Phylogenetics and the Trees of Life

Microbial Ecology
BSCI 464/MEES 698

Overview
- Molecular evolution
- Tree topology
- Building trees
- Genome trees
- Computer software

Taxonomy or Systematic
- Taxonomy is the art of classifying things into groups — a quintessential human behavior — established as a mainstream scientific field by Carolus Linnaeus (1707-1778)
- Provides the essential framework without which we could not recognize or study biological diversity & evolution
Darwin: the first workable hypothesis about evolution

“The natural system [is] based on descent with modification … the characters that naturalists consider as showing true affinity are those which have been inherited from a common parent, and in so far as all true classification is genealogical; that community of descent is the common bond that naturalists have been seeking.”

- Charles Darwin, Origin of Species, 1859

Molecular Phylogenetics

• study of evolutionary relationships among organisms by using molecular data such as DNA and protein sequences
• established by Carl Woese in the 1970s

A phylogenetic classification is based upon evolutionary relationships i.e., upon common ancestry, in contrast to a phenetic classification (phenetics, numerical taxonomy) which is based upon overall similarity

What Sequences to Study?

• DNA sequences have many advantages
  – character states can be scored unambiguously
  – large numbers of characters can be scored for each individual
• Proteins are constrained by natural selection - better for very distant relationships
• Some sequences are highly variable (rRNA spacer regions, immunoglobulin genes), while others are highly conserved (actin, rRNA coding regions)
  – different regions within a single gene can evolve at different rates (conserved vs. variable domains)
Definitions

Homologs: two sequences are homologous if they share a common evolutionary ancestry. THERE ARE NO DEGREES OF HOMOLOGY

Orthologs: are homologs that diverged due to speciation

Paralogs: are homologs that diverged due to duplication

Tree topology

Two alternative representations of a phylogenetic tree:

Unscaled branches: terminal nodes are lined up and nodes are positioned proportionally to time of divergence (e.g. Myear). Branch lengths are meaningless

Scaled branches: lengths of branches are proportional to number of molecular changes along them (e.g. nucleotide or aa substitution). The evolutionary distance between two organisms is the total length of all the branches that connect them
Disclaimers

This area of computational biology is an intellectual minefield!

Neither the theory nor the practical applications of any algorithms are universally accepted throughout the scientific community.

The application of different software packages to a data set is very likely to give different answers; minor changes to a data set are also likely to profoundly change the result.

Genes/Proteins vs. Species

- Relationships calculated from sequence data represent the relationships between sequences, this is not necessarily the same as relationships between organisms.

- Your sequence data may not have the same phylogenetic history as the organism from which they were isolated.

- Different genes evolve at different speeds, and there is always the possibility of horizontal gene transfer.
**Building Trees**

**Neighbor Joining** is the most popular way to build trees from distance measurements:
- The evolutionary distance is calculated based on the number of nucleic acid or AA changes between two sequences.
- Pairwise distances must be calculated between all sequences that will be used to build the tree.

**Maximum parsimony** is the most popular method for reconstructing ancestral relationships:
- Involves evaluating all possible trees and giving each a score based on the number of evolutionary changes that are needed to explain the observed data.
- The best tree is the one that requires the fewest base changes for all sequences to derive from a common ancestor.

---

**Are there Correct trees??**

- It is actually quite simple to use computer programs calculate phylogenetic trees for data sets.
- Provided the data are clean, outgroups are correctly specified, appropriate algorithms are chosen, no assumptions are violated, etc., can the true, correct tree be found and proven to be scientifically valid?
- Unfortunately, it is impossible to ever conclusively state what is the “true” tree for a group of sequences (or a group of organisms); taxonomy is constantly under revision as new data is gathered.

---

**Evaluating trees: Bootstrapping**

Bootstrapping is a resampling of data to determine sampling error or confidence intervals for some estimated parameter.

The computer generates hundreds of version of the tree to confirm that the final tree fits the data best.

The percentage of time that the bootstraps support the original tree topology are noted for each node.
- **Phylogenetic Tree of Life**
  Based on small subunit rRNA sequences, most common conserved molecule to study phylogeny

- **Combined protein sequence trees**
  Trees built using large combined alignments of 23 orthologous proteins conserved across 45 species from all domains of life. Trees highly congruent with SSU rRNA trees in their strong support for the 3 domains (Brown et al., 2001)

- **Gene content trees**
  Huynen et al. 1999; Fitz-Gibbon and House, 1999

Unrooted SSU rRNA tree displaying the 3 domains of life
Gene content tree based on the number of genes shared between two genomes (orthologous genes) divided by the number of genes in the smallest genome. Tree similar to phylogenies based on SSU-rRNA (Huynen et al., 1999)

Computer Software for Phylogenetics

Due to the lack of consensus among evolutionary biologists about basic principles for phylogenetic analysis, it is not surprising that there is a wide array of computer software available for this purpose.

- **PHYLIP** is a free package that includes 30 programs that compute various phylogenetic algorithms on different kinds of data.
- The GCG package (available at most research institutions) contains a full set of programs for phylogenetic analysis including simple distance-based clustering and the complex cladistic analysis program PAUP (Phylogenetic Analysis Using Parsimony).
- **CLUSTALX** is a multiple alignment program that includes the ability to create trees based on Neighbor Joining.
- **MacClade** is a well-designed cladistics program that allows the user to explore possible trees for a data set.