Extension until Monday (Oct 26) 16:00 for Assignment 2
Scientific paper review
Projects
Gotoh optimization for the affine gap function
Inferring phylogenies
  Motivation
  Distance-based methods
  Character-based methods
  Maximum likelihood methods
“Charles Darwin’s famous notebook B containing the first known sketch of an evolutionary tree.”

Prelude: Evidences of evolution
Prelude: Evidences of evolution (cont.)

Genetic Tool Kit
(www.pbs.org/wgbh/evolution/library/03/4/l_034_04.html)
Prelude: Evidences of evolution
The Common Genetic Code
(www.pbs.org/wgbh/evolution/library/04/4/l_044_02.html)
See also:

- Evolving Ideas: How Do We Know Evolution Happens? (www.pbs.org/wgbh/evolution/library/11/2/e_s_3.html)
- Evolution library (www.pbs.org/wgbh/evolution/library)
- Understanding Evolution - Misconceptions about evolution and the mechanisms of evolution
“The objectives of phylogenetic studies are (1) to reconstruct the correct genealogical ties between organisms and (2) to estimate the time of divergence between organisms since they last shared a common ancestor.”

“A phylogenetic tree is a graph composed of nodes and branches, in which only one branch connects any two adjacent nodes.”

“The nodes represents the taxonomic units, and the branches define the relationships among the units in terms of descent and ancestry.”

“The branch length usually represents the number of changes that have occurred in that branch.” (or some amount of time)

Terminology

- A **taxon** (plural **taxa**) or **taxonomic unit** is a species or grouping of species.
- Naming the different taxonomic levels: kingdom; phylum; class; order; family; genus; species.
A **rooted tree** for 5 species. Leaves, $\alpha, \beta, \gamma, \delta$ and $\epsilon$, correspond to **contemporary organisms**, for which data has been collected ($t = 0$). Internal nodes correspond to (inferred) **ancestors** ($t < 0$). **Newick format** of that tree: $(((\alpha, \beta), (\gamma, \delta)), \epsilon)$
Why?

- **Comparative studies**: understanding gene function, adaptation, correlating the appearance of a trait to environmental factors, etc.;
- **Drug design**: designing compounds that are specific to a group of organisms;
- **Bioinformatics**: multiple sequence alignment, protein (secondary) structure prediction, etc.;
“One way of testing such hypotheses is to resurrect the ancestral proteins and study their behavior in the laboratory. To do this, a DNA molecule encoding the ancestral protein is synthesized and expressed in an appropriate host. The ancient protein is then recovered and studied to determine whether its properties are consistent with its inferred ancestral role. (…) digestive ribonuclease emerged near the time when ruminant digestion emerged, in animals in which ruminant digestion developed, at a time where difficult-to-digest grasses emerged, permitting their descendants to exploit a newly available resource emerging at a time of global climatic upheaval.”
The Great Apes

Phylogeny

From the Tree of the Life Website, University of Arizona

Orangutan  Gorilla  Chimpanzee  Human
Example: Antivenins

Venomous Australian Snakes

- Black whip snake
- Taipan
- Fierce snake
- Common brown
- Western brown
- Dugte
- Collatt’s snake
- Spotted Black
- Butler’s snake
- King brown
- Red-bellied black
- Death adder
- Barclick
- Small-eyed snake
- Australian copperhead
- Tiger snake
- Rough-scaled snake
- Broad-headed snake
Scale of The Tree of Life

- 1,5 million described species.
- 10 million to 200 million existing species.
- Reconstruction tools can handle around 500 organisms.
- Reconstruction tools scale exponentially with the amount of data.
Bernard Moret

1980–2006: University of New Mexico
(www.cs.unm.edu/~moret)

2006–: École Polytechnique Fédérale de Lausanne
(people.epfl.ch/bernard.moret)

See also: Tree of Life Web Project (www.tolweb.org).
Summary of the applications

- Study ancestor-descendant relationships
  (Evolutionary biology, adaption, genetic drift, selection, speciation, etc.)
- Paleogenomics: inferring ancestral genomic information from extinct species
  (Comparing Chimpanzee, Neanderthal and Human DNA)
- Origins of epidemics
  (Comparing, at the molecular level, various virus strains)
- Drug design: specifically targeting groups of organisms
  (Efficient enumeration of phylogenetically informative substrings)
- Forensic
  (Relationships among HIV strains)
- Linguistics
  (Languages tree divergence times)

- “Phylogenies, or evolutionary trees, are the basic structures necessary to think clearly about differences between species, and to analyze those differences statistically.”
- “I estimated that there are about 3,000 papers on methods for inferring phylogenies.”
- “The field of inferring phylogenies has been wracked by outrageously excessive controversy (…) there have been many biologists who strove to bring back the field to normality (…).”

Joe Felsenstein is the author of a software package called PHYLIP, which is one the most widely used software system for phylogenetic studies.
Here, the 0s and 1s are indicating the presence or absence of a character (has feathers?, lays eggs?, curved beak?, flies?, ...). A **character** is a measurable feature having well-defined mutually exclusive states.
Based on anatomical and behavioural characters, the panda was classified as a raccoon (1870). Recently, 1985, the panda was re-classified as a bear when an analysis based on molecular data was done.

⇒ Images from www.wikipedia.org
What’s the data? (1/2)

**Platypus** (*Ornithorhynchus anatinus*) “the only **mammals that lay eggs** instead of giving birth to live young”, “[i]t is the sole representative of its family (Ornithorhynchidae) and genus (Ornithorhynchus)”, “[t]he platypus is considered to be one of the strangest specimens of the animal kingdom: a **venomous, egg-laying, duck-billed mammal**”.

(commons.wikimedia.org/wiki/Image:Ornithorhynchidae-00.jpg)

The thylacine (*Thylacinus cynocephalus*) is a now extinct (wolf-like) carnivorous marsupial.
Hard to resolve relationships
using morphology and behaviour alone

1. Similar characteristics can evolve independently in distantly related organisms — **convergent-evolution**;
2. It is often difficult to find characteristics that are common to all the organisms under study.
Nowadays, biologists rely on molecular sequence data, in particular DNA or RNA sequences, which allows the comparison of a broader range of species. **What characters, other than molecular sequence data, would allow to compare E. coli, yeast, clam shell and human?**
Genes or species trees

- Herein, a molecular sequence alignment (DNA, RNA or proteins) is used as input. Each column (site) of this alignment represents a character.
- The taxonomic units (nodes of the tree) can represent genes, species or populations; not all at once obviously.
- A gene tree represents the evolutionary history of a single gene; e.g. the evolution of the globin family (with its numerous gene duplication events).
- A species tree will generally be built using multiple genes (say 100).
- Since our main interest is to study the methods, we will limit our discussion to species trees.
A **rooted tree** not only gives the relationships between the taxonomic units, it also indicates the direction of evolution (time). Such trees can be scaled or unscaled.

![Rooted Tree Diagram]
An unrooted tree specifies the relationships between species.
Biologists generally prefer rooted trees.

- Under the **molecular clock assumption**, the root of the tree would be located at equal distance from all the leaves (contemporary organisms);

- The **outgroup method** consists of including into the analysis an organism that is known to have branched off earlier than the taxa under study (for which paleontological evidences exist, for instance), the root will be placed along the edge connecting the outgroup to the ancestor of the ingroup (taxa under study).
Molecular clock theory

- **Accepted mutations occur at a constant rate.**
- The number of accepted mutations is proportional to the length of the time interval.
- Once the “clock” has been calibrated (using fossil evidences, for instance) the unknown length of some time interval can be deduced from the number of accepted mutations.
- **Note:** different proteins have different clocks (hemoglobin ticks faster than cytochrome c).
- “A great deal of ink (and blood) has been spilt over the molecular clock (…).”

⇒ [3, pages 453–455]
Rooting a tree: molecular clock

Marcel Turcotte (turcotte@site.uottawa.ca)  CSI 5126. Algorithms in bioinformatics
Rooting a tree: outgroup

Marcel Turcotte (turcotte@site.uottawa.ca)  CSI 5126. Algorithms in bioinformatics
Hypothetically, E is the outgroup — e.g. chimpanzee while the ingroup consists of human populations.
Rooting a tree: outgroup (cont.)

Neighbour-joining phylogram based on complete mtDNA genome sequences.
It can be shown that the number of rooted and unrooted trees for a given \( n \) (number of species) are as follows.

\[
\begin{align*}
N_{\text{rooted}}(n) &= \frac{(2n - 3)!}{2^{n-2}(n - 2)!} \\
N_{\text{unrooted}}(n) &= \frac{(2n - 5)!}{2^{n-3}(n - 3)!}
\end{align*}
\]
General paradigm

1. Enumerate trees;
2. Select the “best” tree.
Sequential addition strategy

Given three species, there is a single unrooted tree.
Each branch can serve as an insertion point, adding a new branch off the middle of any existing branch.
Sequential addition strategy

Therefore producing 3 four species unrooted trees.
Sequential addition strategy

The same process is applied to all 3 four species trees.

Marcel Turcotte (turcotte@site.uottawa.ca)  CSI 5126. Algorithms in bioinformatics
A four species unrooted tree has 5 edges, thus leading to 5 new unrooted trees.
There will be 15 five species unrooted trees.
Sequential addition strategy
Methods

- **Distance-based**
  a distance is a measure of the overall differences/similarities between two objects

- **Character-based**
  a character is a characteristic that has well-defined, limited number of states

- **Maximum likelihood**
  Finds a tree such that the likelihood of the data given the tree structure is maximum

“While disputes between the champions of the two approaches [character-based and distance-based] have often been surprisingly intense, it is fair to say that both approaches are widely used and work well with most data sets.” [5, page 85]
I. Distance-based reconstruction

- Let $D_{ij}$ be the pairwise distance between two species; for instance, measured by comparing, in some ways, sequence data from the two species.

- Let $d_{ij}$ be the distance between $i$ and $j$ in some tree; the sum of the length of all the branches on the unique path between $i$ and $j$.

- Distance-based methods seek to find a tree (topology + branch length) such the $D_{ij}$ and the $d_{ij}$, for all $i$ and $j$, are in good agreement.

- For instance, find a tree minimizing

$$Q = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij}(D_{ij} - d_{ij})^2$$

Minimum least squares approach.
I. Distance-based reconstruction

If the tree topology was given, the problem would simply be to find the length of the branches minimizing Equation 1, which is a quadratic function.

The $d_{ij}$ need to be expressed as the sum of the length of the branches, $x$, on the unique path between $i$ and $j$, $\mathcal{P}_{ij}$.

$$Q = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij}(D_{ij} - \sum_{x \in \mathcal{P}_{ij}} x)^2$$

and

$$d_{ij} = \sum_{x \in \mathcal{P}_{ij}} x$$
I. Distance-based reconstruction

\[ d_{AE} = x + y + z \]
I. Distance-based reconstruction

This involves solving the set of linear equations obtained by taking the derivative of $Q$ with respect to the branch lengths and equating those to 0.

$$\frac{dQ}{dx} = -2 \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (D_{ij} - \sum_{x \in P_{ij}} x) = 0$$

Methods for finding exact solutions for the weighted and unweighted least squares branch length equations have been proposed that run in polynomial time (order 3 or less), and iterative/heuristic methods have also been proposed, which in practice converge rapidly towards the correct lengths.
I. Distance-based reconstruction

▶ What’s next?
▶ Well, we assumed that the topology was known!
▶ In principle, it would be possible to enumerate all tree topologies, solve the branch length equations for each one, and select the topology that minimizes the least squares.
▶ This is not practical since the number of topologies grows rapidly w.r.t. the number of species.
[ The same approaches as for parsimony methods for enumerating tree topologies can be applied. However, under certain assumptions, a simple algorithm, UPGMA, performs adequately. ]
Given $n$ species and $m$ characters.

1. Compute a distance matrix between all ($\Theta(n^2)$) pairs, this matrix is symmetrical;
2. Generate the topology of the tree;
3. Estimate the length of the branches.
For now, let’s assume a simple distance measure, e.g. **Hamming distance** or **fractional alignment difference** \( \frac{D}{L} \) where \( D \) is the number of sites that differ (excluding indel containing sites), \( L \) is the number of sites.

More realistic models will be presented together with the Maximum Likelihood approach.
### Distance-based reconstruction

<table>
<thead>
<tr>
<th></th>
<th>alpha</th>
<th>beta</th>
<th>gamma</th>
<th>delta</th>
<th>epsilon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GTGCTGCACGG CTCAGTATA GCATTTACCC TTCCATCTTC AGATCCTGAA</td>
<td>ACGCTGCACGG CTCAGTGCG GTGCTTACCC TCCCCATCTTC AGATCCTGAA</td>
<td>GTGCTGCACGG CTCGCGCGCA GCATTTACCC TCCCCATCTTC AGATCCTATC</td>
<td>GTATCACACGA CTCAGCGCA GCATTTGCCC TCCCCGTCTTC AGATCCTATC</td>
<td>GTATCACATAG CTCAGCGCA GCATTTGCCC TCCCCGTCTTC AGATCCTAAAA</td>
</tr>
</tbody>
</table>

Build a multiple sequence alignment and **select columns**.

⇒ [5, page 87]
Distance-based reconstruction

<table>
<thead>
<tr>
<th>Species</th>
<th>α</th>
<th>β</th>
<th>γ</th>
<th>δ</th>
<th>ε</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>0</td>
<td>9</td>
<td>8</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>β</td>
<td>9</td>
<td>0</td>
<td>11</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>γ</td>
<td>8</td>
<td>11</td>
<td>0</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>δ</td>
<td>12</td>
<td>15</td>
<td>10</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>ε</td>
<td>15</td>
<td>18</td>
<td>13</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

The above matrix has been filled by computing the Hamming distance for all pairs of sequences.

⇒ [5, page 87]
UPGMA = Unweighted Pair Group Method using Arithmetic averages\(^2\).

\[
\{ \text{Initialization} \}
\]
Assign each species \(i\) to its own cluster \(C_i\)

Define one leaf of \(T\) for each species, place it at height zero.

\[
\{ \text{Iterations} \}
\]
Find the pair of clusters \(i\) and \(j\) which minimises \(d_{ij}\).

Define a new cluster \(C_k = C_i \cup C_j\).

Calculate \(d_{kl}\) for all \(l\).

Create the parent node \(k\) of \(i\) and \(j\)
at height \(d_{ij}/2\) in \(T\).

Add \(k\) to the current list of clusters
and remove \(i\) and \(j\).

\[
\{ \text{Termination} \}
\]
Stop when the list of clusters contains only one entry.

\[\Rightarrow \text{Early 1960s, simple and intuitive.}\]

\(^2\)See [1, page 166]
Distance-based reconstruction

<table>
<thead>
<tr>
<th>Species</th>
<th>α</th>
<th>β</th>
<th>γ</th>
<th>δ</th>
<th>ε</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>0</td>
<td>9</td>
<td>8</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>β</td>
<td>9</td>
<td>0</td>
<td>11</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>γ</td>
<td>8</td>
<td>11</td>
<td>0</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>δ</td>
<td>12</td>
<td>15</td>
<td>10</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>ε</td>
<td>15</td>
<td>18</td>
<td>13</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

⇒ Assign each species to its own cluster, place it at height 0.
Distance-based reconstruction

<table>
<thead>
<tr>
<th>Species</th>
<th>α</th>
<th>β</th>
<th>γ</th>
<th>δ</th>
<th>ε</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>0</td>
<td>9</td>
<td>8</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>β</td>
<td>9</td>
<td>0</td>
<td>11</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>γ</td>
<td>8</td>
<td>11</td>
<td>0</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>δ</td>
<td>12</td>
<td>15</td>
<td>10</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>ε</td>
<td>15</td>
<td>18</td>
<td>13</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

⇒ Find the pair of clusters which minimises $d_{i,j}$, define a new cluster $C_k = C_i \cup C_j$, create the parent node $k$ at height $d_{i,j}/2$. 
Distance-based reconstruction

<table>
<thead>
<tr>
<th>Species</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
<th>$(\delta, \epsilon)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>0</td>
<td>9</td>
<td>8</td>
<td>13.5</td>
</tr>
<tr>
<td>$\beta$</td>
<td>9</td>
<td>0</td>
<td>11</td>
<td>16.5</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>8</td>
<td>11</td>
<td>0</td>
<td>11.5</td>
</tr>
<tr>
<td>$(\delta, \epsilon)$</td>
<td>13.5</td>
<td>16.5</td>
<td>11.5</td>
<td>0</td>
</tr>
</tbody>
</table>

$\Rightarrow$ Calculate $d_{k,l}$ for all $l$, where $d_{i,j} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}$. 
Distance-based reconstruction

<table>
<thead>
<tr>
<th>Species</th>
<th>α</th>
<th>β</th>
<th>γ</th>
<th>(δ, ϵ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>0</td>
<td>9</td>
<td>8</td>
<td>13.5</td>
</tr>
<tr>
<td>β</td>
<td>9</td>
<td>0</td>
<td>11</td>
<td>16.5</td>
</tr>
<tr>
<td>γ</td>
<td>8</td>
<td>11</td>
<td>0</td>
<td>11.5</td>
</tr>
<tr>
<td>(δ, ϵ)</td>
<td>13.5</td>
<td>16.5</td>
<td>11.5</td>
<td>0</td>
</tr>
</tbody>
</table>

⇒ Find the pair of clusters which minimises $d_{i,j}$, define a new cluster $C_k = C_i \cup C_j$, create the parent node $k$ at height $d_{i,j}/2$. 
Distance-based reconstruction

<table>
<thead>
<tr>
<th>Species</th>
<th>$(\alpha, \gamma)$</th>
<th>$\beta$</th>
<th>$(\delta, \epsilon)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(\alpha, \gamma)$</td>
<td>0</td>
<td>10</td>
<td>12.5</td>
</tr>
<tr>
<td>$\beta$</td>
<td>10</td>
<td>0</td>
<td>16.5</td>
</tr>
<tr>
<td>$(\delta, \epsilon)$</td>
<td>12.5</td>
<td>16.5</td>
<td>0</td>
</tr>
</tbody>
</table>

⇒ Calculate $d_{k,l}$ for all $l$, where $d_{i,j} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}$.
Distance-based reconstruction

<table>
<thead>
<tr>
<th>Species</th>
<th>$(\alpha, \gamma)$</th>
<th>$\beta$</th>
<th>$(\delta, \epsilon)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(\alpha, \gamma)$</td>
<td>0</td>
<td>10</td>
<td>12.5</td>
</tr>
<tr>
<td>$\beta$</td>
<td>10</td>
<td>0</td>
<td>16.5</td>
</tr>
<tr>
<td>$(\delta, \epsilon)$</td>
<td>12.5</td>
<td>16.5</td>
<td>0</td>
</tr>
</tbody>
</table>

⇒ Find the pair of clusters which minimises $d_{i,j}$, define a new cluster $C_k = C_i \cup C_j$, create the parent node $k$ at height $d_{i,j}/2$. 
Distance-based reconstruction

<table>
<thead>
<tr>
<th>Species</th>
<th>((\alpha, \gamma), \beta)</th>
<th>((\delta, \epsilon))</th>
</tr>
</thead>
<tbody>
<tr>
<td>((\alpha, \gamma), \beta)</td>
<td>0</td>
<td>13.83</td>
</tr>
<tr>
<td>((\delta, \epsilon))</td>
<td>13.83</td>
<td>0</td>
</tr>
</tbody>
</table>

⇒ Calculate \(d_{k,l}\) for all \(l\), where

\[
d_{i,j} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}.\]
Distance-based reconstruction

<table>
<thead>
<tr>
<th>Species</th>
<th>((\alpha, \gamma), \beta)</th>
<th>((\delta, \epsilon))</th>
</tr>
</thead>
<tbody>
<tr>
<td>((\alpha, \gamma), \beta)</td>
<td>0</td>
<td>13.83</td>
</tr>
<tr>
<td>((\delta, \epsilon))</td>
<td>13.83</td>
<td>0</td>
</tr>
</tbody>
</table>

⇒ Find the pair of clusters which minimises \(d_{i,j}\), define a new cluster \(C_k = C_i \cup C_j\), create the parent node \(k\) at height \(d_{i,j}/2\).
Distance-based reconstruction

<table>
<thead>
<tr>
<th>Species</th>
<th>(((α, γ), β), (δ, ε))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(((α, γ), β), (δ, ε))</td>
<td>0</td>
</tr>
</tbody>
</table>

⇒ UPGMA produces **ultrametric** trees, a tree such that the distance from any internal node (including the root) to its descendant leaves is the same. Thus, UPGMA assumes that evolution proceeds at the same rate in all the lineages, this is called the **molecular clock hypothesis**. (An assumption that is often violated)
Distance-based reconstruction

<table>
<thead>
<tr>
<th>Species</th>
<th>α</th>
<th>β</th>
<th>γ</th>
<th>δ</th>
<th>ε</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>0</td>
<td>9</td>
<td>8</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>β</td>
<td>9</td>
<td>0</td>
<td>11</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>γ</td>
<td>8</td>
<td>11</td>
<td>0</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>δ</td>
<td>12</td>
<td>15</td>
<td>10</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>ε</td>
<td>15</td>
<td>18</td>
<td>13</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

Consider $d_T(\alpha, \beta)$ and $d_{\alpha,\beta}$. Is $d_T(i, j) = d_{i,j}$ for all $i, j$?
UPGMA — failure of the molecular clock hypothesis

“True tree”

Felsenstein 2004, page 167
UPGMA — failure of the molecular clock hypothesis

```
A
  B  C
    4

D
  2
```

```
13
  2

10
  2
```

$\Rightarrow$

```
0  17  21  27
17  0  12  18
21 12  0  14
27 18 14  0
```

“True tree”

Felsenstein 2004, page 167
UPGMA — failure of the molecular clock hypothesis

Reconstructed tree

Felsenstein 2004, page 167
UPGMA — failure of the molecular clock hypothesis

UPGMA joins B and C together since both are evolving slowly (short branch attraction).
Distance-based tree reconstruction

- Given a tree $T$, a distance matrix $d_{i,j}$ is additive if $d_T(i,j) = d_{i,j}$, and nonadditive otherwise;
- UPGMA has no guarantee to produce “additive” trees;
- UPGMA produces ultrametric trees, i.e. a tree such that the distance from the root to any leaf is the same;
- UPGMA assumes that evolution proceeds at a fixed constant rate, the so called molecular clock hypothesis;
- Other distance-based methods, such as the neighbour-joining algorithm, do not assume the existence of a molecular clock; good approximation of the least squares methods, fast (works well with hundreds of species).
Distance-based methods are taking as input a multiple sequence alignment ($n$ sequences by $m$ columns). A distance is calculated for each pairwise alignment in order to fill an $n \times n$ (distance) matrix. The distance matrix is used to infer the topology of a tree, as well as the length of its branches.

**Neighbour-joining** is the most widely used distance-based approach. It produces un-rooted, additive (but not necessarily ultrametric) trees. It is an iterative algorithm that requires $O(n^3)$ time.

Distance-based methods are fast!
“However, the real problems arise when the data are not additive. Then, one has to find a tree whose distances best approximate the given data.”

“(…) the problem [distance-based tree reconstruction] remains open as there is no approach that both leads to a provably efficient algorithm and that follows a completely accepted definition of a good approximation.” [3, page 448]

Information is lost when reducing a pairwise alignment to a single number! In particular, the reconstructed tree says nothing about the “characters” of the internal nodes (ancestors) and the evolutionary events that occurred along its branches.

Hence the need for alternatives.
II. Character-based tree reconstruction

Character-based reconstruction algorithms are labelling all the nodes of the tree with characters. **Leaves** are labelled with observed data. While the **internal nodes** are labelled with hypothetical characters (ancestral states).
First, let’s consider a single character (the $i$th nucleotide of a given gene in 5 species). The only observable characters are those at the leaves. Those correspond to the characters in today’s organisms.
Several reconstructions of the ancestral states are possible.

How many events are represented on this tree?
The tree represents 4 events.
Can you find a reconstruction that requires fewer events?
A 3 events tree.
Now considering 3 characters (sites).

<table>
<thead>
<tr>
<th>Species</th>
<th>Sites 1</th>
<th>Sites 2</th>
<th>Sites 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>D</td>
<td>A</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>E</td>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>
A tree for five species and three characters. The reconstruction involves 5 mutations (evolutionary events).
Parsimony

- “Adoption of the simplest assumption in the formulation of a theory or in the interpretation of data, especially in accordance with the rule of Ockham’s razor.” The American Heritage [Online] Dictionary

- Ockham’s Razor: “Plurality should not be posited without necessity.”

- “(1) Mutations are exceedingly rare events and (2) the more unlikely events a model invokes, the less likely the model is to be correct. As a result, the relationship that requires the fewest number of mutations to explain the current state of the sequences being considered is the relationship that is most likely to be correct.” [5, page 98]
Issues

- Reconstructing the ancestral states;
- Counting the number of changes;
- Find all most parsimonious trees;
- Infer branch lengths;
- Is the most parsimonious tree the “real one”?;
- Given several most parsimonious trees, is there a better one?
Small parsimony problem

Problem: Find the most parsimonious labelling of the internal vertices in a given evolutionary tree.

Input: A tree $T$ with each leaf labelled by an $m$-character array.

Output: Labels ($m$-character arrays) for all the internal nodes such that $\sum d_H(u, v)$ for all the edges $(u, v)$ is minimum; $d_H$ is the Hamming distance.
Notice that the characters are independent. The total number of changes is the total number of changes for the first character plus the total number of changes for the second character plus the total number of changes for the third character. Thus, it suffices to develop a method that works for a single character and to apply it to all the characters. Proposals?
Let's define $s_c(u)$ as the minimum parsimony score obtained when $u$ is labelled with $c$. How to compute $s_c(u)$? What do you need to know? What are the dependencies?

$$s_c(u) = \ldots$$
Small parsimony problem

For instance, what would be the most parsimonious score if $u$ was labelled with $A$.

$s_A(v) + ?$  
$s_C(v) + ?$  
$s_G(v) + ?$  
$s_T(v) + ?$

$s_A(w) + ?$  
$s_C(w) + ?$  
$s_G(w) + ?$  
$s_T(w) + ?$
Small parsimony problem

For instance, what would be the most parsimonious score if $u$ was labelled with $A$.

$$
\begin{align*}
  s_A(v) + 0 & \quad s_A(w) + 0 \\
  s_C(v) + 1 & \quad s_C(w) + 1 \\
  s_G(v) + 1 & \quad s_G(w) + 1 \\
  s_T(v) + 1 & \quad s_T(w) + 1
\end{align*}
$$
Small parsimony problem

\[ s_A(u) = \min \left\{ \begin{array}{l} s_A(v) + 0 \\ s_C(v) + 1 \\ s_G(v) + 1 \\ s_T(v) + 1 \end{array} \right\} + \min \left\{ \begin{array}{l} s_A(w) + 0 \\ s_C(w) + 1 \\ s_G(w) + 1 \\ s_T(w) + 1 \end{array} \right\} \]
Weighted small parsimony problem (Sankoff 1975)

\[ s_c(u) = \min_i \{ s_i(v) + \delta_{i,c} \} + \min_j \{ s_j(w) + \delta_{j,c} \} \]

where \( \delta_{j,c} \) is a \( k \times k \) scoring matrix.
**Examples of scoring matrices**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>G</strong></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>T</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td>0</td>
<td>1</td>
<td>0.33</td>
<td>1</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.33</td>
</tr>
<tr>
<td><strong>G</strong></td>
<td>0.33</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>T</strong></td>
<td>1</td>
<td>0.33</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Solving the small parsimony problem

General case.

\[ s_c(u) = \min_i \{ s_i(v) + \delta_{i,c} \} + \min_j \{ s_j(w) + \delta_{j,c} \} \]
**Initialisation.**
For each leaf, \( s_c(v) = 0 \) if character \( c \) is found at that node and infinity otherwise.
Small parsimony problem
Small parsimony problem

Marcel Turcotte (turcotte@site.uottawa.ca)  CSI 5126. Algorithms in bioinformatics
\begin{align*}
S_A(w) &= 2 = \min\{\infty + 0, 0 + 1, \infty + 1, \infty + 1\} + \min\{\infty + 0, 0 + 1, \infty + 1, \infty + 1\} \\
S_C(w) &= 0 = \min\{\infty + 1, 0 + 0, \infty + 1, \infty + 1\} + \min\{\infty + 1, 0 + 0, \infty + 1, \infty + 1\} \\
S_G(w) &= 2 = \min\{\infty + 1, 0 + 1, \infty + 0, \infty + 1\} + \min\{\infty + 1, 0 + 1, \infty + 0, \infty + 1\} \\
S_T(w) &= 2 = \min\{\infty + 1, 0 + 1, \infty + 1, \infty + 0\} + \min\{\infty + 1, 0 + 1, \infty + 1, \infty + 0\}
\end{align*}

\begin{align*}
S_A(x) &= 1 = \min\{0 + 0, \infty + 1, \infty + 1, \infty + 1\} + \min\{\infty + 0, 0 + 1, \infty + 1, \infty + 1\} \\
S_C(x) &= 1 = \min\{0 + 1, \infty + 0, \infty + 1, \infty + 1\} + \min\{\infty + 1, 0 + 0, \infty + 1, \infty + 1\} \\
S_G(x) &= 2 = \min\{0 + 1, \infty + 1, \infty + 0, \infty + 1\} + \min\{\infty + 1, 0 + 1, \infty + 0, \infty + 1\} \\
S_T(x) &= 2 = \min\{0 + 1, \infty + 1, \infty + 1, \infty + 0\} + \min\{\infty + 1, 0 + 1, \infty + 1, \infty + 0\}
\end{align*}

\begin{align*}
S_A(v) &= 2 = \min\{2 + 0, 0 + 1, 2 + 1, 2 + 1\} + \min\{\infty + 0, \infty + 1, 0 + 1, \infty + 1\} \\
S_C(v) &= 1 = \min\{2 + 1, 0 + 0, 2 + 1, 2 + 1\} + \min\{\infty + 1, \infty + 0, 0 + 1, \infty + 1\} \\
S_G(v) &= 1 = \min\{2 + 1, 0 + 1, 2 + 0, 2 + 1\} + \min\{\infty + 1, \infty + 1, 0 + 0, \infty + 1\} \\
S_T(v) &= 2 = \min\{2 + 1, 0 + 1, 2 + 1, 2 + 0\} + \min\{\infty + 1, \infty + 1, 0 + 1, \infty + 0\}
\end{align*}

\begin{align*}
S_A(u) &= 3 = \min\{2 + 0, 1 + 1, 1 + 1, 2 + 1\} + \min\{1 + 0, 1 + 1, 2 + 1, 2 + 1\} \\
S_C(u) &= 2 = \min\{2 + 1, 1 + 0, 1 + 1, 2 + 1\} + \min\{1 + 1, 1 + 0, 2 + 1, 2 + 1\} \\
S_G(u) &= 3 = \min\{2 + 1, 1 + 1, 1 + 0, 2 + 1\} + \min\{1 + 1, 1 + 1, 2 + 0, 2 + 1\} \\
S_T(u) &= 4 = \min\{2 + 1, 1 + 1, 1 + 1, 2 + 0\} + \min\{1 + 1, 1 + 1, 2 + 1, 2 + 0\}
\end{align*}
Small parsimony problem

Marcel Turcotte (turcotte@site.uottawa.ca)  CSI 5126. Algorithms in bioinformatics
Small parsimony problem
\[
S_A(w) = 1 = \min\{\infty + 0, \infty + 1, 0 + 1, \infty + 1\} + \min\{0 + 0, \infty + 1, \infty + 1, \infty + 1\}
\]
\[
S_C(w) = 2 = \min\{\infty + 1, \infty + 0, 0 + 1, \infty + 1\} + \min\{0 + 1, \infty + 0, \infty + 1, \infty + 1\}
\]
\[
S_G(w) = 1 = \min\{\infty + 1, \infty + 1, 0 + 0, \infty + 1\} + \min\{0 + 1, \infty + 1, \infty + 0, \infty + 1\}
\]
\[
S_T(w) = 2 = \min\{\infty + 1, \infty + 1, 0 + 1, \infty + 0\} + \min\{0 + 1, \infty + 1, \infty + 1, \infty + 0\}
\]
\[
S_A(x) = 1 = \min\{0 + 0, \infty + 1, \infty + 1, \infty + 1\} + \min\{\infty + 0, 0 + 1, \infty + 1, \infty + 1\}
\]
\[
S_C(x) = 1 = \min\{0 + 1, \infty + 0, \infty + 1, \infty + 1\} + \min\{\infty + 1, 0 + 0, \infty + 1, \infty + 1\}
\]
\[
S_G(x) = 2 = \min\{0 + 1, \infty + 1, \infty + 0, \infty + 1\} + \min\{\infty + 1, 0 + 1, \infty + 0, \infty + 1\}
\]
\[
S_T(x) = 2 = \min\{0 + 1, \infty + 1, \infty + 1, \infty + 0\} + \min\{\infty + 1, 0 + 1, \infty + 1, \infty + 0\}
\]
\[
S_A(v) = 2 = \min\{1 + 0, 2 + 1, 1 + 1, 2 + 1\} + \min\{\infty + 0, 0 + 1, \infty + 1, \infty + 1\}
\]
\[
S_C(v) = 2 = \min\{1 + 1, 2 + 0, 1 + 1, 2 + 1\} + \min\{\infty + 1, 0 + 0, \infty + 1, \infty + 1\}
\]
\[
S_G(v) = 2 = \min\{1 + 1, 2 + 1, 1 + 0, 2 + 1\} + \min\{\infty + 1, 0 + 1, \infty + 0, \infty + 1\}
\]
\[
S_T(v) = 3 = \min\{1 + 1, 2 + 1, 1 + 1, 2 + 0\} + \min\{\infty + 1, 0 + 1, \infty + 1, \infty + 0\}
\]
\[
S_A(u) = 3 = \min\{2 + 0, 2 + 1, 2 + 1, 3 + 1\} + \min\{1 + 0, 1 + 1, 2 + 1, 2 + 1\}
\]
\[
S_C(u) = 3 = \min\{2 + 1, 2 + 0, 2 + 1, 3 + 1\} + \min\{1 + 1, 1 + 0, 2 + 1, 2 + 1\}
\]
\[
S_G(u) = 4 = \min\{2 + 1, 2 + 1, 2 + 0, 3 + 1\} + \min\{1 + 1, 1 + 1, 2 + 0, 2 + 1\}
\]
\[
S_T(u) = 5 = \min\{2 + 1, 2 + 1, 2 + 1, 3 + 0\} + \min\{1 + 1, 1 + 1, 2 + 1, 2 + 0\}
\]
Problem: Find a tree having the minimum parsimony score.

Input: An $n \times m$ matrix (alignment).

Output: A tree $T$ with $n$ leaves labeled by the $n$ rows ($m$ characters) of the input matrix. The internal nodes are labelled with arrays of $m$ characters such that the overall parsimony score is minimum.

The problem is known to be $\mathcal{NP}$-complete.
Exhaustive approach: 4 to 15 sequences

<table>
<thead>
<tr>
<th># Species</th>
<th># unrooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
</tr>
<tr>
<td>8</td>
<td>10,395</td>
</tr>
<tr>
<td>9</td>
<td>1,351,135</td>
</tr>
<tr>
<td>10</td>
<td>2,027,025</td>
</tr>
<tr>
<td>11</td>
<td>34,459,425</td>
</tr>
<tr>
<td>12</td>
<td>654,729,075</td>
</tr>
<tr>
<td>13</td>
<td>13,749,310,575</td>
</tr>
<tr>
<td>14</td>
<td>316,234,143,225</td>
</tr>
<tr>
<td>15</td>
<td>7,905,853,580,625</td>
</tr>
</tbody>
</table>

For a small number of species, say less than 15, it is be possible to exhaustively enumerate all the trees, and for each tree calculate the minimum parsimony score. The tree that has the overall minimum parsimony score is reported.
Sequential addition strategy

Given three species, there is a single unrooted tree.
Each branch can serve as an insertion point, adding a new branch off the middle of any existing branch.
Sequential addition strategy

Therefore producing 3 four species unrooted trees.
Sequential addition strategy

The same process is applied to all 3 four species trees.
Sequential addition strategy

A four species unrooted tree has 5 edges, thus leading to 5 new unrooted trees.
Sequential addition strategy

There will be 15 five species unrooted trees.
Sequential addition strategy
Generate an initial tree $T$ (using UPGMA method for instance).
Compute $L$ the minimum parsimony score of $T$ (lowest score so far).
Create two empty lists, open and solutions.
Create an unrooted tree for three species and add it open.
While open is not empty
    Remove the front element of the list and call it current.
    Foreach tree $t$ created by a sequential addition to current do
        If the minimum parsimony score of $t$ is larger than $L$ then discard $t$.
        If the minimum parsimony score of $t$ is lower than $L$
            If $t$ has $n$ leaves:
                clear solutions
                add $t$ to solutions
                set $L$ to the minimum parsimony score of $t$
            Else add $t$ to the rear of open.
        Else (equals case)
            If $t$ has $n$ leaves: add $t$ to solutions
            Else add $t$ to the rear of open.
    solutions is the list of all the solutions, their score is $L$. 
Bound = L
Greedy algorithm

1. Generate an initial topology (using UPGMA for instance);
2. Apply nearest neighbour interchange (NNI) transformations to all the internal edges;
3. Select the minimum parsimony tree;
$n > 20$: Nearest-neighbour interchange (NNI) local search heuristic

Other heuristics include: subtree pruning and regrafting (SPR) or tree bisection and reconnection (TBR).
Remarks: distance-based vs character-based

Distance-based methods compute the pairwise sequence distances 1) directly, 2) in isolation, 3) before inferring the tree topology.

Instead, for character-based methods, 1) extant sequences are never compared directly 2) the pairwise distances depend on the reconstructed ancestral sequences, and 3) this process (solving the small phylogeny problem) takes all the sequences into account.
The particular methods that were presented are not modeling the base substitutions accurately.

Specifically, these methods are ignoring the fact that multiple substitutions (for a given site) are likely to occur in any given branch of the tree (time interval).
III. Maximum likelihood methods

Informal discussion!
III. Maximum likelihood methods

$P(D|\Theta)$ denotes the probability of the data given some model $\Theta$ (set of parameters, such as tree topology, branch length, evolutionary model . . . ).

Let $L(\Theta) = P(D|\Theta)$ be the **likelihood function**.

The **maximum likelihood estimate** is the value of $\Theta$ that maximizes $L(\Theta)$. 

Marcel Turcotte (turcotte@site.uottawa.ca)  
CSI 5126. Algorithms in bioinformatics
Let $L(\Theta)$ be the **likelihood** of a phylogenetic tree. $L(\Theta)$ is defined as the probability of the data (generally sequences) for a given tree (topology, branch length, evolutionary model), $P(\text{observed sequences}|\text{tree})$.

A **maximum likelihood** approach finds a tree, amongst all possible trees, with the largest value of $L(\Theta)$. Such tree explains best the data.

[ See Felsenstein 2004, pages 251–253. ]

Assumptions that are generally made:

1. Sites are independent
2. Lineages are independent
Probability of a tree

\[
\sum_{i} \sum_{j} p_{i} q_{iA}(t_{2}) q_{ij}(t_{2} - t_{1}) q_{jA}(t_{1}) q_{jT}(t_{1})
\]
where $k$, $l$, $m$ are nucleotide types found at the given sequence position in the 3 organisms under study. Assuming that the positions (sites) are independent one from another (are evolving independently), the probability of the tree would be the product over all site probabilities.
The $q_{ij}(t)$ terms give the probability of finding the nucleotide type $j$ at a given site knowing that its ancestor had the nucleotide type $i$ at the same position at time $t$ (earlier). Examples of substitution schemes modeling multiple substitutions for a given time interval include Jukes-Cantor one-parameter model and Kimura’s two-parameter model.
Transition rate: blue and transversion rate: red
**Probability of a tree: model of evolution**

**JC69:** Jukes and Cantor 1969; bases are equiprobable; transition rate = transversion rate

**F81:** Felsenstein 1981; variable base composition; transition rate = transversion rate

**K80:** Kimura 1980; bases are equiprobable; transition rate $\neq$ transversion rate

**HKY85:** Hasegawa et al. 1985; variable base composition; transition rate $\neq$ transversion rate; variable transition and transversion rates
\[ p(j|i, t) = \begin{cases} 
\frac{1}{4}(1 + 3e^{-4\alpha t}) & \text{if } j = i \\
\frac{1}{4}(1 - e^{-4\alpha t}) & \text{if } j \neq i 
\end{cases} \]

where \( \alpha \) is the mutation rate parameter.
In addition to the base model, most methods allow for relaxations:

- Variable rates across positions ($+\Gamma$)
- Variable rates across lineages ($+\mathcal{I}$)
HKY85+Γ + I implies variable base composition, transition rate ≠ transversion rate, variable transition and transversion rates, that vary across sites and lineages.
These models are also used to estimate pairwise distances for building phylogenies using distance-based approaches (e.g. Neighbour-joining).
Let $L$ be the **likelihood** of a phylogenetic tree. $L$ is defined as the probability of the data for a given tree, $P(\text{observed sequences}|\text{tree})$.

A **maximum likelihood** approach finds a tree, amongst all possible trees, with the largest value of $L$. Such tree explains best the data.

Furthermore, the length of the branches are unknown and must be estimated as part of this process.

Finding an exact solution to this problem is impractical when the number of input sequences is large, say 5 sequences/species.

Heuristic techniques have been developed to explore the tree space.
Maximum likelihood methods

- Generate an initial tree topology $\mathcal{T}$ (e.g. using NJ)
- Calculate its likelihood $\mathcal{L}$
- For a fixed number of iterations
  - From $\mathcal{T}$, generate new trees using NNI, SPR or TBR
  - For each new tree $\mathcal{T}'$, calculate its likelihood $\mathcal{L}'$
  - $\mathcal{L} = \mathcal{L}'$ and $\mathcal{T} = \mathcal{T}'$ if $\mathcal{L}' > \mathcal{L}$
Remarks: Parsimony vs Maximum Likelihood

- For a given tree topology, maximum parsimony considers all the reconstructions that lead to the same optimal score.
- Maximum parsimony under-estimates the number of evolutionary events (because of the multiple substitutions along the branches of the tree).
- Maximum likelihood, through its evolutionary models, takes into account multiple substitutions, rate variations amongst sites and lineages, etc.
- For a given tree topology, maximum likelihood considers all the reconstructions (and not only the most parsimonious ones).
- This is the most time consuming approach of all three.
Intuitively, the sites (columns) of an alignment that contain a single nucleotide type (invariant sites) provide no useful information for building a phylogenetic tree using a character-based approach.

A site is informative if it allows to discriminate between trees, i.e. the minimum parsimony scores for at least two trees are different for that site, otherwise the site is uninformative.

Clearly, invariant sites are uninformative.
Informative sites

<table>
<thead>
<tr>
<th>Species</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\alpha)</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>(\beta)</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>(\gamma)</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>(\delta)</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>T</td>
</tr>
</tbody>
</table>

⇒ Adapted from [5, pages 99–101].
Given 4 species, there are 3 possible unrooted trees

\[ \alpha \quad \beta \quad \gamma \quad \delta \]

\[ \alpha \quad \beta \quad \gamma \quad \delta \]

\[ \alpha \quad \gamma \quad \beta \quad \delta \]
Fortunately, there is a simple rule to identify informative sites. There has to be at least two types that occur at least twice at that site.

Uninformative sites are discarded prior to the inference of the tree.

Notice that those sites are typically kept by distance-based methods. This partly explains why the methods are producing different results.
Further readings

- Bayesian inference of phylogenetic trees
- Quartet methods
- Evolutionary networks (as opposed to evolutionary trees)
- Bootstraps, consensus, comparing trees
- Matching interior nodes (taxonomic units) with paleontological information, so as to assign time to events


D. E. Krane and M. L. Raymer.  
*Fundamental Concepts of Bioinformatics.*  
Benjamin Cummings, 2003.

W.-H. Li and D. Graur.  
*Fundamentals of Molecular Evolution.*  
Please don’t print these lecture notes unless you really need to!