Bioinformatics

Proteins I. - Phylogenetic Trees
&
Multiple Sequence Alignments

Robert Latek, Ph.D.
Bioinformatics, Biocomputing

Proteins I.-III. - Syllabus

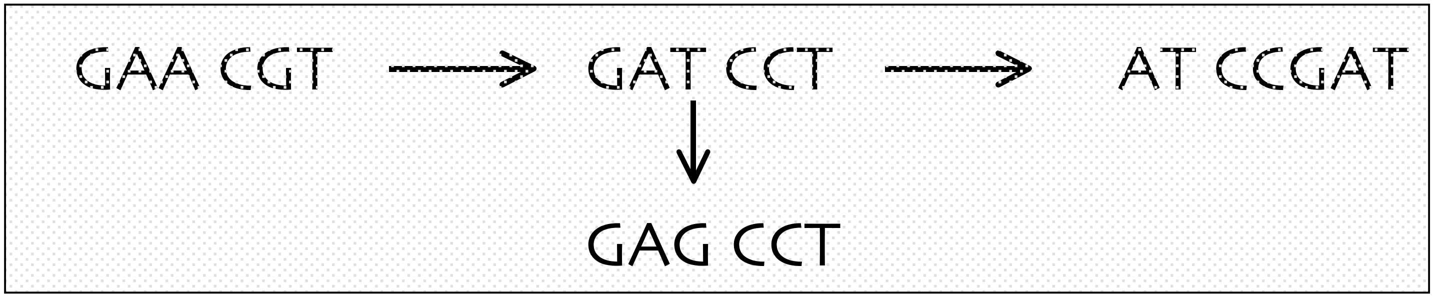
- Proteins I.
  - Phylogenetic Trees
  - Multiple Sequence Alignments
- Proteins II.
  - Profiles, Motifs, and Database Searches
  - Working with Protein Structures
- Proteins III.
  - Comparing Protein Structures
  - Building Structural Models

Proteins I. - Syllabus

- Relationship of MSA and Trees
- Phylogenetic Trees
  - Approaches: Maximum Parsimony, Distance, Maximum Likelihood
- Multiple Sequence Alignments
  - Approaches
    - Global: Dynamic Programming, Progressive, Iterated
    - Local: Profiles, Block-Based, Motif-Based
- Pointers & Demo

Comparative Genomics

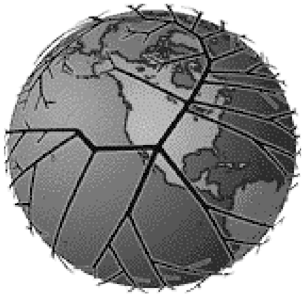
- Identify genes within an organism that are related to each other and across different species
- Generate an evolutionary history of related genes
- Locate insertions, deletions, and substitutions that have occurred during evolution



Homology

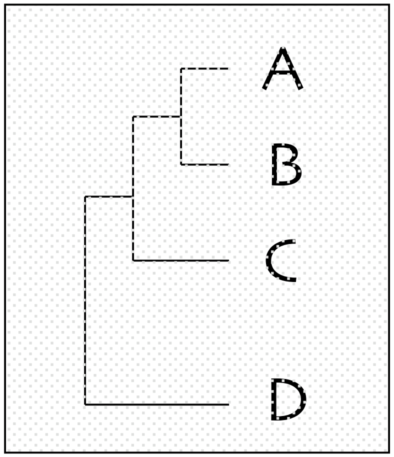
- Homologs: conserved sequences arising from a common ancestor
  - Orthologs: homologous genes that share a common ancestor in the absence of any gene duplication (speciation)
  - Paralogs: genes related through gene duplication (one gene is a copy of another)
- Similarity: genes that share common sequences but that are not necessarily related

Phylogenetic Trees



- A graphs 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

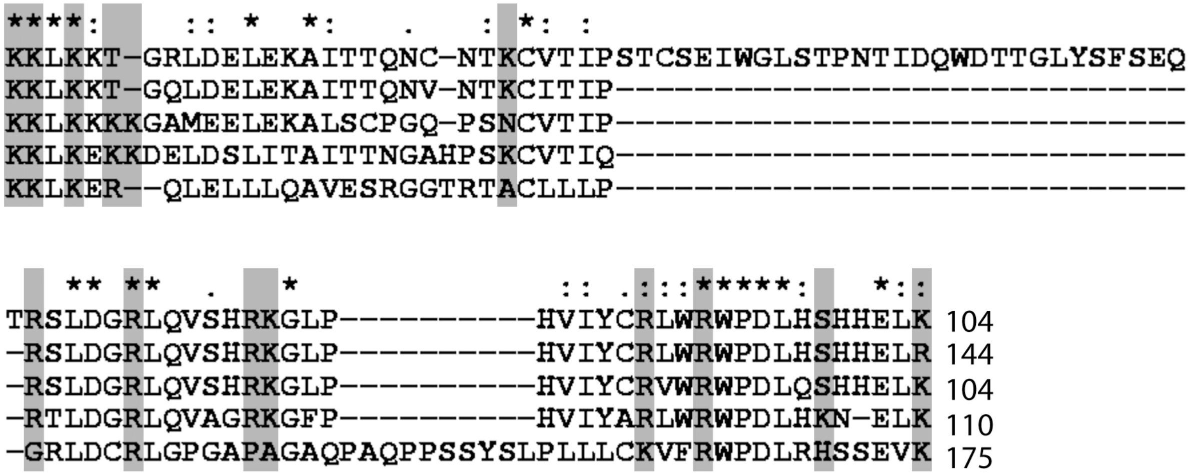
Simple Tree



Multiple Sequence Alignments (MSA)

- 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

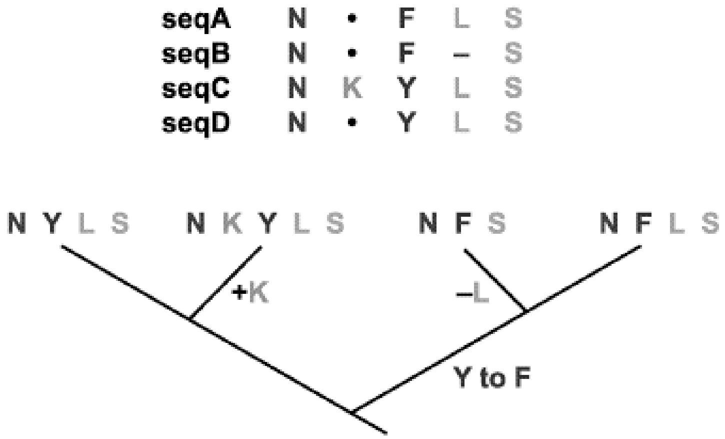


```
****:    :: *  *:   .   : *: :
KKLKKT-GRLDELEKAITTQNC-NTKCVTIPSTCSEIWGLSTPNTIDQWDTTGLYSFSEQ
KKLKKT-GQLDELEKAITTQNV-NTKCITIP-----------------------------
KKLKKKKGAMEELEKALSCPGQ-PSNCVTIP-----------------------------
KKLKEKKDELDSLITAITTNGAHPSKCVTIQ-----------------------------
KKLKER--QLELLLQAVESRGGTRTACLLLP-----------------------------

   ** **  .   *             :: .:::*****:   *::
TRSLDGRLQVSHRKGLP----------HVIYCRLWRWPDLHSHHELK 104
-RSLDGRLQVSHRKGLP----------HVIYCRLWRWPDLHSHHELR 144
-RSLDGRLQVSHRKGLP----------HVIYCRVWRWPDLQSHHELK 104
-RTLDGRLQVAGRKGFP----------HVIYARLWRWPDLHKN-ELK 110
-GRLDCRLGPGAPAGAQPAQPPSSYSLPLLLCKVFRWPDLRHSSEVK 175
```

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)

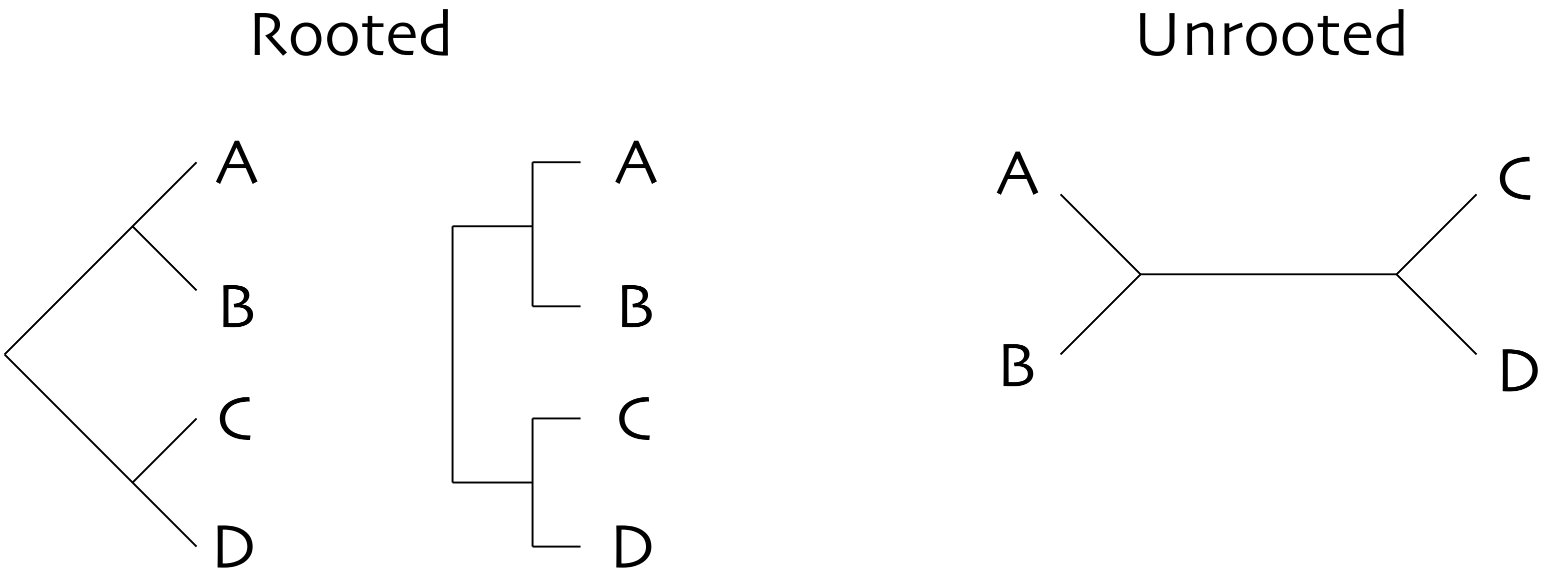


seqA	N	•	F	L	S
seqB	N	•	F	–	S
seqC	N	K	Y	L	S
seqD	N	•	Y	L	S

Proteins I. - Syllabus

- Relationship of MSA and Trees
- Phylogenetic Trees
  - Approaches: Maximum Parsimony, Distance, Maximum Likelihood
- Multiple Sequence Alignments
  - Approaches
    - Global: Dynamic Programming, Progressive, Iterated
    - Local: Profiles, Block-Based, Motif-Based
- Pointers & Demo

Tree Shapes



Rooted

Unrooted

Branches intersect at Nodes

Number of Possible Trees

Leaves	Rooted Trees	Unrooted Trees

(Li)

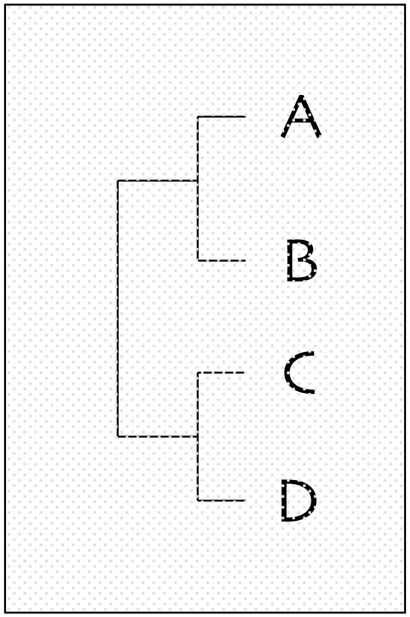
Tree Types

Cladogram: shows the branching order of nodes

Phylogram: shows branching order and distances

Distance: number of changes that have taken place along a branch

Clade: all the descendants of a common ancestor represented by a node



Tree Building Methods

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

Maximum Parsimony

- “Predicts a tree that minimizes the number of steps required to generate the observed variation between sequences” (Mount, 2001)
- Find the tree that changes one sequence into all of the others by the least number of steps
- Only informative sites are analyzed (not gaps or conserved positions)
- Can be misleading when rates of change vary in different tree branches

Maximum Parsimony - Example

Alignment

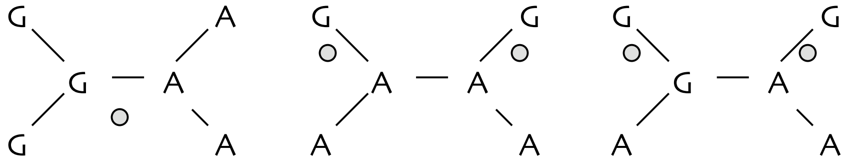
Site

Gene					Site	
One	A	A	G	A	G	T
Two	A	G	C	C	G	T
Three	A	G	A	T	A	T
Four	A	G	A	G	A	T

Informative Trees

Site

Tree I Tree II Tree III



(Select Tree I)

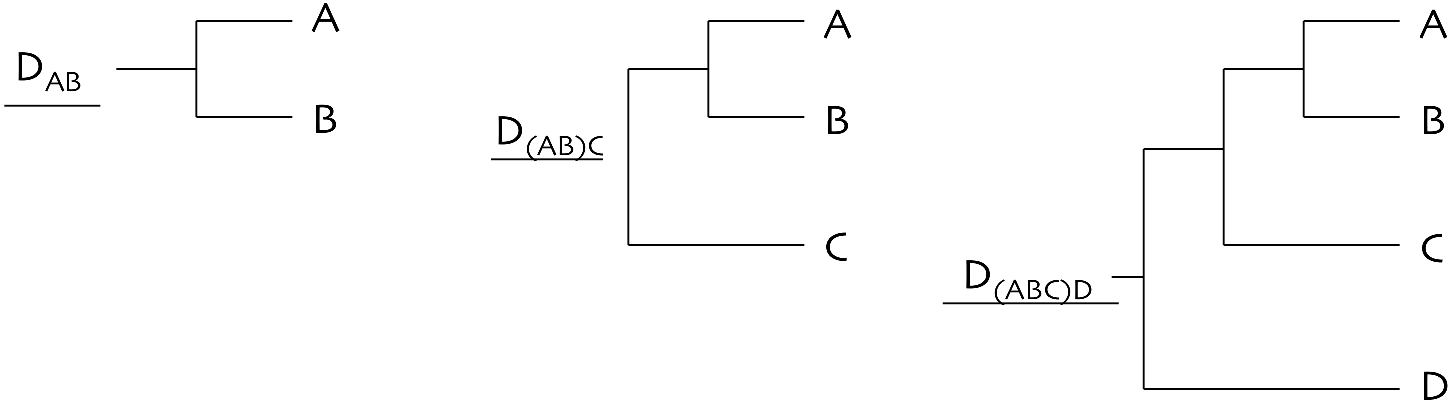
(Li)

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 “neighbors”
- Distance matrix: table generated with distance scores describing the number of changes needed to change one sequence to another
- Build the tree based on a distance matrix derived from multiple alignments

Distance Methods - UPGMA

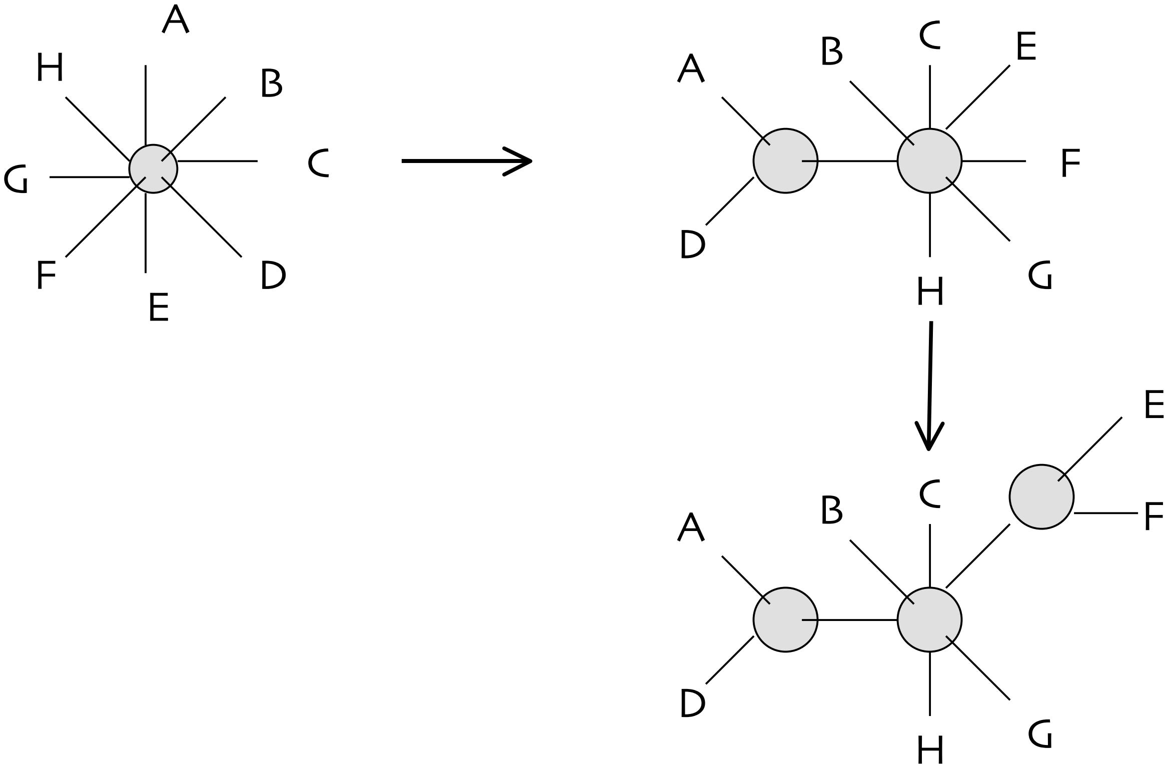
- UPGMA (Unweighted Pair-Group Method with Arithmetic mean)
  - Sequentially find pair of taxa with smallest distance between them, and define branching as midpoint of two
  - Assumes the tree is additive and that rate of change is constant in all of the branches



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
  - Repeat process

Distance Methods - NJ



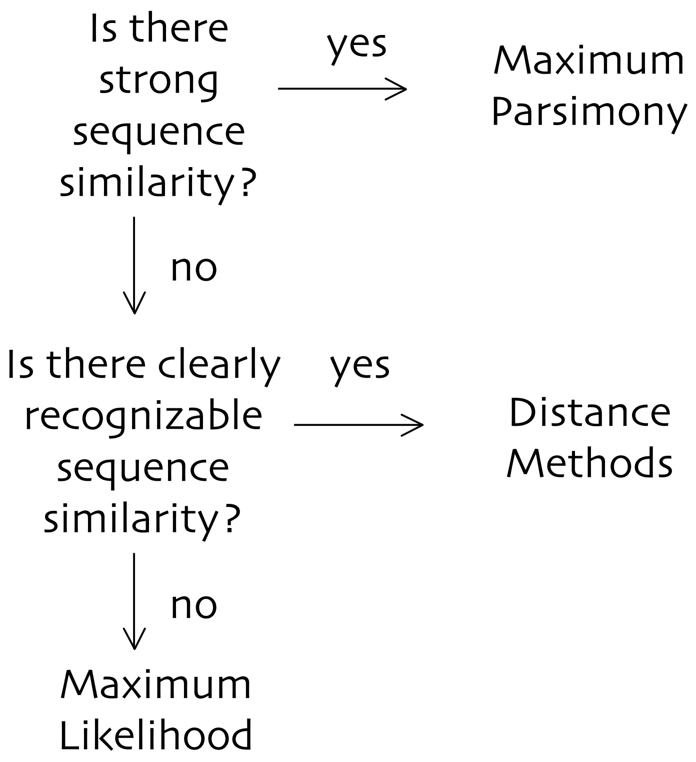
Maximum Likelihood

- 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 site of an alignment is taken for a pseudoalignment
  - A second site is taken randomly, etc. (can take some sites more than once, some not at all)
  - Each set is subjected to the same analysis as the original data set
  - Construct a consensus bootstrap tree with the pseudoalignments & compare to original tree

Tree Building - Methods



Is there strong sequence similarity?

yes → Maximum Parsimony

no ↓

Is there clearly recognizable sequence similarity?

yes → Distance Methods

no ↓

Maximum Likelihood

(Mount)

Proteins I. - Syllabus

- Relationship of MSA and Trees
- Phylogenetic Trees
  - Approaches: Maximum Parsimony, Distance, Maximum Likelihood
- Multiple Sequence Alignments
  - Approaches
    - Global: Dynamic Programming, Progressive, Iterated
    - Local: Profiles, Block-Based, Motif-Based
- Pointers & Demo

MSAs - Global vs. Local Alignments

- Global
  - Search for alignments, matching over entire sequences
- Local
  - Examine regions of sequence for conserved segments
- Matches, Mismatches, Gaps

Approaches

- Optimal Global Alignments
  - Dynamic programming
- Global Progressive Alignments
- Global Iterative alignments
- Local alignments
  - Profiles, Blocks, Patterns

Optimal Global Alignments (Dynamic Programming)

- Used for aligning a small number of sequences
- Build matrices with every possible combination and search for optimal solution
  - Optimal in the mathematical sense
  - Need to implement appropriate parameters
- Problem gets large quickly
  - Length raised to number of sequences
  - Align 10 sequences of 100 aa length

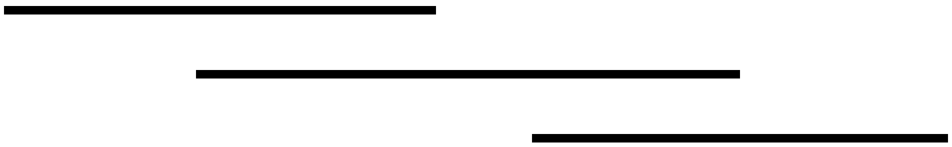
$$100^{10}$$

Global Progressive Alignments

- Compute alignment scores (distances) between all pairs of sequences from pair-wise alignments
- Build a guide tree using the pair-wise alignment distances (NJ)
- Align sequences sequentially following the tree
  - Align each node from leaves to root

Global Progressive Alignments Problems

- Dependence of MSA on the initial pair-wise alignments
- Improper scoring when aligning a set of sequences that have non-overlapping segments



Iterative Multiple Alignments

- “Repeatedly re-align subgroups of sequences into a global alignment to improve alignment score” (Mount, 2001)
- Selection of groups based on tree
- Recalculate pair-wise scores during progressive alignment, use new scores to rebuild the tree, which is used to improve alignments

Localized Alignments

- Profiles
  - A conserved portion of an alignment, includes scores for substitutions and gaps for the conserved region (consensus)
- Blocks
  - Conserved region without gaps
- Patterns
  - Motifs

Profile Analysis

- Perform global MSA on group of sequences
- Move highly conserved regions to smaller MSAs
- Generate scoring table with log odds scores
  - Each column is independent
  - Average Method: profile matrix values are weighted by the proportion of each amino acid in each column of MSA
  - Evolutionary Method: calculate the evolutionary distance (Dayhoff model) required to generate the observed amino acid distribution

Profile Example

C O N	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
I	8	-2	5	4	5	5	-4	**24**	0	15	13	1	1	1	-7	2	22	21	-18	-6
T	13	-5	24	18	-18	19	7	1	7	-7	-4	14	11	10	-1	9	**29**	3	-28	-14
L	5	-5	3	4	13	4	2	8	-4	**14**	12	8	-5	0	-10	0	10	10	-1	5
S	17	17	13	10	-12	29	-5	-5	6	-14	-9	12	10	0	-2	**34**	19	1	-8	-15

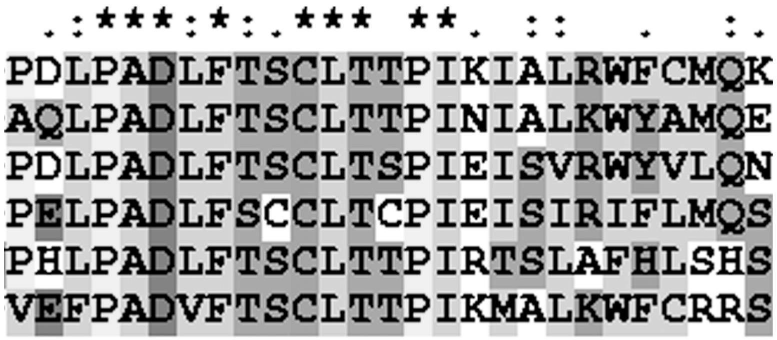
Cons.	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
I	8	-2	5	4	5	5	-4	24	0	15	13	1	1	1	-7	2	22	21	-18	-6
T	13	-5	24	18	-18	19	7	1	7	-7	-4	14	11	10	-1	9	29	3	-28	-14
L	5	-5	3	4	13	4	2	8	-4	14	12	8	-5	0	-10	0	10	10	-1	5
S	17	17	13	10	-12	29	-5	-5	6	-14	-9	12	10	0	-2	34	19	1	-8	-15

(Mount)

Block Analysis

- Represent a conserved region within a MSA
- Contain matches, mismatches, but no gaps
- Serve as anchors to assist in aligning sequences by aligning individual segments

Block Analysis



```
 .:***:*:.*** **. ::  .  :.
PDLPADLFTSCLTTPIKIALRWFCMQK
AQLPADLFTSCLTTPINIALKWYAMQE
PDLPADLFTSCLTSPIEISVRWYVLQN
PELPADLFSCCLTCPIEISIRIFLMQS
PHLPADLFTSCLTTPIRTSLAFHLSHS
VEFPADVFTSCLTTPIKMALKWFCRRS
```

Patterns (Motifs)

- Find patterns like aa1 d1 aa2 d2 aa3
  - Definition of a motif (non-contiguous)
  - Remove low complexity regions
  - Validate with a statistical method
  - i.e. Docking site of a kinase to a receptor

X	X	X	D	A	X	X	X	X	R	Y	X	X	X
			E	V					K	F			
				L					H	W			
				I						H			

Pointers

- When to use each method?
  - – Related sequences = global progressive
  - – Divergent sequences = local alignments
- Applications
  - – MSA viewers: ClustalX, Jalview, Belvu
  - – Annotation: Boxshade, Seqvu, MACAW, PUZZLE

File Formats

- MSF
  - http://web.wi.mit.edu/proteins/education/seq.msf
- ALN
  - http://web.wi.mit.edu/proteins/education/seq.aln
- PIR
  - http://web.wi.mit.edu/proteins/education/seq.pir
- DND
  - http://web.wi.mit.edu/proteins/education/seq.dnd
- PH
  - http://web.wi.mit.edu/proteins/education/seq.ph

Next Week

- Proteins I.
  - Phylogenetic Trees
  - Multiple Sequence Alignments
- Proteins II.
  - Profiles, Motifs, and Database Searches
  - Working with Protein Structures
- Proteins III.
  - Comparing Protein Structures
  - Building Structural Models

Demonstrations

- Multiple Sequence Alignments
  - Clustal (web-based)
    - http://pir.georgetown.edu/pirwww/search/multaln.html
  - ClustalX (local)
  - Jalview
- Tree Building
  - PAUP (UNIX-based)
  - ClustalX
  - Phylodendron http://iubio.bio.indiana.edu/treeapp/treeprint-form.html

References

- Bioinformatics: Sequence and genome Analysis. David W. Mount. CSHL Press, 2001.
- 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.