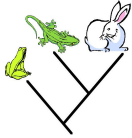


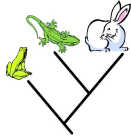
## Phylogenetic Trees

### Five

#### Introduction to Bioinformatics

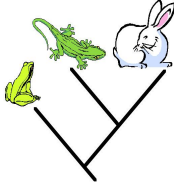


Sami Khuri  
Department of Computer Science  
San José State University  
San José, California, USA  
sami.khuri@sjsu.edu  
June 2016



©2016 Sami Khuri

## Phylogenetic Trees



- ❖ Distance Methods
- ❖ Character Methods
- ❖ Molecular Clock
- ❖ UPGMA
- ❖ Maximum Parsimony
- ❖ Maximum Likelihood
- ❖ Fitch and Margoliash

©2016 Sami Khuri

## Phylogeny Terminology

- **Phylogeny**- the history of descent of a group of organisms from a common ancestor  
From Greek:
  - **phylon** = tribe, race
  - **genesis** = source
- **Taxonomy**- the science of classification of organisms  
From Greek:
  - **taxis** = to arrange, classify

©2016 Sami Khuri

## Phylogeny: Inference Tool

- **Phylogeny** is the inference of evolutionary relationships.
- Traditionally, phylogeny relied on the comparison of morphological features between organisms.
- Today, molecular sequence data are also used for phylogenetic analyses.

©2016 Sami Khuri

## Importance of Phylogeny

- How many genes are related to my favorite gene?
- Was the extinct quagga more like a zebra or a horse?
- Was Darwin correct when he stated that humans are the closest to chimps and gorillas?
- How related are whales and dolphins to cows?
- Where and when did HIV originate?
- What is the history of life on earth?

©2016 Sami Khuri

## Picture of Last Quagga



Died in Amsterdam zoo in 1883.

©2016 Sami Khuri

### Phylogenetic Analysis

- A **phylogenetic analysis** of a family of related nucleic acid or protein sequences is a determination of how the family might have been derived during evolution.
- Two sequences that are very much alike will be located as neighboring outside branches (leaves) and will be joined by a common branch beneath them.

©2016 Sami Khuri

### Aim of Phylogenetic Analysis

- The evolutionary relationships among the sequences are depicted by placing the sequences as outer branches on a tree.
- The branching relationships on the inner part of the tree then reflect the degree to which different sequences are related.
- The **aim of phylogenetic analysis** is to discover all of the branching relationships in the tree and the branch lengths.

©2016 Sami Khuri

### Phylogenetic Trees

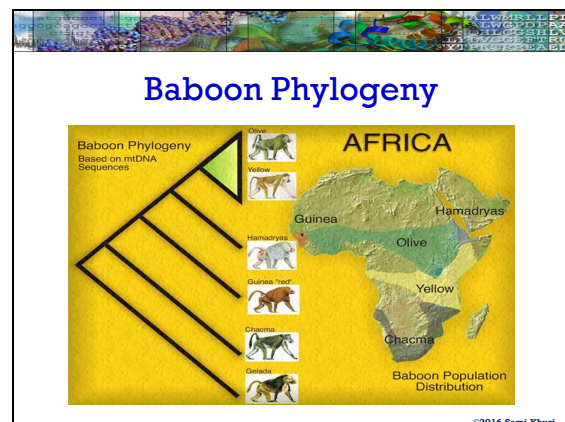
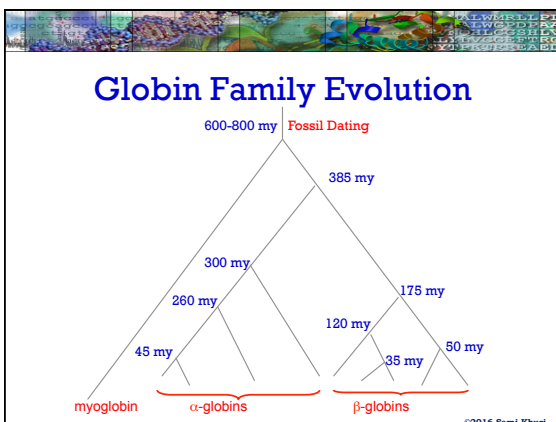
- **Phylogenetic tree**: diagram showing evolutionary paths of species/genes.
- Why do we construct phylogenetic trees?
  - To understand the path (**lineage**) of various species.
  - To understand how various **functions** evolved.
  - To perform **multiple alignment**.

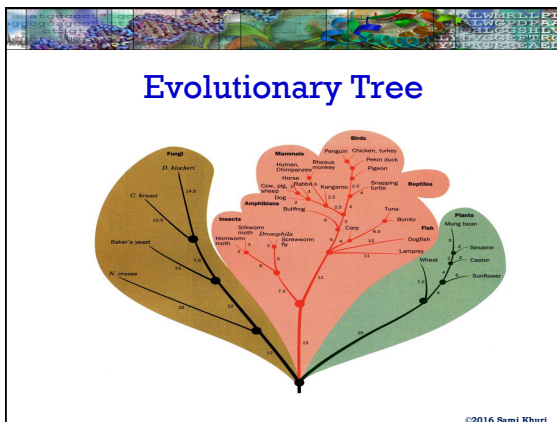
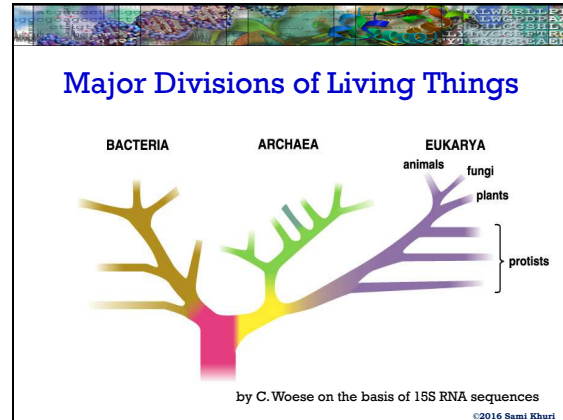
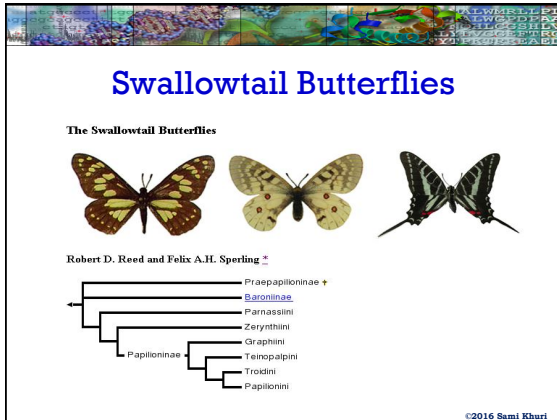
©2016 Sami Khuri

### Additional Uses of Phylogenetic Trees

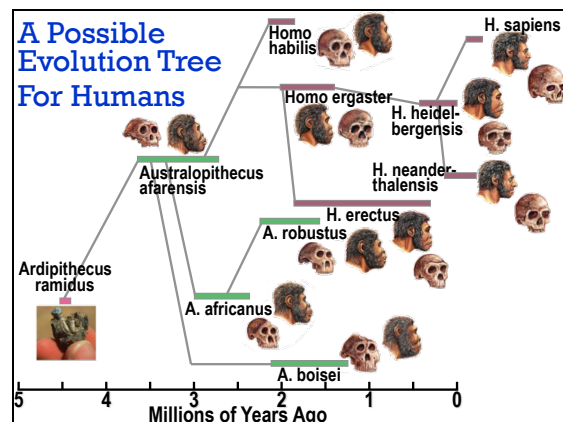
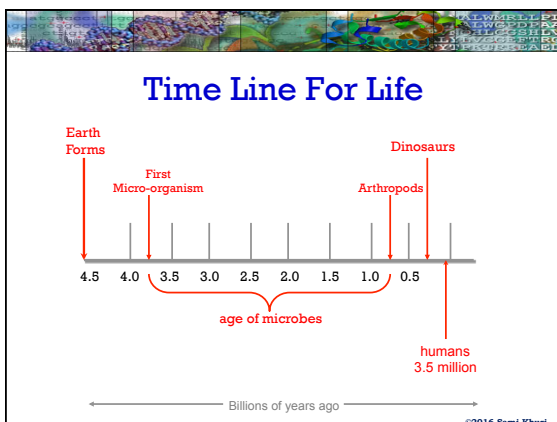
- To study the **evolutionary relationships** of different species and to understand how species relate to one another.
- To **predict** the unknown gene's function according to its phylogenetic relationship to other genes.

©2016 Sami Khuri





- ### Advantages of Molecular Traits
- They directly reflect the underlying process of evolution- changes in the hereditary material
  - There are a vast number of potential traits
  - They can detect differences between very closely related organisms (even those that show no phenotypic difference)
  - They are not affected by the environment (unlike some morphological traits)
  - Since mutations generally occur as random events with specific probabilities, the number of mutations can be used to calibrate evolutionary time (molecular clocks)
- ©2016 Sami Khuri



## More Terminology

- Leaves represent **objects** (genes, species) being compared
  - Taxon** refers to the leaves when they represent species and broader classifications of organisms.
- Internal nodes are hypothetical **ancestral units**
- In a **rooted tree**, the path from root to a node represents an **evolutionary path**.
- An **unrooted tree** specifies **relationships among objects**, but not evolutionary paths.

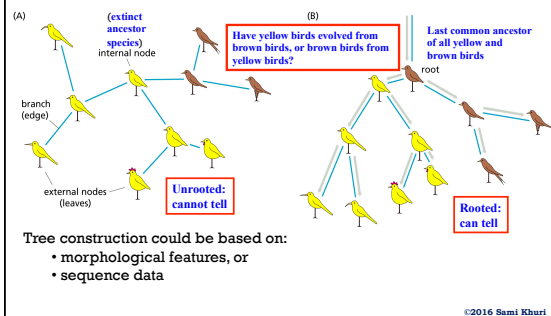
©2016 Sami Khuri

## Rooted and Unrooted Trees

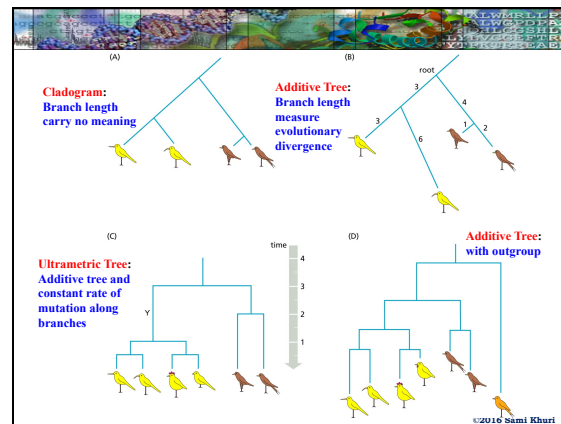
- All objects in a **rooted tree** have a single common ancestor.
  - In general, rooted trees require more information to construct than unrooted ones.
- Objects are leaves in an **unrooted tree** and internal nodes are common ancestors.
  - In general, given any two leaves, we cannot tell if they have a common ancestor.

©2016 Sami Khuri

## Unrooted and Rooted Trees

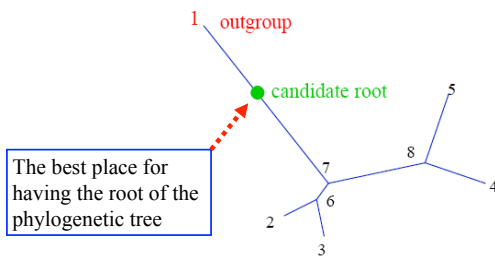


©2016 Sami Khuri



©2016 Sami Khuri

## Rooting a Tree



©2016 Sami Khuri

## Convergent and Parallel Evolution

- Convergent evolution**  
independent evolution of similar traits due to similar selection pressure  
**Example:** wings in birds and bats
- Parallel evolution**- independent evolution of common traits in organisms sharing distant relatives  
**Example:** patterns of butterfly wings.

©2016 Sami Khuri



## Building of a Phylogenetic Tree

- **Sequence Selection:**
  - Identify a DNA or protein sequence.
  - Obtain related sequences by performing a database search.
- Perform multiple alignment.
- Build a phylogenetic tree.
- Check the robustness of the tree.

©2016 Sami Khuri

## Distance and Character Based Trees

The construction of the tree is:

- **distance-based:** measures the distance between species/ genes (eg. mutations, time, distance metric).
  - First calculate the overall distance between all pairs of sequences, then construct a tree based on the distances.
- **character-based:** morphological features (eg. number of legs), DNA/protein sequences.
  - Use the individual substitutions among sequences to determine the most likely ancestral relationships.

The tree is constructed based on the gain or loss of traits.

©2016 Sami Khuri

## Methods for Constructing Phylogenetic Trees

- **Distance-Based Methods:**
  - Unweighted Pair Group Method Using Arithmetic Averages (UPGMA)
  - Fitch Margoliash (FM)
  - Neighbor Joining (NJ)
- **Character-Based Methods:**
  - Maximum Parsimony (MP)
  - Maximum Likelihood (ML)

©2016 Sami Khuri

## Other Methods for Constructing Trees

<p><u>Clustering Methods</u></p> <ul style="list-style-type: none"> <li>• Follow a set of steps (an algorithm) and arrive at a tree.</li> <li>• Use distance data.</li> <li>• Produce a single tree.</li> <li>• Do not use objective functions to compare the current tree to other trees.</li> </ul>	<p><u>Optimality Criterion</u></p> <ul style="list-style-type: none"> <li>• Use objective functions to compare different trees.</li> <li>• First define an optimality criterion, i.e. minimum branch length, and then find the tree with the best value for the objective function.</li> </ul>
---	--

©2016 Sami Khuri

## Clustering Algorithms

- The strength of clustering algorithms is:
  - Their speed
  - Their robustness
  - Their ability to reconstruct trees for very large numbers (thousands) of sequences.
  - Most clustering methods reconstruct phylogenetic trees for a set of sequences on the basis of their pairwise evolutionary distances.

©2016 Sami Khuri

## Classification of Tree Building Methods

		<b>Tree Building Methods</b>	
		Clustering Algorithm	
		Optimality Criterion	
Type of Data	Distance-Based	UPGMA Neighbor Joining	Fitch-Margoliash
	Character-Based		Maximum Parsimony Maximum Likelihood

©2016 Sami Khuri

### Non-graphical Representation of Trees

$= (a,(d,(b,c)))$  or, equivalently  $(a,(d,(c,b)))$

$= ((d,a),(b,(c,e)))$  or  $((b,(e,c),(a,d)))$

©2016 Sami Khuri

### Distance-Based Method

- Given: an  $n \times n$  matrix  $M$ , where  $M(i,j)$  is the distance between objects  $i$  and  $j$
- Build an edge-weighted tree such that the distances between leaves  $i$  and  $j$  correspond to  $M(i,j)$

	A	B	C	D	E
A	0	8	8	5	3
B		0	3	8	8
C			0	8	8
D				0	5
E					0

©2016 Sami Khuri

### UPGMA

- UPGMA is a sequential clustering algorithm.
  - It works by clustering the sequences, at each stage amalgamating two operational taxonomic units (OTUs) and at the same time creating a new node in the tree.
  - The edge lengths are determined by the difference in the heights of the nodes at the top and bottom of an edge.

©2016 Sami Khuri

### The Molecular Clock

- UPGMA assumes that:
  - the gene substitution rate is constant, in other words: divergence of sequences is assumed to occur at the same rate at all points in the tree.
    - Known as the **Molecular Clock**.
  - the distance is linear with evolutionary time.

©2016 Sami Khuri

### Rates of Evolutionary Change

- Different rates throughout genomic DNA base-pair sequence, based mainly on coding.
- ORFs: codon position 3 changes faster than positions 1 and 2.
- Introns change faster than exons.
- Intergenic DNA (especially repeats) changes faster than intragenic (ORF) DNA.
- DNA overall: transition mutations more frequent than transversion mutations.

©2016 Sami Khuri

### UPGMA Algorithm

- The algorithm iteratively picks two clusters and merges them, thus creating a new node in the tree.
- The average **distance** between two clusters is determined by:
 
$$d_{ij} = \frac{1}{|C_i| |C_j|} \sum_{p \in C_i, q \in C_j} d_{pq},$$
 where  $C_i$  and  $C_j$  are clusters.

©2016 Sami Khuri

### The UPGMA Algorithm

- Initialization**
  - Assign each sequence  $i$  to its own cluster  $C_i$ ,
  - Define one leaf of  $T$  for each sequence; place at height zero.
- Iteration while more than two clusters, do**
  - Determine the two clusters  $C_i, C_j$  for which  $d_{ij}$  is minimal.
  - Define a new cluster  $C_k = C_i \cup C_j$ ; compute  $d_{kl}$  for all  $l$ .
  - Define a node  $k$  with children  $i$  and  $j$ ; place it at height  $d_{ij}/2$ .
  - Replace clusters  $C_i$  and  $C_j$  with  $C_k$ .
- Termination**
  - Join last two clusters,  $C_i$  and  $C_j$ ; place the root at height  $d_{ij}/2$ .

©2016 Sami Khuri

### UPGMA: Example (1<sup>st</sup> Iteration)

Sequences A and D are the closest and are combined to create a new cluster V of height  $\frac{1}{2}$  in T.

$d_{ij}$	A	B	C	D	E	F
A	-	6	8	1	2	6
B		-	8	6	6	4
C			-	8	8	8
D				-	2	6
E					-	6

Understanding Bioinformatics by M. Zvelebil and J. Baum  
©2016 Sami Khuri

### UPGMA: Example (2<sup>nd</sup> Iteration)

The table of distances is updated to reflect the average distances from V to the other sequences. V and E are the closest and are combined to create a new cluster W of height 1 in T.

$d_{ij}$	B	C	E	F	V
B	-	8	6	4	6
C		-	8	8	8
E			-	6	2
F				-	6

Understanding Bioinformatics by M. Zvelebil and J. Baum  
©2016 Sami Khuri

### UPGMA: Example (3<sup>rd</sup> Iteration)

After updating the table of distances, B and F are the closest sequences and are combined to create a new cluster X of height 2 in T.

$d_{ij}$	B	C	F	W
B	-	8	4	6
C		-	8	8
F			-	6

Understanding Bioinformatics by M. Zvelebil and J. Baum  
©2016 Sami Khuri

### UPGMA: Example (4<sup>th</sup> Iteration)

Once more the table is updated. W and X are the closest sequences and are combined to create a new cluster Y of height 3 in T.

$d_{ij}$	C	W	X
C	-	8	8
W		-	6

Understanding Bioinformatics by M. Zvelebil and J. Baum  
©2016 Sami Khuri

### UPGMA: Example (Termination)

The remaining 2 sequences, C and Y of distance 8 are combined to create a new cluster Z of height 4 in T.

Understanding Bioinformatics by M. Zvelebil and J. Baum  
©2016 Sami Khuri

### UPGMA Tree: Second Example

	M	N	O	P
M	-	15	26	28
N	-	-	29	31
O	-	-	-	12
P	-	-	-	-

UPGMA assumes a **uniform rate of mutation** in the tree branches. At any given time, the two sequences should have the same number of changes separating them from the common ancestor.

Bioinformatics by David Mount  
©2016 Sami Khuri

### UPGMA's Shortcoming

A tree where closest pair of leaves are not neighboring leaves.

$d(A,C) = 0.3$   
 $d(A,B) = 0.5$

So the **neighboring** pair A and B are further apart than the **non-neighboring** pair A and C.

©2016 Sami Khuri

### Fitch-Margoliash Method

- Fitch-Margoliash does not assume a constant mutation rate.
- With the **Fitch-Margoliash Method**, the sequences are combined in threes to define the branches of the predicted tree and to calculate the branch lengths of the tree.
- This method of averaging distances is most accurate for trees with short branches.

©2016 Sami Khuri

### Introduction to Neighbor-Joining

- Neighbor -Joining does not assume a constant rate of evolution.
- The algorithm is based on the concept of minimum evolution; the true tree is the one for which the total branch length is minimum.
- The resulting tree is not rooted and is additive.

©2016 Sami Khuri

### Limitations of Distance-Based Phylogenetic Trees

The **distance-based phylogenetic tree** is derived from the pairwise distance of aligned sequences and not from the original sequence data. The distance information may not contain all the sequence information.

©2016 Sami Khuri

### Observable Features

- Sometimes we do not have a distance metric between the species we are interested in.
- What we have instead, are **observable features**.
- We then use the **observable features** to build the tree. These trees are called **Character-Based trees**.

©2016 Sami Khuri



## Character-Based Trees

- The building of the tree is based on **morphological features** and not on distances.
- Examples of **morphological features**:
  - has feathers
  - has a backbone
  - has a certain amino acid at a certain position in the sequence
  - whether or not a certain protein regulates another protein.

©2016 Sami Khuri

## Maximum Parsimony Method

- The **maximum parsimony** method predicts the evolutionary tree that minimizes the number of steps required to generate the observed variation in the sequences
  - This method is also known as the **minimum evolution method**.
- The **maximum parsimony** method is used
  - for sequences that are quite similar, and
  - for small number of sequences.

©2016 Sami Khuri

## Maximum Parsimony

- **Maximum parsimony** means fewest evolutionary changes necessary to explain observed taxonomic relationships.
- Fewest postulated steps in evolutionary process.
- Leads to predictions for common ancestor and branch-point ancestors.
- Exhaustive search of trees is possible only for small number of species.

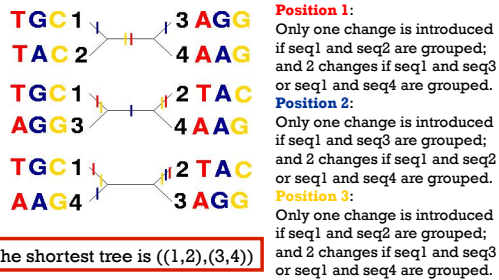
©2016 Sami Khuri

## Parsimony: An Example

- Given four sequences:
  - Sequence 1: TGC
  - Sequence 2: TAC
  - Sequence 3: AGG
  - Sequence 4: AAG
- We want to find the tree with the smallest number of changes that explains the observed data.
- Draw all possible trees with 4 taxa.

©2016 Sami Khuri

## Parsimony Example



©2016 Sami Khuri

## Informative Sites

- A site that provides information for distinguishing between different topologies is said to be an **informative site**.
- Only **informative sites** need to be analyzed.
- A site is phylogenetically informative only when there are at least two different kinds of characters, each represented at least two times.

©2016 Sami Khuri

### Informative Sites: An Example

Taxa	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G

Only sites at columns 5, 7, and 9 are informative sites

©2016 Sami Khuri

### Steps of Tree Reconstruction

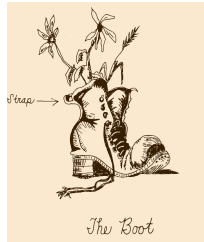
```

    graph TD
      A[Choose a set of related sequences] --> B[Obtain a multiple sequence alignment]
      B --> C{Is there strong sequence similarity?}
      C -- Yes --> D[Maximum Parsimony methods]
      C -- No --> E{Is there a clearly recognizable sequence similarity?}
      E -- Yes --> F[Distance methods]
      E -- No --> G[Maximum Likelihood methods]
      G --> H[Analyze how well the data support prediction]
  
```

David Mount  
©2016 Sami Khuri

### Bootstrapping

Bootstrapping is a statistical technique that uses computer intensive random resampling of data to determine sampling error or confidence intervals for some estimated parameter.



©2016 Sami Khuri

### Bootstrap: An Example (I)

Species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1 Pre (Chimp)	C	T	T	G	A	G	A	A	A	T	T	C	T	T	A	G	A	T	A	
2 Pme (Lizard)	T	C	T	A	A	A	G	A	A	T	T	A	T	A	T	A	G	A	T	A
3 Pma (Human)	T	T	T	A	A	G	G	A	A	A	T	T	C	T	T	A	A	A	T	T
4 Pfa (Human)	T	T	T	G	A	G	A	A	A	A	T	T	C	T	T	A	G	A	T	A
5 Pbe (Rodent)	T	T	T	A	G	A	A	A	A	A	T	T	A	T	A	A	A	T	A	A
6 Plo (Bird)	T	T	T	A	G	A	A	A	A	C	T	C	A	C	A	A	A	T	C	
7 Pfr (Monkey)	C	T	T	A	G	A	G	A	A	T	T	C	T	T	A	G	A	A	A	
8 Pkn (Monkey)	C	T	T	A	G	A	A	A	A	G	T	T	C	T	T	A	G	A	T	A
9 Pcy (Monkey)	C	T	C	A	T	G	A	A	A	A	T	T	C	T	T	A	G	A	T	A
10 Pv (Human)	C	T	T	A	T	G	A	A	A	A	T	C	T	C	G	G	A	T	A	
11 Pga (Bird)	T	T	T	A	G	A	A	A	A	T	T	T	T	C	A	A	A	T	C	

Part of the data matrix of aligned nucleotide sequences for the malaria parasite *Plasmodium*. Only the first 20 columns of the 11 × 221 matrix are shown.

Efron, Bradley et al. (1996) Proc. Natl. Acad. Sci. USA 93, 13429  
©2016 Sami Khuri

### Bootstrap: An Example (II)

Randomly select, with replacement, 221 columns from the original matrix. Tree-building algorithm is applied to give a bootstrap tree. Repeat the process 200 times.

The numbers at the branches are confidence values based on Felsenstein's bootstrap method.

Efron, Bradley et al. (1996) Proc. Natl. Acad. Sci. USA 93, 13429  
©2016 Sami Khuri

### Software Tools

- **PHYLIP**
  - Phylogeny Inference Package.
  - <http://evolution.genetics.washington.edu/phylip.html>
  - Free.
  - Developed by Dr. Joe Felsenstein from the Department of Genome Sciences at the University of Washington.
  - Source code is written in ANSI C.
  - Executables are available for different platforms:
    - Windows, UNIX and Macintosh.
- **PAUP\***
  - Phylogenetic Analysis Using Parsimony.
  - <http://www.lms.si.edu/PAUP/about.html>
  - Developed by Dr. David Swofford of the Laboratory of Molecular Systematics, National Museum of Natural History
  - The most sophisticated parsimony program.

©2016 Sami Khuri