Phylogenetics: Recovering Evolutionary History

COMP 571
Luay Nakhleh, Rice University
The Structure and Interpretation of Phylogenetic Trees

(A) unrooted, binary species tree
- internal node
- branch (edge)
- external nodes (leaves)

(B) rooted, binary species tree
- root
- speciation

Flow of time (direction of descent)

- six extant taxa or operational taxonomic units (OTUs)
The Structure and Interpretation of Phylogenetic Trees
The Structure and Interpretation of Phylogenetic Trees

* In a binary tree on \( n \) taxa, how many nodes, branches, internal nodes and internal branches are there?

* How many unrooted binary trees on \( n \) taxa are there? How many rooted binary trees on \( n \) taxa are there?
The Structure and Interpretation of Phylogenetic Trees

polytomy

Non-binary Multifuracting Partially resolved Polytomous

© six extant taxa or operational taxonomic units (OTUs)
The Structure and Interpretation of Phylogenetic Trees

- A polytomy in a tree can be resolved (not necessarily fully) in many ways, thus producing trees with higher resolution (including binary trees).

- A binary tree can be turned into a partially resolved tree by contracting edges.

- In how many ways can a polytomy of degree $d$ be resolved?

- Compatibility between two trees guarantees that one can back and forth between the two trees by means of node refinement and edge contraction.
The Structure and Interpretation of Phylogenetic Trees

branch lengths have no meaning

ultrametric tree (molecular clock)

Additive tree

Additive tree rooted at an outgroup
The Structure and Interpretation of Phylogenetic Trees

11 clades (4 nontrivial)

9 bipartitions (3 nontrivial)

How many nontrivial clades are there in a binary tree on n taxa?
How many nontrivial bipartitions are there in a binary tree on n taxa?
How many possible nontrivial clusters of n taxa are there?
The Structure and Interpretation of Phylogenetic Trees

Species vs. Gene Tree

(More on this later)
The Structure and Interpretation of Phylogenetic Trees

A tree can be represented as a set of splits

(A)

(B)

<table>
<thead>
<tr>
<th></th>
<th>raccoon</th>
<th>bear</th>
<th>dog</th>
<th>sea lion</th>
<th>seal</th>
<th>weasel</th>
<th>cat</th>
<th>monkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>raccoon</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
The Structure and Interpretation of Phylogenetic Trees

The Newick Format

(monkey, (cat, (weasel, ((seal, sea_lion), (dog, (bear, raccoon))))));
The Structure and Interpretation of Phylogenetic Trees

Branch support and condensed trees

(A) Bootstrap support values (more on this later)

(B) Contracting edges with poor support (bootstrap value < 60%)
The Structure and Interpretation of Phylogenetic Trees

Consensus of sets of trees

- Strict consensus
- Majority consensus (60%)
- Majority consensus (50%)
Molecular Evolution and Its Consequences
Molecular Evolution and Its Consequences

The Darwinian concept of evolution by natural selection concentrates on the consequences of evolutionary changes for the fitness of the organism: its ability to survive and transmit its genes to the next generation by producing offsprings.

Fitness depends on the properties of the organism as a whole, and thus change at the DNA sequence level will be constrained by considerations of how it affects protein expression and function, and how these affect cellular properties and whole-organism physiology and behavior.
Molecular Evolution and Its Consequences

- In the case of sequence-based phylogenetic reconstruction, the changes that occur in genomic DNA are the main focus and, where relevant, their effects on amino acid sequence.
- Changes may affect one or a few nucleotides, entire genes, and even whole genomes.
Multiple Mutations at a Site

- Most related sequences have many positions that have mutated several times.
- The \textit{p distance} between two sequences is the fraction of nonidentical alignment positions.
- The \textit{p distance} is almost always an underestimate of the number of mutations that actually occurred.
- This distance is, therefore, \textit{corrected} to reflect the correct evolutionary distance.
Multiple Mutations at a Site

![Graph showing observed p-distance vs. average number of mutations per site. The graph compares the PAM model with nonoverlapping mutations.]
Transitions vs. Transversions

* The rate of accepted mutation is usually not the same for all types of base substitutions

- Transition mutations have little effect on the DNA structure, and hence are much more commonly observed

- The transition/transversion ratio, $R$, is defined as the number of transitions per transversion during the evolution of the sequences being studied (more on this later)
Synonymous vs. Nonsynonymous Substitutions

Another factor that affects the acceptance rates of mutations in protein-coding sequences is the effect of the mutation on the amino acid sequence, and thus, potentially, on the function of the protein.

Nucleotide mutations that do not change the encoded amino acid are called **synonymous mutations** (most changes at the third codon position are synonymous).

Synonymous mutations are generally considered to be neutral.

Nucleotide mutations that alter the encoded amino acid are called **nonsynonymous mutations**.
Evolutionary models exist that can accommodate variations in mutation rates between transitions and transversions (by setting the value of R) as well as variations in mutation rates at different sites (usually by using the Gamma correction).

In analyzing a data set that involves long evolutionary timescales, it may be useful to remove the third codon sites from the data before further analysis.
Homology

* The term was introduced by Richard Owen in 1843 to designate “the same organ in different animals under every variety of form and function”

* The distinction between orthology and paralogy was introduced by Walter Fitch in 1970
Orthology vs. Paralogy

(A) 
- time 
- gene duplication event 1 (both genes initially identical) 
- genes undergo functional divergence to functions $\alpha$ and $\beta$ 
- speciation 
- gene duplication event 2 (both genes initially identical) 
- $B\alpha$ genes undergo functional divergence to functions $\alpha$ and $\gamma$ 

species A 

species B

(B) 
- gene duplication event 1 
- speciation 
- gene duplication event 2 

$A\alpha$ $B\alpha$ $B\gamma$ $A\beta$ $B\beta$
The Effects of Gene Loss
Species/Gene Tree Reconciliation

Species tree

- human
- chicken
- Xenopus
- Catostomus
- Drosophila
- Artemia
- Hydra

Gene tree

- human 1
- chicken 1
- Xenopus 1
- Catostomus 1
- human 2
- chicken 2
- human 3
- chicken 3
- Drosophila
- Artemia
- Hydra

Reconciled gene tree

- human
- chicken
- Xenopus
- Catostomus
- Drosophila
- Artemia
- Hydra

Reconcile

δ₁
δ₂
δ₃
Horizontal Gene Transfer

The genes in the donor and recipient species are called xenologues.
Common practice: Only orthologous genes should be used to construct species phylogenetic trees

Caveats:

- Sometimes even the gene tree of orthologous genes disagrees with the species tree (e.g., **incomplete lineage sorting**)

- Events such as gene duplication, loss, and HGT do not necessarily result in gene trees that disagree with the species tree

- Under certain evolutionary scenarios (such as hybrid speciation), a tree does not exist; rather, the species phylogeny is a directed, acyclic graph (known as **phylogenetic network**)
Questions?