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

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

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

Tree distance measures

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

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 ...