Phylogenetics

SeqVis: Visualization of compositional heterogeneity in large alignments of nucleotides

Joshua W. K. Ho¹,², Cameron E. Adams¹, Jie Bin Lew¹, Timothy J. Matthews¹, Chiu Chin Ng¹, Arash Shahabi-Sirjani¹, Leng Hong Tan¹, Yu Zhao¹, Simon Easteal¹, Susan R Wilson¹ and Lars S Jermiin¹,²,*

¹School of Biological Sciences, ²Sydney University Biological Informatics and Technology Centre, Sydney, Australia, ³John Curtin School of Medical Research and ⁴Mathematical Sciences Institute, Australian National University, Canberra, Australia

Received on March 20, 2006; accepted on May 26, 2006
Advance Access publication June 9, 2006
Associate Editor: Keith A Crandall

ABSTRACT
Summary: Most phylogenetic methods assume that the sequences evolved under homogeneous, stationary and reversible conditions. Compositional heterogeneity in data intended for studies of phylogeny suggests that the data did not evolve under these conditions. SeqVis, a Java application for analysis of nucleotide content, reads sequence alignments in several formats and plots the nucleotide content in a tetrahedron. Once plotted, outliers can be identified, thus allowing for decisions on the applicability of the data for phylogenetic analysis.
Contact: lars.jermiin@usyd.edu.au

1 INTRODUCTION

Model-based phylogenetic methods usually assume that the aligned nucleotides evolved under stationary, reversible and homogeneous conditions [for definitions, see e.g. Jayaswal et al. (2005)]. If these conditions are violated by data, then the risk of phylogenetic errors is increased (Ho and Jermiin, 2004; Jermiin et al., 2004).

Alignments of nucleotides may vary compositionally in the sense that the composition may vary across sequences and/or across sites. In the first case, the sites would not have evolved under conditions that are stationary, reversible and homogeneous, and in the second case, the sites would have evolved under different stationary, reversible and homogeneous conditions. In both cases, it would be inappropriate to infer a phylogeny assuming that a single time-reversible Markov process underpins variation in the alignment.

Methods to detect compositional heterogeneity in alignments of nucleotides fall into four categories (Jermiin et al., 2004), with those of the first category using graphs or tables to visualize compositional heterogeneity, and those of the other categories producing test statistics that may be evaluated against expected distributions. Methods of the first category, however, are of limited use for surveys of alignments with many species [as in e.g. Hashimoto et al. (1995)] while methods of the other categories are either statistically invalid or not yet accommodated by the wider scientific community.

Inspired by the second problem, Ababneh et al. (2006) described several matched-pairs tests of homogeneity for analysis of aligned nucleotides. The tests are useful because they provide details on the Markov processes that may have operated during the divergence of sequences. However, surveying the results may be impractical if the data include many sequences or impossible for the matched-pairs test of marginal symmetry (Stuart, 1955) and internal symmetry (Ababneh et al., 2006) (because the estimation of the test statistics involves inverting a matrix that sometimes is singular), in which case a visual assessment of the data, preferentially combined with a matched-pairs test of symmetry (Bowker, 1948), may suffice. Here we present a solution to this visual assessment.

2 THE PROGRAM AND ITS FEATURES

We extended the de Finetti plot (Cannings and Edwards, 1968) to a tetrahedral plot with similar properties (i.e. each observation comprises four variables, a, b, c and d, where a + b + c + d = 1 and 0 ≤ a, b, c, d ≤ 1). Each axis in the plot starts at the center of a surface at value 0, and finishes at the opposite corner at value 1 (Fig. 1A). The nucleotide content of a given sequence is simply the list of shortest Markov process underpins variation in the alignment.

Methods to detect compositional heterogeneity in alignments of nucleotides fall into four categories (Jermiin et al., 2004), with those of the first category using graphs or tables to visualize compositional heterogeneity, and those of the other categories producing test statistics that may be evaluated against expected distributions. Methods of the first category, however, are of limited use for surveys of alignments with many species [as in e.g. Hashimoto et al. (1995)] while methods of the other categories are either statistically invalid or not yet accommodated by the wider scientific community.

Inspired by the second problem, Ababneh et al. (2006) described several matched-pairs tests of homogeneity for analysis of aligned nucleotides. The tests are useful because they provide details on the Markov processes that may have operated during the divergence of sequences. However, surveying the results may be impractical if the data include many sequences or impossible for the matched-pairs test of marginal symmetry (Stuart, 1955) and internal symmetry (Ababneh et al., 2006) (because the estimation of the test statistics involves inverting a matrix that sometimes is singular), in which case a visual assessment of the data, preferentially combined with a matched-pairs test of symmetry (Bowker, 1948), may suffice. Here we present a solution to this visual assessment.

We extended the de Finetti plot (Cannings and Edwards, 1968) to a tetrahedral plot with similar properties (i.e. each observation comprises four variables, a, b, c and d, where a + b + c + d = 1 and 0 ≤ a, b, c, d ≤ 1). Each axis in the plot starts at the center of a surface at value 0, and finishes at the opposite corner at value 1 (Fig. 1A). The nucleotide content of a given sequence is simply the list of shortest


distance between its point, P, in the tetrahedron and the each surface. Visual assessment of the spread of points in the tetrahedron shows the extent of compositional heterogeneity.

In order to study the nucleotide content of aligned sequences, we developed SeqVis, a Java application that displays the nucleotide composition of a set of sequences within a tetrahedron. SeqVis requires Java 3D package and Java Runtime Environment (version 5.0 or later). The program was tested on Windows XP and Mandrake Linux, and supports the following features:

- SeqVis reads and writes alignments in the sequential PHYLIP format the NEXUS format and the FASTA format.
- The tetrahedron can be rotated in all directions, animated and manipulated interactively; all items on display can be changed.
- By viewing the points orthogonally through one of the surfaces, the distribution of three nucleotides (e.g. C, G, T) may be assessed while ignoring the fourth nucleotide (i.e. A).
- The nucleotide composition at the codon sites can be surveyed independently and visualized on a single canvas.
LogDet method (Lockhart et al.) imposing the axes of a tetrahedron or a de Finetti plot. We surveyed these data thus corroborate Rokas et al.'s (1994) reason for using the de Finetti diagram.

**3 EXAMPLES**

Rokas et al. (2005) inferred a phylogeny of 32 eukaryotes using an alignment of 12,060 amino acids encoded by nuclear genes and discovered compositional heterogeneity among the sequences. We examined the corresponding alignment of nucleotides using SeqVis to get a better understanding of the data’s complexity (Fig. 1A and B). The spread of points was greatest at the third codon site but also visible at the first codon site. The matched-pairs test of symmetry showed that none of the codon sites could have evolved under stationary, reversible and homogeneous conditions. Given the structure of the genetic code, a similar conclusion must be drawn about the alignment of amino acids. The visual and statistical assessments of these data thus corroborate Rokas et al.’s (2005) reason for using the LogDet method (Lockhart et al., 1994).

Nucleotides may be recoded to reduce compositional heterogeneity (Woese et al., 1991). The effect of this may be visualized by superimposing the axes of a tetrahedron or a de Finetti plot. We surveyed the alignment of 23S ribosomal RNA molecules from Galtier et al. (1999), and we found that the RY- and KM-coding of nucleotides gave a tighter spread of the points than the SW-coding (Fig. 1C), thus indicating that the RY- and KM-coded alignments are more likely than the SW-coded alignment to be consistent with evolution under stationary, reversible and homogeneous conditions.

The two examples show that SeqVis is capable of surveying large sets of data. Compared with other visualization methods, like the one used by Hashimoto et al. (1995), SeqVis permits a more informative exploration of the nucleotide content. However, the spread of points should not be used alone in the assessment because it does not take into account the length of the sequences.

**ACKNOWLEDGEMENTS**

We thank S.-H. Hong and M. A. Charleston for constructive advice, and N. Galtier and A. Rokas for the data. The first eight authors contributed equally to this paper as part of two third-year Bioinformatics Projects at The University of Sydney.

**Conflict of Interest**: none declared.

**REFERENCES**


