BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 19 2009, pages 2611–2612
doi:10.1093/bioinformatics/btp454

Phylogenetics

PhyloDet: a scalable visualization tool for mapping multiple traits to large evolutionary trees

Bongshin Lee*, Lev Nachmanson, George Robertson, Jonathan M. Carlson and David Heckerman

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

Received on May 27, 2009; revised on July 14, 2009; accepted on July 15, 2009

Advance Access publication July 24, 2009

Associate Editor: Martin Bishop

ABSTRACT

Summary: Evolutionary biologists are often interested in finding correlations among biological traits across a number of species, as such correlations may lead to testable hypotheses about the underlying function. Because some species are more closely related than others, computing and visualizing these correlations must be done in the context of the evolutionary tree that relates species. In this note, we introduce PhyloDet (short for PhyloDetective), an evolutionary tree visualization tool that enables biologists to visualize multiple traits mapped to the tree.

Availability: http://research.microsoft.com/cue/phylodet/

Contact: bongshin@microsoft.com.

1 INTRODUCTION

Biomedical research often begins with a large-scale search for correlated traits. As the number of genetic sequences continues its exponential growth, there is an increasing interest in identifying underlying genetic causes of traits by comparing genetic sequences across a large number of species.

Felsenstein (1985) pointed out a key flaw in traditional comparative methods; namely, individual species cannot be considered statistically independent samples because of shared ancestry. For example, it should not be surprising to see a large number of correlations between mice and rats when the other species in the study are reptiles. Rather, it is when a correlation persists across a diverse range of species that we should take notice. Conversely, some traits are expected to be independent of the phylogeny and any clustering with respect to the tree may imply unexpected biological mechanisms or errors in data collection.

The evolutionary tree is typically inferred from genetic data and is annotated by branch lengths that indicate the genetic similarity between two species, with the leaves of the tree representing the species and internal nodes representing (unobserved) speciation events. Thus, the tree structure provides a natural visual representation of which species are generally similar (or different) to each other. Although dozens of visualizations for correlated traits have been developed, the recent explosive growth in the number of traits and species has created a need for a visualization that can scale to dozens of traits mapped to thousands of species and their evolutionary tree, allowing interactive exploration of complex interactions.

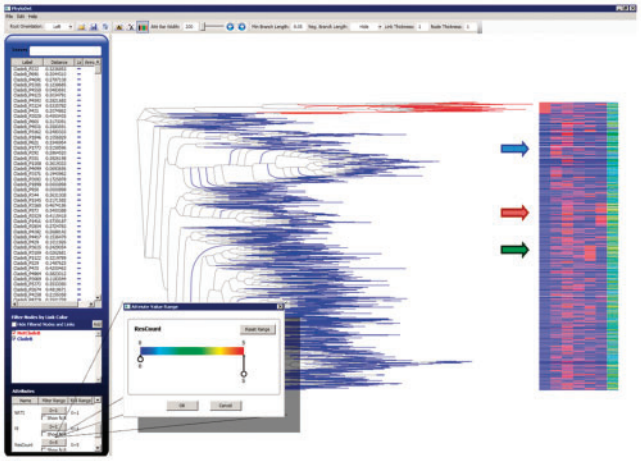

Fig. 1. PhyloDet, here showing six traits mapped to 1134 HIV sequences and their evolutionary tree. The arrows, added for illustration, highlight the clustering of attributes corresponding to IDU (blue), PI (red) and NRTI (green). The dialogue box allows a user to drag the circular tabs to specify a range of values that should be displayed. The current box is for the number of drugs to which an HIV sequence is resistant.

In this article, we present PhyloDet (Fig. 1), a scalable visualization tool for mapping multiple traits to large evolutionary trees. By allowing multiple traits to be visualized on a large tree that preserves branch lengths, biologists can visualize complex interactions and how they relate to the evolutionary history of the species.

2 EXAMPLE: HIV AND DRUG RESISTANCE

To demonstrate PhyloDet, we reanalyzed the HIV anti-retroviral (ARV) drug resistance data from Harrigan *et al.* (2005). In this study, the authors investigated the correlations between several risk factors and drug resistance to highly active anti-retroviral therapy (HAART), a cocktail of several ARV drugs, using a large cohort of 1191 patients who were initiating HAART for the first time. For each patient, the authors sequenced the infecting virus at several time points throughout the course of therapy and measured several attributes, including resistance to several classes of drugs and if the

*To whom correspondence should be addressed.

patient was an injection drug user. Correlations to HAART failure were measured using Cox proportional hazards; the underlying phylogeny that relates the HIV quasi-species was not considered.

Here, we selected one HIV sequence per patient (N = 1134) and constructed a phylogenetic tree relating HIV quasi-species. We then mapped each of the traits provided by Harrigan *et al.* (personal communications) to visualize interactions among the traits and between the traits and the evolutionary history (Fig. 1).

3 PHYLODET FUNCTIONALITIES

3.1 Color and shape coding for leaf nodes

Large cohorts often consist of heterogeneous sources and it can be useful to identify how those sources are distributed across the tree. For example, the extreme diversity of HIV-1 is often simplified by identifying individual sequences as belonging to one of several clades, which roughly map to continental regions. In our drug resistance example, coloring the leaf nodes and their connecting branches based on whether or not the strain is clade B (the predominant clade in North America, in blue) makes it immediately clear which sub-tree corresponds to non-clade B patients (red). The researchers may disregard results from this sub-tree, as they represent a substantially different population from the rest of the sequences. Leaf coloring may also be useful when multiple centers combine data, as it becomes immediately apparent how similar sequences are among different cohorts. In addition, the shape of leaf nodes can be customized to further highlight different leaves. A simple dialogue box allows users to specify colors and shapes that map to leaf names that contain user-defined substrings.

3.2 Visualization of multiple traits

To help biologists gain insight into the underlying mechanism, PhyloDet enables visualization of multiple traits (or attributes), which are mapped to the phylogenetic tree through use of a color-encoded attribute bar in which the columns correspond to the selected attributes and rows line up vertically with the leaves they represent. Each row is semi-transparent so that the attributes of leaves that overlap in vertical space can be distinguished. Preserving relative branch lengths means some leaves may be far from the attribute bar, making it difficult to identify which attributes correspond to a given leaf. Thus, the first column of the attribute bar repeats the leaf link colors and users can drag the entire attribute bar to precisely line up colors with leaves.

Attributes are assumed to be numeric, with colors assigned using a heat map that ranges from blue (smallest value) through green to red (largest value). The currently selected attributes are listed in the attributes list at the bottom of the left panel, which displays the range of observed values for each attribute. In addition, users can hide leaves that are missing data, or can specify a range of values for each attribute that should be displayed (see callout in Fig. 1).

In our example, Harrigan *et al.* tracked several attributes related to drug resistance or else relevant to the demographics of the cohort. In Figure 1, we display attributes corresponding to whether the patient was an injection drug user (IDU, second attribute column) and whether the infecting HIV quasi-species had already selected for one of several resistance mutations (starting at third attribute column: M184, NRTI, NNRTI or PI). The last attribute we display shows the total number of drug classes to which each patient is resistant (values range from 0 to 5). By displaying these attributes simultaneously on the tree, we can visualize several relationships reported by Harrigan *et al.*, as well as observe some new effects. For example, we see that HIV sequences sampled from IDUs tend to cluster together on the tree (blue arrow), an observation that may indicate that infection in the IDU population is largely circulating separately from the rest of the cohort. Interestingly, IDU status does not appear to be strongly correlated with resistance. Other attributes that contain strong tree-based clusters include NRTI (green arrow) and PI (red arrow) resistance mutations, whereas the remaining attributes appear to be evenly dispersed throughout the tree. It should be noted that the clustering of the NRTI and PI attributes may be caused by convergent evolution of the underlying mutations that define those attributes and not from shared ancestry. Thus, a careful analysis would include reconstructing the tree with the resistance sites removed to determine if the clustering is supported by non-resistance-associated mutations (Matthews *et al.*, 2009).

Finally, we note that PhyloDet is intended as a data exploration tool, and any derived hypotheses should be tested with statistical methods that can test for correlations in the context of evolutionary trees (e.g. Carlson *et al.*, 2008). For example, because the definition of which branch is drawn to the left or right for each node is arbitrary, apparent correlations that appear in small groups of closely related sequences may be an artifact of layout decisions.

3.3 Additional features and specifications

PhyloDet allows users to customize the appearance of the tree, allowing users to pick a new root, hide arbitrary sub-trees and specify parameters that determine the display properties of the tree. PhyloDet reads one tree file (in the Newick tree format) and multiple attribute files (in one of several tab-delimited text file formats). PhyloDet requires .NET 3.5, which is available for all windows platforms. Additional details are described at http://research.microsoft.com/cue/phylodet/manual.html and in Lee *et al.* (2008).

ACKNOWLEDGEMENTS

The authors wish to thank Philippa Matthews, Thomas Kuntzen, Toshi Miura, Yaoyu Wang, Jenna Rychert, Zabrina Brumme and Chanson Brumme for their feedback. We also would like to thank Richard Harrigan for providing his valuable data and comments for the HIV drug resistance example, and Meredith Skeels, who initiated the collaboration that made this project possible.

Conflict of Interest: none declared.

REFERENCES

Carlson,J.M., *et al.* (2008) Phylogenetic dependency networks: Inferring patterns of CTL escape and codon covariation in HIV-1 Gag. *PLoS Comput. Biol.*, **4**, e1000225.

Felsenstein,J. (1985) Phylogenies and the comparative method. *Am. Nat.*, **125**, 1–15.

Harrigan,P.R. *et al.* (2005) Predictors of HIV drug-resistance mutations in a large antiretroviral-naive cohort initiating triple antiretroviral therapy. *J. Infect. Dis.*, **191**, 339–347.

Lee,B. *et al.* (2008) Det. (Distance Encoded Tree): a scalable visualization tool for mapping multiple traits to large evolutionary trees. *MSR Tech Report MSR-TR-2008-97*, Microsoft Research.

Matthews,P.C. *et al.* (2009) HLA footprints on HIV-1 are associated with inter-clade polymorphisms and intra-clade phylogenetic clustering. *J. Virol.*, **83**, 4605–4615.

Weiss,R. (2003) HIV and AIDS: Looking Ahead. *Nat. Med.*, **9**, 887–891.