Local Multiple Sequence Alignment (cont')
Sequence File Formats

Gibbs Sampling

- Gibbs Sampling is another statistical method similar in nature to the EM algorithms.
- Gibbs sampling combines both EM and simulated annealing techniques in order to determine a maximal local alignment of multiple sequences.
- Goal: Find most probable pattern by sampling from motif probabilities to maximize ratio of model:background probabilities

- The idea behind Gibbs sampling is to determine the most probable pattern common to all of the sequences by sliding them back and forth until the ratio of the motif probability to the background probability is a maximum.

Predictive Update Step

- random motif start position chosen for all sequences except one
- Initial alignment used to calculate residue frequencies for motif and background
- similar to the Expectation Step of EM

Sampling Step

- ratio of model:background probabilities normalized and weighted
- motif start position chosen based on a random sampling with the given weights
- Different than E-M algorithm

Gibbs Sampling

- process repeated until residue frequencies in each column do not change
- The sampling step is then repeated for a different initial random alignment
- Sampling allows escape from local maxima

Gibbs Sampling

- In order to improve the performance of the Bayesian approach to Gibbs sampling, Dirichlet priors (pseudocounts) are added into the nucleotide counts

- employs a shifting routine that will take a current multiple motif alignment, and shift it a few bases to the left or the right, in order to see if only part of the motif is being found

- A range of motif sizes can be explored in Gibbs sampling as well

Gibbs Sampling Extensions

- Gibbs sampling can be extended to search for multiple motifs in the same set of sequences, and to find a pattern in only a fraction of the sequences.

- In addition, certain model-specific parameters can be enforced, such as palindromic sequences

Gibbs Sampler Web Interface

- http://bayesweb.wadsworth.org/gibbs/gibbs.html

Hidden Markov Models

- Hidden Markov Models (HMMs)
  - probabilistic models for studying sequences of symbols.
- HMMs can model matches, mismatches, insertions and deletions of symbols.
- HMMs have been deeply rooted in speech recognition problems.
- In speech recognition, the problem is the phonemes (or words) that have been spoken in a particular time frame.

Hidden Markov Models

- Consider the difficulty.
  - Everyone you meet has a different voice.
  - Everyone speaks with a slight variation
    - this might be caused by an accent, the person having a cold, or differences in physiological development.
- However, humans are able to distinguish what the speaker is saying.
  - The idea behind speech recognition is to take in a spoken word and to try to fit it to a specific model of possible words.
    - This may in fact be close to what the brain does

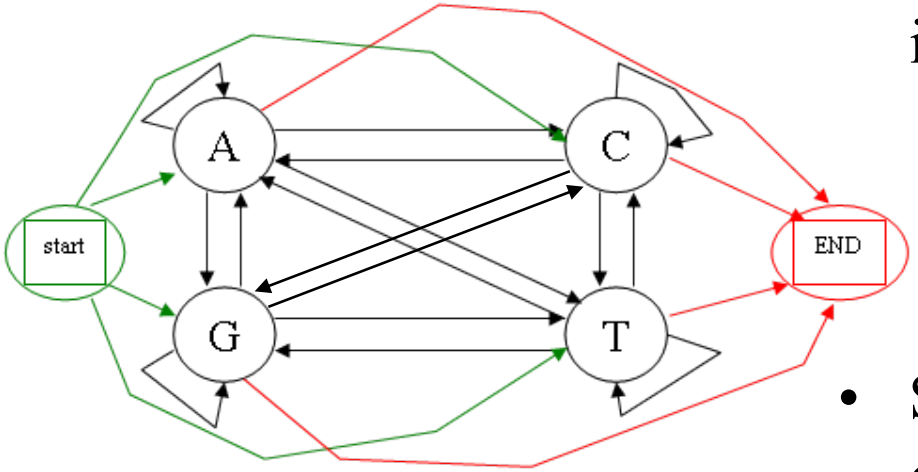
Hidden Markov Models

- Problems in sequence analysis are similar.
- For instance, given an amino acid sequence, we may want to determine the protein family to which it belongs.
- The amino acid sequence can be treated similarly to the speech signal in a given frame, and the amino acids can be treated as the phonemes.

Markov Chain

- a probabilistic model that generates a sequence where the probability of a symbol depends upon the previous symbol.
  - A traffic light is an example of a Markov chain.
- A Markov Chain can be used to model a random DNA sequence, where there are four states:
  - A, C, G, T, one for each letter in the alphabet.
- When we are given a certain state, there is a transition from that state to another state with an associated probability
  - called a transition probability.
- An example Markov Chain can be drawn as follows:

Markov Chain



- The key property of a Markov chain is that
  - the probability of a symbol S at position p, (S_p) depends only upon the previous symbol S at position p–1, (S_{p-1}), and not on the entire previous sequence.
- Since the probability of a symbol is dependent upon the previous symbol, a prime example for the use of Markov chains is in the detection of CpG islands, which are rich in the dinucleotide CG.
  - **CpG** (CG) **islands** is a short stretch of DNA in which the frequency of the CG sequence is higher than other regions.

- "p" simply indicates that "C" and "G" are connected by a phosphodiester bond.

Markov Chain

- The process of methylation in biological systems will typically convert the nucleotide C to a T with a high probability when a CG nucleotide is encountered.
  - As a result, there will be an overabundance of the dinucleotide TG, and an underabundance of the dinucleotide CG.
- If we ignore the start and end states for now, we can see that there are sixteen different transitions.
  - A study of regions of genomic DNA has determined normal genomic transition probabilities to be the following,
    - where the FROM node is labeled along the rows to the left, and the TO node is labeled along the columns above:

Markov Chain

	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

- The model shown above can then assign these weights to the edges of the graph

Markov Chain

- In some regions of the genome, such as the promoter region of genes, methylation is suppressed.
  - In these regions, the dinucleotide CG is found in greater quantities.
- In fact, the nucleotides C and G are found to a greater degree than elsewhere in the genome.
  - A study of regions of genomic DNA where CpG islands exist has determined the transition probabilities to be the following:

Markov Chain

	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

- A new model just like the one above can have its transition properties assigned according to the new table.
- Now we have two different models:
  - the first where CpG islands are absent,
  - the second where CpG islands are present.

Markov Chain

- Let's call the first model the non-CpG model and the second model the CpG model.
- Given a new sequence, how would we determine whether it belongs to the non-CpG model or the CpG model?
- Remember, the key property of a Markov chain
  - the probability of a symbol S at position p, (S_p) depends only upon the previous symbol S at position p–1, (S_{p-1}),
    - not on the entire previous sequence.

Markov Chain

- Therefore, to find the probability that a sequence fits a model,
  - you would multiply all of the conditional probabilities:

$P(x) = P(x_L|x_{L-1})P(x_{L-1}|x_{L-2})\ldots P(x_2|x_1)P(x_1)$

- which can be rewritten as:

$$P(x) = P(x_1)\prod_{i=2}^{L} a_{x_{i-1}x_i}$$

- where $a_{x_{i-1}x_i}$ is the probability from residue at position i-1 to the residue at position i

Markov Chain

- Let's consider for now that in the non-CpG model, P(A) = P(T) = 0.3; P(C) = P(G) = 0.2,
  - so that A and T are more probable.

- In the CpG model, consider P(A) = P(C) = P(G) = P(T) = 0.25.

- Now consider the sequence: GGCGACG
- The probability for this sequence:
  P(G)P(G|G)P(C|G)P(G|C)P(A|G)P(C|A)P(G|C)

Markov Chain

- For the non-CpG model can be calculated as:

  (0.20)(0.298)(0.246)(0.078)(0.248)(0.205)(0.078)
  = 0.000000453499

- For the CpG model can be calculated as:

  (0.25)(0.375)(0.339)(0.274)(0.161)(0.274)(0.274)(0.125) =
  0.0010526

- Given this information, it is more likely that this sequence fits the CpG model.
- One thing to note is how quickly the probability gets to zero.
  - This shows the importance of using log statistics.

Using Markov models for discrimination

- How different the non-CpG and CpG models are in relation to each other?
  - If they are not different enough, then there is not enough information to determine from which model a particular sequence is derived.
- In order to test whether we are able to discriminate between the two models, a log ratio is taken for each of the scores in the two previous tables to create a third table, where each entry, x, in the new table is equal to:

  $\log_2(P(x|\text{CpG model}) / P(x|\text{ non-CpG model}))$
- The resulting table is as follows:

Using Markov models for discrimination

	A	C	G	T
A	-0.740	0.419	0.580	-0.803
C	-0.913	0.302	1.812	-0.685
G	-0.624	0.461	0.331	-0.730
T	-1.169	0.573	0.393	-0.679

- Using this log-odds ratio table as the scores, we can then see that
  - a sequence with a negative score will belong to the non-CpG model,
  - a sequence with a positive score will belong to the CpG model.

Position Specific Scoring Matrix (PSSM)

- Position Specific Scoring Matrices incorporate information theory in order to gain a measure of how much information is contained within each column of a multiple alignment.
- The information contained within a PSSM is a logarithmic transformation of the frequency of each residue in the motif.

PSSMs and Pseudocounts

- One problem with creating a model of a sequence alignment that is then used to search databases is that there is a bias towards the training data
  - Some residues may be underrepresented
  - Other columns may be too conserved

- Solution: Introduce Pseudocounts to get a better indication

Pseudocounts

- Now the estimated probability is changed from a frequency of counts in the data to the following form:

$$P_{ca} = \frac{n_{ca} + b_{ca}}{N_c + B_c}$$

- P_{ca}: Probability of residue a in column c
- n_{ca}: count of a's in column c
- b_{ca}: pseudocount of a's in column c
- N_c: total count in column c
- B_c: total pseudocount in column c

PSSMs and pseudocounts

- These probabilities are then converted into a log-odds form (usually $\log_2$ so the information can be reported in bits) and placed in the PSSM .

Searching PSSMs

- In order to search a sequence against a PSSM, the value for the first residue in the sequence occurring in the first column is calculated by searching the PSSM.
- Similarly, the value for the residue occurring in each column is calculated. These values are added (since they are logarithms) to produce a summed log odds score, S.
- This score can be converted to an odds score using the formula 2^S.
- The odds scores for the motif beginning at each position can be summed together and normalized to produce a probability of the motif occurring at each location.

Information in PSSMs

- Information theory can give an appreciation for the amount of information contained within each sequence.

- When there is no information contained within a column, the amount of uncertainty can be measured as $\log_2 20 = 4.32$ for amino acids, since there are 20 amino acids.

- For nucleic acid sequences, the amount of uncertainty can be measured as $\log_2 4 = 2$.

Information in PSSMs

- If only one amino acid is found in a particular column, then the uncertainty is 0 – there is only one choice.

- If there are two amino acids occurring with equal probability, then there is an uncertainty to deciding which residue it is.

Measure of Uncertainty

- The amount of uncertainty for a particular column is measured as the entropy, as introduced previously

$$H_C = -\sum_{residues(a)} f_{ac} \log(p_{ac})$$

PSSM Uncertainty

- the uncertainty for the whole PSSM can be calculated as a sum over all columns:

$$H = \sum_{allcolumns} H_c$$

Relative Entropy

- In addition to the entropy measure given before, a relative entropy measure could be calculated as well. Relative entropy takes into account not only the data in the columns of the motif, but also the overall composition of the organism being studied. Relative entropy can be measured as:

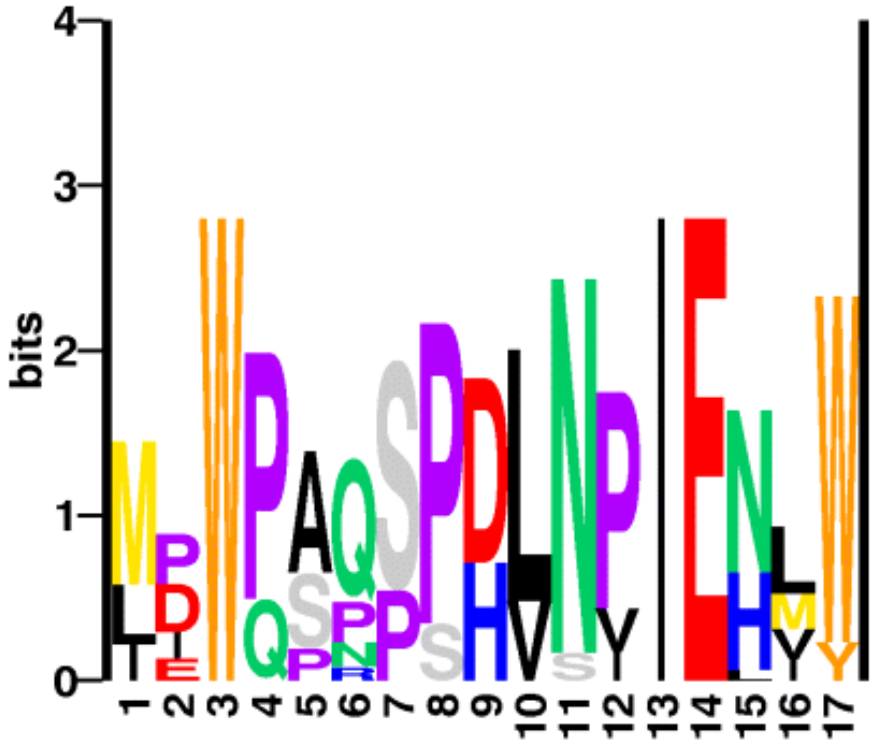
$$R_C = - \sum_{residues(a)} f_{ac} \log_2 (p_{ac} / b_a)$$

- $\boldsymbol{b_a}$ is background frequency of residue ***a*** in the organism

Sequence Logos

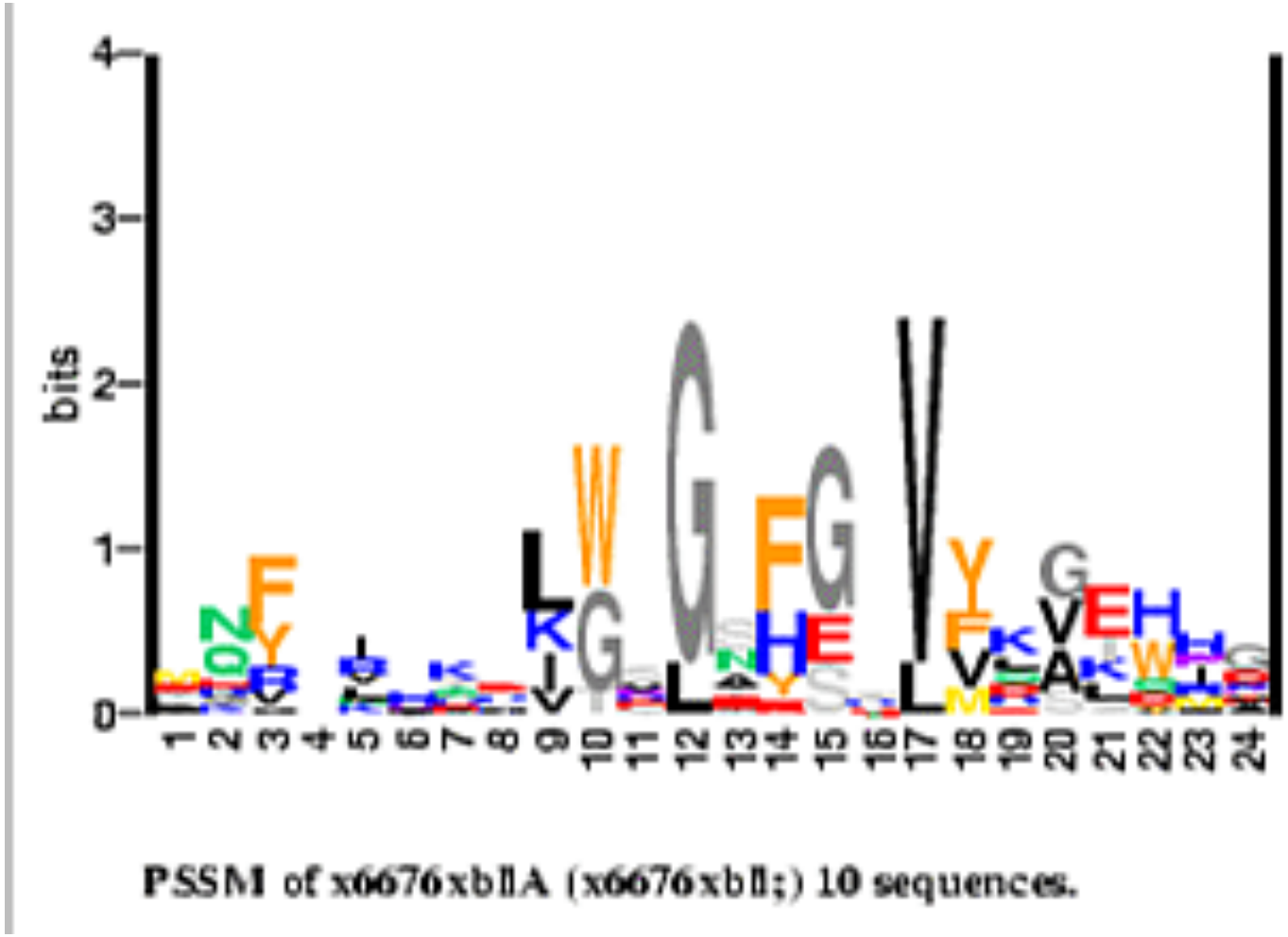
- One way to look at a particular PSSM is to view it visually. Sequence logos are one way to do so, by illustrating the information in each column of a motif.
- Such a graph can indicate which residues and which columns are the most important as far as sequence conservation is concerned.
- The height of the logo is calculated as the amount by which uncertainty has been decreased
- If the frequency in the column is less than the frequency in the background, then a negative relative entropy can be computed, which can be shown by an inverted character in the logo.

Sequence Logos



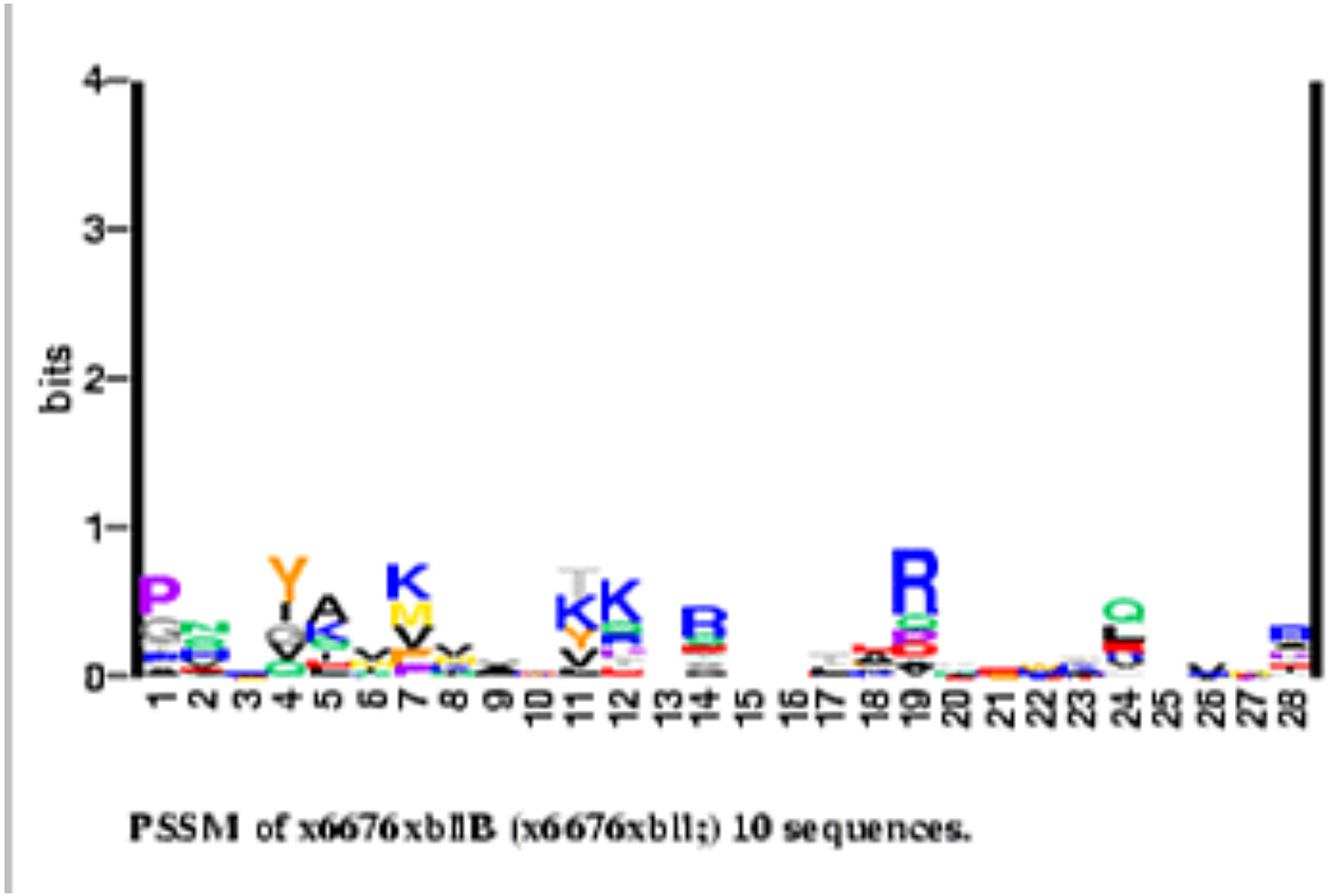
Logo of Gibbs Block D (Tc1) 9 sequences

Sequence Logos



PSSM of x6676xbllA (x6676xbll;) 10 sequences.

Sequence Logos



Sequence Editors

- Allow manual editing of alignments
- Add color to alignments
- Prepare images for publication

Sequence Editors

- CINEMA
- http://www.biochem.ucl.ac.uk/bsm/dbbrowser/CINEMA2.02/kit.html

- GeneDoc
- http://www.psc.edu/biomed/genedoc/

- MACAW
- http://ncbi.nlm.nih.gov/pub/schuler/macaw

- BoxShade
- http://www.ch.embnet.org/software/BOX_form.html

Sequence File Formats

- We have been using DNA and amino acid sequences already
- What is the typical format for these?
- ANSWER: Many different options

Sequence File Formats

- In order to standardize sequence data, The Nomenclature Committee of the International Union of Biochemistry and the *International Union of Pure and Applied Chemistry* (IUPAC)has established a standard code to represent bases that are uncertain or ambiguous. The code, often referred to as the IUPAC code, is as follows:

Standard Codes (IUPAC)

A = adenine

C = cytosine

G = guanine

T = thymine

U = uracil

R = G A (purine)

Y = T C (pyrimidine)

K = G T (keto)

M = A C (amino)

S = G C

W = A T

B = G T C

D = G A T

H = A C T

V = G C A

N = A G C T (any)

- Any other character besides the ones listed above (with the exception of the gap character ‘-‘) represents an error that will not be tolerated by nearly all sequence analysis programs.

- In addition to the nucleic acid codes, a standard single letter and three letter amino acid code has been formulated by IUPAC as well. The table for this code is as follows:

Standard IUPAC Codes

A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic acid
C	Cys	Cysteine
Q	Gln	Glutamine
E	Glu	Glutamic acid
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine
B	Asx	Aspartic acid or Asparagine
Z	Glx	Glutamine or Glutamic acid
X	Xaa or Xxx	Any amino acid

Fasta File Format

- Fasta sequence format is one of the most basic and widespread sequence formats.
- A sequence in fasta format has as its first line a descriptor beginning with a '>' character.
- The proceeding lines contain the sequence (either nucleotide or amino acid) using standard one-letter symbols.
- This format is extremely useful for sequence analysis programs, since it is devoid of numerical and nonsequence characters (with the exception of the newline character).

Fasta File Format

- Example Fasta Sequence:

```
>gi|27819608|ref|NP_776342.1| hemoglobin, beta [beta globin] [Bos taurus]
MLTAEEKAAVTAFWGKVKVDEVGGEALGRLLVVYPWTQRFFESFGDLSTADAVMNNPKVKAHGKKVLDSF
SNGMKHLDDLKGTFAALSELHCDKLHVDPENFKLLGNVLVVVLARNFGKEFTPVLQADFQKVVAGVANAL
AHRYH
```

- first line begins with '>', followed by gi, -- next field surrounded by '|' is GenBank identifier
- the keyword 'ref' -- field will be the reference for the version of this sequence.
- final field is the description

Fasta File Format

- Example Fasta Sequence:

```
>gi|27819608|ref|NP_776342.1| hemoglobin, beta [beta globin] [Bos taurus]
MLTAEEKAAVTAFWGKVKVDEVGGEALGRLLVVYPWTQRFFESFGDLSTADAVMNNPKVKAHGKKVLDSF
SNGMKHLDDLKGTFAALSELHCDKLHVDPENFKLLGNVLVVVLARNFGKEFTPVLQADFQKVVAGVANAL
AHRYH
```

- nearly all sequence based programs treat anything following the '>' as a comment

- a few sequence analysis programs expect sequences to be in a strict fasta format

GenBank

- GenBank is the National Center for Biotechnology Information's nucleic acid and protein sequence database.
- It is the most widely used source of biological sequence data.
- GenBank file format contains information about the sequence, including literature references, functions of the sequence, locations of various features, etc.

GenBank

- information organized into fields, each with an identifier, justified to the farthest left column.
- Some identifiers have additional subfields.
- sequence data lies between the identifier ORIGIN and the '//' which signals the end of a GenBank record.

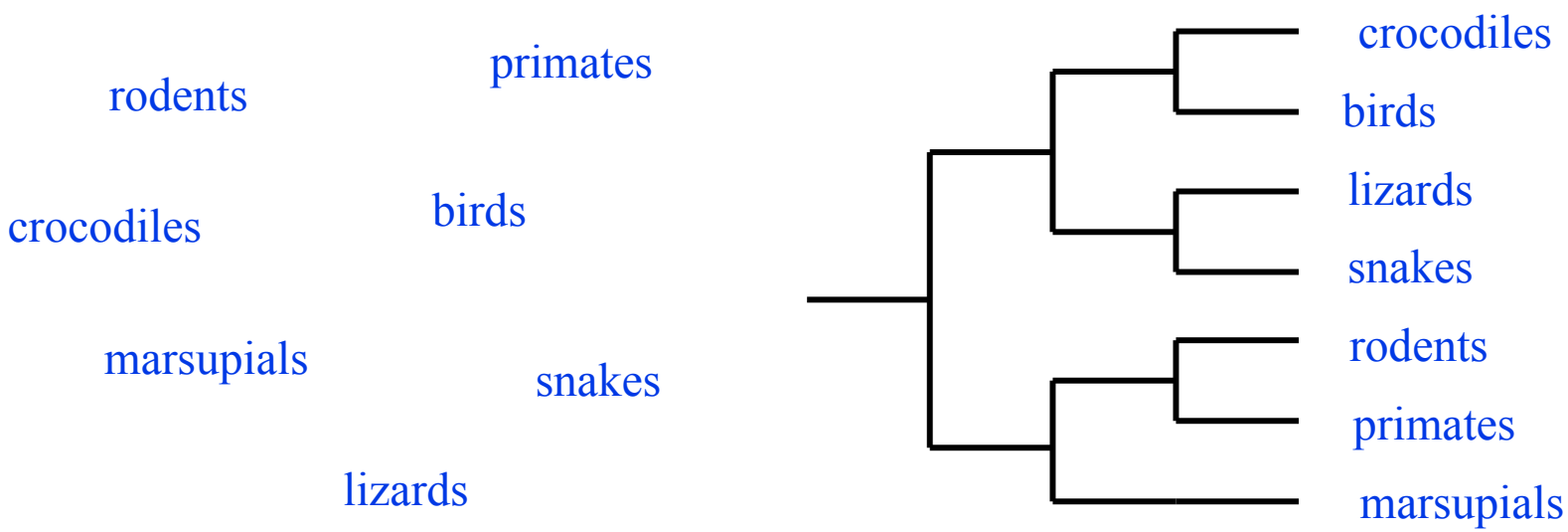
GenBank Record

LOCUS HBB 145 aa linear MAM 22-JAN-2003
DEFINITION hemoglobin, beta [beta globin] [Bos taurus].
ACCESSION NP_776342
VERSION NP_776342.1 GI:27819608
DBSOURCE REFSEQ: accession NM_173917.1
KEYWORDS .
SOURCE Bos taurus (cow)
ORGANISM Bos taurus Eukaryota; Metazoa; Chordata; Craniata;
Vertebrata; Euteleostomi; Mammalia; Eutheria; Cetartiodactyla;
Ruminantia; Pecora; Bovoidea; Bovidae; Bovinae; Bos.
REFERENCE 1 (residues 1 to 145)
AUTHORS Duncan,C.H.
JOURNAL Unpublished (1991)
COMMENT PROVISIONAL REFSEQ: This record has not yet been subject to final
NCBI review. The reference sequence was derived from M63453.1.
FEATURES Location/Qualifiers source 1..145

Phylogenetics

What is Molecular Phylogenetics

- Phylogenetics is the study of evolutionary relationships
- Example:
  - relationship among species



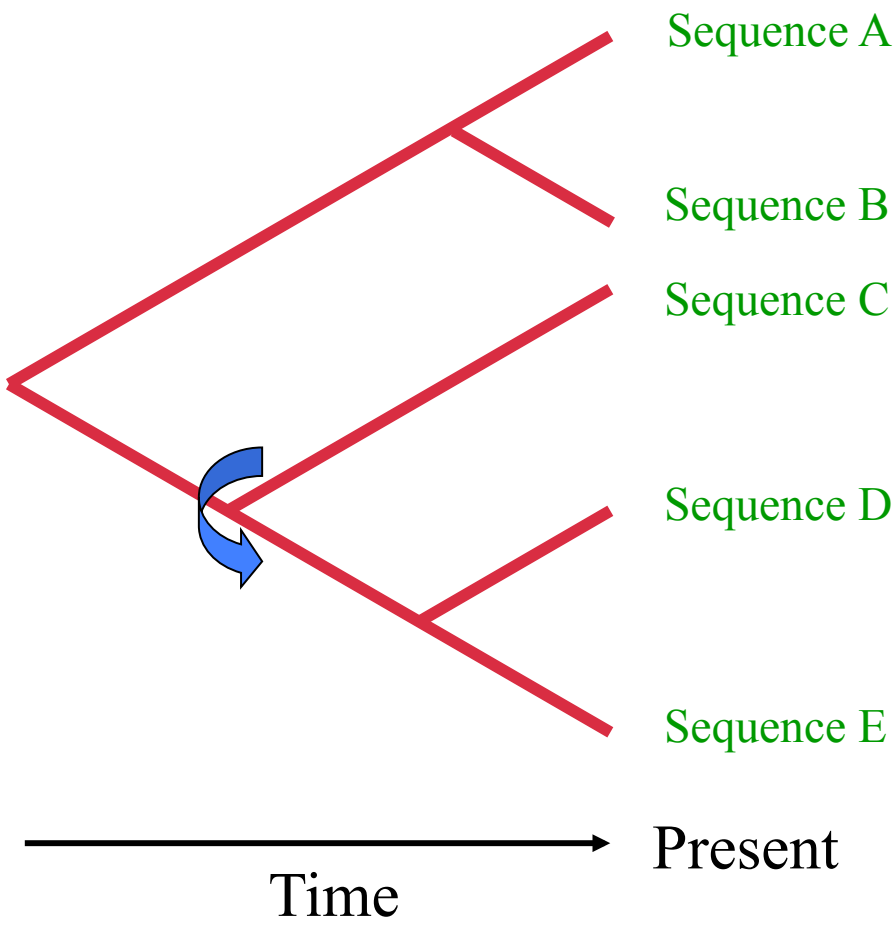
A Brief History of Molecular Phylogenetics

- 1900s
  - Immunochemical studies
    - cross-reactions stronger for closely related organisms
      - Nuttall (1902) - apes are closest relatives to humans!
- 1960s - 1970s
  - Protein sequencing methods, electrophoresis, DNA hybridization and PCR contributed to a boom in molecular phylogeny
- late 1970s to present
  - Discoveries using molecular phylogeny
    - Endosymbiosis - Margulis, 1978
    - Divergence of phyla and kingdom - Woese, 1987
    - Many Tree of Life projects completed or underway

Molecular data vs. Morphology/Physiology

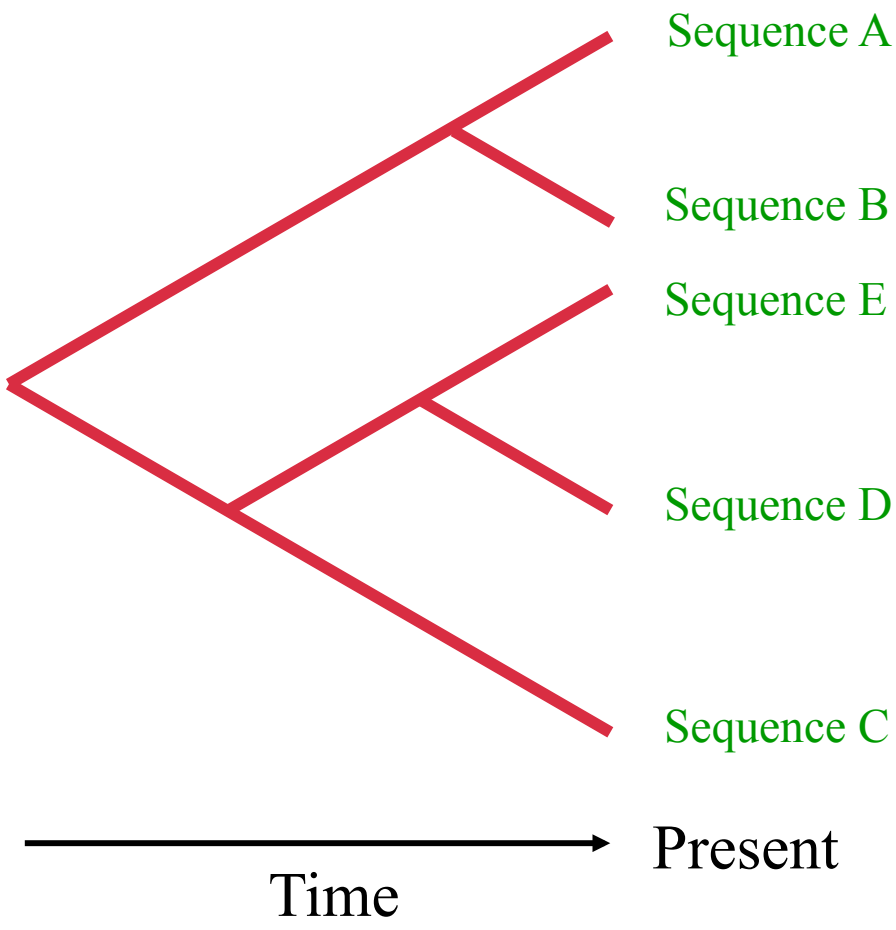
• Strictly heritable entities	• Can be influenced by environmental factors
• Data is unambiguous	• Ambiguous modifiers: “reduced”, “slightly elongated”, “somewhat flattened”
• Regular & predictable evolution	• Unpredictable evolution
• Quantitative analyses	• Qualitative argumentation
• Ease of homology assessment	• Homology difficult to assess
• Relationship of distantly related organisms can be inferred	• Only close relationships can be confidently inferred
• Abundant and easily generated with PCR and sequencing	• Problems when working with micro-organisms and where visible morphology is lacking

Phylogenetic concepts: Interpreting a Phylogeny



- *Physical* position in tree is not meaningful
- Swiveling can only be done at the nodes
- Only tree structure matters

Phylogenetic concepts: Interpreting a Phylogeny



- Physical position in tree is not meaningful
- Swiveling can only be done at the nodes
- Only tree structure matters

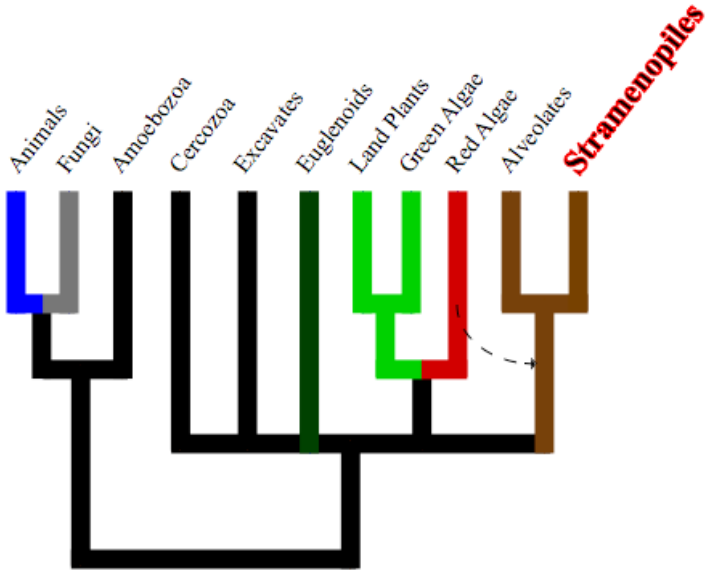
Tree Terminology

- Relationships are illustrated by a phylogenetic tree / dendrogram
  - Combination of Greek dendro/tree and gramma/drawing
  - A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering.
  - Dendrograms are often used in computational biology to illustrate the clustering of genes or samples, sometimes on top of heatmaps.
- A cladogram is a type of phylogenetic tree that only shows tree topology
  - the shape indicating relatedness.
    - It shows that, say, humans are more closely related to chimpanzees than to gorillas, but not the time or genetic distance between the species.
  - Combination of Greek clados/branch and gramma/drawing

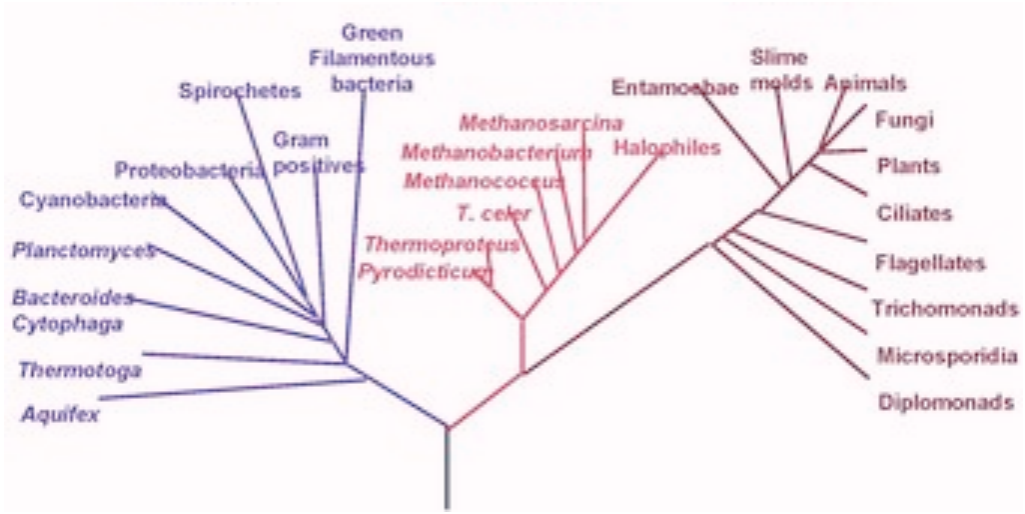
Tree Terminology

- The branching pattern is called the tree's topology
- Trees can be represented in several forms:

Rectangular cladogram

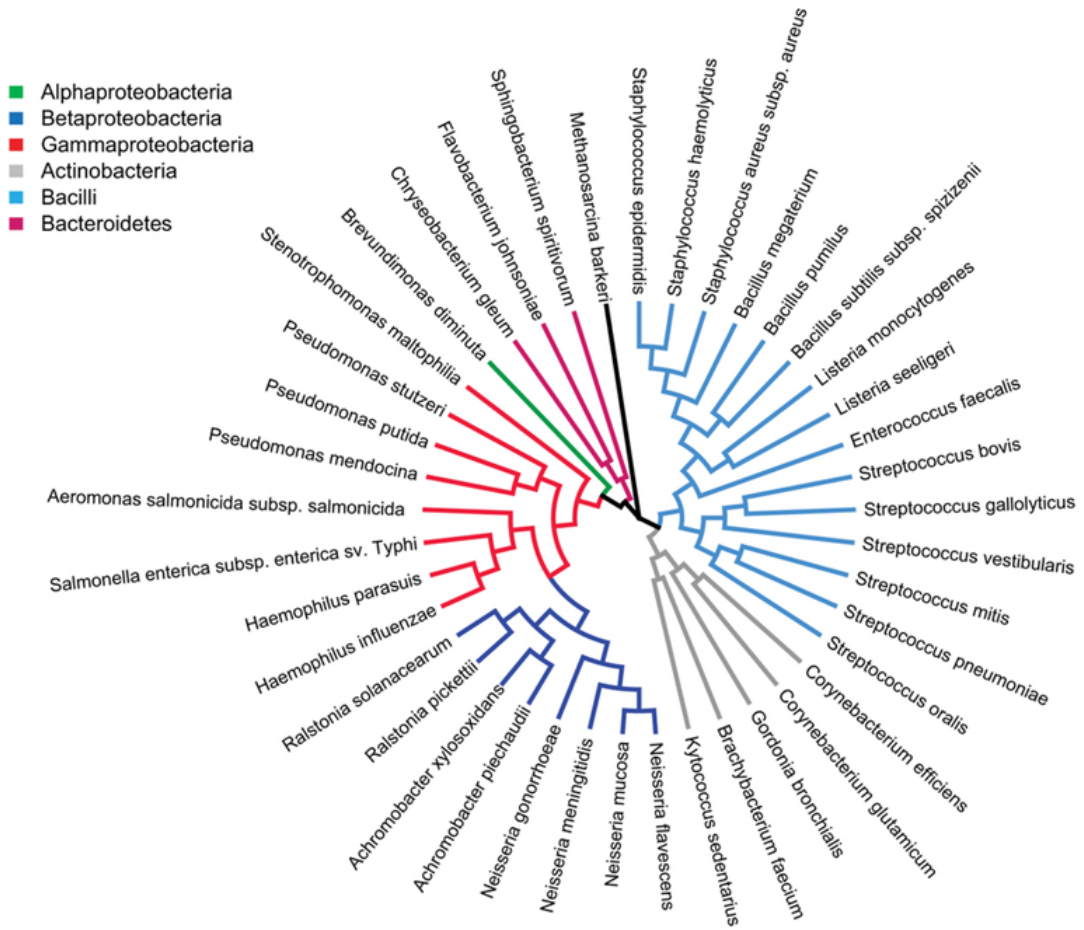


Slanted cladogram



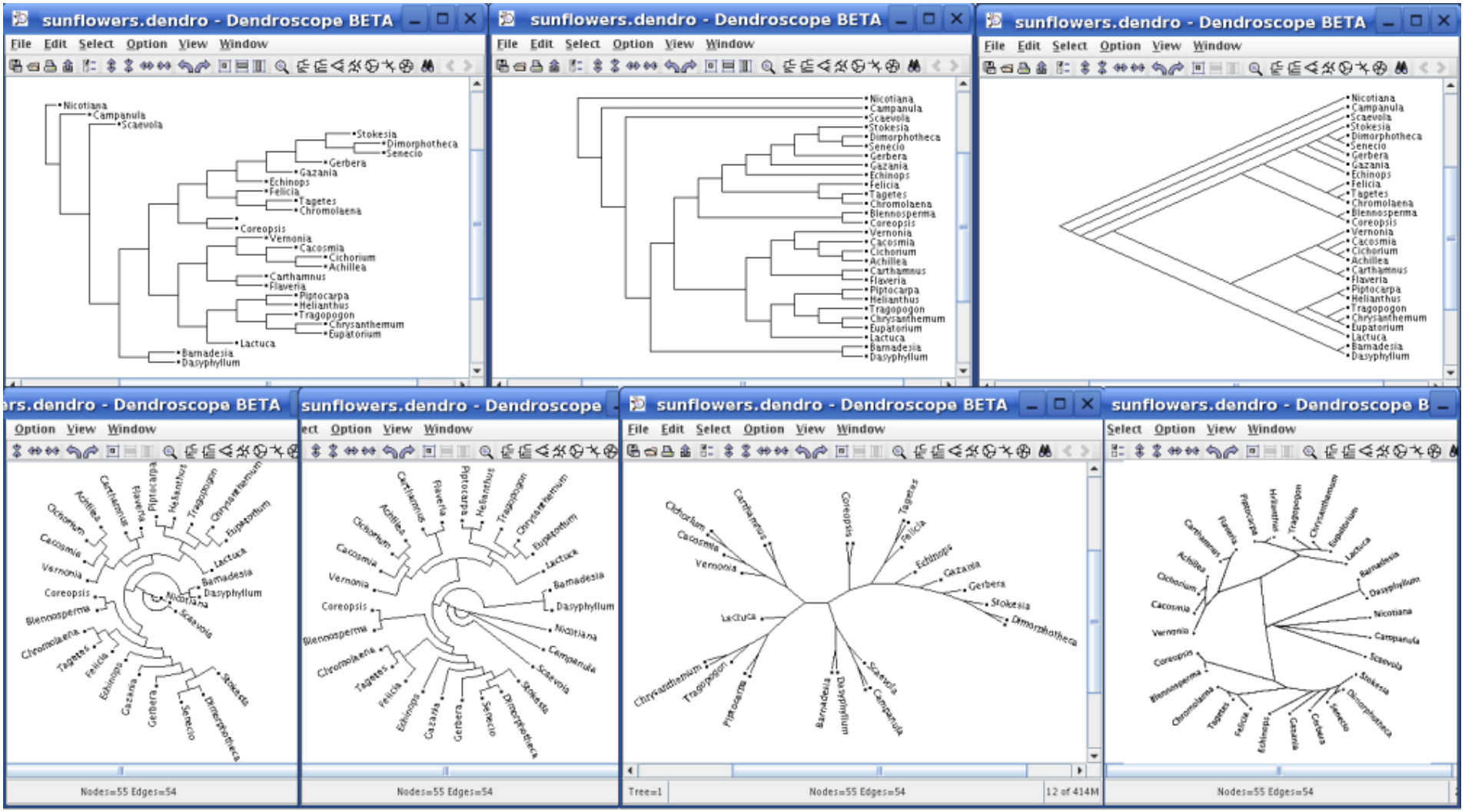
Tree Terminology

Circular cladogram

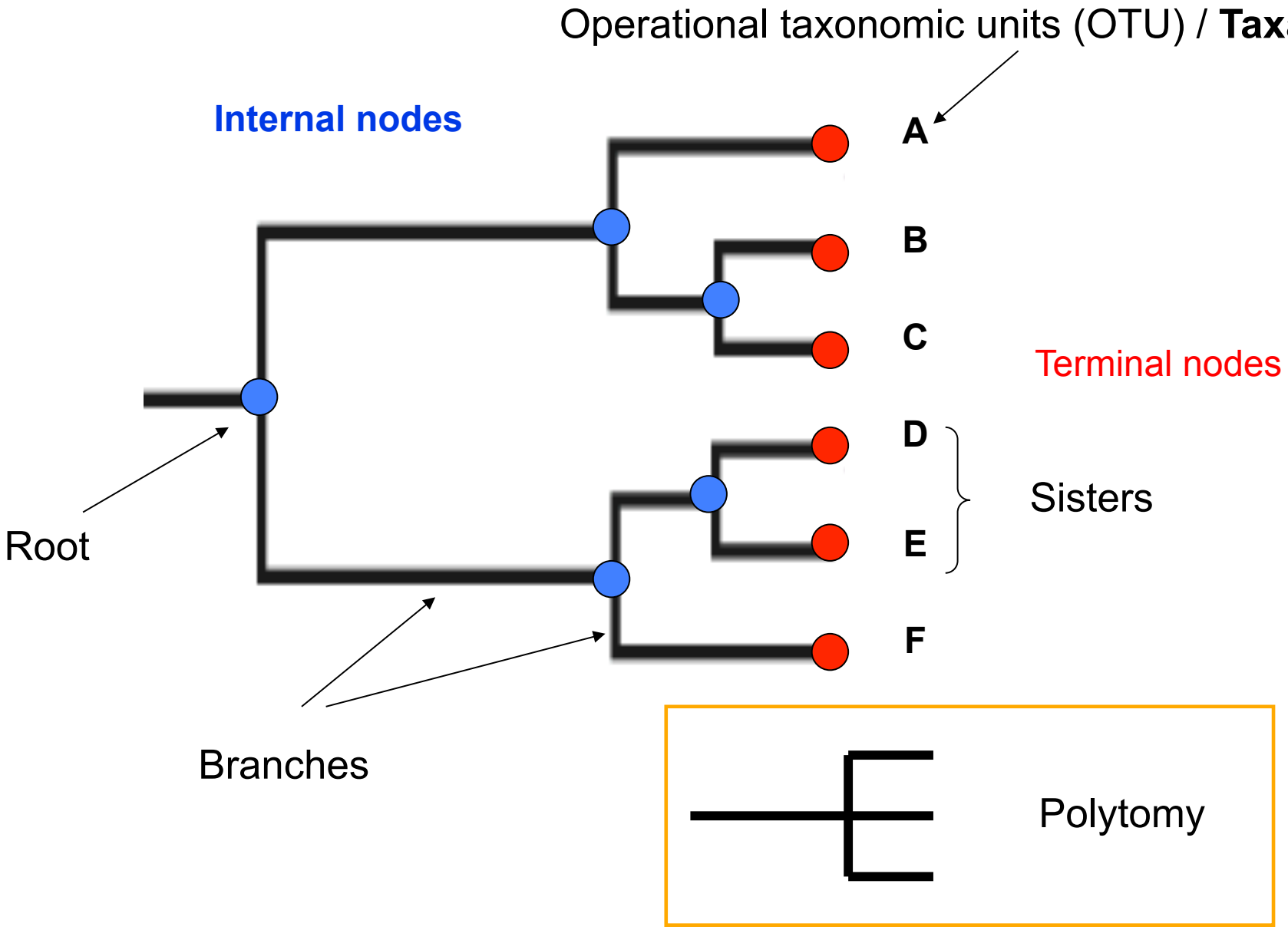


- Same tree - seven different views:

  Rectangular Phylogram, Rectangular Cladogram, Slanted Cladogram, Circular Phylogram, Circular Cladogram, Radial Phylogram and Radial Cladogram

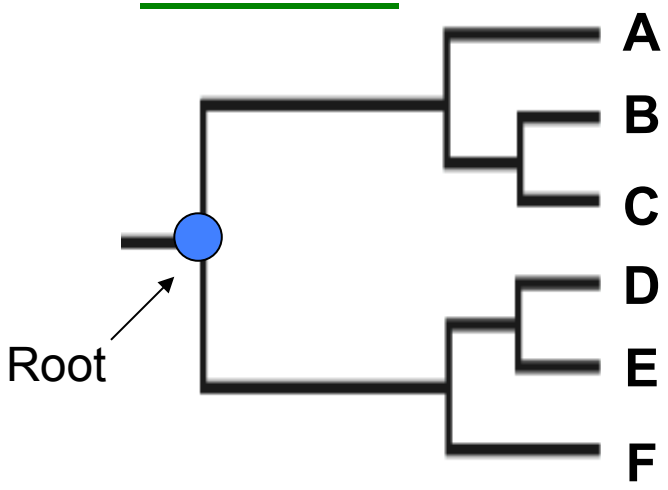


Tree Terminology

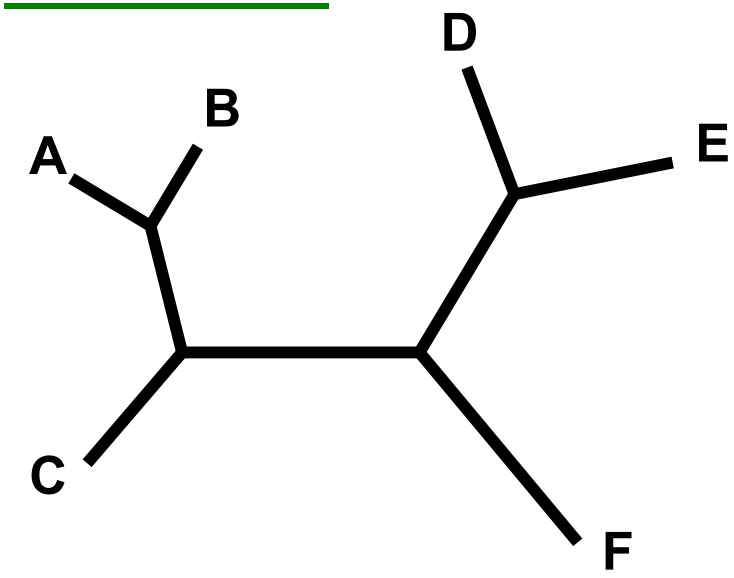


Tree Terminology

Rooted trees



Unrooted trees

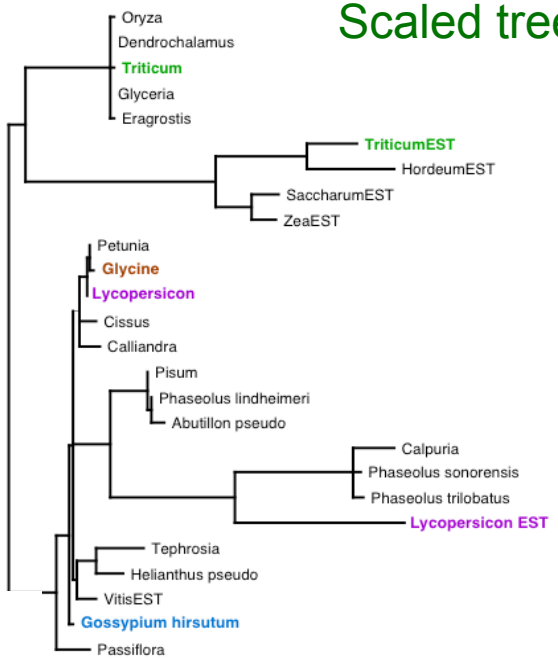


- Rooted trees:
  - Has a root that denotes common ancestry
- Unrooted trees:
  - Only specifies the degree of kinship among taxa but not the evolutionary path

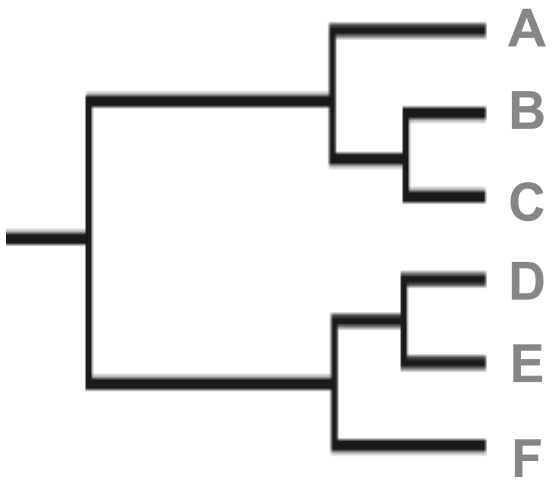
Taxon, plural **taxa**. (taxonomy): Any group or rank in a biological classification into which related organisms are classified.

Tree Terminology

Scaled trees



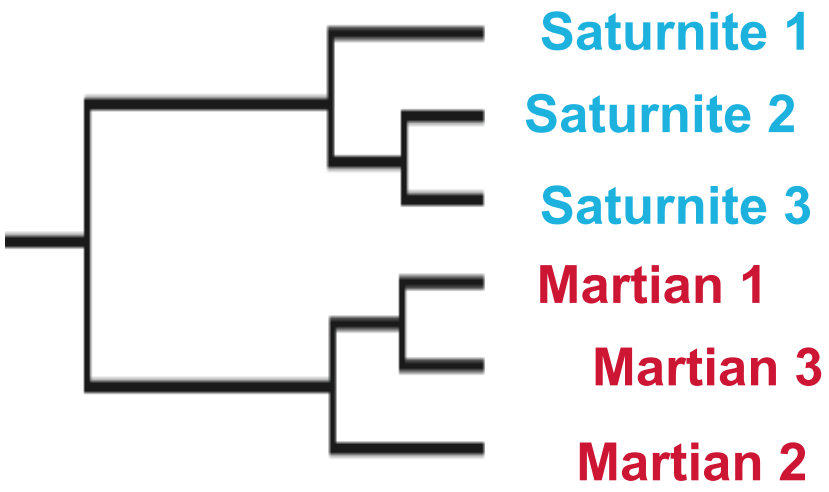
Unscaled trees



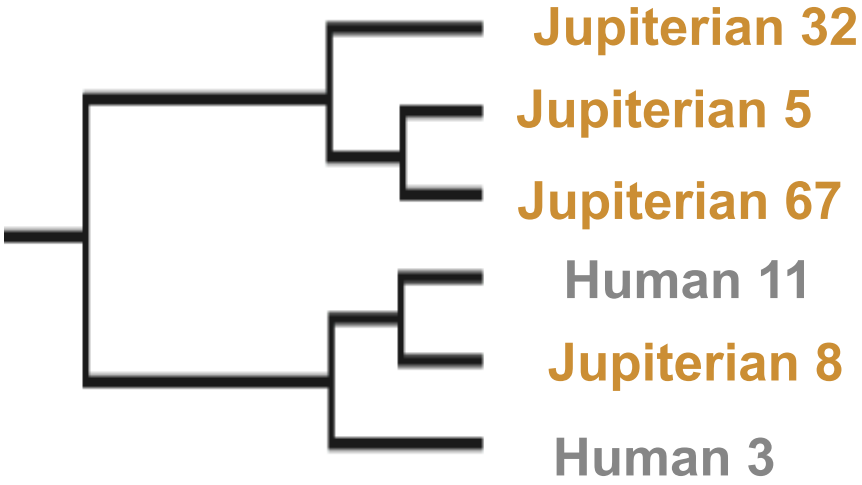
- Scaled trees:
  - Branch lengths are proportional to the number of nucleotide/amino acid changes that occurred on that branch (usually a scale is included).
- Unscaled trees:
  - Branch lengths are not proportional to the number of nucleotide/amino acid changes (usually used to illustrate evolutionary relationships only).

Tree Terminology

Monophyletic groups



Paraphyletic groups



- Monophyletic groups:
  - All taxa within the group are derived from a single common ancestor and members form a natural clade.
- Paraphyletic groups:
  - The common ancestor is shared by other taxon in the group and members do not form a natural clade.

Methods in Phylogenetic Reconstruction

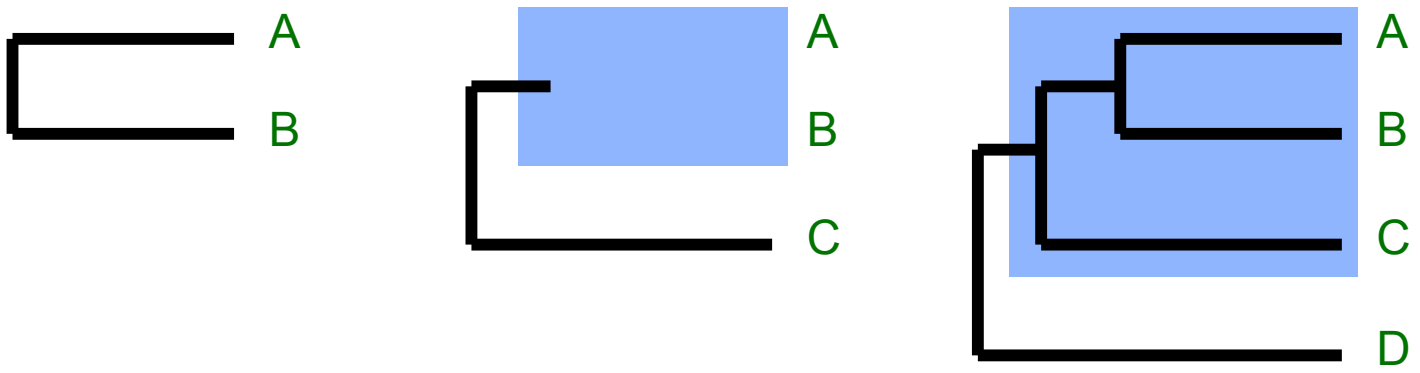
- Distance methods
  - calculate pairwise distances between sequences, and group sequences that are most similar.
  - This approach has potential for computational simplicity and therefore speed
- Maximum Parsimony
  - assumes that shared characters in different entities result from common descent.
  - Groups are built on the basis of such shared characters, and the simplest explanation for the evolution of characters is taken to be the correct, or most parsimonious one.
- Maximum Likelihood
  - compute the probability that a data set fits a tree derived from that data set, given a specified model of sequence evolution.

Comparison of Methods

Distance	Maximum parsimony	Maximum likelihood
• Uses only pairwise distances	• Uses only shared derived characters	• Uses all data
• Minimizes distance between nearest neighbors	• Minimizes total distance	• Maximizes tree likelihood given specific parameter values
• Very fast	• Slow	• Very slow
• Easily trapped in local optima	• Assumptions fail when evolution is rapid	• Highly dependent on assumed evolution model
• Good for generating tentative tree, or choosing among multiple trees	• Best option when tractable (<30 taxa, homoplasy rare)	• Good for very small data sets and for testing trees built using other methods

Methods in Phylogenetic Reconstruction

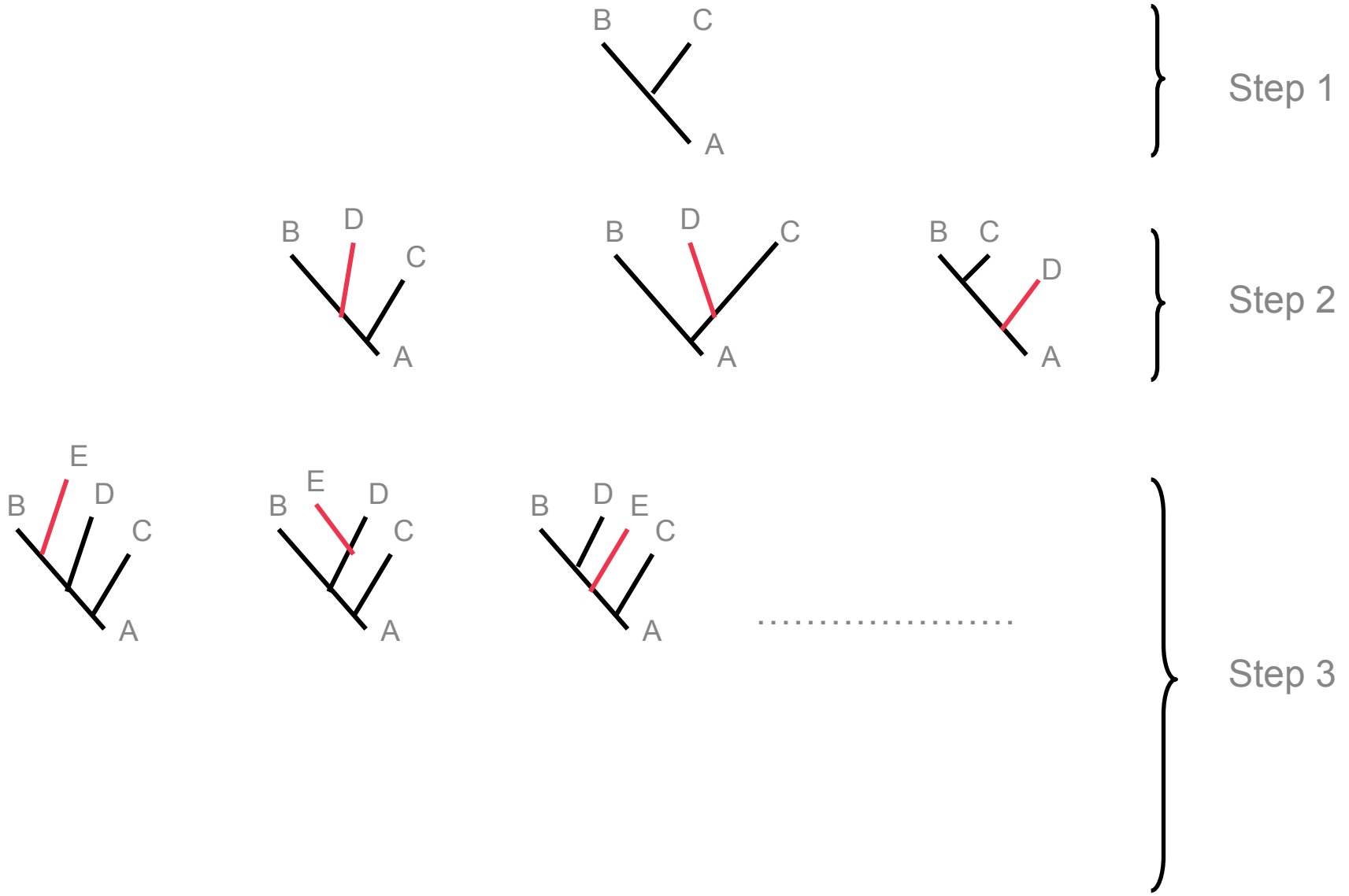
- Distance
  - Using a sequence alignment, pairwise distances are calculated
  - Creates a distance matrix
  - A phylogenetic tree is calculated with clustering algorithms, using the distance matrix.
  - Examples of clustering algorithms include the Unweighted Pair Group Method using Arithmetic averages (UPGMA) and Neighbor Joining clustering.



Methods in Phylogenetic Reconstruction

- Maximum Parsimony
  - All possible trees are determined for each position of the sequence alignment
  - Each tree is given a score based on the number of evolutionary step needed to produce said tree
  - The most parsimonious tree is the one that has the fewest evolutionary changes for all sequences to be derived from a common ancestor
  - Usually several equally parsimonious trees result from a single run.

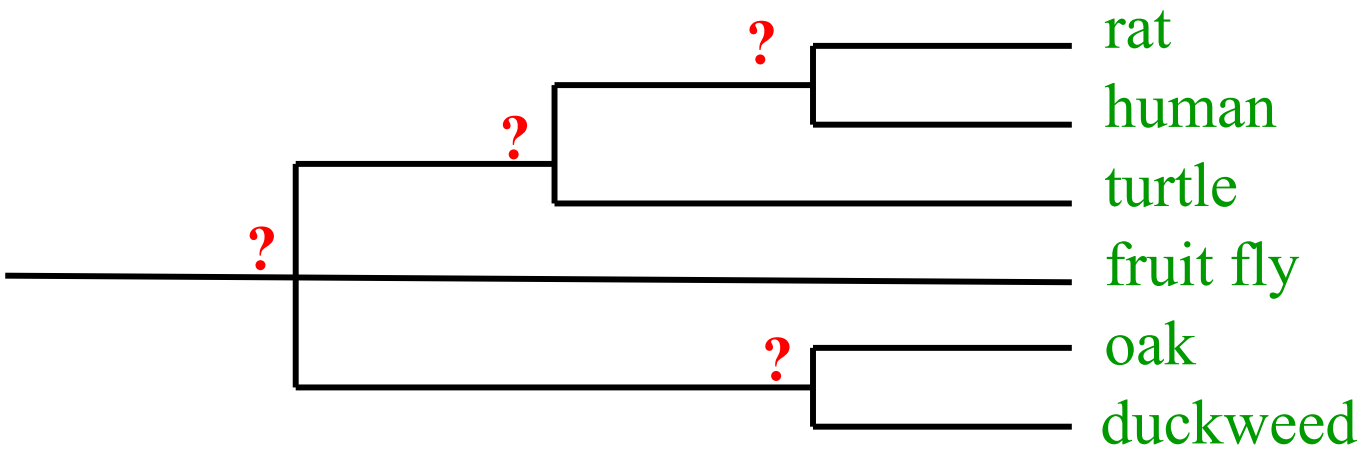
Maximum parsimony: exhaustive stepwise addition



Methods in Phylogenetic Reconstruction

- Maximum Likelihood
  - Creates all possible trees like Maximum Parsimony method but instead of retaining trees with shortest evolutionary steps……
  - Employs a model of evolution whereby different rates of transition/transversion ration can be used
  - Each tree generated is calculated for the probability that it reflects each position of the sequence data.
  - Calculation is repeated for all nucleotide sites
  - Finally, the tree with the best probability is shown as the maximum likelihood tree - usually only a single tree remains
  - It is a more realistic tree estimation because it does not assume equal transition-transversion ratio for all branches.

How confident are we about the inferred phylogeny?



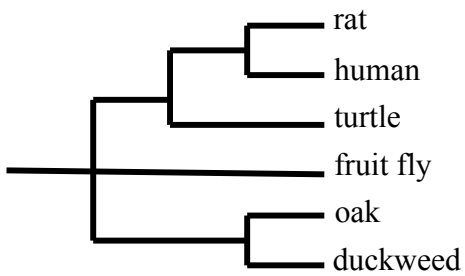
- **Bootstrapping**
- Bootstrap analysis is a kind of statistical analysis to test the reliability of certain branches in the evolutionary tree
- It involves resampling one's own data, with replacement, to create a series of bootstrap samples of the same size as the original data.
- In the case of nucleic acid (amino acid) sequences, the resampled data are the nucleotides (amino acids) of a sequence while the statistical significance of a specific cluster is given by the fraction of trees, based on the resampled data, containing that cluster.

The Bootstrap

- Computational method to estimate the confidence level of a certain phylogenetic tree.

Sample

```
           0123456789
rat        GAGGCTTATC
human      GTGGCTTATC
turtle     GTGCCCTATG
fruitfly   CTCGCCTTTG
oak        ATCGCTCTTG
duckweed   ATCCCTCCGG
```



Inferred tree

Pseudo sample 1

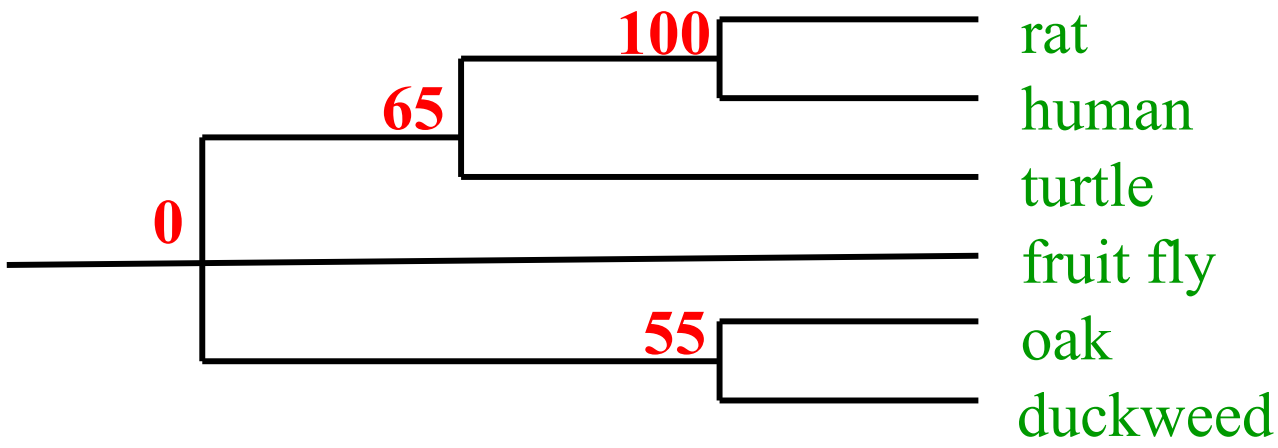
```
           001122234556667
rat        GGAAGGGGCTTTTTA
human      GGTTGGGGCTTTTTA
turtle     GGTTGGGCCCCTTTA
fruitfly   CCTTCCCGCCCTTTT
oak        AATTCCCGCTTCCCT
duckweed   AATTCCCCCTTCCCC
```

Pseudo sample 2

```
           445556777888899
rat        CCTTTTAAATTTTCC
human      CCTTTTAAATTTTCC
turtle     CCCCCTAAATTTTGG
fruitfly   CCCCCTTTTTTTTGG
oak        CCTTTCTTTTTTTGG
duckweed   CCTTTCCCCGGGGGG
```

Many more replicates (between 100 - 1000)

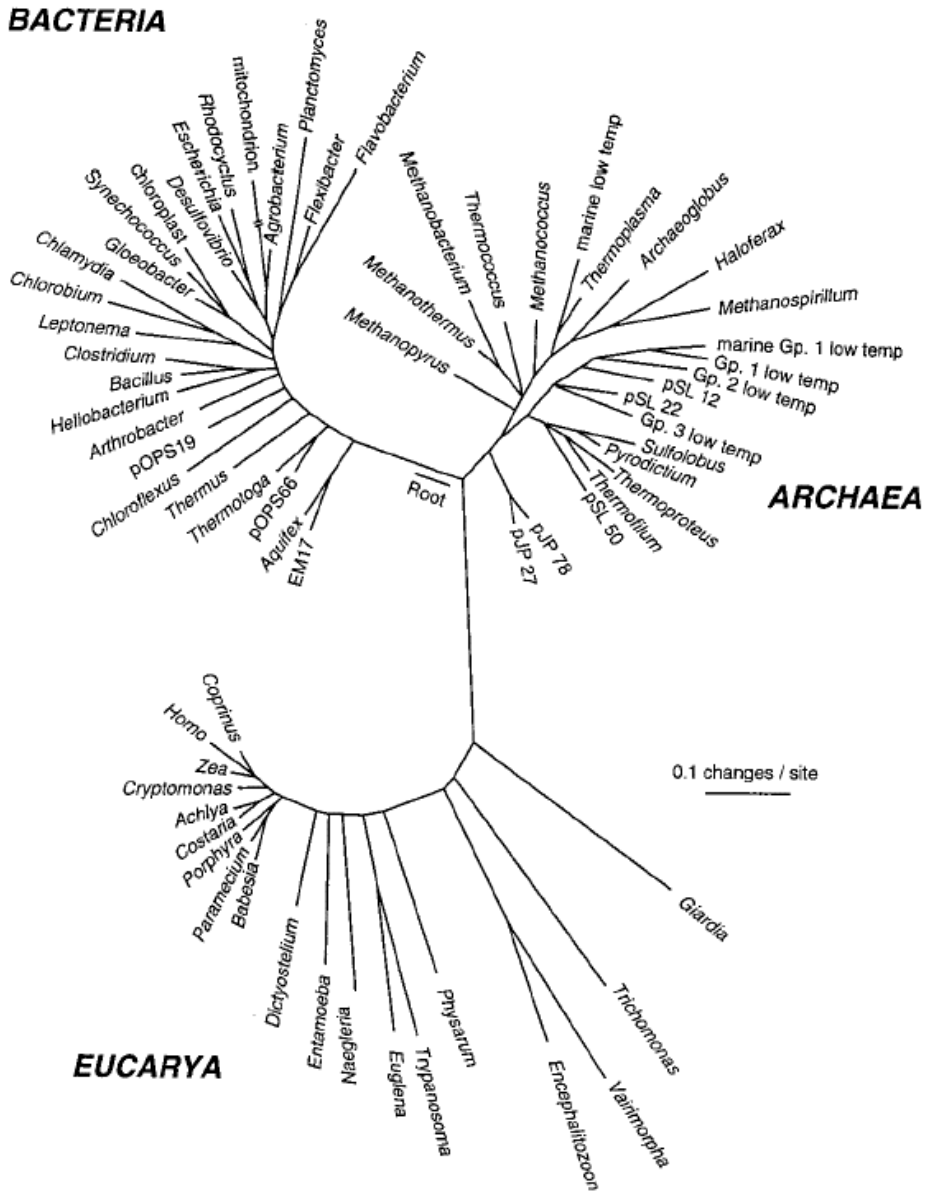
Bootstrap values



- Values are in percentages
- Conventional practice: only values 50-100% are shown

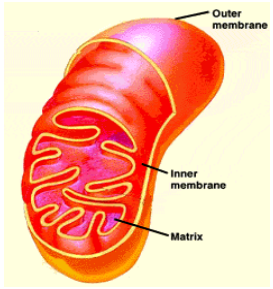
Some Discoveries Made Using Molecular Phylogenetics

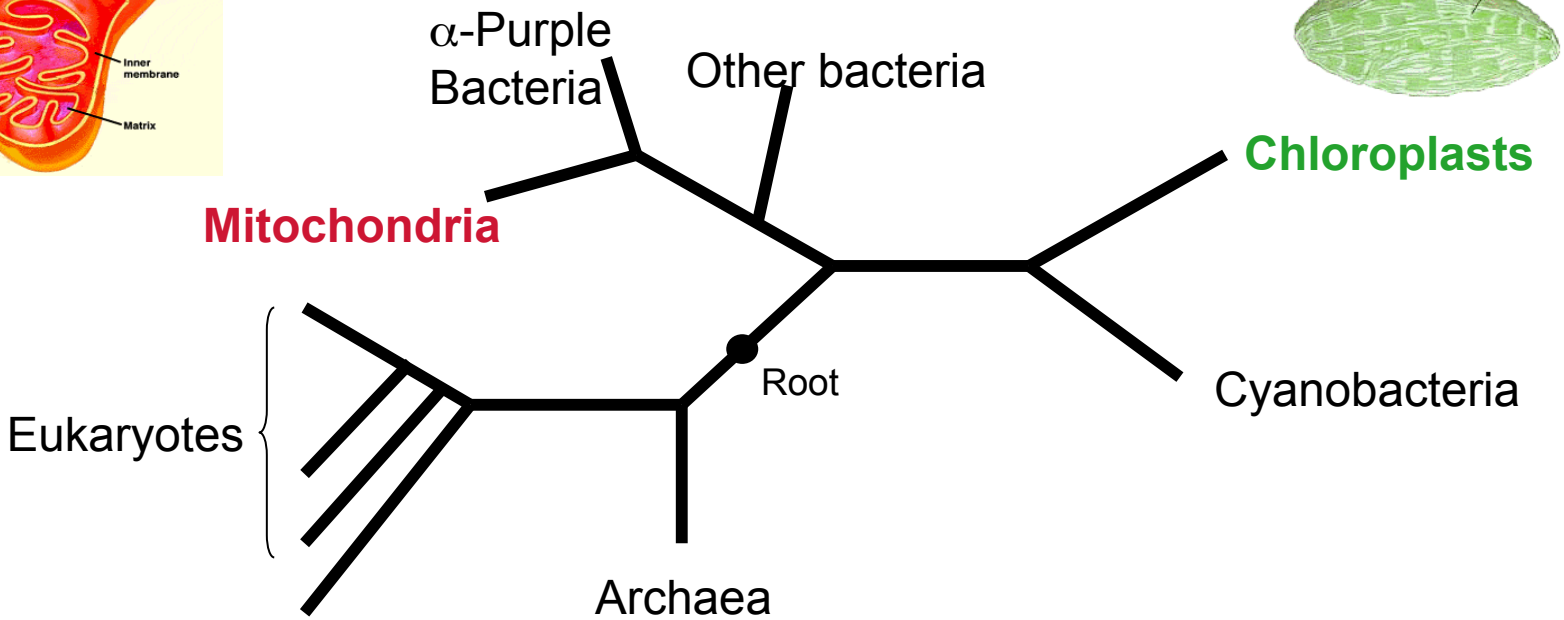
- Universal Tree of Life
  - Using rRNA sequences
  - Able to study the relationships of uncultivated organisms, obtained from a hot spring in Yellowstone National Park



Some Discoveries Made Using Molecular Phylogenetics

- Endosymbiosis: Origin of the Mitochondrion and Chloroplast

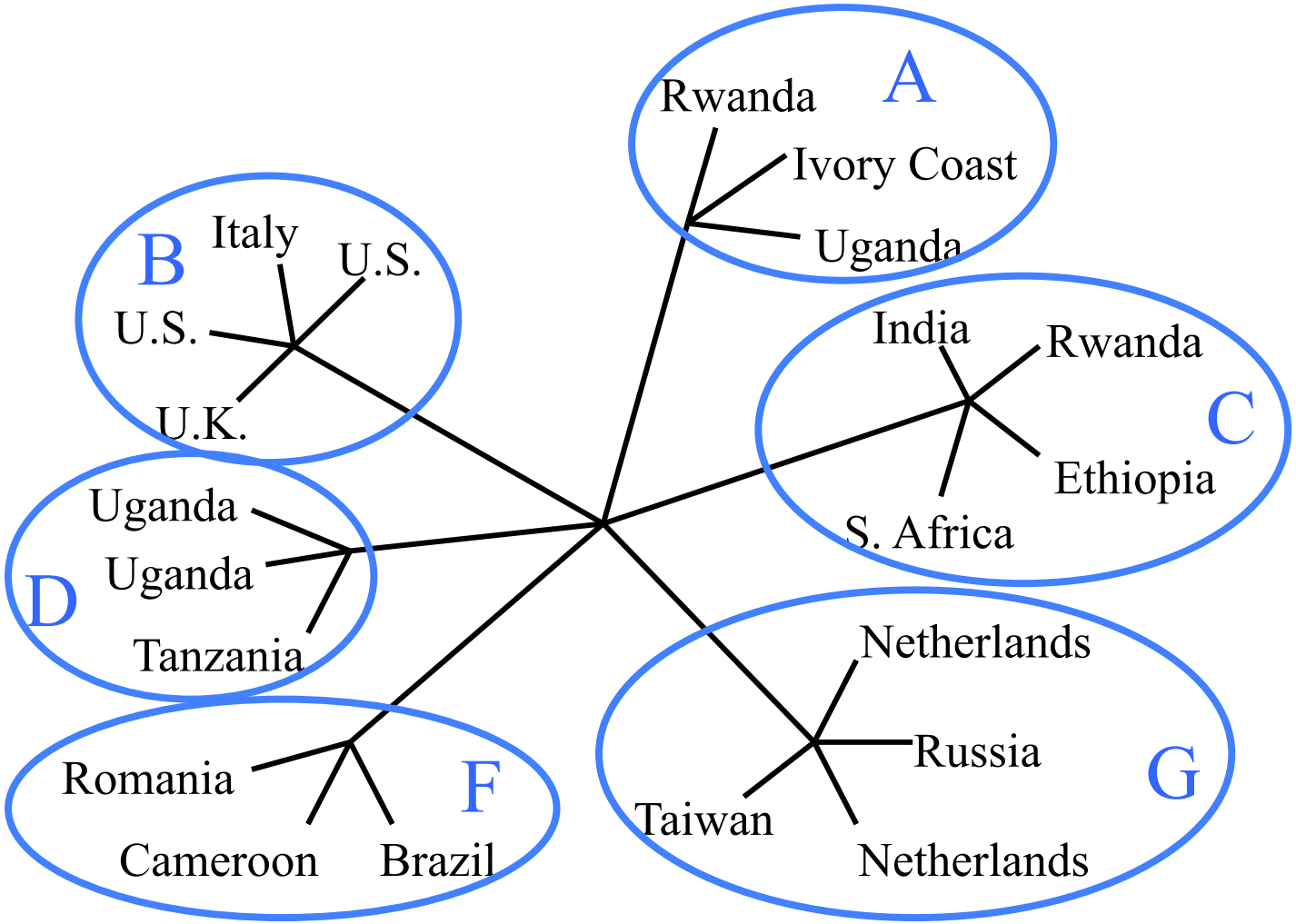




- **Mitochondria** and **chloroplasts** are derived from the α-purple bacteria and the cyanobacteria respectively, via separate endosymbiotic events.

Some Discoveries Made Using Molecular Phylogenetics

- Relationships within species: HIV subtypes



Problems and Errors in Phylogenetic Reconstruction

- Inherent strengths and weaknesses in different tree-making methodologies.
- More is better
  - Errors in inferred phylogeny may be caused by small data sets and/or limited sampling.
- Unsuitable sequences
  - those undergoing rapid nucleotide changes or slow to zero changes overtime may skew phylogenetic estimations

Problems and Errors in Phylogenetic Reconstruction

- Mutations:
  - Duplications, inversions, insertions, deletions etc. can give inaccurate signals
- Genomic hotspots:
  - small regions of rapid evolution are not easily detected
- Homoplasy:
  - nucleotide changes that are similar but occurred independently in separate lineages are mistakenly assumed as inherited changes
- Sample contamination / mislabeling:
  - always a possibility when working with large data sets

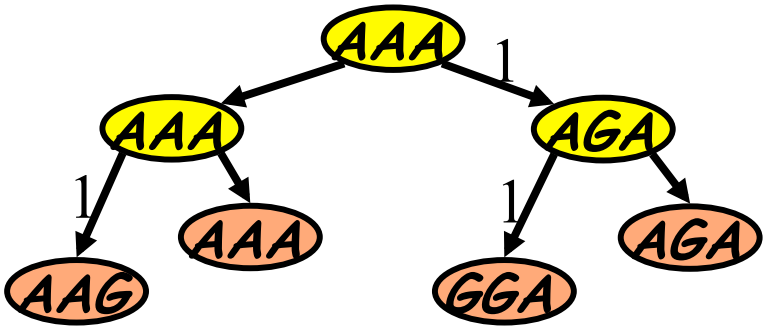
Maximum Parsimony - example

- Maximum parsimony methods predict the evolutionary tree that minimizes the number of steps required to generate the observed variation in the sequences.
  - First, a multiple sequence alignment must be obtained.
- For each aligned position, phylogenetic trees that require the smallest number of evolutionary changes to produce the observed sequence changes are identified.

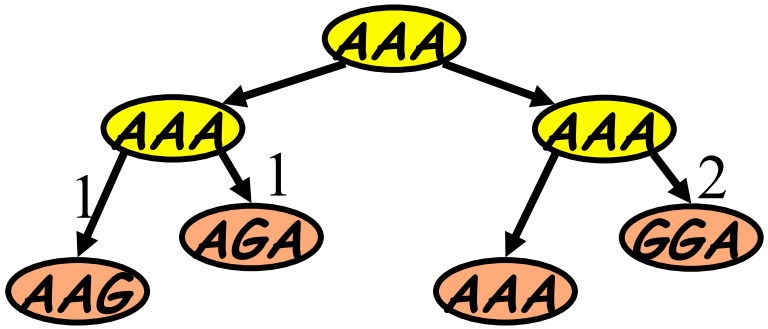
Maximum Parsimony - example

- This continues for each position in the alignment.
- Those trees that produce the smallest number of changes overall for all sequence positions are identified.
  - This is a rather time consuming algorithm that only works well if the sequences have a strong sequence similarity.

Maximum Parsimony - example



Total #substitutions = 3

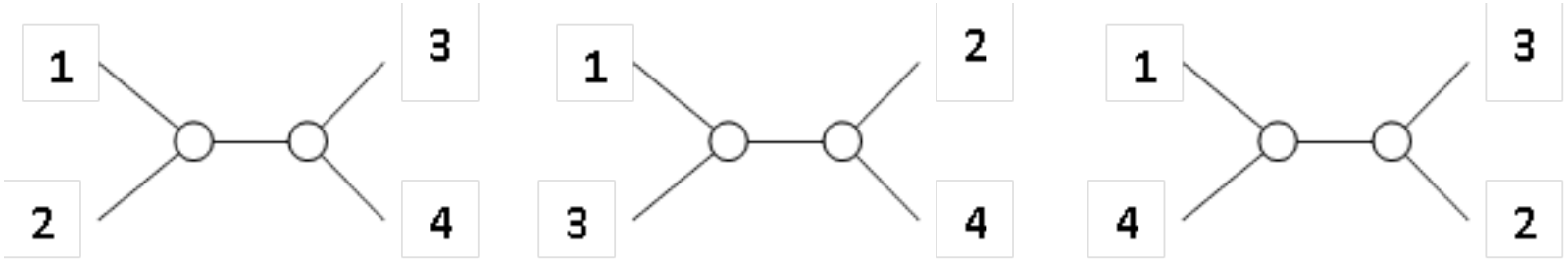


Total #substitutions = 4

- The left tree is preferred over the right tree.

Maximum Parsimony - example

- Assuming we have 4 sequences
  - There are 3 possible trees:



- The optimal tree is obtained by adding the number of changes at each informative site for each tree, and picking the tree requiring the least total number of changes.
  - Informative site, is a site that has at least two characters, each appearing at least in 2 of the sequences of the data set.
- For a large number of sequences the number of trees to examine becomes so large that it might not be possible to examine all possible trees.

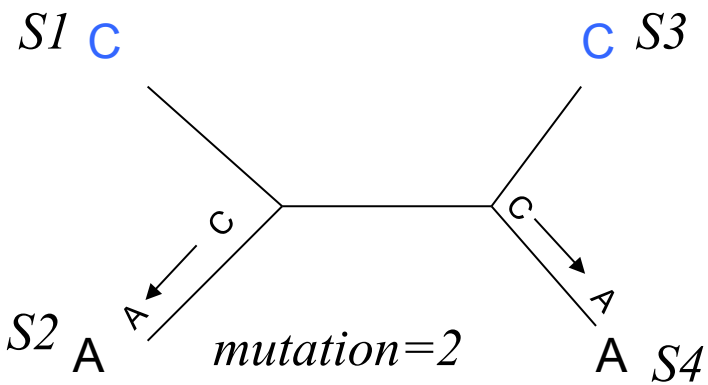
Maximum Parsimony - example

- Consider the following sequences

```
          S1   C A C C C C T T
          S2   A A C C C C A T
          S3   C A C T G C T T
          S4   A A C T G C T A
(S1,S2),(S3,S4)   2 0 0 1 1 0 1 1 6 √
(S1,S3),(S2,S4)   1 0 0 2 2 0 1 1 7
```

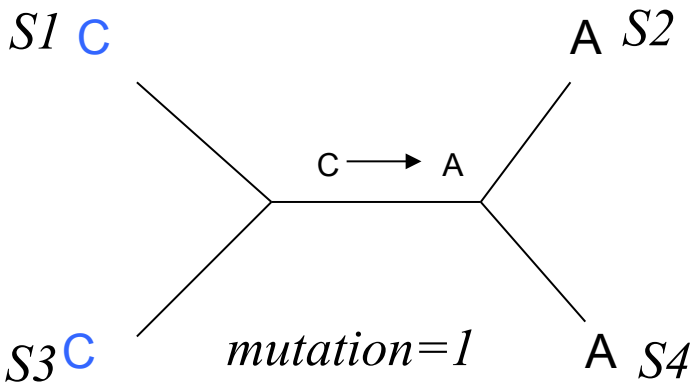
Maximum Parsimony - example

	S1	CACCCCTT	
	S2	AACCCCAT	
	S3	CACTGCTT	
	S4	AACTGCTA	
(S1,S2),(S3,S4)		20011011	**6** √
(S1,S3),(S2,S4)		10022011	**7**



mutation=2

(S1,S2), (S3, S4)



mutation=1

(S1,S3), (S2, S4)

Distance Methods -example

- For phylogenetic analysis, the distance score counted as
  - either the number of mismatched positions in the alignment
  - the number of sequence positions that must be changed to generate the other sequence is used.
- The Fitch and Margoliash method uses a distance table.
  - The sequences are combined in trees to define the branches of the predicted tree and to calculate the branch lengths of the tree.

Distance Methods -example

- Phylogeny reconstruction for 3 sequences
  - There is a single tree topology
  - The branch lengths (a, b, c) :

    $a+b = D_{AB}$

    $b+c = D_{BC}$

    $a+c = D_{AC}$

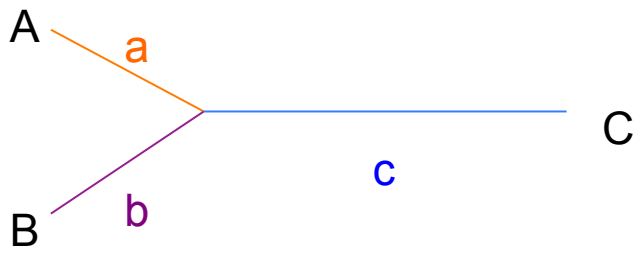
	A	B	C
A	--	a+b	a+c
B	--	--	b+c
C	--	--	--

- Input:
  - D_{AB}, D_{BC} and D_{AC} (pairwise distances)
- Output:

  $a = (D_{AB} + D_{AC} - D_{BC}) / 2$

  $b = (D_{AB} + D_{BC} - D_{AC}) / 2$

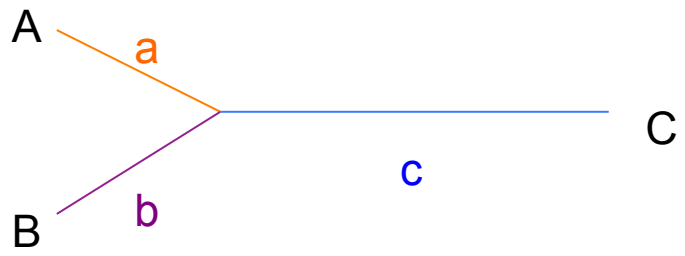
  $c = (D_{AC} + D_{BC} - D_{AB}) / 2$



Distance Methods -example

- Distance matrix of 3 sequences and unrooted tree

	A	B	C
A	--	22	39
B	--	--	41
C	--	--	--



  - distance from A to B = a + b = 22 (1)
  - distance from A to C = a + c = 39 (2)
  - distance from B to C = b + c = 41 (3)
- subtracting (3) from (2) yields:

  b + c – (a + c) = b – a = 41 – 39 = 2 (4)

Distance Methods -example

- adding (1) and (4) yields

  $a + b + b - a = 2b = 22 + 2 = 24$

  $2b = 24$

  $b = 24 / 2 = 12$

- so

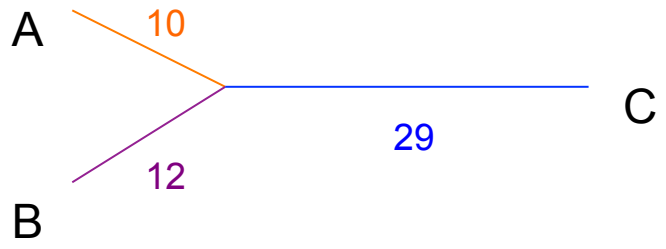
  $a + b = a + 12 = 22;$

  $a = 22 - 12 = 10$

- finally

  $a + c = 10 + c = 39;$

  $c = 39 - 10 = 29$



Distance Methods -example

- Consider the alignment:

```
A   ACGCGTTGGGCGATGGCAAC
B   ACGCGTTGGGCGACGGTAAT
C   ACGCATTGAATGATGATAAT
D   ACACATTGAGTGATAATAAT
```

- The distances between these sequences can be shown as a table:

	A	B	C	D
A	-	3	7	8
B	-	-	6	7
C	-	-	-	3
D	-	-	-	-

Distance Methods -example

- Using this information, an unrooted tree showing the relationship between these sequences can be drawn:

	A	B	C	D
A	-	3	7	8
B	-	-	6	7
C	-	-	-	3
D	-	-	-	-