**Spring 2023**

**BIOS 477/877**

**Bioinformatics and Molecular Evolution**

**Lecture 23**

---

**TODAY'S TOPICS**

- Phylogenetic reconstruction
  - Tree searching (Exhaustive, Branch-and-Bound, Heuristic)
  - Branch support (bootstrap analysis, etc.)

---

**Phylogenetic methods**

- Data types and tree-building methods

  - [Data types]
    - Distances
    - Characters

  - Clustering
    - UPGMA
    - Neighbor joining

  - Optimal criterion
    - Minimum evolution
    - Maximum parsimony
    - Maximum likelihood

  (Bayesian inference)

- Clustering vs. search methods
  - Clustering methods (UPGMA, Neighbor-joining)
    - Do not search all possible topologies
    - Very fast
    - Produce only one tree
  - Search methods
    - Use optimality criterion (minimum evolution, maximum parsimony, maximum likelihood)
    - Exhaustive search for all possible topologies is not possible for a large number of taxa
    - A heuristic search algorithm needs to be used

**How can we search trees?**

---

**Phylogenetic methods: tree searching**

- Number of possible tree topologies

<table>
<thead>
<tr>
<th>Number of OTUs</th>
<th>Number of rooted trees</th>
<th>Number of unrooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

---

**Phylogenetic methods: tree searching**

- Number of possible tree topologies

<table>
<thead>
<tr>
<th>Number of OTUs</th>
<th>Number of rooted trees</th>
<th>Number of unrooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

---

### Phylogenetic methods: tree searching

#### Number of possible tree topologies

<table>
<thead>
<tr>
<th>Number of OTUs</th>
<th>Number of rooted trees</th>
<th>Number of unrooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>954</td>
<td>954</td>
</tr>
<tr>
<td>6</td>
<td>18,595</td>
<td>18,595</td>
</tr>
<tr>
<td>7</td>
<td>3,212,895</td>
<td>3,212,895</td>
</tr>
<tr>
<td>8</td>
<td>63,756,165</td>
<td>63,756,165</td>
</tr>
<tr>
<td>9</td>
<td>1,271,233,945</td>
<td>1,271,233,945</td>
</tr>
<tr>
<td>10</td>
<td>2,542,463,790</td>
<td>2,542,463,790</td>
</tr>
</tbody>
</table>

Impossible to examine all of the possible trees!

---

### Tree-searching methods (unrooted)

#### Exhaustive search

- Possible only for a few taxa (11 or fewer)
- Number of possible unrooted tree:
  \[ B(t) = \prod_{i=1}^{t-1} (2i - 5) \]
  \( t \): number of taxa

  e.g.,
  
  \[ B(7) = 1 \times 3 \times 5 \times 7 \times 9 = 945 \]
  \[ B(10) = B(7) \times 11 \times 13 \times 15 = 2 \times 10^9 \]
  \[ B(20) = 2 \times 10^{18} \]

- An algorithm is required to guarantee generation of all possible trees

---

### Tree-searching methods (unrooted)

#### Exhaustive search (unrooted)

- If there are 2 taxa
  
  Only one topology possible

  There is only one place to add the third taxon

- If there are 3 taxa

  There are three places to add the fourth taxon

  Three topologies are possible

---

### Tree-searching methods (unrooted)

#### Exhaustive search (unrooted)

- If there are 4 taxa

  There are five places to add the fifth taxon

  Five topologies are possible

---

### Tree-searching methods (unrooted)

#### Exhaustive search for 5-taxon trees

- 3 ways to add another branch

  5 ways to add another branch

  Each of the possible topologies (3 x 5 = 15 trees for 5-taxon tree) will be evaluated to identify optimal trees (ME, MP, or ML trees)

- 5 ways to add another branch

  5 ways to add another branch
Tree-searching methods (unrooted)

- **Exhaustive search**
  - searches all possible trees
  - guarantees to find the optimal trees
  - impractical for many taxa

- **Branch-and-bound method**
  - Alternative exact method
  - useful for 25 or so taxa
  - implicitly evaluates all possible trees
  - but cuts off search paths that do not lead to optimal trees
    - reduces the number of trees to evaluate
    - will find the optimal tree(s)

To search 6-taxon MP tree

241 becomes the upper bound ($L_u = 241$) for searching 6-taxon trees

(MP example: searching the shortest tree length, L)

229 is smaller than the current $L_u (241)$
New upper bound: $L_u = 229$
Tree-searching methods (unrooted)

- **Branch-and-bound method**
  - [To search 6-taxon MP tree]
  - \( L = 213 \)  \( \rightarrow \)  \( L_u = 229 \)
  - [Larger than the current upper bound (\( L_u = 229 \)) Stop searching]

- **Branch-and-bound method**
  - [To search 6-taxon MP tree]
  - \( L = 213 \)  \( \rightarrow \)  \( L_u = 229 \)
  - [Larger than the current upper bound (\( L_u = 229 \)) Stop searching]

**B&B method guarantees to find the optimal tree without going through all topologies**

**Heuristic (approximation) method**

- **Stepwise addition**: greedy algorithm
  - Add one branch
  - Evaluate \( L \)
  - Take only the optimal topology at this level

**Heuristic (approximation) method**

- **Stepwise addition**: greedy algorithm
  - No guarantee to find the path that leads to the optimal topology
  - Rarely identifies a globally optimal topology
Tree-searching methods (unrooted)

- **Heuristic (approximation) method**
  - **Stepwise addition**: greedy algorithm
    - Rarely finds the global optima
    - Can be improved (refinement)
  - **Branch-swapping**
    - Cutting off one or more subtrees and reassembling them to generate locally different trees
    - Repeat this process many times to find better topologies
    - Still can be entrapped in local optima
  - **Random addition + branch-swapping (many times)**
    - Multiple optimal islands can be identified

Tree-searching landscape

Reliability of inferred trees

- **Bootstrap analysis**
  - Evaluates the reliability of each cluster
  - A sampling technique to estimate the statistical errors when underlying sampling distribution is unknown
  - Approximates the underlying distribution by resampling from the original dataset
  - First applied by Felsenstein (1985) for phylogenetic analysis

Reliability of inferred trees

- **Bootstrap analysis**
  - Resampling is done on sequence data
  - Each column of the multiple alignment is treated as an independent sample
Bootstrap analysis

Reliability of inferred trees

Columns are randomly sampled

1. Bootstrap replicate with the same length (the same number of sites) as the original alignment
2. Many resamplings are done (~1000 times)
3. Multiple samplings are allowed from the same column
A phylogeny is reconstructed from each pseudoreplicate (~1000 trees).

Bootstrap analysis:

<table>
<thead>
<tr>
<th>Bootstrap support (%)</th>
<th>(~1000 alignments)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>(S1, S2, S3), (S4, S5)</td>
</tr>
</tbody>
</table>

Reliability of inferred trees:

- **Bootstrap analysis**
  - S1 S2 S3 S4 S5
  - (S1, S2, S3), (S4, S5): 3/3 = 100%
  - (S1, S2), (S3, S4, S5): 2/3 = 67%
  - (S3, S4), (S1, S2, S5): 1/3 = 33%

Consensus tree:

- Majority consensus tree reflects the majority relationships

Consensus tree:

- All 100% bootstrap support
- Strict consensus tree includes only clusters supported by all trees
- Most conservative
Interpreting bootstrap support

- Bootstrap values can be considered to measure the probability of the branch length being greater than 0.

- No cluster supported
  - A/B cluster supported
    - C/D/E cluster is also supported
      - Internal branch exists (branch length > 0) → supported by 90%

- Bootstrap values > 95% or 99% (often 70~80% used as a threshold)
- If the internal branches are very short or close to 0, both underestimation and overestimation could happen.

Transfer distance (or R distance):
The minimum number of elements to be transferred (or removed) to transform one partition into the other.

Renewing Felsenstein’s phylogenetic bootstrap in the era of big data

Transfer Bootstrap Expectation (TBE)
- For branch b, TBE: include most similar partitions
- Include only identical partitions TBE: include most similar partitions (no need to be identical)

PHYLIP format for input alignment

Alignment length

Sequence number

Sequence names should be ten characters in length. Add spaces AFTER the sequence name if names are too short.

Do not include space or "_" in the sequence names.
Strict vs. relaxed PHYLIP alignment format

Strict PHYLIP format
4 20
1c2x
lyc2
letp

Relaxed PHYLIP format
4 20
1c2x
lyc2
letp

- Only up to 10 characters are used as the sequence names.
- No space is needed between names and sequences.
- Relaxed format allows to have longer sequence names.
- At least one space is needed between names and sequences.
- Used in PhyML, RAxML, etc.
- Relax format is not supported by PHYLIP (sequence data start at 11th character)

In the strict format: Names > 10 letters will be truncated!!!
Maximum likelihood phylogeny: IQ-TREE

http://www.cipres.org; http://github.com/cipres-exe/iqtree

IQ-TREE web server: fast and accurate phylogenetic trees under maximum likelihood

- Includes: RAXML, MrBayes, FastTree, Bayesian Inference of Phylogeny
- Standalone web interface (Summary, standard bootstrap, and transfer bootstrap)
- Randomized Ancestor Maximum Likelihood (includes rapid bootstrap)
- Faster and more accurate than RAxML, includes transfer bootstrap
- A fast and effective stochastic algorithm for estimating ML phylogenies
- Includes model selection and ultrafast bootstrap; IQ-TREE 2 also includes transfer bootstrap
- Power and coverage comparable to RAxML, including transfer bootstrap
- Approximately maximum likelihood phylogenetic trees (standalone only)
- Faster and more accurate than RAxML, includes transfer bootstrap
- A high performance parallel algorithm: RAxML-NG, MrBayes, GARLI, FastTree, IQ-TREE, etc.
- Multiple Bayesian inferences of Phylogeny
- High performance parallel code: RAxML-NG, MrBayes, GARLI, FastTree, IQ-TREE, etc.
- XG-Phylogeny
- Phylogeny for filtering
- Phylogeny for

More phylogeny programs/websites

- T-REx
- PhyML
- FastTree
- Phylogenetic method programs/websites
- CIPRES
- Phylo toolkits
- Phylogeny program list

Partitions can be specified:
- Each partition can have a different substitution model (e.g., different genes or even DNA and protein alignments can be combined)

DNA substitution model

AA substitution model

Different alignment formats can be used

Default branch support: Ultrafast bootstrap

- SH-aLRT branch test

Substitution model can be chosen automatically

Maximum likelihood phylogeny: IQ-TREE

http://www.cipres.org; http://github.com/cipres-exe/iqtree

IQ-TREE web server: fast and accurate phylogenetic trees under maximum likelihood

- Includes: RAXML, MrBayes, FastTree, Bayesian Inference of Phylogeny
- Standalone web interface (Summary, standard bootstrap, and transfer bootstrap)
- Randomized Ancestor Maximum Likelihood (includes rapid bootstrap)
- Faster and more accurate than RAxML, includes transfer bootstrap
- A fast and effective stochastic algorithm for estimating ML phylogenies
- Includes model selection and ultrafast bootstrap; IQ-TREE 2 also includes transfer bootstrap
- Power and coverage comparable to RAxML, including transfer bootstrap
- Approximately maximum likelihood phylogenetic trees (standalone only)
- Faster and more accurate than RAxML, includes transfer bootstrap
- A high performance parallel algorithm: RAxML-NG, MrBayes, GARLI, FastTree, IQ-TREE, etc.
- Multiple Bayesian inferences of Phylogeny
- High performance parallel code: RAxML-NG, MrBayes, GARLI, FastTree, IQ-TREE, etc.
- XG-Phylogeny
- Phylogeny for filtering
- Phylogeny for

Phylogeny program list

- T-REx
- PhyML
- FastTree
- Phylogenetic method programs/websites
- CIPRES
- Phylo toolkits
- Phylogeny program list

Partitions can be specified:
- Each partition can have a different substitution model (e.g., different genes or even DNA and protein alignments can be combined)

DNA substitution model

AA substitution model

Different alignment formats can be used

Default branch support: Ultrafast bootstrap

- SH-aLRT branch test

Substitution model can be chosen automatically