Phylogenetic Tree Reconstruction

Yuzhen Ye (yye@indiana.edu)
School of Informatics & Computing, IUB
Evolution theory

- Speciation
  - Evolution of new organisms is driven by
    - Mutations
      - The DNA sequence can be changed due to single base changes, deletion/insertion of DNA segments, etc.
    - Selection bias
      - Speciation events lead to creation of different species.
      - Speciation caused by physical separation into groups where different genetic variants become dominant
  - Any two species share a (possibly distant) common ancestor
- The molecular clock hypothesis
- Indiana University (Michael Lynch, Jeff Palmer, Matt Hann et al)
  Lynch: The Origins of Genome Complexity
Phylogeny (phylogenetic tree)

- A phylogenetic tree is a graph reflecting the approximate distances between a set of objects (species, genes, proteins, families) in a hierarchical fashion.

- **Leaves** – current species; sequences in current species.
- **Internal nodes** - hypothetical common ancestors.
- **Branches (Edges) length** - “time” from one speciation to the next (branching represents *speciation* into two new species).
Phylogenetic trees (binary trees)

Rooted tree

Rooted tree satisfying “molecular clock” hypothesis: all leaves at same distance from the root.

Unrooted tree:

Note: 1-5 are called leaves, or leaf nodes. 6-8 are inferred nodes corresponding to ancestral species or molecules. Branches are also called edges. The edge lengths reflect evolutionary distances.

UMPGA
Morphological vs. molecular

- Classical phylogenetic analysis: **morphological** features: presence or absence of fins, number of legs, lengths of legs, etc.

- Modern biological methods allow to use **molecular** features
  - Gene sequences
  - Protein sequences

- Species tree and gene tree
There are many possible types of sequences to use (e.g. Mitochondrial vs Nuclear proteins).
How to choose the best tree?

- To decide which tree is best we can use an optimality criterion.
- Parsimony is one such criterion (the other criteria: Maximum likelihood, minimum evolution, bayesian)
- It chooses the tree which requires the fewest mutations to explain the data.
- The **Principle of Parsimony** is the general scientific principle that accepts the simplest of two explanations as preferable.
Principle of Parsimony

Parsimony based methods look for a tree with the minimum total number of substitutions of symbols between species and their ancestor in the phylogenetic tree.

The left tree is preferred over the right tree.
Maximum parsimony

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CACCCCTT</td>
<td>AACCCTCAT</td>
<td>CACTGCTT</td>
<td>AACTGCTA</td>
</tr>
</tbody>
</table>

\[(S1, S2), (S3, S4)\]: 20011011 \(\mathbf{6}\) \(\checkmark\)

\[(S1, S3), (S2, S4)\]: 10022011 \(\mathbf{7}\)

\(\text{mutation}=2\)

\(\text{mutation}=1\)

(S1, S2), (S3, S4)

(S1, S3), (S2, S4)
We can EASILY get different trees (the “reality check” paper)

- Input sequences
- Multiple alignment programs
- Substitution models
- Phylogenetic tree reconstruction methods
Trees – what might they mean?

Species tree:
- Species A
- Species B
- Species C
- Species D

Gene tree:
- Seq A
- Seq D
- Seq C
- Seq B
Lack of resolution

Weak support, 40% bootstrap for bipartition (AD)(CB)
(typical >80%)
Long branch attraction (LBA)

the two longest branches join together

Strong support, e.g., 100% bootstrap for (AD)(CB)
Horizontal gene transfer

Species tree

Gene tree

Gene Transfer

species A

species B

species C

species D

Seq A

Seq B

Seq C

Seq D
Gene duplication & loss

Species tree

Gene tree

Duplication

Loss

species A
species B
species C
species D

Seq A
Seq B
Seq C
Seq D
Seq B'
Seq C'
Seq D'

A
D
C
B
Orthologs and paralogs
(important for function annotation)

SEQ A
SEQ B
SEQ C
SEQ D

Orthologs:
sequences diverged after a speciation event

Paralogs:
sequences diverged after a duplication event

Xenologs:
sequences diverged after a horizontal transfer

Duplicated genes may have different functions!!
Phylogeny (phylogenetic tree) reconstruction: overview

- Tree topology & branch lengths
- Computational challenge
  - Huge number of tree topology
    - 3 sequences: 1 (unrooted)
    - 4 sequences: 3
    - 5 sequences: 15
    - 10 sequences: 2,027,025
    - 20 sequences: 221,643,095,476,699,771,875
    - n sequences (unrooted & rooted) ?

- Most methods are heuristic
- Two types of methods
  - Distance based (input: distance matrix; UPGMA & NJ)
  - Character based (input: multiple alignment)
Models of evolutionary distance

1. Simplest case: Jukes-Cantor model
   -- equal probability of change to any nucleotide

2. Other models take into account transitions vs. transversion frequencies
   -- Kimura: different probabilities for transitions, transversions
   -- HKY: different probabilities for transitions, transversions &
     takes into account genomic nucleotide biases

Transition: R to R
   Y to Y

Transversion: R to Y
   Y to R

where R = A,G
   Y = C,T
Distance based phylogeny reconstruction

- Phylogeny reconstruction for 3 sequences is EASY
  - There is a single tree topology
  - The branch lengths can be calculated as follows:

To compute: branch lengths $a$, $b$ and $c$, such that

\[
\begin{align*}
  a + b &= D_{AB} \\
  b + c &= D_{BC} \\
  a + c &= D_{AC}
\end{align*}
\]

Input: $D_{AB}$, $D_{BC}$ and $D_{AC}$ (pairwise distances)

Output: $a = \frac{D_{AB} + D_{AC} - D_{BC}}{2}$, $b = \frac{D_{AB} + D_{BC} - D_{AC}}{2}$, and $c = \frac{D_{AC} + D_{BC} - D_{AB}}{2}$
Fitch-Margoliash (FM) algorithm

- For phylogeny reconstruction with more than 3 sequences
- For example, given 5 sequences, A, B, C, D and E. The tree can be reconstructed as follows
  - First choose the closest sequence pair, suppose it is D and E (based on the input pairwise distances; e.g., $D_{DE}=10$)
  - To calculate the branch lengths from D and E to their common ancestor (denoted as $d$ and $e$), we combine the remaining three sequences (A, B and C) and treat them as a single composite sequence (and define $D_{D(ABC)} = \frac{D_{DA} + D_{DB} + D_{DC}}{3}$ and so on) -- so again we are dealing with 3 sequences, and we can easily calculate the branch lengths
  - Then merge D and E into a cluster and treat it as a composite sequence, and update the distance table so that $D_{A(DE)} = \frac{D_{DA} + D_{AE}}{2}$ and so on.
  - Repeat the above steps until no more clusters to merge
Distance based method: UPGMA

- UPGMA: Unweighted Pair Group Method with Arithmetic Mean
- Assume same rate evolution (molecular clock hypothesis)
- The length from root to each leaf is the same (ultra metric).
- It is similar to Fitch-Margoliash algorithm (merge two most similar sequences or clusters first); but the calculation of branch lengths is even simpler.
- For example, for the same example shown above with five sequences, \(d=e=5\) (d and e are the branch lengths from sequence D and E to their common ancestor)
Neighbor-joining (NJ)  
(Seitou & Nei algorithm)

- Minimum evolution -- the least total branch length (distance-based)
- Bottom-up clustering method
- Does not assume same rate evolution
- Fast & produce reasonable trees
NJ method

NJ looks for two sequences (clusters) to merge that minimizes:

$D(C_1, C_2) - u(C_1) - u(C_2)$

Where

$u(C) = \frac{1}{\text{number of clusters} - 2} \sum_{\text{all clusters } C'} D(C, C')$

$C_1$ and $C_2$ are most similar to each other, while they are most dissimilar to the other clusters (far from others)
Comparison of FM, UPGMA and NJ methods

- All are hierarchical clustering methods
- All define the distance between two clusters as the average pairwise distance

\[ D(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{i \in C_1} \sum_{j \in C_2} D(i, j) \]

- FM and NJ do not assume “molecular clock”; UPGMA does and it uses a simpler way to calculate branch lengths.
- UPGMA and FM choose and merge the closest sequence (cluster) pair first, but NJ looks for two sequences (clusters) that are not only close to each other (as in UPGMA and FM) but also far apart from the rest
Parsimony based reconstruction

1. A procedure to find the minimum number of changes needed to explain the data for a given tree topology, where species are assigned to leaves. (Small Parsimony Problem)

2. A search through the space of trees. (hard problem!) (Large Parsimony Problem)
Small Parsimony Problem

- Compute the minimum number of mutations on a GIVEN tree
- Fitch algorithm
- Sankoff algorithm (subtree; DP)
The Fitch Algorithm

- Pick an arbitrary root to work towards (for unrooted tree)
- Work from the tips of the tree towards the root. Label each node with the intersection of the states of its child nodes.
- If the intersection is empty label the node with the union and add one to the cost

Calculate Fitch score

Cost=4

Internal node labeling
Sankoff Algorithm

- More general than the Fitch algorithm.
- Assumes we have a table of costs $c_{ab}$ for all possible changes between states $a$ and $b$ (A, T, C or G for DNA).
- For each node $i$ in the tree we compute $S(i,a)$ the minimal cost given that node $i$ is assigned state $a$.
- In particular we can compute the minimum value over $a$ for $S(root,a)$ which will be the cost of the tree.
Sankoff algorithm: DP

**Initialization:**
\[ S(i,a) = \begin{cases} 0 & \text{i labeled by a, or} \\ \infty & \text{otherwise} \end{cases} \]
\[ a \in \{A,T,C,G\} \]

**Iteration:**
\[ S(i,a) = \min_x (S(j,x) + c(a,x)) + \min_y (S(k,y) + c(a,y)) \]

**Termination:**
\[ \min_a S(\text{root},a) \]
Large Parsimony Problem

- The small parsimony problem – to find the score of a given tree - can be solved in linear time in the size of the tree.
- The large parsimony problem is to find the tree with minimum score.
- It is known to be NP-Hard.
Tree search strategies

- **Exact search**
  - possible for small n only

- **Branch and Bound**
  - Use “cleaver” rules to avoid some branches of trees
  - up to ~20 (25) taxa

- **Local search - Heuristics**
  - Pick a good starting tree and use moves within a “neighbourhood” to find a better tree; e.g., nearest-neighbor interchanges (NNIs)

- **Meta-heuristics**
  - Genetic algorithms
  - Simulated annealing
Branch and bound

Search tree of trees
Branch and bound

same, with parsimony scores in place of trees

This branch can be safely “neglected”!
Nearest neighbor interchange

Crossover – nearest neighbor interchange (nni)

Or, the tree $(t_1, t_4) \ (t_2, t_3)$
Probabilistic approaches to phylogeny

- Rank trees according to their likelihood $P(data|tree)$, or, posterior probability $P(tree|data)$ (Bayesian)
- Maximum likelihood methods
- Sampling methods
Calculate likelihood
Felsenstein’s algorithm for likelihood

- **Initialization:** $i=2n-1$
- **Recursion:** Compute $P(L_i|a)$ for all $a$ as follows
  if $i$ is leaf
  
  \[ P(L_i|a) = \begin{cases} 1 & \text{if } a \text{ is the label of the leaf, otherwise } 0 \end{cases} \]
  
  else
  
  Compute $P(L_j|a)$, $P(L_k|a)$ for all $a$
  
  \[ P(L_i|a) = \sum_{b,c} P(b|a,t_j)P(L_j|b)P(c|a,t_k)P(L_k|c) \]
How can one tell if a tree is significant

*Biological knowledge*

**Bootstrapping:** how dependent is the tree on the dataset
1. Randomly choose $n$ objects from your dataset of $n$, *with replacement*  
   *(picking columns from the alignment at random with replacement)*
2. Rebuild the tree based on the subset of the data
3. Repeat many (1,000 – 10,000) times
4. How often are the same children joined?

**Jackknifing:** how dependent is the tree on the dataset
1. Randomly choose $k$ objects from your dataset of $n$, *without replacement*
2. Rebuild the tree based on the subset of the data
3. Repeat many (1,000 – 10,000) times
4. How often are the same children joined?
Commonly used phylogeny packages

  - Phylip (general package, protdist, NJ, parsimony, maximum likelihood, etc)
  - PAUP (parsimony)
  - PAML (maximum likelihood)
  - TreePuzzle (quartet based)
  - PhyML (maximum likelihood)
  - MyBayes
  - MEGA (biologist-centric)
What a surprise

“..because of their intrinsic reliance on summary statistics, distance-matrix methods are assumed to be less accurate than likelihood-based approaches. In this paper [Science, 327:1376–1379, 2010], pairwise sequence comparisons are shown to be more powerful than previously hypothesized. A statistical analysis of certain distance-based techniques indicates that their data requirement for large evolutionary trees essentially matches the conjectured performance of maximum likelihood methods—challenging the idea that summary statistics lead to suboptimal analyses.”
Readings

- Chapter 7 (Revealing Evolutionary History)
- Chapter 8 (Building Phylogenetic Trees)