta}$ is acyclic. □

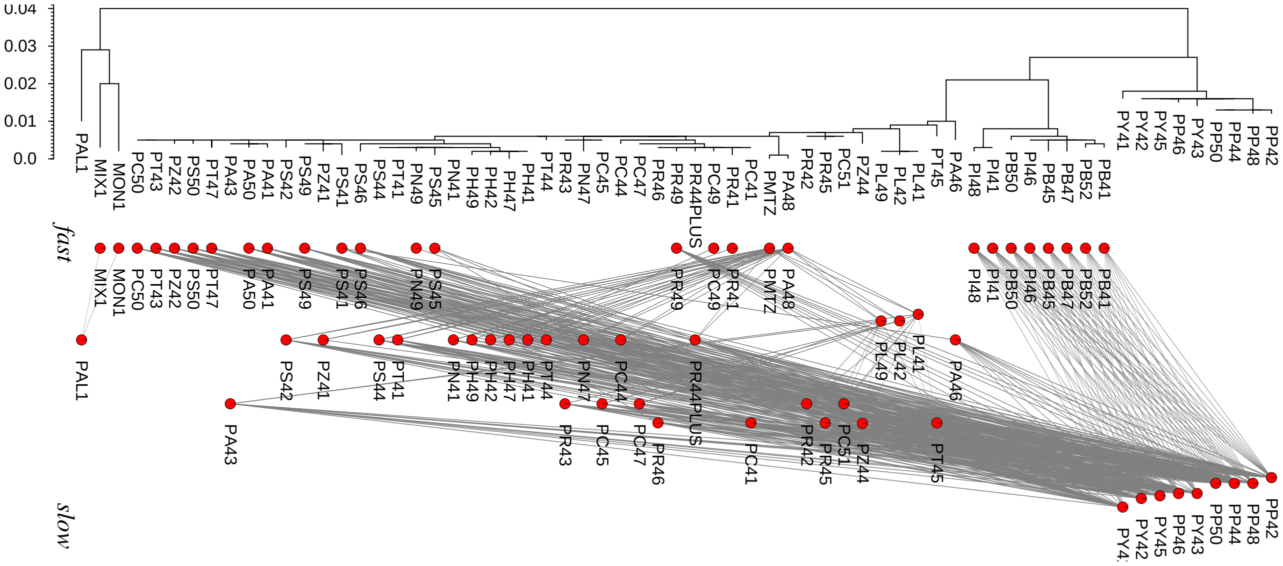


Figure 2. Phylogenetic tree (neighbor joining) and Hasse diagram of the relative-rate poset of mtND1 nucleotide sequence data of wolf spiders of the *Pardosa saltuaria* group [7]. Significance level for Tajima tests $p \leq 0.1$ ($\chi^2 = 2.706$), test results of all subtrees included. Labels refer to geographic locations: North/South Scandinavia PN, PS; Eastern/Western Riesengebirge PC, PR; Tatra Mountains PT; Alps PA, PL, PZ; Eastern/Western Pyrenees PP, PY; Balkans PB, PI; Bohemia PH; Lago di Garda area PMTZ. Outgroup: *P. palustris* PAL, *P. monticola* MON1, *P. mixta* MIX.

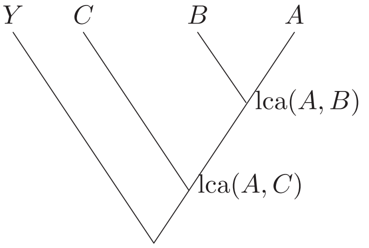
In order to work with real data, we have to relax the assumption that d is an additive tree metric. The estimates for a and b will then depend explicitly on the outgroup C. Note, however, that these variations are small as long as the data are at least approximately tree-like. We can therefore estimate η_{AB} as an *average* over all those triples (A, B, C) for which the Tajima test demonstrates a significant rate difference. The χ^2 value obtained from the Tajima test can be used as weight of the individual estimates. Numerically, we observe that $\boldsymbol{\eta}$ is indeed acyclic even when small χ^2 significance thresholds for the Tajima test are used.

The construction of the matrix $\boldsymbol{\eta}$ starting from a sequence alignment using Tajima's relative rate test has been implemented in a software prototype. It either uses a phylogenetic tree $\mathfrak{T}$ as additional input, or tests for all triples (A, B, C) with outgroup C if $d_{AC}, d_{BC} > d_{AB}$. In order to facilitate the interpretation of the data, it produces a graphical out that compares the phylogenetic tree with the Hasse diagram of the po-set derived from $\boldsymbol{\eta}$, Fig. 1. Points are positioned so that differences along the rate-axis are approximately proportional to differences in $\boldsymbol{\eta}$-values.

2.3. Loss of Phylogenetic Footprints

Relative rate tests can also be designed for more complex settings than substitution rates in homologous sequences. For example, the quantitative analysis of dynamical aspects of footprint loss and acquisition is complicated by the fact that individual regulatory DNA regions cannot be observed independently of sequence conservation. The reason is that phylogenetic footprinting [8, 9, 10, 11] always detects regulatory elements in (at least) pairs of sequences. As a consequence, even very simplistic models of footprint loss lead to rather sophisticated inference.

In the approach proposed in [12], *two* outgroups are required to first identify conserved sequence positions, before one tests for differential loss rates among two ingroup species. More precisely, consider a sub-tree of the following form:



(7)

Restricting the sequences to those positions for which $Y_i = C_i$ holds, we define

$$\begin{aligned} c_{CA} &= |\{i|Y_i = C_i = A_i\}|, \\ c_{CB} &= |\{i|Y_i = C_i = B_i\}|, \\ c_{CAB} &= |\{i|Y_i = C_i = A_i = B_i\}|\,. \end{aligned} \quad (8)$$

Note that $c_{CA} \geq c_{CAB}$ and $c_{CB} \geq c_{CAB}$ always holds. The number of conserved positions exclusively lost along the edge $A, \text{lca}(A, B)$ is $m'_A = c_{CB} - c_{CAB}$ and similarly, for $B, \text{lca}(A, B)$ we have $m'_B = c_{CA} - c_{CAB}$. One now tests whether m'_A and m'_B are significantly different. The corresponding matrix $\boldsymbol{\eta}$ has entries $\eta_{AB} = \ln(m'_A/m'_B)$ provided the difference is statistically signficant, and $\eta_{AB} = 0$, otherwise. For a fixed combination of outgroups Y, C, we immediately check that $m'_A - m'_{A'} > 0$ and $m'_{A'} - m'_{A''} > 0$ implies $m'_A - m'_{A''} > 0$. We therefore expect $\boldsymbol{\eta}$ to be acyclic. Since the choice of a different outgroup pair may lead to the selection of different conserved position, we cannot logically rule out contradictory test results in this case, however. The implementation of this test is currently in progress.

3. EXAMPLE

The expansion of a species in a heterogeneous environment can be correlated with relative rates of evolution in geographically separated subpopulations. The rate variation may be due to adaptation to different environmental conditions and due to changes in population size or structure [13]. Slowly evolving populations are typically large and stable, while small unstable populations exhibit higher evolution rates. Multiple waves of migration thus may lead to rate variations that show little correlation with phylogenetic position.

As an example of a real-life data set we consider here a recent comprehensive European-wide phylogeographical study of the arctic-alpine distribution of wolf spiders of the *Pardosa saltuaria* group [7]. The data, mitochondrial ND1 gene sequences, show a complex picture of rate differences, with some clear regularities.

For instance, the substitution rates are increased in almost all lineages relative to the samples from the the Pyrenees. This suggests that the Pyrenees served as glacial refugia. The rate correlation between the sequences of the Pyrenees and the Balkan individuals indicates a secound glacial refugium in the Balkan mountains. However, the data indicate migration out of the Pyrenees refugia only. The data set also reflects one further cold period with refugia in the Alps, Sudeten Mountains, and the Upper Tatra.

4. DISCUSSION

We have introduced here an a convenient way to visualize and summarize information on significant rate differences across larger phylogenetic data sets. The poset-approach seems convenient for the exploratory phase of data analysis. As it stands our tool does not attempt to correct for multiple testing, although a strategy such as Bonferroni's correction could easily be incorporated. We also note that the $\mathcal{O}(N^3)$ RRTs that can be performed within a given tree are of course not independent from each other. It might therefore be desirable to restrict attention to a less redundant set of tests.

5. ACKNOWLEDGMENTS

This work was supported in part by the DFG Bioinformatics Initiative and the 6th Framwork Programme of the European Union as part of the EDEN project (contract no. 043251).

6. REFERENCES

[1] J. Felsenstein, "Phylogenies from molecular sequences: inference and reliability," *Annu. Rev. Genet.*, vol. 22, pp. 521–565, 1988.

[2] F. Tajima, "Simple methods for testing molecular clock hypothesis," *Genetics*, vol. 135, pp. 599–607, 1993.

[3] C.-I. Wu and W.-H. Li, "Evidence for higher rates of nucleotide substitution in rodents than in man," *Proc. Natl. Acad. Sci. USA*, vol. 82, pp. 1741–1745, 1985.

[4] Z. Yang, "Maximum-likelihood models for combined analyses of multiple sequence data," *J. Mol. Evol.*, vol. 42, pp. 587–596, 1996.

[5] P. Li and J. Bousqet, "Relative-rate test for nucleotide substitutions between two lineages," *Mol. Biol. Evol.*, vol. 9, pp. 1185–1189, 1992.

[6] M. Robinson, M. Gouy, C. Gautier, and D. Mouchiroud, "Sensitivity of relative-rate tests to taxonomic sampling," *Mol. Biol. Evol.*, vol. 15, pp. 1091–1098, 1998.

[7] C. Muster and T. U. Berendonk, "Divergence and diversity: lessons from an arctic-alpine distribution (*Pardosa saltuaria* group, lycosidae)," *Mol. Ecol.*, vol. 15, pp. 2921–2933, 2006.

[8] D. A. Tagle, B. F. Koop, M. Goodman, J. L. Slightom, D. L. Hess, and R. T. Jones, "Embryonic epsilon and gamma globin genes of a prosimian primate (galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints," *J. Mol. Biol.*, vol. 203, pp. 439–455, 1988.

[9] C. Mayor, M. Brudno, J. R. Schwartz, A. Poliakov, E. M. Rubin, K. A. Frazer, L. S. Pachter, and I. Dubchak, "VISTA: visualizing global DNA sequence alignments of arbitrary length," *Bioinformatics*, vol. 16, pp. 1046–1047, 2000.

[10] M. Blanchette and M. Tompa, "Discovery of regulatory elements by a computational method for phylogenetic footprinting," *Genome Research*, vol. 12, pp. 739–748, 2002.

[11] S. Prohaska, C. Fried, C. Flamm, G. Wagner, and P. F. Stadler, "Surveying phylogenetic footprints in large gene clusters: Applications to Hox cluster duplications," *Mol. Phyl. Evol.*, vol. 31, pp. 581–604, 2004.

[12] G. P. Wagner, C. Fried, S. J. Prohaska, and P. F. Stadler, "Divergence of conserved non-coding sequences: Rate estimates and relative rate tests," *Mol. Biol. Evol.*, vol. 21, pp. 2116–2121, 2004.

[13] C. Stringer and R. McKie, *African Exodus: The Origins of Modern Humanity*, J.Macrae/H.Holt, New York, 1996.