What we will cover today

• Multiple Sequence Alignment (MAS)
• Motifs
• Phylogenetic Trees
What is Multiple Sequence Alignment (MSA)?

- An extension of a pairwise alignment
- Can be local or global
- The “inputs” are the same
  - A set of amino acid or nucleic sequences
  - Substitution (scoring) matrices
  - Gap penalties
- The objectives are similar: find an alignment of more than two sequences
- Discussed in earlier lecture

Multiple Sequence Alignment

<table>
<thead>
<tr>
<th></th>
<th>Wheat</th>
<th>Rice</th>
<th>Soy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSADKPSAYMLWLSNARESIKRENPDSGIL</td>
<td>MKADKPSAYML- - - NARESI- - ENPDSGRL</td>
<td>MPADKPSMFML- - - NPSESI- - NPDSARL</td>
</tr>
</tbody>
</table>
Why Do Multiple Sequence Alignment?

- Characterize protein families by identifying shared regions of homology
- Determine the consensus sequence of several aligned sequences
- Establish relationships and phylogenies
  - Clustering analysis
  - Structural modeling
  - Evolutionary analysis
- Use in a database search of protein families

Multiple Alignment programs
align several protein sequences

- ClustalW
  - http://www.ch.embnet.org/software/ClustalW.html
  - Multiple sequence alignment program
- T-Coffee
  - Alignment program that often gives better results, especially when dealing with divergent sequences and long insertions
Multiple sequence alignment

• ClustalW
  – Does alignment and phylogenic tree
  – www.ebi.ac.uk/clustalw/index.html
• Dialign
  – Bibiserv.techfak.uni-bielefeld.de/dialign/
• Tcoffee
  – Igs-seerver.cnrs-mrs.fr/Tcoffee

MSA Algorithms

• As with the pairwise sequence comparisons, there are two types of multiple alignment algorithms
  – Optimal
  – Heuristic
Optimal MSA

- Extension of dynamic programming to multiple dimensions
- Exhaustive search
- Guaranteed to find an optimal score
- Need an n-dimensional matrix for scoring
- Computationally expensive
- Time complexity for pairwise comparisons is $O(m_1 \times m_2)$; for multiple alignment should be $O(m^n)$
- Not feasible for $n>10$ sequences of length $m>200$ residues

Heuristic MSA

- Limit the exhaustive search
- Attempt to rapidly find a good, but not necessarily optimal alignment
- Most popular methods:
  - Tree alignments
  - Star alignments
Heuristic approaches to MSA

- Progressive global alignment starting from the most similar sequences:
  - CLUSTALW
    - Pairwise alignment: calculate distance matrix
    - Neighbor joining: draw guide tree
    - Progressive alignment: align following guide tree
- Iterative methods: make initial crude alignment, then revise it: DIALIGN
- Alignment based on locally conserved patterns found in sequences in the same order: BLOCKS, eMOTIF, GIBBS, MEME
- Use statistical methods and probabilistic models of the sequences: HMMER, SAM

What is a motif?

- A short conserved region in DNA, RNA or protein sequence
- Corresponds to a structural or functional feature in proteins
- Shared by several sequences and can be generated by MSA
- Can be represented using position-specific scoring matrices
What is a profile?

• A position-specific scoring matrix, or matrix of scores representing a motif
• 22 columns, one for each of the 20 amino acids, and 2 for the penalties of opening and extending gaps
• The rows of the profile: aligned amino acid residues of a group of sequences
• Residues with the highest scores define a consensus

What is a protein family?

• A set of evolutionarily related proteins
• Members of a protein family may range from very similar to quite diverse
• Often share domains. Domain is a part of a protein (greater than a motif) that can fold and carry out a function independently.
Motif- and domain-oriented databases

- Secondary databases, small compared to GenBank
- Contain representations of conserved sequences shared by a sequence family
- Are primarily used for annotation of unknown sequences
- Examples: Pfam, Blocks, PRINTS, Prodom, PROSITE

Motifs and conserved domains
Pfam

- **Protein FAMily (Pfam)** is a large collection of multiple sequence alignments of sequence motifs or domains
- Made up of two parts: Pfam-A and Pfam-B
- Pfam-A: curated database of gapped profiles
- Pfam-B: generated automatically from sequences taken from the Prodom database that do not overlap with Pfam-A
- Use a Hidden Markov Model (HMM) to define domains or to align a set of sequences

Blocks

- Multiple sequence alignments without gaps that were used to construct the BLOSUM substitution matrices
- Generated automatically
- Correspond to the most conserved regions of a proteins
- Better used to identify protein sequence domains or families rather than identify motifs
**PROSITE**

- A database of sequence patterns (~motifs) associated with protein family membership
- Developed by largely manual process of seeking the patterns that best fit particular protein families
- Patterns may be useful in assigning distant homologs to sequence families
- PROSITE patterns are very short => may result in false positive occurrences in unrelated sequences

**PRINTS**

- A compendium of protein fingerprints
- A **fingerprint** is a group of conserved motifs used to characterize a protein family
- The motifs do not overlap (separated along a sequence)
**Prodom**

- An automatically generated collection of protein domains
- Better described as a software tool to visualizes a protein’s sequence domain structure

**Profile searches**

- Numerical representations of multiple sequence alignments
- Depends upon patterns or motifs containing conserved residues
- Represent the common characteristics of a protein family
- Can find similarities among sequences with little or no sequence identity
- Allows for the analysis of distantly-related proteins
ProfileScan

- Search sequence against a collection of profiles and patterns
- Databases available
  - PROSITE profiles
  - PROSITE patterns
  - PfamA
  - PfamB
- http://hits.isb-sib.ch/cgi-bin/PFSCAN

Profile Construction

- Which residues are seen at each position?
- What is the frequency of observed residues?
- Which positions are conserved?
- Where can gaps be introduced?

Position-Specific Scoring Table
Patterns

\[ [FY] - x - C - x(2) - [VA] - x - H(3) \]

reads as:

- Phe or Tyr followed by
- any amino acid followed by
- Cys followed by
- any two amino acids followed by
- Val or Ala followed by
- any amino acid followed by
- three His

ProfileScan

- http://hits.isb-sib.ch/cgi-bin/PFSCAN
- Find all known motifs in a sequence
Protein sequence

LAQNPRSTLLPKARGFSRL
QLPEMASVSLAKYKLVFLG
DQSVGKTSIIITRFMYDKFDN
TYQATIGIDFLSKTMYLEDRTVRQLWDTAGQERFRSLIPSYIRDSSVAIVYDVASRQTFLNTAKWEESVRTERGSDVI
IVLVGNKTDLVEKRQVSIEEGEAKARELNVMFIETSAKAGFNKALFRKIAALPGMETLSSAKQEDMVNDVLKSTNGSA
QSQPQSSGCAC*

Motif or conserved domain searches

• A pattern retained through evolution not randomly changed by mutation
• Retained to help perform a specific function
• Domains can be found in databases
  SMART
  http://smart.embl-heidelberg.de/
Pfam
  http://www.sanger.ac.uk/Software/Pfam/
COGs Clusters of Orthologous Groups
Conserved Domain in Beta Globin

cdd1040.1

globin

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>CDD</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>Human</td>
</tr>
<tr>
<td>Bank</td>
<td>Gene Bank</td>
</tr>
<tr>
<td>Primary accession</td>
<td>cdd1040.1</td>
</tr>
</tbody>
</table>

**Feature 1:** Home-binding site

**Description:** Structural alignment of beta globin with bound heme and oxygen release. View structure with CDBox 4.1.

**Interaction:**
- Name: [Enzyme Name]
- Source: [Organism Name]
- Template: [Template Name]
- Sequence: [Sequence Information]
- Structure: [Structure Information]
- Alignment: [Alignment Information]

**Additional Information:**
- Protein ID: [Protein ID]
- Accession: [Accession]
- Length: [Length]
- Score: [Score]
- E-value: [E-value]
Motif scanning means finding all known motifs that occur in a sequence. This form lets you paste a protein sequence, select the collections of motifs to scan for, and launch the search. Some general documentation is available about the Prosite and Pfam collections of motifs. Another documentation deals with the interpretation of the match scores. You should consult the home pages of Prosite on EMBASE, Pfam and InterPro for additional information.

A pre-computed list of matches is also available on our server (File). If your proteins of interest are already in the databases, the Query by Protein form is much faster, and the Protein Hub provides you a collection of tools that you might find useful.
Result

- Summary:

  fam.ATP_BND
  post: 156
  E-value: 0.73

  fam.REF
  post: 21
  E-value: 3.3e-05

  fam.CITP_BND
  post: 56
  E-value: 4.1

  fam.RAS
  post: 35
  E-value: 3.3e-03

- Match Location:

  [Image of match location with sequence alignment]

- Prosite patterns (weekly-updated):
  no match

- Prosite profiles (weekly-updated):
  [Image of prosite profiles]
Constructing Phylogenetic Trees

Why use phylogenetics

• Determine closest relative of an organism
• Discover the function of a gene
  – Identify orthologous, well-characterized gene in another species
• Retrace the origin of a gene
  – Mutations, deletions, gene duplications, gain- or loss-of-function, inactivation
Important

• Data quality
  – Highly accurate multiple sequence alignment that contains properly chosen sequences

Types of genes

• Orthologs
  – Separated only by speciation
    • A common ancestor gives birth to two subgroups that slowly drift away to become distinct species

• Paralogs
  – A gene is duplicated. The resultant two genes slowly diverge in sequence

• Xenologs
  – Result from a lateral transfer of a gene from one organism into the genome of another organism
How do you know two genes coming from two different species are orthologs or paralogs?

No simple solution

Strategy

• 1) Choose a sequence from genome A
• 2) GLAST search sequence A against every sequence in complete genome of B
• 3) BLAST search returns sequence B as a top hit
• 4) BLAST search B against every sequence in genome A
• Is sequence B the top hit???
• This does not prove – but provides support for
Tips

• Avoid sequence fragments
• Avoid xenologs
• Avoid recombinant sequences (especially seen in viruses)
• Avoid very large complex families containing repeats
• Keep your data set small
• Add an out group to root your tree
  – ie. a sequence that you know is a member, but has diverged long ago from the rest of the set

How to improve your multiple sequence alignment for phylogenetics

• Remove gaps
  – Gaps cause problems

Wheat  MSADKPSAYMLWLSNARESIKRENPSGIL
Rice    MKADKPSAYML- - - NARESI- - ENPSGRL
Soy     MPADKPSMFML- - - NPSSESI - - NPDSARL
How to improve your multiple sequence alignment for phylogenetics

- Remove extremities of your multiple sequence alignment
- N-terminus and C-terminus tend to be poorly conserved
- Remove gap-rich regions
- Keep most informative blocks
  - Typically 20 to 30 amino acids long
  - Contains a few conserved positions

Tree software

- ClustalW
  - easy
- Phylip
  - sophisticated
**Example: tree alignment of four sequences**

A
B
C
D

• Compare all six pairs of sequences
• Define and compute distances between the sequences
• Then use cluster analysis
• The number of pairs of N segments = N(N-1)/2=4(3)/2=6

**Clustal W**

• Format is important
• Can past or upload sequences
• Each sequence must have a unique name
• No empty lines
• No white spaces
• No control characters
• Limited to 500 sequences or 10MD, which ever is smaller)
>Arabidopsis
MVMAGASSLDEIRQAQRADGPAGILAGTANPENHVLQAEYHDYYFRTINSEHKT
DLKEFKRMCDKSTIKRHKHMLTEEFLKENPHMCAYMAPSLDRQDVMVVEVPKLI
GEKAAAKAIKEWQPGPSKTHIVVFCNTGVDMPGADYOLTLKGLHRVPSYKRMML
YQQCGFAATGTVLRLAKDLAENNRGARVLVVCSEIATAVTRFQPSDTHLDSLVQAF
FSDGAALIVGSDPLTSVGEKFPEMVSAAQTLPDSDGAAIDHLLREVGLTFHILKD
VPGLLSKNIVKLWDEAEKFGPSDGLNSFHIWAPGAPVAILQDVEAKLGKEEMQAM
RVHLSEYGNMSSACVLFILDEMRRKSAKDOVATGEEGLEGWVGFPGPLTVLHSEY

>SoyCH5
MVSVEERQARERPGATVMAIGTATPNCVDQSTYPDYYFRITINSEHMTELKEK
FRMCDKSMIKKRYMYLNEEILKENPSGCAYMAPSLDRQDVMVVEVPKLGKEEA
ATKAIKEWQPGPSKTHILFCTSGVDMPGADYQLKTLGLRPSKRYMMYQGQC
FAGGTVLRLAKDLAENNRGARVLVVCSEIATAVTRFQPSDTHLDSLVQALFDGA
AAVWDSPEPVEKLFVQVTAQILTPESEGAIDHLLREVGLTFHILKDVPGLISIK
NIEKALVEAFFQPLGDINSELFWIAPGAPVAILQDVEAKLGKEEMQAM

>SoyCH6
MVSVEERQARERPGATVMAIGTATPNCVDQSTYPDYYFRITINSDNHMELEKEF
FRKRMCDKSMIKKRYMYLNEEILKENPSGCAYMAPSLDRQDVMVVEVPKLGKEAA
TKAIKEWQPGPSKTHILFCTSGVDMPGADYQLKTLGLRPSKRYMMYQGQC
AGGTVLRLAKDLAENNRGARVLVVCSEIATAVTRFQPSDTHLDSLVQALFDGA
AAVWDSPEPVEKLFVQVTAQILTPESEGAIDHLLREVGLTFHILKDVPGLISIK
NIEKALVEAFFQPLGDINSELFWIAPGAPVAILQDVEAKLGKEEMQAM
• Phylogram
  – Branch length represent real distances
• Cladogram
  – Branches indicate only branching order
Parts of a Phylogenetic Tree

- **Root**
- **Node**
- **Branch length**
- **Leaves**
- **Branches**

Tree building

- Trees are also called **dendrograms**
- Nodes represent different organisms and links are used to show lines of descent
- Two basic types of questions about a tree:
  a) its **topology**: how its interior nodes connect to one another and to the leaves
  b) **distance between pairs of nodes**, which is an estimate of an evolutionary distance
- Tree may or may not have a root. A tree with root implies ancestral relationship between interior nodes
Phylogenetic tree evaluation

- How reliable phylogenetic tree is?
- One criterion: if different methods of tree construction give the same result, this is good evidence that the tree is reliable
- Another criterion (bootstrapping): data are randomly sampled from any position within MSA, and are built into new artificial alignments, which are then tested by tree building
- Third criterion (jackknife): similar to bootstrapping, but instead of generating new datasets with replacement, it re-samples the original data set by dropping one or more alignment positions in each replicate

What we learned

- Multiple Sequence Alignment (MAS)
- Motifs
- Phylogenetic Trees