New Tools for Visualizing Genome Evolution

Lutz Hamel
Dept. of Computer Science and Statistics
University of Rhode Island

J. Peter Gogarten
Dept. of Molecular and Cell Biology
University of Connecticut

Motivation

- Early life on Earth has left a variety of traces that can be utilized to reconstruct the history of life
  - the fossil and geological records
  - information retained in living organisms
- Our research focuses on how information can be gained from the molecular record:
  - information about the history of life that is retained in the structure and sequences of macromolecules found in extant organisms
- The analyses of the mosaic nature of genomes using phylogenetics will be a key ingredient to unravel the life's early history.

Relevance to NASA

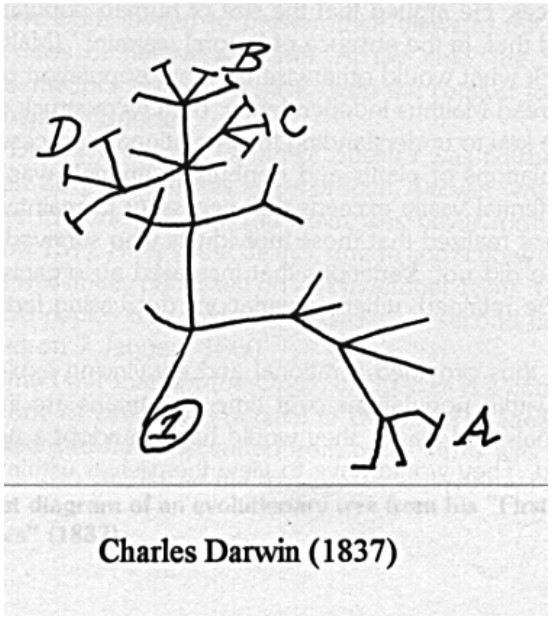
- The analyses are relevant in the context of NASA's Origin theme:
  - Understand the origin and evolution of life on Earth.
- We address questions that are central to NASA's Astrobiology program:
  - Understand how past life on Earth interacted with its changing planetary and Solar System environment.
  - Understand the evolutionary mechanisms and environmental limits of life.

Phylogenetics

Phylogenetics (Greek: *phylon* = race and *genetic* = birth) is the taxonomical classification of organisms based on how closely they are related in terms of evolutionary differences.

Phylogenetics: Classic View

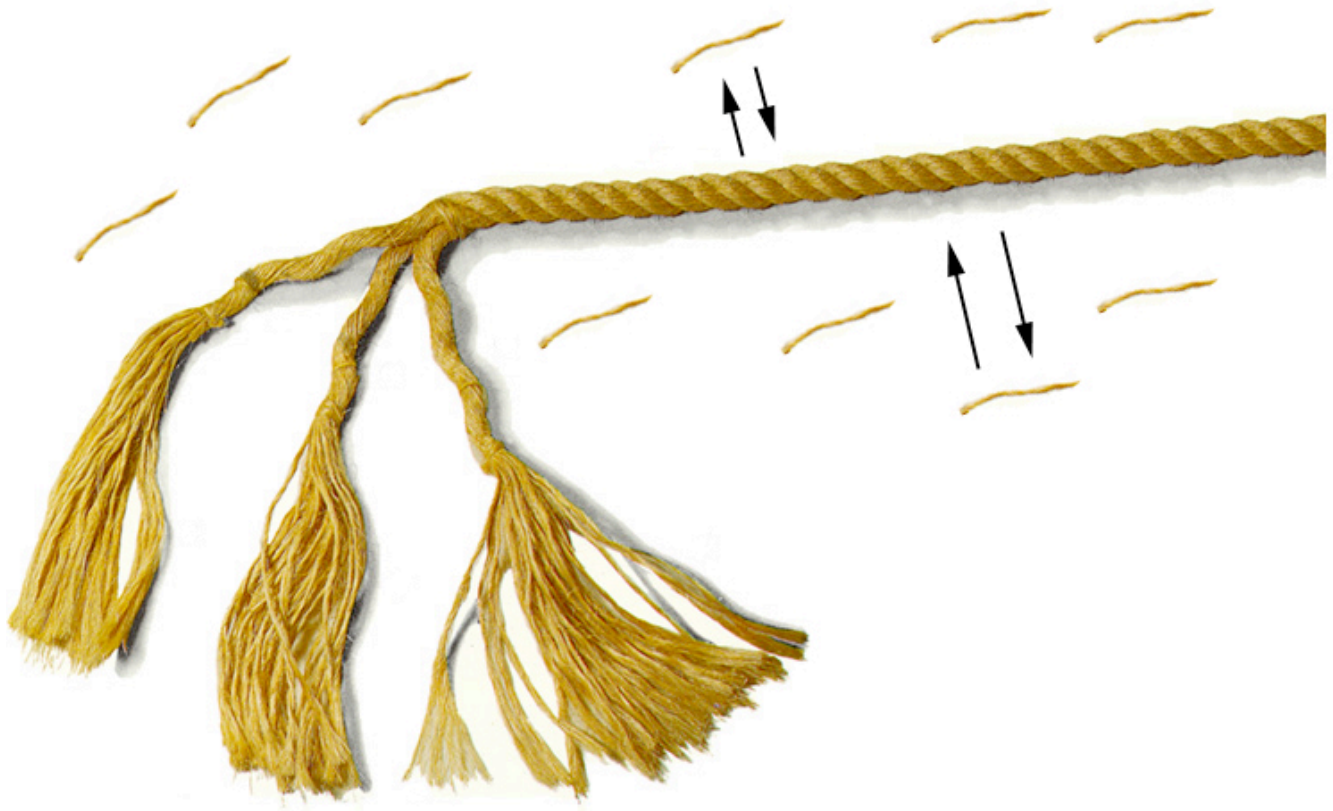
- All genes are inherited from ancestor.
- Branching reflects speciation events.
- Evolutionary tree follows very closely the SSU-rRNA tree.



Charles Darwin (1837)

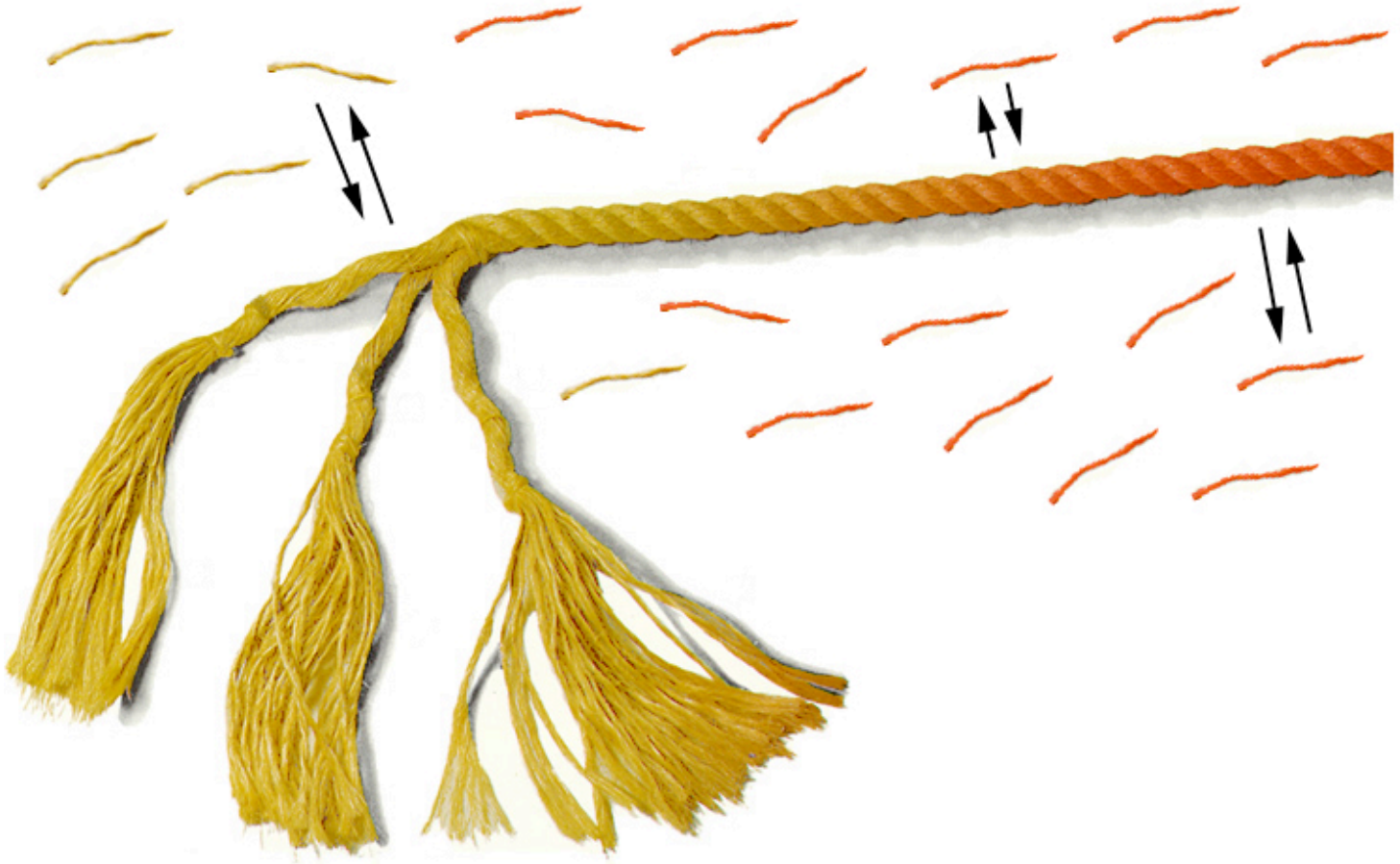
Rope as a metaphor to describe an organismal lineage

Individual fibers = genes that travel for some time in a lineage.



While no individual fiber present at the beginning might be present at the end, the rope (or the organismal lineage) nevertheless has continuity.

However, the genome as a whole will acquire the character of the incoming genes (the rope turns solidly red over time).

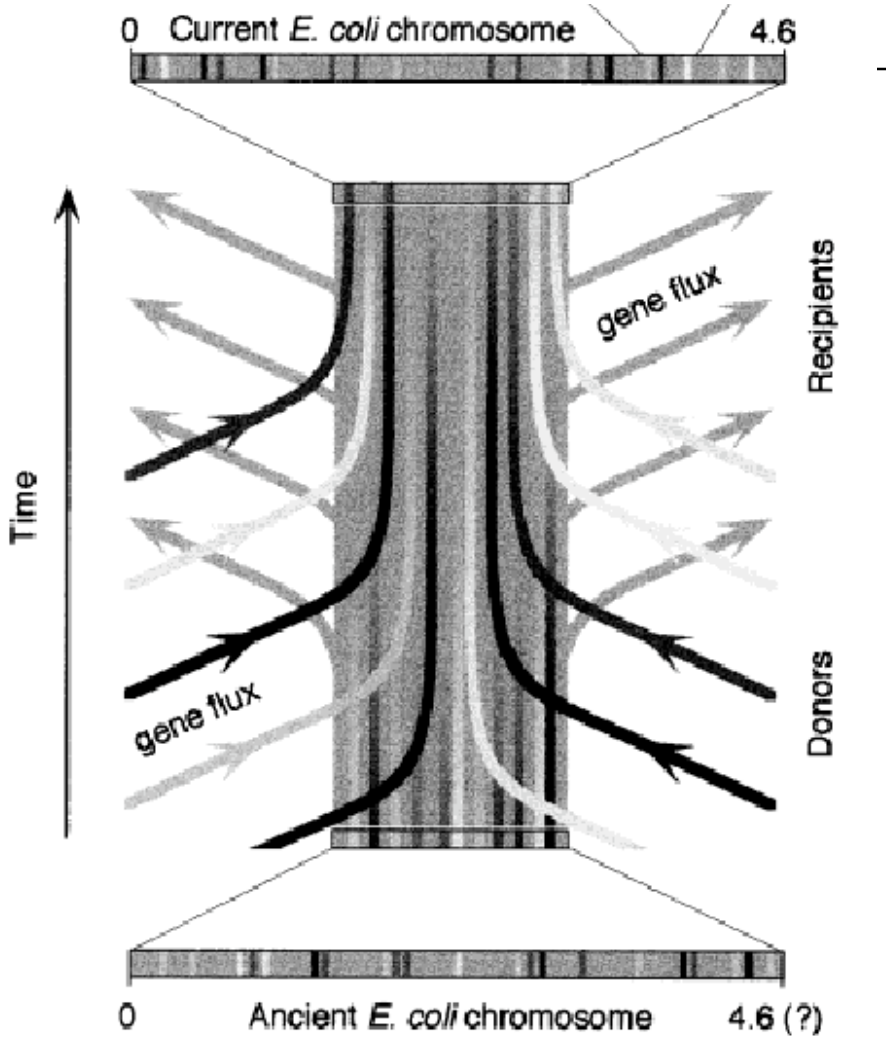


Transferred genes can be detected using:

(a) unusual composition,

(b) the comparison between closely related species, or

(c) *conflicting molecular phylogenies*.



From Bill Martin
BioEssays 21 (2), 99-104.

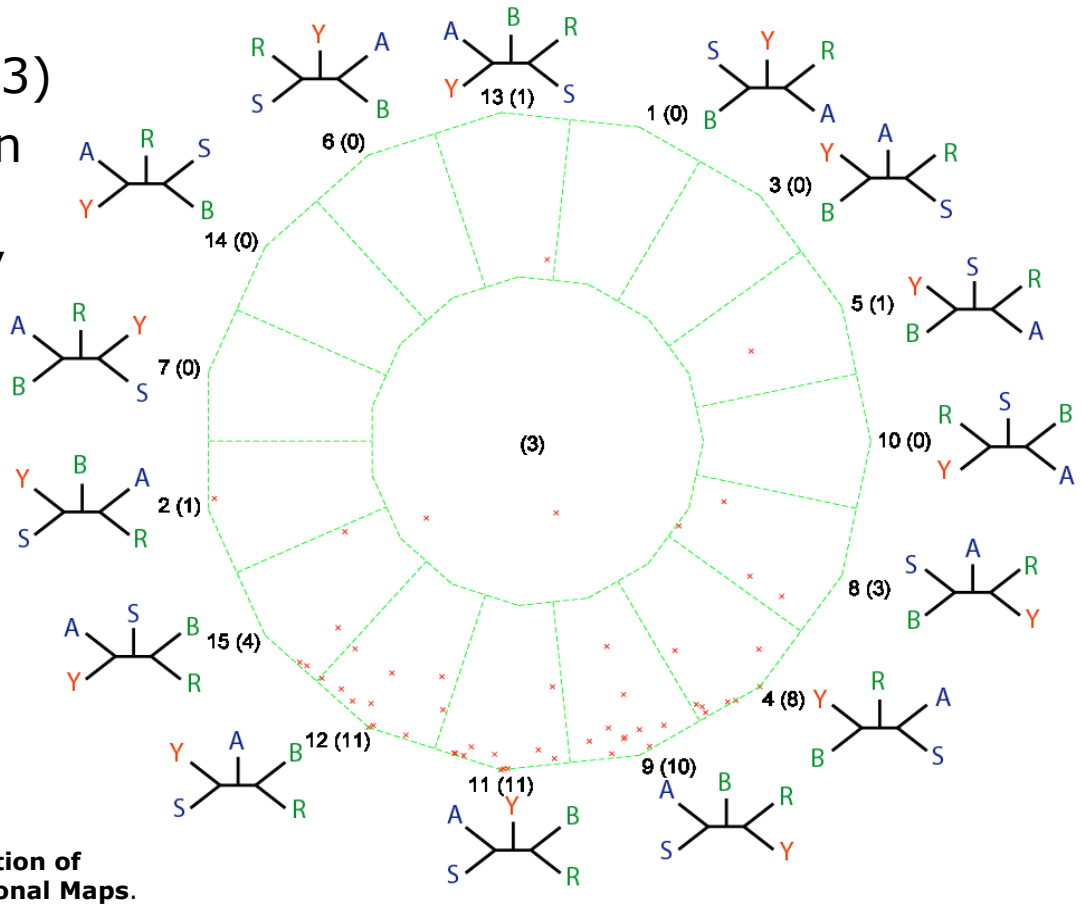
Genome Evolution

- Given the previous discussions
  - A phylogenetic tree based on a few genes is no longer an appropriate model for microbial evolution
  - A better approach is to visualize the evolution of as many genes as possible with the construction of consensus trees when necessary

Visualizing Genome Evolution

- Five genomes (15 unrooted trees)
- Multiple genes (53)
- Genome evolution represented by multiple, possibly conflicting trees.

A: Archaeoglobus
S: Sulfolobus
Y: Yeast
R: Rhodobacter
B: Bacillus



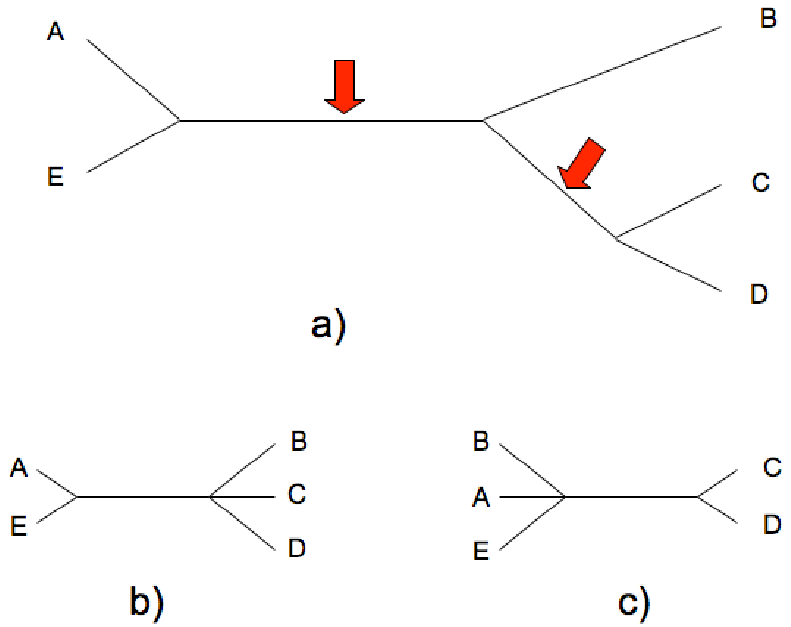
Zhaxybayeva O, Hamel L, Raymond J, Gogarten JP: **Visualization of Phylogenetic Content of Five Genomes with Dekapentagonal Maps**. *Genome Biology* 2004, 5:R20

Bipartitions

- Trees are computationally and representationally inefficient
- Use bipartitions instead

Number of genomes	Number of trees	Number of bipartitions
4	3	3
6	105	25
8	10,395	119
10	2,075,025	501
13	1.37E + 10	4,082
20	2.22E + 20	5.24E + 05
50	2.84E + 74	5.63E + 14

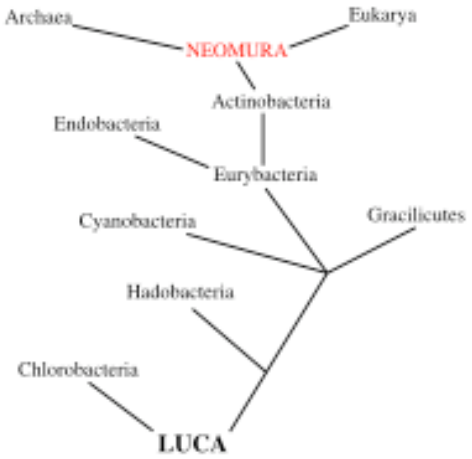
Bipartitions



a) An unrooted tree with 5 genomes.
b) A bipartition due to the left arrow above.
c) A bipartition due to the right arrow above.

Bipartition Space

- Given n genomes we can form 2^n-1 bipartitions



$$(b_1, b_2, b_3, \ldots, b_{2^n-1})$$

- A tree can be represented as a vector of bipartitions - *spectrum*
- A tree is a point in bipartition space, *e.g.*, (0,1,0,…,1)

Visualizing Genome Evolution

- Given *n* genomes, select *k orthologous genes* (genes that appear in all *n* genomes)
- Construct spectra for the *k* evolutionary trees
- Visualize the spectra with *self-organizing maps*
- Straight-forward algorithms exist to detect *bipartition conflicts* (incompatible evolutionary trees)
- *Consensus trees* are easily computed from a collection of tree spectra

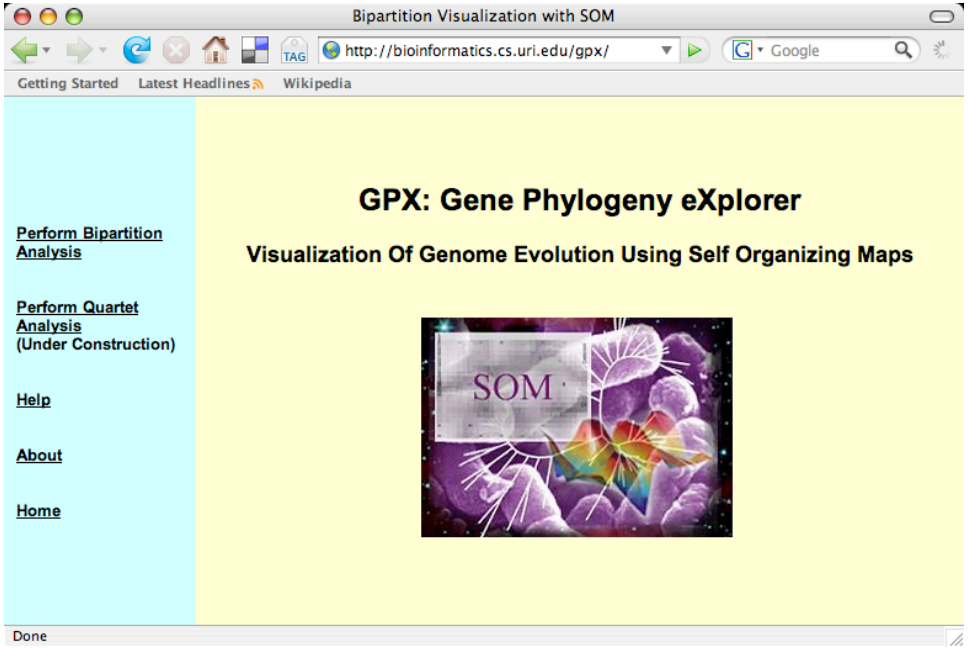
Unsupervised Learning in Detection of Gene Transfer, Lutz Hamel, Neha Nahar, Maria S. Poptsova, Olga Zhaxybayeva, and J. Peter Gogarten. Journal of Biomedicine and Biotechnology, J Biomed Biotechnol. 2008; 2008: 472719. Published online 2008 April 1. doi: 10.1155/2008/472719.

Unsupervised Learning in Spectral Genome Analysis, Lutz Hamel, Neha Nahar, Maria S. Poptsova, Olga Zhaxybayeva, and J. Peter Gogarten. Proceeding of the IEEE Conference Frontiers in the Convergence of Bioscience and Information Technologies (FBIT 2007), October 2007, pp317 - 321, IEEE Press, ISBN 0-7695-2999-2.

GPX: A Tool for the Exploration and Visualization of Genome Evolution, Neha Nahar, Maria S. Poptsova, Lutz Hamel, and J. Peter Gogarten. Proceedings of the IEEE 7th International Symposium on Bioinformatics & Bioengineering (BIBE07), Oct 14th-17th 2007, Boston, pp1338 - 1342, IEEE Press, ISBN 1-4244-1509-8.

GPX

- Genome Phylogeny eXplorer
- Highly interactive tool
- Web 2.0 application that supports bipartition based spectral analysis using self-organizing maps.

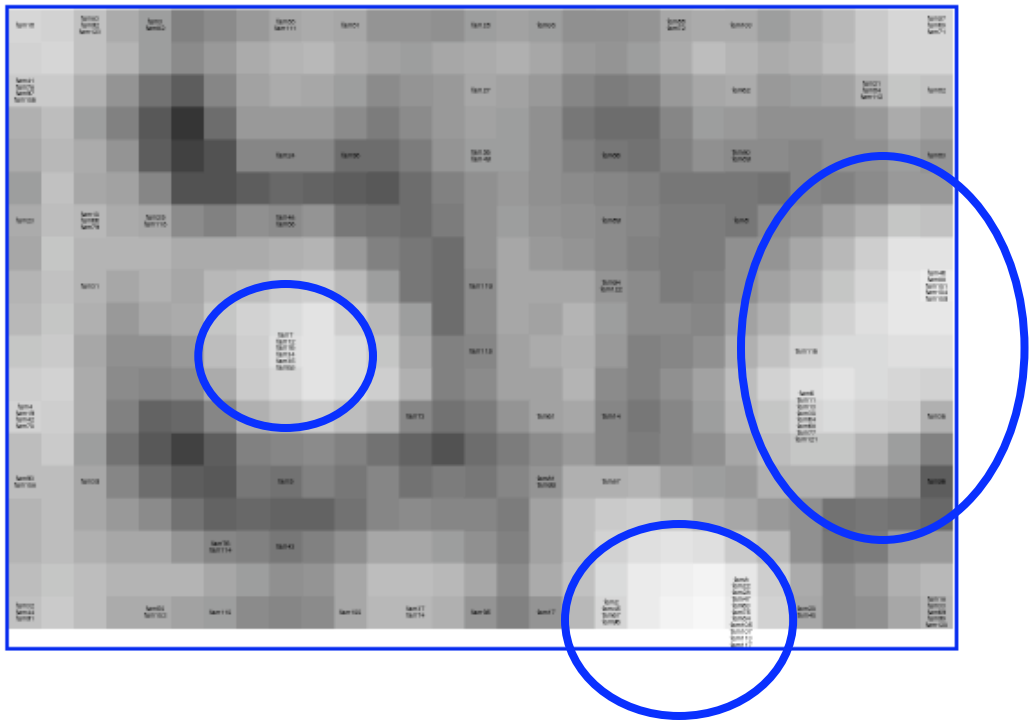


http://bioinformatics.cs.uri.edu/gpx/

Gene Maps

Upload
Map
Clusters
Bipartitions

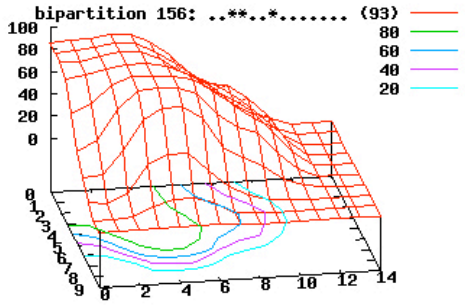
Home



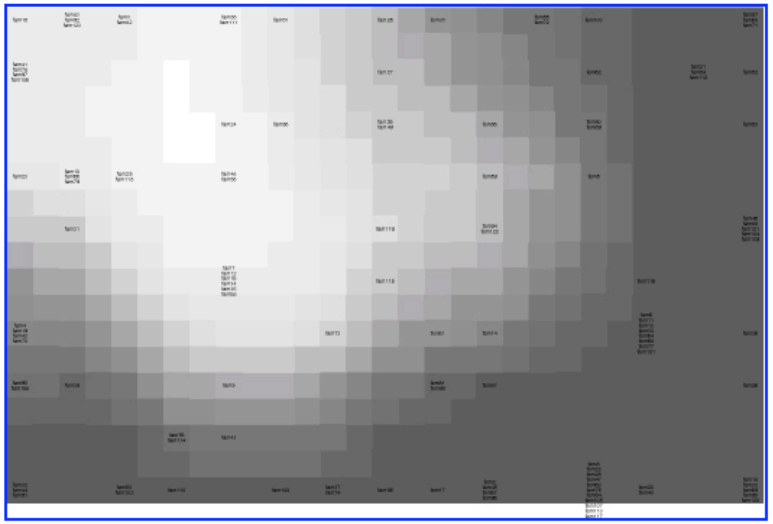
Strongly Supported Bipartitions

Upload
Map
Clusters
Bipartitions
Home

bipartition 156: ..**..*..... (93)



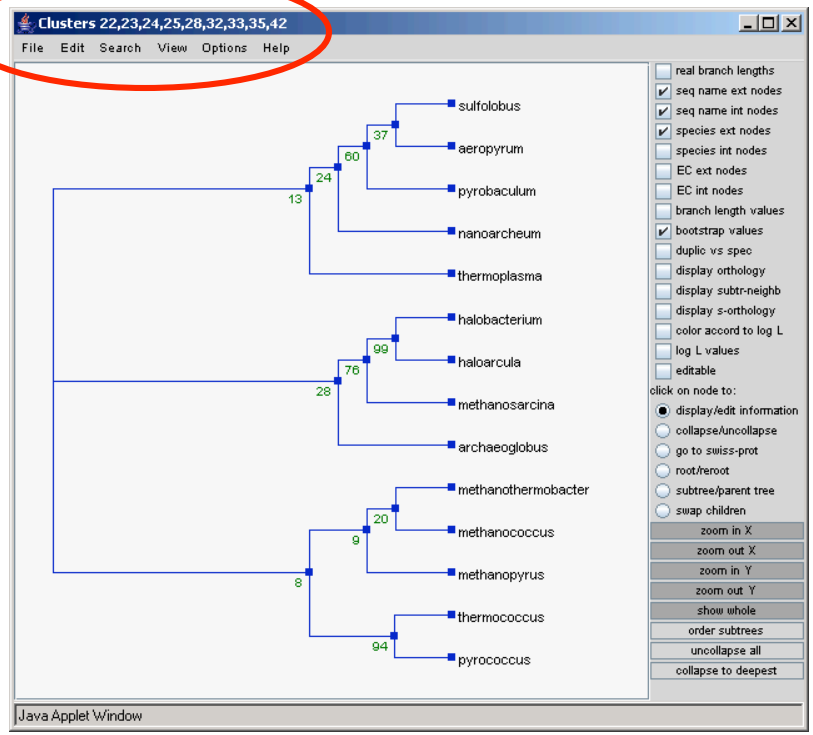
Halobacterium, Haloarcula and *Methanosarcina* groups together



Consensus Trees

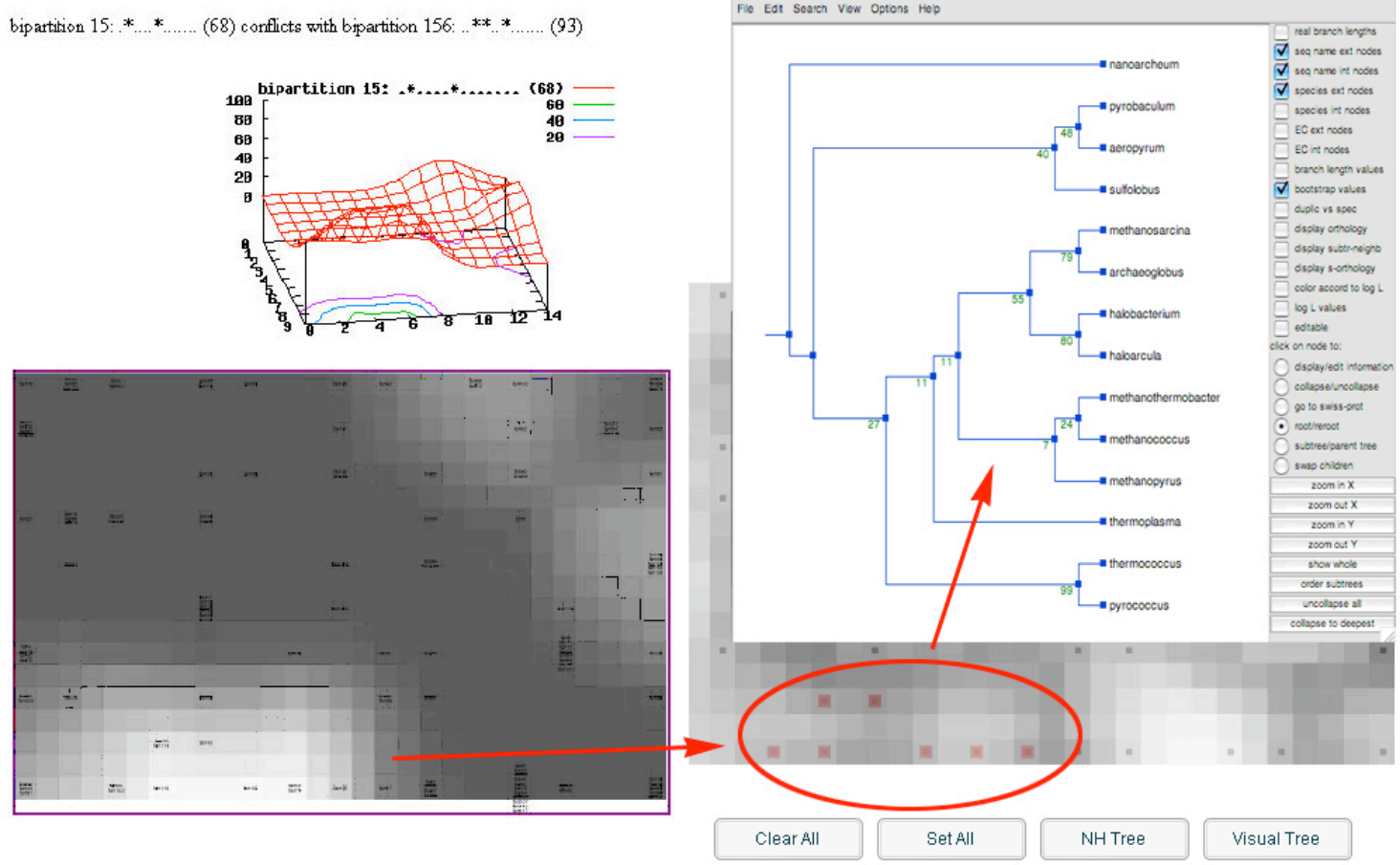
Select neurons 23-25, 28, 32-34 and 42 and visualize tree

ATV tree viewer displays plurality consensus for selected clusters



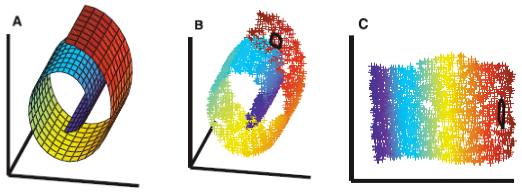
Conflicting Bipartitions

Bipartition 15 corresponds to a split where *Archaeoglobus* groups together with *Methanosarcina*. This is a bipartition that is in conflict with the consensus phylogeny of conserved genes.

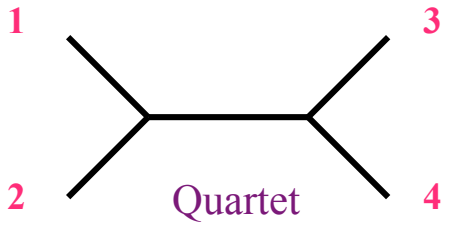


Future Work

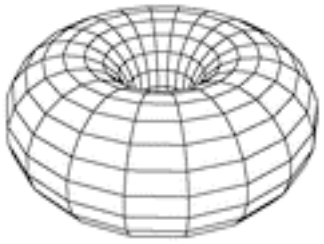
- Explore Locally Linear Embedding (LLE) as opposed to SOM.
- Explore quartets as opposed to bipartitions.
- Use boundless maps to avoid border effects.



LLE



Quartet



Toroidal map

Acknowledgements

- Maria Poptsova, Dept. of Molecular and Cell Biology, University of Connecticut
- Neha Nahar, Dept. of Computer Science and Statistics, University of Rhode Island
- Olga A. Zhaxybayeva, Department of Biochemistry and Molecular Biology, Dalhousie University

…and of course the NASA Applied Information System Research Program (NNG04GP90G).

Thank You!