Bioinformatics at TIGR in support of AToL

Jonathan Badger

The Institute for Genomic Research

(TIGR)

Two main systems

- APIS – Automatic Phylogenetic Inference System
  - Automates sequence similarity, alignment, phylogenetic inference, and tree interpretation.
- ECfinder
  - Uses Eisen's concept of "phylogenomics" to make functional assignments based on phylogenetic analyses.

What is APIS?

- APIS automatically generates and summarizes phylogenetic trees for each gene in a genome.
- Large scale analyses help answer (in a way that individual gene trees cannot) questions about gene function, species phylogeny, and the origin of gene families.
- Results are viewable on an internal TIGR web server.
- Portable version for installation elsewhere being developed.

APIS Outline

BLASTP all proteins in genome against ComboDB

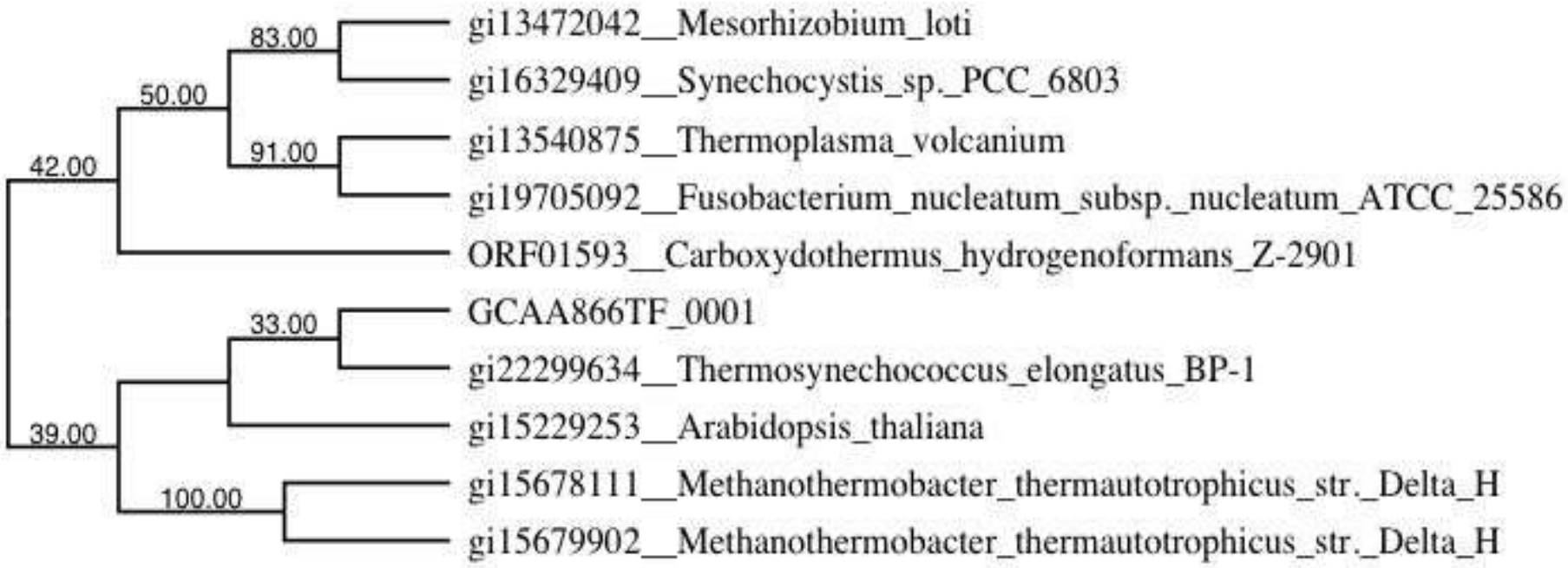
Extract full length homologs from ComboDB

Multiple Sequence Alignment (MUSCLE)

Phylogenetic Inference (currently bootstrapped NJ using QuickTree)

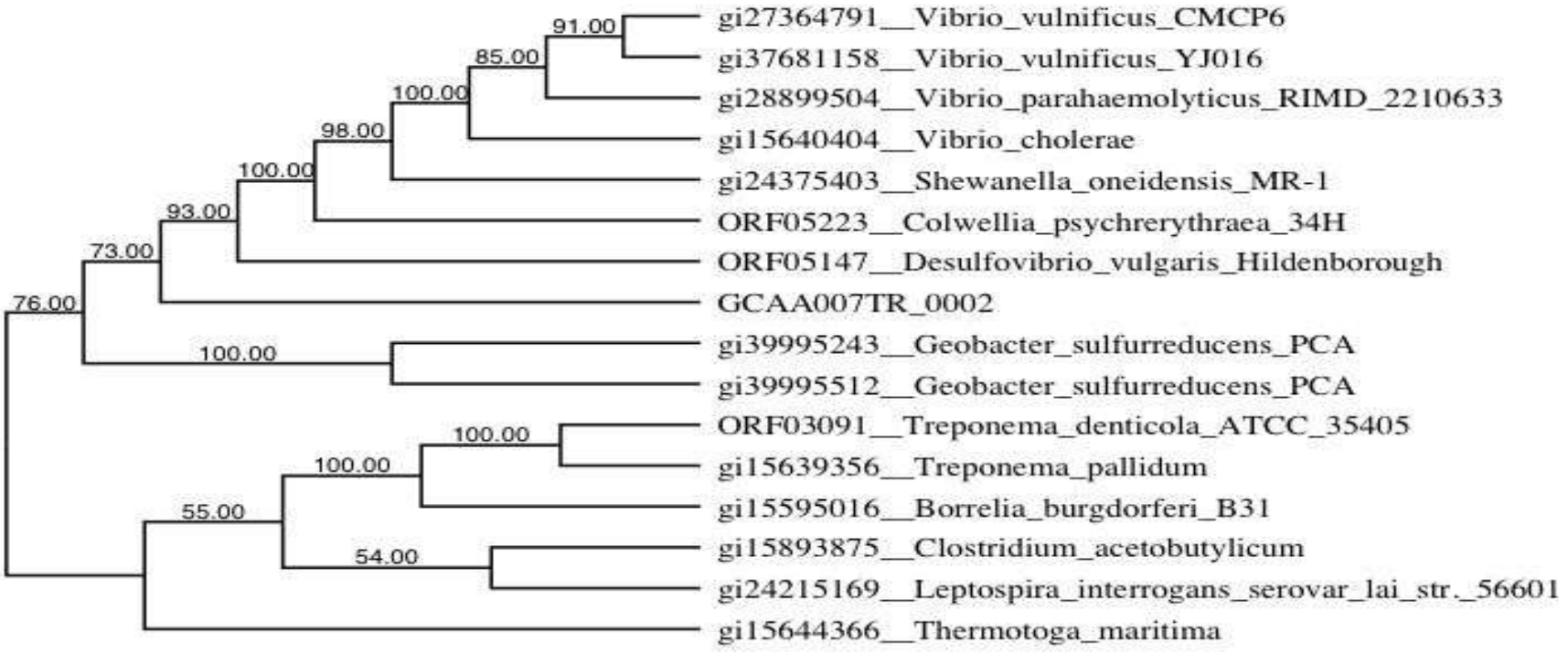
Determination of most related groups

Determination of Most Related Group - I



GCAA866TF_0001 is “Outgroup of Bacteria”

Determination of Most Related Group - II



GCAA007TR_0002 is “Contained in Bacteria”

APIS: Phylogenomic breakdown of proteins from *Thermodesulfobacterium commune*

Search for ORF # or description

Paralogs

Kingdom

Contained within Bacteria 1050 (75.6%)
Outgroup of Bacteria 210 (15.1%)
Closest relative is unresolved at kingdom level 65 (4.7%)
Outgroup of Archaea 40 (2.9%)
Contained within Archaea 11 (0.8%)
Outgroup of Eukaryota 2 (0.1%)
Outgroup of Yellowstone_Cyanobacteria 1 (0.1%)

Phylum

Closest relative is unresolved at phylum level 322 (23.2%)
Outgroup of Nitrospirae 270 (19.4%)
Outgroup of Proteobacteria 203 (14.6%)
Contained within Proteobacteria 105 (7.6%)
Outgroup of Aquificae 104 (7.5%)
Outgroup of Thermodesulfobacteria 71 (5.1%)
Outgroup of Firmicutes 60 (4.3%)

What is ECfinder?

- System similar to APIS but uses as database the EC (Enzyme Commission) annotated database of the Japanese KEGG project.
- EC numbers are “call numbers” for enzymes, allowing consistent annotation
- Allows functional characterizations to be made using Jonathan Eisen’s concept of “phylogenomics”

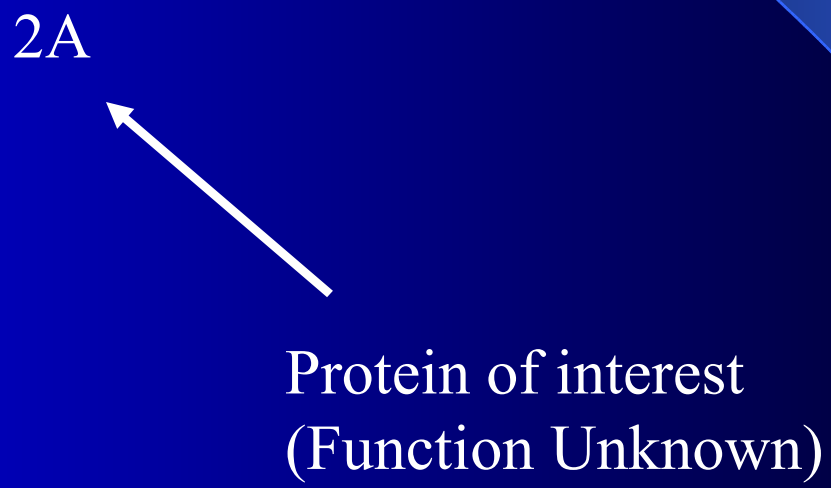
Why not just use BLAST?

- BLAST can't handle:
  - Gene duplication / diversification
    - Many genes in a genome known to be related to each other. After duplication, one copy retains old function and other copy is free to acquire new function (because not under any evolutionary constraint).
  - Domain shuffling
    - Domains are functional units of proteins (for example, DNA binding sites). By partial gene duplication and rearrangement, new functions can be generated out of existing parts.

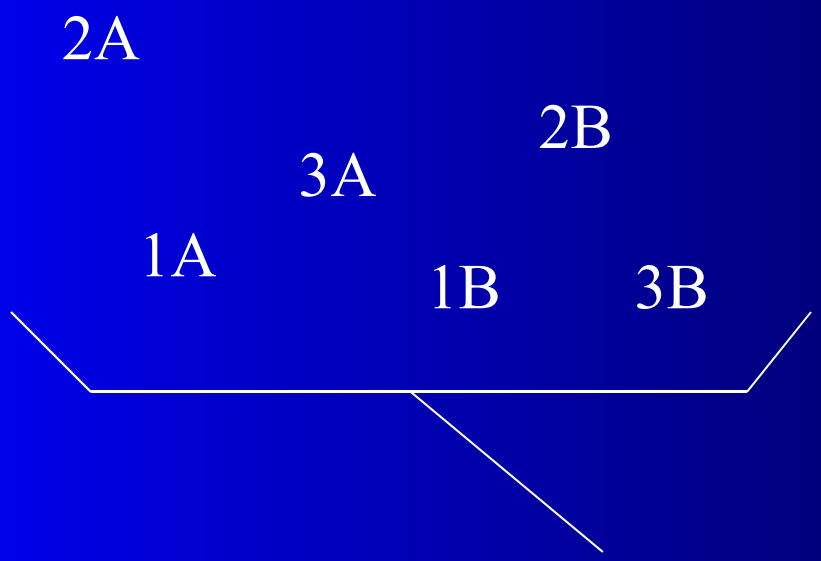
Enter “Phylogenomics”

- Term coined by Jonathan Eisen in 1998
- Basic assumptions
  - Protein function is conserved by natural selection over time (with new functions arising through gene duplications).
  - We can reconstruct evolutionary history of sequences.
- Conclusion
  - We can predict the function of a protein by inferring a phylogenetic tree of it and related proteins. By looking at known functions, we can make a functional prediction consistent with the tree.

Phylogenomics: An Example

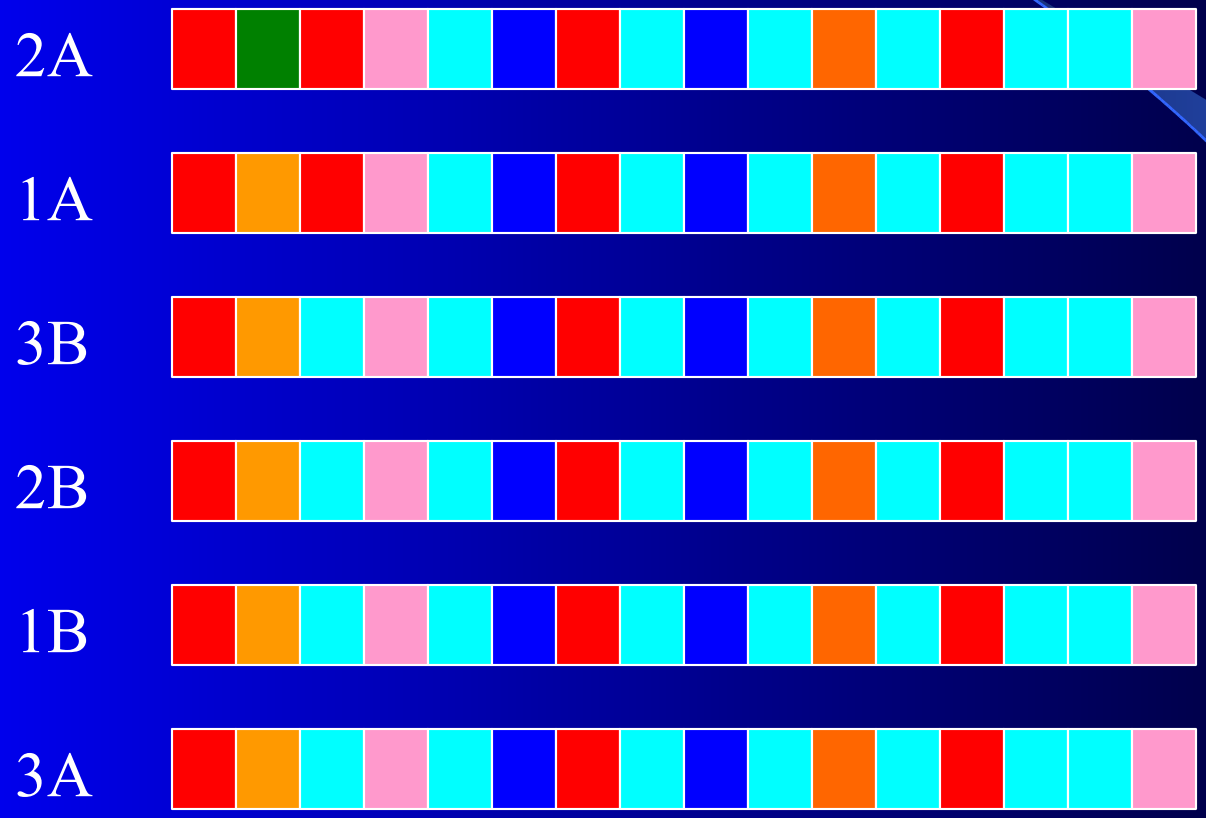


Phylogenomics: An Example



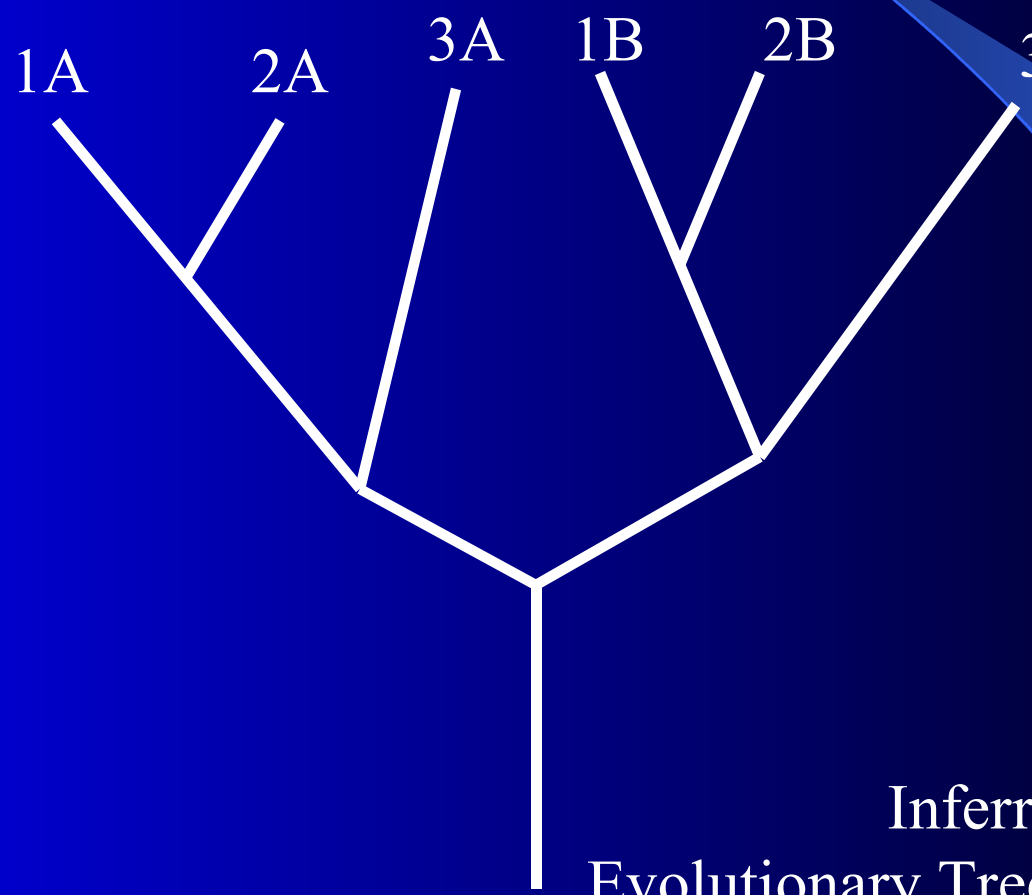
Homologs to 2A found by BLAST (Functions mostly known), from three species (1, 2, and 3)

Phylogenomics: An Example



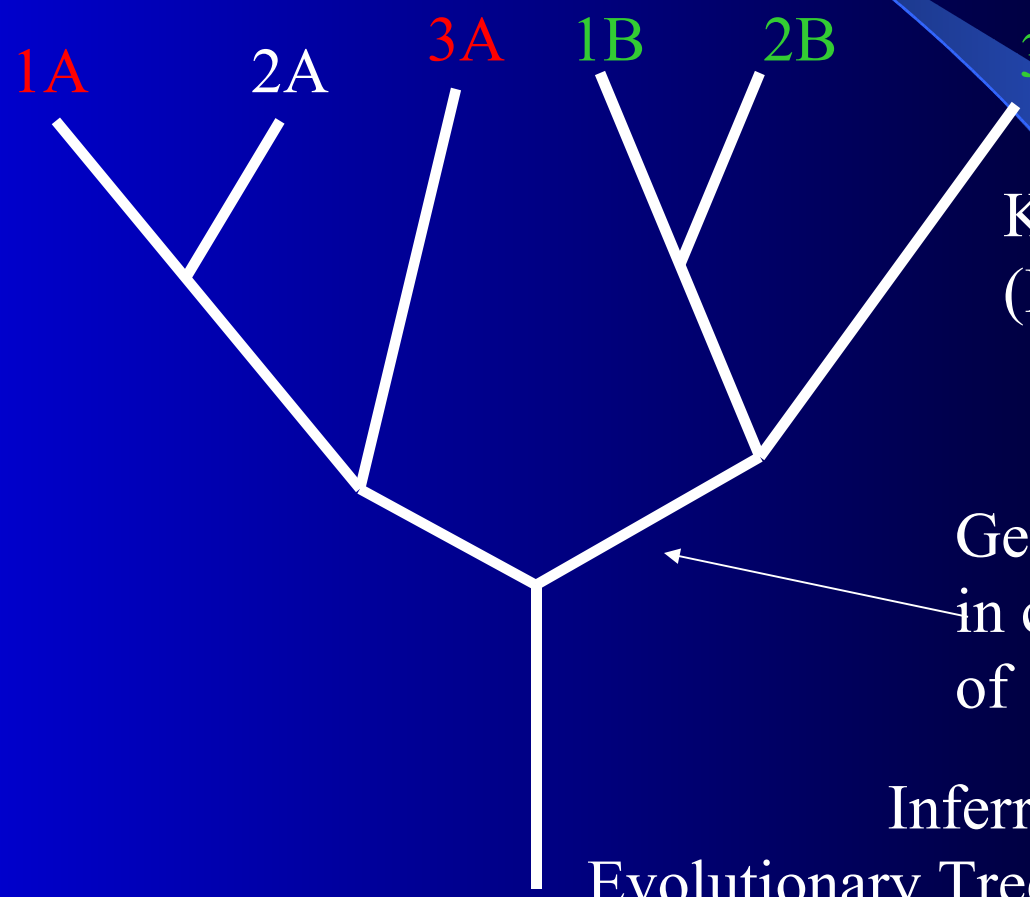
Multiple Sequence Alignment
(Created by CLUSTALW, MUSCLE, etc.)

Phylogenomics: An Example



Inferred
Evolutionary Tree of Sequences

Phylogenomics: An Example

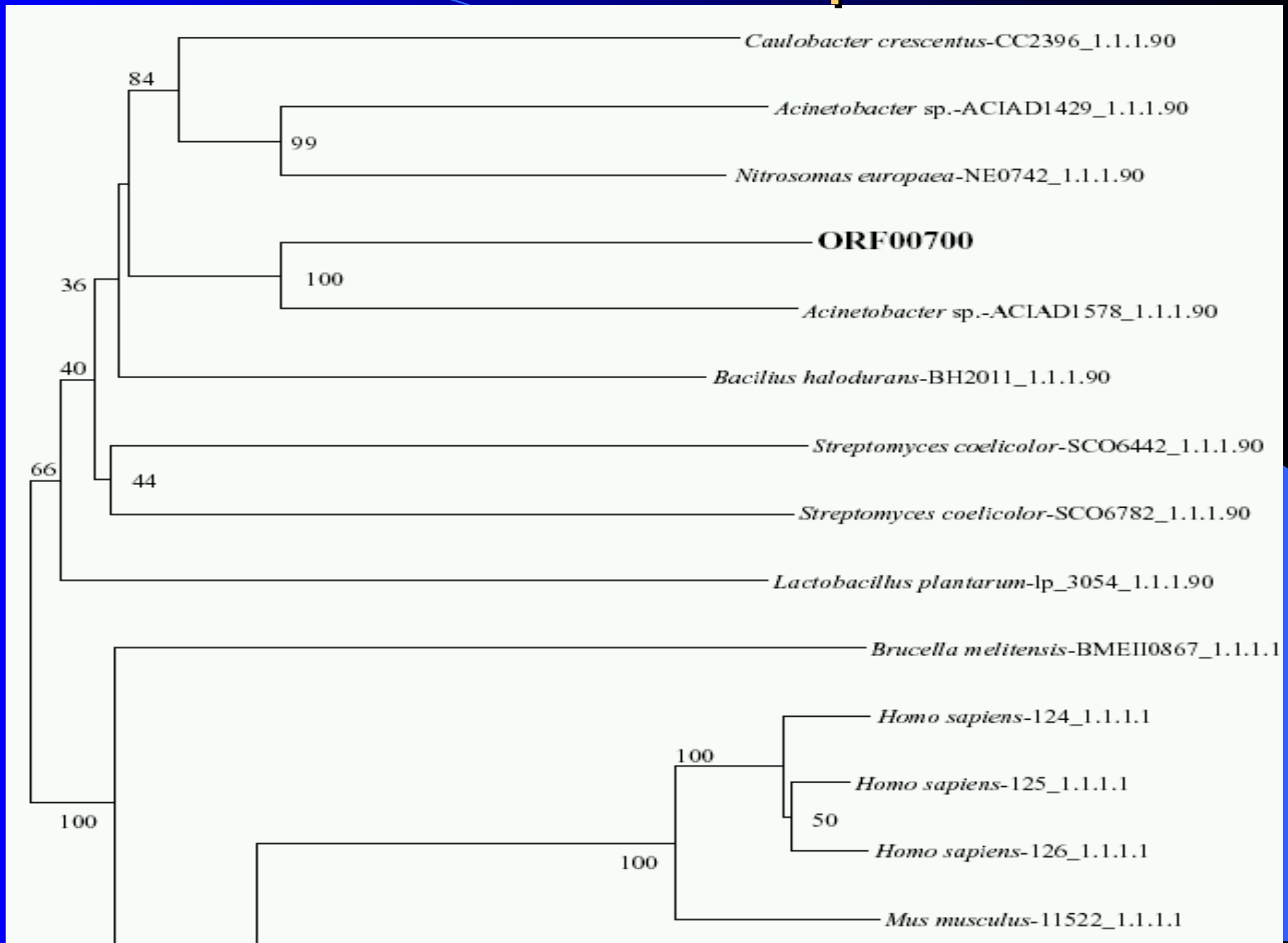


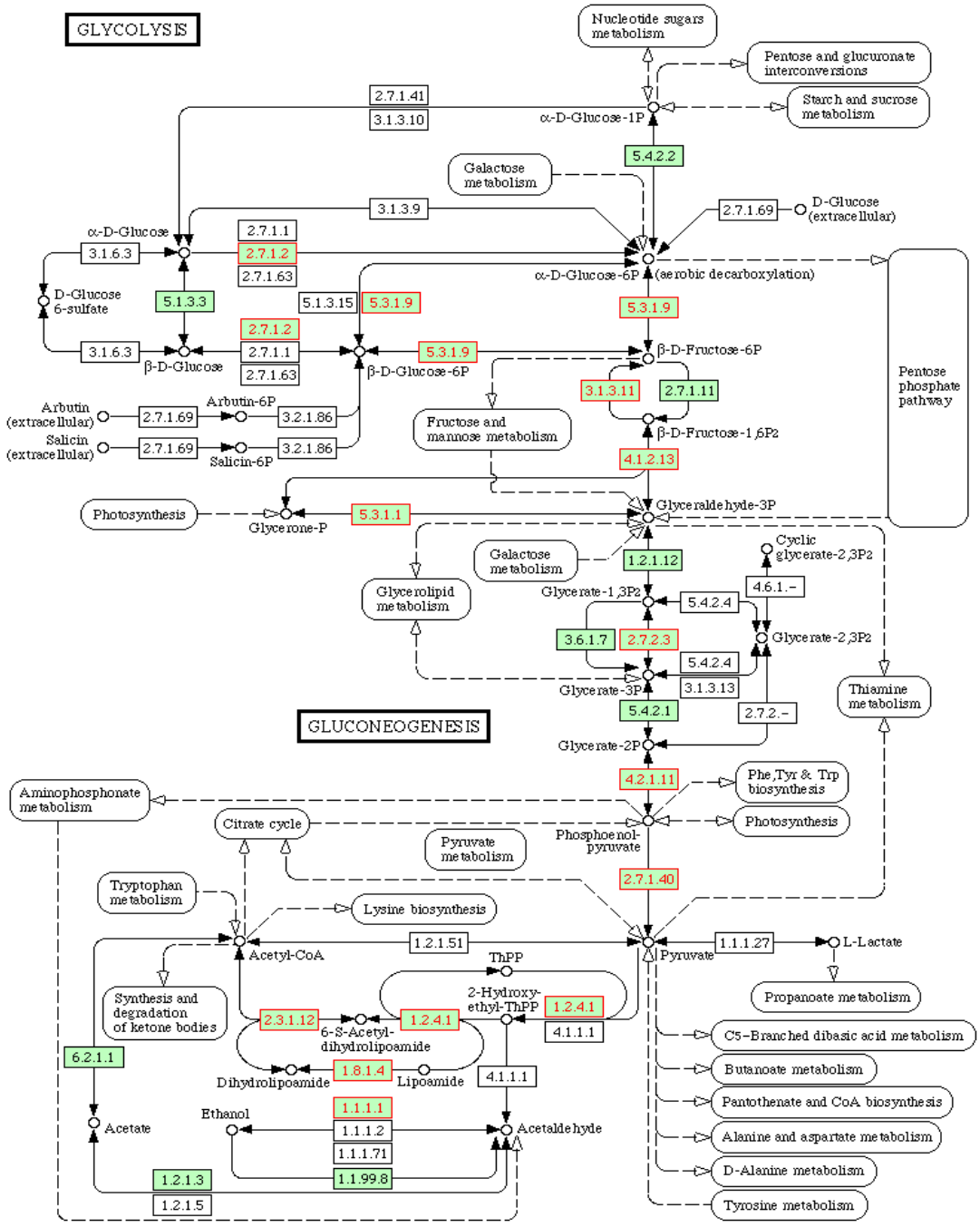
Known Functions
(Red and Green)

Gene Duplication
in common ancestor
of 1, 2, 3

Inferred
Evolutionary Tree of Sequences

A Real Example




GLYCOLYSIS
GLUCONEOGENESIS
Nucleotide sugars metabolism
Pentose and glucuronate interconversions
Starch and sucrose metabolism
Galactose metabolism
α-D-Glucose-1P
α-D-Glucose
D-Glucose (extracellular)
α-D-Glucose-6P
(aerobic decarboxylation)
D-Glucose 6-sulfate
β-D-Glucose
β-D-Glucose-6P
β-D-Fructose-6P
β-D-Fructose-1,6P2
Pentose phosphate pathway
Fructose and mannose metabolism
Arbutin (extracellular)
Arbutin-6P
Salicin (extracellular)
Salicin-6P
Photosynthesis
Glycerone-P
Glyceraldehyde-3P
Cyclic glycerate-2,3P2
Glycerolipid metabolism
Glycerate-1,3P2
Glycerate-2,3P2
Glycerate-3P
Glycerate-2P
Thiamine metabolism
Aminophosphonate metabolism
Citrate cycle
Pyruvate metabolism
Phosphoenol-pyruvate
Phe,Tyr & Trp biosynthesis
Photosynthesis
Tryptophan metabolism
Lysine biosynthesis
Acetyl-CoA
Pyruvate
L-Lactate
ThPP
2-Hydroxy-ethyl-ThPP
Propanoate metabolism
Synthesis and degradation of ketone bodies
6-S-Acetyl-dihydrolipoamide
Dihydrolipoamide
Lipoamide
C5-Branched dibasic acid metabolism
Butanoate metabolism
Pantothenate and CoA biosynthesis
Alanine and aspartate metabolism
D-Alanine metabolism
Tyrosine metabolism
Acetate
Ethanol
Acetaldehyde
2.7.1.41
3.1.3.10
5.4.2.2
3.1.3.9
2.7.1.69
2.7.1.1
2.7.1.2
2.7.1.63
3.1.6.3
5.1.3.3
5.1.3.15
5.3.1.9
3.1.3.11
2.7.1.11
4.1.2.13
3.2.1.86
5.3.1.1
1.2.1.12
4.6.1.-
5.4.2.4
3.6.1.7
2.7.2.3
3.1.3.13
5.4.2.1
2.7.2.-
4.2.1.11
2.7.1.40
1.2.1.51
1.1.1.27
2.3.1.12
1.2.4.1
1.2.4.1
4.1.1.1
6.2.1.1
1.8.1.4
1.1.1.1
1.1.1.2
1.1.1.71
1.1.99.8
1.2.1.3
1.2.1.5