reconciling trees

Stefanie Hartmann
postdoc, Todd Vision’s lab
University of North Carolina

the data

• alignments and phylogenies for ~27,000 gene families from 140 plant species

www.phytome.org

publicly available DNA sequence data

813,840 protein sequences

26,870 gene families and 19,483 singletons
the goal

- reconcile the species tree with the gene trees to infer duplication / speciation nodes
the major challenge

• species tree and/or gene tree is unresolved

<table>
<thead>
<tr>
<th>species tree</th>
<th>gene tree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
minor challenges

• gene tree(s) must be rooted
  - midpoint rooting assumes a molecular clock
  - molecular clock assumption is often violated
• species tree must be rooted
• species names/IDs must be substrings of sequence names/IDs
• labeling and visualization of results?

available software
(gene tree / species tree reconciliation)

SDI
• Java, Perl
• gui, command line
• gene tree and species tree have to be fully resolved
• input/output: rooted trees in newick, NHX format
• http://www.phylogenomics.us/forester/
available software
(gene tree / species tree reconciliation)

Notung
- Java
- gui, command line
- input/output: rooted trees in newick, NHX, or Notung format
- species tree has to be resolved,
- the gene tree can be unresolved if it contains support values, output gene tree is resolved

available software
(gene tree / species tree reconciliation)

softparsmap
- Java, Jooc (XML tags)
- command line
- allows both gene tree and species tree to be unresolved
- input species tree: the NCBI taxonomy
- input gene trees: unrooted newick trees with support values
- input sequence: darwin database file (including GI number of GenBank entry)
- additional format requirements
- http://www.cbu.uib.no/~steffpar/softparsmap/
available software
(gene tree / species tree reconciliation)

softparsmap

through a speciation event.

With some additional Java programming the package can be adapted to almost any source of data. However, existing interrace retrieve species trees from NCBI Taxonomy, sequence data from an XML file, the gene trees from one file per family, and the results are written to one or two files per family.

available software
(gene tree / species tree reconciliation)

Eulenstein software
• species tree has to be fully resolved, gene tree can be unresolved (?)

GeneTree
• http://taxonomy.zoology.gla.ac.uk/rod/genetree/genetree.html
• available only for Mac OS 9 PPC, not scriptable

Mesquite component (http://mesquiteproject.org/mesquite/mesquite.html)

gtp (http://ginger.ucdavis.edu/gtp/gtp.html)

primetv, reconcile (http://prime.sbc.su.se/primetv/index.html)
Motivation
• reconcile a species tree with a gene tree to infer duplication/speciation nodes

Key Challenges
• gene tree(s) and/or species tree are often unresolved
• most software requires rooted input trees

Precondition
• user has gene tree and species tree in newick (or NHX or Schreiber) format. Species names or identifiers in species tree are substrings of sequence names

Steps
• adjust input format for gene tree(s), species tree, sequences, etc
• customize software (e.g., softparsmap) for data input (e.g., remove requirement for GI Number for sequences, allow sequence names to consist of integers and numbers)
• run software, parse output if necessary

Results
• labeling of nodes on gene tree as "duplication", "speciation", "unknown"

literature

SDI

Notung
• Durand D, Halldorsson BV, Vernot B (2005) A hybrid micro-macroevolutionary approach to gene tree reconstruction. RECOMB2005 link to the paper on Dannie Durand's website
literature

softparsmap


Eulenstein software


GeneTree


literature

reviews or other relevant articles