MCB 372 #12:
Tree, Quartets and Supermatrix Approaches

Edvard Munch, The Dance of Life (1900)

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

Collaborators:
Olga Zhaxybayeva (Dalhousie)
Jinling Huang (ECU)
Tim Harlow (UConn)
Pascal Lapierre (UConn)
Greg Fournier (UConn)

Funded through the NASA Exobiology and AISR Programs, and NSF Microbial Genetics

In the Felsenstein Zone

0.8
0.1
0.8
0.1
0.1

A
B
C
D

long branches attract

true tree

inferred tree

ML reconstructions with alignment step

long branch attraction artifact

the two longest branches join together

seq. from A
seq. from B
seq. from C
seq. from D

100% bootstrap support for bipartition (AD)(CB)
What could you do to investigate if this is a possible explanation?
use only slow positions,
use an algorithm that corrects for ASRV

seq. from B
seq. from A
seq. from C
seq. from D

Consensus of all trees from all bootstrap samples

Consensus of all collapsed (<95%) consensus trees

METAGENOME
Genomic Islands

Gene frequency in a typical genome
- Pick a random gene from any of the 293 genomes
- Search in how many genomes this gene is present
- Sampling of 15,000 genes

\[ F(x) = \sum (A_n \cdot \exp(K_n \cdot x)) \]

Character genes
Accessory pool
Extended core

Kézdy-Swinbourne Plot

Gene frequency in individual genomes
Character Genes
Accessory Pool
Extended Core

Approximate number of genes sampled in 200 bacterial genomes:
- 25,160 core genes
- 453,781 extended core genes
- 156,259 accessory genes

The Phylogenetic Position of Thermotoga

(a) concordant genes,
(b) all genes & according to 16S
(c) according to phylogenetically discordant genes


Supertree vs. Supermatrix

Schematic of MRP supertree (left) and parsimony supermatrix (right) approaches to the analysis of three data sets. Clade C+D is supported by all three separate data sets, but not by the supermatrix. Synapomorphies for clade C+D are highlighted in pink. Clade A+B+C is not supported by separate analyses of the three data sets, but is supported by the supermatrix. Synapomorphies for clade A+B+C are highlighted in blue. E is the outgroup used to root the tree.