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 <u>similar</u> or even identical function, and at other times, mutating or rearranging to perform an <u>altered</u> 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.

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

$\Rightarrow$ **few inferences can be drawn.**

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

$\Rightarrow$ **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:

- **Identify 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**

Inferences from MSA

Some examples:

- **Secondary structure prediction:**

- Conserved pattern of hydrophobicity with spacing **2** with intervening residues more variable and including hydrophilic residues – **suggests β-strand**

- Conserved pattern of hydrophobicity with spacing **4 suggests a helix**

12 glutamyl tRNA reductase sequences

13 - 16

JPRED server

38 – 45 α-helix i, i+3, i+4, i+7

β-strand

hydrophobic

Non-hydrophobic (active sites)

Conservation

Consensus

Relationship between MSA & Phylogenetic Analysis

SeqA N – F L S
SeqB N – F – S
SeqC N KY L S
SeqD N – Y L S
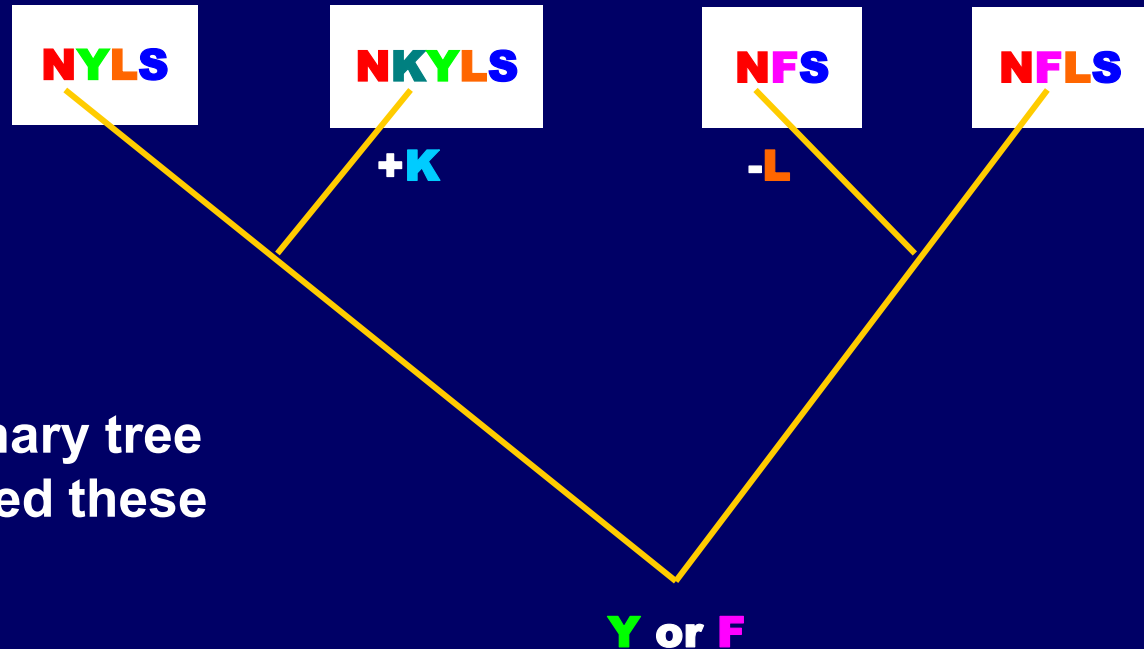
A hypothetical evolutionary tree that could have generated these sequence changes.

NYLS   NKYLS   NFS   NFLS

+K   -L

Y or F

MSA of 8 fragments of immunoglobin sequences

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

Conserved residues, regions, patterns

alternating hydrophobicity pattern, conserved Cys, W

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

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

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

Multiple Alignment: Evaluation

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

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
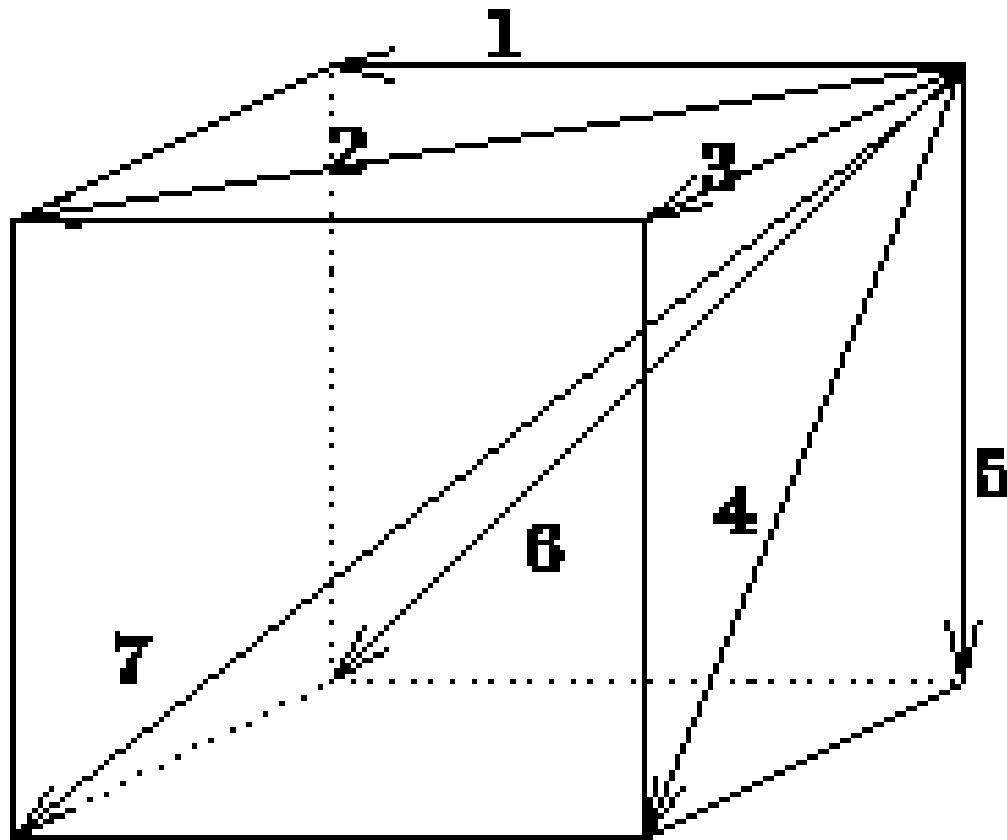
current
visit

i, j-1, k
i, j, k-1
i, j-1, k-1
i-1, j, k
i-1, j-1, k
i-1, j, k-1
i-1, j-1, k-1

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

$\boxed{\alpha_{i_1,i_2,...,i_N}}$ – maximum score of an alignment up to the subsequences ending with $\boxed{x_{i_1}^1, x_{i_2}^2, ..., x_{i_N}^N}$

Recursive relation for multiple sequences:

$$\alpha_{i_1,i_2,...,i_N} = max \begin{cases} \alpha_{i_1-1,i_2-1,...,i_N-1} & + & S(x_{i_1}^1, x_{i_2}^2, ..., x_{i_N}^N), \\ \alpha_{i_1,i_2-1,...,i_N-1} & + & S(-, x_{i_2}^2, ..., x_{i_N}^N), \\ \alpha_{i_1-1,i_2,i_3-1,...,i_N-1} & + & S(x_{i_1}^1, -, ..., x_{i_N}^N), \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ \alpha_{i_1-1,i_2-1,...,i_N} & + & S(x_{i_1}^1, x_{i_2}^2, ..., -), \\ \alpha_{i_1,i_2,i_3-1,...,i_N-1} & + & S(-, -, ..., x_{i_N}^N), \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ \alpha_{i_1,i_2-1,...,i_{N-1}-1,,i_N} & + & S(-, x_{i_2}^2, ..., -), \\ & \cdot & \\ & \cdot & \\ & \cdot & \end{cases}$$

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

$$\alpha_{i_1,i_2,\dots,i_N} = \max_{\Delta_1+\dots+\Delta_N>0} \left\{ \alpha_{i_1-\Delta_1,i_2,-\Delta_2\dots,i_N-\Delta_N} + S(\Delta_1 \cdot x_{i_1}^1, \Delta_2 \cdot x_{i_2}^2,\dots,\Delta_N \cdot x_{i_N}^N) \right\}$$

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

DP Hyperlattice: Example

Figure 1: Alignment Path for 3 Sequences.
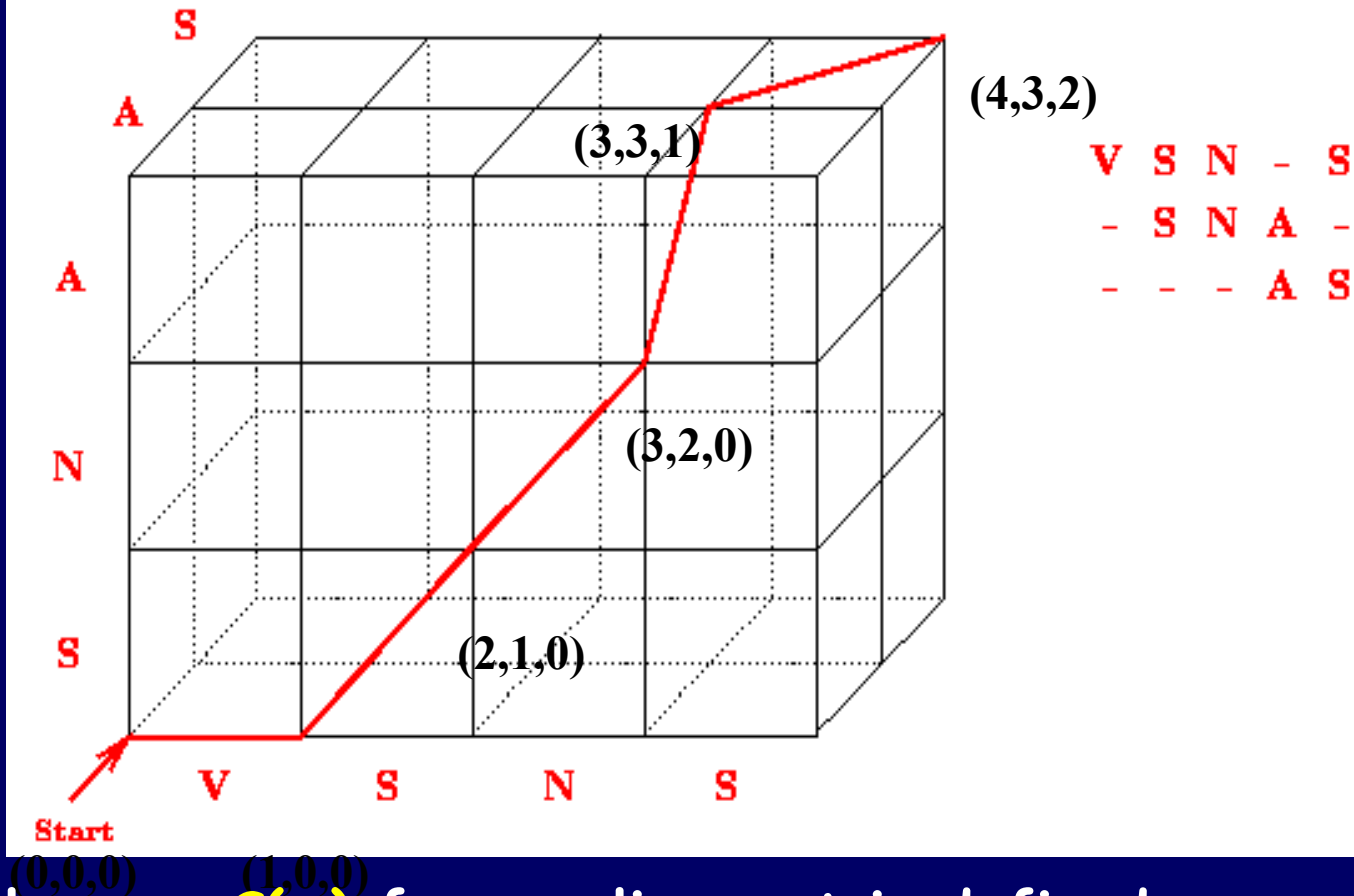
Overall score $S(m)$ for an alignment is defined as a sum of scores $S(m_i)$ for each column i: $S(m) = \Sigma_i S(m_i)$

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!

Carrilo-Lipman/MSA Algorithm

- **Carrillo and Lipman proposed a heuristic method for accelerating the search, implemented in the program called MSA.**

- **It is based on the property that if the strings are relatively similar, the alignment path would be close to the main diagonal.**

- ⇒ **not all values in the multi-dimensional cube need to be calculated.**

- **Their central idea is - every multiple alignment imposes a pw-alignment on each pair of sequences**

- **For each pair, it sets the upper bound equal to the cost of the imposed alignment.**
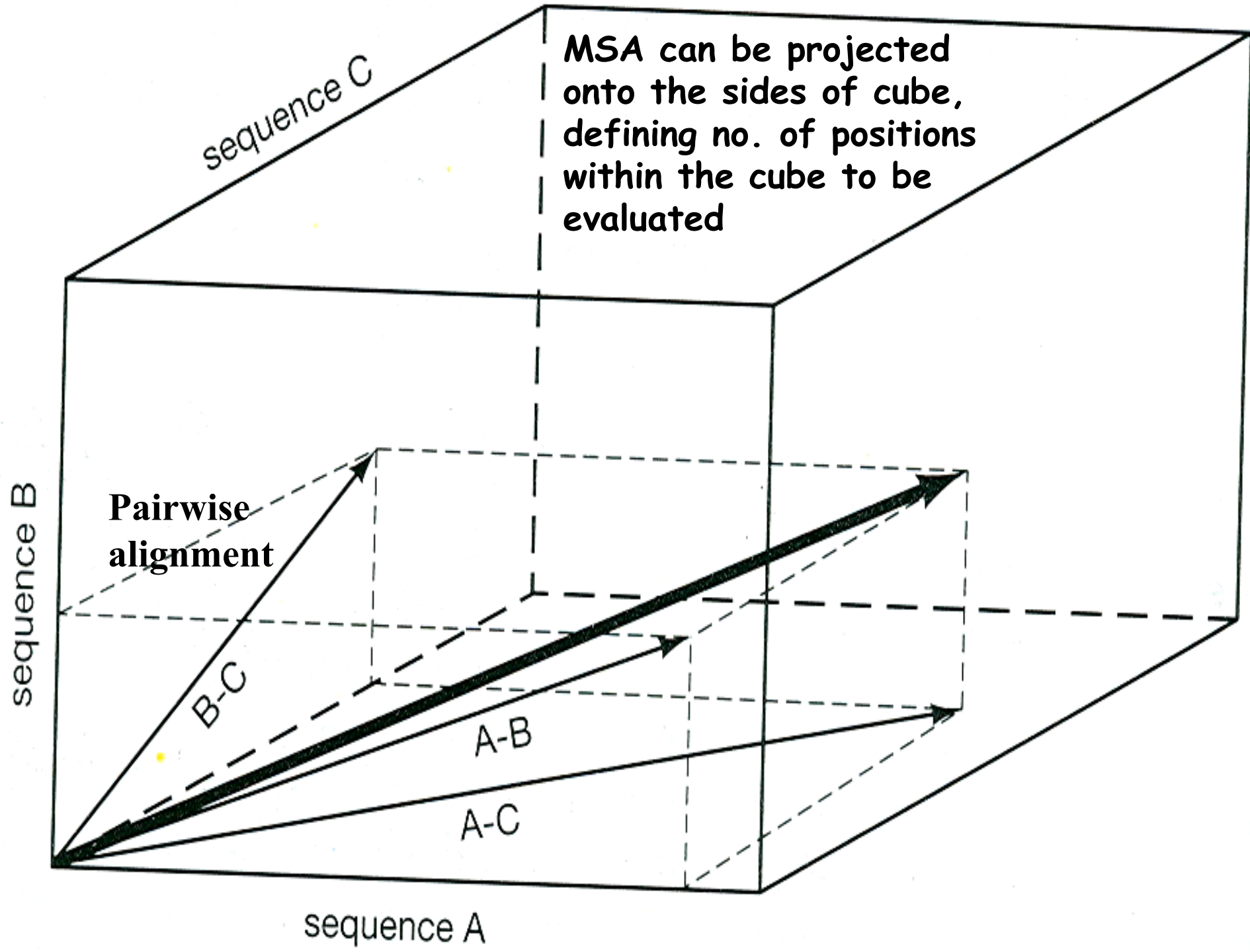
sequence C

MSA can be projected onto the sides of cube, defining no. of positions within the cube to be evaluated

sequence B

Pairwise alignment

B-C

A-B

A-C

sequence A

Bounds within which an optimal alignment will be found by MSA

Coloured areas on each cube surface - 2d projections of the gray volume

it is only necessary to calculate optimal alignment scores within the gray volume

Multiple alignment DP: Heuristics

- Optimal alignment path will lie around the main diagonal (MSA, by Carillo-Lipman, 1988)
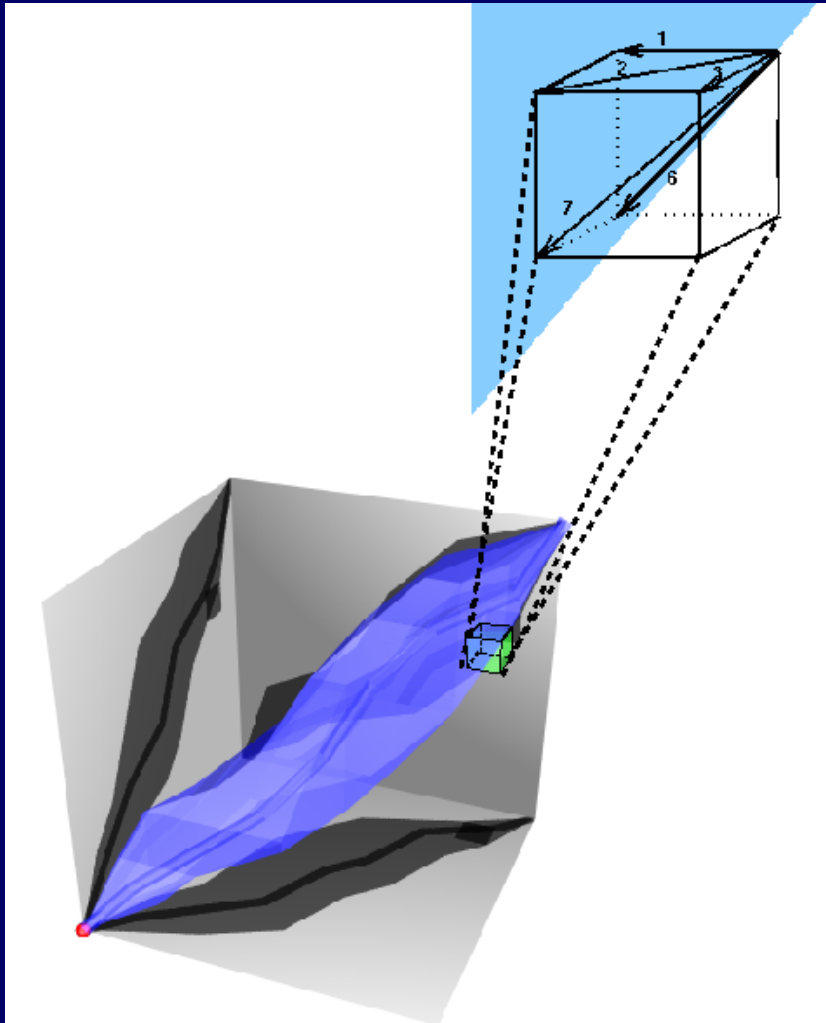
While visiting a node and looking for the minimum along all the incoming edges - ignore the edges "coming from outside polyhedron". Edges 1, 2, 3, 6, 7 are coming from inside and edges 4 and 5 (not shown) can be ignored

Multiple alignment DP: Heuristics

- **Optimal alignment path will lie around the main diagonal (MSA, by Carillo-Lipman, 1988)**

- **It uses a heuristic approach to choose upper bounds for the pairs.**

- **Works for very similar sequences**

- **MSA can find optimal multiple alignments for ~ 6-8 sequences, 200-300 length**

Alignment by Consensus

- **Aim of alignment by Consensus: to minimize the consensus error**

- **Align sequence 1 & 2 and obtain a consensus sequence**

- **Building consensus: an example**

$$
\begin{array}{cccccc}
A & V & - & C & - & D \\
A & V & C & C & - & D \\
A & V & C & D & E & - \\
\hline
A & V & C & S & E & D
\end{array}
$$

- **As is clear from above, the consensus need not be similar to any of the aligned sequences**

Consecutive consensus

- **Align sequence 3 to the consensus of first two sequences to obtain a new consensus**

- **Continue until the consensus alignment converges to a global consensus**

- **Problem - Depends on the order in which the sequences are introduced**

Solution ?

Consecutive consensus

- **Align sequence 3 to the consensus of first two sequences to obtain a new consensus**

- **Continue until the consensus alignment converges to a global consensus**

- **Problem - Depends on the order in which the sequences are introduced**

- **Solution   - one may use a guide tree obtained by p-w alignment scores to build the consensus**

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-ENKFH

Pig   STCVLSAYWRNELNNFH
Rat   STCMLGTY-QD-LNKFH

Pig   STCVLSAYWRNELNNFH
Salmon   STCVLGKL-SQELHKLQ

Human   STCMLGTYQDFNKFH
Rat   STCMLGTYQDLNKFH

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**

pairwise alignment
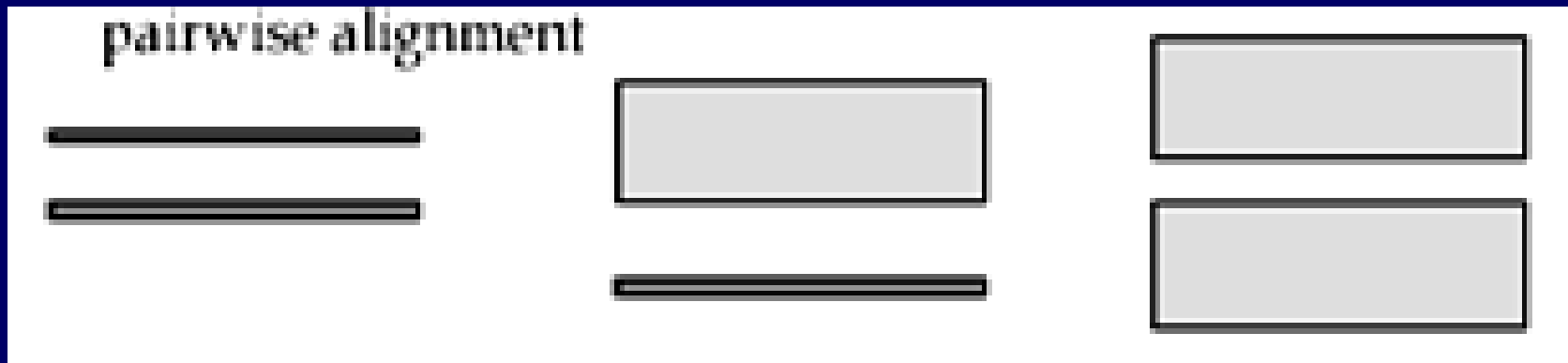
Note - computationally this is always a PW

Aligning Alignments

- Pairwise alignment of alignments is also called <u>profile</u> 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									

Aligning Alignments

```
Alignment 1:    ATA
                CCA
                        |
                        ↓
Alignment 2:    TCAFE
                TAT-E
                TATF-
                AGTFD
```

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 <u>hierarchical</u>**

- **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

Sheep     STCVLSA--YWK-DLNNYH
Cattle    STCVLSA--YWK-DLNNYH
Pig       STCVLSA--YWRNELNNFH
```

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

Weighting factor

A 0.2 + 0.3/2 = 0.35

B 0.1 + 0.3/2 = 0.25

C 0.5

(tree branch lengths: 0.2, 0.3, 0.1, 0.5)
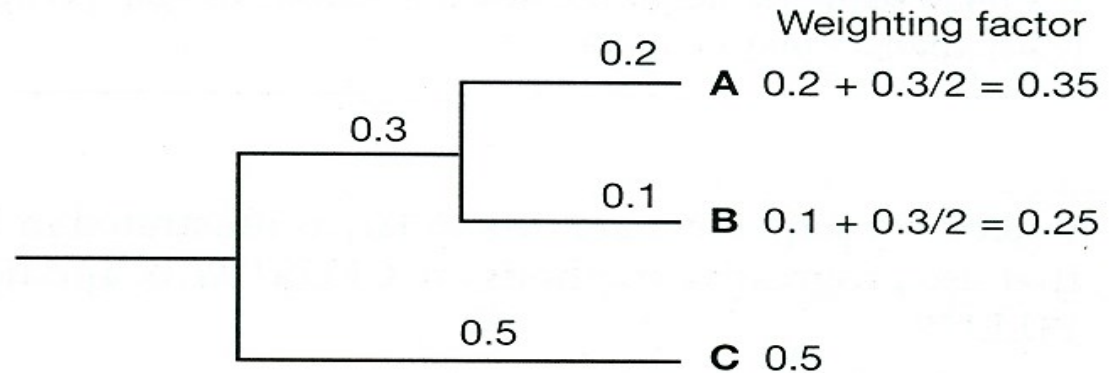
B. Use of sequence weights

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
```

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

Example
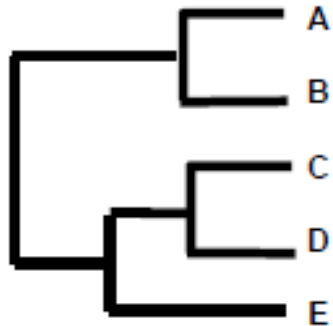
| C | PADKTNVKAAWGKVGAHAGEYGA |
| D | AADKTNVKAAWSKVGGHAGEYGA |

| 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
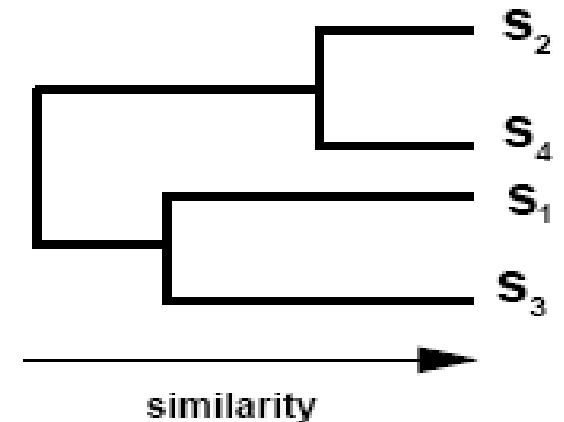
(B) Multiple alignment following the tree from A

S_2 ───────────────── ─── ─────

S_4 ── ── ─────────────────

Gaps to optimize alignment
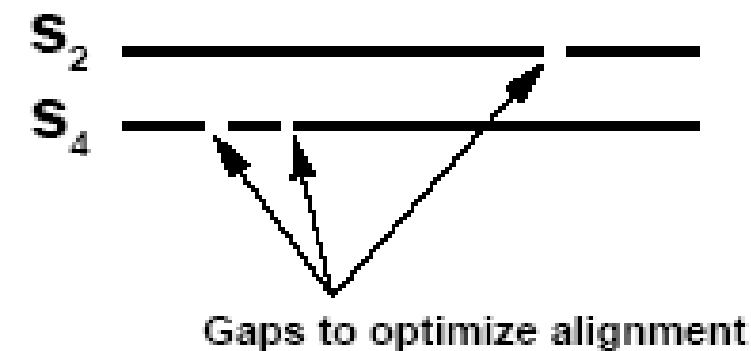
align most similar pair

S_1 ── ── ───── ── ──── ──

S_3 ── ── ───────── ── ─────

align next most similar pair

New gap to optimize
alignment of (s s) with (s s)
                 2 4   1 3

S_2 ─────────── ── ─

S_4 ── ── ── ─────── ──

S_1 ── ── ──────── ── ──

S_3 ── ── ──────── ── ───
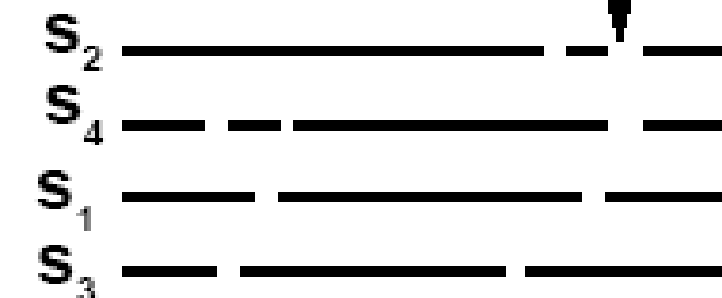
align alignments - preserve gaps

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 conserved domains, based on observation that structurally related proteins align in a way to have gaps between secondary structural elements**

- **Frequency of gaps next to each amino acid computed in these regions is used (Pascarella and Argos)**

		1	2	3	4	5	6
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

Pairwise alignment: Calculate distance matrix

MSA of 7 globins by CLUSTALW

Hbb_Human
Hbb_Horse
Hba_Human
Hba_Horse
Myg_Phyca
Glb5_Petma
Lgb2_Luplu
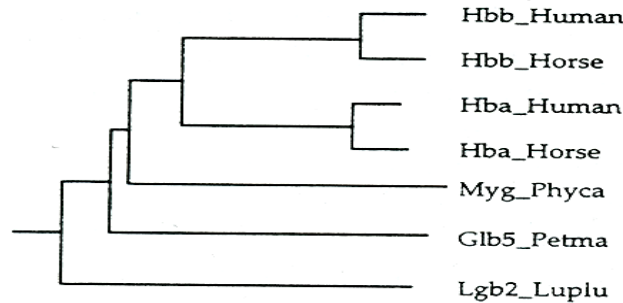
Rooted Neighbor Joining tree (guide tree)

Progressive alignment: Align following the guide tree

```
--------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
--------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
--------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS--
--------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLS--
--------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
--------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
        *                   *                    *

PDAVMGNPKVKAHGKKVLGAFSDGLAHLD-----NLKGTFATLSELHCDKLHVDPENFRL
PGAVMGNPKVKAHGKKVLHSFGEGVHHLD-----NLKGTFAALSELHCDKLHVDPENFRL
----HGSAQVKGHGKKVADALTNAVAHVD-----DMPNALSALSDLHAHKLRVDPVNFKL
----HGSAQVKAHGKKVGDALTLAVGHLD-----DLPGALSNLSDLHAHKLRVDPVNFKL
EAEMKASEDLKKHGVTVLTALGAILKKKG-----HHEAELKPLAQSHATKHKIPIKYLEF
ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-VADAHFPV
        *                        *              *

LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
LAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
```
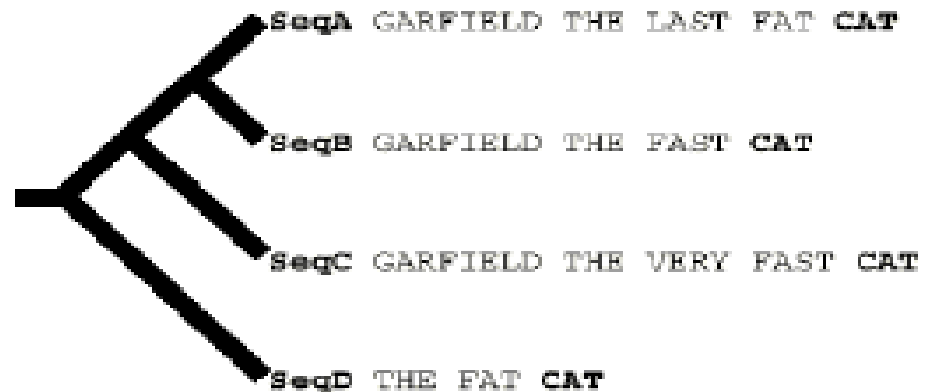
Known locations of 7 α-helices in the structure of this group shown in boxes

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

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

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.

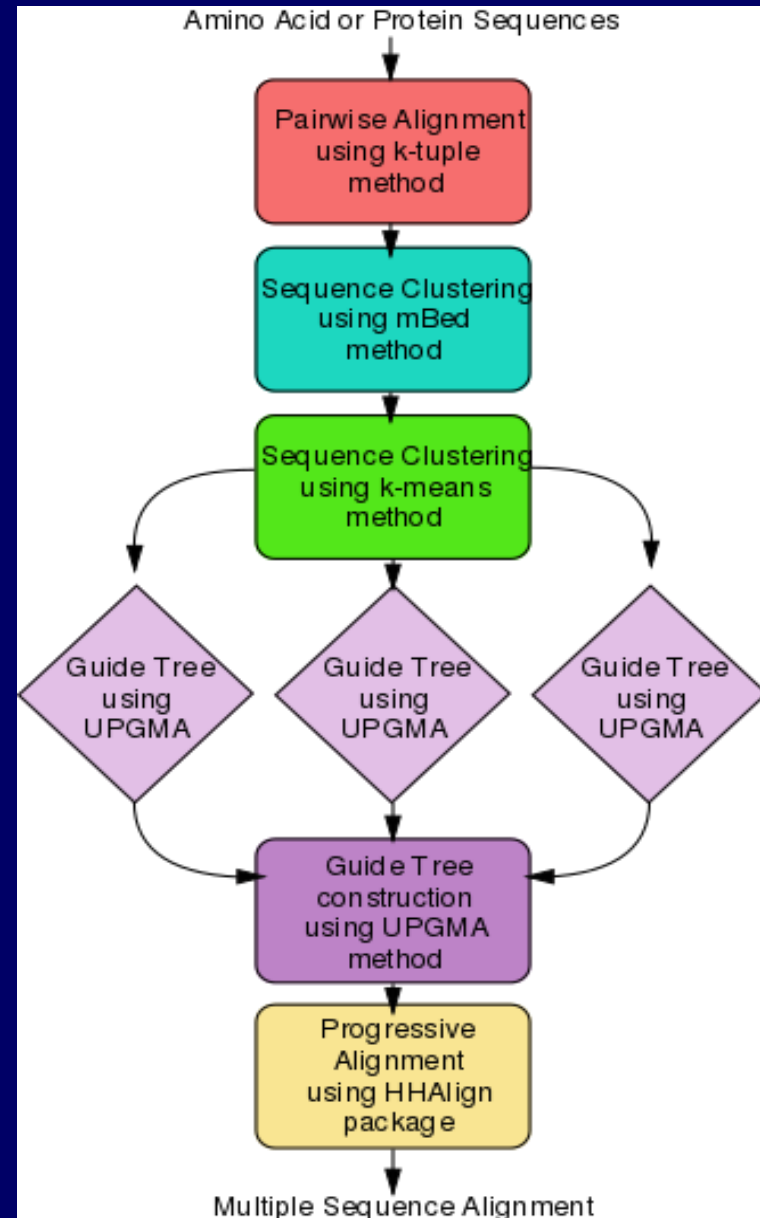

Not suitable for sequence assembly!

ClustalW / ClustalΩ online

- **EBI:**
  https://www.ebi.ac.uk/Tools/msa/

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

- one of the fastest multiple sequence alignment tools

- **DDBJ:**
  http://clustalw.ddbj.nig.ac.jp/

Iterative Methods of MSA

- To correct for errors introduced by initial alignment, use <u>iterative</u> 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

Other Methods

Genetic Algorithm: **It is a machine learning approach, implemented in the program SAGA, and produces alignments by attempted simulation of the evolutionary changes in sequences.**

Basic idea behind this method is to try to generate many different MSAs by rearrangements that simulate gap insertion & recombination events during replication in order to generate a higher and higher score for the MSA.

Hidden Markov Models (HMM): **HMMs are trained to recognize a specific family of proteins; require different HMM for each protein family**

Examine alignments by eye

Some artifacts observed in the output of algorithms:
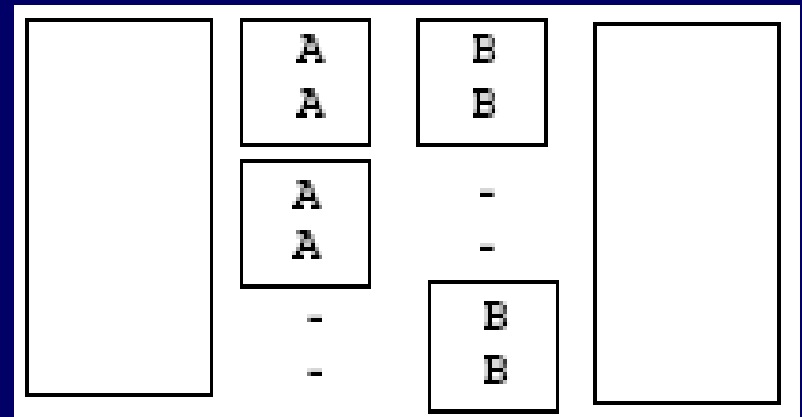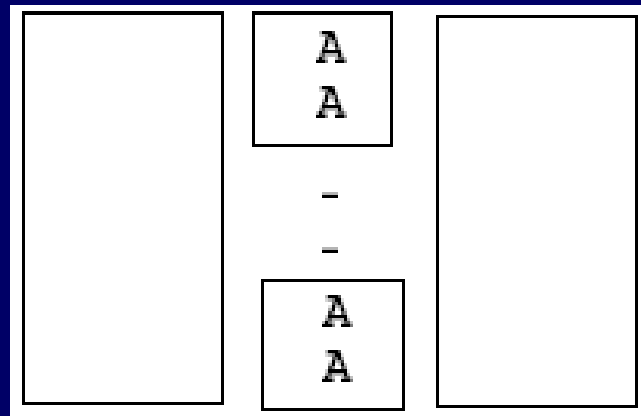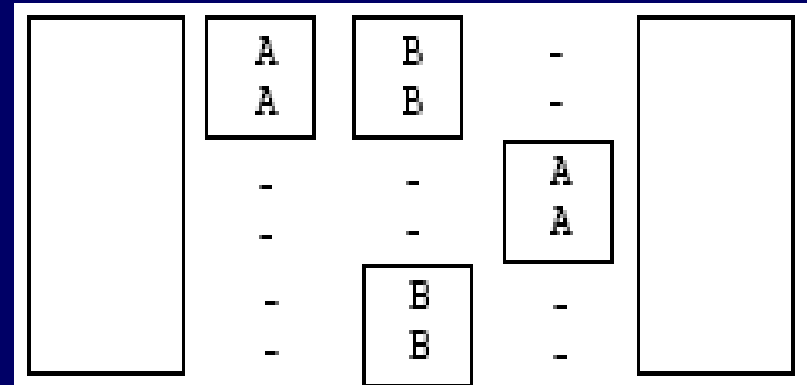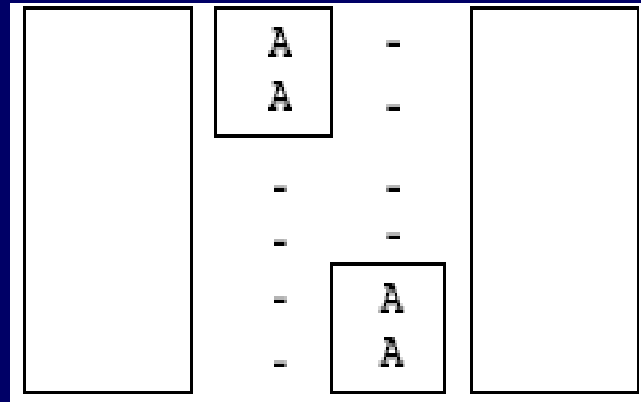
Always examine your alignment manually to see if it can be improved.

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).

- Carrillo, H. and Lipman, D. "The multiple sequence alignment problem in biology" *SIAM J. Appl. Math.* 48, 1073-82 (1988).