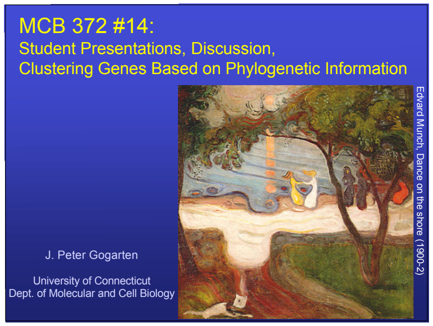


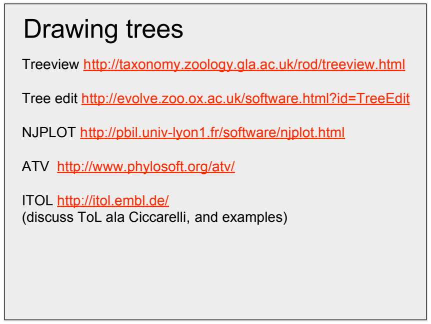
MCB 372 #14:

Student Presentations, Discussion, Clustering Genes Based on Phylogenetic Information

Edvard Munch, Dance on the shore (1900-2)

J. Peter Gogarten

University of Connecticut
Dept. of Molecular and Cell Biology



Drawing trees

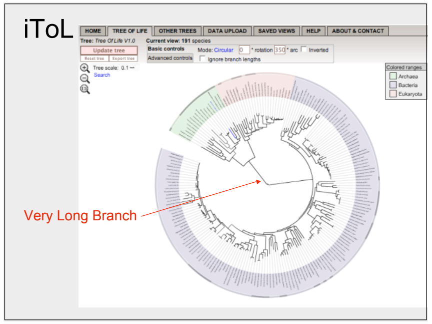
Treeview http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

Tree edit http://evolve.zoo.ox.ac.uk/software.html?id=TreeEdit

NJPLOT http://pbil.univ-lyon1.fr/software/njplot.html

ATV http://www.phylosoft.org/atv/

ITOL http://itol.embl.de/
(discuss ToL ala Ciccarelli, and examples)



iToL

Tree: *Tree Of Life V1.0*

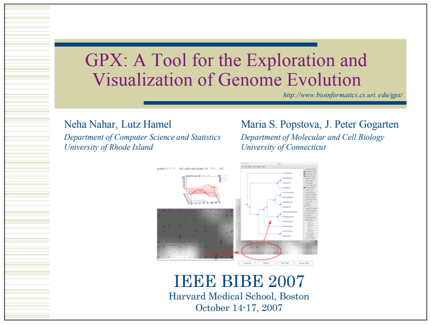
Current view: 191 species

Mode: Circular | 0 ° rotation | 350 ° arc | □ Inverted

□ Ignore branch lengths

Tree scale: 0.1

Very Long Branch



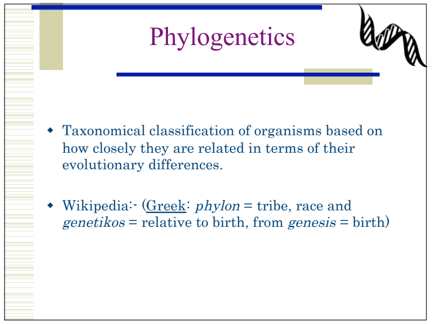
GPX: A Tool for the Exploration and Visualization of Genome Evolution

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

Neha Nahar, Lutz Hamel
Department of Computer Science and Statistics
University of Rhode Island

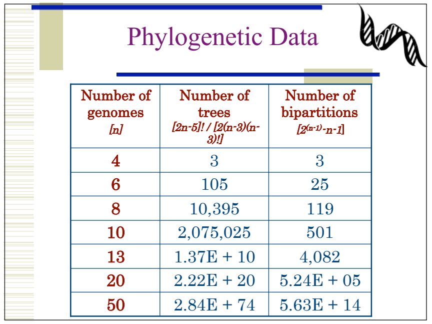
Maria S. Popstova, J. Peter Gogarten
Department of Molecular and Cell Biology
University of Connecticut

IEEE BIBE 2007
Harvard Medical School, Boston
October 14-17, 2007



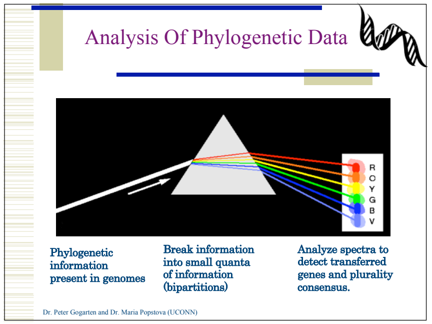
Phylogenetics

- Taxonomical classification of organisms based on how closely they are related in terms of their evolutionary differences.
- Wikipedia:- (Greek: *phylon* = tribe, race and *genetikos* = relative to birth, from *genesis* = birth)



Phylogenetic Data

Number of genomes *[n]*	Number of trees *[2n-5]! / [2(n-3)(n-3)!]*	Number of bipartitions *[2(n-1)-n-1]*
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



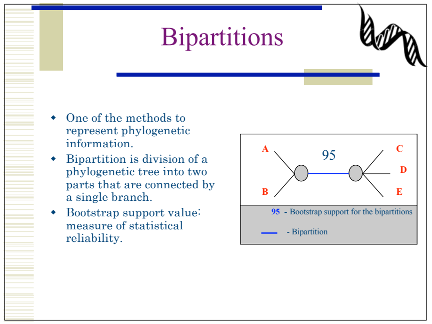
Analysis Of Phylogenetic Data

Phylogenetic information present in genomes

Break information into small quanta of information (bipartitions)

Analyze spectra to detect transferred genes and plurality consensus.

Dr. Peter Gogarten and Dr. Maria Popstova (UCONN)

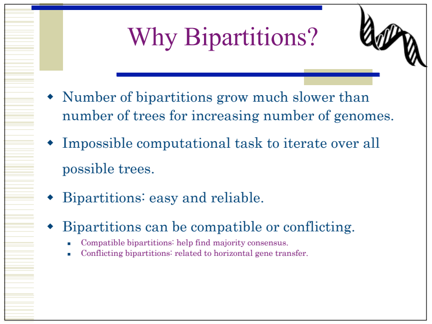


Bipartitions

- One of the methods to represent phylogenetic information.
- Bipartition is division of a phylogenetic tree into two parts that are connected by a single branch.
- Bootstrap support value: measure of statistical reliability.

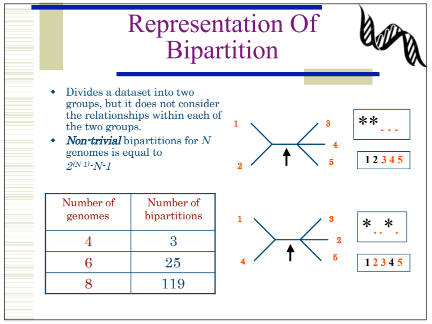
95 - Bootstrap support for the bipartitions

— - Bipartition



Why Bipartitions?

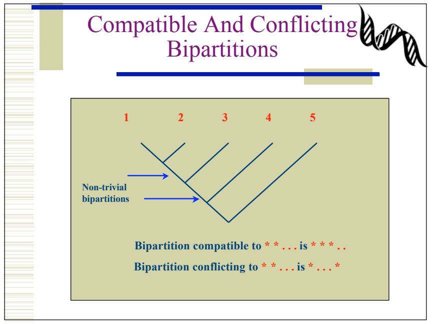
- Number of bipartitions grow much slower than number of trees for increasing number of genomes.
- Impossible computational task to iterate over all possible trees.
- Bipartitions: easy and reliable.
- Bipartitions can be compatible or conflicting.
  - Compatible bipartitions: help find majority consensus.
  - Conflicting bipartitions: related to horizontal gene transfer.



Representation Of Bipartition

- Divides a dataset into two groups, but it does not consider the relationships within each of the two groups.
- ***Non-trivial*** bipartitions for *N* genomes is equal to *$2^{(N-1)}$-N-1*

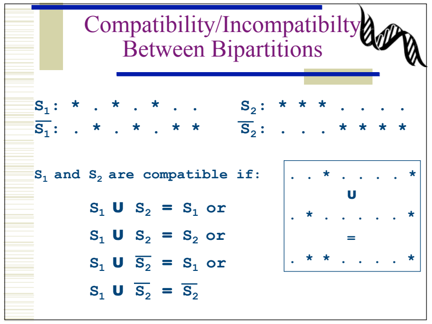
Number of genomes	Number of bipartitions
4	3
6	25
8	119



Compatible And Conflicting Bipartitions

Bipartition compatible to * * . . . is * * * . .

Bipartition conflicting to * * . . . is * . . . *



Compatibility/Incompatibilty Between Bipartitions

S_1: * . * . * . . S_2: * * * . . . .

$\overline{S_1}$: . * . * . * * $\overline{S_2}$: . . . * * * *

S_1 and S_2 are compatible if:

S_1 **U** S_2 = S_1 or

S_1 **U** S_2 = S_2 or

S_1 **U** $\overline{S_2}$ = S_1 or

S_1 **U** $\overline{S_2}$ = $\overline{S_2}$

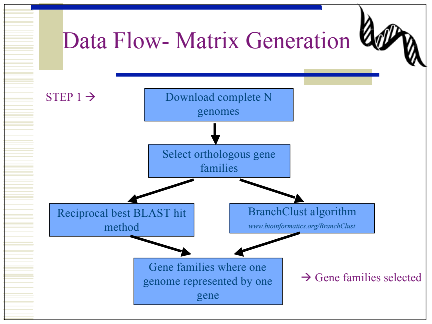
. . * . . . . *

U

. * . . . . . *

=

. * * . . . . *



Data Flow- Matrix Generation

STEP 1 →

Download complete N genomes

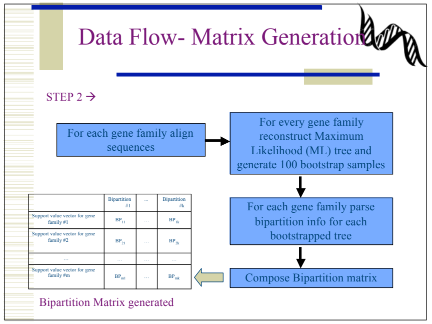
Select orthologous gene families

Reciprocal best BLAST hit method

BranchClust algorithm *www.bioinformatics.org/BranchClust*

Gene families where one genome represented by one gene

→ Gene families selected



Data Flow- Matrix Generation

STEP 2 →

For each gene family align sequences

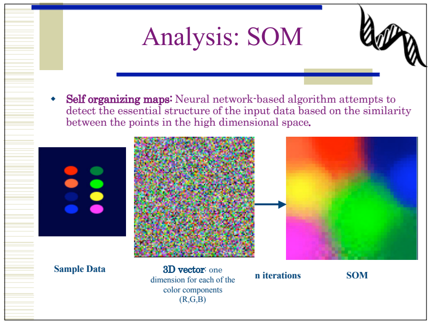
For every gene family reconstruct Maximum Likelihood (ML) tree and generate 100 bootstrap samples

For each gene family parse bipartition info for each bootstrapped tree

Compose Bipartition matrix

	Bipartition #1	...	Bipartition #k
Support value vector for gene family #1	BP_{11}	...	BP_{1k}
Support value vector for gene family #2	BP_{21}	...	BP_{2k}
...	...	...	...
Support value vector for gene family #m	BP_{m1}	...	BP_{mk}

Bipartition Matrix generated



Analysis: SOM

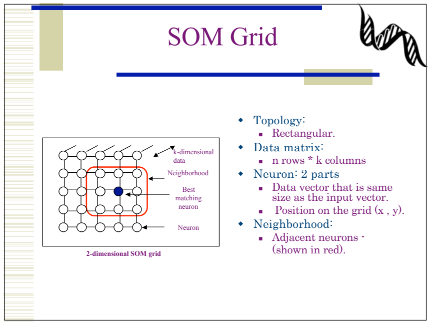
- **Self organizing maps:** Neural network-based algorithm attempts to detect the essential structure of the input data based on the similarity between the points in the high dimensional space.

Sample Data

3D vector: one dimension for each of the color components (R,G,B)

n iterations

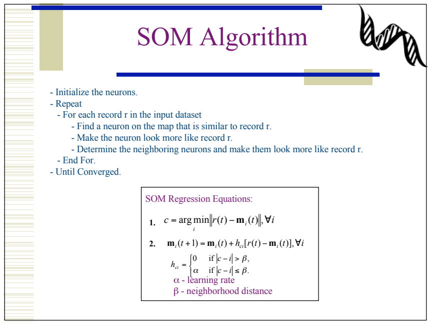
SOM



SOM Grid

2-dimensional SOM grid

- Topology:
  - Rectangular.
- Data matrix:
  - n rows * k columns
- Neuron: 2 parts
  - Data vector that is same size as the input vector.
  - Position on the grid (x , y).
- Neighborhood:
  - Adjacent neurons - (shown in red).



SOM Algorithm

```
- Initialize the neurons.
- Repeat
  - For each record r in the input dataset
      - Find a neuron on the map that is similar to record r.
      - Make the neuron look more like record r.
      - Determine the neighboring neurons and make them look more like record r.
  - End For.
- Until Converged.
```

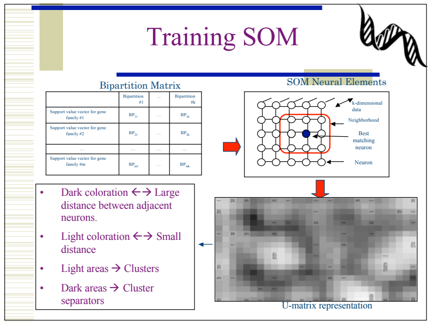
SOM Regression Equations:

1. $c = \arg\min_i \|r(t) - \mathbf{m}_i(t)\|, \forall i$
2. $\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}[r(t) - \mathbf{m}_i(t)], \forall i$

$$h_{ci} = \begin{cases} 0 & \text{if } |c-i| > \beta, \\ \alpha & \text{if } |c-i| \le \beta. \end{cases}$$

α - learning rate

β - neighborhood distance



Training SOM

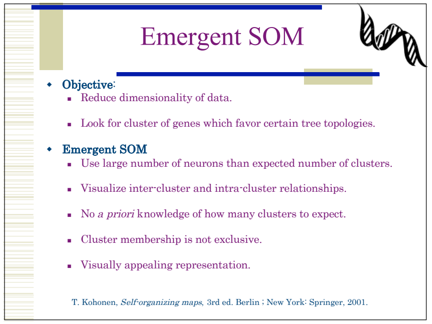
Bipartition Matrix

	Bipartition #1	...	Bipartition #k
Support value vector for gene family #1	BP_{11}	...	BP_{1k}
Support value vector for gene family #2	BP_{21}	...	BP_{2k}
...	...	...	...
Support value vector for gene family #m	BP_{m1}	...	BP_{mk}

SOM Neural Elements

- Dark coloration ←→ Large distance between adjacent neurons.
- Light coloration ←→ Small distance
- Light areas → Clusters
- Dark areas → Cluster separators

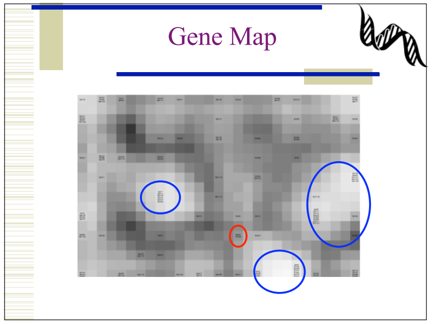
U-matrix representation



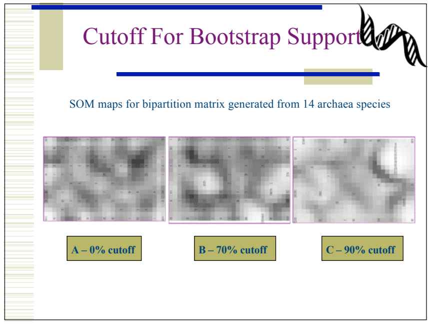
Emergent SOM

- **Objective**:
  - Reduce dimensionality of data.
  - Look for cluster of genes which favor certain tree topologies.
- **Emergent SOM**
  - Use large number of neurons than expected number of clusters.
  - Visualize inter-cluster and intra-cluster relationships.
  - No *a priori* knowledge of how many clusters to expect.
  - Cluster membership is not exclusive.
  - Visually appealing representation.

T. Kohonen, *Self-organizing maps*, 3rd ed. Berlin ; New York: Springer, 2001.



Gene Map



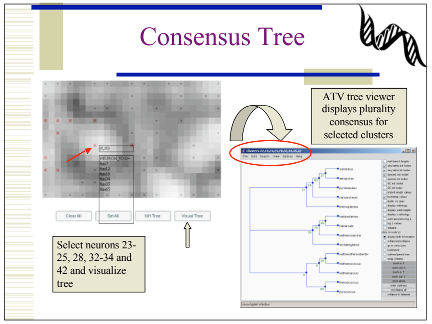
Cutoff For Bootstrap Support

SOM maps for bipartition matrix generated from 14 archaea species

A – 0% cutoff

B – 70% cutoff

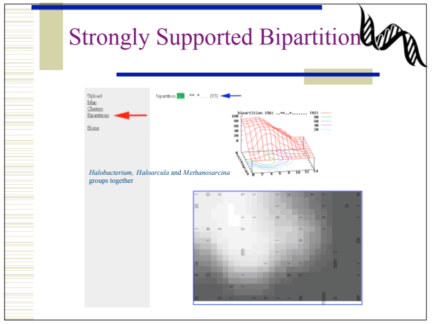
C – 90% cutoff



Consensus Tree

ATV tree viewer displays plurality consensus for selected clusters

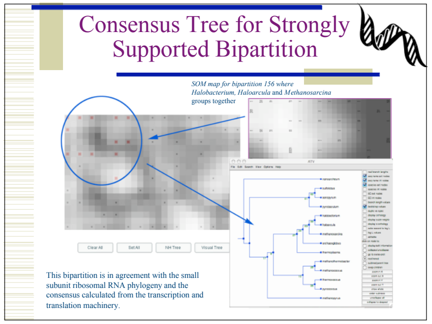
Select neurons 23-25, 28, 32-34 and 42 and visualize tree



Strongly Supported Bipartition

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

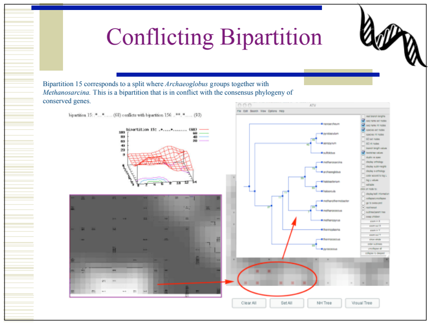
Halobacterium, Haloarcula and *Methanosarcina* groups together



Consensus Tree for Strongly Supported Bipartition

SOM map for bipartition 156 where Halobacterium, Haloarcula and *Methanosarcina* groups together

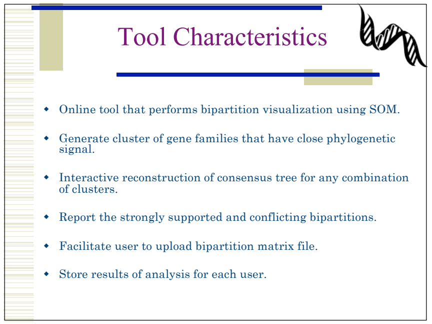
This bipartition is in agreement with the small subunit ribosomal RNA phylogeny and the consensus calculated from the transcription and translation machinery.



Conflicting Bipartition

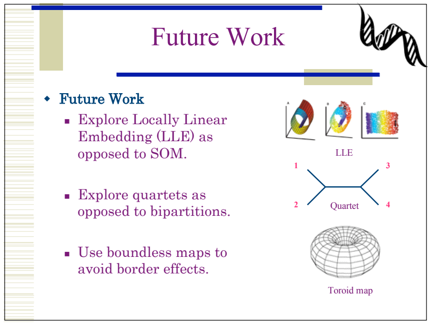
Bipartition 15 corresponds to a split where *Archaeoglobus* groups together with *Methanosarcina.* This is a bipartition that is in conflict with the consensus phylogeny of conserved genes.

bipartition 15: .*....*...... (68) conflicts with bipartition 156: ..**..*...... (93)



Tool Characteristics

- Online tool that performs bipartition visualization using SOM.
- Generate cluster of gene families that have close phylogenetic signal.
- Interactive reconstruction of consensus tree for any combination of clusters.
- Report the strongly supported and conflicting bipartitions.
- Facilitate user to upload bipartition matrix file.
- Store results of analysis for each user.



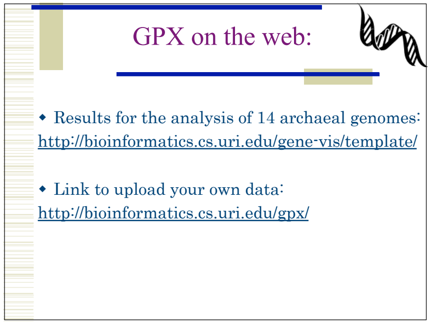
Future Work

- **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

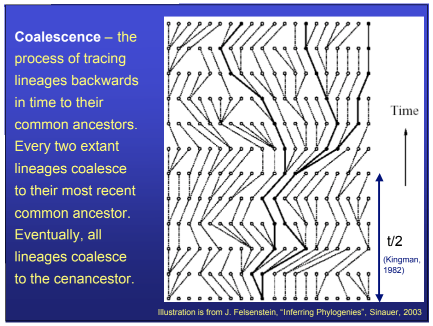
Quartet

Toroid map



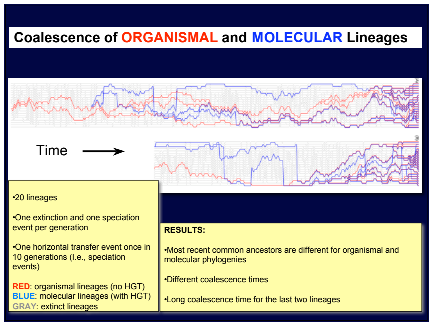
GPX on the web:

- Results for the analysis of 14 archaeal genomes: http://bioinformatics.cs.uri.edu/gene-vis/template/
- Link to upload your own data: http://bioinformatics.cs.uri.edu/gpx/



Coalescence – the process of tracing lineages backwards in time to their common ancestors. Every two extant lineages coalesce to their most recent common ancestor. Eventually, all lineages coalesce to the cenancestor.

Illustration is from J. Felsenstein, "Inferring Phylogenies", Sinauer, 2003



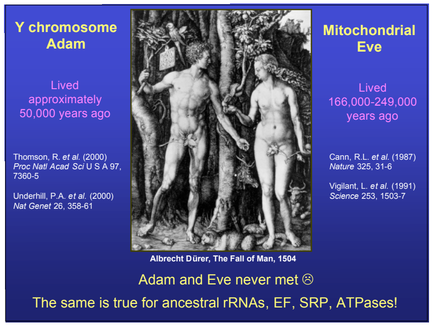
Coalescence of ORGANISMAL and MOLECULAR Lineages

- •20 lineages
- •One extinction and one speciation event per generation
- •One horizontal transfer event once in 10 generations (I.e., speciation events)

RED: organismal lineages (no HGT)
BLUE: molecular lineages (with HGT)
GRAY: extinct lineages

RESULTS:

- •Most recent common ancestors are different for organismal and molecular phylogenies
- •Different coalescence times
- •Long coalescence time for the last two lineages



Y chromosome Adam

Lived approximately 50,000 years ago

Thomson, R. *et al.* (2000) *Proc Natl Acad Sci* U S A 97, 7360-5

Underhill, P.A. *et al.* (2000) *Nat Genet* 26, 358-61

Mitochondrial Eve

Lived 166,000-249,000 years ago

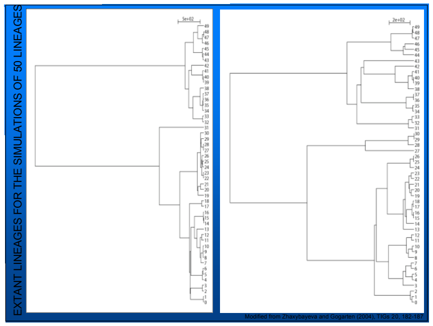
Cann, R.L. *et al.* (1987) *Nature* 325, 31-6

Vigilant, L. *et al.* (1991) *Science* 253, 1503-7

Albrecht Dürer, The Fall of Man, 1504

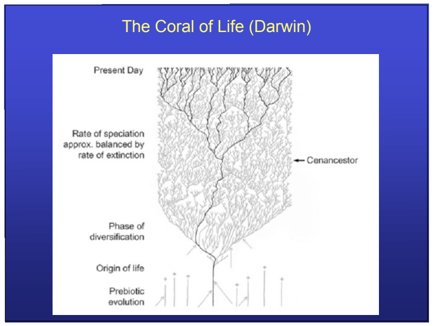
Adam and Eve never met ☹

The same is true for ancestral rRNAs, EF, SRP, ATPases!

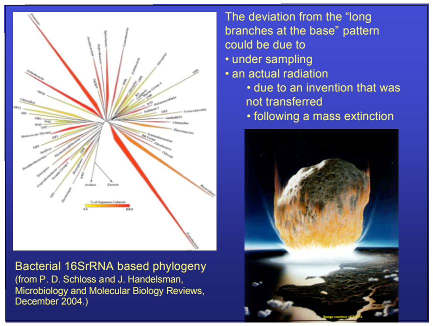


EXTANT LINEAGES FOR THE SIMULATIONS OF 50 LINEAGES

Modified from Zhaxybayeva and Gogarten (2004), TIGs 20, 182-187



The Coral of Life (Darwin)



The deviation from the "long branches at the base" pattern could be due to

- under sampling
- an actual radiation
  - due to an invention that was not transferred
  - following a mass extinction

Bacterial 16SrRNA based phylogeny (from P. D. Schloss and J. Handelsman, Microbiology and Molecular Biology Reviews, December 2004.)