Tree of Life Grand Challenge Project: Gene Tree Reconciliation

Todd Vision

iPlant Engagement Team Meeting

1 July 2009

iPlant™ Collaborative

Empowering a new plant biology

Outline

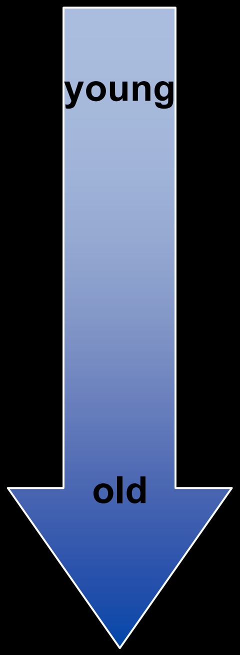
- Key motivations and problems for tree reconciliation
- Bottlenecks and checkpoints
- Applicable existing tools and software components
- Role of the postdoc/superuser
- iPToL expectations of the engagement team and iPlant developers

Two post-tree analysis priorities from Nov 08 GC workshop

- Trait evolution
  - See previous presentation by Brian O'Meara
  - Target audience: ecology, evolution & organismal biologists
- Tree reconciliation
  - Use a species tree to interpret a gene family tree (or vice versa)
  - Target audience: molecular, cellular, developmental biologists

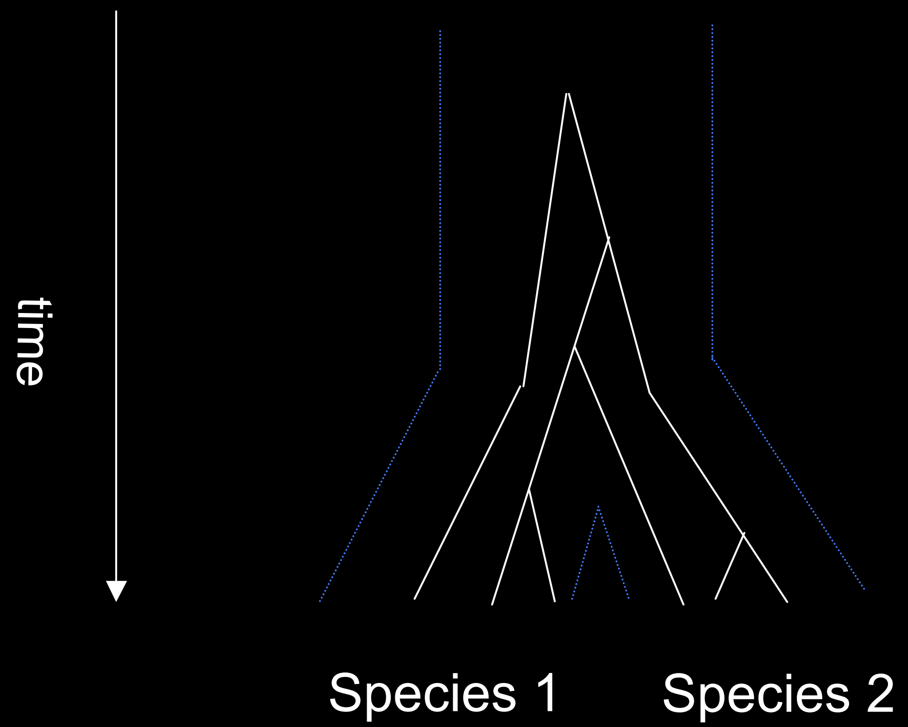
Incongruence: when gene trees differ from species trees

Timescale

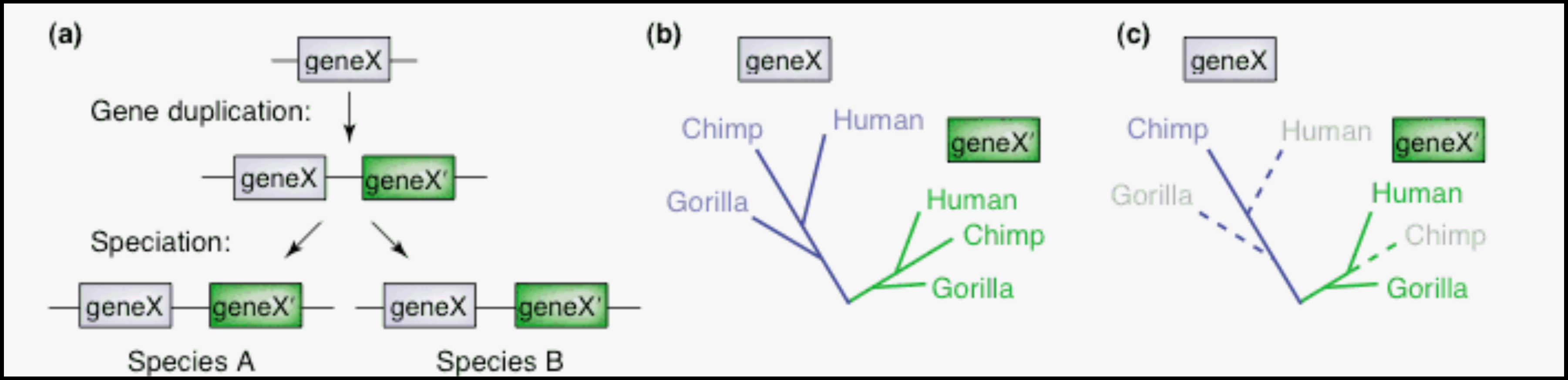


- Lineage sorting & hybridization
- Gene duplication (and loss)
- Horizontal transfer (incl. endosymbiotic transfer)

Lineage sorting

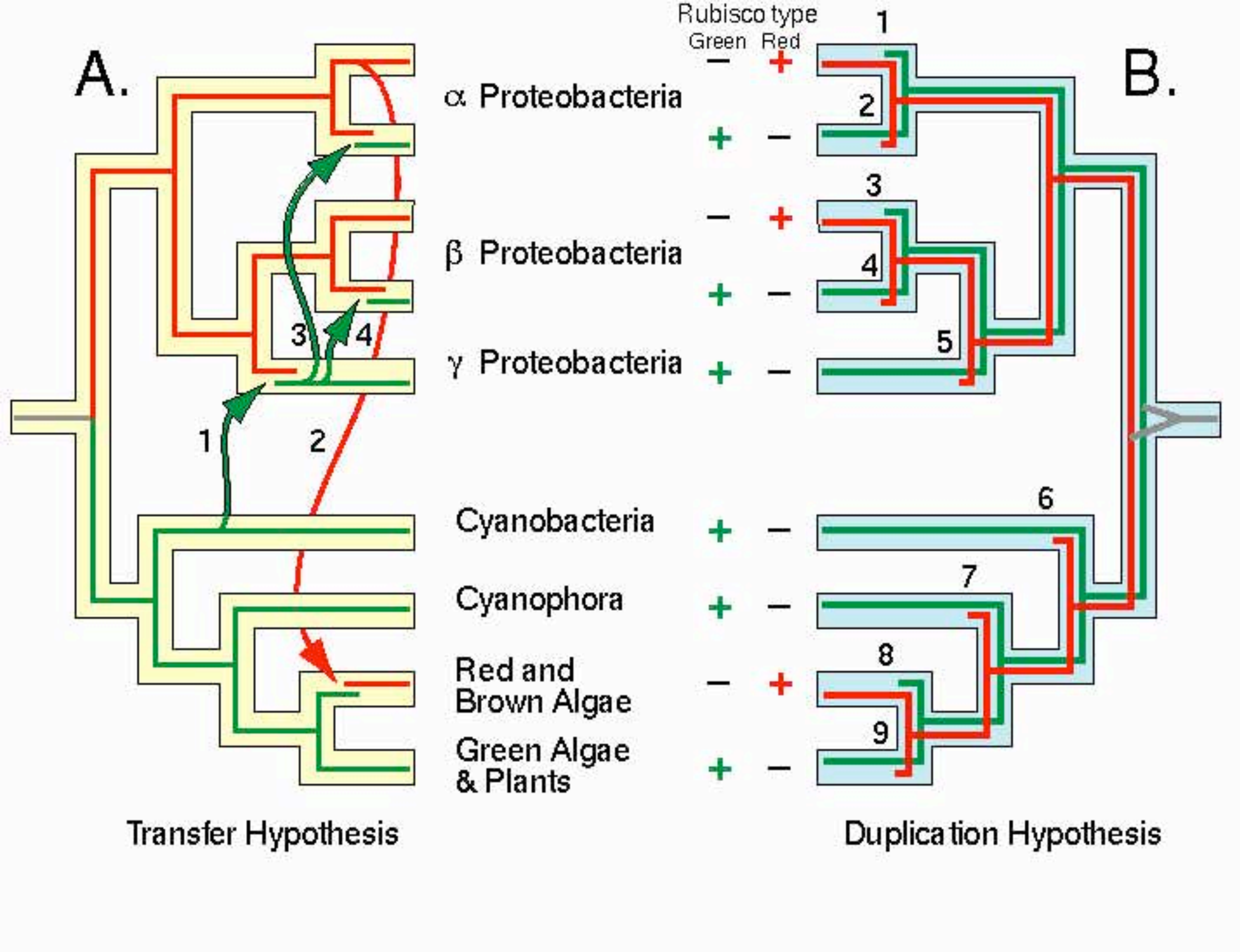


Gene duplication and loss



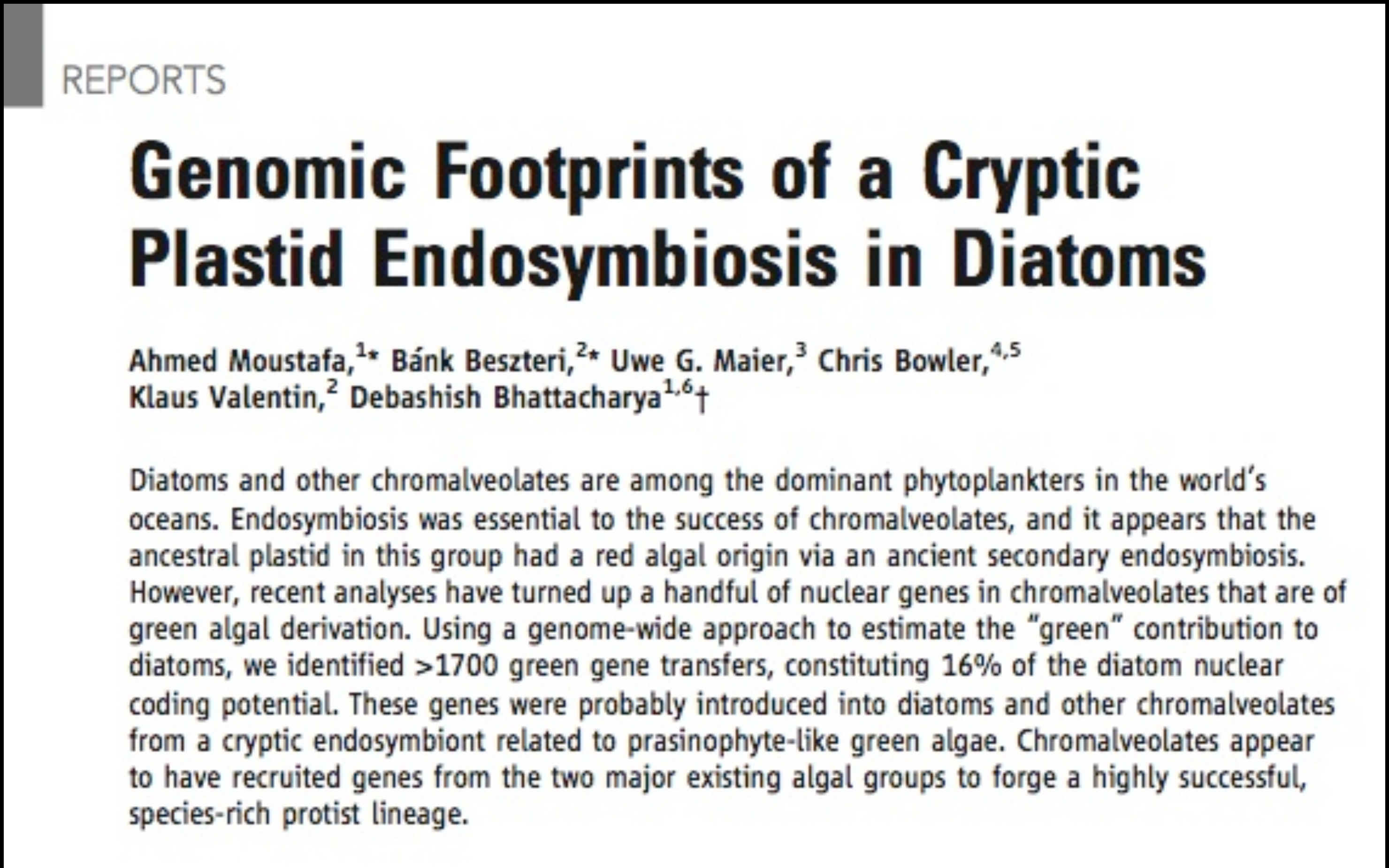
- Homologs: genes descended from a common ancestor
- Reconciliation allows you to distinguish two kinds of homologs
  - Orthologs: diverged through speciation
  - Paralogs: diverged through duplication, whether or not they are in the same genome

Horizontal transfer



Delwiche CF and Palmer JD (1996) Mol Biol Evol: 873-882.

Endosymbiotic transfer



REPORTS

Genomic Footprints of a Cryptic Plastid Endosymbiosis in Diatoms

Ahmed Moustafa,[1]* Bánk Beszteri,[2]* Uwe G. Maier,[3] Chris Bowler,[4,5]
Klaus Valentin,[2] Debashish Bhattacharya[1,6]†

Diatoms and other chromalveolates are among the dominant phytoplankters in the world's oceans. Endosymbiosis was essential to the success of chromalveolates, and it appears that the ancestral plastid in this group had a red algal origin via an ancient secondary endosymbiosis. However, recent analyses have turned up a handful of nuclear genes in chromalveolates that are of green algal derivation. Using a genome-wide approach to estimate the "green" contribution to diatoms, we identified >1700 green gene transfers, constituting 16% of the diatom nuclear coding potential. These genes were probably introduced into diatoms and other chromalveolates from a cryptic endosymbiont related to prasinophyte-like green algae. Chromalveolates appear to have recruited genes from the two major existing algal groups to forge a highly successful, species-rich protist lineage.

Science, 26 June 2009

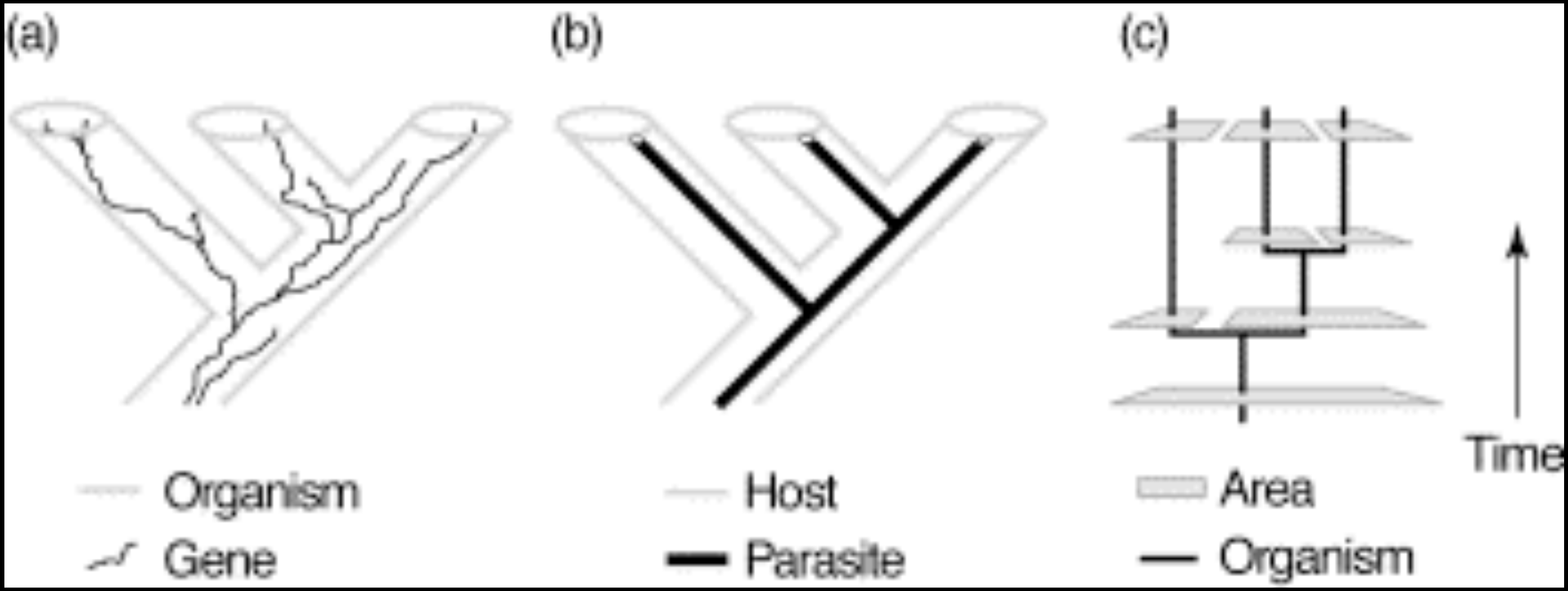
Gene tree reconciliation (GTR)

- Projection of a species tree onto a gene tree
- Inferring duplications (and optionally losses)
  - With incomplete genomes, losses are ambiguous
  - Most frequent objective function is parsimony
  - Probabilistic methods not yet fast enough for practical applications
- Lineage sorting, horizontal transfer
  - Good recent algorithmic work, but a (mostly) separate literature
  - The former mostly of interest to biogeographers, the latter to microbiologists

Some applications of GTR

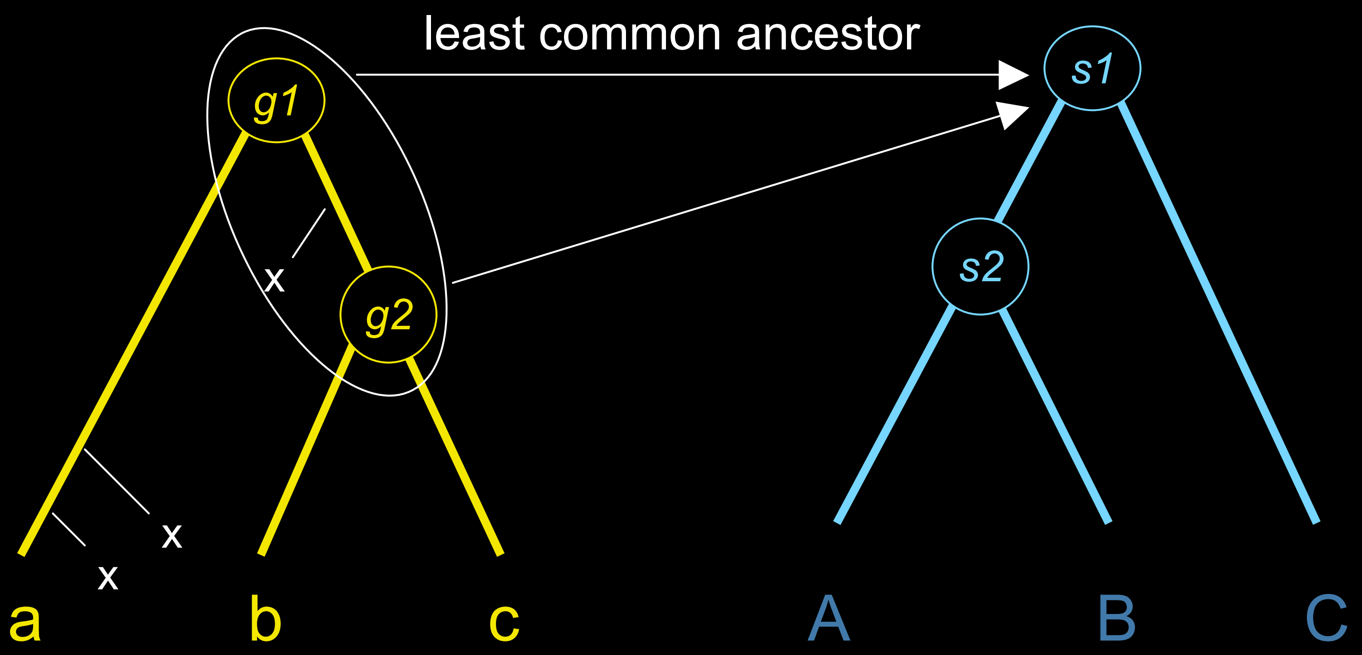
- What nodes are duplications (and what missing nodes represent losses)?
- Which sets of genes are orthologous?
- What was the complement of genes in a given ancestral species?
- What is the (rooted) species tree?
- Where are the phylogenetic positions of ancient polyploidy events (and how many duplicates have survived)?
- Are gene families coevolving (and thus potentially interacting)?

Viewing reconciliation more generally



Page and Charleston 1998

g1 is a duplication, and there were 3 losses



gene tree

species tree

Complications

- Polytomies (in either species or gene tree)
- Uncertainty in rooting (particularly the gene tree)
- Algorithm performance has not been thoroughly tested
  - Speed on large trees
  - Accuracy (particularly if incongruence is not due to one factor alone)
- Confidence measures are generally lacking

Bottlenecks & checkpoints

- Obtaining conservatively resolved, rooted species tree(s), possibly w/ branch lengths
- Obtaining gene tree(s) with confidence values, optionally rooted, from online sources (or calculating them…)
  - Determining user needs for gene tree metadata (to enable search & retrieval)
  - Enabling user upload
- Aligning taxonomic identifiers between the species & gene trees
- Determining algorithm(s) for reconciliation
  - Rooting
  - Confidence values
  - Objective function (dup only, dup+loss, lineage sorting, hybridization, horizontal transfer, etc)
  - Speed and accuracy
  - Deciding on extent of user options
- Determining user needs for analysis results (orthology, ancestral gene content, domain evolution, etc)
- Formatting, visualization, exchange of results

By date	(Very) preliminary project milestone
10/09	Assemble/recruit team, gather requirements, thoroughly review existing tools & data sources, design simulation engine
1/10	Benchmark scalability and accuracy of existing tools, begin system and interface design (eg mockups, user feedback), define work needed for scalability/visualization/data exchange technology/other needs.
4/10-7/10	Implementation begins
10/10	Scalability/visualization/data exchange solutions implemented.
1/11	Begin analysis on comprehensive gene family dataset ("marquee analysis" Begin user testing
4/11	Submission of "marquee analysis". Begin work on training materials
7/11	Disseminate at workshops & conferences

Applicable existing tools and software components

- NOTUNG (Dannie Durand’s group) - most full featured and well-maintained software
  - Includes a version of Zmasek’s ATV viewer for visualization
- Zmasek & Eddy (2001) fast heuristic algorithm implemented in a few tools (e.g. TreeFam)
- SoftParsMap (from David Liberles group) - deals with polytomies differently

Existing visualization approaches

Notung / ATV

http://www.cs.cmu.edu/~durand/Notung/

e.g. princeton protein orthology DB

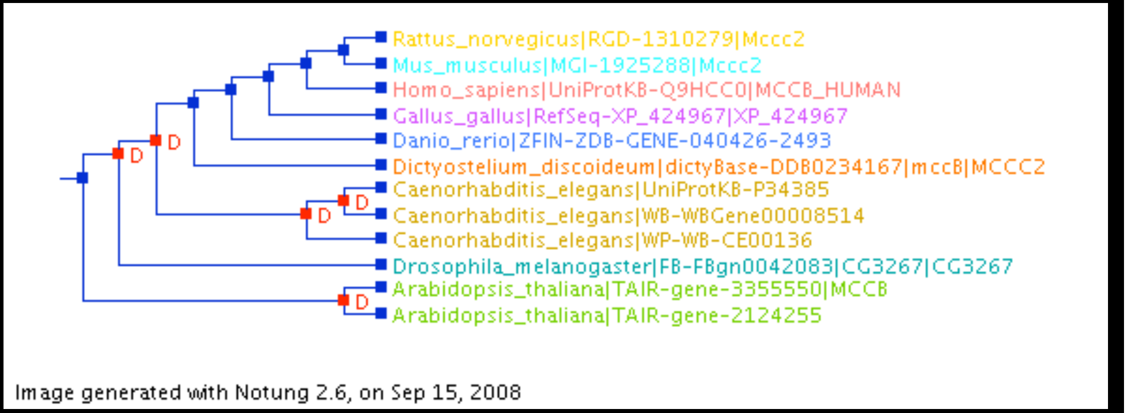
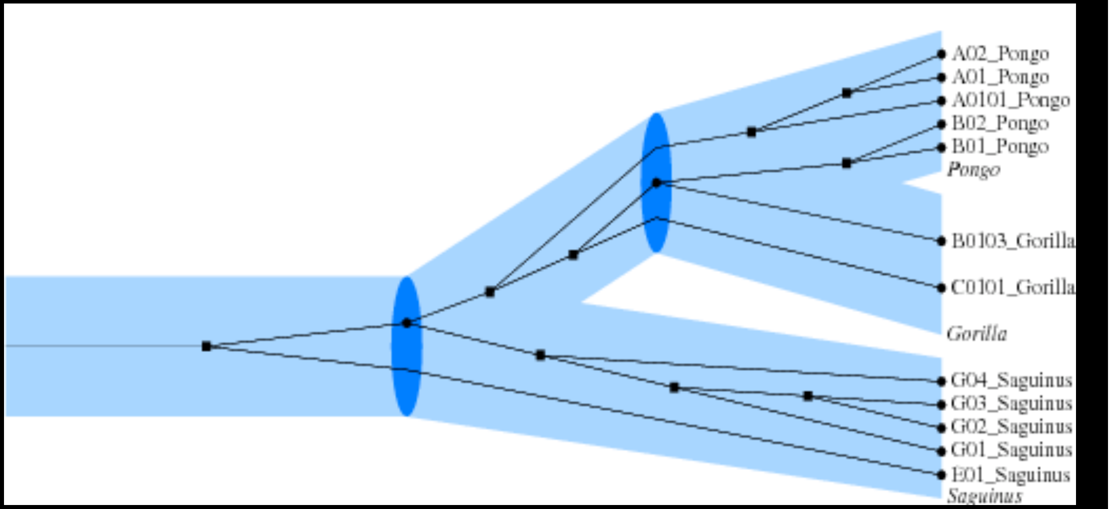

Rattus_norvegicus|RGD-1310279|Mccc2
Mus_musculus|MGI-1925288|Mccc2
Homo_sapiens|UniProtKB-Q9HCC0|MCCB_HUMAN
Gallus_gallus|RefSeq-XP_424967|XP_424967
Danio_rerio|ZFIN-ZDB-GENE-040426-2493
Dictyostelium_discoideum|dictyBase-DDB0234167|mccB|MCCC2
Caenorhabditis_elegans|UniProtKB-P34385
Caenorhabditis_elegans|WB-WBGene00008514
Caenorhabditis_elegans|WP-WB-CE00136
Drosophila_melanogaster|FB-FBgn0042083|CG3267|CG3267
Arabidopsis_thaliana|TAIR-gene-3355550|MCCB
Arabidopsis_thaliana|TAIR-gene-2124255
D
D
D
D
D

Image generated with Notung 2.6, on Sep 15, 2008

PrimeTV

http://prime.sbc.su.se/



Plant gene tree databases

- Comprehensive for plants (*i.e.* includes EST data)
  - Phylota http://loco.biosci.arizona.edu/pb/
  - Phytome http://phytome.org
  - PlantTribes http://fgp.bio.psu.edu/tribedb/

- Non-comprehensive
  - PhyloFacts http://phylogenomics.berkeley.edu/phylofacts
  - Phytozome http://www.phytozome.net
  - TreeFam (Metazoan only) http://www.treefam.org/

Role of the postdoc

- Consultant and client to developers
  - Help in requirement gathering and drafting specs
  - Testing
  - Contributing to end-user documentation
  - Benchmarking speed and accuracy of different algorithms and implementation options
- Performing “marquee” analysis using comprehensive gene family set
- Presenting outcomes at conferences and in papers

Expectations of the engagement team and iPlant developers

- Of iPlant as a whole
  - Aggressive project management & cross-WG coordination
- Of the developers
  - Expertise in technologies to be deployed
  - Acquisition of requisite domain knowledge
  - Design charettes and feedback cycles with external users
  - Open & iterative development
    - Mailing lists, a public website with syndicated news, etc.
    - Weekly status checks and frequent opportunities for feedback
    - Releasing software (incl. source code) early & often
- Of the WG scientists
  - Service orientation
  - Algorithm agnosticism
  - An interest in rigorously benchmarking scalability and accuracy

Suggested reading

- Gene duplication/loss parsimony: Durand et al (2006) J. of Comp. Biol. 13(2): 320. (& recent paper by Vernot et al on non-binary species trees)
  - Alternative heuristic: Zmasek & Eddy (2001) Bioinformatics 17, 821.
- Fitting a probabilistic gene duplication/loss model: Arvestad et al. (2004) Proc. 8th Ann. Int. Conf. on Computational Mol Biol 326.
- Horizontal transfer: Hallett, M et al. (2004) Proc..8th Ann. Int. Conf. on Computational Mol Biol, 347. (includes duplication)
- Lineage sorting: see recent work by Lacey Knowles, Scott Edwards, & Cecile Ane.
- Species trees: Page & Cotton (2002) Pacific Symposium on Biocomputing 7, 536.
- Genome duplication: Bansal and Eulenstein(2008) Bioinformatics doi:10.1093/bioinformatics/btn150 (& Burleigh et al, submitted)
- Comparisons to host-parasite & biogeographic reconciliation: Page & Charleston (1998) Tr. Ecol Evol. 13, 356.