EE550
Computational Biology

Week 8 Course Notes

Instructor: Bilge Karaçalı, PhD
Topics

• Inter-species evolutionary relationships via phylogenetic trees
  – Trees
    • Rooted trees
    • Unrooted trees
    • Tree topology
  – Sequences
  – Distance metrics
  – Clustering schemes
Phylogenetic Trees

• From molecular sequences to evolutionary relationships
  – Molecular sequences express how close or far apart different species are in terms of accumulated differences
    • Mutations in terms of substitutions and indels
  – Organisms with more similar sequences can be thought of descending from a more recent common ancestor
  – The evaluation of the ancestral history between different species is studied by molecular phylogenetics
    • Shared common ancestors how far ago and between which species
Phylogenetic Trees

- Trees provide the ideal scheme for representing the evolutionary relationships between different species
  - The species (at the present time) start out as the leaves
  - In the time past, leaves are merged into branches indicating common ancestry
    - Leaves with more similar sequences are merged first
  - The branches are merged into more ancient species
  - …
Example

- Phylogenetic tree of dogs (by Wayne et al., University of California)
  - Several parameters taken into account
    - characteristics of skulls, skeletons and chromosomes,
    - genetic analysis of mitochondrial DNA,
    - non-coding and coding nuclear DNA,
    - protein analysis
  - Corroboration with fossil evidence was also sought for validation

Source:
http://www.nbii.gov/portal/community/Communities/Ecological_Topics/Genetic_Diversity/Taxonomy_Phylogenetics_Systematics/
Rooted Phylogenetic Trees

• The root of a phylogenetic tree (if it exists) indicates the original species that is ancestral to all.
• Rooted phylogenetic trees thus have:
  – A root where all studied species have descended from.
  – A graded time axis indicating the evolutionary time it took for each species to differentiate from the common origin.
• Rooted phylogenetic trees provide information on
  – when a given pair of species had a common ancestor.
  – which pair of species diverged earlier or later than another pair.
  – how much evolutionary time has passed between the bifurcations.
Rooted Phylogenetic Trees

- Notes:
  - Most tree construction algorithms produce rooted trees by default
  - Just because a tree is shown like it has a root does not make it a rooted tree
    - The fine print under the figure must be read carefully!!
Unrooted Phylogenetic Trees

• It is not always possible to deduce a temporal order of events from the molecular data
  – The mutation rates may vary from species to species
• In such cases, the direction of changes are questionable among ancestral species
• Without a clear understanding of which species are descending from which ancestor, it is not possible/feasible/realistic to establish a common ancestral species

→ Unrooted trees
Unrooted Phylogenetic Trees

- The species at the present time are at the outer rims of the tree
  - Leaves
- The ancestral species are located inwards
- The hypothetical root is possibly somewhere around the ancestral species
  - On a link between a pair of ancestral species
  - But the exact location is unknown
Unrooted Phylogenetic Trees

- Rooted trees can easily be converted into unrooted trees
  - Simply ignore the root and spread out the leaves
- Unrooted trees cannot be converted to rooted trees so easily
  - One would have to identify the root
  - Identifying the root of an unrooted tree requires a priori information
  - This can be addressed by including an outgroup
    - sufficiently close to the species’ of interest
    - sufficiently far so that the root would be on its link
Unrooted Phylogenetic Trees
Tree Topology

- The topology of a phylogenetic tree indicates the set of bifurcations observed among ancestral species’
  - Roughly the “shape” of a tree

- When one or more ancestral relationships change, one obtains a different phylogenetic tree
  - The topology of the tree is different

- Note:
  - The actual drawing of the tree may change without affecting the topology
  - One obtains a different tree only when the ancestral relationships are altered
Sequences

• Evolutionary distances are inferred from sequence differences
  – Nucleic acid sequences
  – Amino acid sequences

• Constructing phylogenetic trees over a set of species requires sets of homologue sequences – sequences with common ancestry
  – Homologue genes
  – Homologue proteins

• Phylogenetic trees can be constructed on any homologue set
  – The results obtained on different homologue sets can vary!!

• A selection must be made with regard to the biological question at hand
  – Studying the evolution of a particular gene of interest
  – Studying the evolution of gene families
    • With lots of sequences available from many species
Sequences – Selection

• Selection of sequences are also important for practical purposes
  – **If the sequences are too similar**, the study will lack reliable evolutionary differentiation information
    • High similarity between sequences is an indication that the evolutionary process has not had adequate time to induce significant sequence alterations
      – A slow mutation rate and/or a short time period
    • This is contrary to the goal of producing a phylogeny
  – **If the sequences are too distinct**, the errors in evolutionary distances will take over
    • The standard deviation of the error in evolutionary distance estimation increases with the extent of sequence differences
      – This will cause dramatic changes in the resulting tree topology
    • Large differences between sequences are also suggestive of the lack of necessary homology among the sequences
Sequences – Alignment

• Phylogenetic tree construction is guided by multiple sequence alignment
  – Multiple sequence alignment algorithms have a set of innate choices
    • Similarity scores between aligned nucleotides or amino acids
    • Gap penalties
    • Sequence to cluster alignment procedures
  – Variations in these choices produces different multiple alignments
  – It is imperative that the multiple alignments obtained at the beginning of tree construction are valid and reasonable
    • Manual verification likely to be worth the effort, if at all feasible
  – Choice of the set of sequences upon which to construct a phylogeny directly determines how reliable the alignments will be
    • Highly variable sequence segments are not suitable choices
      – High variability implies high noise and little phylogeny information
Distance Metrics

• Multiple sequence alignment determines
  – the locations and extents of gaps to be inserted into each sequence in the set
  – so that all sequences are jointly aligned

• The resulting alignment must then be used to compute the evolutionary distances between the sequences
  – Multiple sequence alignment algorithms make use of a substitution model to determine the rates at which they will evaluate the matches and the mismatches
    • Evolutionary model in nucleic acid sequences
      – Jukes-Cantor
      – Kimura’s two-parameter model
    • Substitution matrices in amino acid sequences
      – PAM
      – BLOSUM
  – This substitution model determines the relationship between sequence differences and evolutionary distances
Distance Matrices

• Let $S$ be a set of $K$ homologous nucleic acid sequences, aligned to serve as the basis for phylogenetic tree construction

$$S = \{S_1, S_2, \ldots, S_K\}$$

- $S_k(i) \in \{A, G, T, C, _\}$ for $k = 1,2, \ldots, K$ and $i = 1,2, \ldots, N$ (for nucleic acid sequences)
- $S_k(i) \in \{A, R, N, \ldots, C, _\}$ for $k = 1,2, \ldots, K$ and $i = 1,2, \ldots, N$ (for amino acid sequences)
  - Multiple sequence alignment produces same length aligned sequences, where the common length (after the insertion of gaps) is denoted by $N$

• Let $d_{k,m}$ denote the evolutionary distance between the $k$’th and the $m$’th sequence
  - The computation of $d_{k,m}$ depends on the underlying substitution model
    - J-K model: $d_{k,m} = -\frac{3}{4} \log \left(1 - \frac{3}{4} D_{k,m}\right)$, where $D_{k,m}$ denotes the substitution ratio
    - Evolutionary distances are computed between the sequences typically by ignoring the gaps

• The resulting distance matrix $d = [d_{k,m}]$ is then to be used to derive the phylogenetic tree
Clustering Schemes

• Constructing a phylogenetic tree requires grouping the most similar species together into clusters earlier than the others

→ hierarchical clustering
  – Grouping species is easy:
    • Find the ones with the smallest evolutionary distance
    • Make a group containing the two closest species
  – Grouping clusters is not so easy:
    • Requires defining a distance for clusters

• Once a cluster distance is decided upon, a general strategy goes as follows:
  – Start with each species in a distinct cluster
  – Find the most similar pair of clusters among the available set
  – Merge the two clusters into a larger cluster
  – Update the distance matrix by
    (i) replacing the original two clusters with the newly formed cluster and
    (ii) updating the cluster-to-cluster distances involving the new cluster
      • The size of the distance matrix is reduced by one
  – Repeat until only one cluster remains
Cluster Distances

- Clusters are collections of sequences
- Distances between clusters can therefore be defined using the pairwise distances between their elements
- Let $C_i$ and $C_j$ be two clusters of size $m$ and $n$ respectively
  - $C_i = \{A_1, A_2, \ldots, A_m\}$
  - $C_j = \{B_1, B_2, \ldots, B_n\}$
- Let also $d(A, B)$ denote the inferred evolutionary distance between any two aligned sequences $A$ and $B$
- Consider the following quantities:
  - $\rho_{\text{min}}(C_i, C_j) = \min_{k, \ell} d(A_k, B_\ell)$
  - $\rho_{\text{max}}(C_i, C_j) = \max_{k, \ell} d(A_k, B_\ell)$
  - $\rho_{\text{mean}}(C_i, C_j) = \frac{1}{mn} \sum_{k, \ell} d(A_k, B_\ell)$ (UPGMA)
- These quantities all satisfy the distance properties and can therefore be used to determine the distances between clusters
Example

• Phylogeny of blue tongue viruses
  – A set of 5 sequences obtained from the NCBI Nucleotide database
    • Bluetongue virus isolate BTV-9/RSAvvv1/09 VP6 protein gene, complete cds
    • Bluetongue virus isolate BTV-9/BOS2002/02 VP6 protein gene, complete cds
    • Bluetongue virus isolate BTV-4/RSAvvv3/04 VP6 protein gene, complete cds
    • Bluetongue virus isolate BTV-4/ARG2002/01 VP6 protein gene, complete cds
    • Bluetongue virus isolate BTV-1/GRE2001/01 VP6 protein gene, complete cds
  – Multiple alignment carried out using the Clustal Omega software package
    • Clustal Omega webserver at the EMBL website
  – Evolutionary distances between sequences identified using the Kimura two-parameter model
  – Phylogenetic tree constructed using the minimum distance metric
Example

- Multiple sequence alignment

<table>
<thead>
<tr>
<th>SeqA Name</th>
<th>Len (nt)</th>
<th>SeqB Name</th>
<th>Len (nt)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene_34</td>
<td>1052</td>
<td>Gene_35</td>
<td>1049</td>
<td>79</td>
</tr>
<tr>
<td>Gene_34</td>
<td>1052</td>
<td>Gene_36</td>
<td>1049</td>
<td>78</td>
</tr>
<tr>
<td>Gene_34</td>
<td>1052</td>
<td>Gene_37</td>
<td>1051</td>
<td>94</td>
</tr>
<tr>
<td>Gene_34</td>
<td>1052</td>
<td>Gene_38</td>
<td>1049</td>
<td>78</td>
</tr>
<tr>
<td>Gene_35</td>
<td>1049</td>
<td>Gene_36</td>
<td>1049</td>
<td>91</td>
</tr>
<tr>
<td>Gene_35</td>
<td>1049</td>
<td>Gene_37</td>
<td>1051</td>
<td>78</td>
</tr>
<tr>
<td>Gene_35</td>
<td>1049</td>
<td>Gene_38</td>
<td>1049</td>
<td>92</td>
</tr>
<tr>
<td>Gene_36</td>
<td>1049</td>
<td>Gene_37</td>
<td>1051</td>
<td>78</td>
</tr>
<tr>
<td>Gene_36</td>
<td>1049</td>
<td>Gene_38</td>
<td>1049</td>
<td>94</td>
</tr>
<tr>
<td>Gene_37</td>
<td>1051</td>
<td>Gene_38</td>
<td>1049</td>
<td>77</td>
</tr>
</tbody>
</table>
Example

- Pairwise distances:

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
<th>$S_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>0</td>
<td>0.2749</td>
<td>0.2828</td>
<td>0.0593</td>
<td>0.2824</td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.2749</td>
<td>0</td>
<td>0.1015</td>
<td>0.2783</td>
<td>0.0901</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0.2828</td>
<td>0.1015</td>
<td>0</td>
<td>0.2679</td>
<td>0.0615</td>
</tr>
<tr>
<td>$S_4$</td>
<td>0.0593</td>
<td>0.2783</td>
<td>0.2679</td>
<td>0</td>
<td>0.2766</td>
</tr>
<tr>
<td>$S_5$</td>
<td>0.2824</td>
<td>0.0901</td>
<td>0.0615</td>
<td>0.2766</td>
<td>0</td>
</tr>
</tbody>
</table>

- These distances have been computed using Kimura’s two-parameter model with

$$d = -\frac{1}{2}\log(1 - 2S - V) - \frac{1}{4}\log(1 - 2V)$$

where $S$ is the average substitutions between (A-G) or (T-C), and $V$ is the average substitutions between purines and pyrimidines.
Example

• Clustering:
  – Hierarchical clustering using the minimum distance definition for cluster distances

• Step 1:
  – The minimum distance is between $S_1$ and $S_4$
  – $S_1$ and $S_4$ are merged into a new cluster $S_{1,4}$
  – Its distances to the remaining clusters ($S_2$, $S_3$, and $S_5$) are computed using the minimum distance definition
    • $\rho(S_{1,4}, S_2) = \min(0.2749, 0.2783) = 0.2749$
    • $\rho(S_{1,4}, S_3) = \min(0.2828, 0.2679) = 0.2679$
    • $\rho(S_{1,4}, S_2) = \min(0.2824, 0.2766) = 0.2766$
Example

- Resulting distance matrix at the end of Step 1:

<table>
<thead>
<tr>
<th></th>
<th>$S_{1,4}$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{1,4}$</td>
<td>0</td>
<td>0.2749</td>
<td>0.2679</td>
<td>0.2766</td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.2749</td>
<td>0</td>
<td>0.1015</td>
<td>0.0901</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0.2679</td>
<td>0.1015</td>
<td>0</td>
<td>0.0615</td>
</tr>
<tr>
<td>$S_5$</td>
<td>0.2766</td>
<td>0.0901</td>
<td>0.0615</td>
<td>0</td>
</tr>
</tbody>
</table>
Example

• Step 2:
  – The minimum distance is between $S_3$ and $S_5$
  – $S_3$ and $S_5$ are merged into a new cluster $S_{3,5}$
  – Its distances to the remaining clusters ($S_{1,4}$, $S_2$) are computed using the minimum distance definition
    • $\rho(S_{3,5}, S_{1,4}) = \min(0.2679, 0.2766) = 0.2679$
    • $\rho(S_{3,5}, S_2) = \min(0.1015, 0.0901) = 0.0901$
Example

- Resulting distance matrix at the end of Step 2:

\[
\begin{array}{ccc}
S_{1,4} & S_2 & S_{3,5} \\
\hline
S_{1,4} & 0 & 0.2749 & 0.2679 \\
S_2 & 0.2749 & 0 & 0.0901 \\
S_{3,5} & 0.2679 & 0.0901 & 0 \\
\end{array}
\]

- Step 3:
  - The minimum distance is between \( S_2 \) and \( S_{3,5} \)
  - \( S_2 \) and \( S_{3,5} \) are merged into a new cluster \( S_{2,3,5} \)
  - Its distance to the remaining cluster (\( S_{1,4} \)) is computed using the minimum distance definition
    - \( \rho(S_{2,3,5}, S_{1,4}) = \min(0.2749, 0.2679) = 0.2679 \)
Example

- Resulting distance matrix at the end of Step 3:

<table>
<thead>
<tr>
<th></th>
<th>$S_{1,4}$</th>
<th>$S_{2,3,5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{1,4}$</td>
<td>0</td>
<td>0.2679</td>
</tr>
<tr>
<td>$S_{2,3,5}$</td>
<td>0.2679</td>
<td>0</td>
</tr>
</tbody>
</table>

- The merger of the remaining two clusters are inevitable
  - and forms the common ancestor of all five species’
- Note that this is a rooted tree!!
Remarks

- The tree construction procedure determines the order in which similar species and clusters are merged together.
- However, the evolutionary time between successive mergers are not well defined:
  - The distances are not necessarily additive
    - At the time of estimation from the sequence distances using a substitution model
  - The minimum distance method does not seek the additive nature of evolutionary distances
- Clearly, different measures of cluster distance are likely to produce different trees:
  - Which one is best for a particular application is subject for debate.
Summary

• Phylogenetic trees express the evolutionary relationships between a set of species
• The relationship is inferred from the similarities and differences between homologue sequences
• The inferred relationships are subject to influences from
  – The choice of homologue sequences
  – The multiple alignment method used
  – The method for converting sequences differences into evolutionary distances
  – The cluster distance definitions for assessing the similarities between clusters