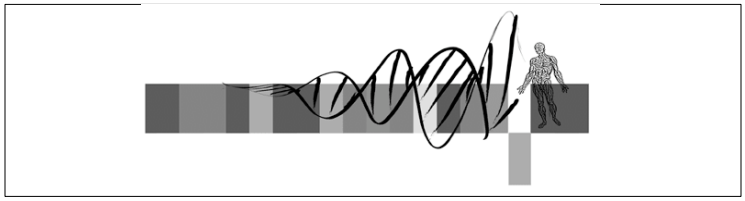


Bioinformatics for Biologists

Comparative Protein Analysis: Part I. Phylogenetic Trees and Multiple Sequence Alignments

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

Comparative Protein Analysis

Overview 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 definition

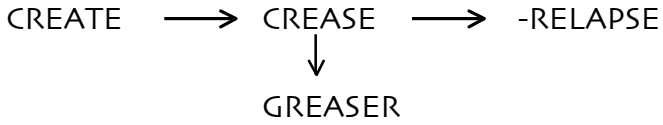
- What are they?
- What are their functions?
- Why are they important?

Syllabus

- **Comparative Protein Analysis**
- Phylogenetic Tree Techniques and Application
- Multiple Sequence Alignment Techniques and Application
- Demonstration - Putting Trees and MSAs to Work

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



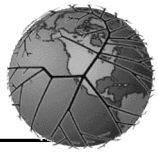
Homology

- **Homology**: 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 are not necessarily related

Syllabus

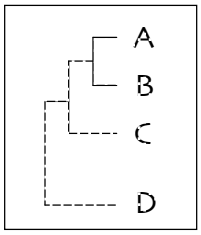
- Comparative Protein Analysis
- **Phylogenetic Tree Techniques and Application**
- Multiple Sequence Alignments Techniques and Application
- Demonstration - Putting Trees and MSAs to Work

Phylogenetic Trees



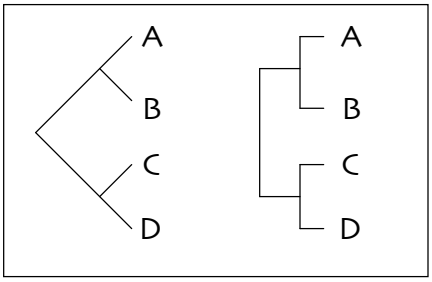
- A graph representing the evolutionary history of a sequence
- Relationship of one sequence to other sequences
- Dissect the order of appearance of insertions, deletions, and mutations
- Predict function, observe epidemiology, analyzing changes in viral strains

Simple Tree

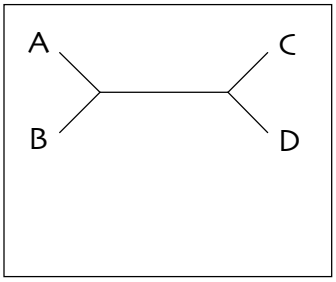


Tree Shapes

Rooted



Un-rooted



Branches intersect at Nodes
Leaves are the topmost branches

Number of Possible Trees

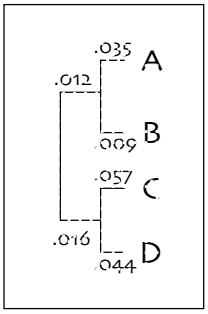
Leaves	Rooted Trees	Un-rooted Trees
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10,395	954
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025

(Li, 1997)

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

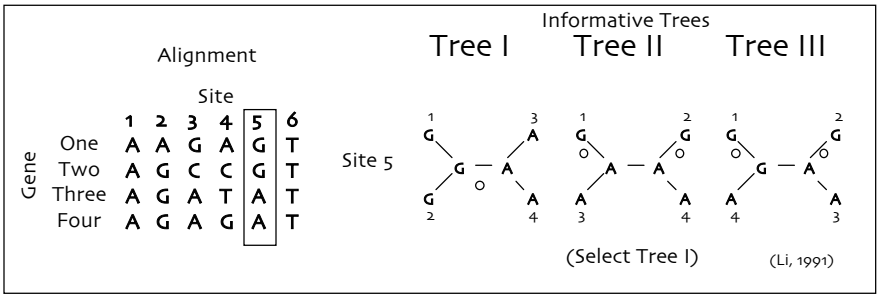
Phylogram



Tree Building Methods

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

Maximum Parsimony



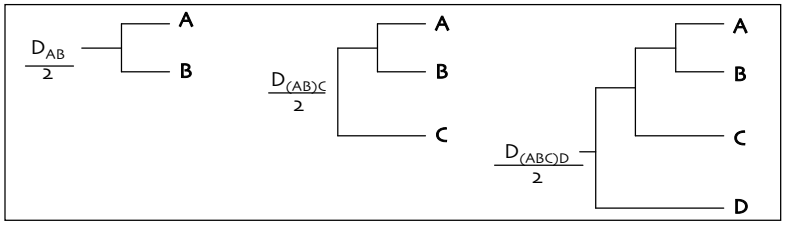
- Find the tree that changes one sequence into all of the others by the least number of steps
- Only informative sites are analyzed (no gaps or conserved positions)
- Can be misleading when rates of change vary in different tree branches

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"

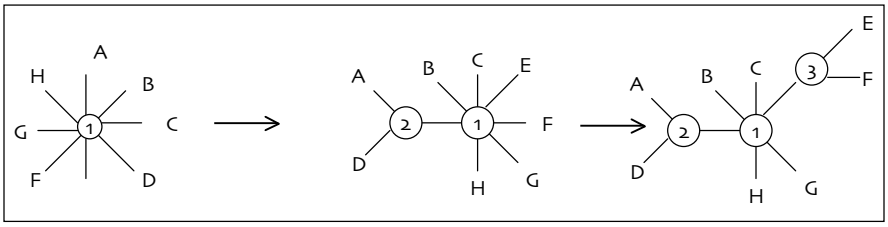
Distance Methods - UPGMA

- **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



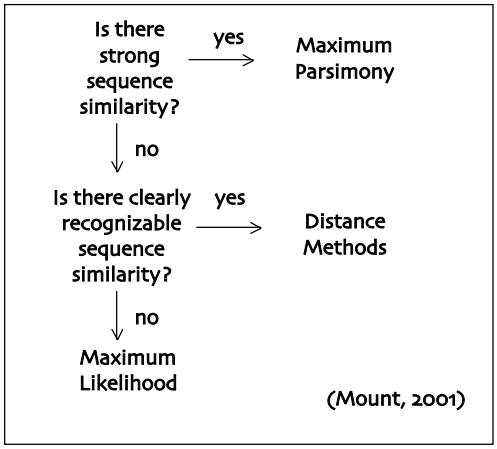
Maximum Likelihood

- 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?

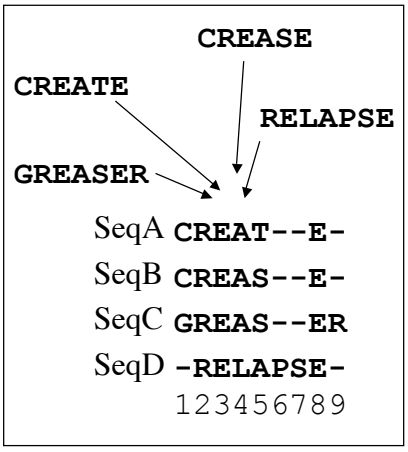


Syllabus

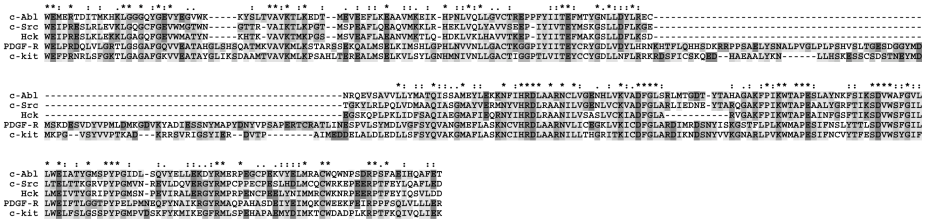
- Comparative Protein Analysis
- Phylogenetic Tree Techniques and Application
- **Multiple Sequence Alignments Techniques and Application**
- Demonstration - Putting Trees and MSAs to Work

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



Global vs. Local Alignments

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

Approaches

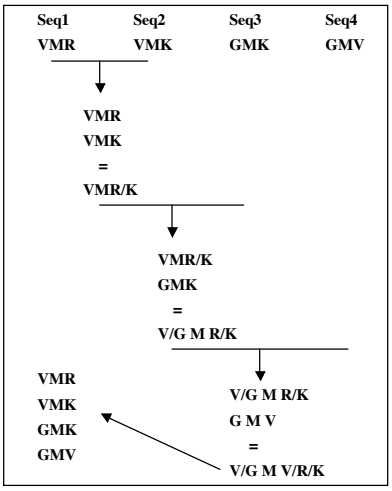
- Optimal Global Alignments
  - Dynamic programming
- Global Progressive Alignments
- Global Iterative Alignments
- Local alignments
  - Profiles, Blocks, Patterns

Optimal Global Alignments

- Dynamic programming is used for aligning a small number of sequences
- Build matrices with every possible combination and search for optimal solution
  - Optimal in the mathematical sense
- Problem gets large quickly
  - Length raised to number of sequences
  - Align 10 sequences of 100 aa length $\mathbf{100^{10}}$

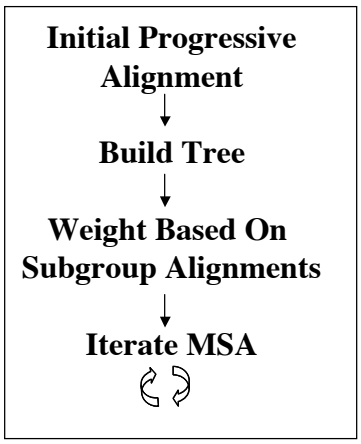
Global Progressive Alignment

- 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



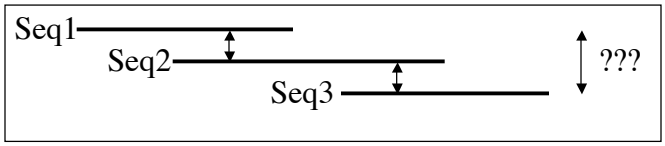
Iterative Multiple Alignment

- "Repeatedly re-align subgroups of sequences into a global alignment to improve alignment score" (Mount, 2001)
- 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



Global Alignment Errors

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

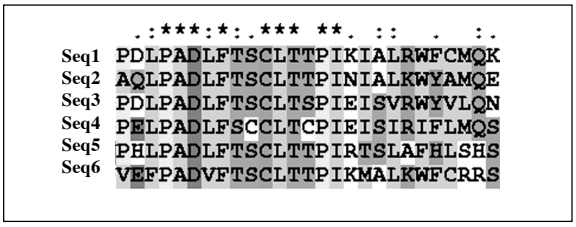


Localized Alignments

- Blocks
  - Conserved region without gaps
- Patterns
  - a deterministic syntax that describes multiple combinations of possible residues within a protein string
- Profiles
  - probabilistic generalizations that assign to every segment position, a probability that each of the 20 aa will occur. Includes scores for substitutions and gaps for the conserved region (consensus)

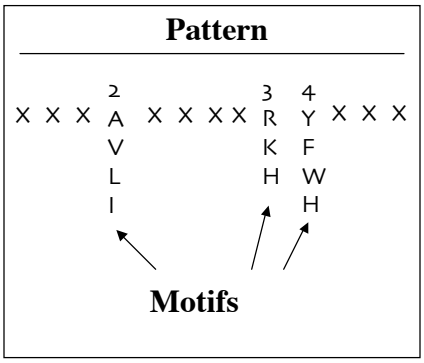
Block Analysis

- 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



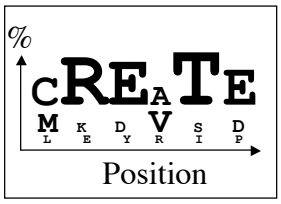
Patterns (Motifs)

- Patterns are a string of non-contiguous motifs
  - Remove low complexity regions
  - i.e. Docking site of a kinase to a receptor



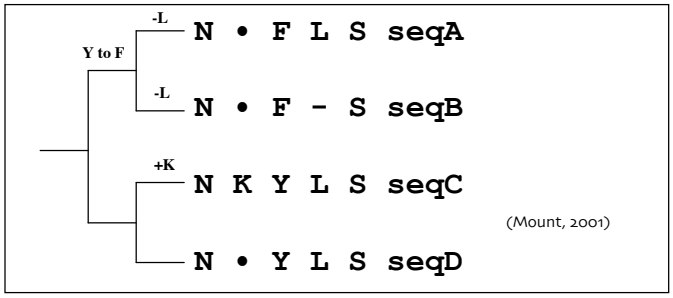
Profile Analysis

- 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



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)



Pointers

- When to use each method?
  - Related sequences = global alignments
  - Divergent sequences = local alignments
  - Use together to build the 'biologically relevant' alignment
- Applications
  - MSAs: ClustalX, Jalview, Belvu
  - Annotation:

File Formats

- MSF
  - http://jura.wi.mit.edu/bio/education/bioinfo/session3/seq.msf
- ALN
  - http://jura.wi.mit.edu/bio/education/bioinfo/session3/seq.aln
- PIR
  - http://jura.wi.mit.edu/bio/education/bioinfo/session3/seq.pir
- DND
  - http://jura.wi.mit.edu/bio/education/bioinfo/session3/seq.dnd
- PH
  - http://jura.wi.mit.edu/bio/education/bioinfo/session3/seq.ph

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 Ouellette. Wiley Interscience, 2001.
- Bioinformatics: Sequence, structure, and databanks. Des Higgins and Willie Taylor. Oxford University Press, 2000.