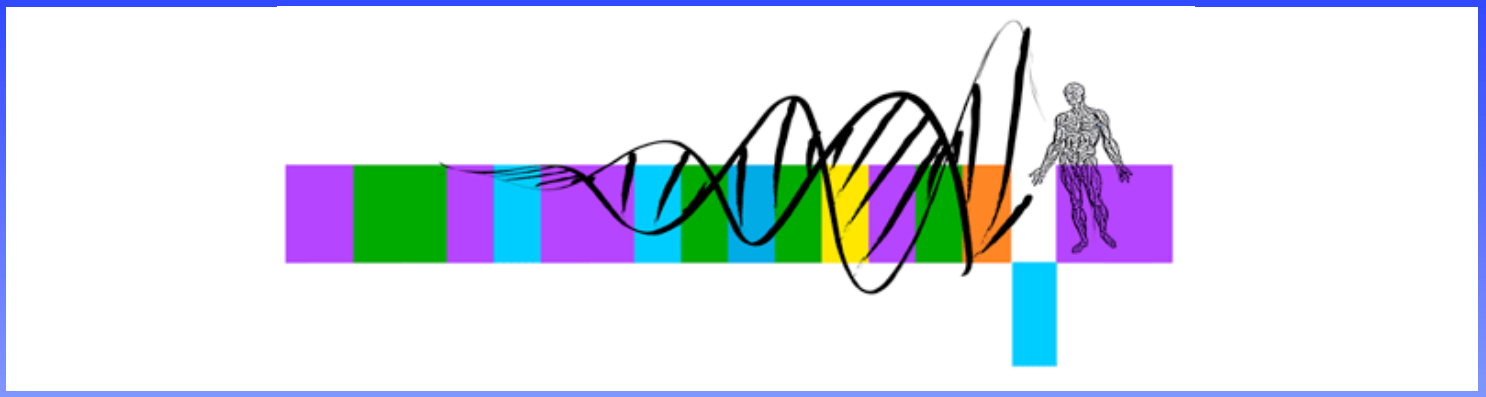


Getting To Know Your Protein

Comparative Protein Analysis:

Part I. Phylogenetic Trees and Multiple Sequence Alignments

Robert Latek, PhD
Sr. Bioinformatics Scientist
Whitehead Institute for Biomedical Research

Meeting Your Protein *In Silico*

- Define and characterize your favorite sequence
  - Identify homologous sequences
  - Predict function
  - Examine potential mutations
  - Study in 3D
  - Make manuscript figures :-)

Comparative Protein Analysis

Definition

Use information regarding a group of sequences to determine the function of an undefined sequence.

Extract novel information about a protein, or a series of proteins, through comparisons with other, related sequences.

Application

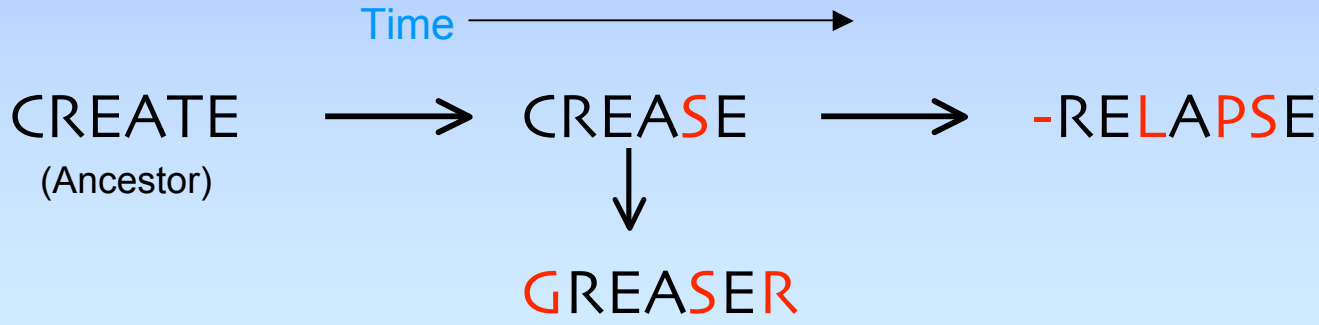
What are they?

What are their functions?

Why are they important?

Comparative Protein Analysis

- Identify proteins within an organism that are related to each other and across different species
- Generate an evolutionary history of related genes
- Locate insertions, deletions, and substitutions that have occurred during evolution



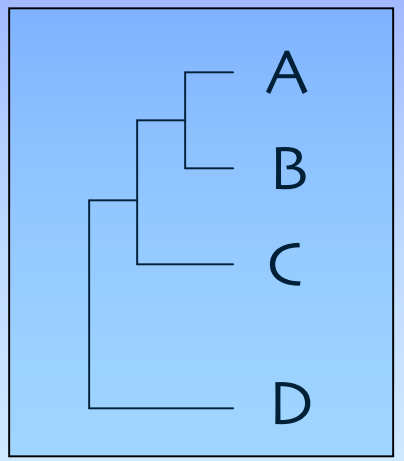
Syllabus

- **Phylogenetic Trees**
- Multiple Sequence Alignments
- From Trees and MSAs to Manuscript Figures
- Exercises

Phylogenetic Trees

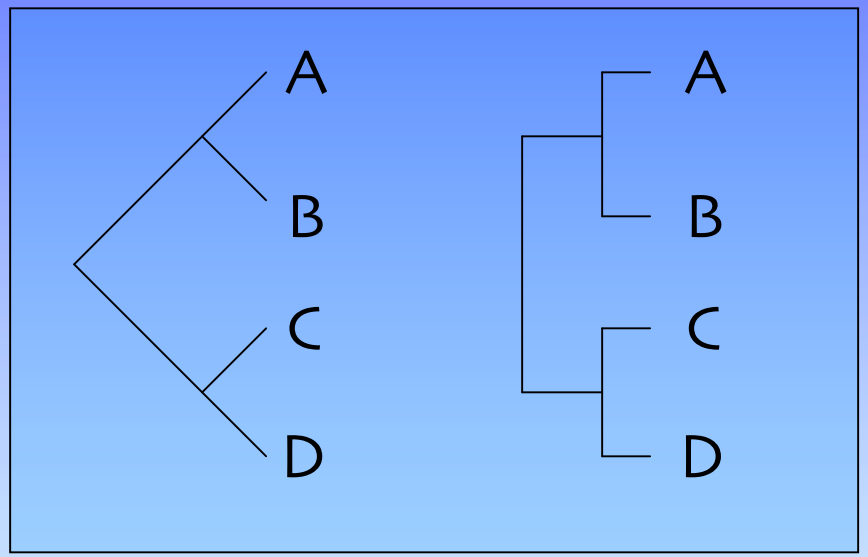
- A graph representing the evolutionary history of a sequence
- Relationship of one sequence to other sequences
- Dissect the order of appearance of insertions, deletions, and mutations
- Predict function, observe epidemiology, analyzing changes in viral strains
- Tree of Life http://tolweb.org/tree/phylogeny.html

Simple Tree

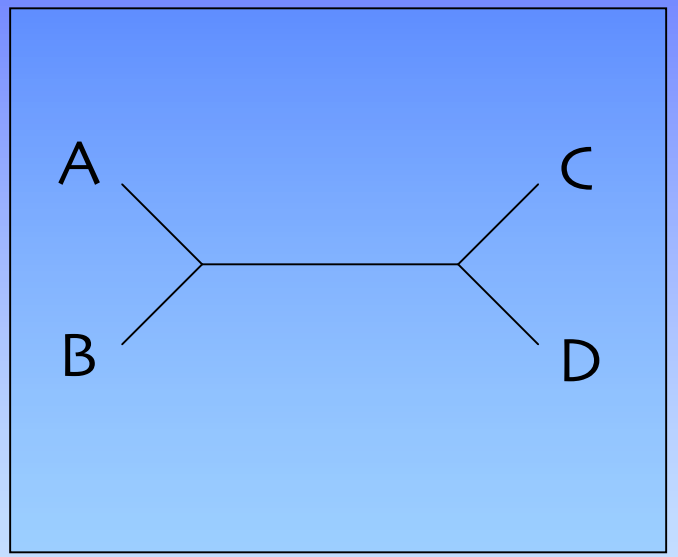


Tree Shapes

Rooted



Un-rooted

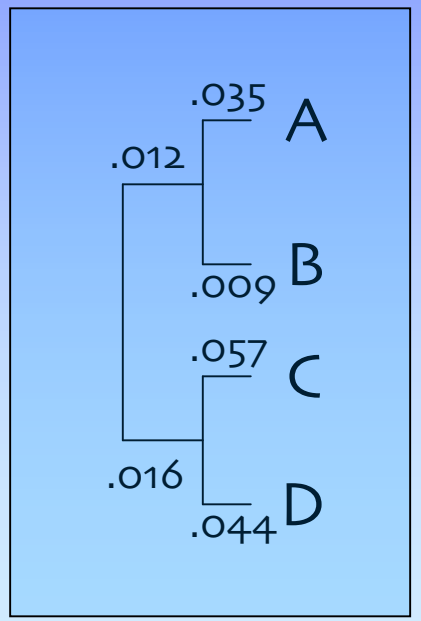


Branches intersect at Nodes
Leaves are the topmost branches

Tree Characteristics

- **Tree Properties**
  - **Clade**: all the descendants of a common ancestor represented by a node
  - **Distance**: number of changes that have taken place along a branch
- **Tree Types**
  - **Cladogram**: shows the branching order of nodes
  - **Phylogram**: shows branching order and distances

Phylogram



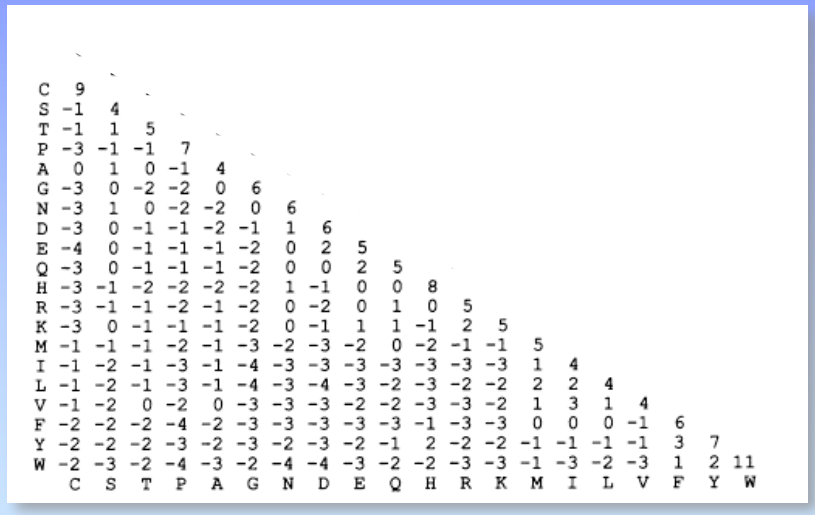
Tree Building Algorithms

- Maximum Parsimony
- Distance Methods
  - UPGMA
  - Neighbor Joining
- Maximum Likelihood

Amino Acid Substitutions

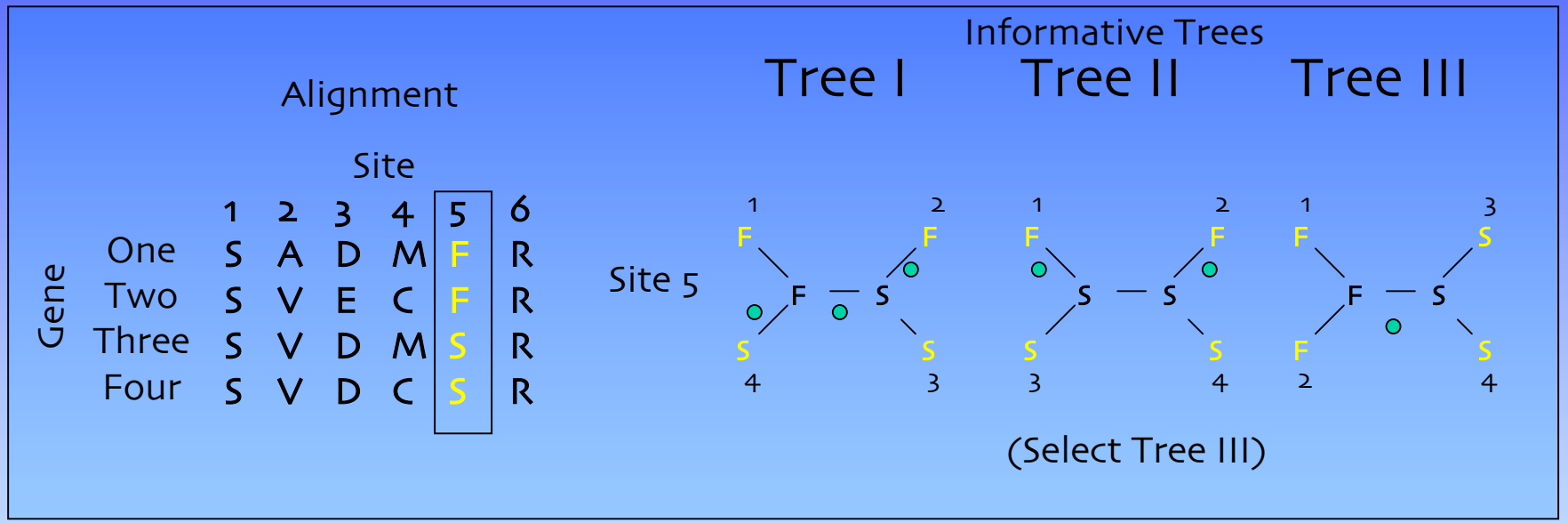
- Substitute one aa with another based on:
  - Chemical property
  - Size
  - Evolution (matrices)

BLOSUM 62
(Blocks Amino Acid Substitution Matrices)
(62% Identical Sequences)



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

Maximum Parsimony



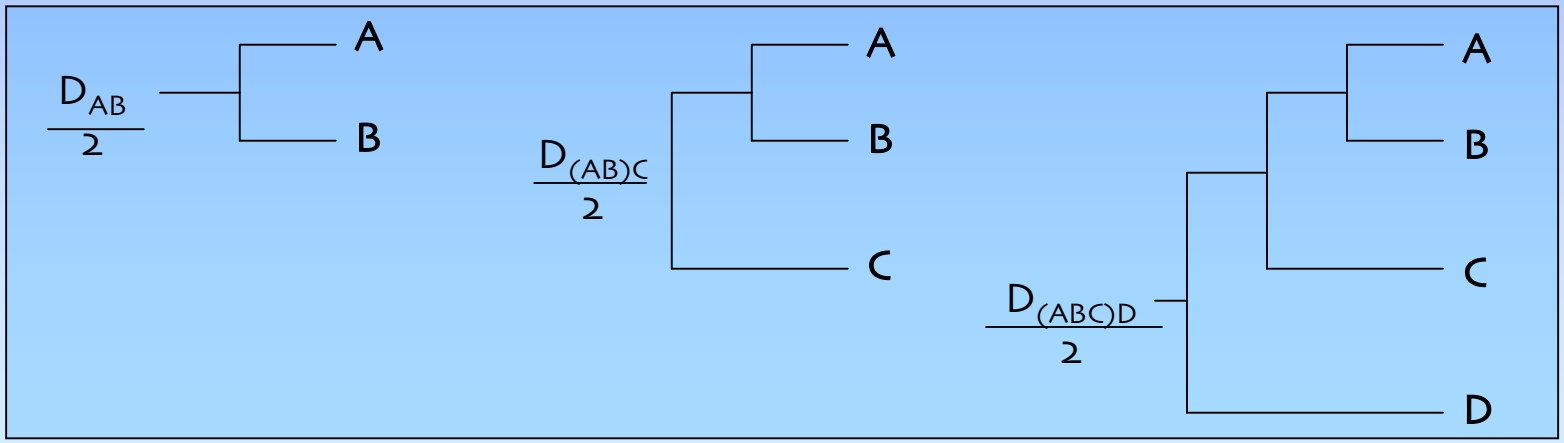
- Find the tree that changes one sequence into all of the others by the least number of steps [Focus solely on end product sequences, ignore evolutionary history]
- Only informative sites are analyzed (no gaps or conserved positions)
- Can be misleading when rates of change vary in different tree branches

Distance Methods

- **Distance** is expressed as the fraction of sites that differ between two sequences in an alignment
- Sequences with the smallest number of changes (shortest distance) are "related taxa"

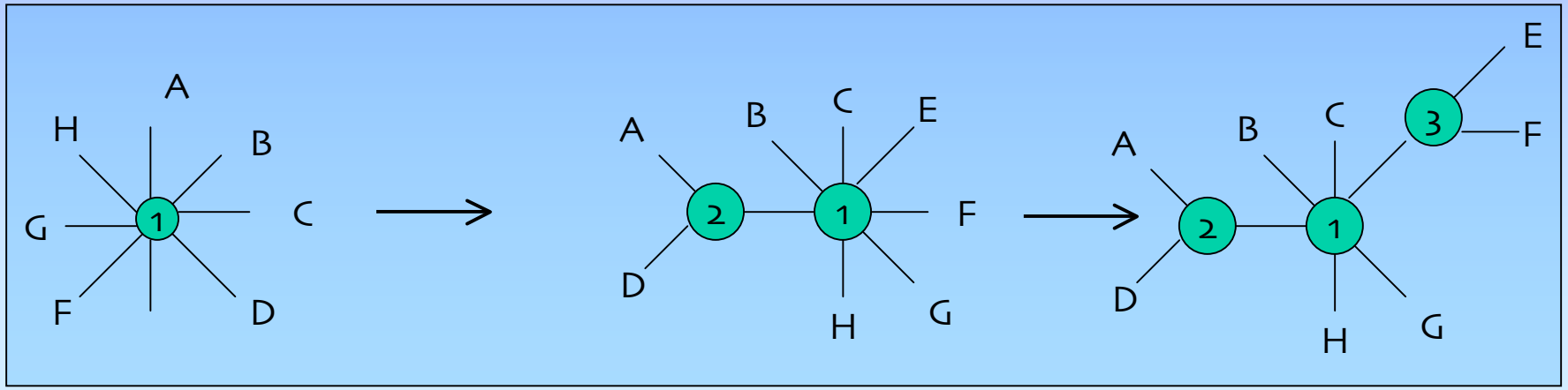
Distance Methods - UPGMA

- **UPGMA** (Unweighted Pair-Group Method with Arithmetic mean)
  - Sequentially find pair of taxa with smallest distance between them, and define branching as midpoint of two
  - Assumes the tree is additive and that rate of change is constant in all of the branches



Distance Methods - NJ

- **Neighbor-Joining** (NJ): useful when there are different rates of evolution within a tree
  - Each possible pair-wise alignment is examined. Calculate distance from each sequence to every other sequence
  - Choose the pair with the lowest distance value and join them to produce the minimal length tree
  - Update distance matrix where joined node is substituted for two original taxa and then repeat process



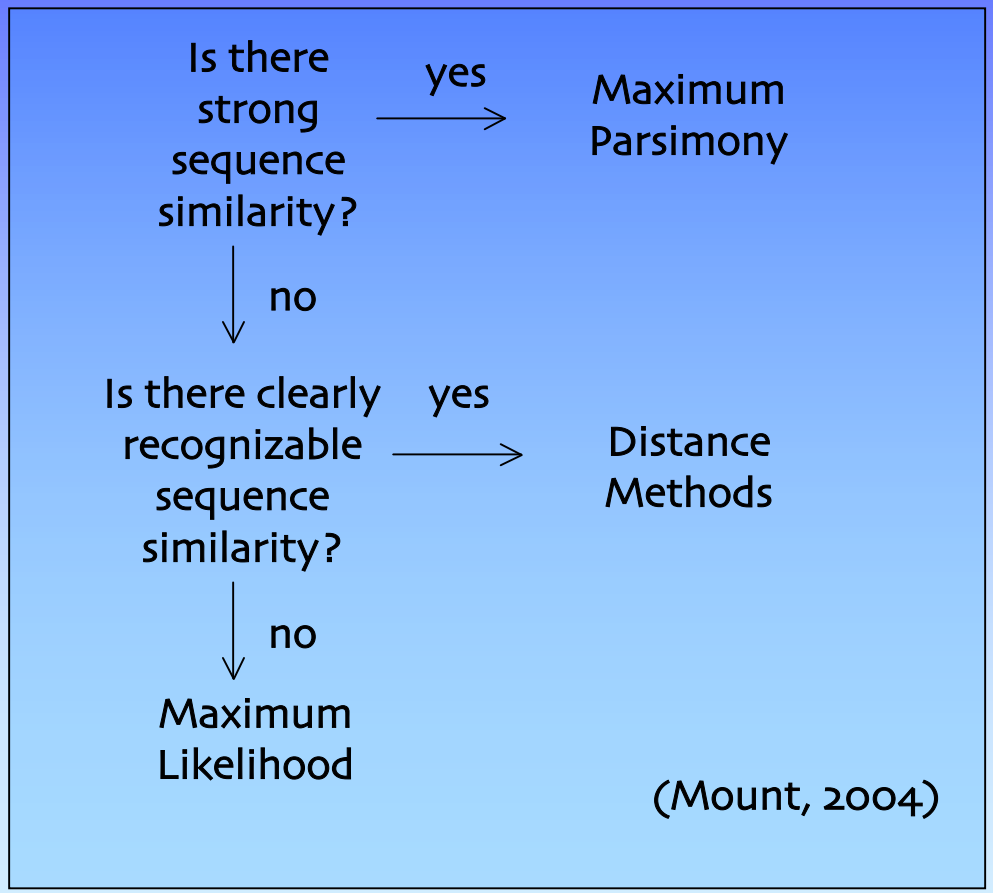
Maximum Likelihood

- Best accounts for variation in sequences
- Establish a **probabilistic model** with multiple solutions and determine which is most likely
- All possible trees are considered, therefore, only suitable for small number of sequences
  - Maximizes probability of finding optimal tree

Tree Reliability

- Probability that the members of a clade are always members of that clade
- Sample by **Bootstrapping**
  - Random sites of an alignment are randomly sampled so as to create a dataset the same size as the original. The same analysis as applied to the original data set is performed on the bootstrap dataset
  - Construct a consensus bootstrap tree and compare to the original tree

Which Method to Use?

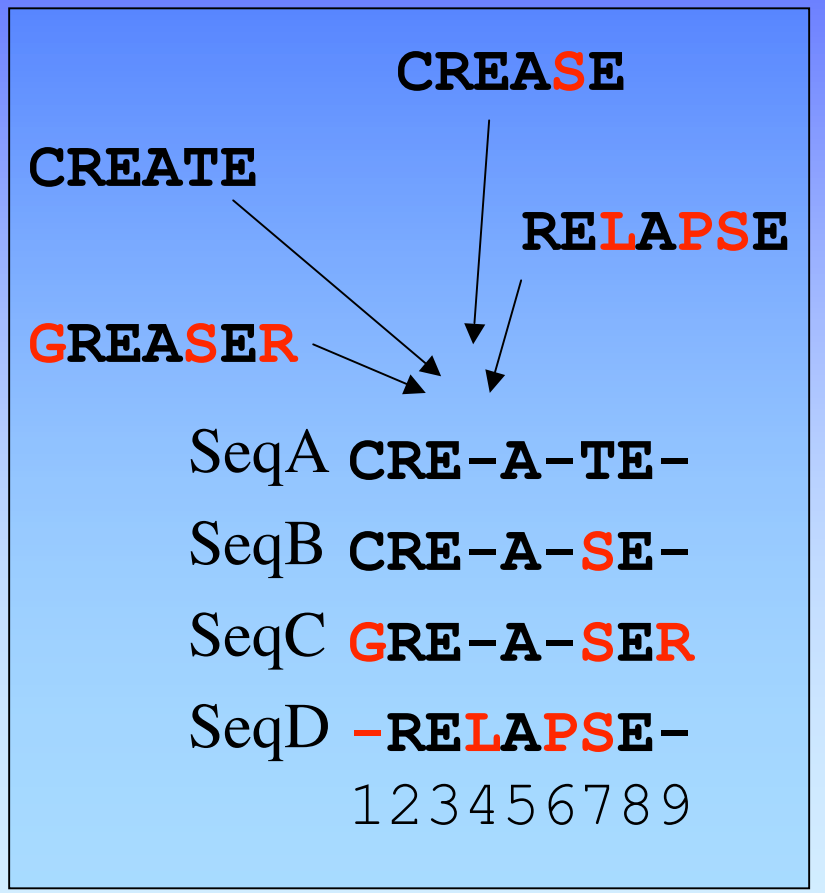


Syllabus

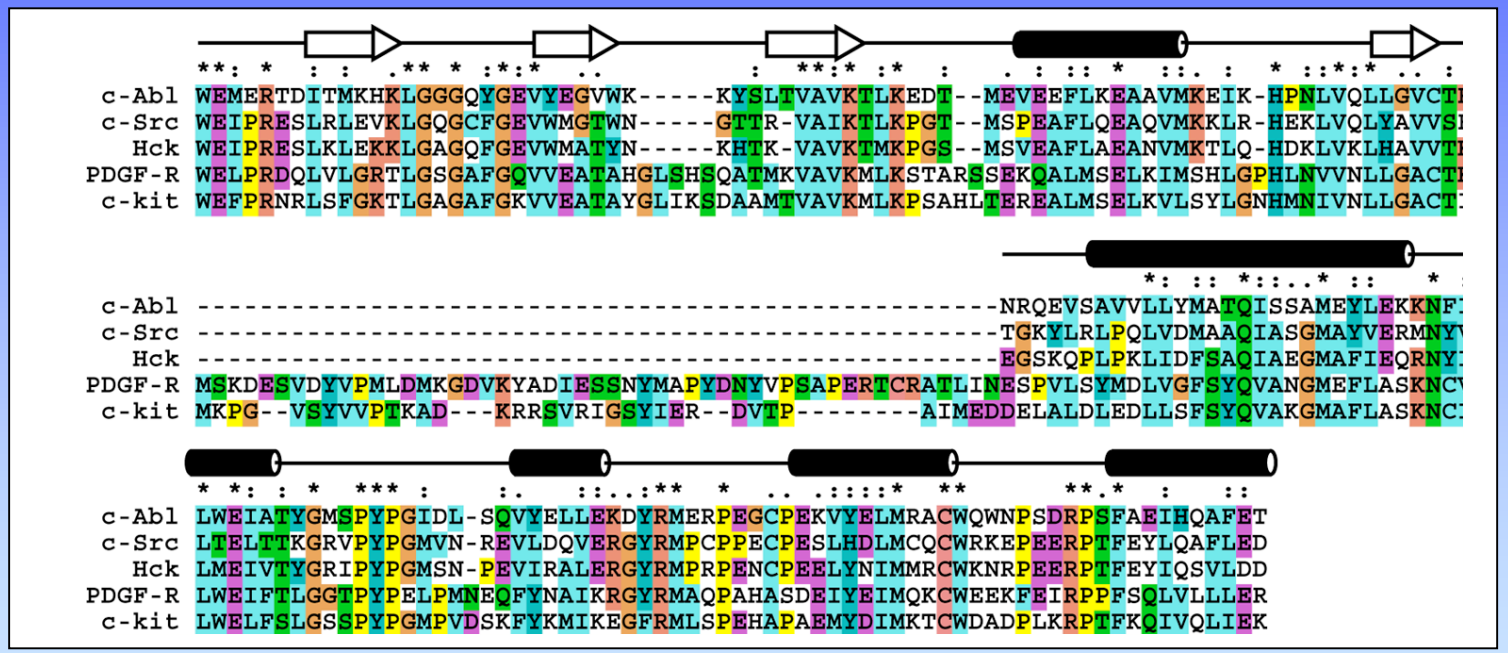
- Phylogenetic Trees
- **Multiple Sequence Alignments**
- From Trees and MSAs to Manuscript Figures
- Exercises

Multiple Sequence Alignments

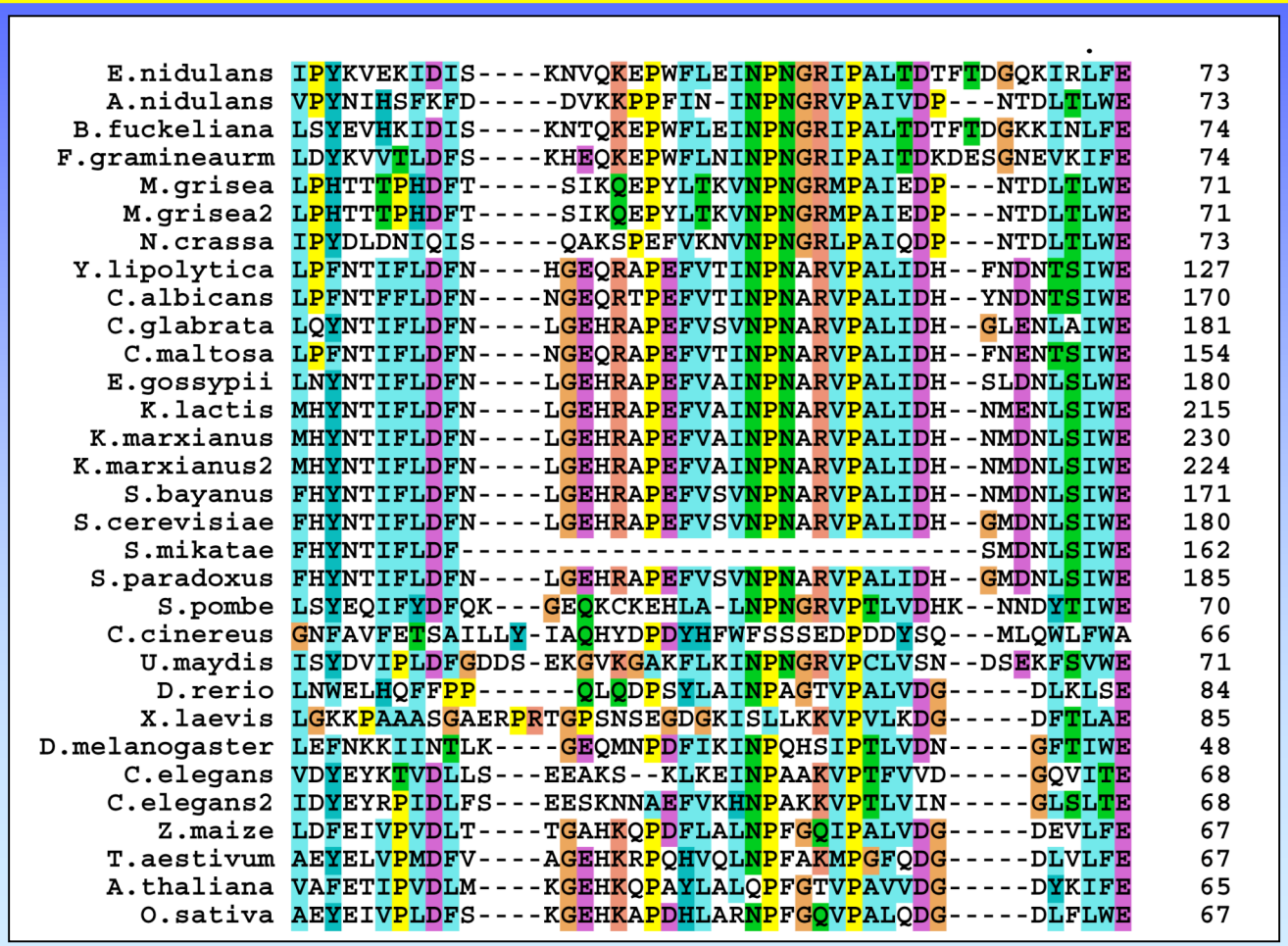
- Place residues in columns that are derived from a common ancestral residue
- MSA can reveal sequence patterns
  - Demonstration of homology between >2 sequences
  - Identification of functionally important sites
  - Protein function prediction
  - Structure prediction
  - Search for weak but significant similarities in databases
  - Design PCR primers for related gene identification
  - Genome sequencing: contig assembly



Multiple Sequence Alignment



Multiple Sequence Alignment



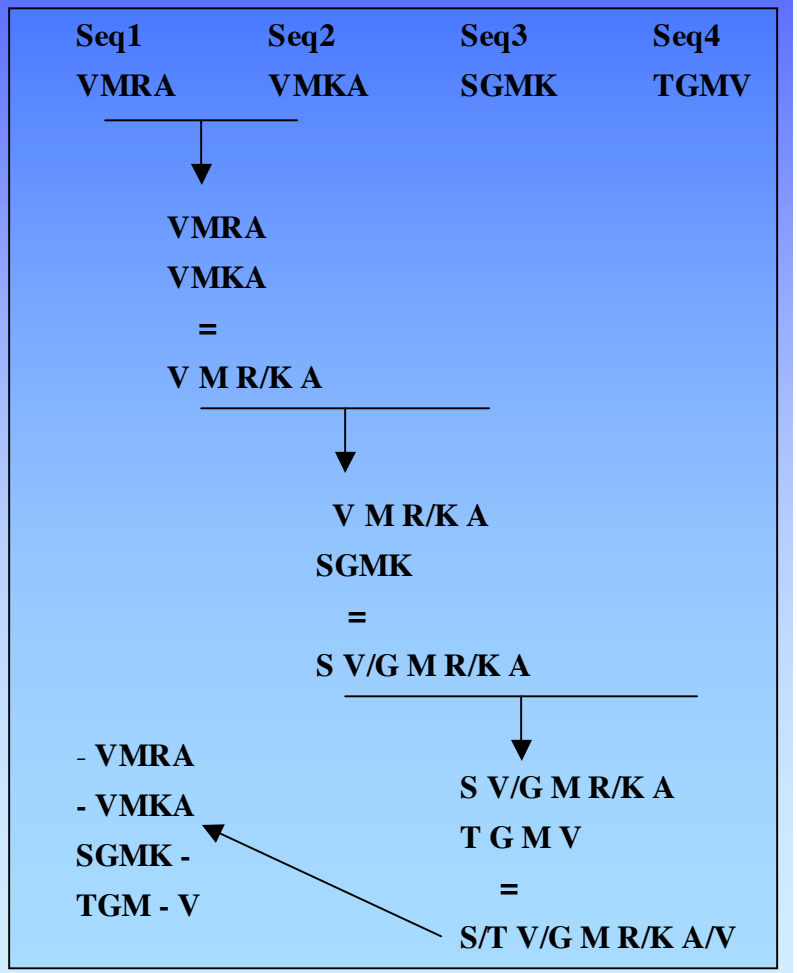
Approaches

- **Optimal Global Alignments** -Dynamic programming
  - Build matrices with every possible combination and search for optimal solution
  - Align 10 sequences of 100 aa length
  - Optimal in the mathematical sense

  $= 100^{10}$ possibilities
- **Global Progressive Alignments** - Match most common sequences together
- **Global Iterative Alignments** - Multiple re-building attempts to find best alignment
- **Local alignments**
  - Profiles, Blocks, Patterns

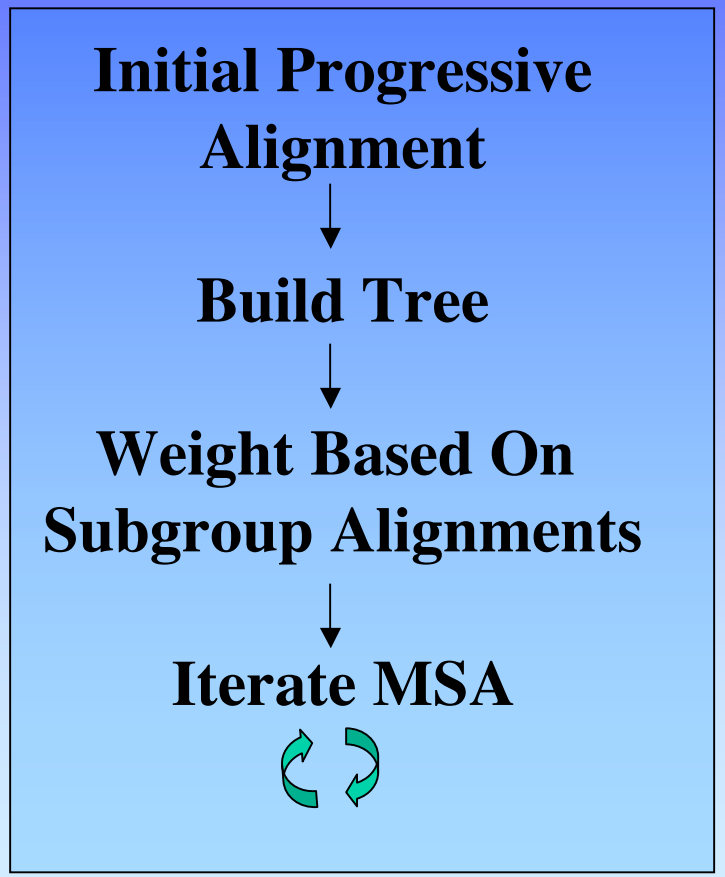
Global Progressive Alignment

- A heuristic approach that utilizes phylogenetic information to assist in routing the alignment (clustalw/clustalx)
- Feng & Doolittle1987, Higgins and Sharp 1988
- Most alike sequences are aligned together in order of their similarity (tree-based), a consensus is determined and then aligned to next most similar sequence



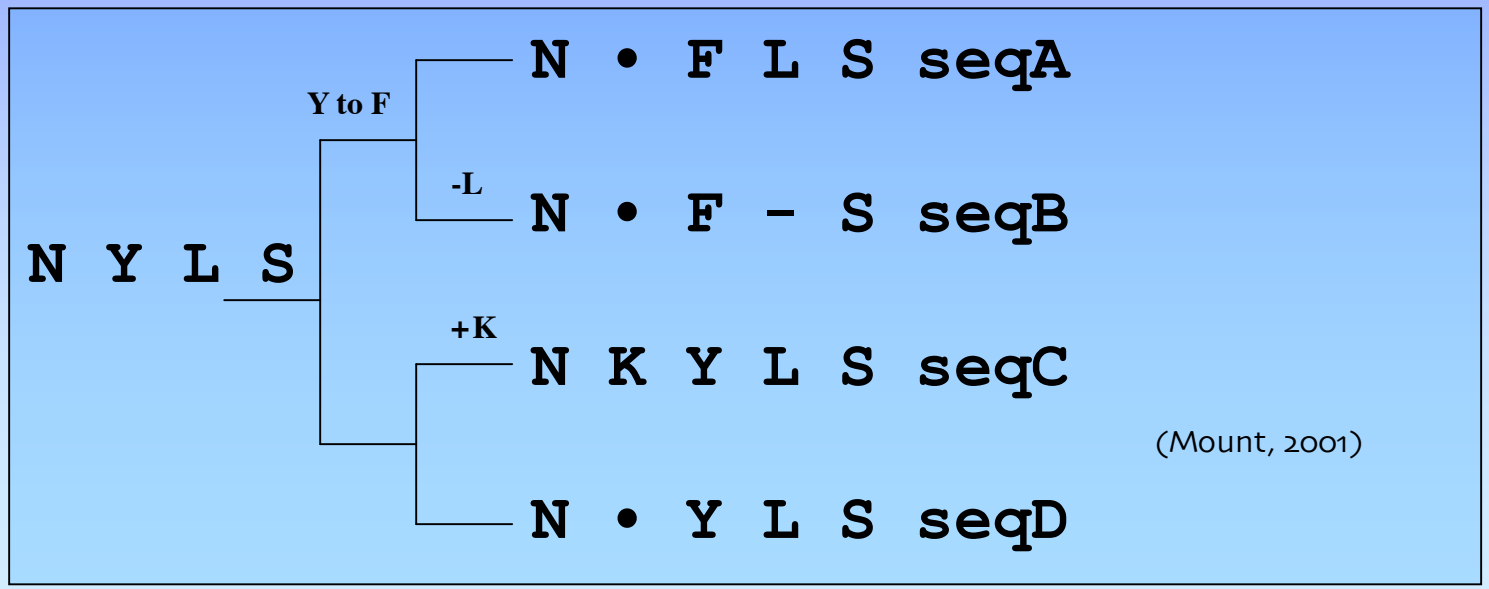
Iterative Multiple Alignment

- "Repeatedly re-align subgroups of sequences into a global alignment to improve alignment score" (Mount, 2004)
- Start with a progressive alignment and tree
- Recalculate pair-wise scores during progressive alignment, use new scores to rebuild the tree, which is used to improve alignments



MSA and Tree Relationship

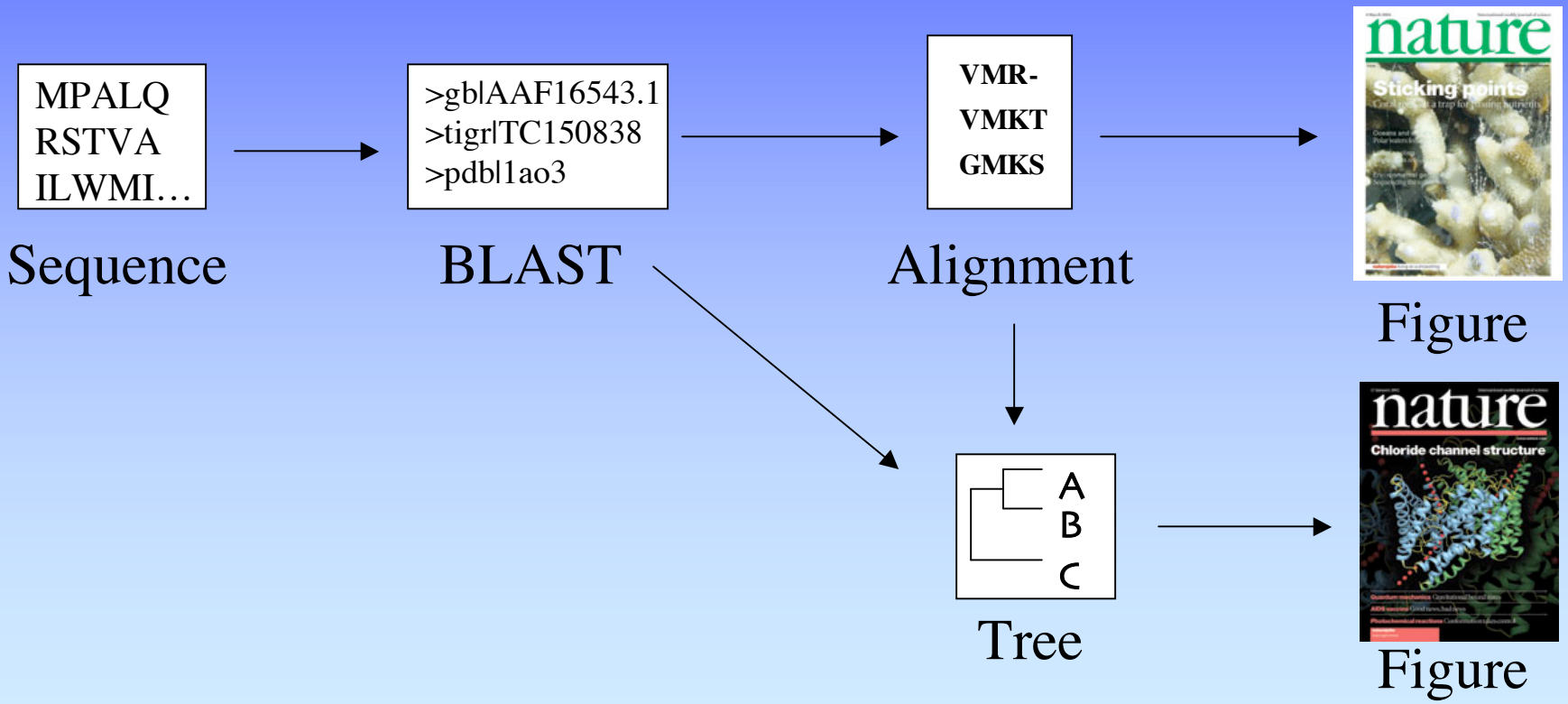
- "The optimal alignment of several sequences can be thought of as minimizing the number of mutational steps in an evolutionary tree for which the sequences are the leaves" (Mount, 2001)



Syllabus

- Phylogenetic Trees
- Multiple Sequence Alignments
- **From Trees and MSAs to Manuscript Figures**
- Exercises

Manuscript Figures 101



1. Find Related Sequences

- **BLAST**
  - www.ncbi.nih.gov/BLAST

MLEICLKLVGCKSKKGLSSSSSCYLEEALQRPVASDFEPQGLSEAARWNSKENLLAGPSENDPNLFVALY
DFVASGDNTLSITKGEKLRVLGYNHNGEWCEAQTKNGQGWVPSNYITPVNSLEKHSWYHGPVSRNAAEYL
LSSGINGSFLVRESESSPGQRSISLRYEGRVYHYRINTASDGKLYVSSESRFNTLAELVHHHSTVADGLI
TTLHYPAPKRNKPTVYGVSPNYDKWEMERTDITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDTMEV
EEFLKEAAVMKEIKHPNLVQLLGVCTREPPFYIITEFMTYGNLLDYLRECNRQEVNAVVLLYMATQISSA
MEYLEKKNFIHRDLAARNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAKFPIKWTAPESLAYNKFSIKS
DVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERPEGCPEKVYELMRACWQWNPSDRPSFAEIH
QAFETMFQESSISDEVEKELGKQGVRGAVSTLLQAPELPTKTRTSRRAAEHRDTTDVPEMPHSKGQGESD
PLDHEPAVSPLLPRKERGPPEGGLNEDERLLPKDKKTNLFSALIKKKKKTAPTPPKRSSSFREMDGQPER
RGAGEEEGRDISNGALAFTPLDTADPAKSPKPSNGAGVPNGALRESGGSGFRSPHLWKKSSTLTSSRLAT
GEEEGGGSSSKRFLRSCSASCVPHGAKDTEWRSVTLPRDLQSTGRQFDSSTFGGHKSEKPALPRKRAGEN
RSDQVTRGTVTPPPRLVKKNEEAADEVFKDIMESSPGSSPPNLTPKPLRRQVTVAPASGLPHKEEAGKGS
ALGTPAAAEPVTPTSKAGSGAPGGTSKGPAEESRVRRHKHSSESPGRDKGKLSRLKPAPPPPPAASAGKA
GGKPSQSPSQEAAGEAVLGAKTKATSLVDAVNSDAAKPSQPGEGLKKPVLPATPKPQSAKPSGTPISPAP
VPSTLPSASSALAGDQPSSTAFIPLISTRVSLRKTRQPPERIASGAITKGVVLDSTEALCLAISRNSEQM
ASHSAVLEAGKNLYTFCVSYVDSIQQMRNKFAFREAINKLENNLRELQICPATAGSGPAATQDFSKLLSS
VKEISDIVQR

2. Compile & Align Sequences

- **Compile** Sequences into FASTA format

```
>Human
MPALGYKFSTW…
>Mouse
MDGSTDYGILQINS…
>Rat
MKKP..
>Murine_Leukemia_Virus
MTSR….
```

- **Align**
  - PC: www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html
  - OS X: www.embl.de/~chenna/clustal/darwin/
  - Web: pir.georgetown.edu/pirwww/search/multaln.html
  - Jalview: http://www.jalview.org/Web_Installers/install.htm

3. Build Tree

- **Create** tree
  - Clustalx Neighbor Joining method
- **Draw** tree
  - TreeView:
    - taxonomy.zoology.gla.ac.uk/rod/treeview.html
  - Web:
    - iubio.bio.indiana.edu/treeapp/treeprint-form.html

4. Create Figures

- MSAs are often multipage
  - **Convert** to PDF with **Acrobat Distiller** or open with **Ghostview** (http://www.cs.wisc.edu/~ghost/ or http://www.kiffe.com/macghostview.html)
  - Extract pages individually and save as separate PDF/PS files
  - **Open** images in favorite illustration program
  - **Export** annotated alignments/trees to Powerpoint

- **Publish** paper, give award-winning presentation!

Exercise I

- **BLAST your sequence**
  - NCBI BLAST
  - Collate and edit sequences in a text editor
- **Perform multiple sequence alignment**
  - Clustalx
- **Build Phylogenetic Tree**
  - Clustalx and TreeView
- **Manage Postscript Files**
  - Adobe Acrobat Distiller/Ghostview
- **Create Figure**
  - Illustrator - > Powerpoint

References

- Bioinformatics: Sequence and genome Analysis. David W. Mount. CSHL Press, 2001 and 2004.
- Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Andreas D. Baxevanis and B.F. Francis Ouellete. Wiley Interscience, 2001.
- Bioinformatics: Sequence, structure, and databanks. Des Higgins and Willie Taylor. Oxford University Press, 2000.