EE550
Computational Biology

Week 9 Course Notes

Instructor: Bilge Karaçalı, PhD

Topics

- Evaluation and optimality in phylogenetic tree construction
  - Bootstrapping
  - Tree fitness measures
  - Optimization methods

Error in Phylogenetic Tree Construction

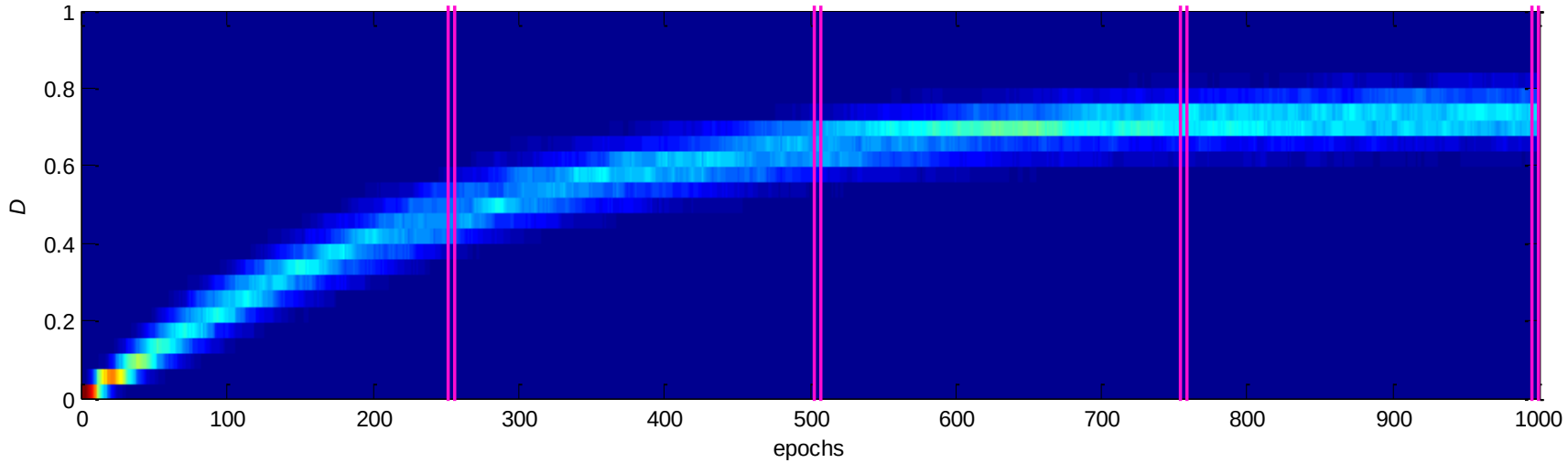
- Pairwise evolutionary distances between organisms are computed from the respective sequences
- While evolutionary models characterize the average substitution rates, the actual observations are subject to random fluctuations
- These random fluctuations directly affect the inferred evolutionary distances

Random Fluctuations by the Evolutionary Models

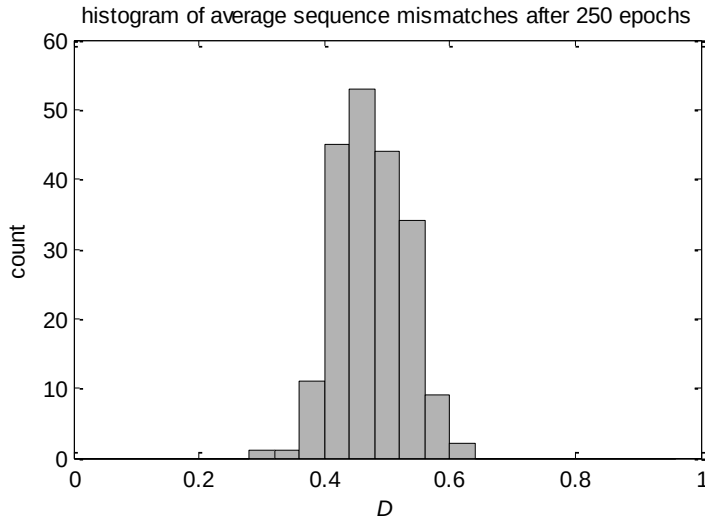
- Evolutionary models characterize the “average” behavior
  - Suppose two identical sequences are let to evolve for a certain period several times
  - The average/expected number of substitutions is governed by the evolutionary model
  - The number of substitutions observed in any one of the experiments may vary substantially from the mean
- Example:
  - Consider
    - the number of nucleotide mismatches
    - between sequences descending from the same ancestor
    - observed under the Jukes-Cantor model
    - across different time spans

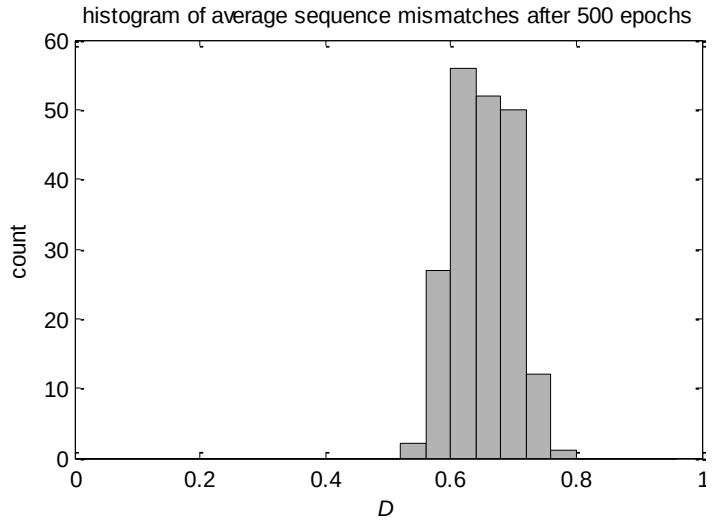
Random Fluctuations by the Evolutionary Models

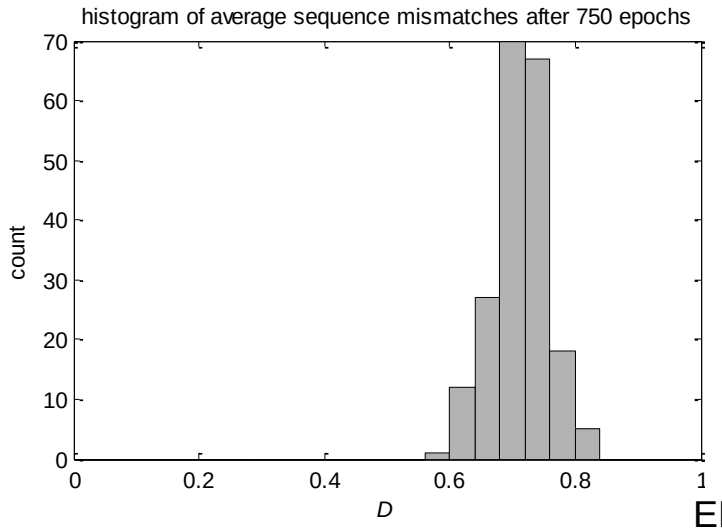
- Numeric illustration of noise in distance calculation:
  - Two nucleic acid sequences of length 100 were evolved for 1000 epochs under the Jukes-Cantor model with $\alpha = 0.0005$
  - The experiment was repeated 200 times and the average mismatch between the sequences were measured
  - The histograms of these distances were computed and plotted for increasing epochs

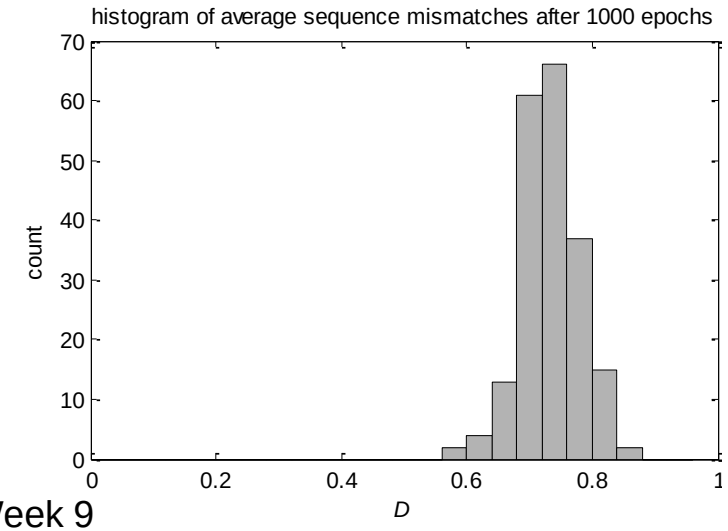


Random Fluctuations by the Evolutionary Models









Phylogenetic Tree Stability

- Phylogenetic trees are constructed based on similarity and dissimilarities between the sequences in consideration
  - Similar sequences are merged first
  - Dissimilar sequences are merged last
- Strong similarities and strong dissimilarities are clear indicators of if/when the corresponding sequences should be merged
- Lesser scores, however, are subject to heavy interference from random noise
- The end result is a tree with an unstable topology
  - Small changes in distances producing alternative tree topologies
    - Alternative order in which sequences are merged

Bootstrapping for Assessing Phylogenetic Tree Stability

- Bootstrapping
  - Suggested by Bradley Efron in 1979 to assess the reliability of parameter estimates from small sample sets
  - Operates on the principle that additional datasets can be constructed by **re-sampling** the actual observations **with replacement**
  - Estimation is carried out on every **bootstrap dataset**
  - The distribution of the estimate is determined using the values estimated on the bootstrap datasets
    - The more dispersed the distribution, the less stable the estimated value on the actual dataset

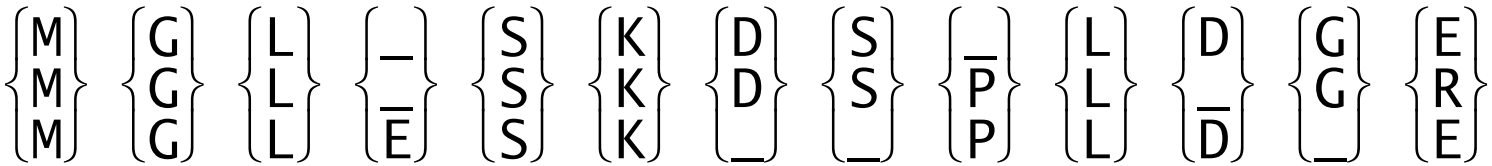
Bootstrapping for Assessing Phylogenetic Tree Stability

- Bootstrapping techniques are use to assess the stability and reliability of phylogenetic trees constructed from molecular data
  - **Bootstrap datasets** are constructed by randomly selecting aligned sites
  - **Bootstrap trees** are constructed using the bootstrap sequences
  - The number of times (in percentages) the nodes of the actual tree are observed in the bootstrap trees are noted
    - The higher the percentage, the more reason to believe in the validity of the associated ancestral node

Bootstrapping for Assessing Phylogenetic Tree Stability

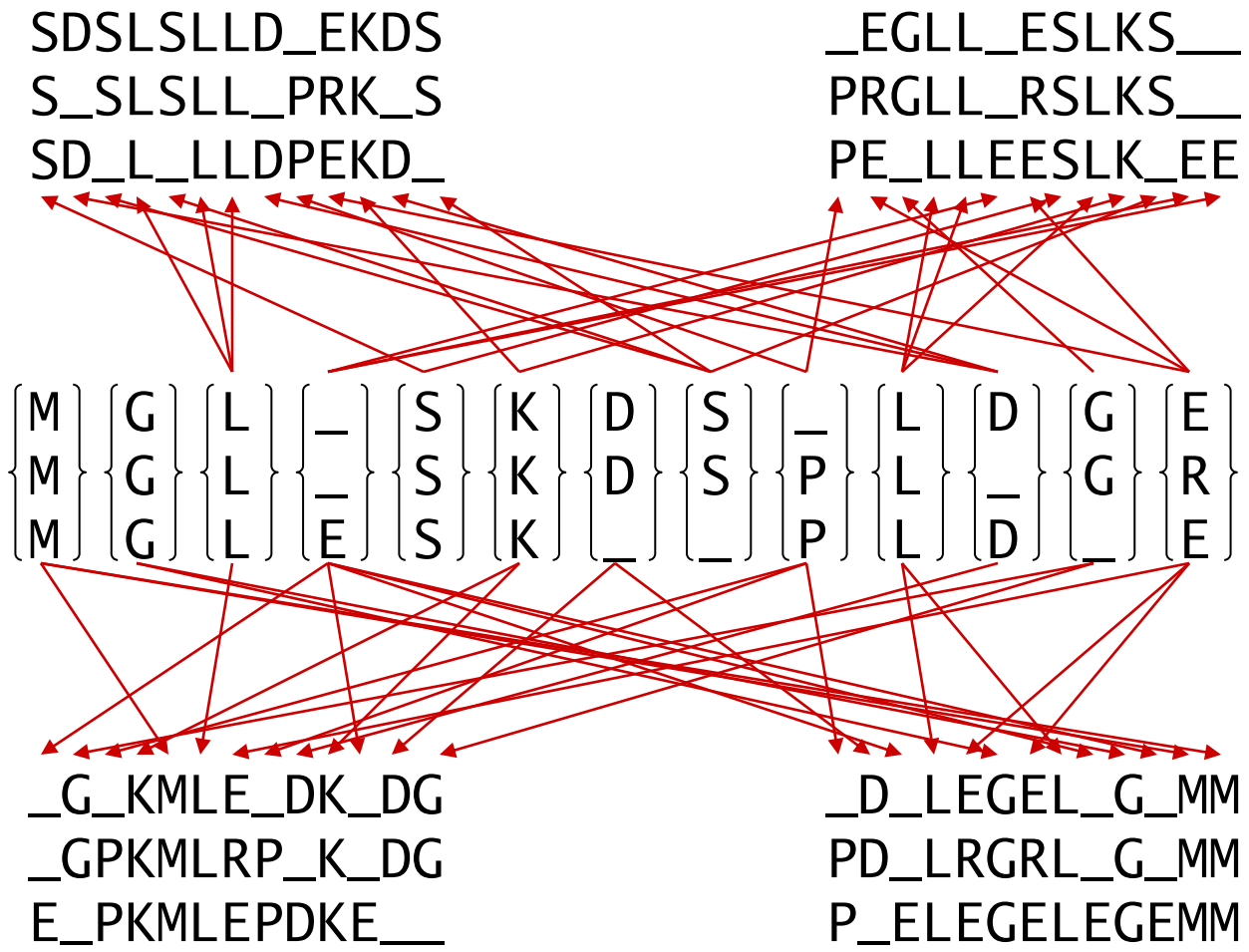
- Example:
  - Consider the 13-site multiple sequence alignment

    MGL_SKDS_LDGE
    MGL_SKDSPL_GR
    MGLESK__PLD_E
  - Each column of the multiple sequence alignment becomes a potential sample in the dataset



  - A bootstrap multiple sequence alignment is constructed by randomly selecting 13 samples from the dataset and organizing them sideways

Bootstrapping for Assessing Phylogenetic Tree Stability



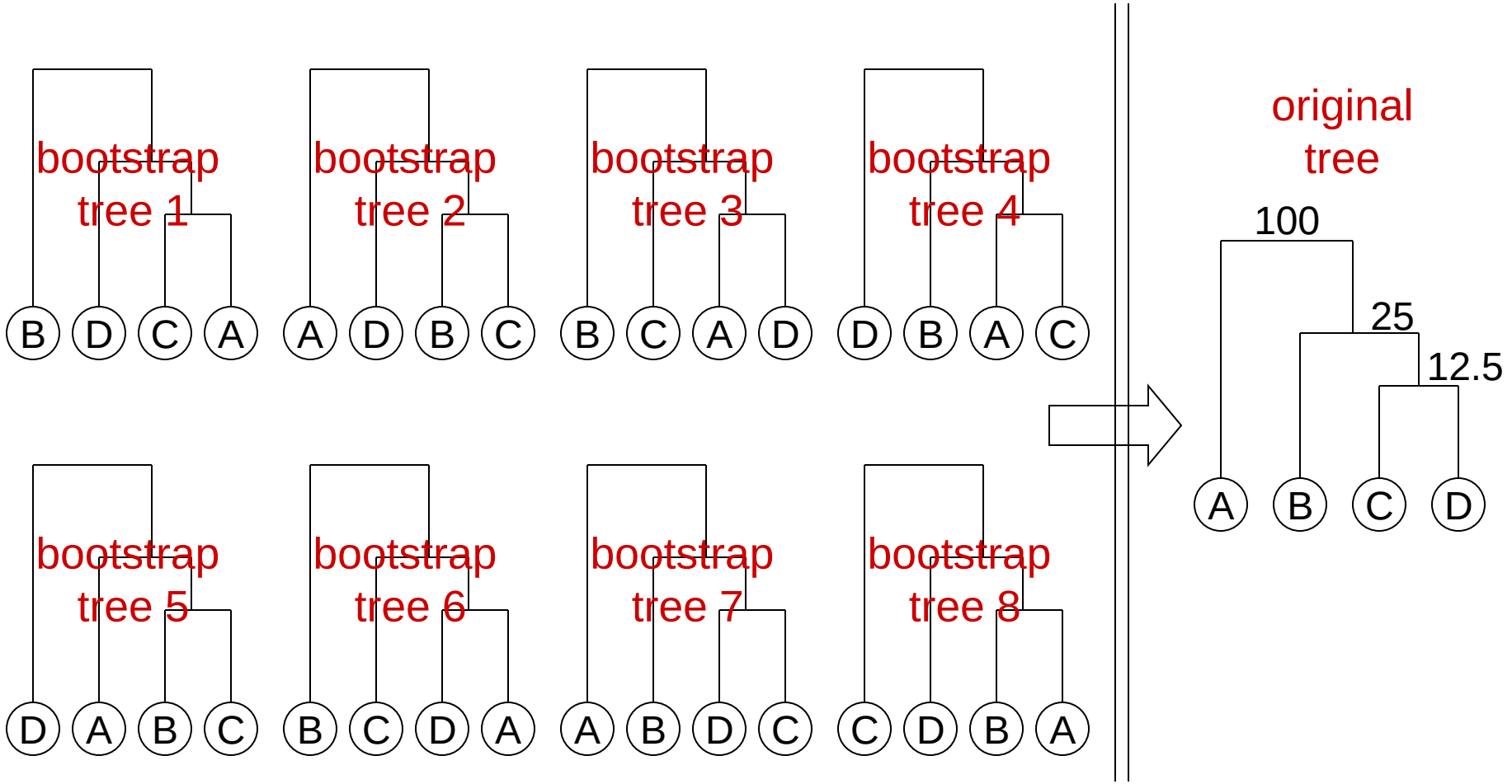
Bootstrapping for Assessing Phylogenetic Tree Stability

- Notes:
  - Each bootstrap multiple sequence alignment has the same length as the original
  - The sample dataset is of size equal to the number of sites in the original multiple alignment
  - However, bootstrap multiple sequence alignments are not merely a re-ordering of the original:
    - some points may be selected more than once
    - others may not be selected at all

Bootstrapping for Assessing Phylogenetic Tree Stability

- From each bootstrap multiple sequence alignment, a new phylogenetic tree is constructed
  - Each bootstrap multiple sequence alignment produces a different set of distances between the organism pairs
- The presence of each ancestral node of the original tree in the bootstrap tree is noted
  - An ancestral node is present if there is a node in the bootstrap tree under which the exact same organisms are grouped
  - The particular organization of these organisms further down may vary
- The average number of times an ancestral node is observed in the bootstrap sequences is marked on the original tree
  - Higher percentages indicate that the ancestral node is supported very strongly by the phylogenetic information in the sequences

Bootstrapping for Assessing Phylogenetic Tree Stability



Tree Fitness Measures

- Bootstrapping determines how stable and statistically reliable a constructed phylogenetic tree is
  - Phylogenetic trees are constructed using one of many tree construction algorithms
  - A given tree construction algorithm produces one and only one tree given a set of distances
- On the other hand, it does not evaluate how “good” the constructed tree is
  - Goodness of a constructed tree is to be measured using a scalar-valued function ➔ **a tree fitness measure**
  - The fitness measure incorporates all the desired and undesired properties expected from a phylogenetic tree

Example: Additive Trees

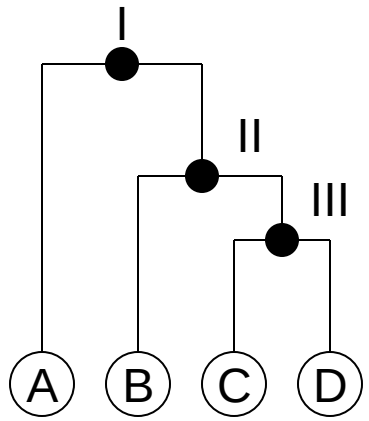
- A desired property in a phylogenetic tree is for the distances to be additive
  - The original set of distances are computed using a particular relationship between the sequence differences and the evolutionary distance
  - After these sequences are organized in a phylogenetic tree, the tree-based distances are expected to reproduce the original evolutionary distances
  - This property is called the additivity of the phylogenetic tree
- A fitness measure that penalizes the deviation from additivity is

$$E = \sum_{i,j} \left(d_{i,j} - d_{i,j}^{tree}\right)^2 / d_{i,j}^2$$

where

$d_{i,j}$ denotes the predicted evolutionary distance between the i'th and the i'th sequences, and

$d_{i,j}^{tree}$ is the distance reproduced by the tree



$$d_{A,B}^{tree} = d_{A,I} + d_{I,II} + d_{II,B}$$

$$d_{A,C}^{tree} = d_{A,I} + d_{I,II} + d_{II,III} + d_{III,C}$$

$$d_{A,D}^{tree} = d_{A,I} + d_{I,II} + d_{II,III} + d_{III,D}$$

$$d_{B,C}^{tree} = d_{B,II} + d_{II,III} + d_{III,C}$$

$$d_{B,D}^{tree} = d_{B,II} + d_{II,III} + d_{III,D}$$

$$d_{C,D}^{tree} = d_{C,III} + d_{III,D}$$

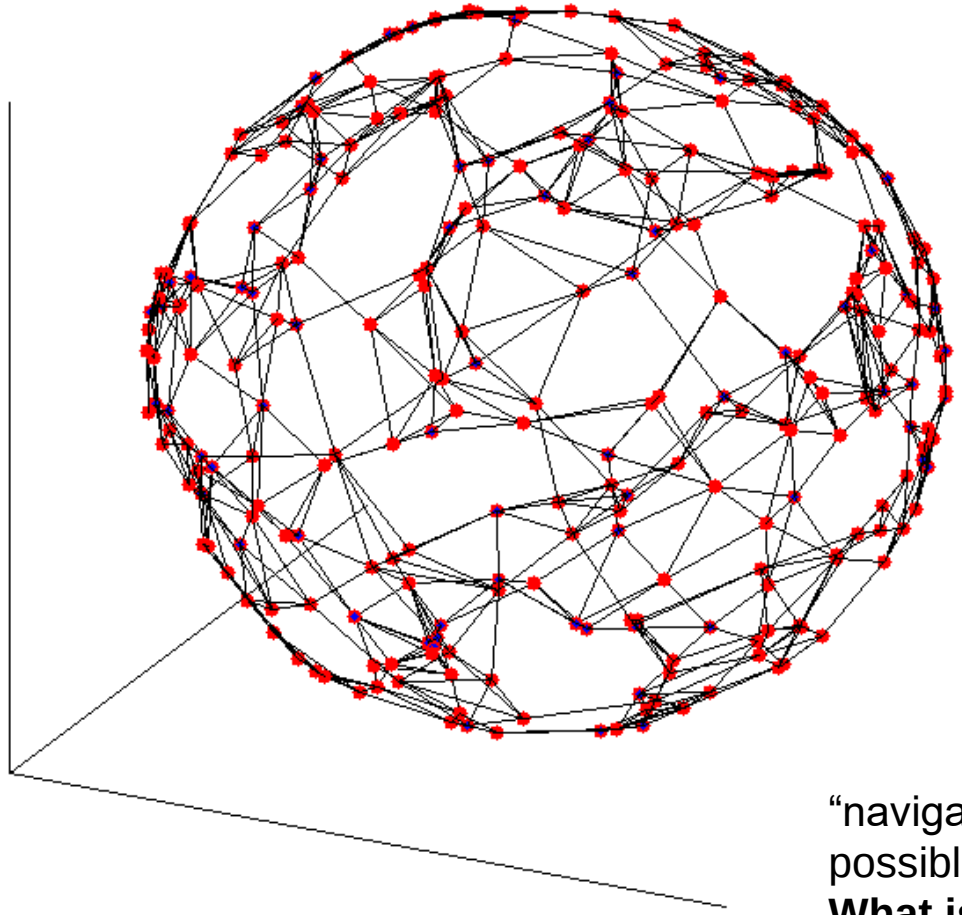
Neighbor Joining Method

- An alternative to hierarchical clustering of sequences is provided by the neighbor joining method
  - Saitou, N., Nei, M., "The neighbor-joining method: a new method for reconstructing phylogenetic trees," Molecular Biology and Evolution, 4(4):406-25 (1987).
  - Studier, J. A., Keppler, K. J., "A note on the neighbor-joining algorithm of Saitou and Nei," Molecular Biology and Evolution, 5(6):729-31 (1988).
- The method differs from hierarchical clustering in two principal ways:
  1) “nearest” nodes are those whose merger produces the tree with minimal overall distances
  2) the distances between the parent node and the remaining nodes are calculated to produce an additive tree

Search-Based Phylogenetic Methods

- An alternative to tree construction methods is provided by methods that evaluate many candidate trees and identify the most adequate one
  - The set of candidate trees contains potentially all possible tree topologies that can be constructed from the given set of sequences
  - The search-based methods employ a fitness measure to be optimized over this set
- Optimization of the fitness function of choice over the set requires the ability to move from one tree to another
  - The set of all possible trees must have a structure
  - This structure determines the neighborhood and distance over different trees
    - The distance from one tree to another
  - In this way, a better organized search scheme can be formulated compared to a random search
    - The neighborhood of an initial tree is evaluated to seek a better tree

Search-Based Phylogenetic Methods



"navigating" the set of all possible trees:
What is the criterion?

Search-Based Phylogenetic Methods

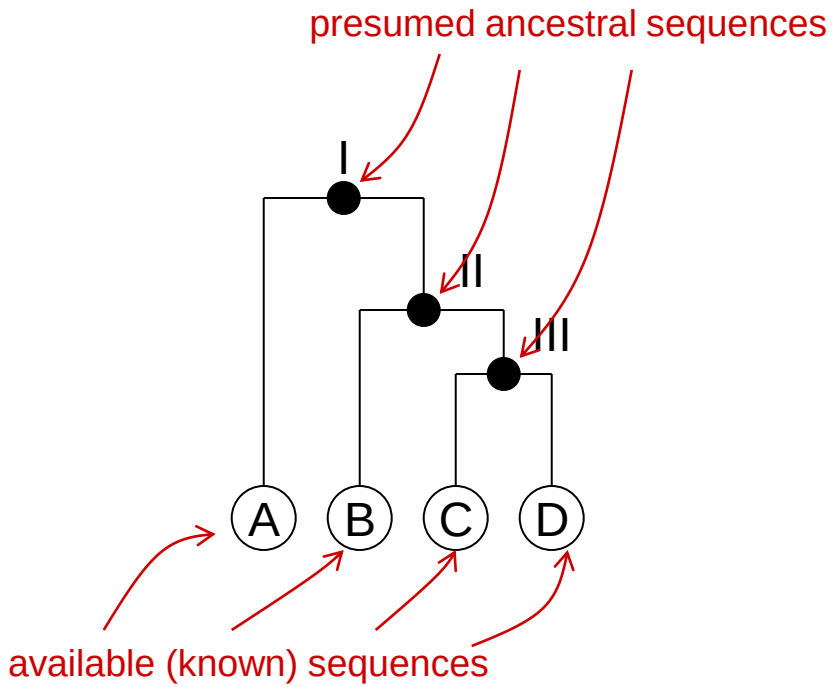
- The maximum likelihood criterion:
  - Given a sequence evolution model, one can calculate the likelihood of a given tree
    - Likelihood is a notion from the probability theory
    - Technically, it is defined as the underlying probability distribution function conditional to a hypothesis evaluated at the observations
    - It characterizes the odds associated with the particular hypothesis
    - The maximum likelihood estimation scheme simply picks out the hypothesis with the greatest likelihood
  - The maximum likelihood phylogenetic tree estimate of a given set of sequences is the tree that maximizes the likelihood function

Search-Based Phylogenetic Methods

- The likelihood of a phylogenetic tree is defined as

  the **probability** with which the specific evolutionary relationship dictated by **the tree** would **produce the observed sequences**

- Example:
  - Given
    - the phylogenetic tree T relating the sequences $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ and
    - a model $P(d)$ governing the substitution probabilities in evolutionary distance/time d
  - Consider the sequence elements (nucleotides or amino acids) at the n'th site, $\boldsymbol{A}_n$, $\boldsymbol{B}_n$, $\boldsymbol{C}_n$, and $\boldsymbol{D}_n$
    - Site correspondences established via multiple sequence alignment



The phylogenetic tree T

Search-Based Phylogenetic Methods

- Example (continued):
  - Then, the likelihood $L_{\boldsymbol{III}_n}(X) = \Pr\{\boldsymbol{III}_n = X\}$ for any letter X is given by

$$L_{\boldsymbol{III}_n}(X) = P_{X\to\boldsymbol{C}_n}\left(d_{\boldsymbol{III},\boldsymbol{C}}\right) \cdot P_{X\to\boldsymbol{D}_n}\left(d_{\boldsymbol{III},\boldsymbol{D}}\right)$$

    where $P_{X\to\boldsymbol{C}_n}\left(d_{\boldsymbol{III},\boldsymbol{C}}\right)$ is the probability that the letter X at $\boldsymbol{III}_n$ is converted to $\boldsymbol{C}_n$ after an evolutionary time of $d_{\boldsymbol{III},\boldsymbol{C}}$ and so on, for all X
  - As for $L_{\boldsymbol{II}_n}(Y)$, we have

$$L_{\boldsymbol{II}_n}(Y) = P_{Y\to\boldsymbol{B}_n}\left(d_{\boldsymbol{II},\boldsymbol{B}}\right) \cdot \sum_X L_{\boldsymbol{III}_n}(X) \cdot P_{Y\to X}\left(d_{\boldsymbol{II},\boldsymbol{III}}\right)$$

Search-Based Phylogenetic Methods

- Example (continued):
  - Finally, the expression for $L_{\boldsymbol{I}_n}(Z)$ becomes

$$L_{\boldsymbol{I}_n}(Z) = P_{Z\to \boldsymbol{A}_n}\left(d_{\boldsymbol{I},\boldsymbol{A}}\right) \cdot \sum_{\boldsymbol{Y}} L_{\boldsymbol{II}_n}(Y) \cdot P_{Z\to Y}\left(d_{\boldsymbol{I},\boldsymbol{II}}\right)$$

  - The log likelihood of the whole tree can then be computed by

$$\log\left(L(T)\right) = \sum_{n} \log\left(\sum_{Z} \pi_Z \, L_{\boldsymbol{I}_n}(Z)\right)$$

- Note:
  - The likelihood of a tree is computed over all possible occupants of all sites
    - Summations are carried out over all possible X's, Y's, Z's, and for all n

Search-Based Phylogenetic Methods

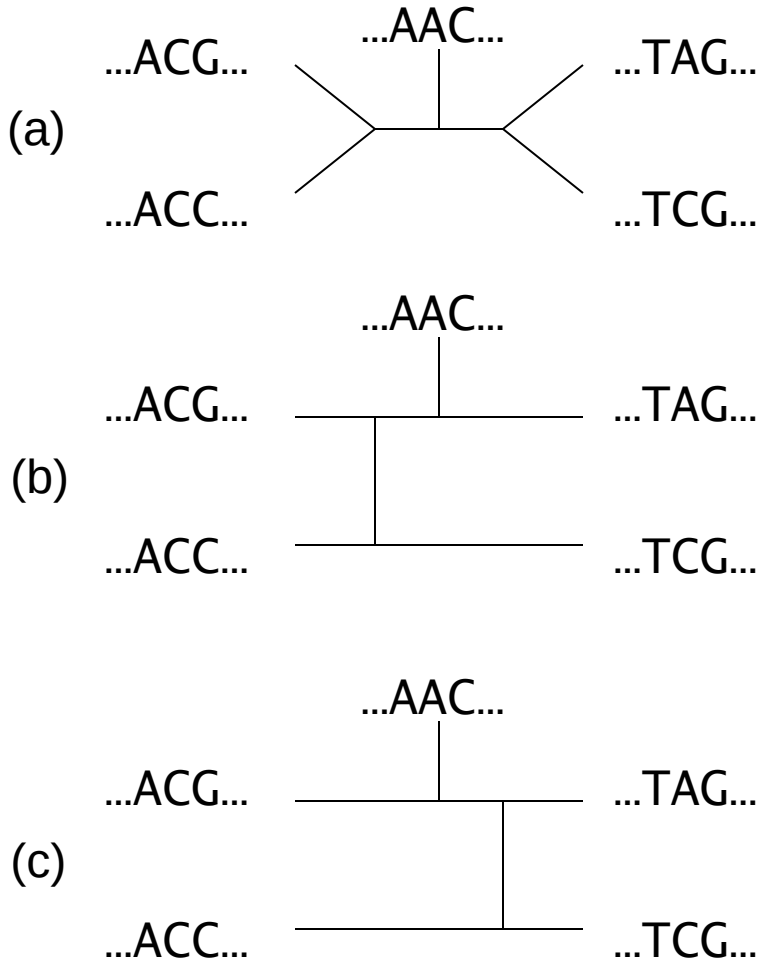
- The parsimony criterion:
  - Several corollaries:
    - “All else being equal, the best solution is the simplest” – Occam’s razor
    - “Things must be made as simple as possible, but not simpler” – Albert Einstein
    - “The most accurate indicator function comes from the most restricted indicator function sets that still fit the data” – Structural Risk Minimization principle
  - The parsimony criterion seeks to provide the least number of property changes
    - Each bifurcation effectively divides the descendant organisms into two groups:
      - those that possess a given property, and
      - those that do not
  - Before the rise of molecular phylogenetics, the parsimony method was the most popular phylogenetic tool used on morphological characteristics

Search-Based Phylogenetic Methods

- The parsimony criterion in molecular phylogenetics:
  - The phylogenetic characters in terms of nucleic acid or amino acid sequences are the sequence elements occupying a given site
  - In this context, the parsimony criterion favors the trees that offer the least number of nucleotide or amino acid substitutions
    - The penalty for substitutions between specific characters can be weighed according to a substitution rate mechanism
      - Evolutionary model
      - Scoring matrices

Search-Based Phylogenetic Methods

- According to the **first** characters,
  - Tree (a) entails 1 substitution
  - Tree (b) entails 2 substitutions
  - Tree (c) entails 2 substitutions
- According to the **second** characters,
  - Tree (a) entails 2 substitutions
  - Tree (b) entails 1 substitution
  - Tree (c) entails 2 substitutions
- According to the **third** characters,
  - Tree (a) entails 2 substitutions
  - Tree (b) entails 2 substitutions
  - Tree (c) entails 2 substitutions
- …



Remarks

- Parsimony and maximum likelihood are closely related
  - Both evaluate the candidate tree structure according to preferred attributes
    - In parsimony, fewer substitutions are preferred
    - In maximum likelihood, substitutions with higher probabilities are preferred
- On the other hand, they exhibit a key difference on their use of an evolutionary model
  - The evolutionary model is crucial in the maximum likelihood model
  - Parsimony essentially ignores the details of the model and uses only the part where fewer substitutions are more likely
- Consequently, the choice between the two depends on how accurate the model is expected to behave
  - If highly accurate, than maximum likelihood
  - Otherwise, parsimony

Case Study: Phylogeny over EGFR

- Procedure:
  - Download amino acid sequences of EGFR proteins from different species
  - Carry out multiple sequence alignment using Clustal Omega at the url
    https://www.ebi.ac.uk/Tools/msa/clustalo/
  - Obtain a phylogenetic tree using the neighbor joining algorithm on the EBI website at the url
    https://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny/
  - Obtain another phylogenetic tree using the maximum likelihood criterion using the resource at the url
    http://iqtree.cibiv.univie.ac.at
  - Obtain additional phylogenetic trees using the resources available at the url
    http://www.phylogeny.fr/

Summary

- There are many ways in which sequence similarity can be put into the form of a phylogenetic tree
- Depending on the choice of the algorithm and/or the associated algorithmic parameters, the resulting trees may vary in their topologies
- Other phylogenetic methods make use of this multiplicity of candidate trees to pick out the best one according to a given criterion
- If the phylogenetic information available in the sequences is strong enough, one topology would be expected to stand out and capture the most agreeable phylogeny