Multiple Sequence Alignment (MSA)

Multiple Sequence Alignment

Most important contribution of MB to **evolutionary analysis** is the discovery that DNA sequences of different organisms are often **related.**

i.e., genes are **conserved** across widely divergent species, often performing a **similar** or even identical function, and at other times, mutating or rearranging to perform an **altered** function.

Through **simultaneous alignment** of gene sequences, sequence patterns that have been subject to alteration may be analyzed.

Multiple Sequence Alignment

Aligning more than two sequences

A C - - B C D B

- C A D B - D -

A C A - B C D -

In an MSA, homologous residues among a set of sequences are aligned together in columns.

'Homologous' is meant in both the structural and evolutionary sense.

Motivation for MSA

- MSA helps identify conserved regions and those allowed to vary – regions resistant to change are functionally most important to the molecule.
- Carries more information than mere pair-wise alignment
- Multiple sequence similarity suggests common structure for the protein, a common function or evolutionary origin
- MSA requirements are different in the various applications

Motivation for MSA

Multiple alignments can improve pairwise alignments:

```
(A) p110α          TFILGIGDRHNSNIMVKDDG-QLFHIDFGHFLDHKKKKFGYKRERVPFVLT--QDFLIVI 142
    cAMP-kinase    QIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAPE 179

(B) p110β          SYVLGIG----------DRHSDNINVKKTGQLFHIDFGHILGNFKSKFGIKRERVPFILT 136
    p110δ          TYVLGIG----------DRHSDNIMIRESGQLFHIDFGHFLGNFKTKFGINRERVPFILT 136
    p110α          TFILGIG----------DRHNSNIMVKDDGQLFHIDFGHFLDHKKKKFGYKRERVPFVLT 135
    p110γ          TFVLGIG----------DRHNDNIMITETGNLFHIDFGHILGNYKSFLGINKERVPFVLT 135
    p110_dicti     TYVLGIG----------DRHNDNLMVTKGGRLFHIDFGHFLGNYKKKFGFKRERAPFVFT 135
    cAMP-kinase    QIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCG--TPEYLA 177
```

Catalytic domains of 5 P13-kinases and cAMP-dependent protein kinase

MSA for DNA Sequences

In DNA sequences MSA is used in

- Genome sequence assembly - shotgun sequencing
- Discovering new regulatory elements
- Inferring evolutionary relationships
- DNA barcoding
- SNP identification
- Develop primers & probes - use conserved regions to develop
  - Primers for PCR
  - Probes for DNA microarrays

In which of these applications do we look for similarity/differences?

MSA for Protein Sequences

In protein sequences, MSA is used in

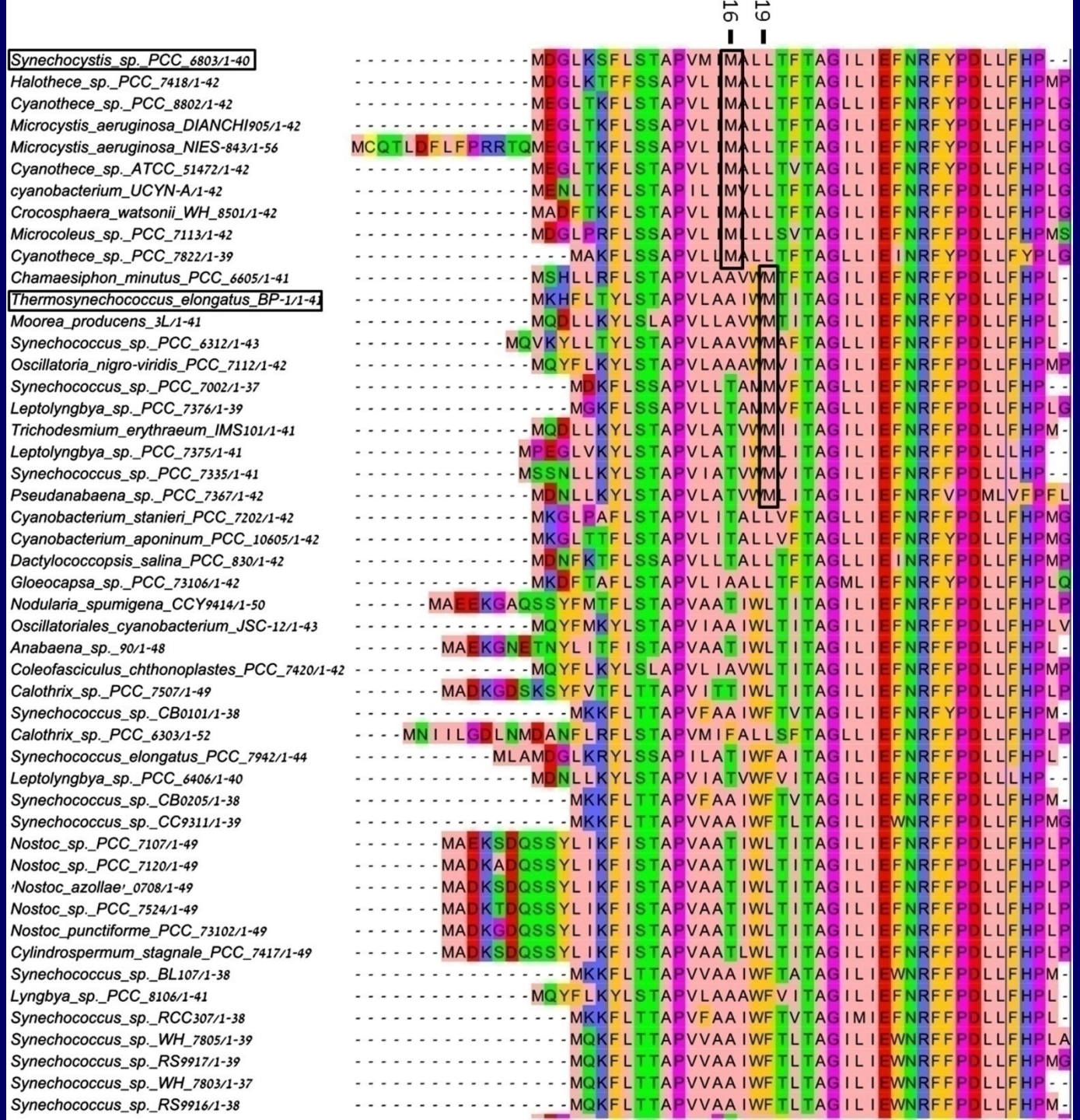
- Homology modeling of proteins
- Building phylogenetic tree
- Constructing scoring matrices – PAM, BLOSUM
- Predicting secondary & tertiary structures of new sequences
- Identifying conserved patterns, motifs, blocks in protein sequences - to characterize protein families
- Identify related proteins in database searches, e.g., Profiles, PSI-BLAST, HMMs

MSA

Visual alignment of MSA tables use different colours for displaying AAs of different physico-chemical type - aids in identifying conserved patterns:

Colour	Residue Type	Amino acids
Yellow	Small nonpolar	Gly, Ala, Ser, Thr
Green	Hydrophobic	Cys, Val, Ile, Leu, Pro, Phe, Tyr, Met, Trp
Magenta	Polar	Asn, Gln, His
Red	Negatively charged	Asp, Glu
Blue	Positively charged	Lys, Arg

Structure prediction tools also give more reliable results when based on MSAs than on single sequences.


116
119
Synechocystis_sp._PCC_6803/1-40
Halothece_sp._PCC_7418/1-42
Cyanothece_sp._PCC_8802/1-42
Microcystis_aeruginosa_DIANCHI905/1-42
Microcystis_aeruginosa_NIES-843/1-56
Cyanothece_sp._ATCC_51472/1-42
cyanobacterium_UCYN-A/1-42
Crocosphaera_watsonii_WH_8501/1-42
Microcoleus_sp._PCC_7113/1-42
Cyanothece_sp._PCC_7822/1-39
Chamaesiphon_minutus_PCC_6605/1-41
Thermosynechococcus_elongatus_BP-1/1-41
Moorea_producens_3L/1-41
Synechococcus_sp._PCC_6312/1-43
Oscillatoria_nigro-viridis_PCC_7112/1-42
Synechococcus_sp._PCC_7002/1-37
Leptolyngbya_sp._PCC_7376/1-39
Trichodesmium_erythraeum_IMS101/1-41
Leptolyngbya_sp._PCC_7375/1-41
Synechococcus_sp._PCC_7335/1-41
Pseudanabaena_sp._PCC_7367/1-42
Cyanobacterium_stanieri_PCC_7202/1-42
Cyanobacterium_aponinum_PCC_10605/1-42
Dactylococcopsis_salina_PCC_830/1-42
Gloeocapsa_sp._PCC_73106/1-42
Nodularia_spumigena_CCY9414/1-50
Oscillatoriales_cyanobacterium_JSC-12/1-43
Anabaena_sp._90/1-48
Coleofasciculus_chthonoplastes_PCC_7420/1-42
Calothrix_sp._PCC_7507/1-49
Synechococcus_sp._CB0101/1-38
Calothrix_sp._PCC_6303/1-52
Synechococcus_elongatus_PCC_7942/1-44
Leptolyngbya_sp._PCC_6406/1-40
Synechococcus_sp._CB0205/1-38
Synechococcus_sp._CC9311/1-39
Nostoc_sp._PCC_7107/1-49
Nostoc_sp._PCC_7120/1-49
'Nostoc_azollae'_0708/1-49
Nostoc_sp._PCC_7524/1-49
Nostoc_punctiforme_PCC_73102/1-49
Cylindrospermum_stagnale_PCC_7417/1-49
Synechococcus_sp._BL107/1-38
Lyngbya_sp._PCC_8106/1-41
Synechococcus_sp._RCC307/1-38
Synechococcus_sp._WH_7805/1-39
Synechococcus_sp._RS9917/1-39
Synechococcus_sp._WH_7803/1-37
Synechococcus_sp._RS9916/1-38

MSA

To be informative a MSA should

- contain a distribution of closely- and distantly-related sequences.

If all closely-related - information contained is largely redundant

⇒ few inferences can be drawn.

If all very distantly-related - difficult to construct an accurate alignment

⇒ quality of results & inferences might be questionable

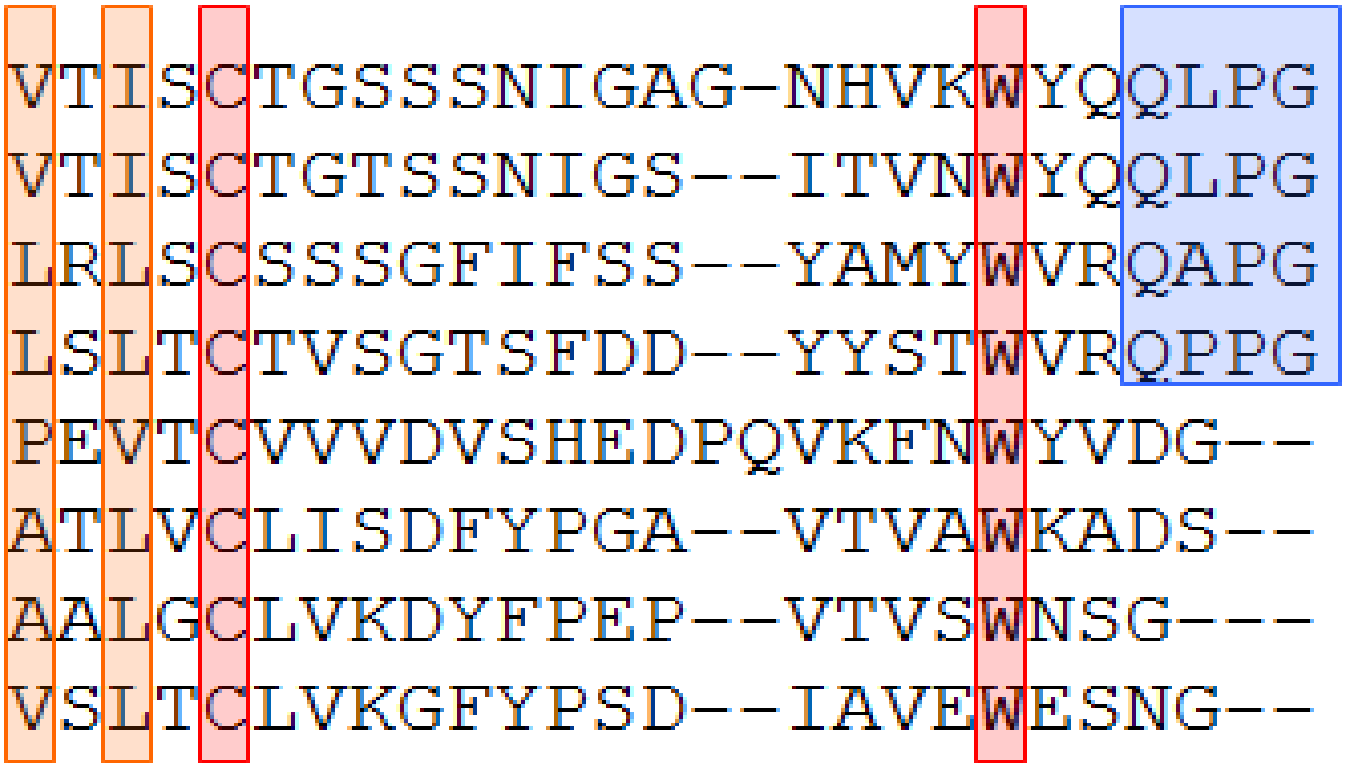
Ideally, one should have a complete range of similarities, including distantly-related examples linked through chains of close relationships

Inferences from MSA

Some examples:

- Highly conserved regions - likely to be essential sites for structure/function, e.g. active site
- Regions rich in insertions/deletions – may correspond to loops/turns in proteins
- Build gene/protein families – use conserved regions to guide search
- Basis for phylogenetic analysis – infer evolutionary relationships between genes

MSA of 8 fragments of immunoglobin sequences



Conserved residues, regions, patterns

Multiple Alignment: Evaluation

VTISCTGSSSNIG-AGNHVKWYQQLPG	VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIG--SITVNWYQQLPG	VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSSSGFIFS--SYAMYWVRQAPG	LRLSCS-SSGFIFSS-YAMYWVRQAPG
LSLTCTVSGTSFD--DYYSTWVRQPPG	LSLTCT-VSGTSFDD-YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNW--YVDG	PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCLISDFYPG--AVTVAW--KADS	ATLVCLISDFYPGA--VTVAWKADS--
AALGCLVKDYFPE--PVTVSW--NS-G	AALGCLVKDYFPEP--VTVSWNSG---
VSLTCLVKGFYPS--DIAVEW--ESNG	VSLTCLVKGFYPSD--IAVEWESNG--

It is not enough to just get a multiple alignment; we need to score the alignment

Multiple Alignment: Evaluation

A simple way to evaluate a multiple alignment is to evaluate the cost column by column

Sum of Pairs (SP)

$$= \sum_{i<j} D(S_i, S_j)$$

Using unit cost:
mismatch costs 1,
match 0, and
indel costs 1

$$ColumnCost\begin{pmatrix} \mathrm{L} \\ \mathrm{L} \\ \mathrm{A} \\ \mathrm{P} \\ \mathrm{G} \\ \mathrm{S} \\ \_ \\ \mathrm{G} \end{pmatrix} = ?$$

Summing the scores of all possible combinations of *AA* pairs in a column of *MSA*

Multiple Alignment: Evaluation

$$ColumnCost\begin{pmatrix} \mathtt{L} \\ \mathtt{L} \\ \mathtt{A} \\ \mathtt{P} \\ \mathtt{G} \\ \mathtt{S} \\ \mathtt{-} \\ \mathtt{G} \end{pmatrix} = 26.$$

Assumes a model for evolutionary change in which any of the sequence could be the ancestor of others

SP Scoring Method

- There are problems with SP scoring system as illustrated in the example:

Sequence	Col. A	Col. B	Col. C
1	…N…	…N…	…N…
2	…N…	…N…	…N…
3	…N…	…N…	…N…
4	…N…	…N…	…C…
5	…N…	…C…	…C…
Score	60	24	9

(Using Blosum62):
N-N: 6, N-C: - 3, C-C: 9

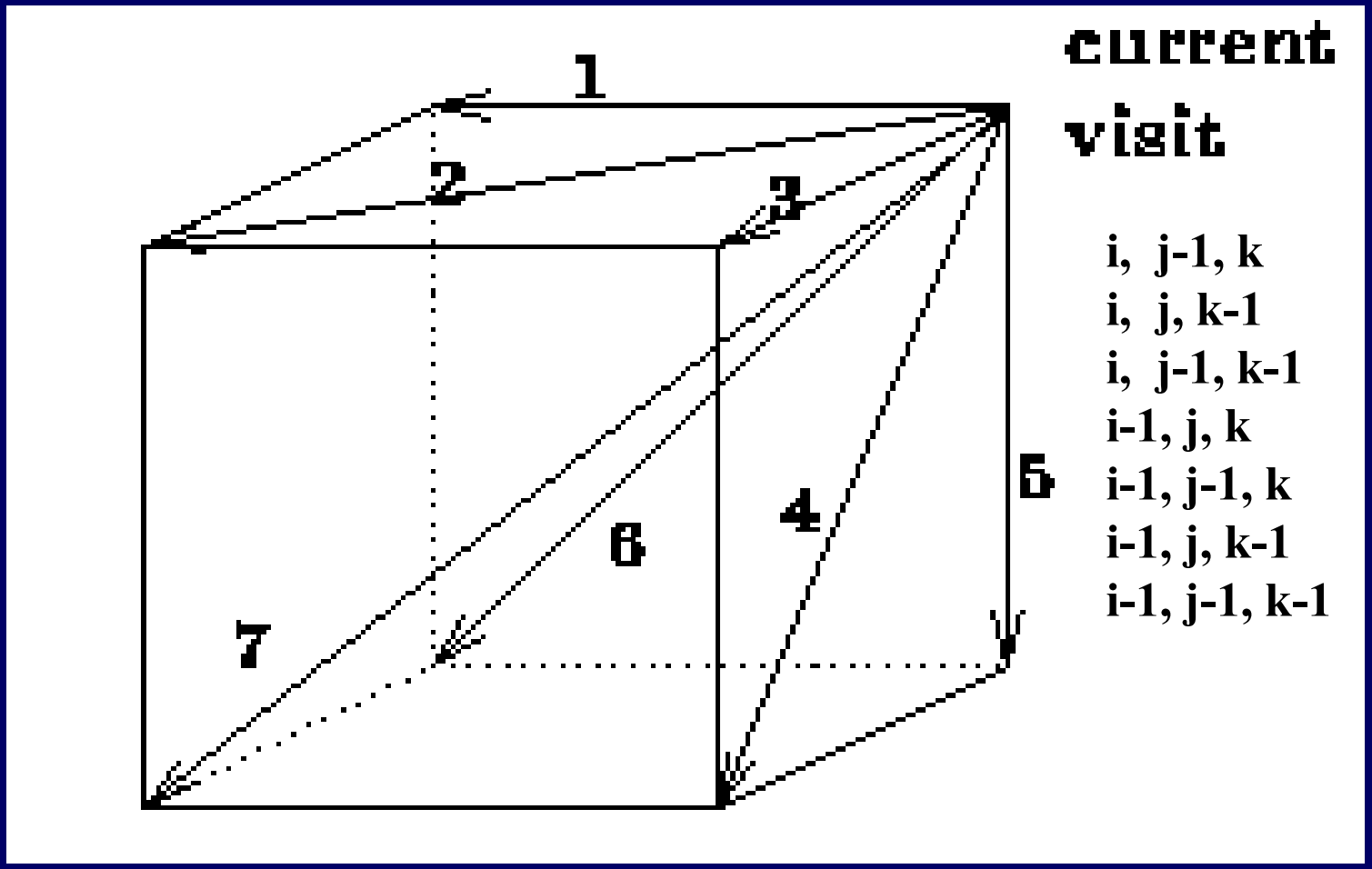
What's the problem?
Score for N = 10 seq

Multiple Alignment: DP

- Pair-wise alignment: DP involves a L x L (L^2) matrix
- Multiple alignment: DP involves an L^N matrix, an N-dimensional hyper-lattice, *N* - no. of sequences of length *L* each, to align simultaneously
- Computationally not feasible: for 5 sequences, each ~100 bp long, 10^{10} matrix elements need to be computed

  - equivalent to a pairwise alignment of two 100,000 bp sequences

The Recursive Relation



For 3 sequences, to assign a value to a node (i,j,k), we need to consider 7 values; for 2 seqs, we needed only 3!

Multiple Sequence Alignment

$\alpha_{i_1,i_2,\ldots,i_N}$ - maximum score of an alignment up to the subsequences ending with $x^1_{i_1}, x^2_{i_2}, \ldots, x^N_{i_N}$

Recursive relation for multiple sequences:

$$\alpha_{i_1,i_2,\ldots,i_N} = max \begin{cases} \alpha_{i_1-1,i_2-1,\ldots,i_N-1} + S(x^1_{i_1}, x^2_{i_2}, \ldots, x^N_{i_N}), \\ \alpha_{i_1,i_2-1,\ldots,i_N-1} + S(-, x^2_{i_2}, \ldots, x^N_{i_N}), \\ \alpha_{i_1-1,i_2,i_3-1,\ldots,i_N-1} + S(x^1_{i_1}, -, \ldots, x^N_{i_N}), \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{i_1-1,i_2-1,\ldots,i_N} + S(x^1_{i_1}, x^2_{i_2}, \ldots, -), \\ \alpha_{i_1,i_2,i_3-1,\ldots,i_N-1} + S(-, -, \ldots, x^N_{i_N}), \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{i_1,i_2-1,\ldots,i_{N-1}-1,,i_N} + S(-, x^2_{i_2}, \ldots, -), \\ \cdot \\ \cdot \\ \cdot \end{cases}$$

Multiple Sequence Alignment

To calculate each entry, need to maximize over all $2^N - 1$ combinations of gaps in a column, excluding the case where all Δ_k are zero.

Introducing the notation Δ_i which is 0 or 1 and define the 'product'

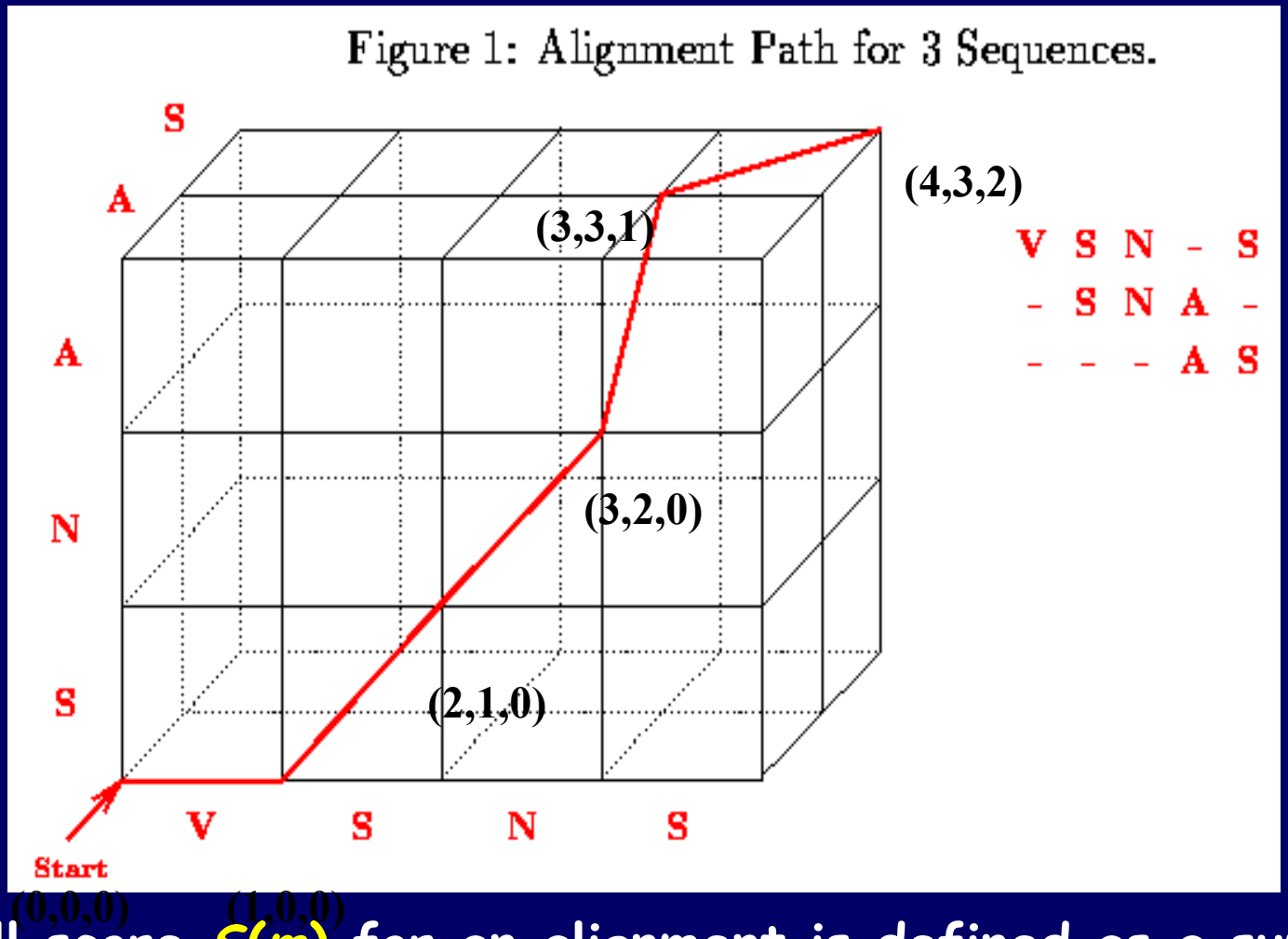
$$\Delta_i \cdot x = \begin{cases} x & if \quad \Delta_i = 1, \\ - & if \quad \Delta_i = 0. \end{cases}$$

Recursion relation can now be written as

$$\alpha_{i_1,i_2,\ldots,i_N} = \max_{\Delta_1+\ldots+\Delta_N>0} \left\{ \alpha_{i_1-\Delta_1,i_2,-\Delta_2\ldots,i_N-\Delta_N} + S(\Delta_1 \cdot x^1_{i_1}, \Delta_2 \cdot x^2_{i_2},\ldots,\Delta_N \cdot x^N_{i_N} \right\}$$

For N = 3, 4, 5, and 10, $2^N - 1$ = ?

DP Hyperlattice: Example



Overall score $S(m)$ for an alignment is defined as a sum of scores $S(m_i)$ for each column i: $S(m) = \Sigma_i S(m_i)$

Time and Space complexity

Assuming all sequences of roughly same length L, memory complexity of the multi-dimensional DP algorithm is $O(L^N)$ and time complexity is $O(2^N L^N)$

- impractical for more than a few sequences.

For 6 sequences, each 100bp long, time taken will be $2^6 \times 100^6 \times 10^{-9}$ = 64000 seconds (~ 18 hrs)

Add 2 more sequences of same length and the no. is 2.56×10^9 seconds (over 81 yrs)

Even worse is memory space requirement - 10^{12} for 6 sequences!

Progressive approach

- Align each sequence to every other pair-wise
- Compute distances between each aligned pair (*e.g.* no. of mismatches)
- Construct a phylogenetic tree
- Cluster closely related sequences
- Align closely related sequences first
- Gaps inserted in closely related sequences are propagated throughout

Progressive alignment - involves constructing a succession of pairwise alignments.

Tools: ClustalW,T-Coffee, MUSCLE

Pairwise alignments

```
Sheep    STCVLSAYWKDLNNYH
Cattle   STCVLSAYWKDLNNYH

Sheep    STCVLSAYWK-DLNNYH
Pig      STCVLSAYWRNELNNFH

Sheep    STCVLSAYWKDLNNYH
Human    STCMLGTY-QDFNKFH

Sheep    STCVLSAYWKDLNNYH
Rat      STCMLGTY-QDLNKFH

Sheep    STCVLSAYWKD-LNNYH
Salmon   STCVLGKL-SQELHKLQ

Pig      STCVLSAYWRNELNNFH
Human    STCMLGTY-QD-FNKFH

Pig      STCVLSAYWRNELNNFH
Rat      STCMLGTY-QD-LNKFH

Human    STCMLGTY-QDFNKFH
Salmon   STCVLGKLSQELHKLQ

Rat      STCMLGTY-QDLNKFH
Salmon   STCVLGKLSQELHKLQ
```

Hierarchy of Addition

Sheep-Cattle	0	Pig-Rat	8
Sheep-Pig	4	Pig-Salmon	10
Sheep-Human	8	Human-Rat	1
Sheep-Rat	7	Human-Salmon	9
Sheep-Salmon	11	Rat-Salmon	8
Pig-Human	9		

- Align Sheep and Cattle first
- Align Human and Rat
- Align Pig to Sheep and Cattle
- Align these two clusters to each other
- …
- Align Salmon to large alignment last

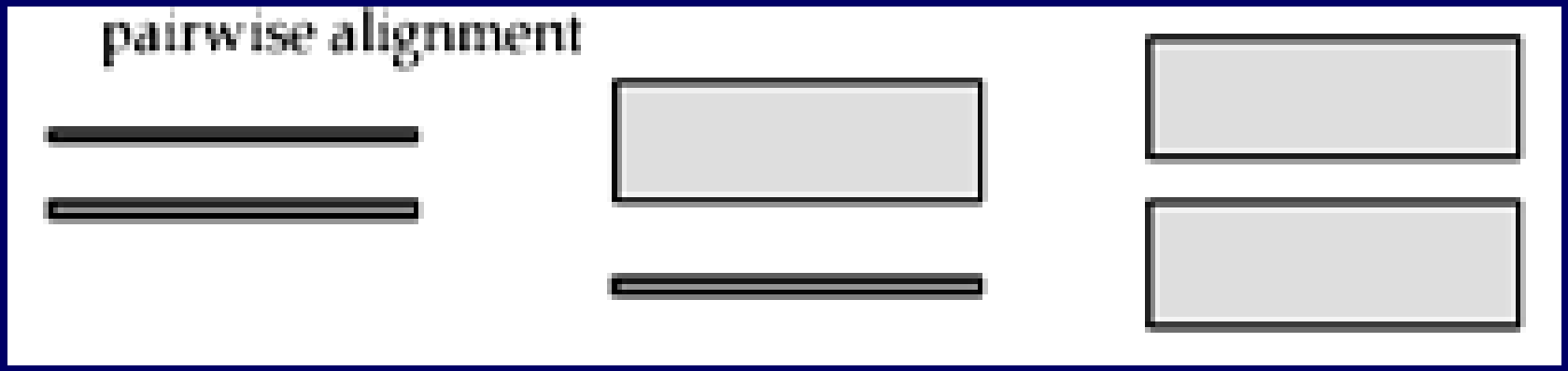
Progressive alignment

```
step 1   Sheep    STCVLSAYWKDLNNYH
         Cattle   STCVLSAYWKDLNNYH

step 2   Human    STCMLGTYQDFNKFH
         Rat      STCMLGTYQDLNKFH

step 3   Sheep    STCVLSAYWK-DLNNYH
         Cattle   STCVLSAYWK-DLNNYH
         Pig      STCVLSAYWRNELNNFH

...      Salmon   STCVLGKLSQE-LHKLQ
```

Aligning Alignments

All possible cases that arise in progressive alignment approach:

- Align two sequences to each other
- Align a sequence to an existing alignment
- Align two alignments to each other



Note - computationally this is always a PW

Aligning Alignments

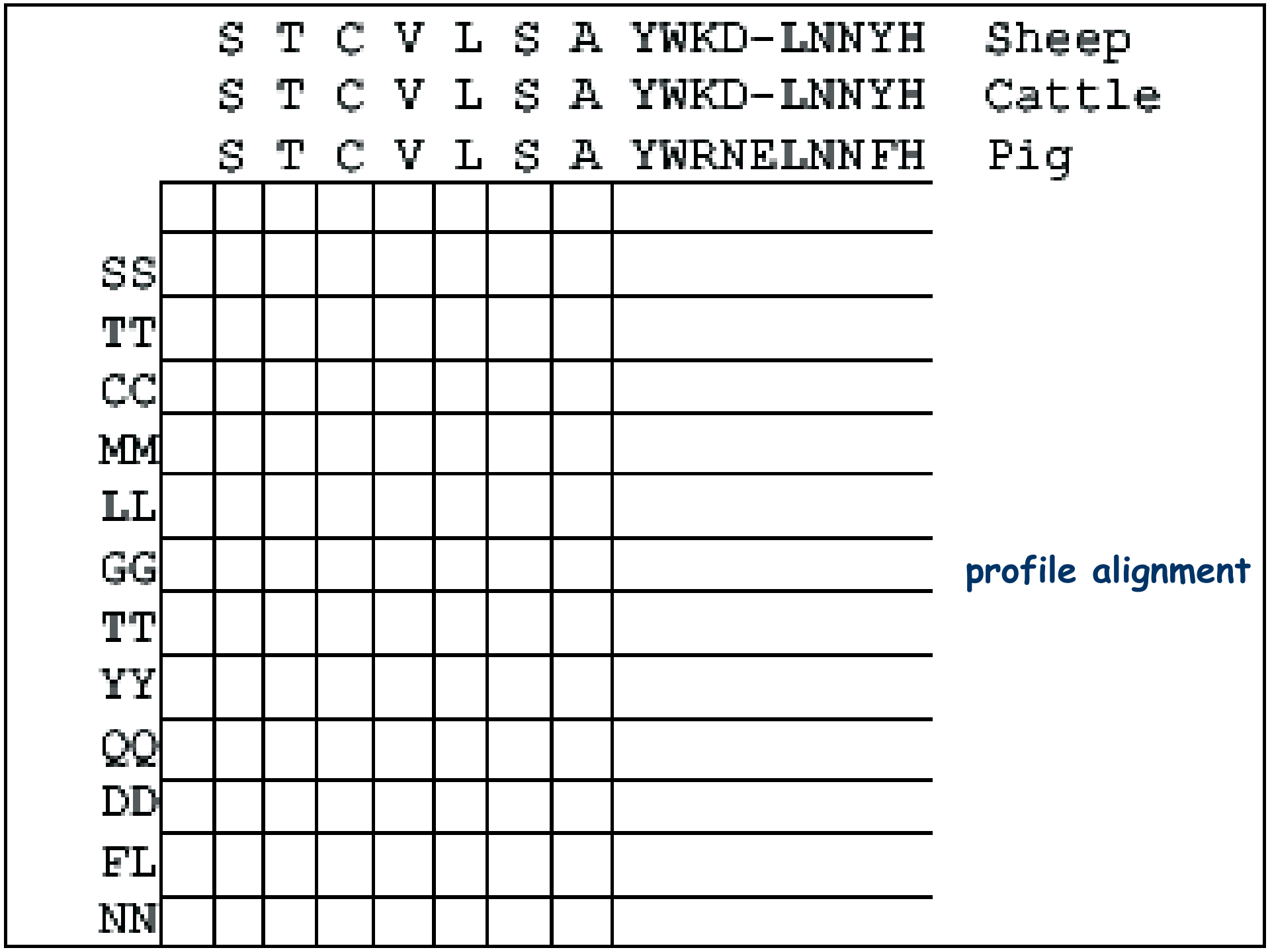
- Pairwise alignment of alignments is also called a profile alignment

Again, we can use DP

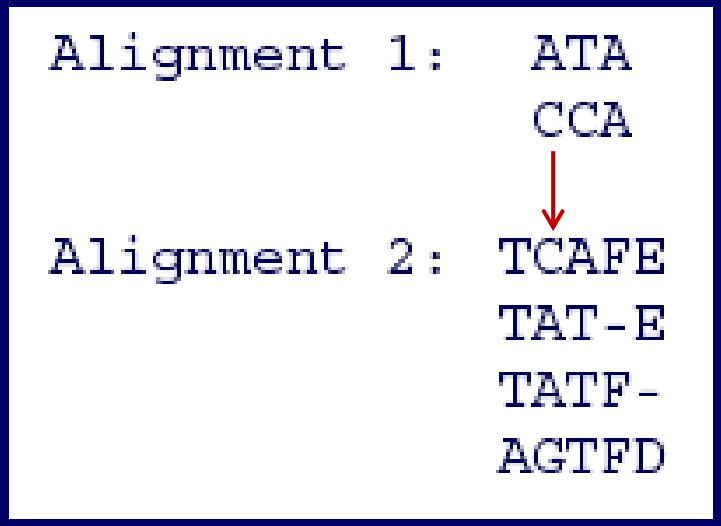
$$S(i,j) = \max[S(i-1,j-1)+m(i,j),\ S(i-1,j)+g,\ S(i,j-1)+g]$$

$m(i,j)$ – similarity score averaged over characters at that position, here i and j no longer stands for single sequences, but an average of several, when aligning alignments

g – gap penalty


S T C V L S A YWKD-LNNYH Sheep
S T C V L S A YWKD-LNNYH Cattle
S T C V L S A YWRNELNNFH Pig
SS
TT
CC
MM
LL
GG
TT
YY
QQ
DD
FL
NN
profile alignment

Aligning Alignments



Score 1st column of 1st alignment against 2nd column in the other alignments using:

= 1/8(score(A,C) + score(A,A) + score(A,A) + score(A,G) + score(C,C) + score(C,A) + score(C,A) + score(C,G))

Aligning Alignments

- Once sequences are aligned & gaps introduced, these are not altered – the alignment method is **hierarchical**
- ClustalW finds a local optimum as early alignment decisions are "locked in" by the "greedy" algorithm
- Early errors will be propagated and cause the final alignment to be worse

```
Sheep     STCVLSAYWK-DLNNYH
Cattle    STCVLSAYWK-DLNNYH
Pig       STCVLSAYWRNELNNFH
```

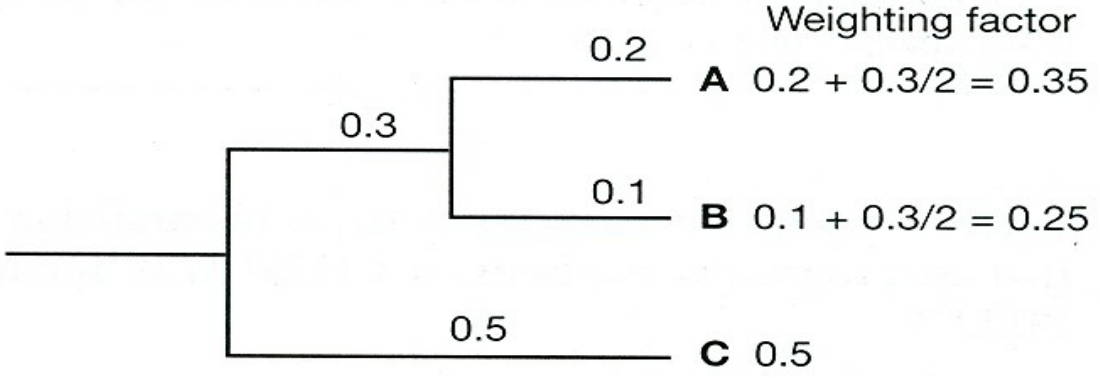
ClustalX/ClustalW

- Performs pair-wise alignments of all the sequences
  - k-tuple based alignment (fast/approx.), full DP (slow/accurate)
- Uses alignment scores to produce a phylogenetic tree
  - Genetic distance: no. of mismatches/no. of matches (positions against gaps not considered)
- Aligns sequences sequentially, guided by the phylogenetic relationships indicated by tree.
  - Sequence contributions are 'weighted' by their relationship on the predicted tree, weights based on the distance of each sequence from root

Alignment scores between two positions in msa calculated using resulting weights as multiplication factors

Weights are normalized so that largest weight is 1

A. Calculation of sequence weights



B. Use of sequence weights

Column in alignment 1

Sequence A (weight a) .....K.....

Sequence B (weight b) .....I.....

Column in alignment 2

Sequence C (weight c) .....L.....

Sequence D (weight d) .....V.....

Score for matching these two column in an msa =

[a x c x score (K,L) +
a x d x score (K,V) +
b x c x score (I,L) +
b x d x score (I,V)] / 4

Basic idea in Progressive Heuristic Approach:

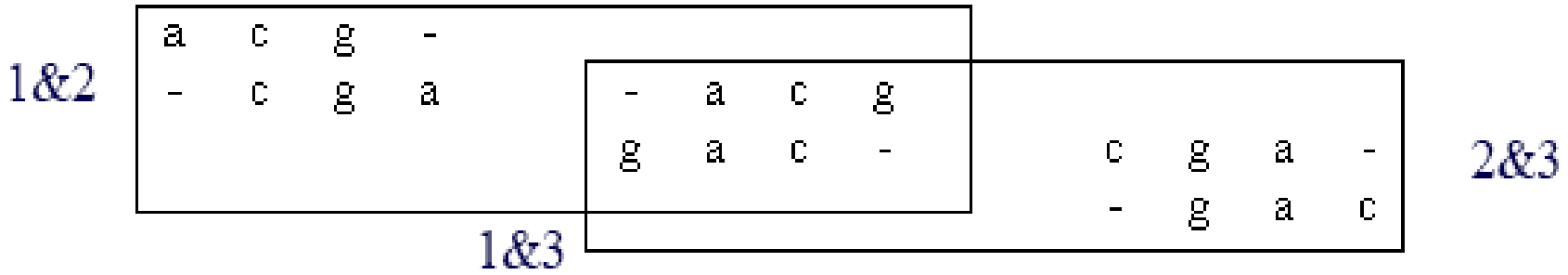
- compute pairwise alignments and merge alignments consistently

Consider alignment of 3 sequecens:

acg, cga, gac

Get optimal pairwise alignments:

a c g -	- a c g	c g a -
- c g a	g a c -	- g a c



Merge using alignments with 1st sequence

```
-acg-
--cga
gac--
```

Merge using alignments with 3rd sequence

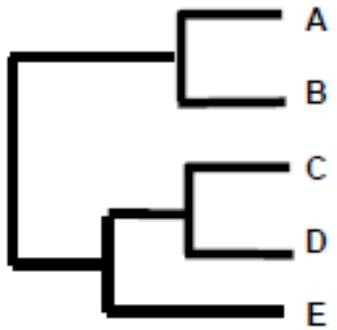
```
--acg
cga--
-gac-
```

Merge using alignments with 2nd sequence

```
acg--
-cga-
--gac
```

Order of merging matters !
Note once a gap, always a gap …

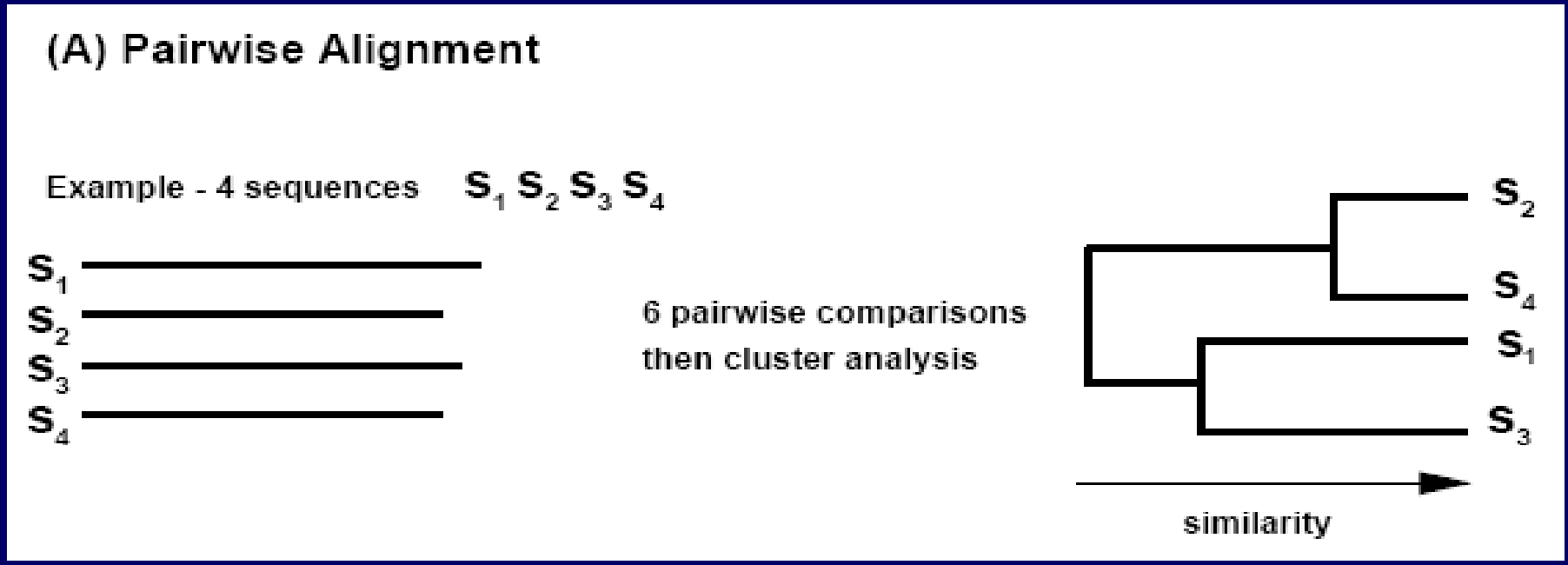
Example



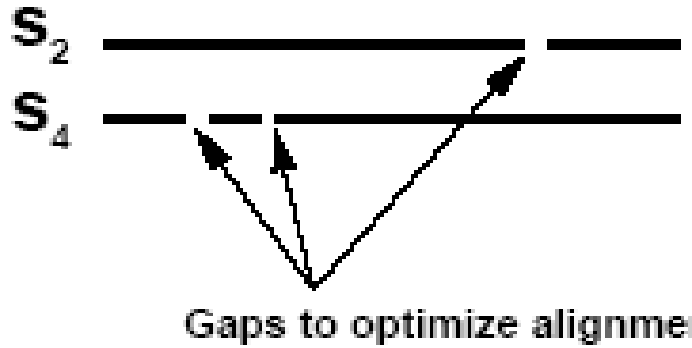

```
C   PADKTNVKAAWGKVGAHAGEYGA
D   AADKTNVKAAWSKVGGHAGEYGA

A   PEEKSAVTALWGKVNVDEYGG
B   GEEKAAVLALWDKVNEEEYGG

C   PADKTNVKAAWG_KVGAHAGEYGA
D   AADKTNVKAAWS_KVGGHAGEYGA
E   AA__TNVKTAWSSKVGGHAPA__A

A   PEEKSAV_TALWG_KVN__VDEYGG
B   GEEKAAV_LALWD_KVN__EEEYGG
C   PADKTNVKAA_WG_KVGAHAGEYGA
D   AADKTNVKAA_WS_KVGGHAGEYGA
E   AA__TNVKTA_WSSKVGGHAPA__A
```

Once a gap, always a gap

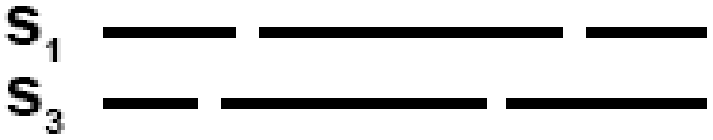
Steps in Progressive Alignment



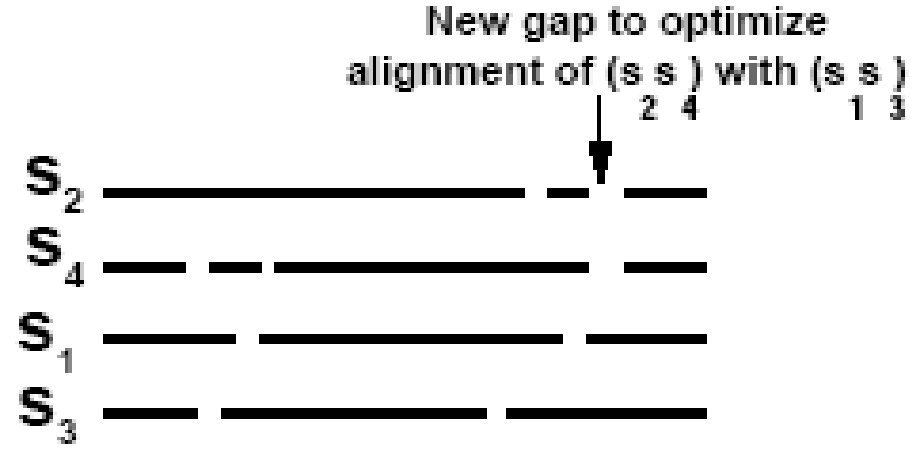
(B) Multiple alignment following the tree from A



align most similar pair



align next most similar pair



align alignments - preserve gaps

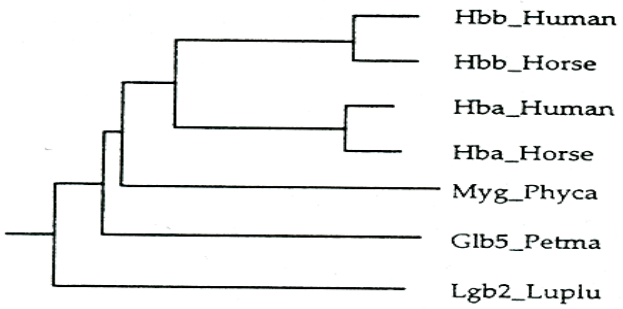
ClustalW: Refinements

- Different BLOSUM matrices used based on similarity of sequences – to reflect evolutionary changes better
- Recall: BLOSUM80/ BLOSUM62 are based on sequences that are 80% / 62% identical, *i.e.* lower numbers for more distant sequences
- ClustalW calculates gaps in a way to place them between secondary structural elements
- Frequency of gaps next to each amino acid is based on the table provided by Pascarella and Argos.

Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90
		1	2	3	4	5	6

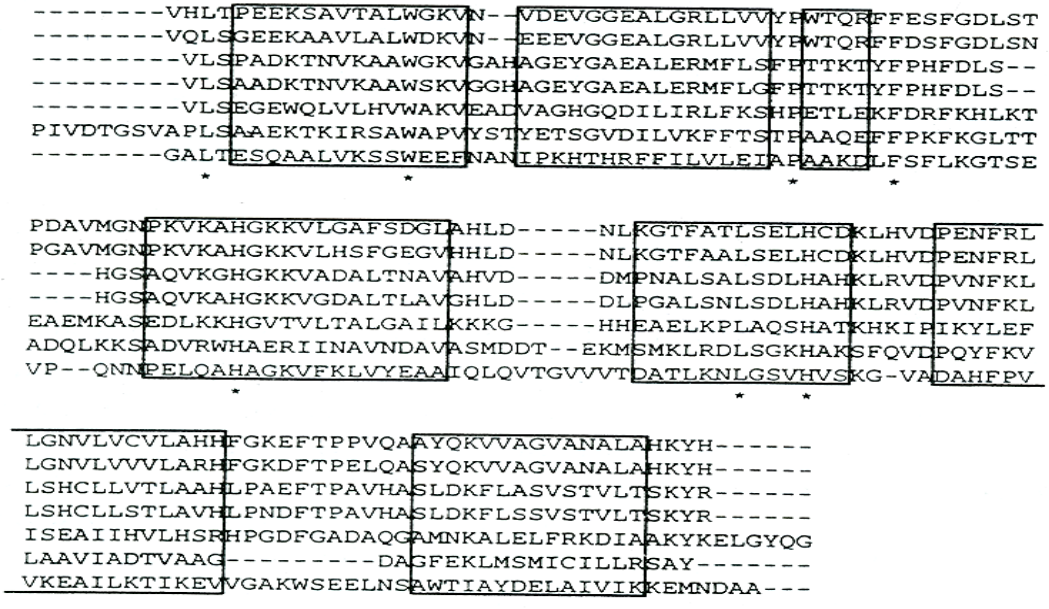
Pairwise alignment: Calculate distance matrix

MSA of 7 globins by CLUSTALW



Rooted Neighbor Joining tree (guide tree)

Progressive alignment: Align following the guide tree

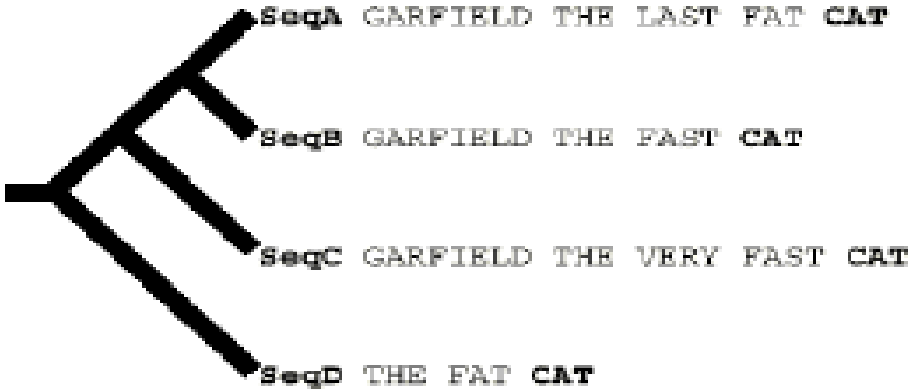


Known locations of 7 α-helices in the structure of this group shown in boxes

Problems with Progressive Alignment

Gaps at the ends are penalized less, so CAT is aligned with FAT in sequence 2

The greedy approach results in efficiency of the algorithm at the cost of accuracy

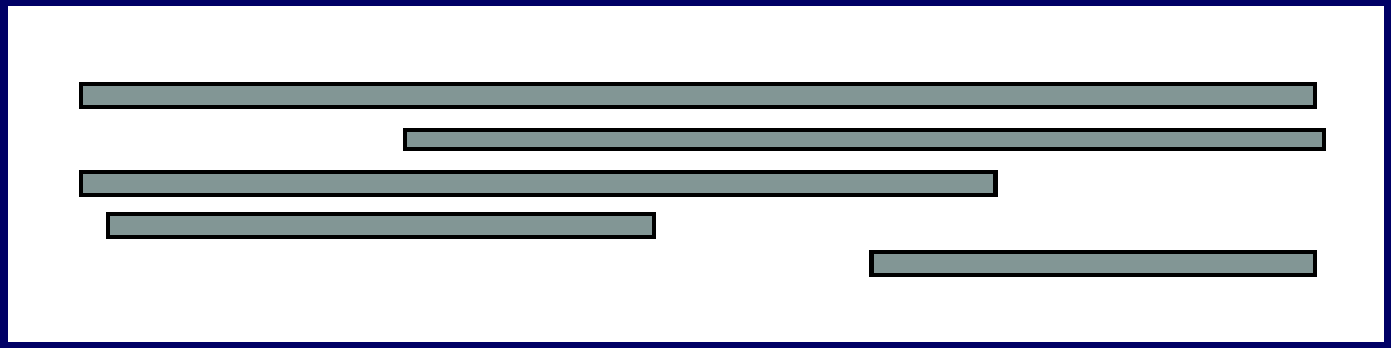



```
CLUSTALW (Score=20, Gop=-1, Gep=0, M=1)

SeqA GARFIELD THE LAST  FA-T CAT
SeqB GARFIELD THE FAST  CA-T ---
SeqC GARFIELD THE VERY  FAST CAT
SeqD -------- THE ----  FA-T CAT
```

```
CORRECT (Score=24)

SeqA GARFIELD THE LAST  FA-T CAT
SeqB GARFIELD THE FAST  ---- CAT
SeqC GARFIELD THE VERY  FAST CAT
SeqD -------- THE ----  FA-T CAT
```

ClustalW Misapplied

ClustalW and other algorithms that include an initial pair-wise comparison step should not be used to align sequences that do not all share a common block.

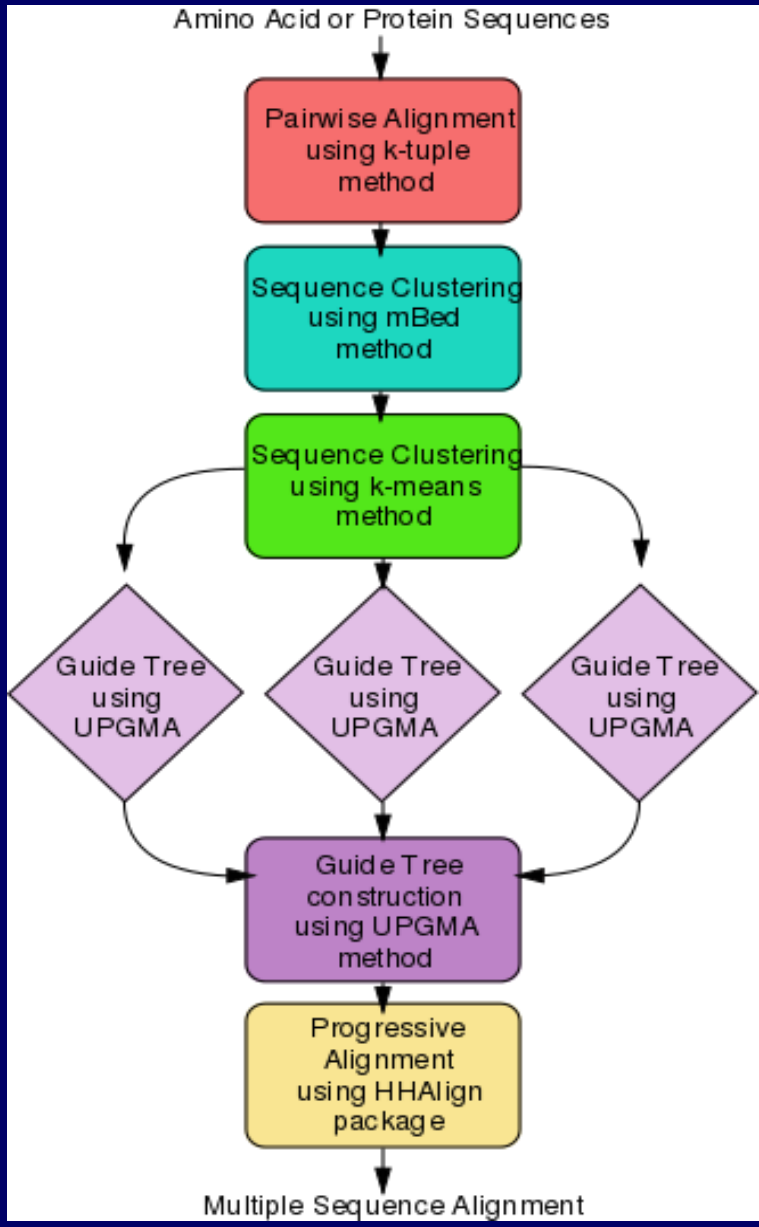


Not suitable for sequence assembly!

The latest version, called **Clustal Omega**, uses seeded guide trees and HMM profile-profile techniques to generate alignments

ClustalW / ClustalΩ online

1. **Produce pairwise alignment using k-tuple method**
2. **Sequences are clustered using mBed method, which calculates pairwise distance using sequence embedding.**
3. **This is followed by k-means clustering method.**
4. **Next, the guide tree is constructed using the UPGMA method.**
5. **Finally, MSA is produced using HHAlign package from the HH-Suite, which uses two profile HMM's.**



https://www.ebi.ac.uk/Tools/msa/

Problems with Progressive Alignment

- Depends on the very first closely related sequences used for constructing the multiple alignment
- If these sequences align well, there will be few errors
- More distantly related these sequences are, more errors will get propagated through the alignment
- Using Bayesian methods such as Hidden Markov models (HMMs) may be useful for aligning more distantly related sequences.

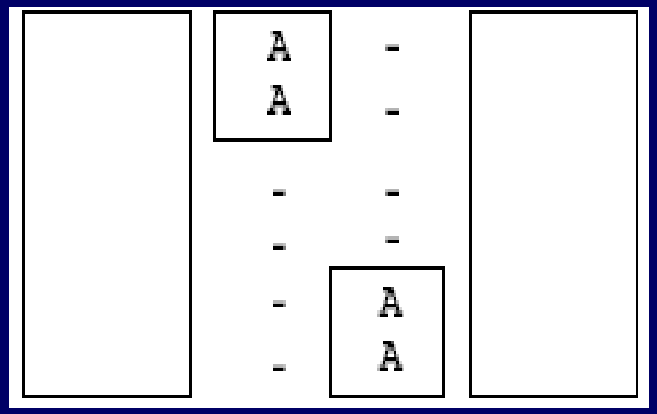
Solution: Stochastic or Iterative Methods

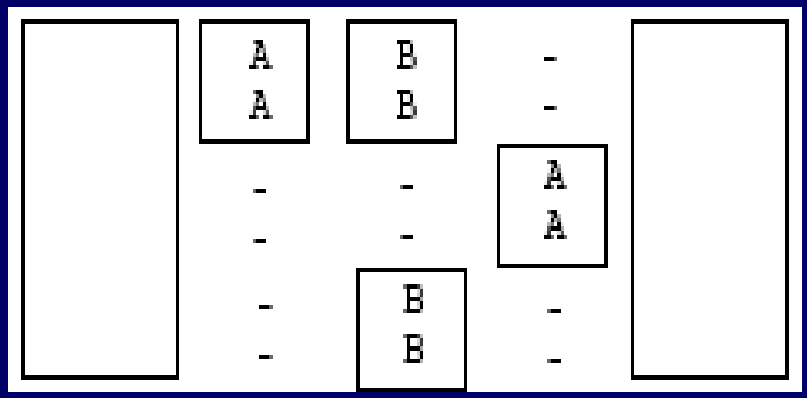
Iterative Methods of MSA

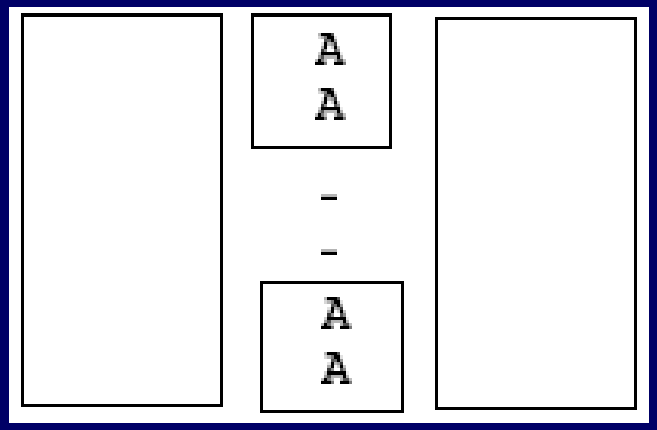
- To correct for errors introduced by initial alignment, use iterative methods: re-align sub-groups of sequences and then align these sub-groups into a global alignment
- Objective – to improve overall alignment score
- Selection of groups may be based on phylogenetic tree, separation of one or two sequences from the rest, or a random selection of the groups.
- Programs using iterative methods – MultiAlin, PRRP and DIALIGN

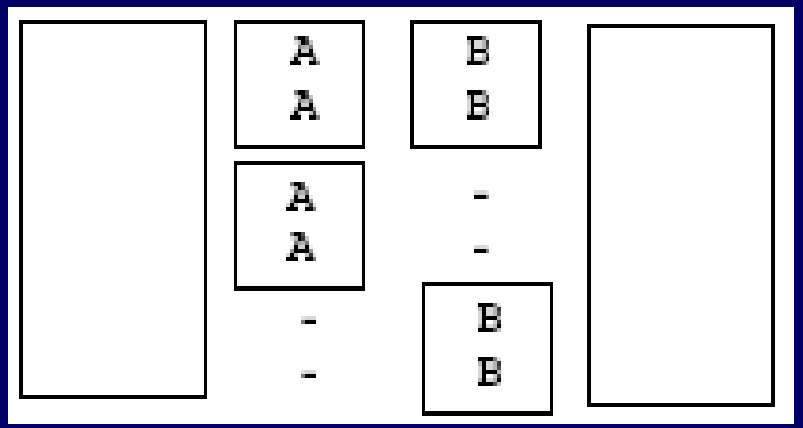
Examine alignments by eye

Some artifacts observed in the output of algorithms:









Always examine your alignment manually to see if it can be improved.

Edit alignments manually

Multiple sequence alignment tools

- Viewers
  - ClustalX, Jalview, Cinema, Sequence logos
- Editors / annotation
  - SeqVu, MACAW, BioEdit
- BioEdit available at:
  https://bioedit.software.informer.com/7.2/

Work with proteins

- Twenty symbols to match as against four for DNA
- No noise resulting from the degeneracy of genetic code
- More sensitive scoring matrices
- Requires less of manual editing

Choose genes judiciously

- When inferring phylogeny choose genes carefully
- For closely related organisms choose genes which mutate fast
- For distantly related species choose slowly mutating genes
- Compare orthologous genes between species and paralogous ones within an organism

Summary

- Treat the output of multiple alignment programs as a first alignment
- Examine it by eye and edit it manually to improve it
- Get rid of low confidence or highly divergent regions
- Ensure you have started with a sensible evolutionary hypothesis

Summary

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

References

- David W. Mount, Bioinformatics: Sequence and Genome Analysis, CBS Publishers & distributors, New Delhi, and references therein.
- Thompson, J.D., Higgins, D.G. and Gibson. T.J. "CLUSTALW: improving the sensitivity of multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice" *Nuc. Acids Res.* 22, 4673-80 (1994).