MICB 405 Bioinformatics

Lecture 5.1

Multiple Sequence Alignments

FSC 1221

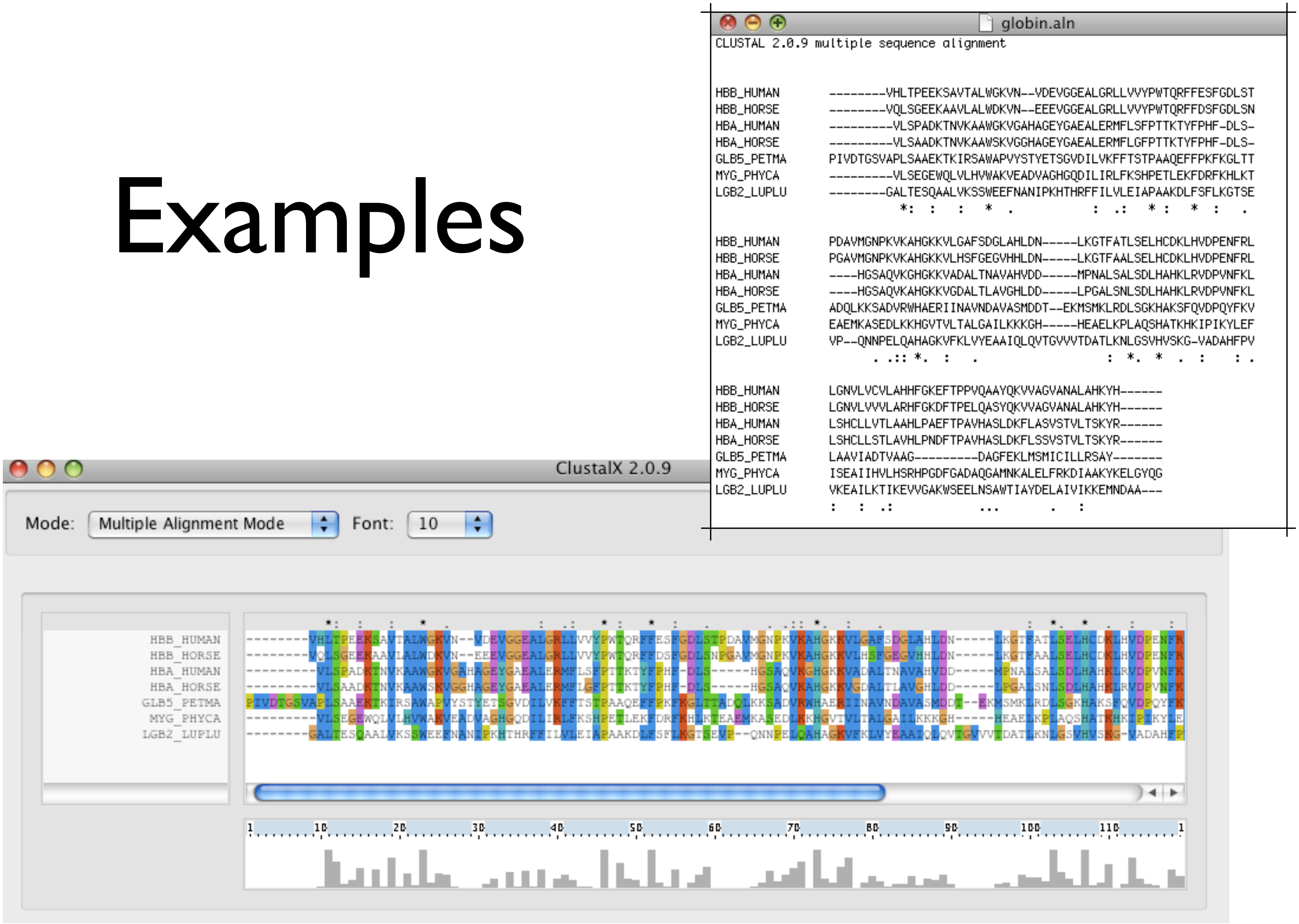
September 30th, 2008

Objectives

By the end of today's lecture:

- You will be able to compare and contrast pairwise vs. multiple sequence alignments.
- You will be able to describe the method of progressive multiple sequence alignments.
- You will be able to explain how the CLUSTAL algorithm works.
- You will list examples of uses and applications of multiple sequence alignments.

Examples



Mode: Multiple Alignment Mode

Font: 10

```
CLUSTAL 2.0.9 multiple sequence alignment


HBB_HUMAN       --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
HBB_HORSE       --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
HBA_HUMAN       ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
HBA_HORSE       ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
GLB5_PETMA      PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
MYG_PHYCA       ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
LGB2_LUPLU      --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                          *:  :   :   *  .           :  .:   * :   *  :   .

HBB_HUMAN       PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
HBB_HORSE       PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
HBA_HUMAN       ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
HBA_HORSE       ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
GLB5_PETMA      ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
MYG_PHYCA       EAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATKHKIPIKYLEF
LGB2_LUPLU      VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-VADAHFPV
                      . .:: *.  :   .                  :  *.  *  .  :    : .

HBB_HUMAN       LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
HBB_HORSE       LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
HBA_HUMAN       LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
HBA_HORSE       LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
GLB5_PETMA      LAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
MYG_PHYCA       ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
LGB2_LUPLU      VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
                :   :  .:            ...       .   :
```

Multiple Sequence Alignment

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSSSGFIFSS--YAMYWVRQAPG
LSLTCTVSGTSFDD--YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCLISDFYPGA--VTVAWKADS--
AALGCLVKDYFPEP--VTVSWNSG---
VSLTCLVKGFYPSD--IAVEWESNG--
```

The sole purpose of multiple sequence alignments is to place *homologous positions* of *homologous sequences* into the *same column.*

Pairwise vs. MSA

Pairwise

- Can use dynamic programming method
    - very fast to find optimal alignment
- Given scoring matrix and gap penalties
    - exact solution to optimal alignment is possible to compute

MSA

- Optimize alignment of every sequence with every other sequence
    - Slow
- Use heuristics
    - example - progressive alignment heuristic
- Approximate solution
    - biologically significant

Clustal

- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994)
  - CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice.
    - Nucleic Acids Research, 22:4673-4680.

What is a Progressive Alignment?

Build up multiple sequence alignment by iteratively adding new sequences to an existing alignment.

✓ Example – simple progressive alignment with random sequence selection

- Starting with N sequences to align:

  1. Create initial pairwise seed alignment

  2. Randomly select next sequence to align

  3. Align sequence to existing alignment

  4. Return to step 2 until all N sequences are aligned

Limitations

- Crucial that early sequence alignments are correct
  - Every new sequence aligned can introduce errors; worsens with sequence divergence
- Order that sequences are aligned may alter final alignment
  - Random selection may not be best
- Alignments can often be improved by hand

CLUSTAL - an improved progressive method

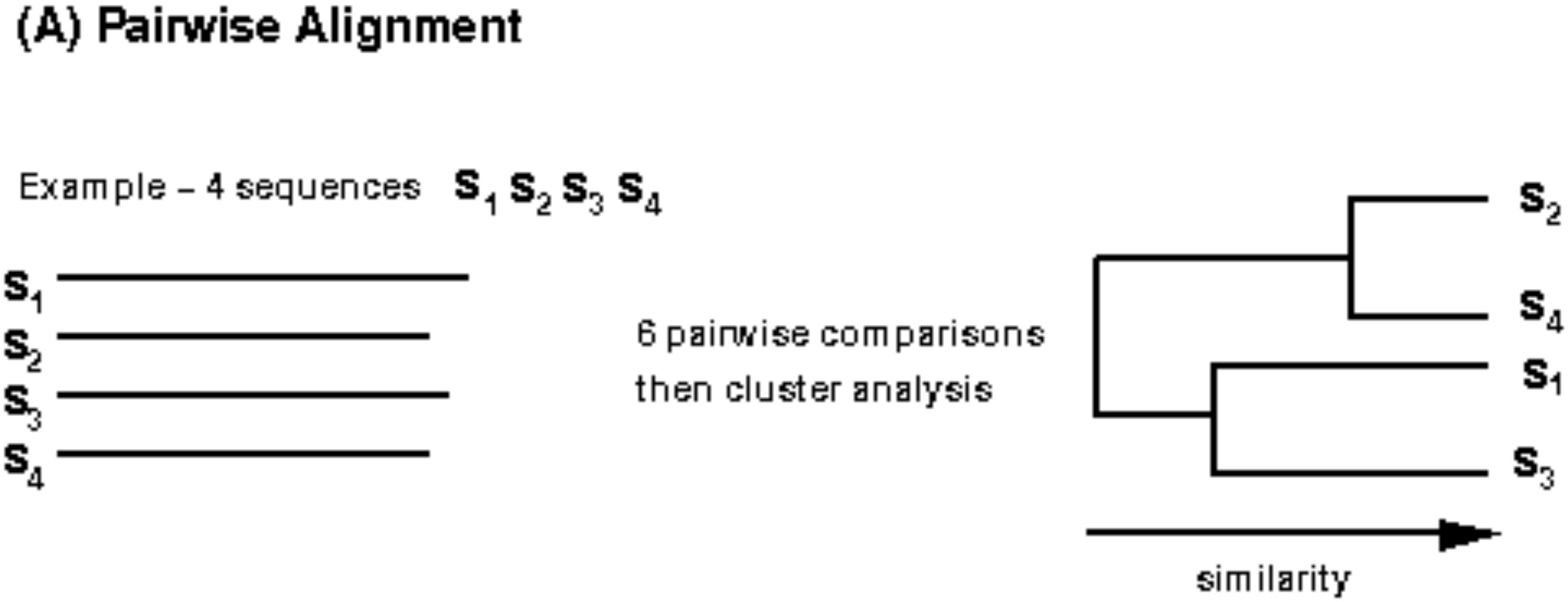
✓Incorporate biological information

- Assume sequences are homologous
- Align most similar sequences first
  - avoid introduction of unnecessary gaps
  - use existing alignment information
- Position Specific Gap Penalties

CLUSTAL Algorithm Steps

1. Pairwise alignment of each sequence pair
   - Number of comparisons depends on how many sequences
2. Compute distance matrix
   - Percent non-identity between each alignment pair
   - Lower distance means more similar
3. Construct a sequence similarity tree
   - Cluster sequences according to distance (similarity)
4. Progressive alignment of sequences according to a tree

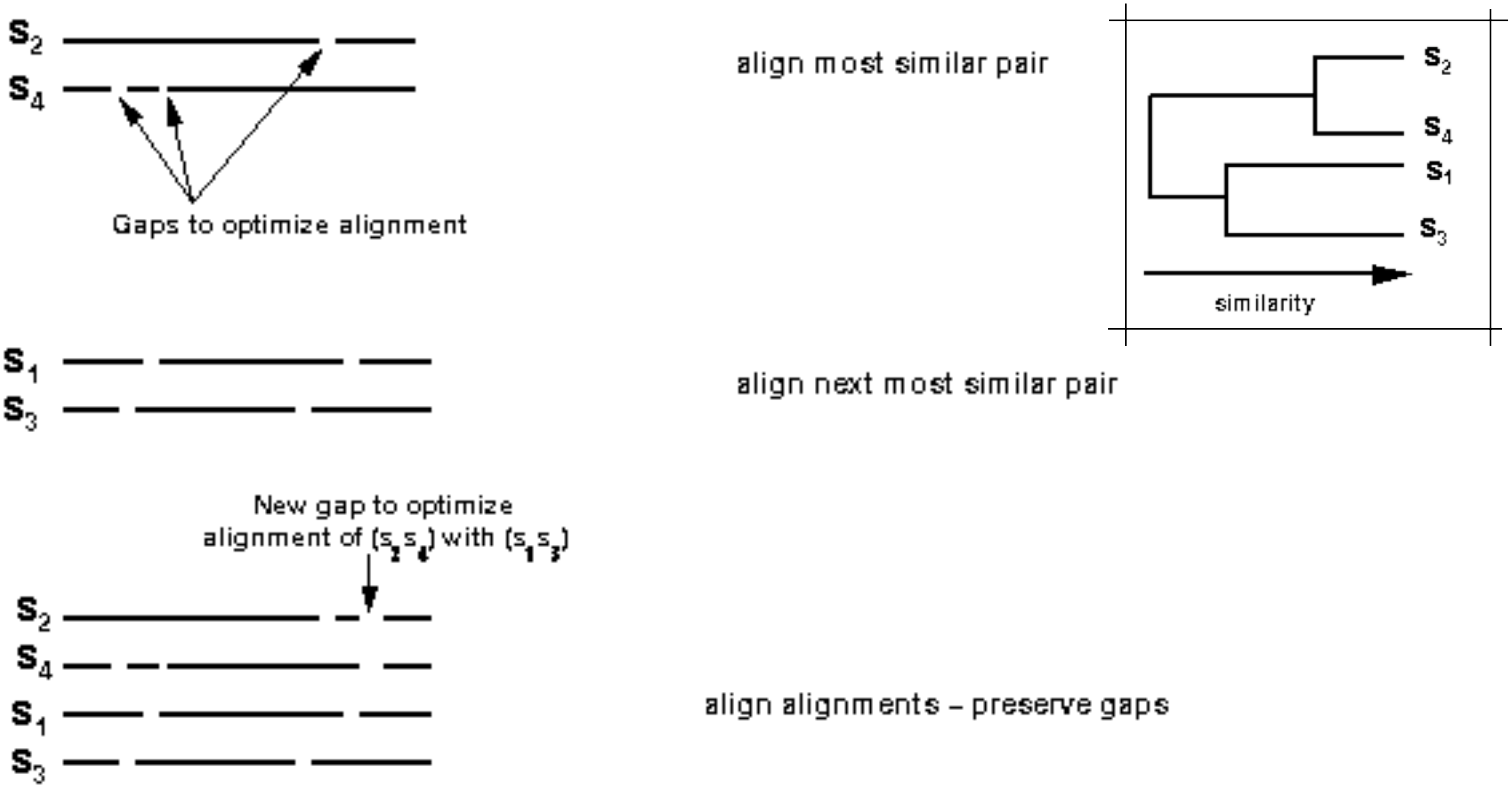
How does the Clustal algorithm actually work?



Which sequences would be aligned first?

Steps in a Multiple Sequence Alignment continued …

(B) Multiple alignment following the tree from A


s2
s4
Gaps to optimize alignment
align most similar pair
s2
s4
s1
s3
similarity
s1
s3
align next most similar pair
New gap to optimize alignment of (s2 s4) with (s1 s3)
s2
s4
s1
s3
align alignments – preserve gaps

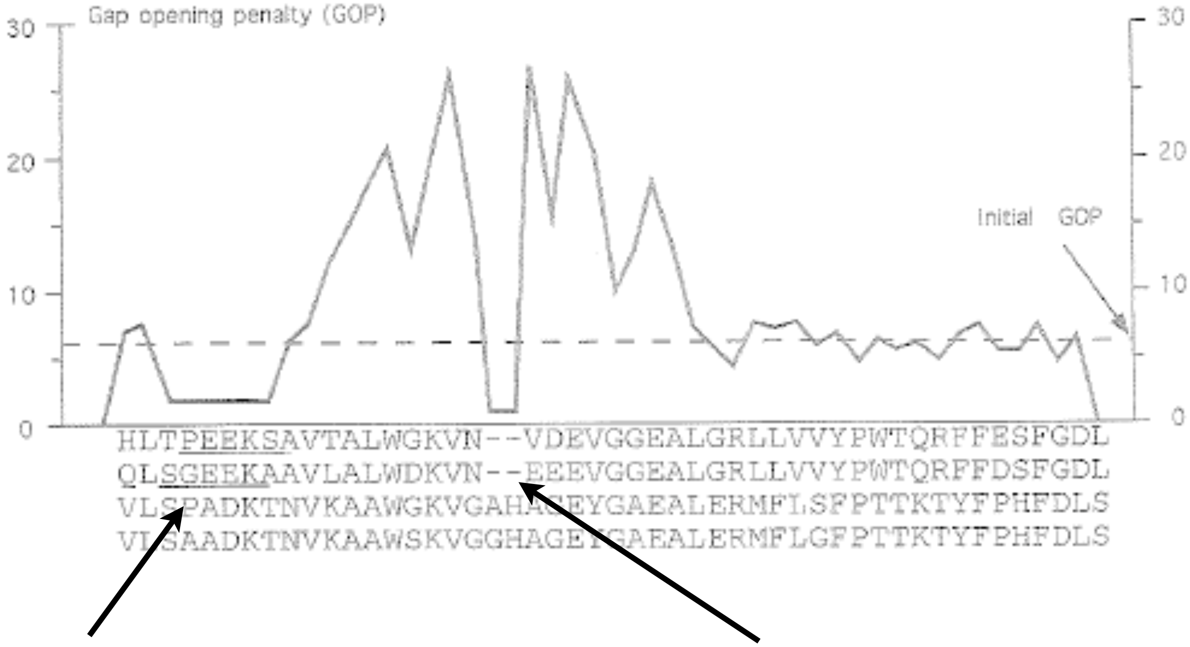
Gap Penalities

- Gaps are introduced as alignment progresses
  - Gaps in new sequences to be aligned
  - Gaps in existing multiple alignments
- Insertions and deletions (indels) are rare events
  - Want indels to be aligned
  - Indels usually occur in loop structures of proteins
- There are two type of gap opening penalities: gap opening and gap extension
  - Determined empirically by user

Position Specific Gap Penalities

- Decrease penalties where gaps already occurs
- Increase penalties in adjacent positions to where gap already occurs
  - Encourage extension of gaps in loop regions vs. introduction of new gaps
- Increase or decrease gap penalties according to amino acid type
  - Increase penalties in stretches of hydrophobic residues
  - Discourage the disruption of secondary structure elements

Gap Penalties Example



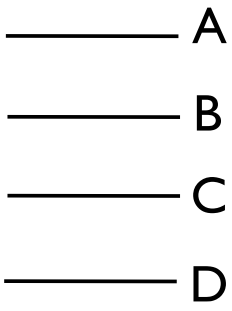
hydrophobic stretch

existing gap

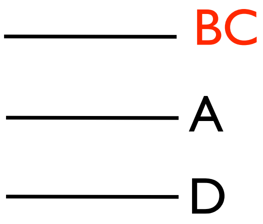
Figure from Higgens et al, Methods in Enzymology 266: 383

The order of your input sequences affects your resulting multiple sequence alignment

Let’s try and illustrate this with an example……



	A	B	C	D
A	-			
B	0.6	-		
C	0.2	0.1	-	
D	0.7	0.1	0.8	-



	A	BC	D
A	-		
BC	0.4	-	
D	0.7	0.45	-

- BC and BD are both equally similar
- However the BC and BD consensus sequences can be quite different:

```
B= ELVIS   BC= ELVIS   BD= ELVIS
C= LIVES       LIVES       EVILS
D= EVILS       --V-S       E---S
```

Sequence Order

The order of your input sequences could affect your resulting multiple sequence alignment

✓ What should you do?

- Try aligning your sequences with different input orders to see if there is any significant difference in the alignments.
- Always examine your alignment

Applications of MSA

Differences between CLUSTAL and BLAST?

CLUSTAL	BLAST
• global alignment method	• local alignment method
- Align complete sequence	- Search for HSP
• Assumes homology	• Test for homology
• Complex gap penalties	• Simple gap penalties
• Slower	• Fast
• Align protein-protein or nucleotide-nucleotide only	• Translated searches

Standard Multiple Sequence Alignment Approach

- Be as sure as possible that the sequences included are homologous
- Know as much as possible about the gene/ protein in question before trying to create an alignment (secondary structure, domains etc..)
- Start with an automated alignment: preferably one that utilizes some evolutionary theory such as CLUSTAL

Standard Multiple Sequence Alignment Approach

Examine alignment:

- Are you confident that aligned residues/bases evolved from a common ancestor?
- Are domains of the proteins/predicted secondary structures, etc. aligning correctly?
- Are most indels outside of known motifs or secondary structure?

→ No? May need to edit sequences and redo…

The Take Home Message

Why perform an MSA?

- Visualize trends between homologous sequences
    - Shared regions of homology
    - Regions unique to a sequence within a family
    - Consensus sequence
- As the first step in a phylogenetic analysis

The Take Home Message

How does one perform an MSA?

- By hand: too hard!
- Automated alignment: Fast, but doesn't necessarily produce the "correct" alignment

Best approach = Automated alignment with manual editing

Summary

- Uses of Multiple Sequence Alignments (MSA)
- Pairwise vs. MSA
- CLUSTAL = progressive alignment method
- CLUSTAL method involves use of:
  - distance matrix
  - guide tree
  - optimized gap penalties

Links

- CLUSTALW @ EBI
  - http://www.ebi.ac.uk/clustalw/
- Download CLUSTALX
  - ftp://ftp.ebi.ac.uk/pub/software/clustalw2/

For more information:

- *Baxevanis & Ouellette (3rd Edition)*
  - Chapter 12: p326 – p331
- *Westhead, Parish & Twyman*
  - Sections F1