

reconciling trees

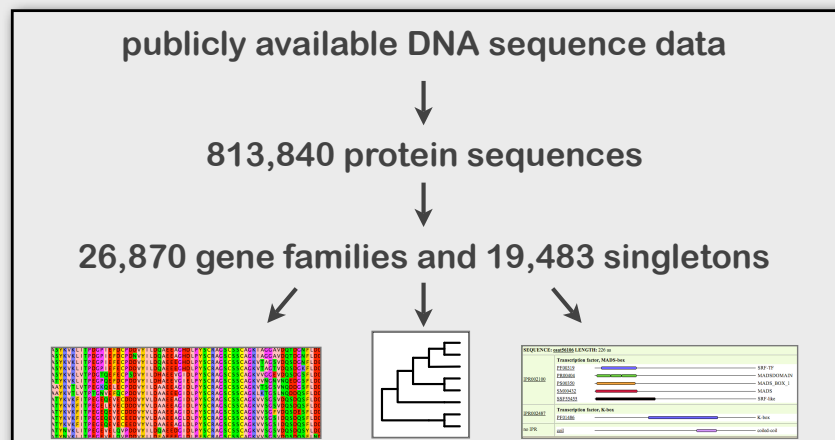
Stefanie Hartmann
postdoc, Todd Vision's lab
University of North Carolina

1

the data

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

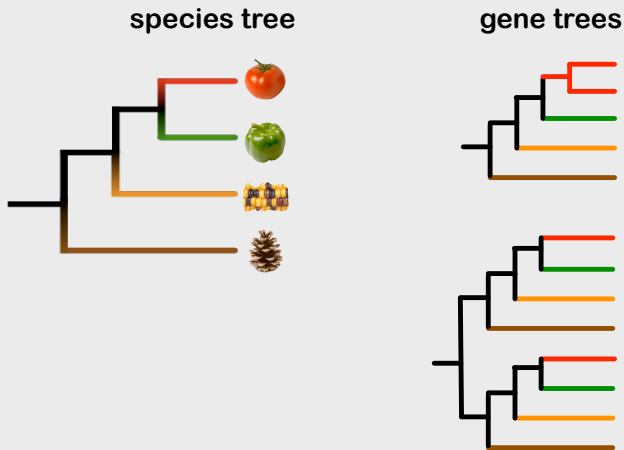
www.phytome.org



2

the goal

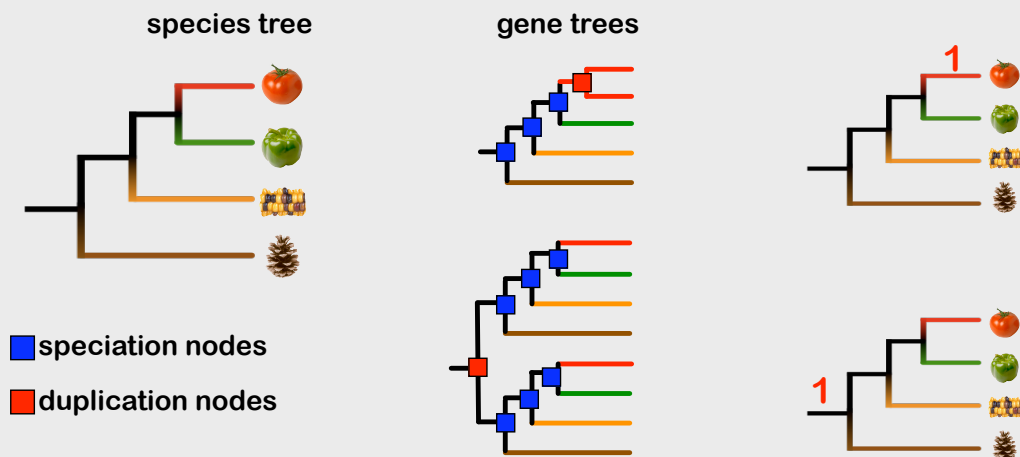
- reconcile the species tree with the gene trees to infer duplication / speciation nodes



3

the goal

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



3

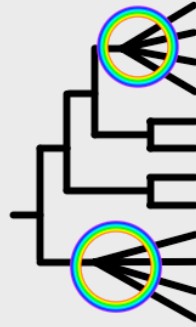
the major challenge

- species tree and/or gene tree is unresolved

species tree



gene tree



4

the major challenge

- species tree and/or gene tree is unresolved

species tree	gene tree
<p>A phylogenetic tree with a root on the left and two main branches extending to the right. Each main branch further divides into two smaller branches, resulting in a fully resolved tree.</p>	<p>A phylogenetic tree with a root on the left and two main branches extending to the right. Each main branch further divides into two smaller branches, resulting in a fully resolved tree.</p>
<p>A phylogenetic tree with a root on the left and two main branches extending to the right. Each main branch further divides into two smaller branches, resulting in a fully resolved tree.</p>	<p>A phylogenetic tree with a root on the left and two main branches extending to the right. Each main branch further divides into three smaller branches, resulting in two polytomies. One of the polytomies is highlighted with a multi-colored circle (red, green, blue).</p>
<p>A phylogenetic tree with a root on the left and three branches extending to the right. The three branches are highlighted with a multi-colored circle (red, green, blue), indicating an unresolved node.</p>	<p>A phylogenetic tree with a root on the left and two main branches extending to the right. Each main branch further divides into two smaller branches, resulting in a fully resolved tree.</p>
<p>A phylogenetic tree with a root on the left and three branches extending to the right. The three branches are highlighted with a multi-colored circle (red, green, blue), indicating an unresolved node.</p>	<p>A phylogenetic tree with a root on the left and two main branches extending to the right. Each main branch further divides into three smaller branches, resulting in two polytomies. One of the polytomies is highlighted with a multi-colored circle (red, green, blue).</p>

5

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?

6

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/>

7

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
- <http://goby.compbio.cs.cmu.edu/notung/index.html>

8

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/>

9

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 interfaces 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.

10

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>)

11

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"

12

literature

SDI

- Zmasek CM, Eddy SR (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 17: 821-828.

Notung

- Chen C, Durand D, Farach-Colton M (2000) Notung: dating gene duplications using gene family trees. RECOMB2000, Fourth Annual International Conference on Computational Molecular Biology
- Chen K, Durand D, Farach-Colton M (2000) NOTUNG: A program for dating gene duplications and optimizing gene family trees. *Journal of Computational Biology* 7: 429-447.
- 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
- Durand D, Halldorsson BV, Vernot B (2006) A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol* 13: 320-335.)

13

literature

softparsmap

- Berglund-Sonnhammer AC, Steffansson P, Betts MJ, Liberles DA (2006) Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *J Mol Evol* 63: 240-250.

Eulenstein software

- Chang, Wen-Chieh and Eulenstein, Oliver (2005) Reconciling Gene Trees with Apparent Polytomies. Technical Report, Computer Science, Iowa State University. link to technical report at IA State

GeneTree

- Page RDM (1998) GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics* 14: 819-820.

14

literature

reviews or other relevant articles

- Arvestad L, Berglund A-C, Lagergren J, Sennblad B (2004) Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. *Recomb 2004* link to pdf of article on prime website
- Cotton JA, Page RDM (2004) Tangled tales from multiple markers. In: Bininda-Emonds ORP, editor. *Phylogenetic Supertrees. Combining information to reveal the Tree of Life*. Springer. link to pdf of article on James Cotton's website
- Driskell AC, Ane C, Burleigh JG, McMahon MM, O'meara BC, et al. (2004) Prospects for building the tree of life from large sequence databases. *Science* 306: 1172-1174.
- Eulenstein O, Chen DH, Burleigh JG, Fernandez-Baca D, Sanderson MJ (2004) Performance of flip supertree construction with a heuristic algorithm. *Systematic Biology* 53: 299-308.
- Arvestad L, Berglund A-C, Lagergren J, Sennblad B (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* 19: 7i-15.
- Rydin C, Kallersjo M (2002) Taxon sampling and seed plant phylogeny. *Cladistics-the International Journal of the Willi Hennig Society* 18: 485-513.
- Simmons MP, Freudenstein JV (2002) Uninode coding vs gene tree parsimony for phylogenetic reconstruction using duplicate genes. *Molecular Phylogenetics and Evolution* 23: 481-498.
- V'Yugin VV, Gelfand MS, Lyubetsky VA (2002) Tree reconciliation: Reconstruction of species phylogeny by phylogenetic gene trees. *Molecular Biology* 36: 650-658.
- Simmons MP, Donovan Bailey C, Nixon KC (2000) Phylogeny Reconstruction Using Duplicate Genes. *Molecular Biology and Evolution* 17: 469-473.
- Slowinski JB, Knight A, Rooney AP (1997) Inferring species trees from gene trees: A phylogenetic analysis of the elapidae (Serpentes) based on the amino acid sequences of venom proteins. *Molecular Phylogenetics and Evolution* 8: 349-362.

15