Tree concordance analysis: tree distance measures, clustering techniques, and model selection criteria (working title)

Kevin Gori[1], Nick Goldman[1], and Christophe Dessimoz*[1]

[1] EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Email: Christophe Dessimoz - dessimoz@ebi.ac.uk;

*Corresponding author

Abstract

Text for this section.

Background

1. generalities about tree incongruences
2. previous approaches & their limitations
   (a) process specific approaches
      i. LGT $->$ cite a bunch of methods
      ii. Incomplete lineage sorting $->$ e.g. Rannala & Yang 2003
      iii. BUT: limited to these specific processes and typically on other strong assumptions
   (b) Bayesian concordance analysis $->$ Bucky
      i. BUT: computationally expensive. Cannot go beyond a few taxa
   (c) Clustering approaches ("Tree of Trees")
      i. Tom Nye
      ii. Leigh et al. 2011
      iii. Darlu and Genoche 2011

iv. BUT: no statistical framework to work out the optimal distance/clustering method, and no insights into model selection

3. In this work

   (a) introduce a statistical framework for clustering-based concordance analysis

   (b) investigate which combination of distance and clustering method performs best in simulation and real data

   (c) investigate 3 model selection criteria to identify the number topologies (clusters):

Methods

Statistical framework

- show how the traditional ML model (Felsenstein 1981) can be extended to accomodate > 1 tree topology: index tree topologies from 1..K, where K is the number of cluster (or "classes"), the likelihood becomes

$$L_{global} = \prod_{\text{aligned gene } g \in G} L(D_g | M, T, \theta_g) = \prod_{\text{aligned gene } g \in G} L(D_g | T_g, \theta_g)$$

  where T is the set of K cluster topologies, M is a map between marker genes and T, T_g the topology in T corresponding to $M(g)$, and D_g the aligned sequences of marker gene g. Note that $L(D_g|T_g, \theta_g)$ is the traditional likelihood formula.

Tree distance measures

- list the various tree distance measures and explain how we computed them

We begin by calculating a distance matrix between the marker genes in the dataset, based on the ML tree obtained from each alignment. Distances were calculated using partition-based distance metrics

- Unweighted Robinson-Foulds
- Weighted Robinson-Foulds
- Euclidean Distance, and
- Geodesic Distance

The basis of these methods is to examine the set of splits defined by each tree. Each edge in the tree can be considered to partition the set of species the tree describes, S, into two non-overlapping subsets, S_A and S_B, where $S_A \cup S_B = S$. The union of these partitions is the set of splits for the tree. Given two trees on the same S a distance between them can be defined by the symmetric difference of the two sets of splits. This is the unweighted Robinson-Foulds distance between the trees, and takes an integer value between 0 and $2\,(|S| - 3)$. This measure is based solely on the topologies of the two trees, and takes no account of branch lengths. The weighted Robinson-Foulds distance and the euclidean distance extend this approach into the real-valued domain by taking branch-lengths into consideration. In the weighted Robinson-Foulds, the measure is the sum of the absolute differences in branch lengths for equivalent splits between both trees, with the branch length taken to be zero when the split is present in one tree and not the other. The euclidean distance is calculated similarly, but sums the squared difference between branch lengths, and takes the square root of the result. The geodesic distance of Billera, Holmes and Vogtmann is a more complex measure that explicitly takes topology and branch length into account. Under this scheme, a tree-space for $|S|$ species is embedded as a hypercube in real-valued space. All possible tree topologies are represented uniquely as points on the faces, called *orthants*, of the hypercube

Clustering methods

- list the various clustering methods and explain how we computed

Determining the number of classes

- Permutation analysis
- Goldman-Cox method
- Cluster-based approach (à la Leigh et al. 2011)

Results
Discussion and Outlook

Author's contributions

Text for this section ...

Acknowledgements

Text for this section ...