“Although the technical side of tree building may appear to be a matter of pure graph theory and combinatorial optimization, the fundamental issues that determine the validity of these methods are sometimes discussed in terms more suited for religion.”  (Gusfield, 1997)
Methods for reconstructing phylogeny

Hennigian phylogenetic inference and parsimony
characters, novelties and clades
Hennigian inference, basic ‘algorithm’ for parsimony
justification for parsimony as criterion

‘Modern’ tree building/searching methods

clustering methods
follow a set of steps (algorithm) and arrive at a tree
fast, easy to implement, almost always produce single tree
limitations: result often depends on order of taxa added to make tree,
more important- they do not allow evaluation of competing hypotheses
e.g., UPGMA, neighbor-joining (NJ)

optimality methods
choose among set of all possible trees (i.e., solutions)
requires an explicit function of relationship between data and tree
criterion used to assign each tree a score - allows evaluation of trees
computationally expensive, difficult to determine if ‘optimal’ tree is found
e.g., parsimony, maximum likelihood, certain distance methods
## Phylogenetic inference methods

Some common phylogenetic methods classified by the method used to build the tree, and by type of data.*

<table>
<thead>
<tr>
<th>Type of data</th>
<th>Molecular sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distances</td>
<td>UPGMA</td>
</tr>
<tr>
<td></td>
<td>Neighbor joining</td>
</tr>
<tr>
<td></td>
<td>Minimum evolution</td>
</tr>
<tr>
<td>Maximum parsimony</td>
<td>Maximum likelihood/ Bayesian</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tree-building method</th>
<th>Clustering algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPGMA</td>
<td></td>
</tr>
<tr>
<td>Neighbor joining</td>
<td></td>
</tr>
<tr>
<td>Minimum evolution</td>
<td></td>
</tr>
</tbody>
</table>

### Desirable properties a tree-building method should have:

- **efficiency** - how fast is it?
- **power** - how much data does the method need to produce a reasonable result?
- **consistency** - will it converge on the right answer given enough data?
- **robustness** - will minor violations of the method’s assumptions result in poor estimates of phylogeny?
- **falsifiability** - will the method tell us when its assumptions are violated, i.e., is it appropriate to the question, data?

*Page & Holmes (1998)*

*Page & Holmes (1998)*
## Comparison of phylogenetic inference methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantage(s)</th>
<th>Disadvantage(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighbor joining</td>
<td>Fast</td>
<td>Information is lost in converting sequences into distances; reliable estimates of distances may be hard to obtain with divergent sequences</td>
</tr>
<tr>
<td>Minimum evolution</td>
<td>Uses models to correct for multiple hits (unseen changes)</td>
<td>Distance corrections problematic when distances are large</td>
</tr>
<tr>
<td>Parsimony</td>
<td>Fast enough for analyses of 100s of sequences; robust if branches are not hugely different in length</td>
<td>Can perform poorly if there is substantial variation in branch lengths</td>
</tr>
<tr>
<td>Maximum likelihood</td>
<td>The likelihood fully “captures” what the data says about the phylogeny under a given model</td>
<td>Can be prohibitively slow (depending on # of taxa and thoroughness of search)</td>
</tr>
<tr>
<td>Bayesian</td>
<td>Has strong connection to maximum likelihood method; often faster way to assess clade support for trees than ML bootstrapping; produces multiple trees at or near maximum</td>
<td>“Prior” distributions for parameters must be specified; can be difficult to determine whether Markov chain Monte Carlo approximation have run for enough time</td>
</tr>
</tbody>
</table>
Optimization methods

Optimality methods like parsimony pose two problems that must be solved in searching for the “best” tree:

(i) For a given data set and a given tree, what is the value of the optimality criterion for that tree?
(ii) Which of all possible trees has the maximum value of this criterion?

Of the two, the second is more difficult, and belongs to a class of problems called “NP-complete” problems for which no efficient algorithms for their solution are known to exist. It is thought, however, that if one problem could be solved efficiently all of them could be.

In practice, for any ‘reasonable’ number of sequences (c. 20) it is often impossible to guarantee that the optimal tree has been found.

Consequently, we must rely on heuristic methods - “quick and dirty” strategies to explore subsets of all possible trees in a process likened to ‘hill climbing’ to figure out if one is on a local peak or not; trees found by such methods may turn out to be far from optimal.
Optimization methods: likened to “hill climbing”

The general problem of optimization. In this figure, it is assumed that higher values of the objective function are “better”, i.e., closer to optimal.
Tree searching strategies: practical issues

1. **When number of taxa is < 20**
   - **exact solutions** - algorithms guaranteed to find the optimal tree exist
   - **exhaustive search** is a brute-force strategy
   - **branch-and-bound** similar to exhaustive but does not have to examine every tree that is descendant from initial “search tree” and suboptimal trees

2. **When N > 20 taxa** we can’t possibly examine all possible trees, so how do we find the best tree or trees?
   - **approximate or “heuristic” solutions**, most of which have two phases:
     - **sequential taxon addition** phase to produce one or more trees with all taxa
     - **branch swapping** phase that rearranges these trees to find more optimal solutions

   *however, heuristic strategies are essentially greedy, in that they look for the quickest route to the top of often the nearest peak (local optima)
   the algorithm can become trapped in local optima (“islands”) of equally parsimonious trees that are not the shortest

3. **Important issues in considering these kinds of simple optimization problems:**
   - The objective function *may not* have a maximum
   - There may be more than one local maximum, making it difficult to find the ‘global’ one
   - Even local maxima may be difficult to find because the objective function may not be smooth or may exhibit other “pathological” behavior, or there may be constraints on the values of the unknown parameters

   **Note:** exactly the same strategies and issues that we discuss in terms of parsimony are not unique to parsimony, they arise with other methods and criteria for inferring phylogenies
Heuristic methods: step 1, making initial tree, taxon addition sequence

Taxa are always added sequentially to make a tree in this phase. The simplest order of addition is known as “ASIS” addition; here taxa are added in the order they appear in the matrix. The first three taxa are joined into an unrooted three-taxon tree, then the fourth taxon in the matrix is added. It can be added in one of three places, so the length of the tree is determined for each possibility and the placement that is optimal at that point in time is selected. Next, the fifth taxon is added, and so on, until a complete tree is built. Other addition sequence implemented in software such as PAUP* include RANDOM (random order addition) and CLOSEST (which chooses next taxon to be added by finding the one that would add the fewest number of steps to the new tree).
Heuristic methods: step 2, branch swapping

Branch swapping by tree bisection and reconnection (TBR). The tree is initially bisected along a branch, yielding two disjoint subtrees. The subtrees are then reconnected by joining a pair of branches, one from each subtree, with all possible bisections and reconnections evaluated. The shortest is saved and branch swapping proceeds again until a shorter tree is found.

(after Swofford et al. 1996)
On a landscape of trees, tree building by random addition sequences are used to find multiple optima, or ‘tree islands’. Branch swapping moves search nearer to top of local optima. New random addition sequences may find additional local optima.
Characters for phylogenetic inference

There are two fundamental assumptions ‘required’ of characters that are common to most character based methods of phylogenetic analysis:

1. first is the assumption of **independence** among characters. This assumption enables us to treat each position in a data matrix separately in certain time-intensive computational algorithms, thereby allowing problems to be subdivided into a number of much simpler ‘sub-problems’.

2. second is the assumption that the characters be **homologous**. The concept of **homology** is complicated by a variety of meanings, depending on the kind of character. For molecular data, two nucleotides in different gene sequences (or gene products) are homologous if and only if the two sequences acquired that state directly from their common ancestor.

Character data are either:
- **qualitative**, in which case the possible states are two or more discrete values, or
- **quantitative**, in which the characters can vary continuously and are measured on an interval scale

Qualitative characters may be **binary**, with two possible states
  e.g., presence or absence of something, “0” or “1”)
or **multistate**, with two to many possible states
  e.g., nucleotide data with A, C, G, T, or N)

Quantitative characters are less commonly used as character data in molecular systematics and evolution, with a few exceptions (e.g., mtDNA haplotypes coded as frequencies).
Characters for phylogenetic inference

Multi-state characters are of two types, **ordered** or **unordered**. This depends on whether a relationship in ordering (rank or polarity) is imposed on the possible states. Consider the examples of character state transitions . . .

(a) \begin{align*}
\text{0} & \quad \text{1} & \quad \text{2} & \quad \text{3} \\
\end{align*}

(b) \begin{align*}
\text{A} & \quad \text{B} & \quad \text{C} & \quad \text{D} \\
\end{align*}

**Ordered multistate character** (transformation between any two states that are not directly connected implies passage through one or more intermediate states)

**Unordered multistate character** (any state can transform directly into any other state)

(c) \begin{align*}
\text{2} & \quad \text{0} & \quad \text{0} & \quad \text{2} \\
\text{1} & \quad \text{1} & \quad \text{1} & \quad \text{1} \\
\text{0} & \quad \text{2} & \quad \text{0} & \quad \text{2} \\
\end{align*}

**Ordered multistate character** in which the polarity is indicated (the ordering relation is the same in all three cases but the ancestral state differs)

From Swofford et al. (1996)
DNA and protein sequences are generally considered *unordered, multistate* characters (a) since there is no *a priori* reason to assume that any one state is intermediate between any other two; for example A → T → G, A mutates to G only after first changing to T. However, from observations of many sequences, it is clear that there are often biases in the frequencies at which one nucleotide changes to another. For example, in many sequences, the number of transitions exceeds that of transversions (b). Sometimes compositional bias (unequal base frequencies) cause altered patterns of substitution.

![Diagram of nucleotide changes](image)

Equal probabilities of change from one state to another

Higher rate of transitions vs. transversions

A+T >> G+C

We also know that not every single position in a molecular sequence undergoes substitution at the same rate. In most protein-coding DNA sequences (genes) higher rates of nucleotide substitution are observed in 3rd codon positions, sometimes approaching 10-20 fold higher, compared to 1st or 2nd codon positions.

**Rate of substitution:** 3rd position >> 1st position > 2nd position
Maximum parsimony (MP)

- one of two major methods that operate directly on discrete characters or on functions derived from them, rather than on pairwise distances, the other being maximum likelihood (ML)
- use of parsimony as a modern method for building trees date to Edwards & Cavalli-Sforza’s work on human population histories (1960s) invoked it as a practical approximation to more complex ML methods
- methods based on the principle of parsimony have been the most commonly used for inferring phylogenies

Principle of parsimony essentially maintains that simpler hypotheses are preferable to more complicated ones, and that ad hoc hypotheses should be avoided whenever possible. In general, parsimony methods for inferring phylogenies operate by selecting trees that minimize the number of evolutionary steps (character state transformations) required to explain a given set of data; this is usually expressed as total tree length. For example, the steps might be base or amino acid substitutions for molecular sequence data. In most real data sets (and those of reasonable size) there will be conflicts (homoplasy) among informative characters, and in parsimony analysis hypotheses for individual characters are examined with respect to the entire set of characters analyzed. Obviously, a tree that minimizes the total number of steps also minimizes the number of extra steps (homoplasies) need to explain the data.
Parsimony

This data matrix contains character conflict. For example, character 4 suggests \{B,C\} is a monophyletic group, but characters 2 and 3 suggest \{C,D\} is monophyletic. They cannot both be true. How do we reconstruct phylogeny when the characters do not all agree?

<table>
<thead>
<tr>
<th>Taxa</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Phylogenetic analysis using parsimony is a procedure by which individual hypotheses of shared, derived characters (synapomorphies) are “tested” against one another for their overall explanatory power. The tree reconstruction with the fewest number of character state changes (sum of # of changes, here \textbf{length=5}) is considered the most parsimonious of the three possible solutions.
Parsimony

Any discussion of parsimony methods must distinguish between the optimality criterion (e.g., minimize tree length) and the actual algorithm used to search for optimal trees. Algorithms for estimating minimum-length trees are constantly being invented so it is best not to become mired in the algorithmic details. Parsimony analysis actually comprise a group of related methods united by the goal of minimizing some evolutionarily significant quantity but differing in their underlying evolutionary assumptions. For example . . .

**Fitch and Wagner Parsimony**
- the simplest parsimony methods, imposing no (Fitch) or minimal (Wagner) constraints on permissible character state changes
- Wagner parsimony assumes that any transformation from one character state to another also implies a transformation through any intervening states (Kluge & Farris, 1969)
- Fitch (1971) parsimony generalized the Wagner method to allow unordered, multi-state characters like nucleotide and protein sequences allows any state to transform directly to any other state

- both methods permit free reversibility; that is, change of character states in either direction is assumed to be equally probable, changing from one to another and back again
Parsimony

An example of Fitch algorithm applied to a single site. Consider a 5-taxon tree, A, rooted as shown. At a particular nucleotide site we observe the bases C, A, C, A, and G in the terminals. As we move down the tree we create a set containing those nucleotides (states) that are observed or compatible with the observation, as shown in tree B. In algorithmic terms, this process is known as a **postorder tree reversal**. At each interior node we create a set that is the intersection of sets at the two descendant nodes.

e.g., because \( \{C\} \cap \{A\} = \emptyset \), \( \{C\} \cap \{A\} = \{AC\} \)
and count 1 change of state (*)

At bottom node, \( \{AC\} \cap \{ACG\} = \{AC\} \)
Note three changes of state overall.

After Felsenstein (2004)
Parsimony

Tree B shows state sets computed for interior nodes. Trees C and D represent alternative, equally parsimonious reconstructions; branches on which character state change occur are indicated in bold, changes in green.

(B)

1. \{C\}
2. \{A\}
3. \{C\}
4. \{A\}
5. \{G\}

(A)

1. \{C\}
2. \{A\}
3. \{C\}
4. \{A\}
5. \{G\}

(C)

1. \{C\}
2. \{A\}
3. \{C\}
4. \{A\}
5. \{G\}

(D)
Parsimony

Justifications for parsimony center around two main arguments, including:

1. “Parsimony is a methodological convention that compels us to maximize the amount of evolutionary similarity that we can explain as homologous similarity” (Page and Holmes, p. 190), or maximize the similarity that can be attributed to common ancestry. Put another way, **parsimony strives to minimize the number of assumptions of homoplasy needed to explain the character data**. Hypotheses of homoplasy (e.g., convergence or parallel evolution) that are not needed to explain the data are considered *ad hoc* in that they attempt to explain why data does not fit a hypothesis. The most parsimonious tree minimizes the number of such *ad hoc* hypotheses required, and for that reason is preferred.

2. **Parsimony is based on an implicit assumption about evolution.** It is arguable whether the use of parsimony to choose among trees requires an implicit assumption about whether evolutionary change is common or rare. This is the most widely stated fallacy regarding parsimony in phylogenetics, that “because parsimony selects the hypothesis requiring the fewest evolutionary changes, the method assumes something about the process of evolution that might not be true – that evolution is parsimonious and change is rare”.
Parsimony

A third reason, parsimony just plain “works” under certain known conditions, i.e., it has certain desirable properties. Although its behavior under more general conditions are largely unknown, this could be said about all phylogenetic algorithms. One such desirable property is statistical consistency, defined here as the convergence to the “true” tree as more character data are gathered. Parsimony is known to be consistent under some, if not many, conditions (although inconsistent under others).
Objections to Parsimony

The principal objection to parsimony has to do with its **consistency** as a method of tree reconstruction. **Under some conditions and models of evolution it is not consistent**, that is, even if more data is added it is possible to obtain the wrong tree. Felsenstein called the behavior of parsimony in situations like this “positively misleading” because as the number of characters (sequence length) increases, we become more and more certain to infer an incorrect tree. At this point we have entered the **Felsenstein zone** (of inconsistency).

(A) Consider this hypothetical 4-taxon tree containing 2 long peripheral branches, with all other branches being very short. (B) Incorrect tree selected by maximum parsimony. In this case inconsistency is due to strongly unequal rates of change along different branches - the term “**long branch attraction**” suggested for this general phenomenon.
Distance matrix methods

A family of phylogenetic methods introduced by Cavalli-Sforza & Edwards (1967) and Fitch & Margoliash (1967), influenced by clustering algorithms of Sokal and Sneath from early 1960s.

General idea is that if one knew the actual evolutionary distance between all members of a set of sequences, one could reconstruct the evolutionary history of those sequences. Approach is to calculate a measure of the distances between each pair of species (sequences), and then find a tree that predicts the observed set of distances as closely as possible.

As Felsenstein (2004) discusses, the best way of thinking about distance methods is to consider distances as estimates of the path length separating each pair of species, and in effect, what we have then is a large number of estimated two-species ‘trees’ and we are trying to find the full, N-species tree that is implied by these. The difficulty is that the individual distances are not exactly the path lengths in the N-species tree between those two species, rather they depart from it, and the need then is to find the full tree that does the best job of estimating these two species trees. This approach is found in so-called ‘goodness of fit’ methods. A second class of methods seek the tree whose sum of branch lengths is the minimum overall (‘minimum evolution’), an approach reminiscent of parsimony.
**Distance matrix methods**

**What is a distance?**

Distance is a measure or estimate related to overall dissimilarity between two taxa, hence the term ‘pairwise distance’ is often used:

- distances can be obtained directly in the case of certain kinds of data e.g., DNA-DNA hybridization, immunological cross-reactivities
- distances are calculated (or estimated) from other kinds of data, either continuous or discrete, e.g., gene frequencies, morphometric data, molecular sequence data

Distance is the number or **estimate of the actual number** of evolutionary changes between two taxa along the lineages that separate them from their MRCA, sometimes referred to as ‘path length’ distance. For example, consider two homologous sequences . . .

<table>
<thead>
<tr>
<th>taxon 1:</th>
<th>ACGTGCTTTTAACGTCTCTC</th>
</tr>
</thead>
<tbody>
<tr>
<td>taxon 2:</td>
<td>ACGTGCTTTTGCGTCTTAC</td>
</tr>
</tbody>
</table>

The overall dissimilarity between them can be quantified as the proportion of sites with different bases, here, 4 of 20 for a $Q = 0.20$. $Q$ is one estimate of the true or actual number of evolutionary changes per site along the lineages that separate 1 and 2, which is what we call the **distance, $K$.** $Q$ is a **minimum estimate of $K,** it can be higher but cannot be lower. This bias can be ‘corrected’ through use of models of sequence evolution.
Distances and trees

**Metric** distances

In order for a distance measure to be used to reconstruct phylogenies, it must satisfy some basic requirements - it must be a **metric** and it must be **additive**. Let $d(A,B)$ be the distance between two sequences, $A$ and $B$. A distance $d$ is a metric if it satisfies these properties:

1. $d(A, B) > 0$ (non-negativity)
2. $d(A, B) = d(B, A)$ (symmetry)
3. $d(A, C) \leq d(A, B) + d(B, C)$ (triangle inequality)
4. $d(A, B) = 0$ if and only if $A = B$ (distinctiveness)

**Ultrametric** distances

A metric is an **ultrametric** if it satisfies the additional criterion that:

$$d(A, B) \leq \text{maximum } [d(A, C), d(B, C)]$$

This criterion implies that the two largest distances are equal. Ultrametric distances have the useful evolutionary property of implying a **constant rate of evolution**. The familiar ‘relative rate’ test for a molecular clock is really a test of how far the pairwise distances between 3 sequences depart from ultrametricity.

Ultrametric trees have the total branch length from the root up to the tip of each branch equal.
Distances and trees

Additive distances
Being a metric or ultrametric is a necessary, but not sufficient condition for being a valid measure of evolutionary change; a measure must also satisfy the four-point condition:

\[ d(A, B) + d(C, D) \leq \text{maximum } [d(A, C) + d(B, D), d(A, D) + d(B, C)] \]

This is equivalent to requiring that the distances between tips equal the sum of path distances connecting them, such that for the three sums of distances one of these must be less than or equal to the other two and the other two are equal to each other.

Tree distances
An additive distance measure defines a tree. Consider tree (((A, B), C), D) from Page and Holmes (1998; 27): If it is ultrametric, sequence D is equidistant from all other sequences, and C is equidistant from A and B.

If distances are not ultrametric, they can still represent a tree, an additive tree. Note that while sequences B and C are the most similar \([d(B, C) = 3]\) they are not most closely related.
Distance matrix methods

The fundamental idea of distance matrix methods is that we have a matrix of actual or observed distances ($D_{ij}$) from the sequences, and that any particular tree topology that has branch lengths leads to a predicted set of tree distances (denoted by $d_{ij}$). For real data, the observed and predicted tree distances rarely match. This leads to a discrepancy between the observed and the expected distances. One approach to resolving this problem is exemplified in least squares methods, which are some of the best justified methods statistically. The measure that is used to define this discrepancy in least squares methods is given by $Q$, where $Q$ is

$$Q = \prod_{i=1}^{n} \prod_{j=1}^{n} W_{ij} (D_{ij} - d_{ij})^2$$

and where the $W_{ij}$ are weights that differ between different least squares methods. We then search for the tree topology and it’s associated branch lengths that minimize the value of $Q$. For any given topology it is possible to solve for the branch lengths that minimize $Q$ using standard least squares methods.
Distance matrix methods

Branch lengths and time. Branch lengths are not simply a function of time, they reflect expected amounts of evolution in different branches of the tree. Two branches may reflect the same elapsed time (e.g., sister lineages in rooted tree) but they can have different expected amounts of evolution, resulting from different rates of evolution.

A tree and the distance matrix it predicts, generated by adding up the lengths of branches between each pair of species.

\[ d_{\text{A-D}} = v_1 + v_7 + v_4 \]

\[ v \text{ = variable used to describe amount of evolution along each branch} \]
Distance matrix methods

**Clustering methods** - these share the feature that they sequentially build up clusters from original set of taxa *without* searching through a space of possible trees trying to minimize some function as most optimization methods do, for this reason they can be extremely fast.

**UPGMA and related methods (WPGMA)**

**UPGMA = unweighted pair-group method using arithmetic averages**
- first step is calculation of a pairwise distance matrix (w/ or w/o model)
- clusters are formed sequentially beginning with closest pair of taxa
- depth of common ancestor node for the pair (A and B) is taken to be exactly 1/2 the pairwise distance
- forces this to be an ultrametric method \[d(A, \text{node}) = d(B, \text{node})\]
- new pairwise matrix is constructed in which first 2 taxa removed and replaced by one taxon whose distance to remaining taxa is constructed in similar way
- new distances are taken as either unweighted or weighted averages of the distances between original two taxa and all remaining taxa
- weights are based on number of taxa in the cluster
- because of assumption of ultrametricity, methods produce rooted trees
Distance matrix methods

Clustering starts with the pair of taxa with the smallest calculated distance, say A and B, and creates a new group (AB). A and B are connected to a new node such that the branches connecting A to (AB) and B to (AB) are of equal length \( d(A,B)/2 \). The composite taxon (AB) added to remaining taxa and distances between this new group (cluster) and all other groups (except A and B) are computed. This process is repeated until only one group remains in the data matrix.

This method takes about \( n^3 \) operations to infer a phylogeny with \( n \) taxa or species. Methods like UPGMA can be used to infer phylogenies if one can assume that evolutionary rates are the same in all lineages, that is, that they satisfy a "molecular clock". If trees are truly ultrametric, it is extremely simple to find the least squares branch lengths and reconstruct a tree.

The main disadvantage of UPGMA is that it can give seriously misleading results if the distances actually reflect a substantially non-clock-like tree.
Distance matrix methods

Neighbor Joining (NJ; Saitou and Nei, 1987)

- does not assume a molecular clock, approximates minimum evolution
  this approximation is quite good
- widely used due to speed and usually produces a single tree, requires
  mere additivity
- also builds clusters, but attempts to take rate variation among
  lineages into account
- begins with unrooted star phylogeny, then examines each possible
  two taxon clusters and calculates the total path length over the
  whole tree for each scenario using pairwise distances
- selects the pair that minimizes this total path length, replace them
  with a node and constructs new pairwise distance matrix
  without original two taxa and repeats until all taxa are joined
- NJ is ‘guaranteed’ to recover the true tree if the distance matrix
  happens to be an exact representation of the tree

NJ trees often used as starting trees that can be improved by searching
using other criteria, e.g., maximum likelihood.
Distance matrix methods

Advantages
- easiest to program, and computational speed guarantee popularity
- computational speed due to clustering rather than tree searching
- corrections for “multiple hits” (based on evolutionary assumptions) possible
- if correction for multiple hits is correct, methods are statistically consistent, and converge on “correct” tree with increasing amounts of data
- some data are only in the form of distances (e.g., DNA-DNA hybridization) thus, we have no other option but to use distance methods

Disadvantages
- **loss of information in conversion to distances**, leading to lowered efficiency or accuracy calculated branch lengths occasionally negative, raising logical dilemma having to do with evolutionary interpretability
- this becomes especially acute when variation in rates of evolution is large
- lack of direct logical correspondence between evolution and distance matrix (distances don’t evolve, characters do)
- cannot directly combine with other/new data sets w/o going back to original data matrices
Phylogenetic inference methods

Distances vs. discrete characters: distance methods first convert aligned sequences into a pairwise distance matrix, then input that matrix into a tree building method whereas discrete methods consider each nucleotide site directly. Consider the following example:

Trees obtained by parsimony and minimum evolution maybe identical in topology and branch lengths. However, parsimony gives us information about which site contributes to the length of each branch. Once sequences are converted into distances, that information is lost.

(after Page & Holmes 1998)
Maximum Likelihood methods

“Maximum likelihood methods of phylogenetic inference **evaluate a hypothesis** about evolutionary history in terms of the **probability** that a proposed **model** of the evolutionary process and the hypothesized **history** would give rise to the **observed data**.” (Swofford et al., 1996)

*evaluate a hypothesis* - monophyly of angiosperms?
*probability* - statistically, how likely is it?
*model* - for rates of sequence evolution and frequency of bases in sequences
*history* - i.e., given a particular tree
*observed data* - sequences of chloroplast *rbcL* genes
**Maximum Likelihood methods**

**A short history**

Maximum likelihood (ML) is one of the standard tools of statistics. ML was developed by R. A. Fisher in the early 1900s, and first used in phylogeny reconstruction by Cavalli-Sforza and Edwards (1967), but not on sequence data. Felsenstein (1981) brought the ML framework to molecular sequence-based phylogenetic inference. Only in last 10-15 years that ML was applied to amino acid sequence data (Kishino et al., 1990), and only in last 3-5 years has ML been applied to discrete morphological data (Lewis, 2001).

**A short definition**

Phylogenetic analysis seeks to infer the history(s) that are most consistent with a set of observed data. In our case, the data are observed nucleotide (or protein) sequences, and the unknowns are the branching order and branch lengths of the tree. A concrete model of the evolutionary process that accounts for the conversion of one sequence to another must be specified. A maximum likelihood approach to phylogenetic inference then evaluates the probability that the chosen model will have generated the observed data; and phylogenies are inferred by finding those trees that yield the highest likelihoods.
Maximum Likelihood methods

This is often written more formally as

\[ L = \Pr(D|H) \], the probability of the data \((D)\) given hypothesis \(H\)
or sometimes, the “likelihood of the hypothesis”

Note that this formula is not the probability of the hypothesis, whichwould be expressed as \(\Pr(H|D)\).

Likelihoods \((L)\) are often very small numbers, and are often expressed asnatural logarithms and referred to as log-likelihoods.

ML permits the inference of phylogenetic trees using complex evolutionarymodels. In addition to its statistical consistency, what makes ML soattractive for phylogenetic inferences is its power as a tool for testinghypotheses. ML provides the means not only for estimating modelparameters (from the data) but also for comparing competing models andtrees - and so make inferences simultaneously about the patterns andprocesses of evolution.
Maximum Likelihood methods

There are two main objectives in likelihood methods for phylogenetic inference:

(i) computing the likelihood of a tree (given a topology and parameters)

(ii) finding the maximum likelihood tree, with optimized branch lengths

There are different strategies for achieving the objectives mentioned above:

(i) simultaneous search for both tree with highest likelihood and optimized parameter values
done by heuristic search w/ parameters empirically derived or estimated from data matrix (often takes much longer)

(ii) begin with initial topology (provided by MP or NJ), heuristic search for tree with highest likelihood using parameters values estimated from the data using the input tree
To computing the likelihood of a tree, we start with a set of aligned DNA sequences (of $m$ sites), and are given a phylogeny with branch lengths and an evolutionary model that allows us to compute the probabilities of changes of states along the branches of that tree. The model allows us to compute transition probabilities $P_{ij}(t)$, i.e., the probability that state $j$ will exist at the end of a branch of length $t$ if the state at the beginning of the branch is $i$. Note that `$t$' measures branch length, not time.

This calculation assumes a Markov model, in which the probability of change from state $i$ to state $j$ at a given site does not depend on the history of the site prior to its possession of state $i$ - knowing what state the site possessed prior to state $i$ is irrelevant to the probability.

Two assumptions that are central to computing likelihoods need to be made:

1. evolution (nucleotide substitution) in different sites on a given tree is independent
2. evolution (nucleotide substitution) in different lineages is independent

These assumptions allow us to take the likelihood and ‘decompose’ it into a product of simpler terms and that makes the computation of probabilities more practicable.
Maximum Likelihood methods

Computing the likelihood of a tree, continued.

The first of these allows us to take the likelihood and decompose it into a product, one term for each site in the sequence:

\[
L = \text{Prob}(D|T) = \prod_{i=1}^{m} \text{Prob}(D^{(i)}|T) \tag{1}
\]

where \(D^{(i)}\) is the data at the \(i\)th site. This means we can compute the likelihood for each site separately, and combine the likelihoods from all sites into a total value at the end. The likelihood of a tree for one site is the sum (over all possible nucleotides that may have existed at the internal nodes of a tree) of the probabilities of all possible scenarios by which the tip sequences could have evolved, with each summation over all four nucleotides:

\[
\text{Prob} \left( D^{(i)} | T \right) = \sum_{x} \sum_{y} \sum_{z} \sum_{w} \text{Prob} \left( A, C, C, G, x, y, z, w | T \right) \tag{2}
\]
Maximum Likelihood methods

Computing the likelihood of a tree, continued.

The second of these allows us to decompose the probability on the right side of equation [2] into a product of terms. An example given a tree with branch lengths \( t \) and data at a single site.

\[
A, C, C, C, G = \text{observed character states} \\
x, y, z, w = \text{ancestral character states}
\]

\[
\text{Prob } (A, C, C, C, G, x, y, z, w | T) = \\
\text{Prob } (x) \cdot \text{Prob } (y|x, t_6) \cdot \text{Prob } (A|y, t_1) \cdot \text{Prob } (C|y, t_2) \\
\text{Prob } (z|x, t_8) \cdot \text{Prob } (C|z, t_3) \\
\text{Prob } (w|z, t_7) \cdot \text{Prob } (C|w, t_4) \cdot \text{Prob } (G|w, t_5) \quad [3]
\]

The probability of \( x \) is essentially taken to be the ‘equilibrium’ probability that, at a random point on an lineage, we would see base \( x \) (where \( x = A, C, G, \text{ or } T \)) under the particular model of base substitution we are using. The other probabilities are similarly derived from the same model.
Maximum Likelihood methods

The expression in equation [3] looks difficult to compute. The individual probabilities are not hard to compute (depending on which model of DNA evolution we are using), but the problem is that there are a great many terms in [3]. **For each site in a sequence, we would have to sum $4^4$ or 256 terms which does not sound difficult but the number of terms rises exponentially with the number of sequences!** On a tree with $N$ species, there are $n - 1$ internal nodes, and each can have one of 4 states (A, C, G, T), so we will need $4^{n-1}$ terms. For $N = 10$, there will be 262,144 terms to compute. For $N = 20$, there are 274,877,906,944 terms. (you get the picture . . . .)

Fortunately, Felsenstein has come to the rescue (Felsenstein 1973, 1981) - there is a method he calls ‘pruning’ that makes the whole computation economical. In this method, an algorithm that enables a flow of computations in a manner corresponding to the flow of information down a tree exists. It makes use of a quantity he calls the **conditional likelihood** of a subtree, $L_k^{(i)}(s)$. It is the probability of everything that is observed from node $k$ on the tree up to the tips, at site $i$, conditional on node $k$ having state $s$. 
Maximum Likelihood methods

In equation [3] above, the term

\[ \text{Prob (C|w, t_4) \cdot Prob (G|w, t_5)} \]

is one of these quantities, being the probability of everything seen at or above the node having base w in our example. There will be 4 such quantities, corresponding to the different values of w (A, C, G, T). Once these 4 probabilities have been computed, they need not continually be recomputed - this is the economy inherent in the pruning method.

This algorithm is applied starting at the node(s) that has all of its immediate descendants being tip taxa, and is then applied successively to nodes further down the tree until all nodes and descendants have been ‘processed’. Once the likelihood for each site is computed, the overall likelihood of the tree is the product of these.
Models of sequence evolution

Observed distances may underestimate the actual amount of evolutionary change. The extent of differences between two sequences is not linear with time (as might be expected if the rate of molecular evolution is approximately constant) but curvilinear due to ‘multiple hits’ at various sites. A number of methods exist, all with various assumptions about the nature of the molecular evolutionary process, whose goal is to ‘correct’ the observed distances by estimating the true evolutionary distance that has been ‘overprinted’ or superimposed by multiple hits. Most of the correction methods are interrelated, differing only in the number of parameters they include.
Models of DNA sequence evolution

Almost all DNA substitution models proposed to date are special cases of a general matrix.

Since it is usually assumed that the overall rate of change from base \( i \) to base \( j \) in a given length of time is the same as the rate of change from base \( j \) to base \( i \), such models are said to be time-reversible. This corresponds to rate parameter restrictions \( g=a, h=b, i=c, j=d, k=e, \) and \( l=f \).

The most general model is the general time-reversible model (GTR).

In this model, there are six different rate parameters (rate for A to C, A to G, A to T, C to G, C to T, and G to T) and unequal base frequencies (A does not = C does not = G does no = T) are assumed. Most of the remaining models commonly used either for estimation of pairwise evolutionary distances or maximum likelihood inference can be obtained by restricting the parameters in this matrix.
Models of DNA sequence evolution

Mathematical expression of a substitution model is a table of rates (substitutions per site per unit of evolutionary distance) at which each nucleotide is replaced by each alternative nucleotide. For DNA sequences, these rates can be expressed in 4 x 4 instantaneous rate matrix $Q$ in which each element $Q_{ij}$ represents the rate of change from base $i$ to base $j$ during some infinitesimal time period $dt$. The most general form of this matrix is

$$
Q = \begin{pmatrix}
-a_G + b_G + c_T & a_C & b_G & c_T \\
-g_A + d_G + e_T & -g_A + d_G & d_G & e_T \\
-h_A & j_C & -h_A + j_T & f_T \\
i_A & k_C & l_G & -i_A + k_C + l_G
\end{pmatrix}
$$

where the rows and columns correspond to the bases A, C, G, and T, respectively. Factor $\square$ represents the mean instantaneous substitution rate, which is modified by relative rate parameters $a, b, c, ... l$, that corresponds to each and every possible transformation from one base to another and back again (A to T and T to A). The product of $\square$ and relative rate parameter constitutes a rate parameter, while $\square_A, \square_C$, and so on, are frequency parameters that correspond to the frequencies of the bases.
Models of DNA sequence evolution

Almost all models of DNA substitution models proposed to date are special cases of the previous matrix. Since it is usually assumed that the overall rate of change from base $i$ to base $j$ in a given length of time is the same as the rate of change from base $j$ to base $i$, such models are said to be time-reversible. This corresponds to rate parameter restrictions $g=a$, $h=b$, $i=c$, $j=d$, $k=e$, and $l=f$. The most general time-reversible model (GTR) is then represented by

$$Q = \begin{pmatrix}
-a_c(a_c + b_c + c_t) & a_c & b_G & c_T \\
 a_A & -a_A(a_A + d_G + e_T) & d_G & e_T \\
 b_A & d_C & -a_A(a_A + d_G + e_T) & f_T \\
 c_A & e_C & f_G & -a_A(a_A + c_C + f_G)
\end{pmatrix}$$

In this model, there are six different rate parameters (A to C, A to G, A to T, C to G, C to T, and G to T) and unequal base frequencies are assumed. Most of the remaining models commonly used either for estimation of pairwise evolutionary distances or maximum likelihood inference can be obtained by restricting the parameters in this matrix.
Models of DNA sequence evolution

The Jukes-Cantor (1969) model, one of the first proposed and perhaps the simplest model of sequence evolution, can be represented by using the following substitution probability matrix and base composition vector:

\[
Q = \begin{bmatrix}
\frac{3}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
\frac{1}{4} & \frac{3}{4} & \frac{1}{4} & \frac{1}{4} \\
\frac{1}{4} & \frac{1}{4} & \frac{3}{4} & \frac{1}{4} \\
\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{3}{4}
\end{bmatrix}
\]

The quantity \( \square \) = rate of substitution, all substitutions occur at same rate between all pairs of bases, staying the same nucleotide is equal to \( 1 - 3/4 \square \). The equilibrium frequencies of all four bases are equal, \( p_A = p_G = p_C = p_T = 1/4 \).

The base frequency (all equal) and substitution rate are typically combined into a single parameter \( a = \square/4 \), leading to the simpler form of the matrix shown at left.
Models of DNA sequence evolution

Under the JC model, the distance between two sequences may be calculated by a simple formula

\[ d = -\frac{3}{4} \ln \left(1 - \frac{4}{3}p\right) \]  

[1]

where \( p \) is the proportion of nucleotides that are different in the two sequences. Felsenstein (2004; p. 156-158) approaches at this problem slightly differently, using a maximum likelihood estimation. The probability that a particular site will be different at the two ends of a branch is the sum of three equally likely events (change of an A to any of three other nucleotides),

\[ D_S = \frac{3}{4} \left(1 - e^{-4/3 \Delta t}\right) \]  

[2]

The value \( \Delta t \) is the product of the rate of substitution or change and the time, the value of \( \Delta t \) is the branch length. The difference per site is then used to estimate the branch length, and the resulting estimate is in effect the distance corrected for all events that are likely to have occurred, seen or not,

\[ D = -\frac{3}{4} \ln \left(1 - \frac{4}{3} D_S\right) \]  

[3]

Note that equations 1 and 3 are very similar.
Even though one may expect transversions to be more common than transitions, the reverse is more typically true of most real data. Kimura (1980) introduced a model that allows different rates of transitions and transversions, while still being symmetrical. The two rate parameters, $a$ and $b$, of the **Kimura 2-parameter model** (K2P) allows not only the overall rate of substitution per site per unit time to vary, but also the fraction of them that are transitions as opposed to transversions. So, what does this mean?

From any one nucleotide, there is one rate of change to transitions and a second that causes transversions, and $a$ does not equal $b$. The total rate of substitution per site will be $a + 2b$. The transition:transversion ratio $a/b$ is often represented by the letter kappa ($\kappa$); Note Felsenstein (2004) uses “$R$”.

Simplification of the K2P rate matrix is shown at right. Note how this rate matrix only differs from the JC model in having different rates for transitions and transversions. Of course, if $a = b$, then the K2P model becomes the JC model.

$$Q = \begin{bmatrix}
 a & 2b & 0 & 0 & 0 \\
 2b & a & 0 & 0 & 0 \\
 0 & 2b & a & 0 & 0 \\
 0 & 0 & 2b & a & 0 \\
 0 & 0 & 0 & 2b & a \\
\end{bmatrix}$$
Models of DNA sequence evolution

In simple models like JC, we compute the distance by expressing the fraction of difference between two sequences in terms of the distance, and then solving that equation for the distance. With the K2P model, which allows a transition/transversion inequality of rates, we now have two observations, the fraction $P$ of transition differences between the two sequences, and the fraction $Q$ of transversion differences. A simplified expression for calculating the number of substitutions per site is given by

$$d = \frac{1}{2} \ln \left[ \frac{1}{1-2P-Q} \right] + \frac{1}{4} \ln \left[ \frac{1}{1-2Q} \right]$$

where $P$ and $Q$ are the proportional differences between the two sequences due to transitions and transversions, respectively.
Models of DNA sequence evolution

Relationship between special cases of the GTR family of substitution models. Arrow labels indicate restrictions that convert from more general model to a more specific one*. More common models in boldface.

*After Swofford et al. (1996)
Models of DNA sequence evolution

Observed and expected patterns of nucleotide substitutions for three different models, an example from human and chimpanzee mtDNA sequences*. As the models add parameters they more closely approximate the actual observed pattern of rate and base frequency differences between the sequences. In this example, it is apparent that the observed frequencies of the bases are not equal (sizes along diagonal) and that transitions are more common than transversions (size and shading differences in other elements).

Observed:

```
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>C</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>G</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>T</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>
```

*After Page & Holmes (1998)