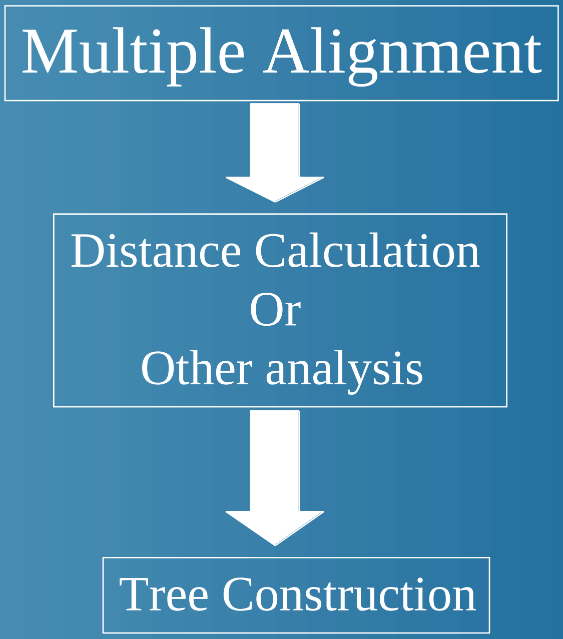
Multiple Alignment and Phylogenetic Analysis

Steps in Phylogenetic Analysis



Multiple Alignment

Distance Calculation
Or
Other analysis

Tree Construction

Basic Assumptions in Phylogenetic Analysis

What people agree about -

Proteins evolve over time

What people disagree about - Everything else

How proteins evolve
How fast proteins evolve
What is the best method of measuring evolution
How to construct phylogenetic trees

Alignment is Key

The multiple alignment is critical.

If you start with a bad alignment
the phylogenetic tree will be incorrect.

A good alignment will have no gaps
Chop off regions that don’t align before
sending the alignment for phylogenetic analysis.

Clustal W

W stands for Weighted

Different weights are given to sequences and parameters in different parts of the alignment.

Position Specific Gap Penalties

- The goal is to insert gaps only in “loop” regions
- Higher penalties in the middle of helices and strands

Large penalty for closely related sequences
Small penalty for divergent sequences

Step 1 - Pairwise Alignments

Compare each sequence with each other
Calculate a distance matrix

A	-		
B	.87	-	
C	.59	.60	-
	A	B	C

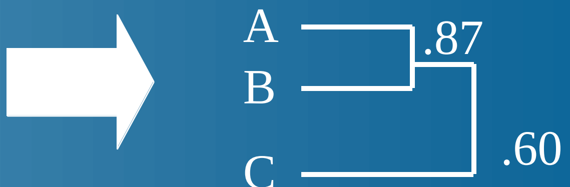
Distance =
Number of
exact matches
divided by the
sequence length
(ignoring gaps)

.87 = 87% identical

Step 2 - Create Guide Tree

Use the results of the Distance Matrix to create a Guide Tree to help determine in what order the sequences are aligned.

A	-		
B	.87	-	
C	.59	.60	-
	A	B	C

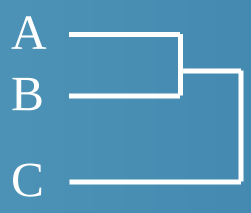


Guide Tree

Guide Tree, or Dendogram has no phylogenetic meaning
Cannot be used to show evolutionary relationships

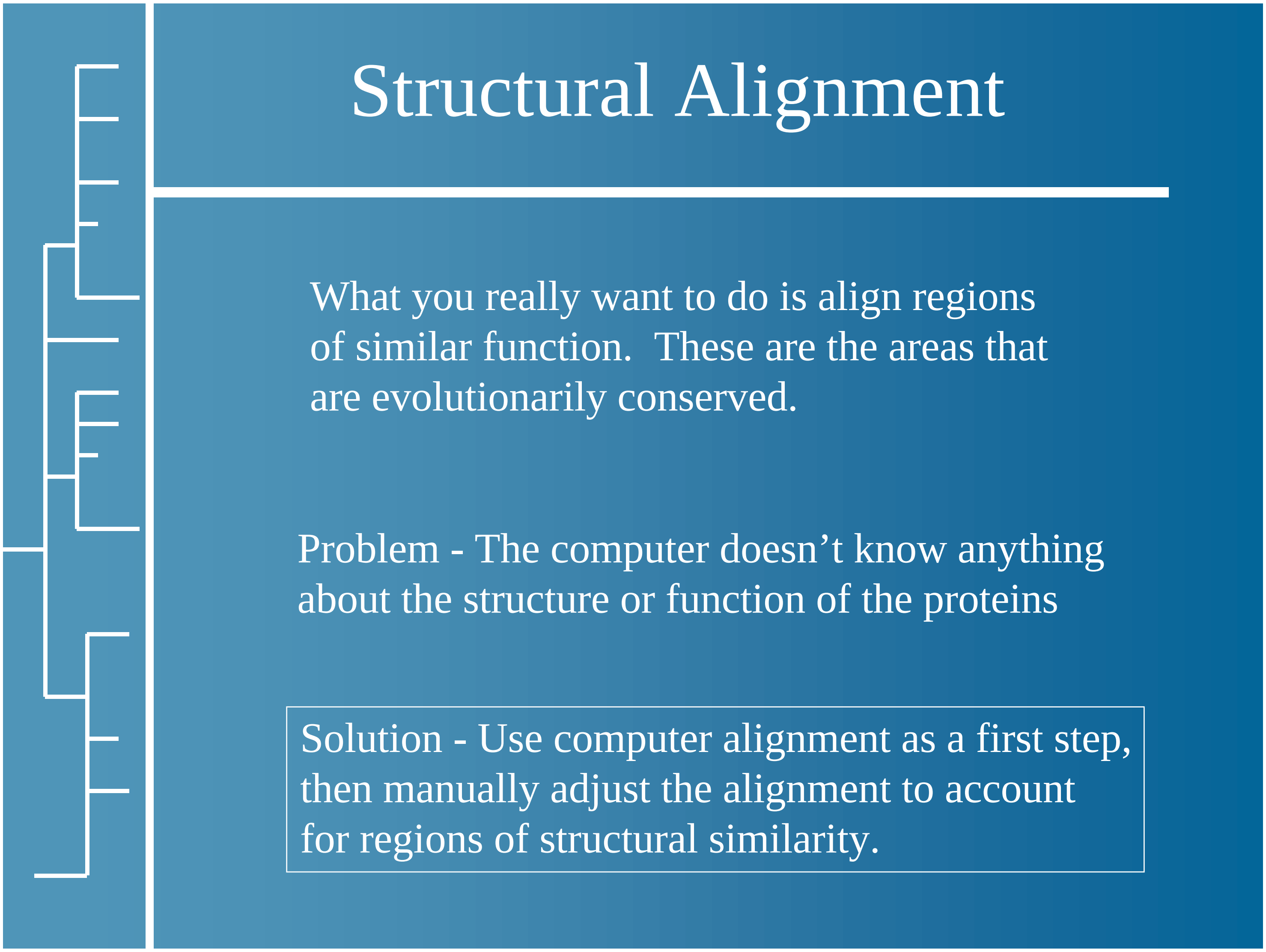
Step 3 - Progressive Alignment

Follow the Guide Tree and align the sequences



- Align A and B first
- Then add sequence C to the previous alignment

Align the most closely related sequences first, then add in the more distantly related ones and align them to the existing alignment, inserting gaps if necessary



Structural Alignment

What you really want to do is align regions of similar function. These are the areas that are evolutionarily conserved.

Problem - The computer doesn’t know anything about the structure or function of the proteins

Solution - Use computer alignment as a first step, then manually adjust the alignment to account for regions of structural similarity.

Structural Regions to Align

- Helices
- Sheets
- Active Sites or other functional regions
- Disulfide Bonds

Where to find this structural information

- Primary literature, enzymatic or mutational studies
- Protein Sequence Databases
- Structural Databases `http://www.rcsb.org`

DNA or Protein in Evolutionary Analysis?

DNA can show underlying mutations that won't appear in proteins. "Wobble" base mutations seldom change protein sequence.

Glycine = GGA, GGC, GGG, GGT

Many phylogenetic programs ignore the 3rd base

Counter Argument -
These silent mutations don't reflect true rate of evolution since they are not under evolutionary pressure like amino acids are. Most protein mutations are silent.

DNA may change, but protein remains the same

Sequence 1 CTA GCT AGA GGA AGC CCA ACA GTA

Sequence 2 TTG GCG CGT GGG TCT CCG ACC GTT

LARGSPT

The protein could be under evolutionary pressure. If you only use protein data, you won’t see all the mutational events going on in the background.

Practical Considerations

When to use Clustal

Can be used to align any group of protein or nucleic acid sequences that are related to each other over their entire lengths.

Clustal is optimized to align sets of sequences that are entirely colinear, i.e. sequences that have the same protein domains, in the same order.

Multiple Alignment Problems

- Does the quality of the guide tree matter?

  Not for closely related sequences, but perhaps for distantly related ones.

- Single Parameter problem

  You are using one weight matrix, and one set of penalties for all the sequences. The best set of parameters for one part of the alignment may not be the best for another part.

- Local minimum problem

  If the initial alignments have a problem, they can’t be removed during subsequent steps.

When Not To Use Clustal

- Sequences do not share common ancestry
- Sequences have large, variable, N- and C-terminal overhangs
- Sequences are partially related
- Sequences include short non overlapping fragments.

Garbage in …. Garbage out

Two Basic Methods For Phylogenetic Analysis

- Distance Methods

Multiple Alignment -> Evolutionary Distances -> Tree

- Character Based Methods (Parsimony)

Multiple Alignment -> Tree(s)

Underlying Rate of Mutation

Normal mutation rate is 1 in 10^{-8} nucleotides

Normal Polymorphic Variance

Approximately 1 in every 1000 nucleotides

This is the background on which evolutionary changes are analyzed.

Types of Mutations

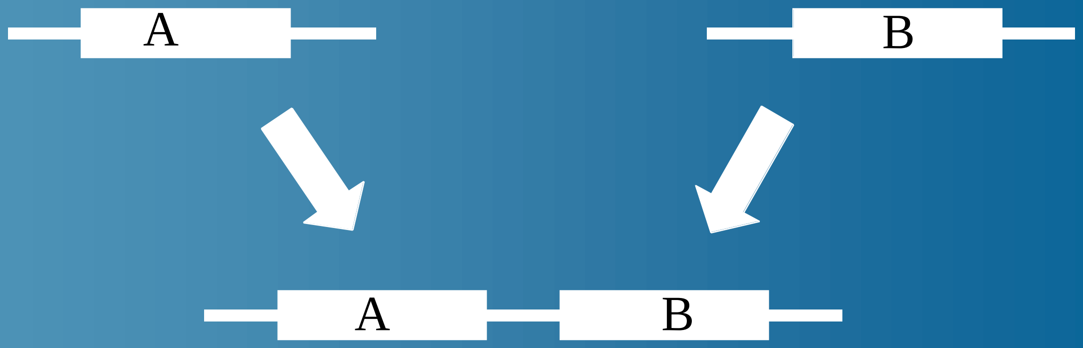
These mutations are easily detected by multiple alignment

- Base Substitutions
- Indel - Insertions & Deletions

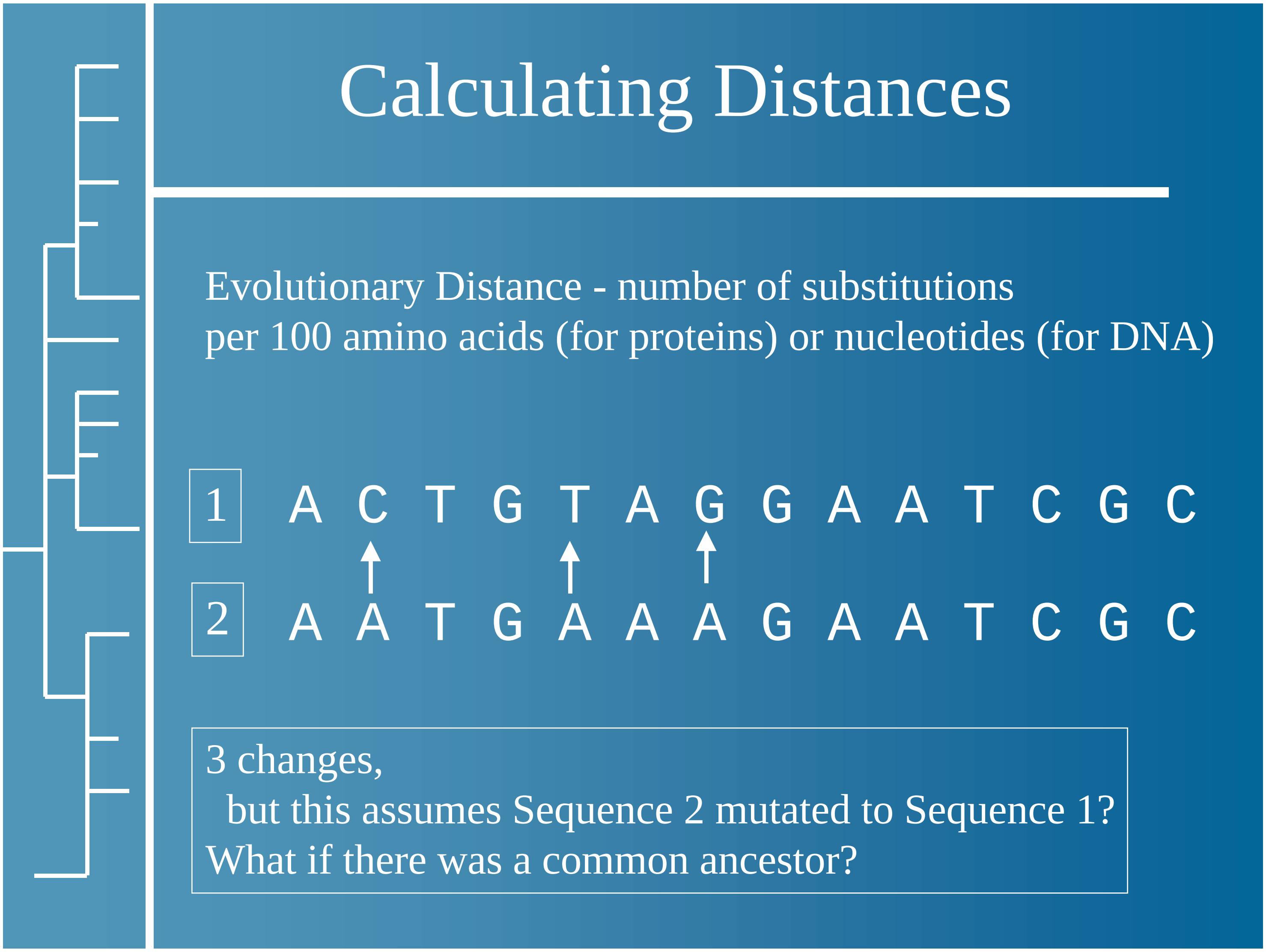
These mutations are not easily detected by multiple alignments

- Transposition
- Exon (domain) Shuffling

What about domain shuffling?



You would have to analyze each domain separately
Each domain may have evolved at a different rate.



Calculating Distances

Evolutionary Distance - number of substitutions per 100 amino acids (for proteins) or nucleotides (for DNA)

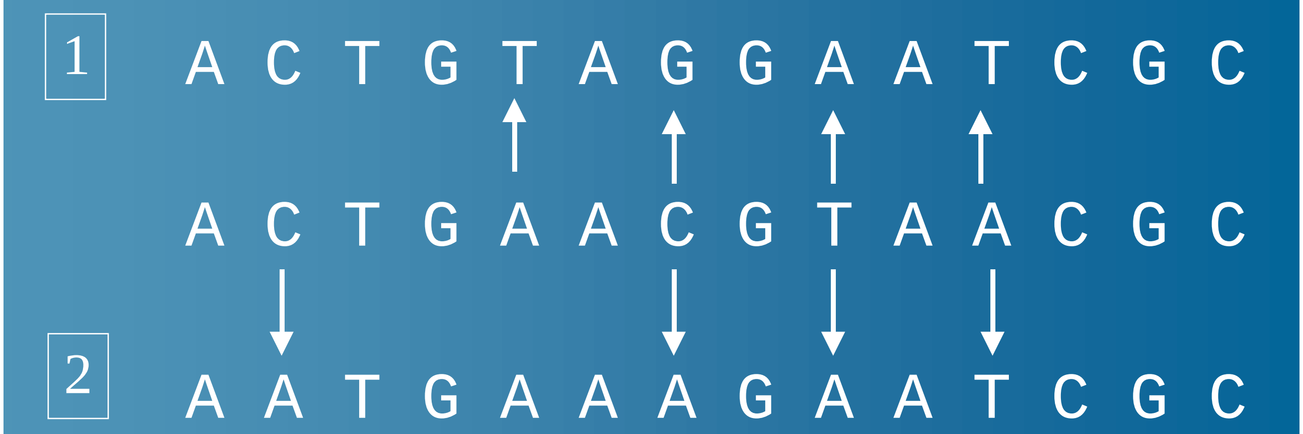
```
1  A C T G T A G G A A T C G C
     ↑     ↑   ↑
2  A A T G A A A G A A T C G C
```

3 changes,

but this assumes Sequence 2 mutated to Sequence 1?

What if there was a common ancestor?

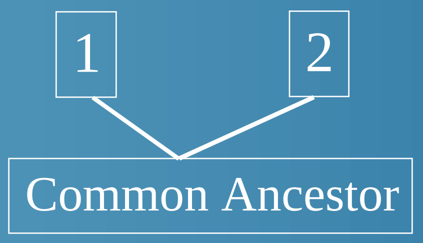
Distances



1 ACTGTAGGAATCGC

ACTGAACGTAACGC

2 AATGAAAGAATCGC

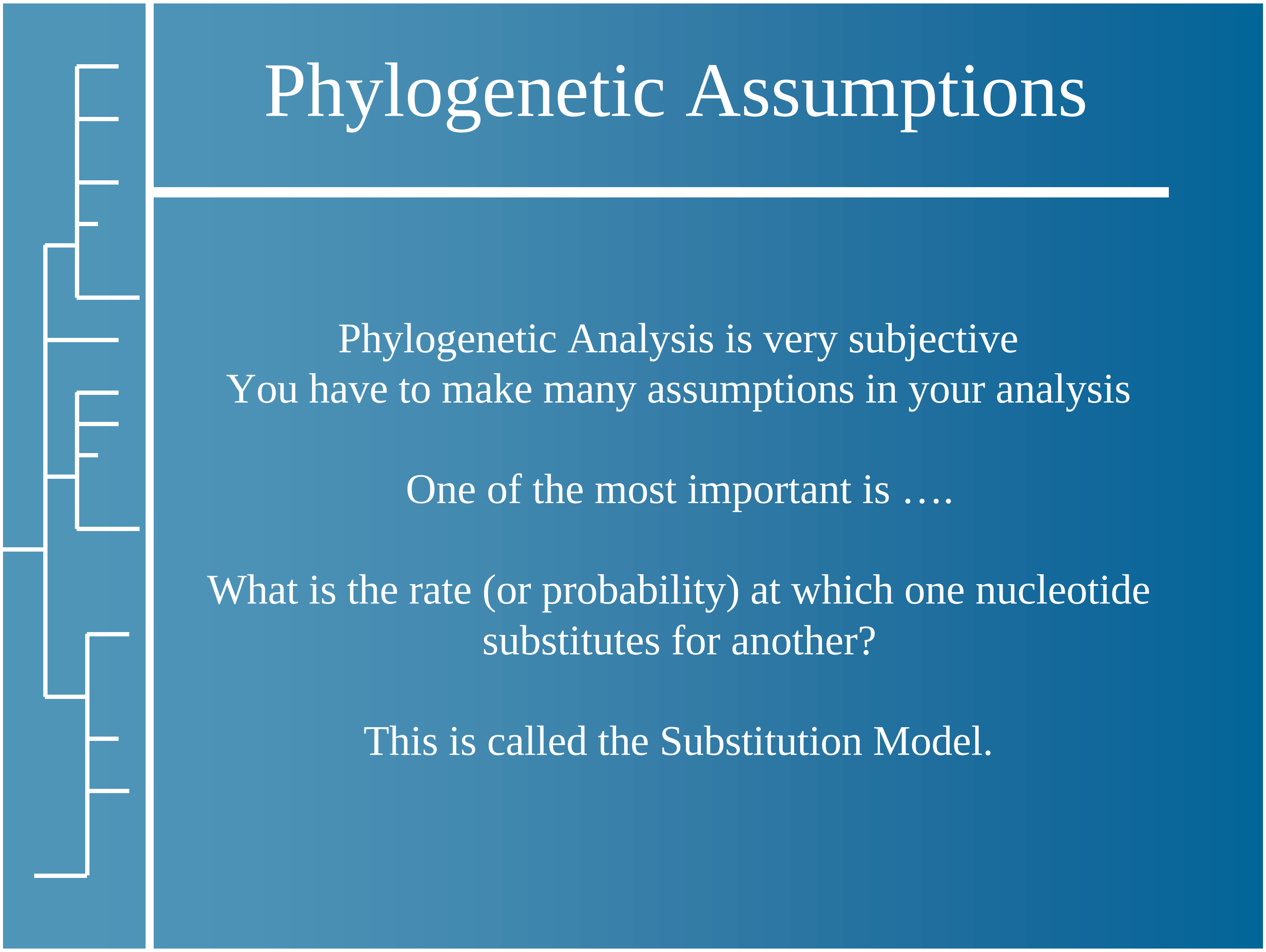


1

2

Common Ancestor

In reality, there are 8 changes



Phylogenetic Assumptions

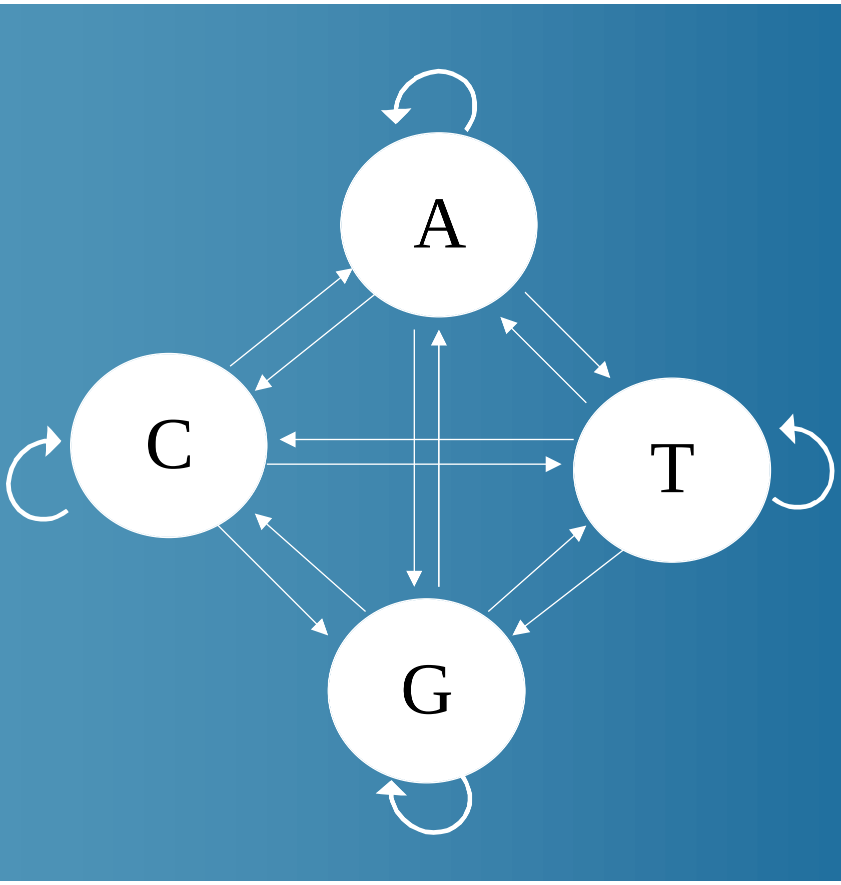
Phylogenetic Analysis is very subjective
You have to make many assumptions in your analysis

One of the most important is ….

What is the rate (or probability) at which one nucleotide substitutes for another?

This is called the Substitution Model.

Mutation Possibilities



What are the probabilities for each possibility?

Distance Corrections

What if the rates are not equal?

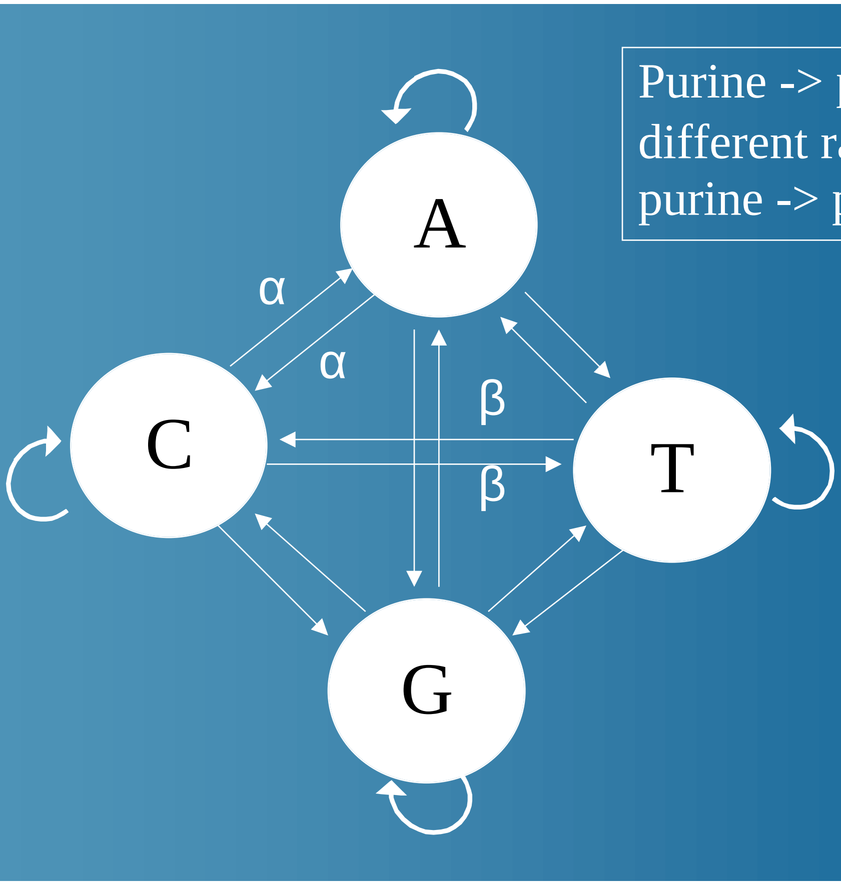
Jukes - Cantor - No bias. Substitutions occur randomly.
Equal probability of mutation for all nucleotides.
$D = -b \ln (1-D^*/b)$ b = .75 for DNA .95 for protein

Kimura (2 parameter)-
Transitions (C->T or A->G) more frequent than transversions (A ->T, C->G)

There are many other possibilities.

This is called a “substitution model”
The model for how one nucleotide substitutes for another

Kimura 2 Parameter



Purine -> purine (β) have a different rate than purine -> pyrimidine (α)

Tree Calculation Methods

Distance Methods - Evolutionary distances are used to construct trees (UPGMA & Neighbor Joining). Fast, easy to handle large numbers of sequences.

Character Methods

Parsimony Methods - trees are created to minimize the number of changes that are needed to explain the data.

Maximum Likelihood - Using a model for sequence evolution, create a tree that gives the highest likelihood of occurring with the given data.

Problems With Distance Methods

Distance methods build trees by grouping OTUs according to overall similarity.

There is the possibility that apparent overall similarity and true evolutionary relationship are not necessarily the same thing.

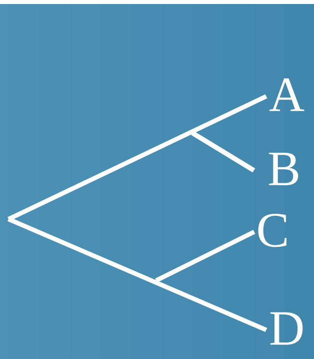
Tree Calculation

Once you calculate the distances

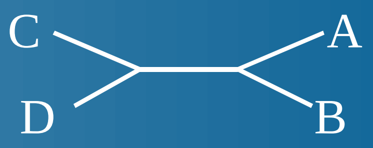
You then have to cluster the data together in a tree

There are many different ways to create trees
and to display the data

Tree Definitions



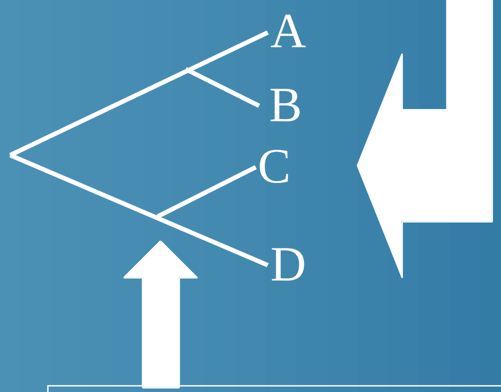
Rooted Tree
A common ancestor is defined



Unrooted Tree
No common ancestor

Definitions

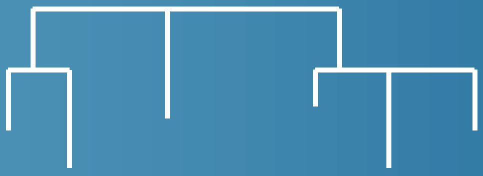
External Nodes, Operational Taxanomic Unit (OTU) - These are the molecules you are studying



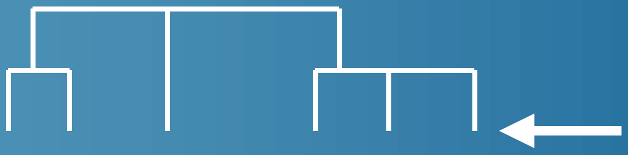
Internal Nodes - Hypothetical ancestral units

Types of Trees

Phylogram - Branch Lengths Proportional To Distance



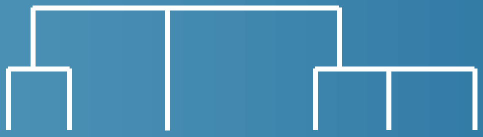
Cladogram - All Branch Lengths Equal



These are Unrooted Trees

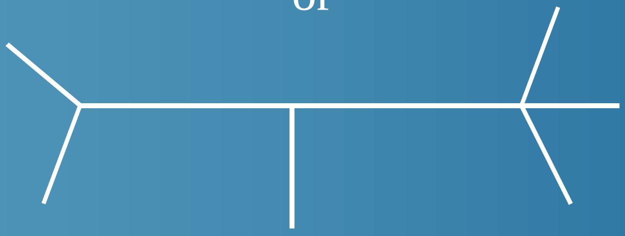
Alternative Methods of Drawing Trees

Trees can be drawn in many styles



These are Unrooted Trees

or



Outgroup

Outgroup - An OTU that is the least related to the group of taxa that you are studying.

Defining an outgroup is one way of rooting a tree.

How to Construct A Tree From Distance Data

There are two basic methods

- UPGMA
- Neighbor Joining

UPGMA

Unweighted Pair Gap Method with Arithmetic Mean

Simplest Method for Tree Construction

Sequential clustering method - Start with one pair of OTUs and sequentially add other OTUs

UPGMA Advantages & Disadvantages

Advantage - Fast, can handle many sequences

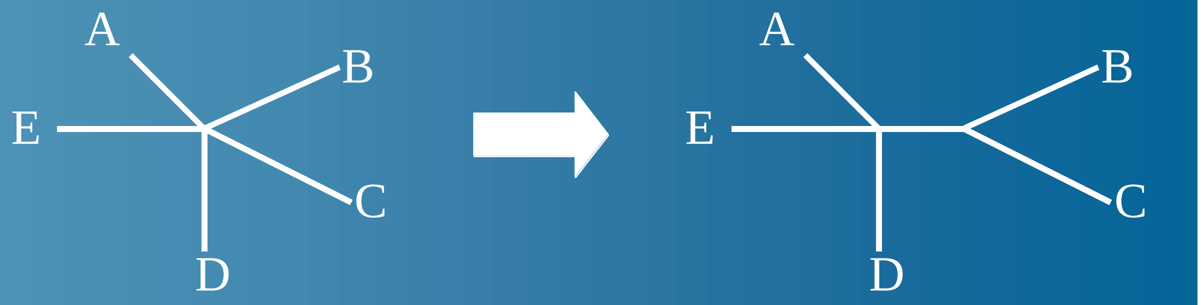
Disadvantage -

Gives different branch orders when rates of substitution vary greatly in different lineages.

Not used anymore with Clustal to create dendrogram
ClustalW now uses neighbor joining.

Neighbor Joining

Most commonly used method



Combine Nodes until you find a combination that gives the smallest sum of branch lengths.

Long Branch Attraction

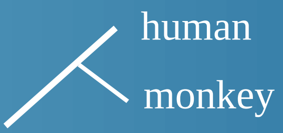
This is a good diagnostic for tree-building errors.
Various errors such as misalignment will cause this.

Symptom -

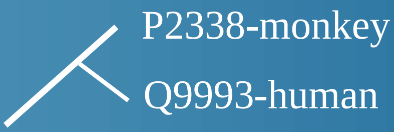
Rapidly evolving sequences (with long branches), will be placed with other rapidly evolving sequences even if the sequences are only distantly related.

Species Tree vs Gene Trees

Species Tree - a tree that shows the evolution of a species.



Gene Tree - a tree that shows the evolution of a gene.



Be careful using one gene to infer the evolution of a species. Certain genes may evolve independently of the species evolution.

Palentological Evidence

How do you know if your tree is correct?

Look at the palentological record

This is the only real proof you have
that a phylogenetic tree for a species
is correct.

Maximum Parsimony

Parsimony - the minimum number of means to achieve an end.

The most parsimonious tree, or shortest tree is one that requires the fewest total evolutionary events (for example, substitutions).

Parsimony methods may lead to an incorrect tree if the amount of evolutionary change is sufficiently divergent in different branches.

Occam's Razor

Given a choice between

- a hard way of doing things
- an easy way of doing things

Nature will always pick the easiest way
Simple is always preferred over complex.

This is the underlying philosophy behind parsimony

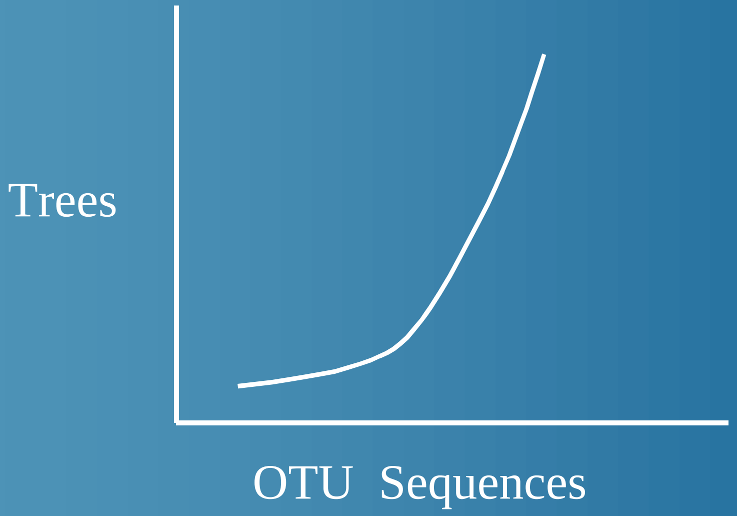
Multiple Trees

Distance methods produce only one “optimal” tree

Character methods can produce more than one tree that is “optimal”

If the data does not contain enough phylogenetically informative sites then you can get multiple trees.

How Many Trees?



Taxa	Trees
3	1
4	3
5	15
6	105
7	945
10	10^{6}
20	10^{21}
100	10^{182}
1000	10^{2860}

Finding the most parsimonious tree is computationally intensive if you use exhaustive search methods. A shortcut is needed.

Tree Construction Shortcuts

Heuristic Search -

Fastest method of finding trees
Uses a variation of Neighbor-Joining
Not guaranteed to find the optimal tree

Branch and Bound -

- Calculate an initial tree using N-J as a reference
- Now start building other partial trees.
- As you add limbs to the tree, calculate the “cost”
- If the cost is greater than the initial tree, stop
- Try building another tree

Statistical Methods to Evaluate Trees

Is there a way of objectively determining the best tree?

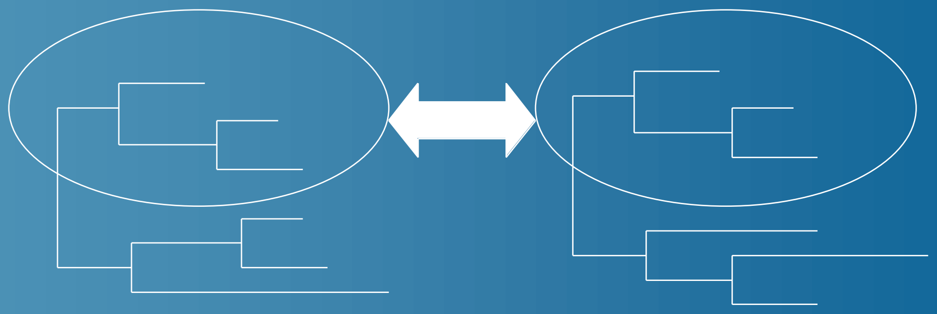
Bootstrapping - commonly used for estimating statistics when the distribution is difficult to derive analytically.

Method - resample and reanalyze single row of characters
Look for groupings that appear frequently as a measure of confidence in a particular tree.

Jacknife - Remove one sequence and reanalyze

Consensus Trees -

Consensus Tree



If you get multiple trees, look for regions that are similar. Those are the regions that you can be more confident are correct.

PAUP

PAUP - Phylogenetic Analysis Using Parsimony

Common software program for analyzing sequences using parsimony.   Can also create trees using distance methods.

Mac & PC versions.
UNIX version is part of GCG command line & X