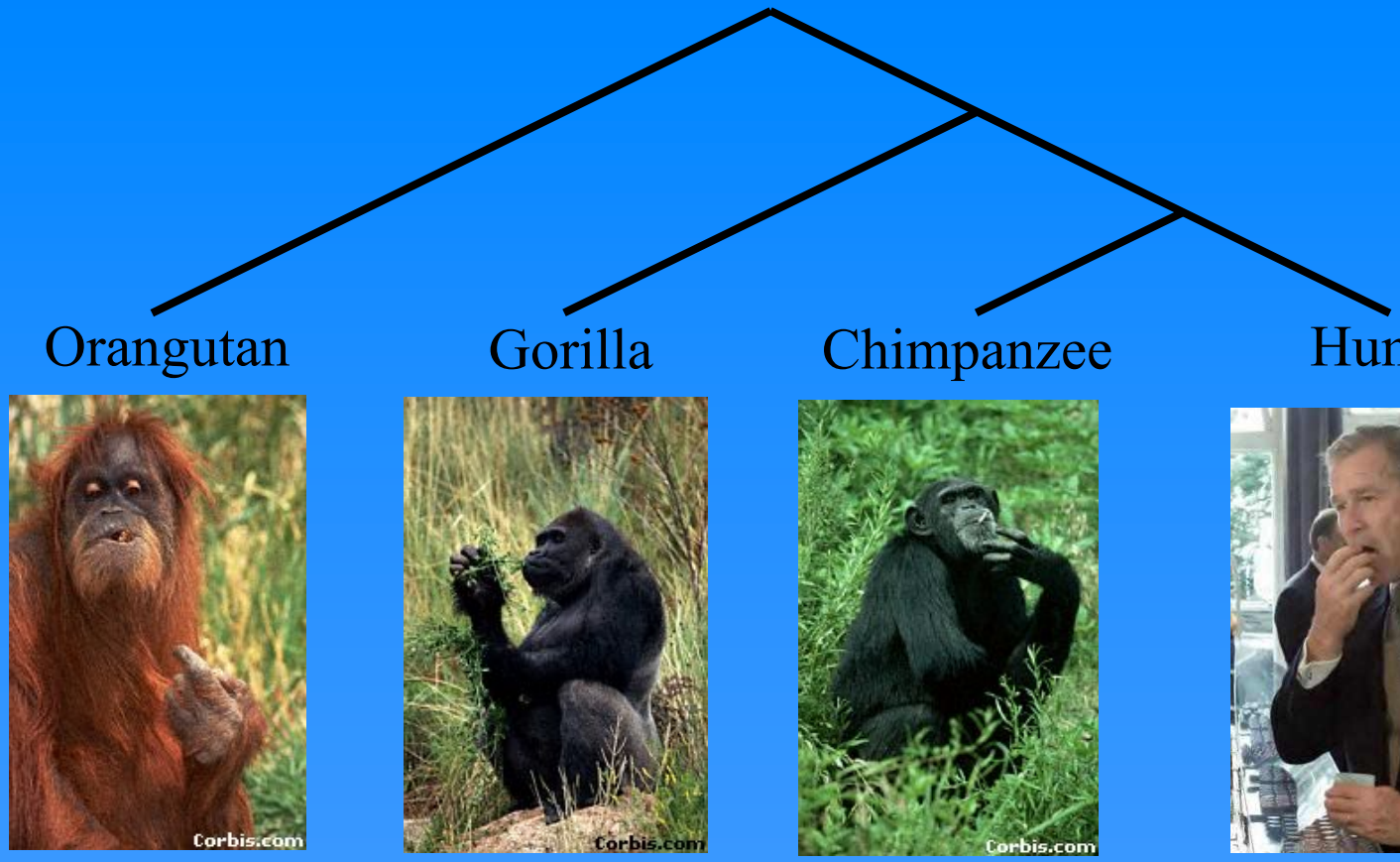
Introduction to Phylogenetics I



From the Tree of the Life Website, University of Arizona

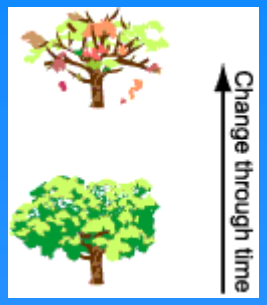
Sagi Snir

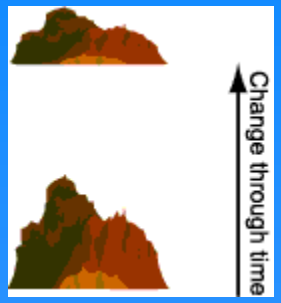
Dept. of Evol. Env. Biol. and The Inst. of Evolution,

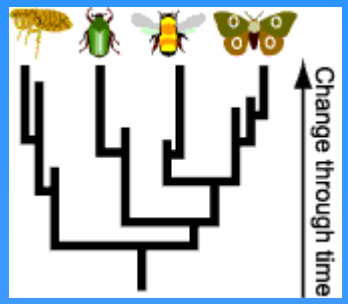
University of Haifa

Introduction to Phylogenetics

background and basic concepts

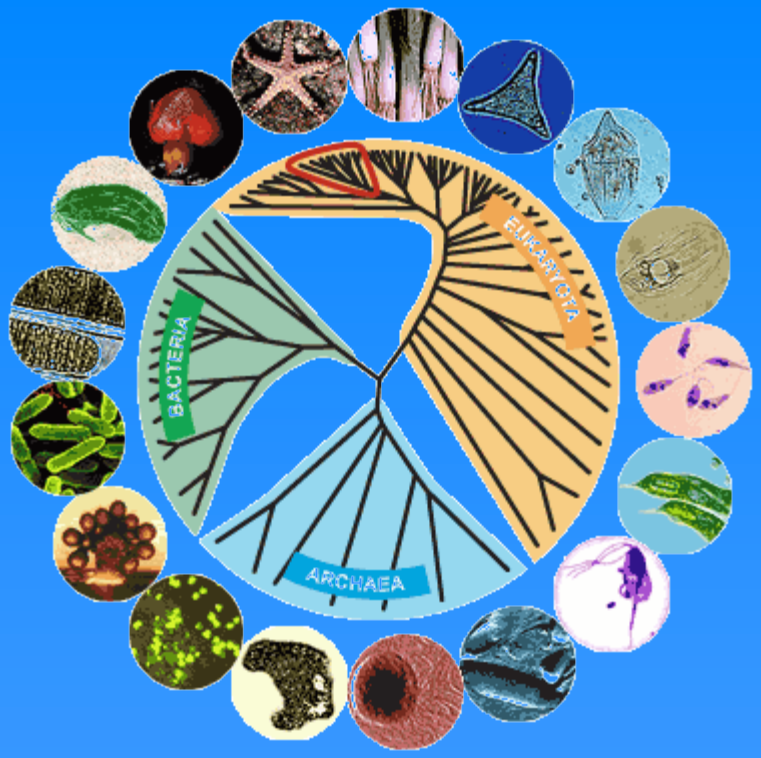






1. Biological Evolution: inheritance through changes
    1. Within a species – evolution through generations
    2. Between species – evolution from a common ancestor
2. In contrast to arbitrary changes through time (geological, ecological), changes occur through genetic inheritance
3. Underlying assumption: universal common ancestor
4. Via a long continual process of changes, the current diversity of life was formed.

עץ תורשה



The three domains

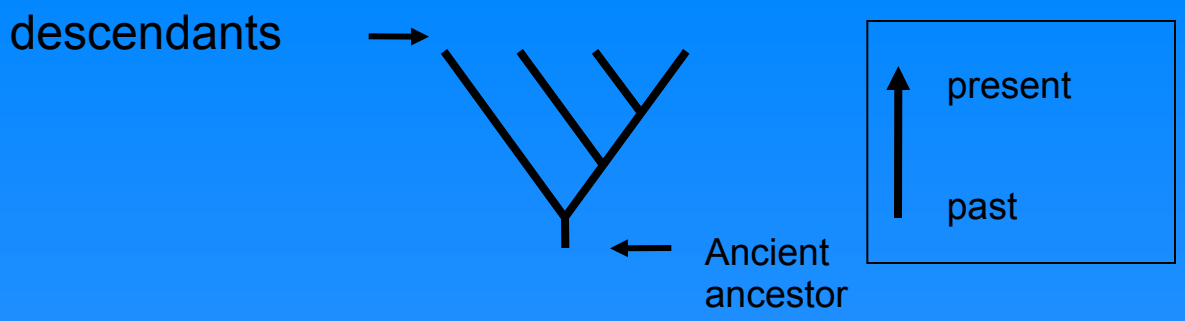
The single common ancestor property creates a tree like structure

By studying common properties to species, we can hypothesize the evolutionary history of there species

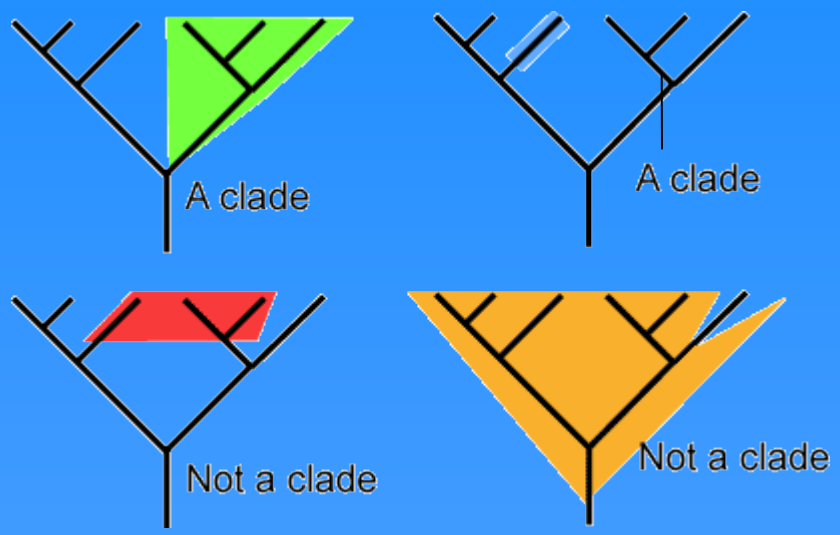
This history is merely a hypothesis

The tree in the figure represents the primordial division of life into three kingdoms

Understanding Phylogenies



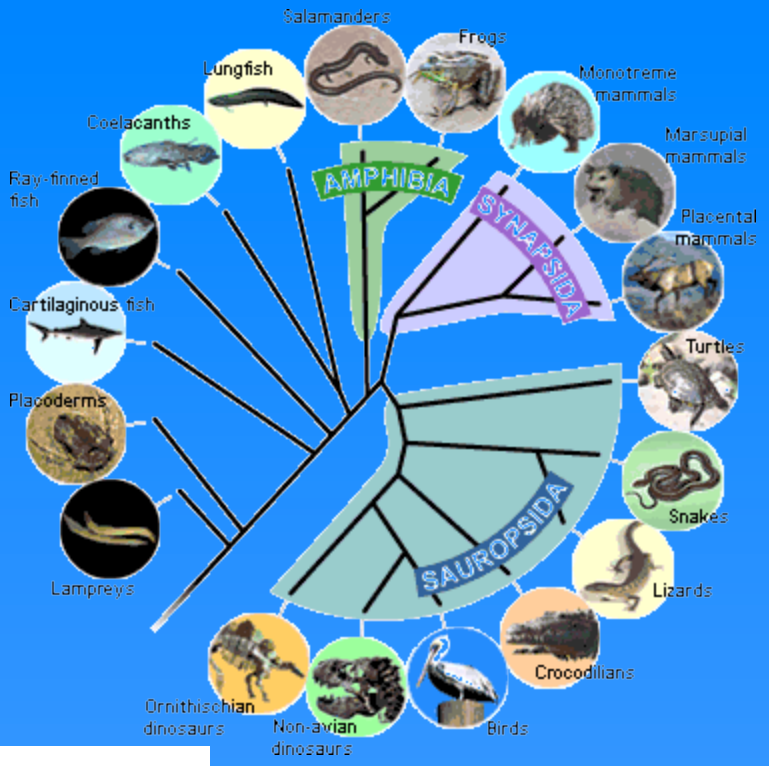
1. Development through time

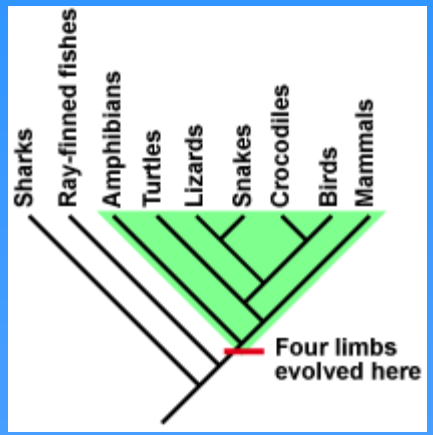


1. A clade and its members

Constructing the tree

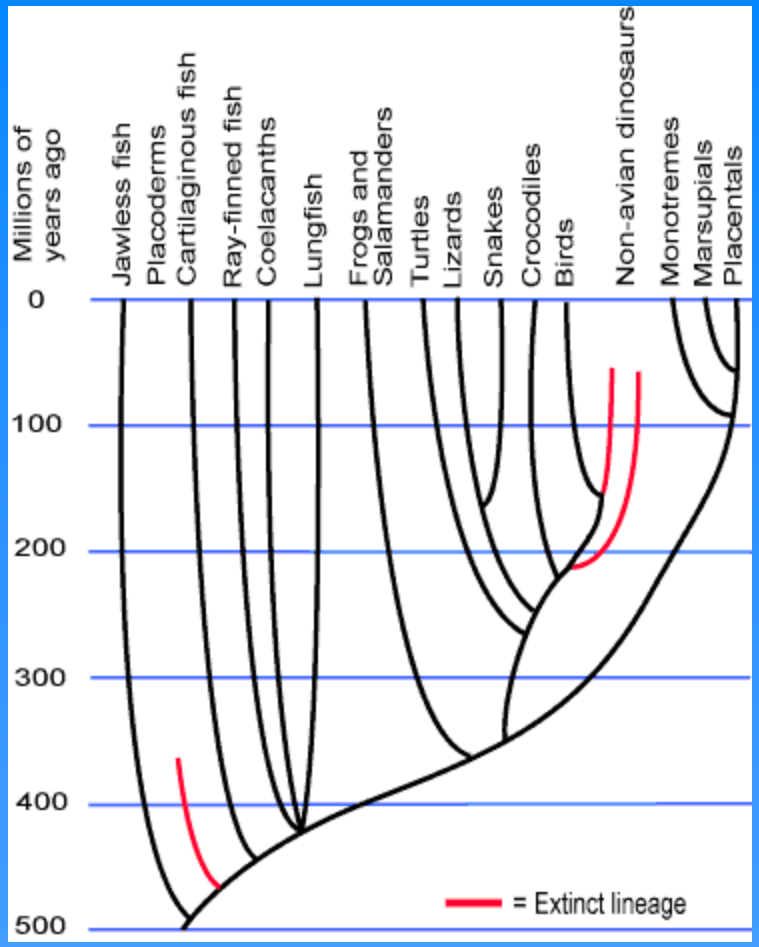
- In order to construct the evolutionary tree, we need to collect common properties in all species under investigation
- In order to construct the tree, we need to find properties common to all organisms.
- If we find a characteristic distinguishing some organisms from the others, we can use it for our classification.





What about Time?

1. Biologists tend to mark time on the tree by assigning lengths to branches proportional to the respective period length
2. Branches that have not survived, terminate before present time



The Evolutionary process

1. The basis to Evolution are genetic changes on which the various forces operate
2. Evolutionary Mechanisms:
    1. Mutations – crates individuals with different genetic code
    2. Migrations – individuals from one population move to a new environment
    3. Natural Selection – traits in one species become advantageous over another species

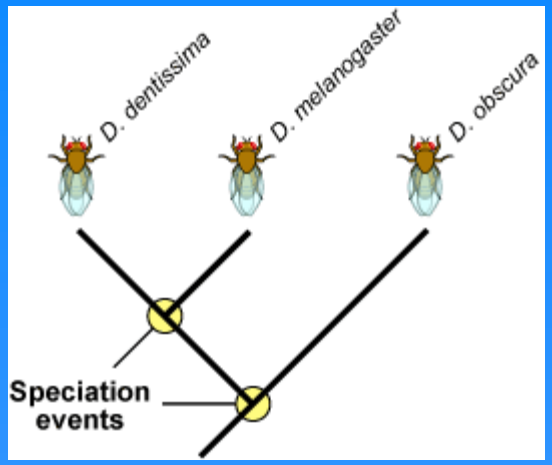
Mutations

1. Changes in DNA – the genetic code of every living organism
2. Mutations are random
3. Not all mutations are meaningful – only those passing to offspring play an evolutionary role
4. These occur in germlines
5. Some are lethal, some are synonymous, and some survive
6. Their source are errors in cell replication

Natural Selection

1. Some traits are beneficial and cause a group of organisms to survive
2. In rare cases selection is immediate and observable
3. The finches in the Galapagos developed strong beak as a result of a succession of droughts.

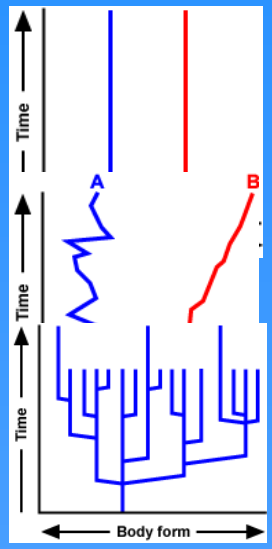
speciation



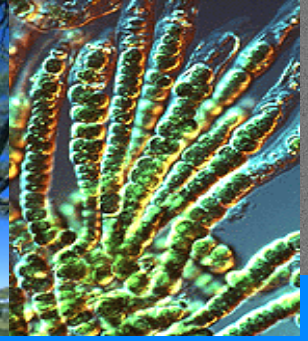
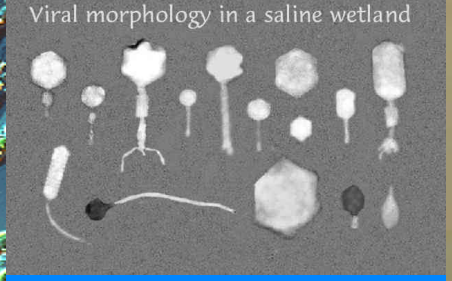
1. A species – group of individuals capable of inter breeding.
2. A speciation event is a point in time where a single species splits into two or more.
3. Causes of speciation:
4. Geographical separation – continents split, a river changes its trajectory, a mountain rises
5. A decrease in gene mobility – a wider span of a population
6. Specialization in different niches

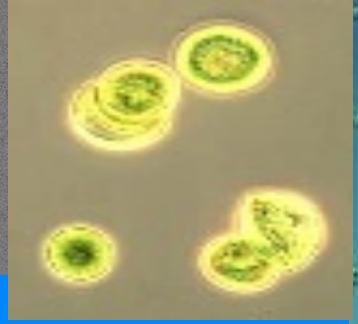
Macro-Evolution

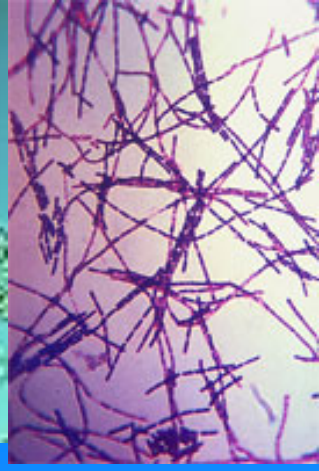
Mutation
Gene Flow
Genetic Drift
Natural Selection
+ 3.8 billion years = Macroevolution



1. The major processes led to the creation of the Tree of Life:
    1. Point mutations accompanied with natural selection, genetic drift, over 3.8 billion years
2. Forms of Evolution:
    1. Freeze – no change over time
    2. Directionality – traits appear and disappear
    3. Extinction – 99% of the species ever lived exist nowadays

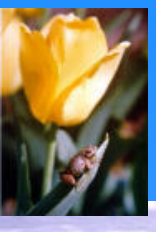




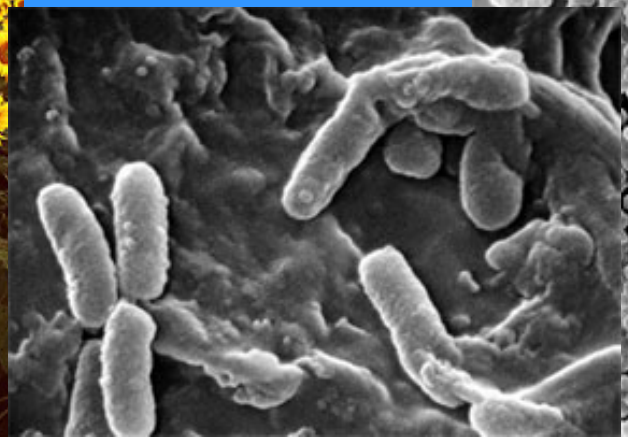
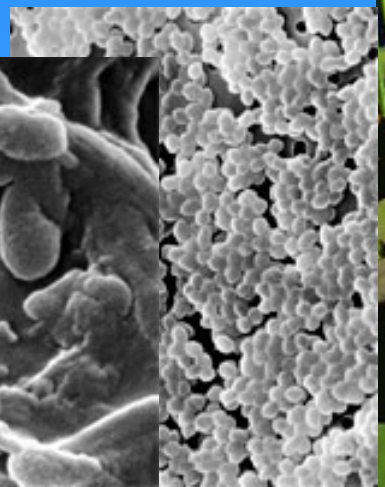




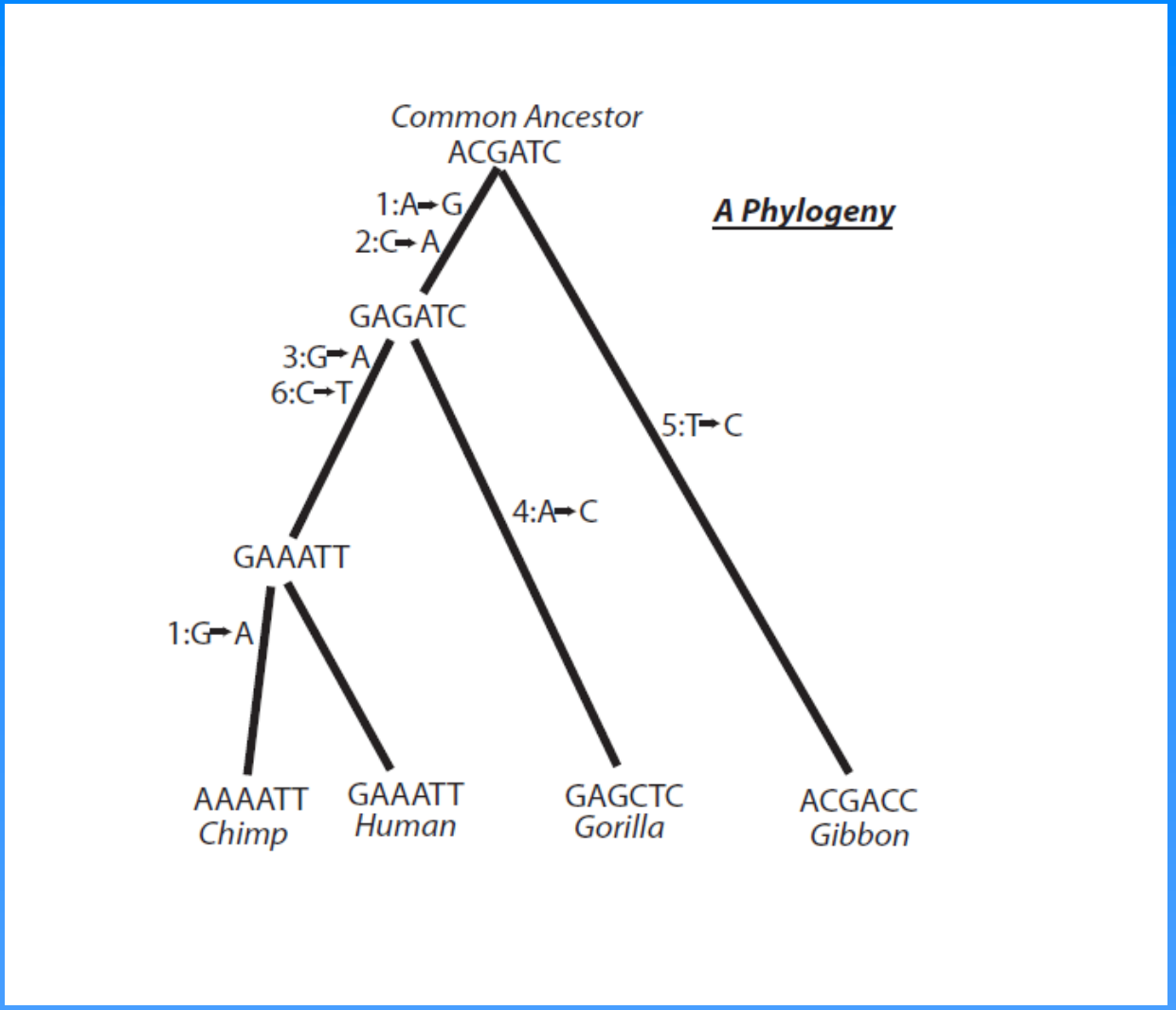
Phylogenetics

- Taxonomical classification of organisms based on distinctions
- Phylogeny – Evolutionary history (mostly identified with an evolutionary tree)

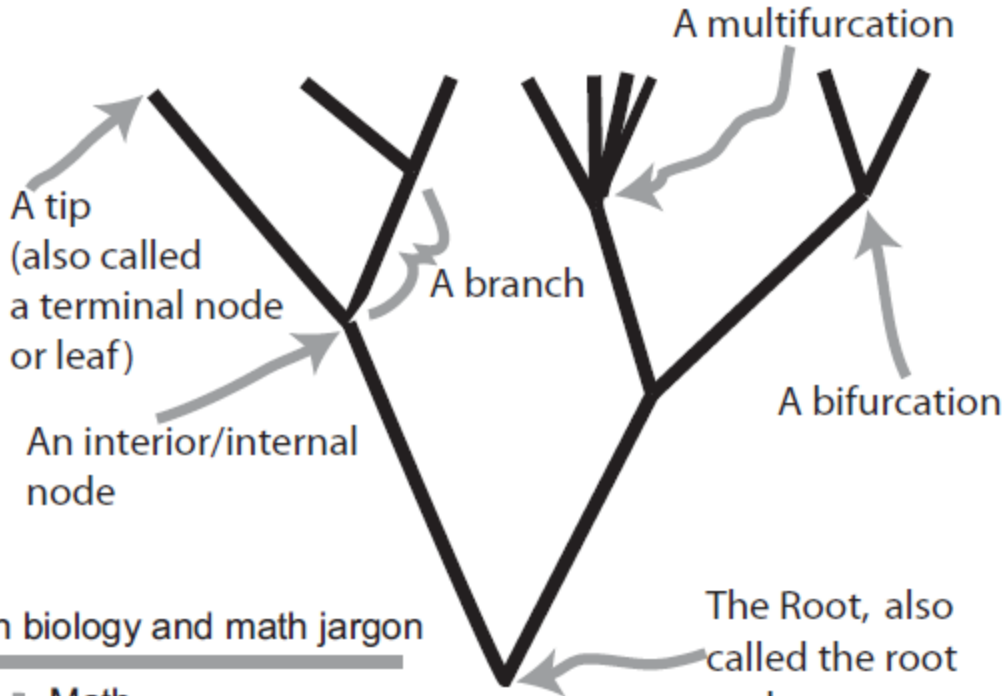



Basic Concepts



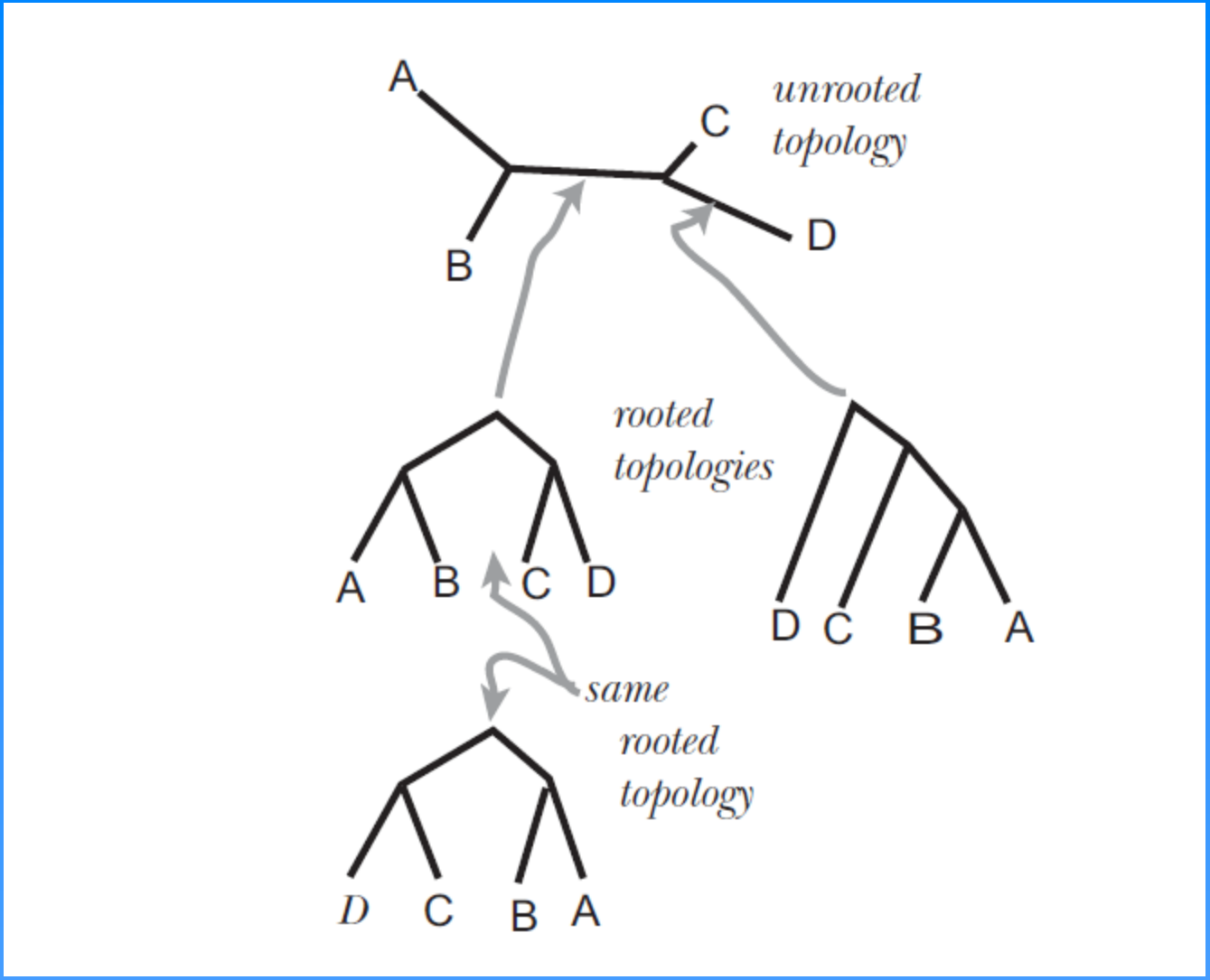
Tree Anatomy



Translating between biology and math jargon

Biology	Math
Tree	Graph
Branch	Edge
Node	Vertice

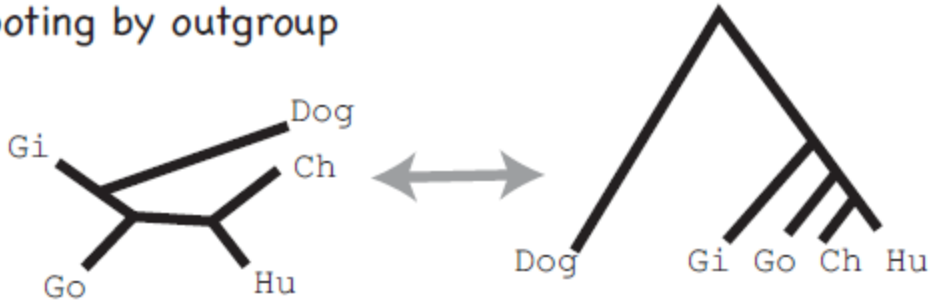
Rooted vs Unrooted topologies



Rooting unrooted trees

The two common ways phylogenies are rooted:

1. Rooting by outgroup



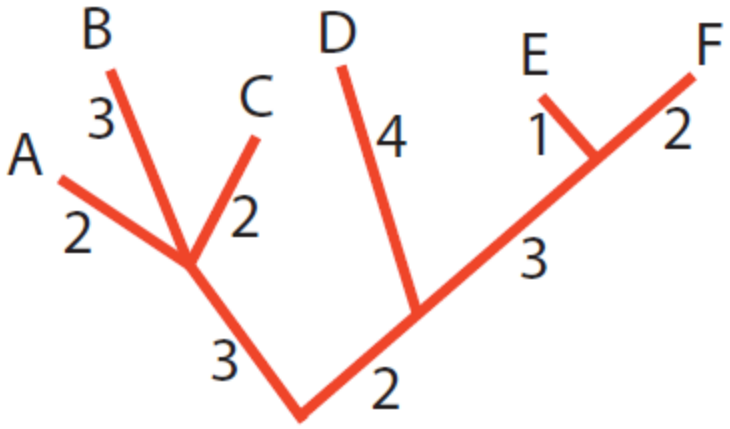
"Outgroup" = Dog "Ingroup" = Gi & Go & Ch & Hu

2. Rooting by molecular clock

All "tips" should be equally far from root

Tree Representation

Newick Tree Format: a computer-readable representation of trees



above can be expressed as

((A:2,B:3,C:2):3,(D:4,(E:1,F:2):3):2);

or

((D:4,(F:2,E:1):3):2,(B:3,A:2,C:2):3,);

or ...

For more detail, see http://evolution.genetics.washington.edu/phylip/newicktree.html

Interpreting Newick Format

Skeleton of Implementation ...

Start at Root and
Create Root Node

"(" means create branch and
create node to end branch

Newick Tree Format: a computer-readable representation of trees

above can be expressed as

((A:2,B:3,C:2):3,(D:4,(E:1,F:2):3):2);

or

((D:4,(F:2,E:1):3):2,(B:3,A:2,C:2):3,);

or ...

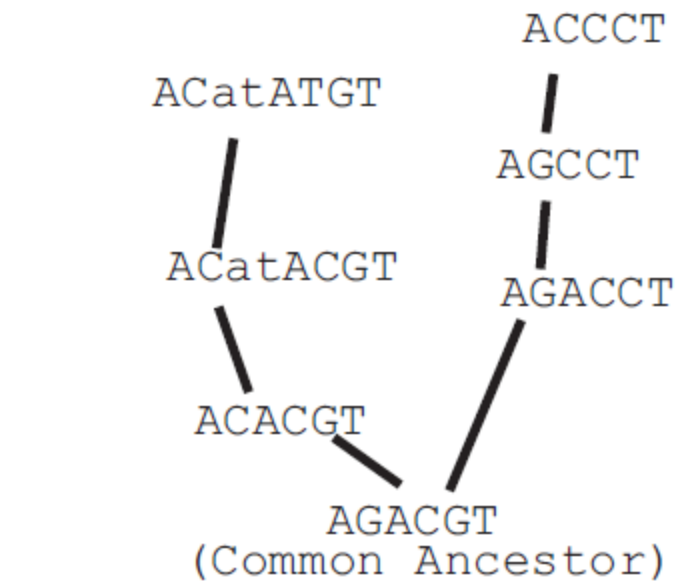
"," means backtrack 1 branch and then create branch and create node to end branch

")" means backtrack 1 branch

":" means get ready to read length of branch that ends at current node

(taxon names also need to be handled, convenient to convert underscore in Newick Representation to a blank space in a taxon name)

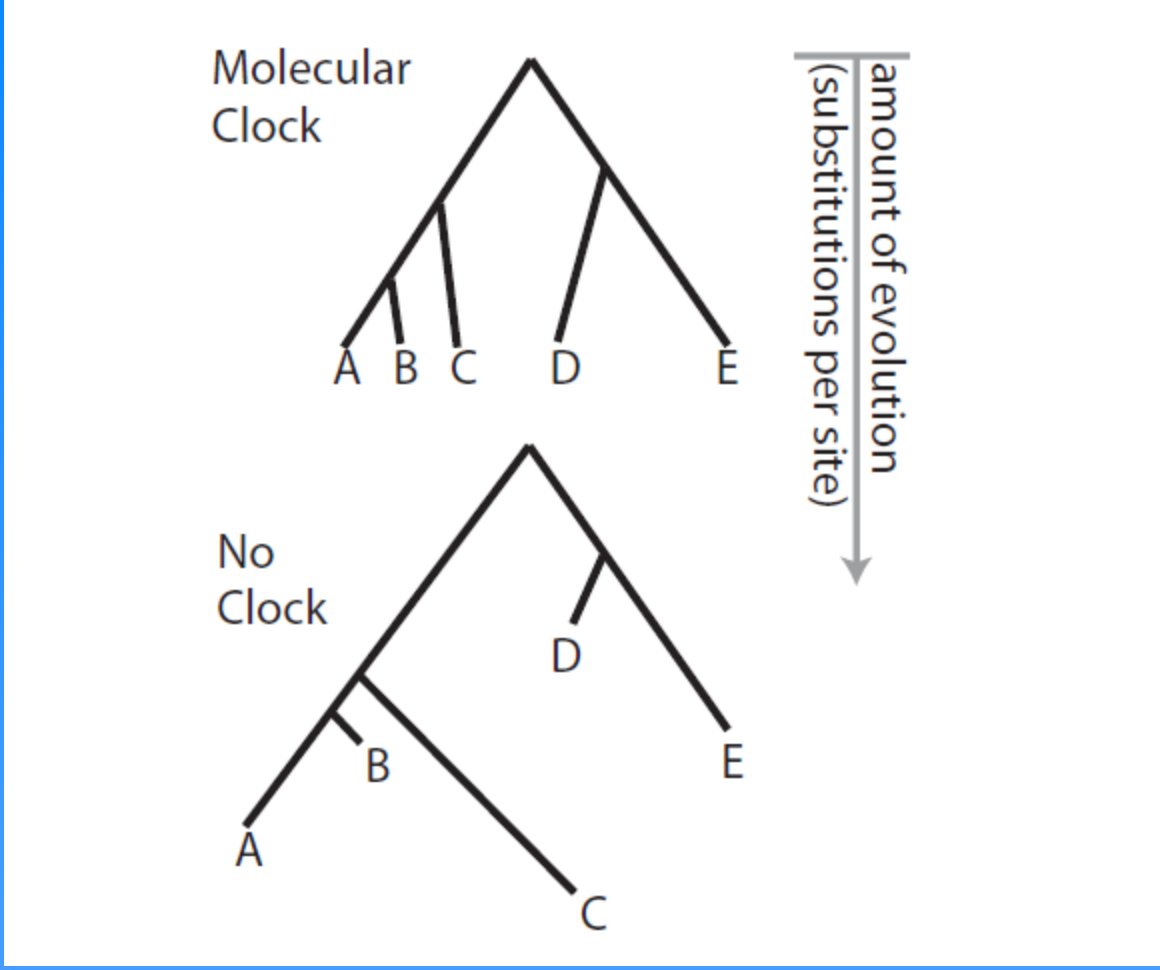
Insertions and Deletions



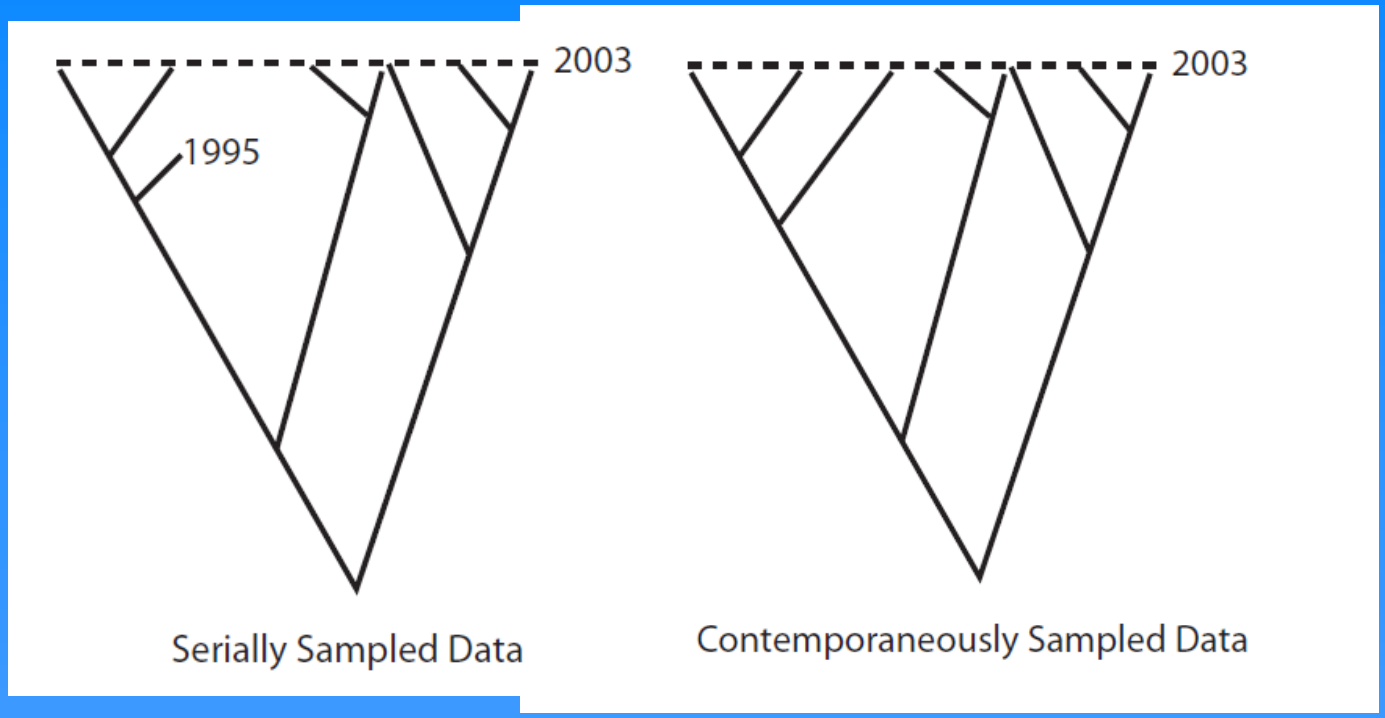
The "true" alignment:

ACATATGT
AC---CCT

Time vs Rates



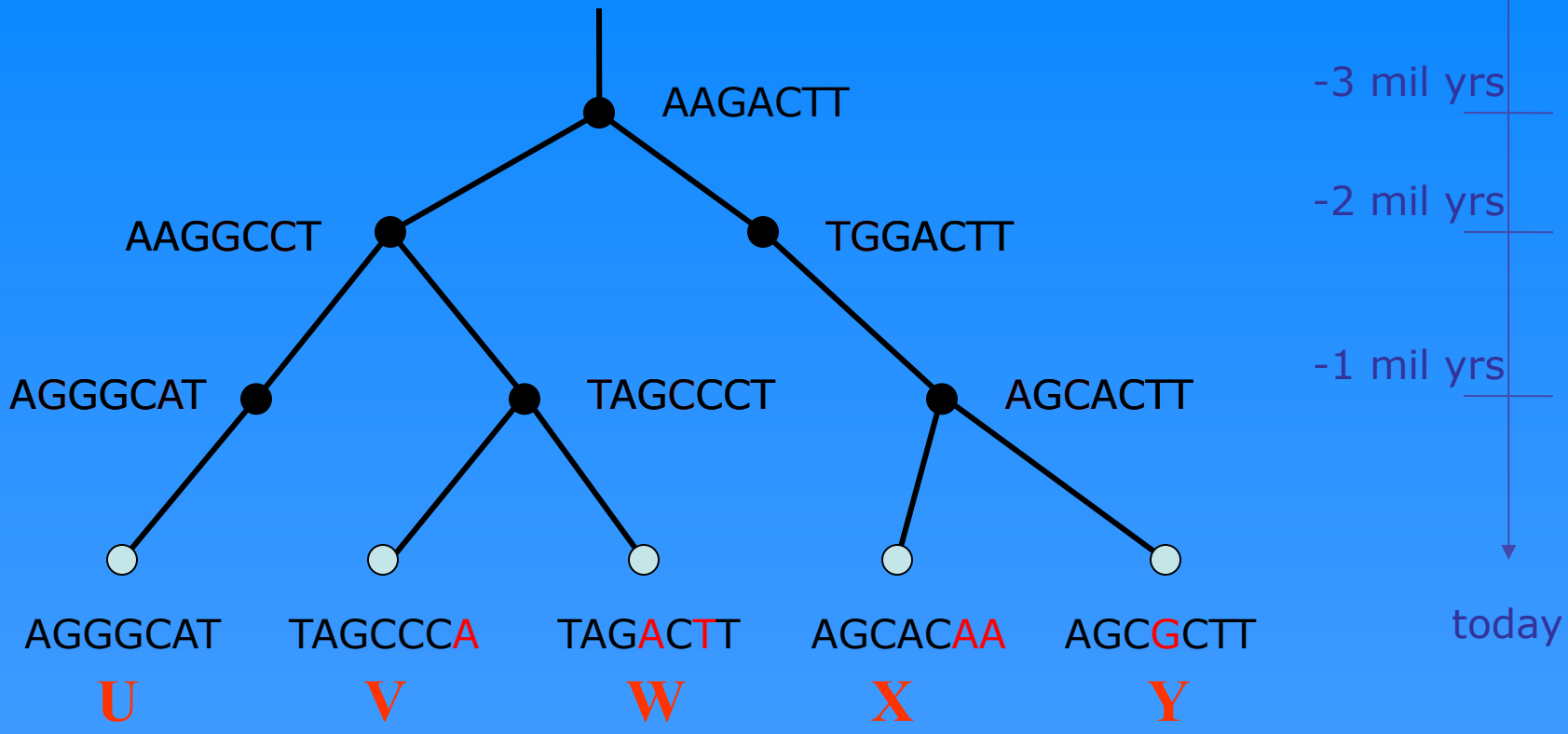
Sampling Times



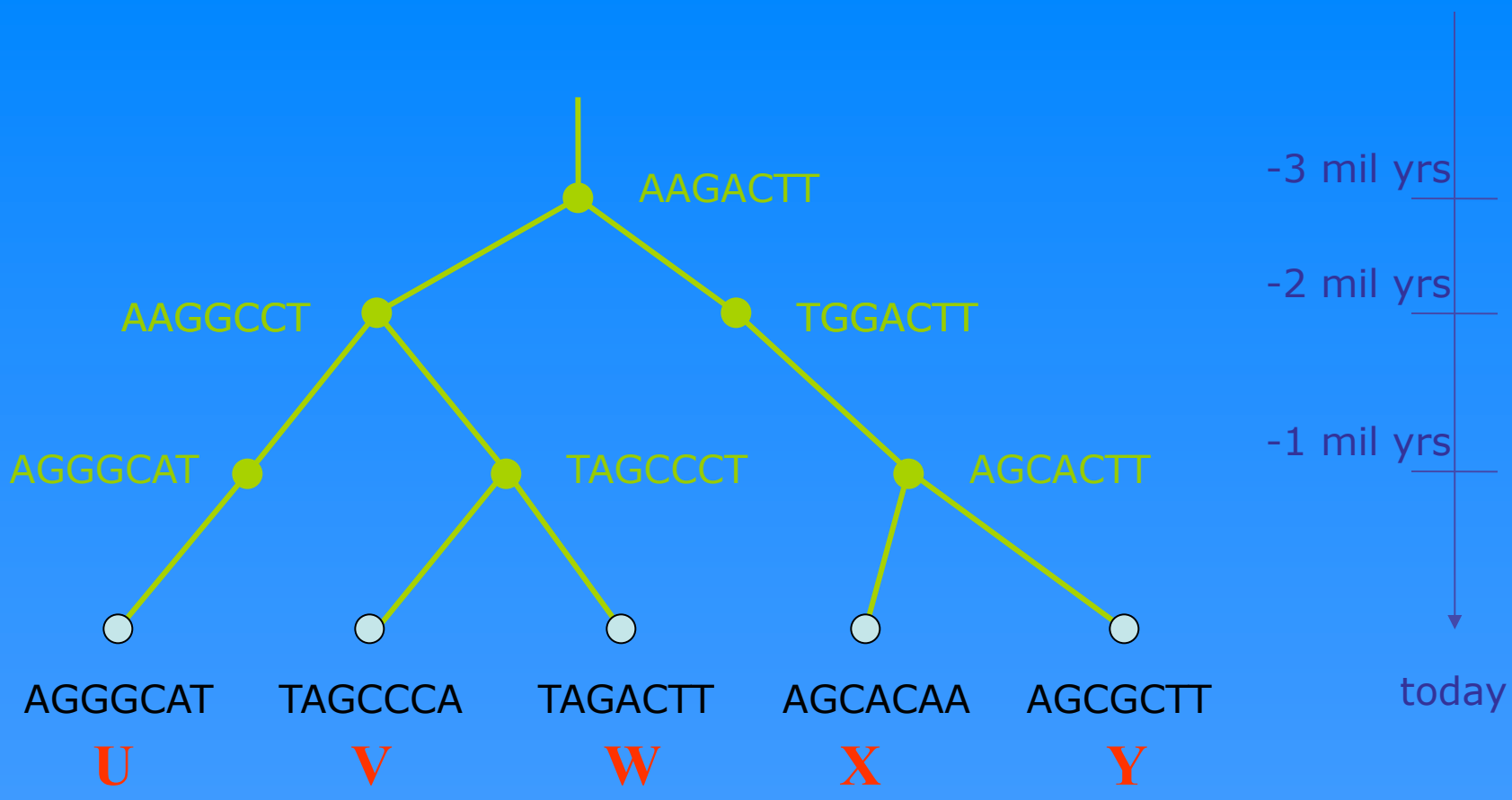
Jargon

- Mutation -- A change in hereditary material (i.e.,DNA or RNA). Change may be 1 nucleotide type instead of another (i.e., a point mutation) or may be insertion or may be deletion
- Fixation (by descent) -- When a new mutation later becomes ancestor of all gene copies in population
- Nucleotide Substitution -- a point mutation that gets fixed.
- Amino Acid Replacement -- Change in protein sequence that results from nucleotide substitution.

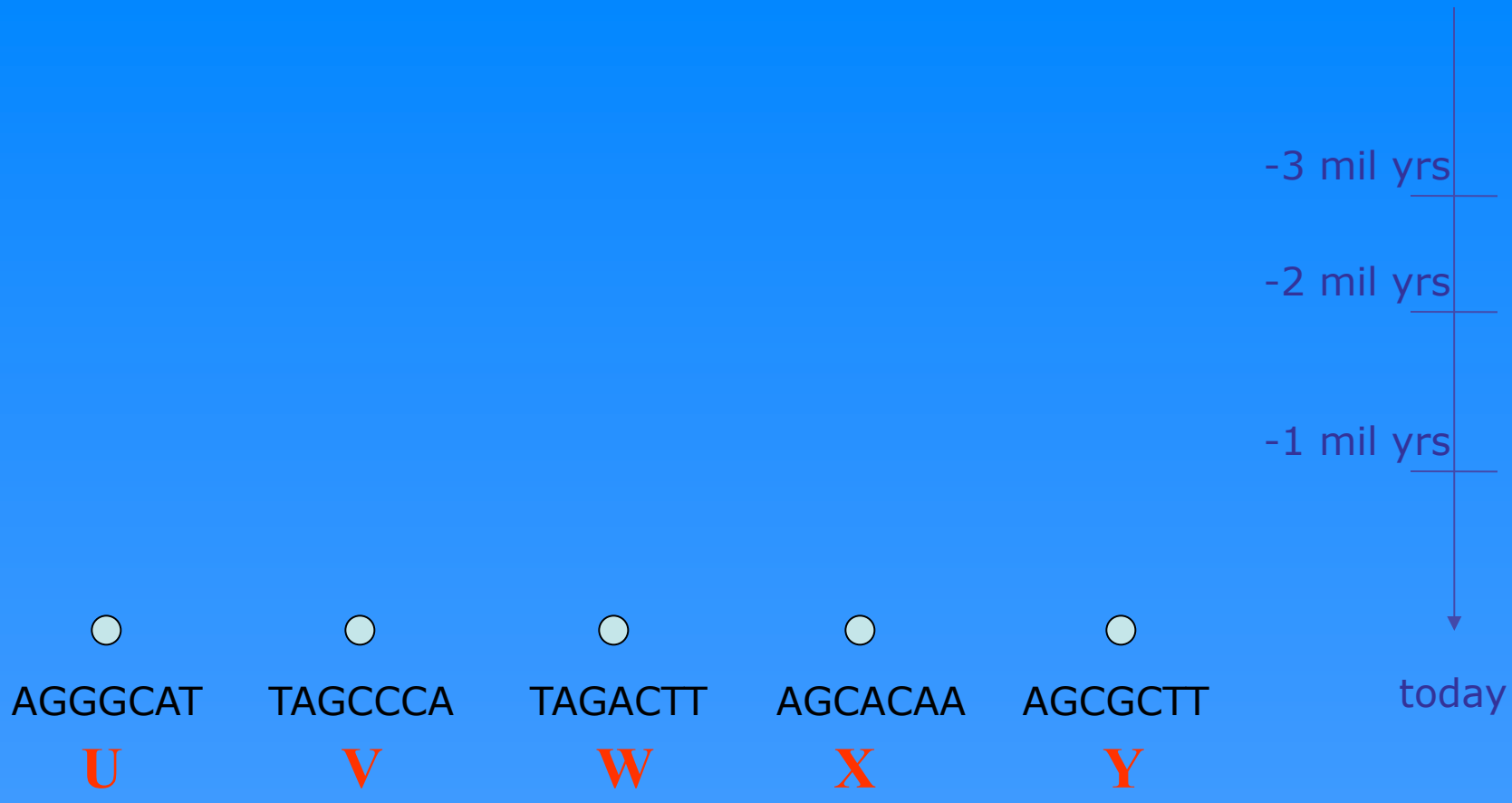
Sequence Evolution
(substantially simplified)



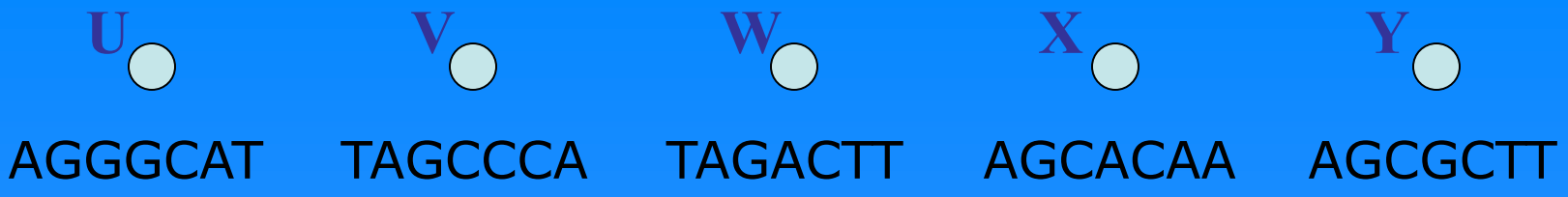
Sequence Evolution

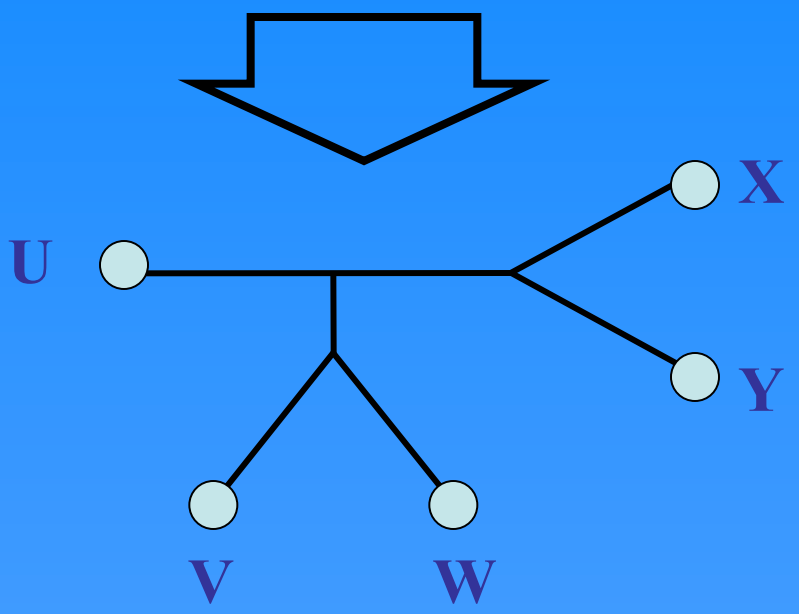


Sequence Evolution



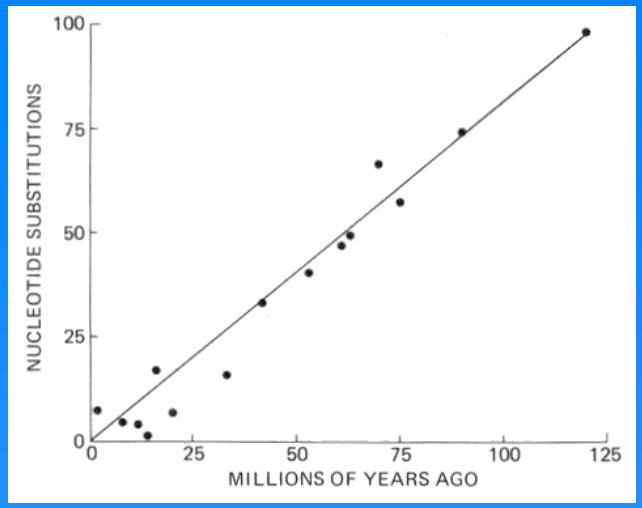
The Phylogenetic Problem





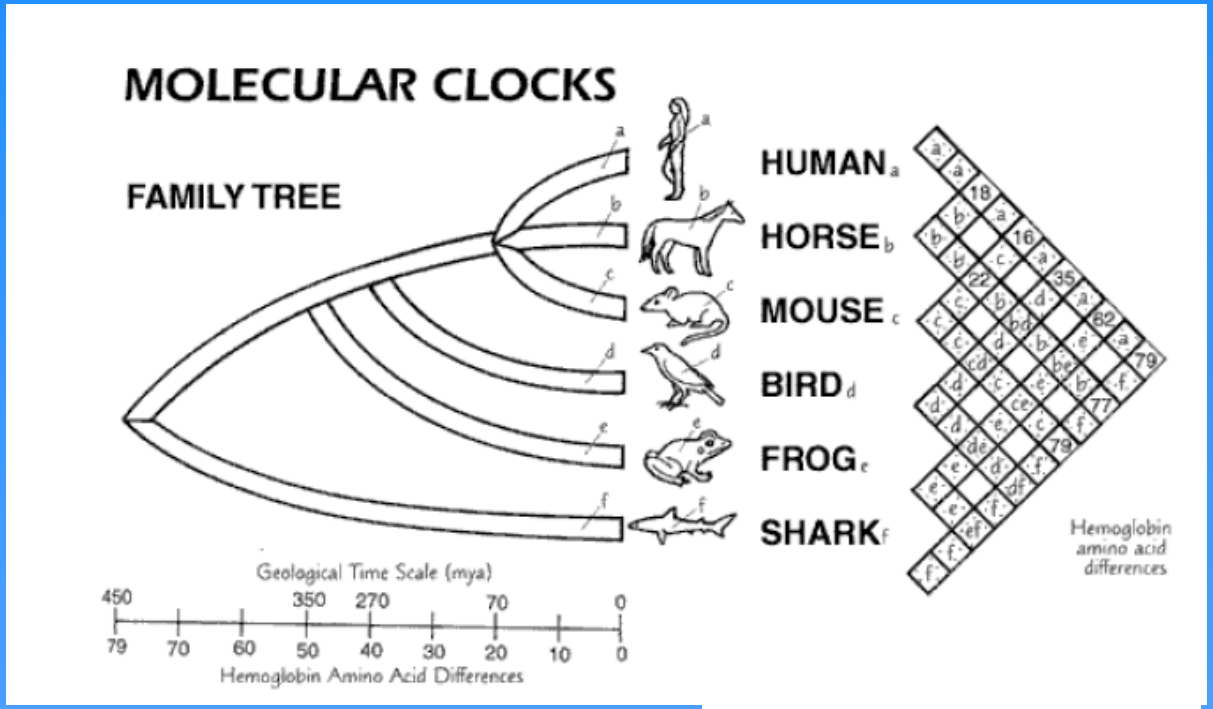
Unrooted trees!

Molecular Phylogenetics



First time to link molecular differences to evolutionary rate:

Zuckerkandl and Pauling
Molecules as Documents of Evolutionary History (1965)



First Molecular Phylogenetics

- Fitch & Margoliash: First *molecular* phylogenetic reconstruction over 20 species.

  Fitch &. Margoliash *Construction of phylogenetic trees*. 1967

  - Used least squares to best fit a tree to the observed data (Amino Acid differences).
  - However, "*We were lucky that cytochrome c was so slowly evolving, or our first tree would have been garbage*."
- Fitch later introduced his more famous method: "*Maximum Parsimony* ". Fitch, *Toward defining the course of evolution: minimum change for a specific tree topology,* 1971.
- based on the "*Ockham's Razor*" principle that evolution is parsimonious.
- A very widespread technique in biology.

Standard Phylogenetic Analyses

- Step 1: Identify orthologous (originated from a common ancestor) genes in the set of species to be analysed (e.g. Cytochrom C, 16s-rRNA).
- Step 2: Align the sequences to pose orthologous positions in one column.
- Step 3: Estimate the evolutionary history from the multiple alignment. (This can result in *many* trees.)
- Step 4: Assign confidence values to the inferred tree edges.

Issues in reconstructing evolutionary histories

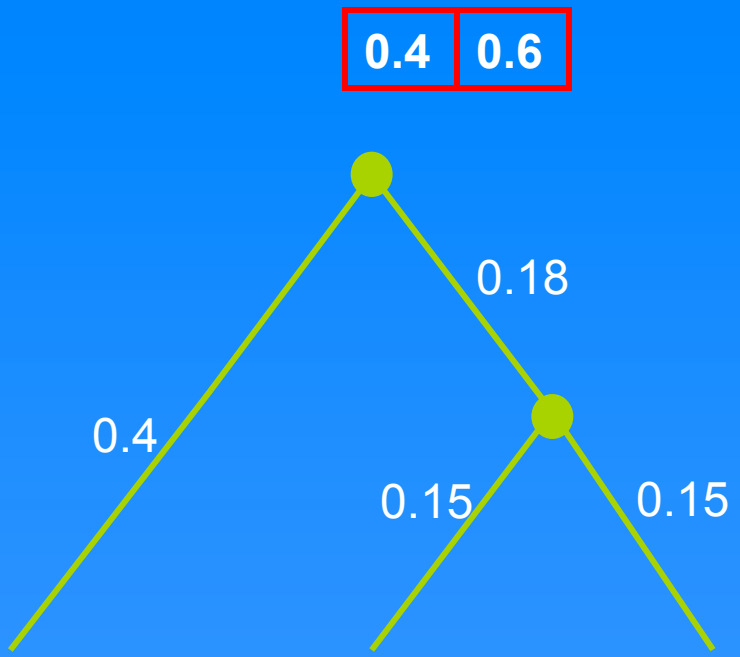
- Tree is unknown; how can we tell if we have the right answer?
- We model Evolution as a stochastic process operating on an unknown tree.
- This modelling allows us to study phylogenetics reconstruction as a statistical inverse problem

The Evolutionary Model

- Each *site* is a position in a sequence
- The *state* (i.e., *nucleotide*) of each site at the root is determined by the model.
- The sites evolve independently and identically (i.i.d.)
- For every edge *e*, (a matrix) $p^e(i,j)$ is defined, which is the probability of change from state *i* to state *j* along *e*.
- Along with the topology of *T* the probability of every *character* (a site pattern) is well defined.

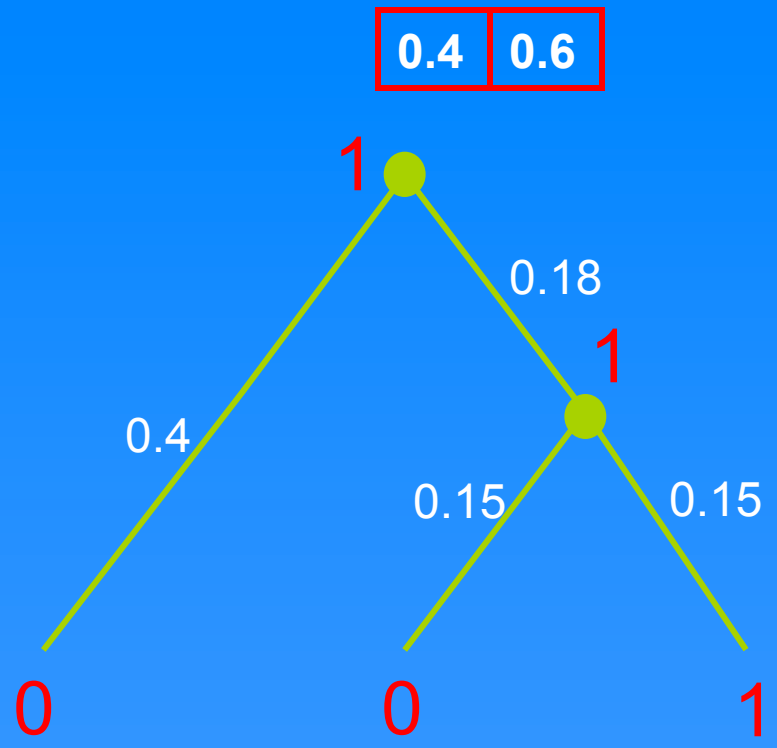
Evolutionary Model - Example

- Binary characters {0,1}
- Substitutions matrices are just a single probability *p* – *Pr*(change)
- We now evolve a character.
  - We first set a state at the root (according to the root probability.
  - Next, top down, by the edge probabilities, we swap or retain.



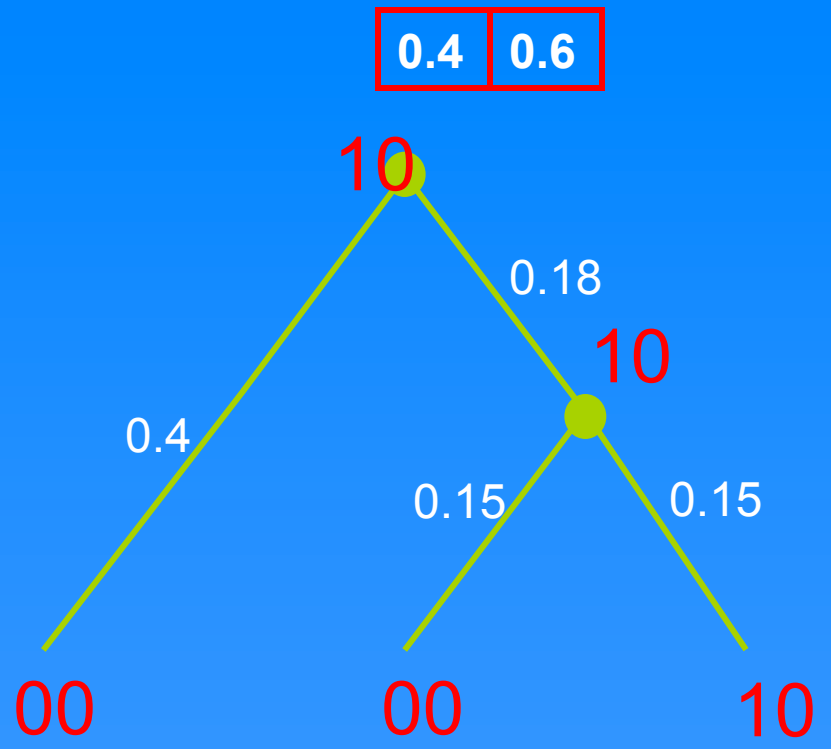
Evolutionary Model - Example

- Tossed 1 at the root.
- Swapped to 0 at left leaf.
- Remained 1 at right child.
- Swapped to 0 at middle leaf.
- Remained 1 at right leaf.
- Now another character.



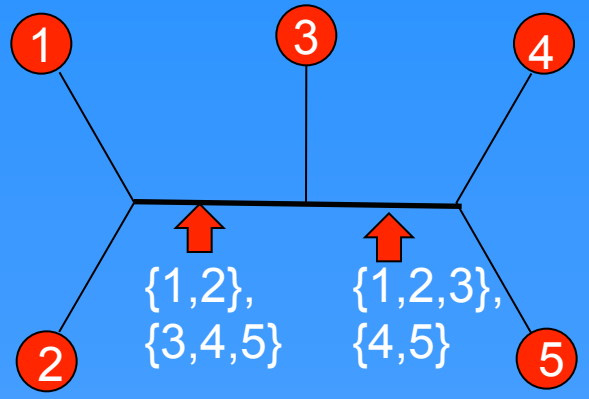
Evolutionary Model - Example

- Tossed 1 at the root.
- Swapped to 0 at left leaf.
- Remained 1 at right child.
- Swapped to 0 at middle leaf.
- Remained 1 at right leaf.
- Now another character.
- If repeated long enough, we converge to the edge probs.



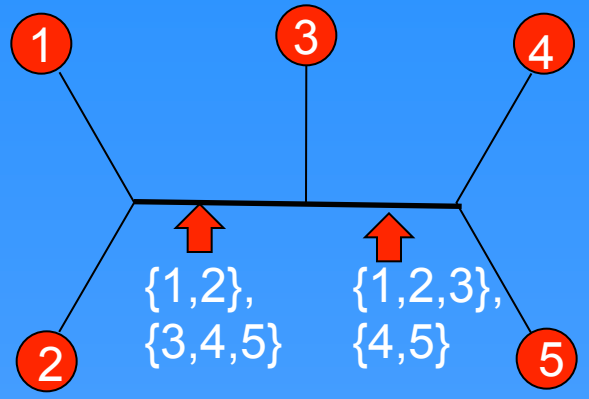
Tree Metric and Distances

- Tree metric is very useful when comparing trees, assigning confidence values, and other tasks.
- The procedure treats a tree as a combinatorial object, regardless of the statistical model underlying the tree.
- When we remove an edge from a tree, we remain with two subtrees



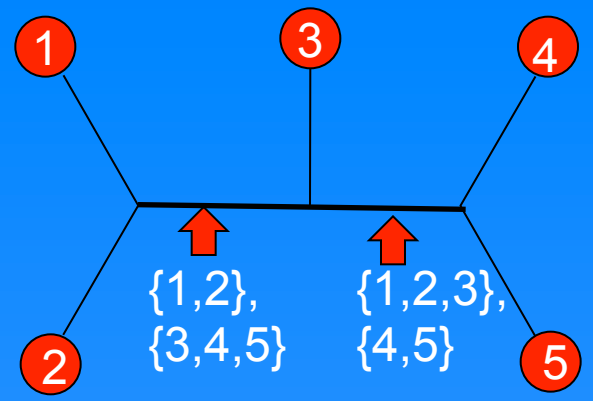
Tree as a Split System

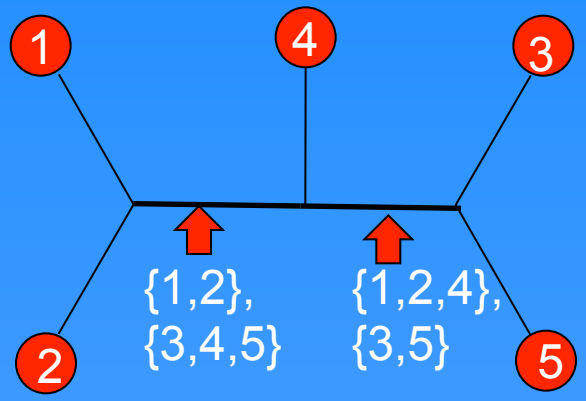
- When we remove an edge from a tree, we remain with two subtrees on two complementing taxa sets.
- We call this a *split* or a *partition*.
- We can identify that edge with the induced split.
- A tree is fully and uniquely identified by its induced split system.



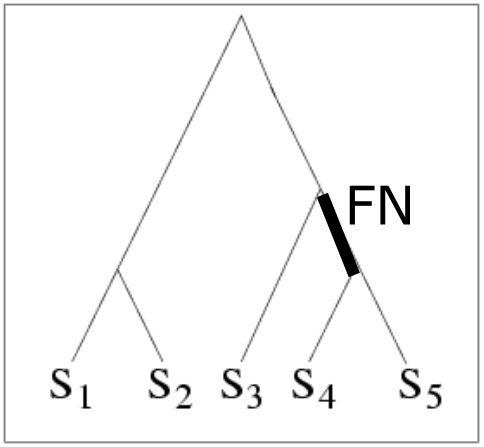
Tree Metric

- T_1 and T_2 are two trees on the same taxa set.
- The *symmetric difference* between T_1 and T_2 is the number of splits in exactly one tree.
- We normalize by the number of edges (here 2/4 = 0.5).
- When one tree is the model and the other is inferred, we call it *false positive/negative.*





Quantifying Topological Error



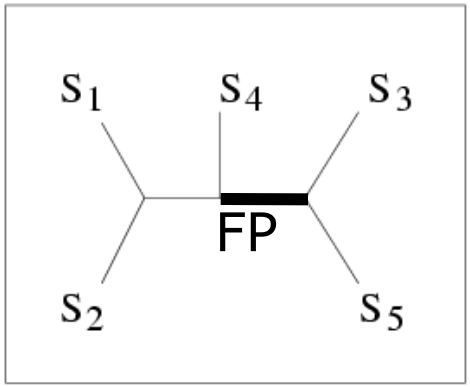
TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

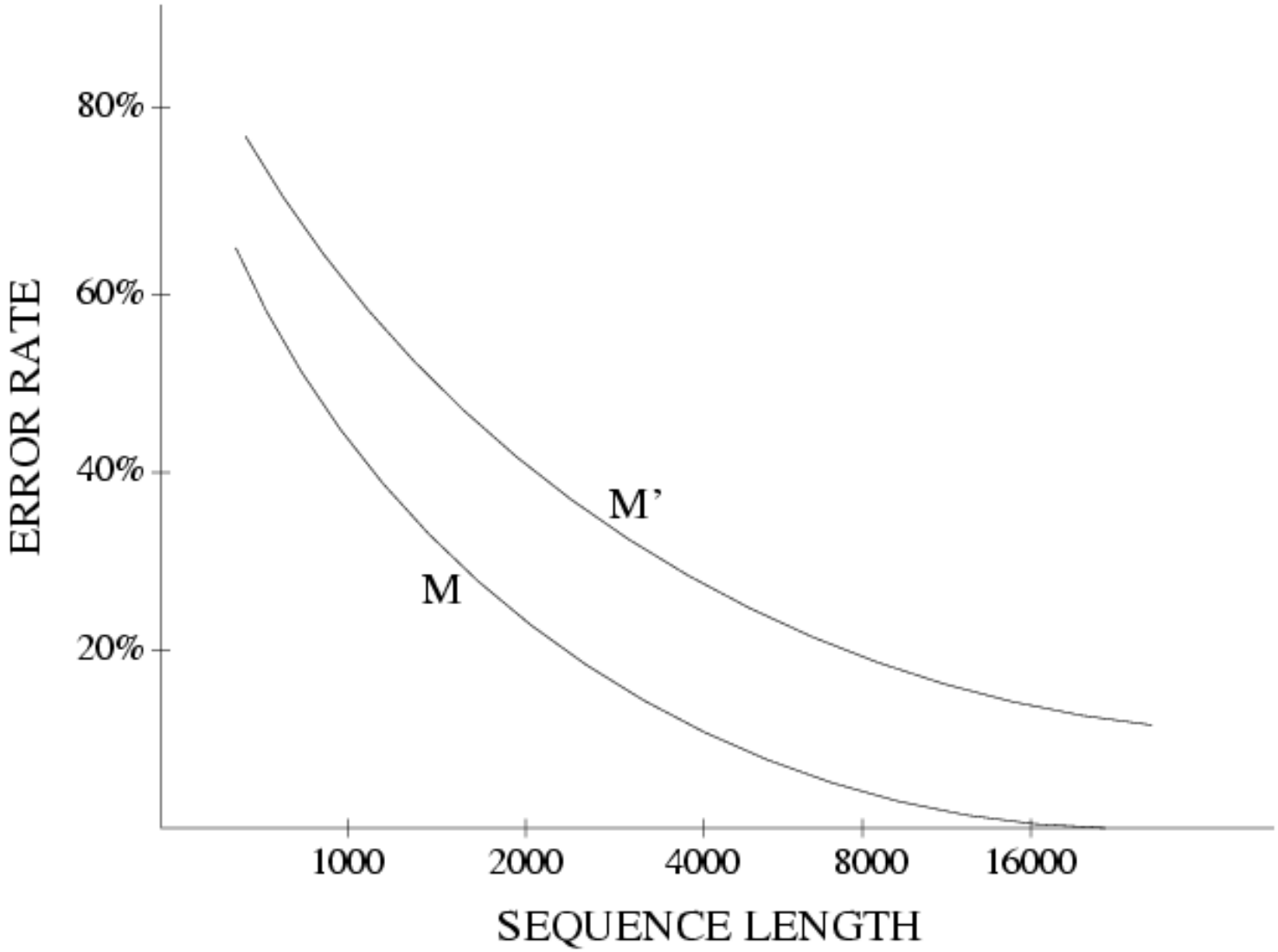
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



INFERRED TREE

Statistical Consistency And Convergence Rates



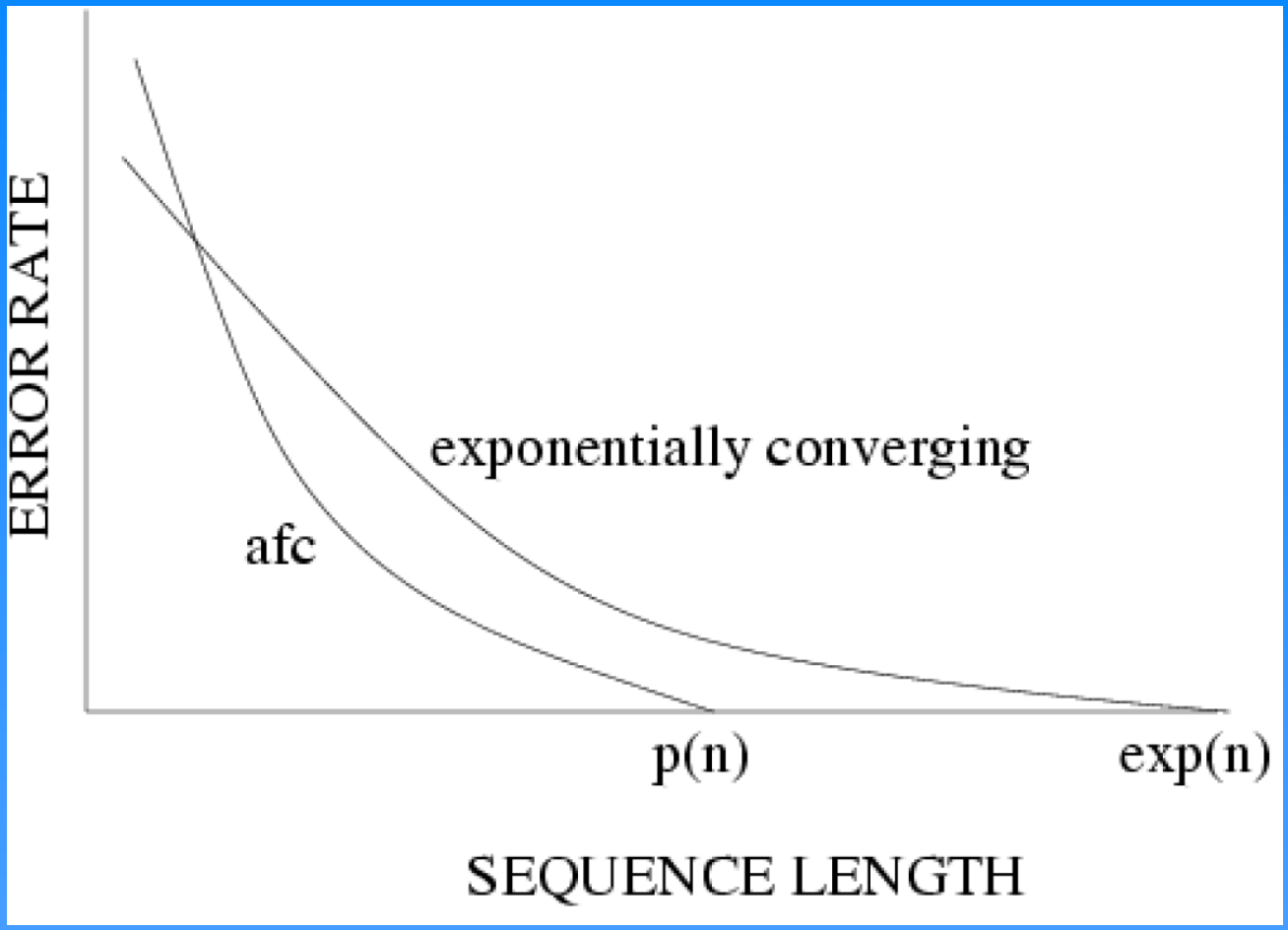
Statistical Performance Issues

- Intuition: Consider a coin. By looking at it we might not be able to determine if the coin is fair or biased.
- This will not change with the more time we look at it.
- Now we start to toss it.
- The first toss yielded a head. Can we say it is biased?
- After 100 tosses we have more idea.

Statistical Performance Issues

- An estimation method is *statistically consistent* under a model if the probability that the method returns the true tree goes to 1 as the sequence length goes to infinity.
- Convergence rate: the amount of data that a method needs (or simply how fast) to return the true tree.
- That amount of data is naturally proportional to the tree size – number of taxa.

Absolute Fast Convergence vs. Exponential Convergence



Reconstruction Approaches

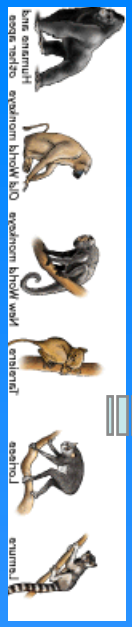
- Sequence based methods:
  - Two main categories:
    - *Character based methods*: Trees are constructed by comparing the characters of the corresponding sequences. Characters are mainly molecular (nucleotides in homologous DNA)
    - *Distance based methods*: Input is a square symmetric distance matrix. Seeks trees (edge-weighted) best-describing these distances.
- Supertree methods:
  - Construct small (reliable) trees form any data and combine it to a complete tree by combinatorial algorithms.
  - Quartet based methods.

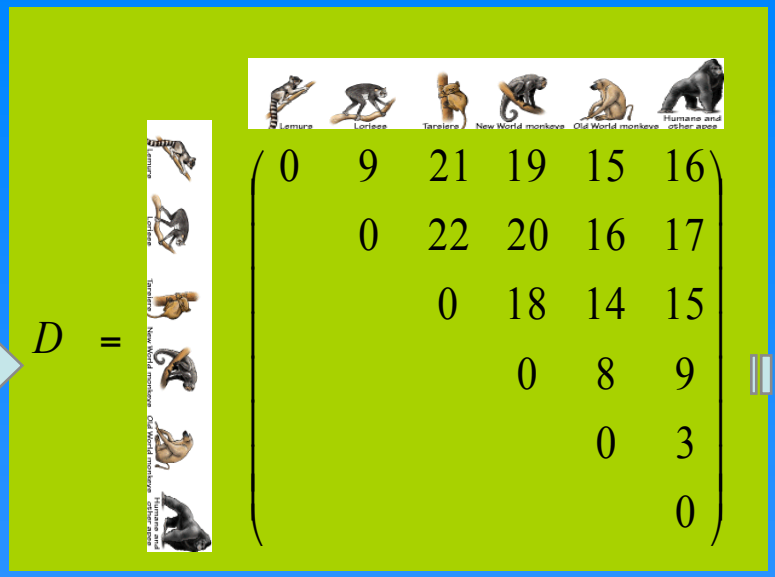
Character Based Methods

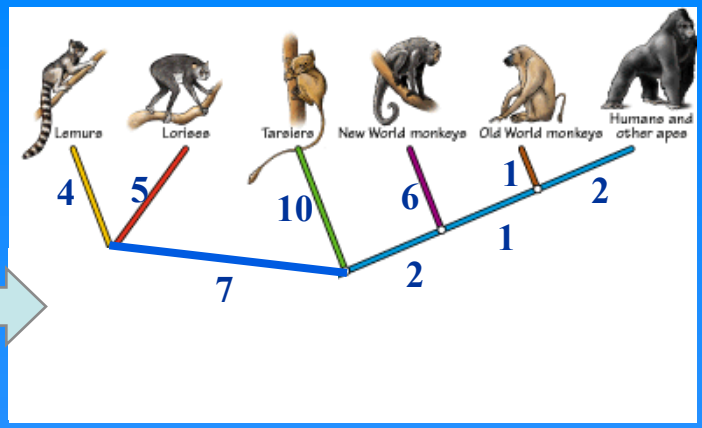
species	C1	C2	C3	C4	…												Cm
dog	A	A	C	A	G	G	T	C	T	T	C	G	A	G	G	C	C	C
horse	A	A	C	A	G	G	C	C	T	A	T	G	A	G	A	C	C	C
frog	A	A	C	A	G	G	T	C	T	T	T	G	A	G	T	C	C	C
human	A	A	C	A	G	G	T	C	T	T	T	G	A	T	G	A	C	C
pig	A	A	C	A	G	T	T	C	T	T	C	G	A	T	G	G	C	C
	*	*	*	*	*			*	*			*	*				*	*

1. Input: A n*m matrix.
2. Each character (column) is processed independently.
3. Task: Find a tree that best explains simultaneously all characters.

Distance Based Methods







1. Input: A n*n matrix.
2. Each entry represents the *observed* distance between the corresponding species.
3. Task: Find a *weighted* tree that best approximates the input distances.