BCB 5200 Introduction to Bioinformatics

Multiple Sequence Alignment

Bioinformatics and Computational Biology
Saint Louis University

Outline

- Multiple sequence alignment (MSA)

- Scoring MSA

- MSA algorithms
  - Exact approach - Dynamic programming
  - Progressive alignments - ClustalW
  - Iterative approach -MUSCLE
  - Profile-based approach - Promals

- Warning

- One of the most important contributions of biological sequences to biology and evolution is the discovery that sequences of different organisms are often related.

- They are called homologies

Multiple sequence alignment: definition

- A collection of three or more protein (or nucleic acid) sequences that are partially or completely aligned

- A model

- Indicates relationship between residues of different sequences (similarity/dissimilarity)

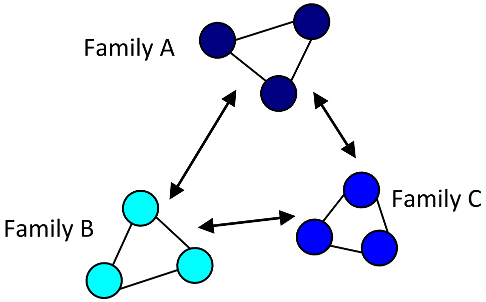
Multiple sequence alignment

```
BSUB00  RMAHYDSLTDLPNRRHAISHLTKVLNREHSLHYNTVVFFLDLNRFKVINDAL
ECU738  VMSTRDGMTGVYNRRHWETMLRNEFDNCRRHNRDATLLIIDIDHFKSINDTW
D90790  HEVGMDVLTKLLNRRFLPTIFKREIAHANRTGTPLSVLIIDVDKFKEINDTW
SYCSLL  QISSLDALTQVGNRYLFDSTLEREWQRLQRIREPLALLLCDVDFFKGFNDNY
ECAE00  NIAHRDPLTNIFNRNYFFNEL--TVQSASAQKTPYCVMIMDIDHFKKVNDTW
AF0348  QAANVDSLTGLANRAAYNAHM-ERLTAADAPS--IGLLLIDVDRLKQVNDIL
D90796  IRSNMDVLTGLPGRRVLDESFDHQLRNAEPLN--LYLMLLDIDRFKLVNDTY
Y4LL_R  HMARHDALTGLPNRQFLREEF-ERLSDHIAPSTRLAILCLDLDGFKAINDAY
Y07I_M  YLADHDDLTGLHNRRALLQHLDQRLAPGQPGP--VAALFLDLDRLKAINDYL
......
```

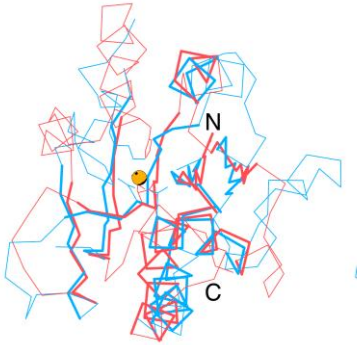
Multiple sequence alignment

```
BSUB00  RMAHYDSLTDLPNRRHAISHLTKVLNREHSLHYNTVVFFLDLNRFKVINDAL
ECU738  VMSTRDGMTGVYNRRHWETMLRNEFDNCRRHNRDATLLIIDIDHFKSINDTW
D90790  HEVGMDVLTKLLNRRFLPTIFKREIAHANRTGTPLSVLIIDVDKFKEINDTW
SYCSLL  QISSLDALTQVGNRYLFDSTLEREWQRLQRIREPLALLLCDVDFFKGFNDNY
ECAE00  NIAHRDPLTNIFNRNYFFNEL--TVQSASAQKTPYCVMIMDIDHFKKVNDTW
AF0348  QAANVDSLTGLANRAAYNAHM-ERLTAADAPS--IGLLLIDVDRLKQVNDIL
D90796  IRSNMDVLTGLPGRRVLDESFDHQLRNAEPLN--LYLMLLDIDRFKLVNDTY
Y4LL_R  HMARHDALTGLPNRQFLREEF-ERLSDHIAPSTRLAILCLDLDGFKAINDAY
Y07I_M  YLADHDDLTGLHNRRALLQHLDQRLAPGQPGP--VAALFLDLDRLKAINDYL
......
```

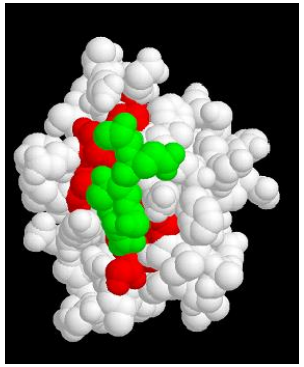
Protein similarity search and family identification/classification (conserved motif)



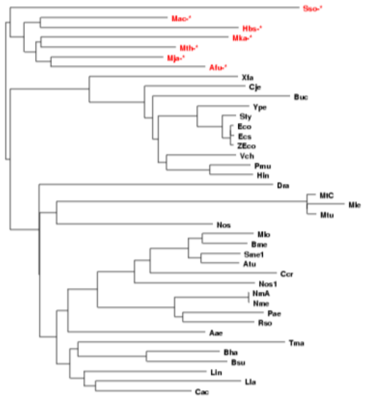
Structure modeling (Protein,RNA)



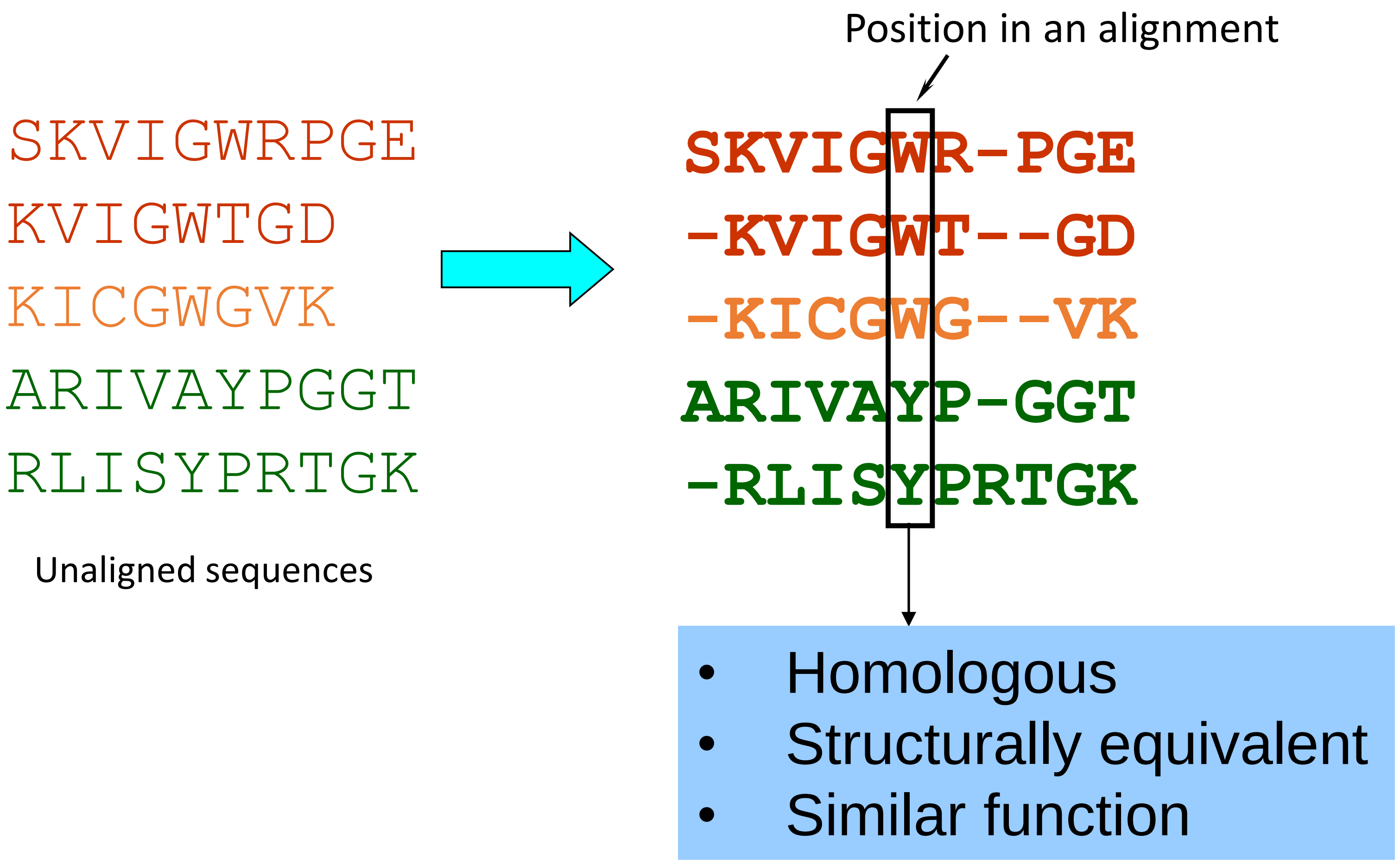
Active site prediction experimental design



Phylogenetic analysis



Meaning of alignments



```
SKVIGWRPGE
KVIGWTGD
KICGWGVK
ARIVAYPGGT
RLISYPRTGK
```

Unaligned sequences

Position in an alignment

```
SKVIGWR-PGE
-KVIGWT--GD
-KICGWG--VK
ARIVAYP-GGT
-RLISYPRTGK
```

- Homologous
- Structurally equivalent
- Similar function

Outline

- Multiple sequence alignment (MSA)

- Scoring MSA

- MSA algorithms
    - Exact approach - Dynamic programming
    - Progressive alignments - ClustalW
    - Iterative approach -MUSCLE
    - Profile-based approach - Promals

- Warning

Scoring a multiple sequence alignment

```
A T _ G C G A    A T _ G C G A    A T _ G C G A
A _ C G T _ A    A C _ G T _ A    A C _ G T A _
A T C A C _ A    A T C A C _ A    A T C A C A _
```

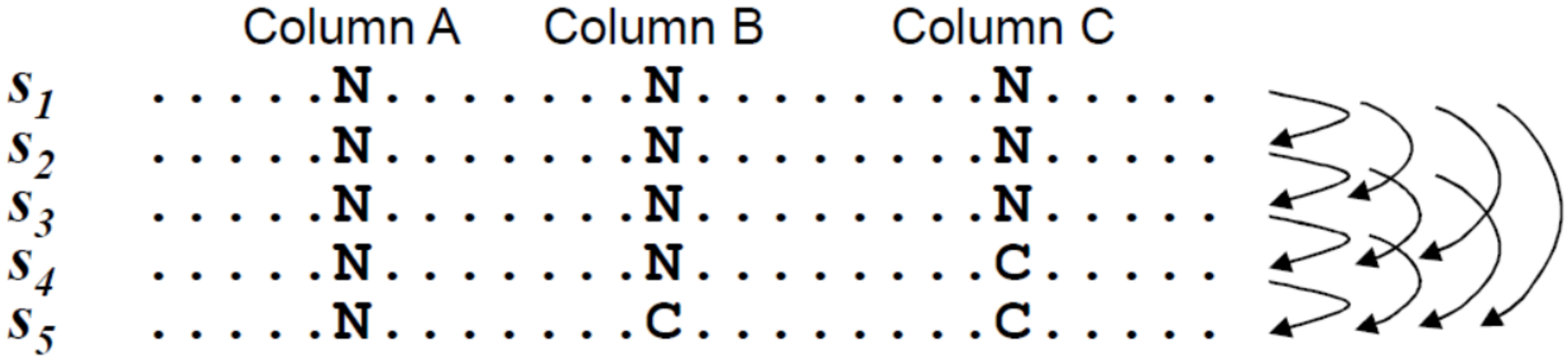
- Score: more conserved columns, better alignment

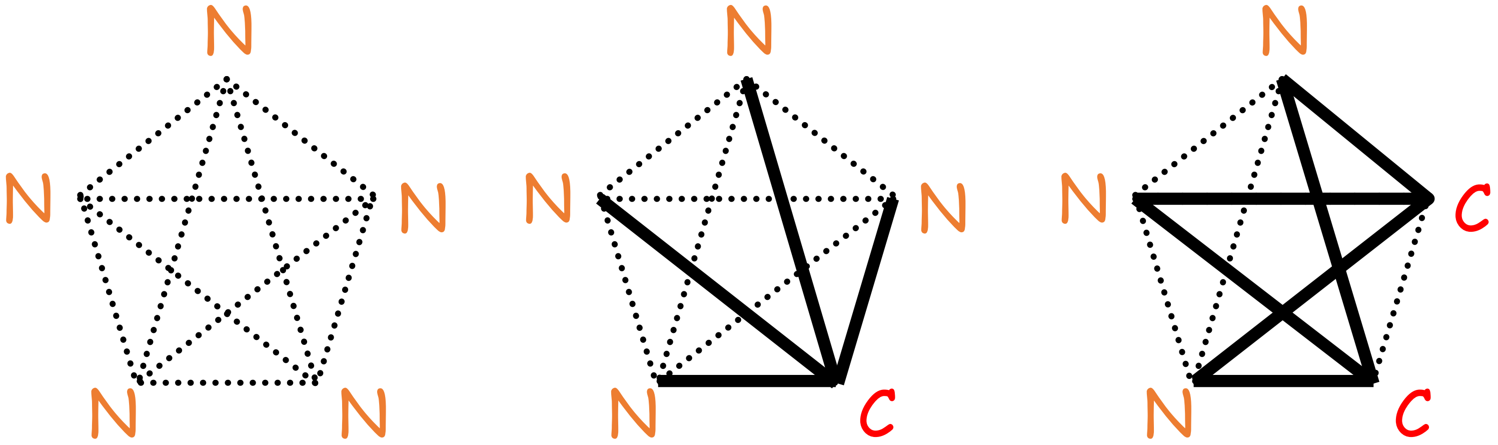
- To find alignment that maximizes a score function

Sum-of-pairs (SP) scoring

- Standard MSA scoring method
- Assumes column independence
- SP is a column-by-column cost/weight function
- SP scored using a **substitution matrix** (e.g PAM or BLOSUM)
- MSAs *maximize* total alignment score by *maximizing* each column SP score

Scoring a multiple sequence alignment: sum-of-pairs (SP) scoring


Column A
Column B
Column C
s_1 . . . . . .N. . . . . .N. . . . . .N. . . . . .
s_2 . . . . . .N. . . . . .N. . . . . .N. . . . . .
s_3 . . . . . .N. . . . . .N. . . . . .N. . . . . .
s_4 . . . . . .N. . . . . .N. . . . . .C. . . . . .
s_5 . . . . . .N. . . . . .C. . . . . .C. . . . . .



		Column		
Alignmt	Score	A	B	C
N – N	6	10	6	3
N – C	-3	0	4	6
C – C	9	0	0	1
		60	24	9

History of MSA

1975 Sankoff

Formulated multiple alignment problem and gave DP solution

1988 Carrillo-Lipman

Branch and Bound approach for MSA

1990 Feng-Doolittle

Progressive alignment

1994 Thompson-Higgins-Gibson-ClustalW

Most popular multiple alignment program

1998 DIALIGN (*Segment-based multiple alignment)*

2000 T-coffee *(consensus-based)*

2004 MUSCLE

2005 Kalign

2005 ProbCons (uses Bayesian consistency)

2006 M-Coffee (consensus meta-approach)

2006 Expresso (3D-Coffee; use structural template)

2007 PROMALS (profile-profile alignment)

2009 FastTree

Outline

- Multiple sequence alignment (MSA)

- Scoring MSA

- MSA algorithms
  - Exact approach - Dynamic programming
  - Progressive alignments - ClustalW
  - Iterative approach -MUSCLE
  - Profile-based approach - Promals

- Recommendations

Exact approach

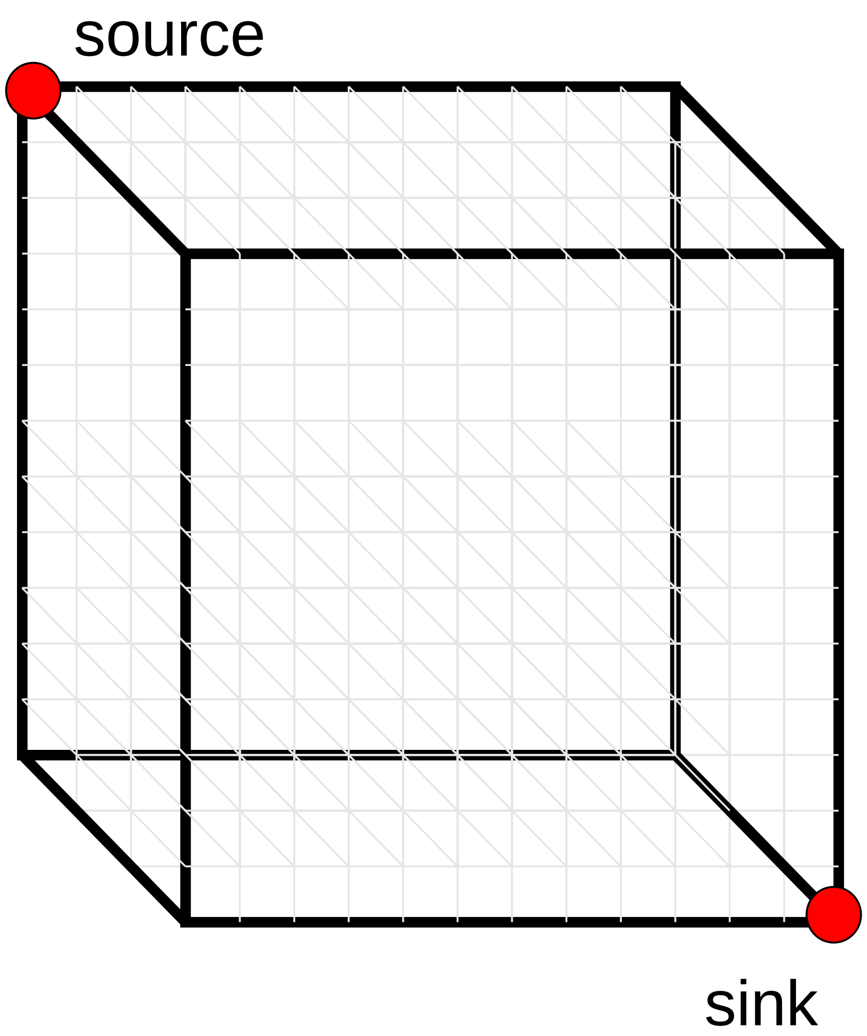
- Exact methods of multiple alignment use dynamic programming (Generalization of Needleman-Wunsch)
- Guaranteed to find optimal solutions
- Computationally expensive and so impractical
    - Time grows as product of sequence lengths

From pairwise to multiple alignment

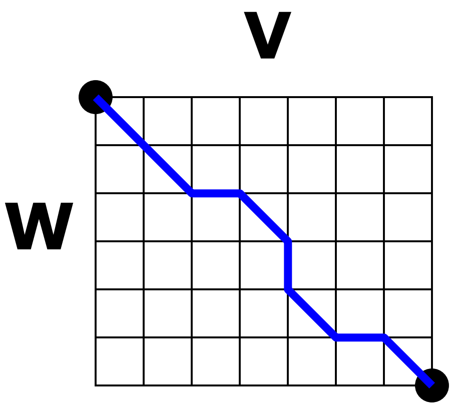
- Alignment of 2 sequences is represented as a 2-row matrix

- In a similar way, we represent alignment of 3 sequences as a 3-row matrix (or a 3-D “Manhattan Cube”, with each axis representing a sequence to align)

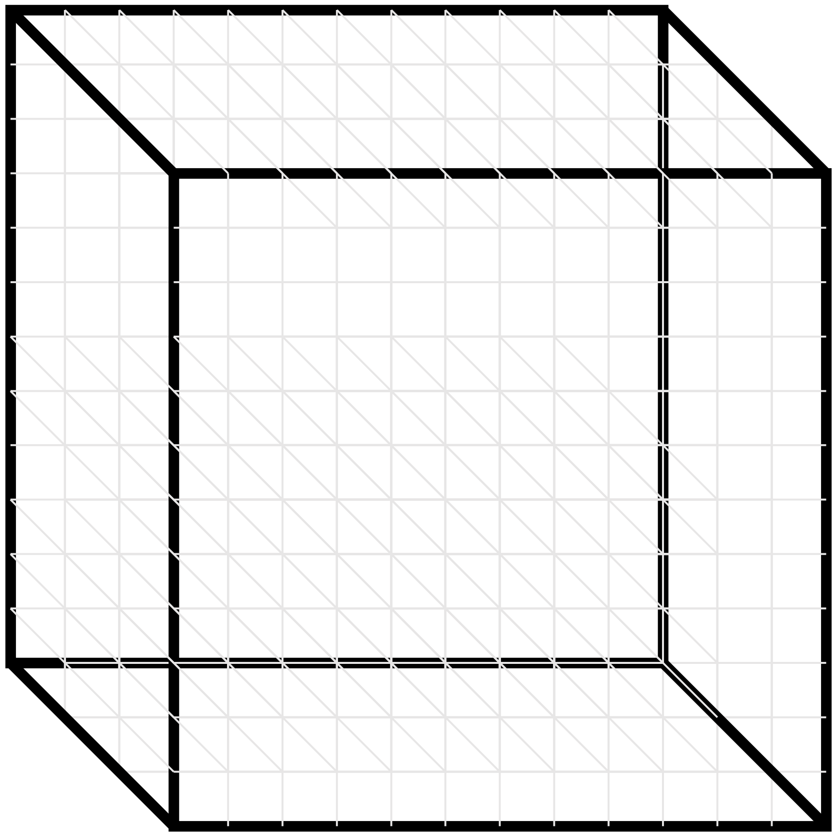
- For global alignments, go from source to sink



2D vs 3D alignment cells



2D table



3D graph

0 1 2 2 3 4 5 6

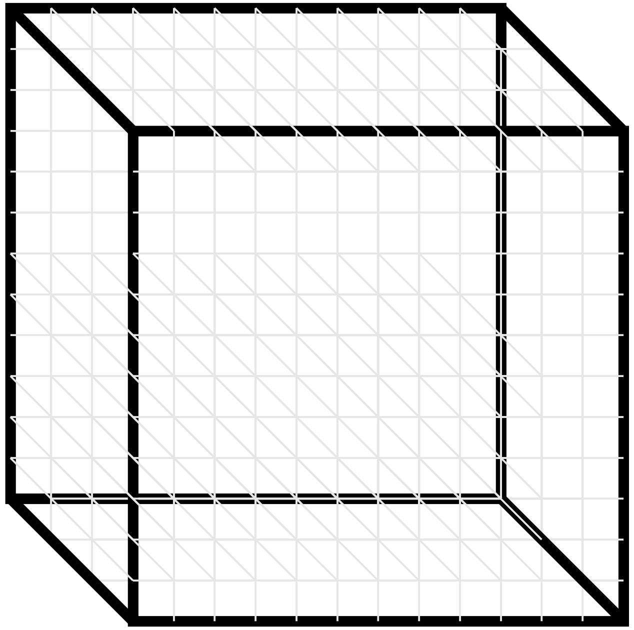
A T _ G C G A

0 1 1 2 3 4 4 5

A _ C G T _ A

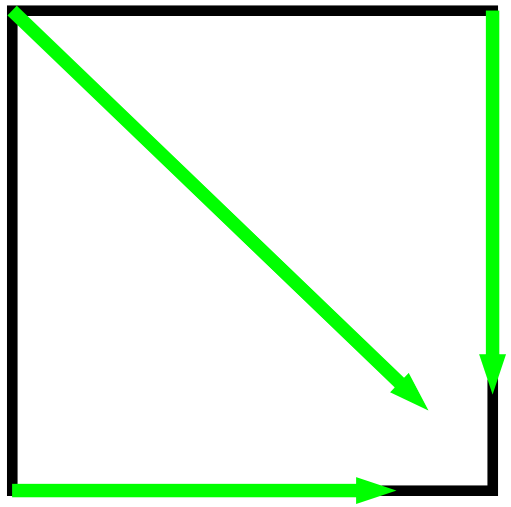
0 1 2 3 4 5 5 6

A T C A C _ A

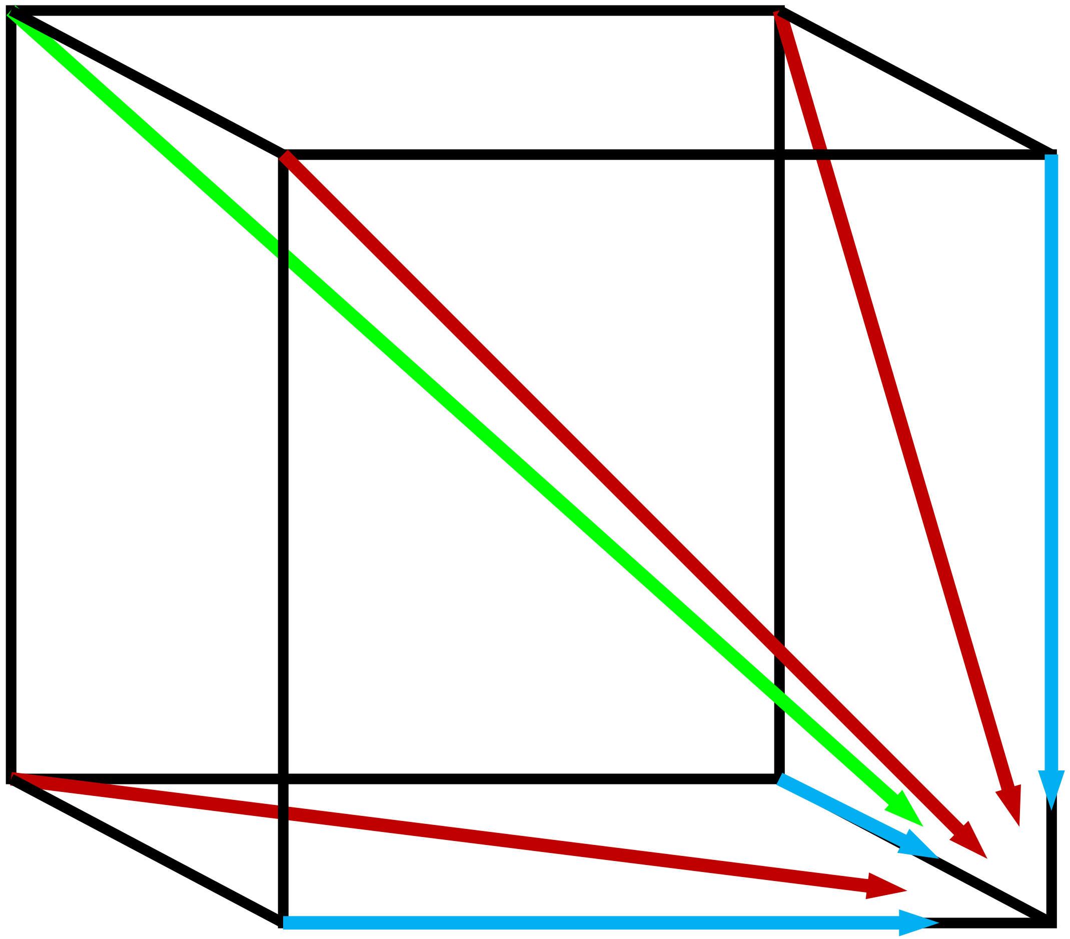


(0,0,0)->(1,1,1)-> (2,1,2)->(2,2,3)->(3,3,4)->(4,4,5)->(5,4,5)->(6,5,6)

2D vs 3D alignment cell: 3 paths vs 7 paths



Pairwise: 3 possible paths (match/mismatch, insertion, and deletion)



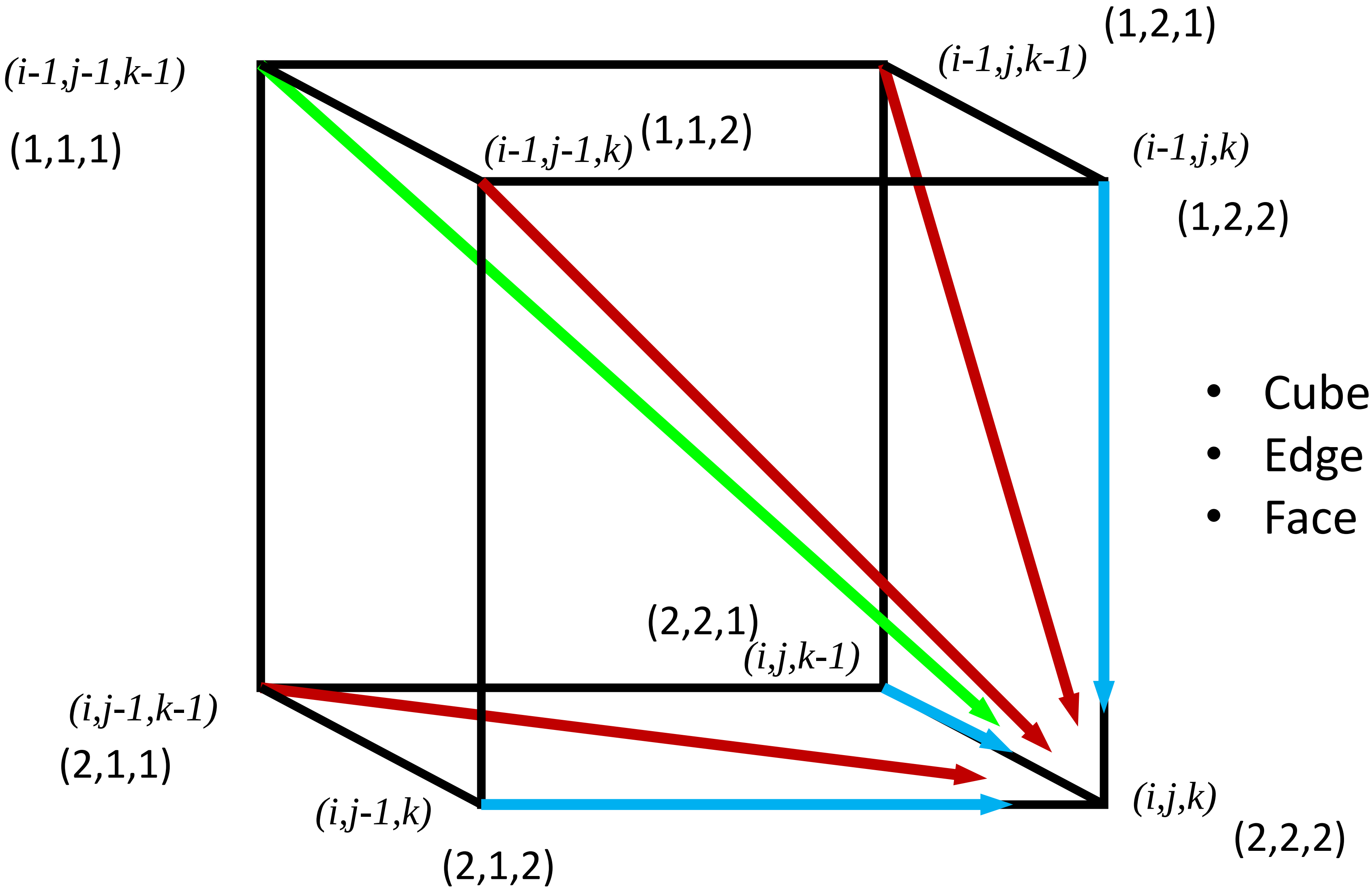
In **3-D**, 7 edges in each unit cube

There are seven cases when aligning three sequences

1	2	3	4	5	6	7
I	**I**	**I**	**-**	**I**	**-**	**-**
J	**J**	**-**	**J**	**-**	**J**	**-**
K	**-**	**K**	**K**	**-**	**-**	**K**

2^3 -1 to choose the maximum similarity

Architecture of 3D alignment cell



- Cube diagonal line (no gaps)
- Edge diagonal line (one gap)
- Face diagonal line (two gaps)

Multiple alignment: dynamic programming

- $$s_{i,j,k} = \max \left\{ \begin{array}{l} \left. s_{i-1,j-1,k-1} + \delta(v_i, w_j, u_k) \right\} \text{cube diagonal: no gaps} \\ \left. \begin{array}{l} s_{i-1,j-1,k} + \delta(v_i, w_j, \_) \\ s_{i-1,j,k-1} + \delta(v_i, \_, u_k) \\ s_{i,j-1,k-1} + \delta(\_, w_j, u_k) \end{array} \right\} \text{face diagonal: one gap} \\ \left. \begin{array}{l} s_{i-1,j,k} + \delta(v_i, \_, \_) \\ s_{i,j-1,k} + \delta(\_, w_j, \_) \\ s_{i,j,k-1} + \delta(\_, \_, u_k) \end{array} \right\} \text{edge diagonal: two gaps} \end{array} \right.$$

- $\delta(x, y, z)$ is an entry in the 3D scoring matrix

MSA: running time

- For 3 sequences of length ***n***, operation time is $7\boldsymbol{n}^3$; O($\boldsymbol{n}^3$)
- For ***k*** sequences, build a ***k***-dimensional Manhattan, with operation time (2^k-1)($\boldsymbol{n}^k$); O($2^k\boldsymbol{n}^k$)
  - 32 thousand years for 10 seqs of 100 residues!
- Conclusion: although dynamic programming approach for alignment between two sequences is easily extended to ***k*** sequences (simultaneous approach), it is impractical due to exponential running time.
- Heuristic sequence alignment algorithm is needed, which doesn't guarantee to find the optimal solution

Outline

- Multiple sequence alignment (MSA)

- Scoring MSA

- MSA algorithms
  - Exact approach - Dynamic programming
  - Progressive alignments - ClustalW
  - Iterative approach -MUSCLE
  - Profile-based approach - Promals

- Recommendations

Progressive alignment

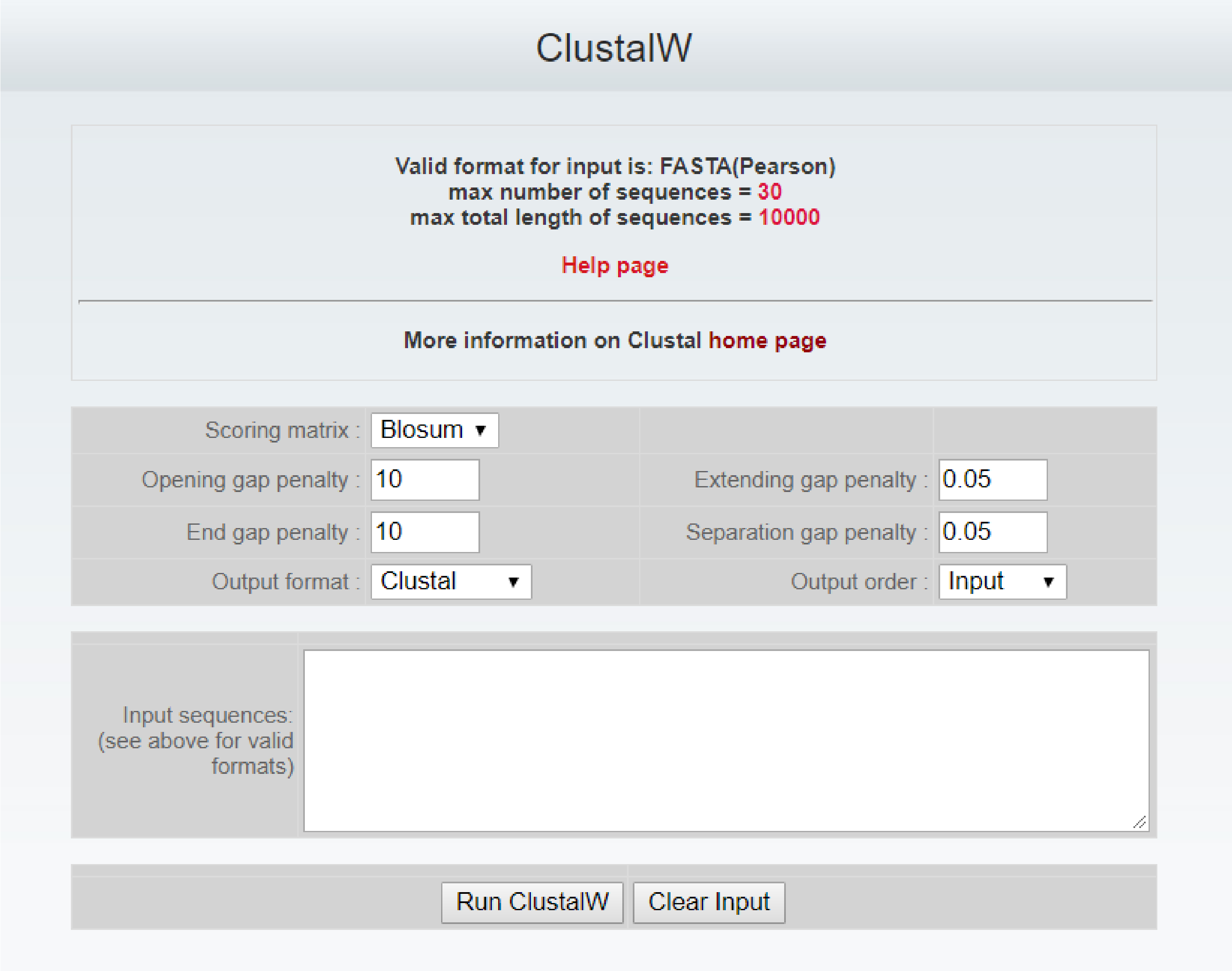
- Feng & Doolittle 1987, Higgins and Sharp 1988

- Concept: to build the alignment of larger number of sequences from partial alignments of subsets of sequences

- A guide tree (related to a phylogenetic tree) is used to determine how to combine pairwise alignments one by one to create a multiple alignment.

- Examples: ClustalW

ClustalW – the most widely used program



ClustalW

Valid format for input is: FASTA(Pearson)
max number of sequences = 30
max total length of sequences = 10000

Help page

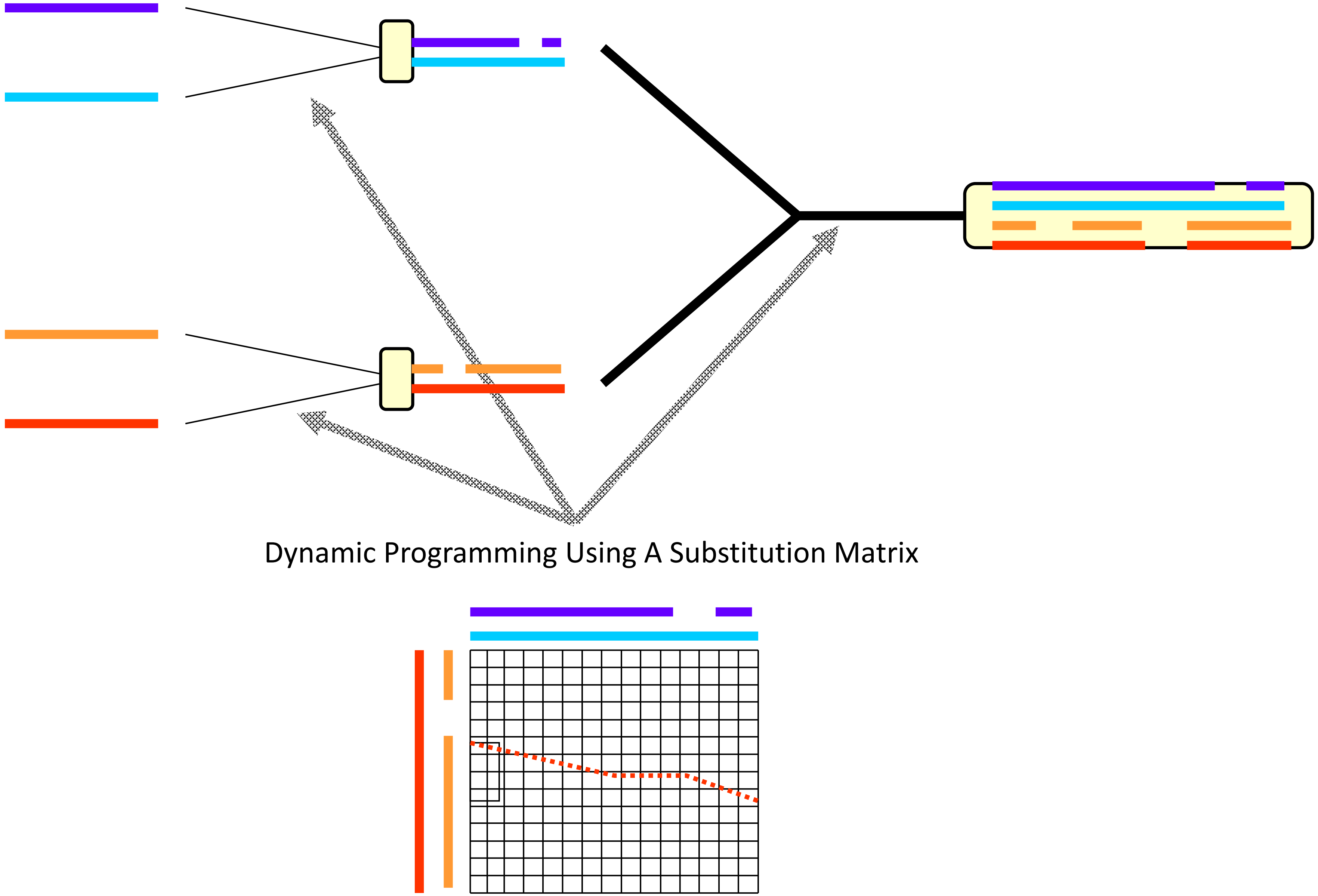
More information on Clustal home page

Scoring matrix :	Blosum ▾		
Opening gap penalty :	10	Extending gap penalty :	0.05
End gap penalty :	10	Separation gap penalty :	0.05
Output format :	Clustal ▾	Output order :	Input ▾

Input sequences:
(see above for valid formats)

Thompson et al. (1994). http://www.ch.embnet.org/software/ClustalW.html

ClustalW algorithm



Dynamic Programming Using A Substitution Matrix

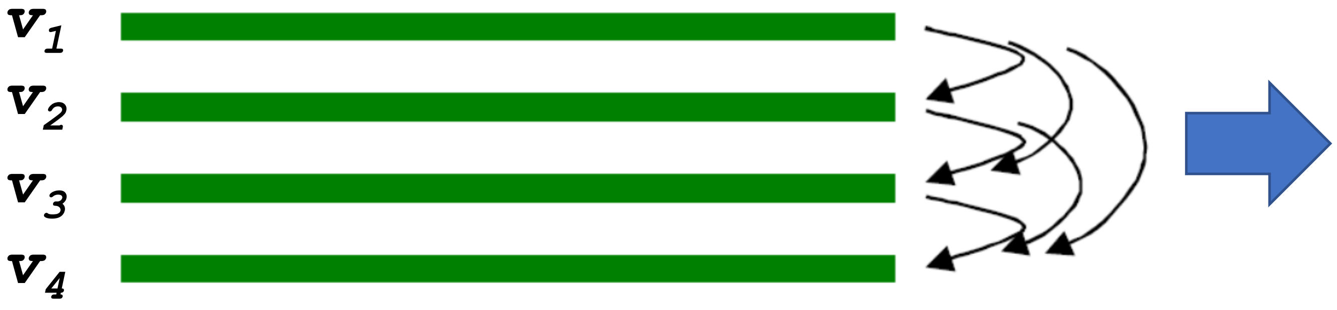
ClustalW

The three basic steps in the CLUSTAL W approach are shared by all progressive alignment algorithms:

A. Calculate a matrix of pairwise distances based on pairwise alignments between the sequences
B. Use the result of A to build a guide tree, which is an inferred phylogeny for the sequences
C. Use the tree from B to guide the progressive alignment of the sequences

Step 1: Pairwise alignment

- Aligns each sequence against each other using dynamic programming
- a similarity or distance measure for the pair is calculated using the aligned portion (gaps excluded) - for example, percent identity.
- Similarity = exact matches / sequence length (percent identity)

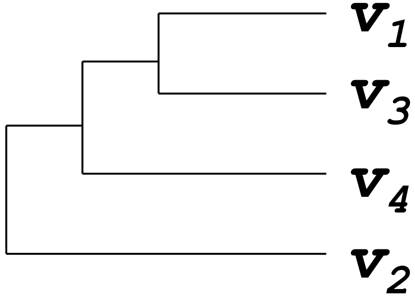


	v_1	v_2	v_3	v_4
v_1	–			
v_2	.17	–		
v_3	.87	.28	–	
v_4	.59	.33	.62	–

(.17 means 17 % identical)

Step 2: Guide tree by clustering

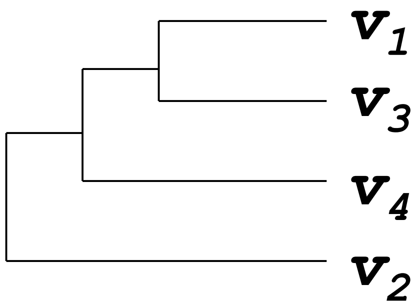
	v_1	v_2	v_3	v_4
v_1	–			
v_2	.17	–		
v_3	.87	.28	–	
v_4	.59	.33	.62	–



- To build guide tree
  - Neighbour-Joining (NJ)
  - Unweighted pair group method using arithmetic averages (UPGMA)

- Guide tree roughly reflects evolutionary relations

Step 3: progressive alignment of the sequences



Calculate:

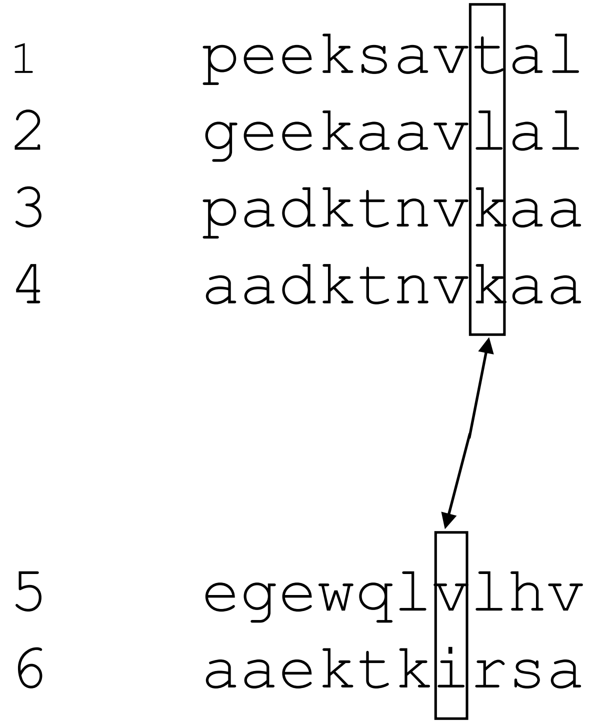
$v_{1,3}$ = alignment (v_1, v_3)

$v_{1,3,4}$ = alignment$((v_{1,3}),v_4)$

$v_{1,2,3,4}$ = alignment$((v_{1,3,4}),v_2)$

- Partial alignment was generated (profile)
- In the past we were aligning a **sequence against a sequence**
- Can we align a **sequence against a profile?**

Scoring an alignment of two partial alignments



```
1     peeksavtal
2     geekaavlal
3     padktnvkaa
4     aadktnvkaa

5     egewqlvlhv
6     aaektkirsa
```

Sequence weights

$w_1, \ldots, w_6$

'W' stands for 'weighted' (sequences are weighted differently).

Score: $$\frac{1}{8}\left[M(t,v)w_1w_5 + M(t,i)w_1w_6 + \ldots + M(k,i)w_4w_6\right]$$

Potential problems with ClustalW

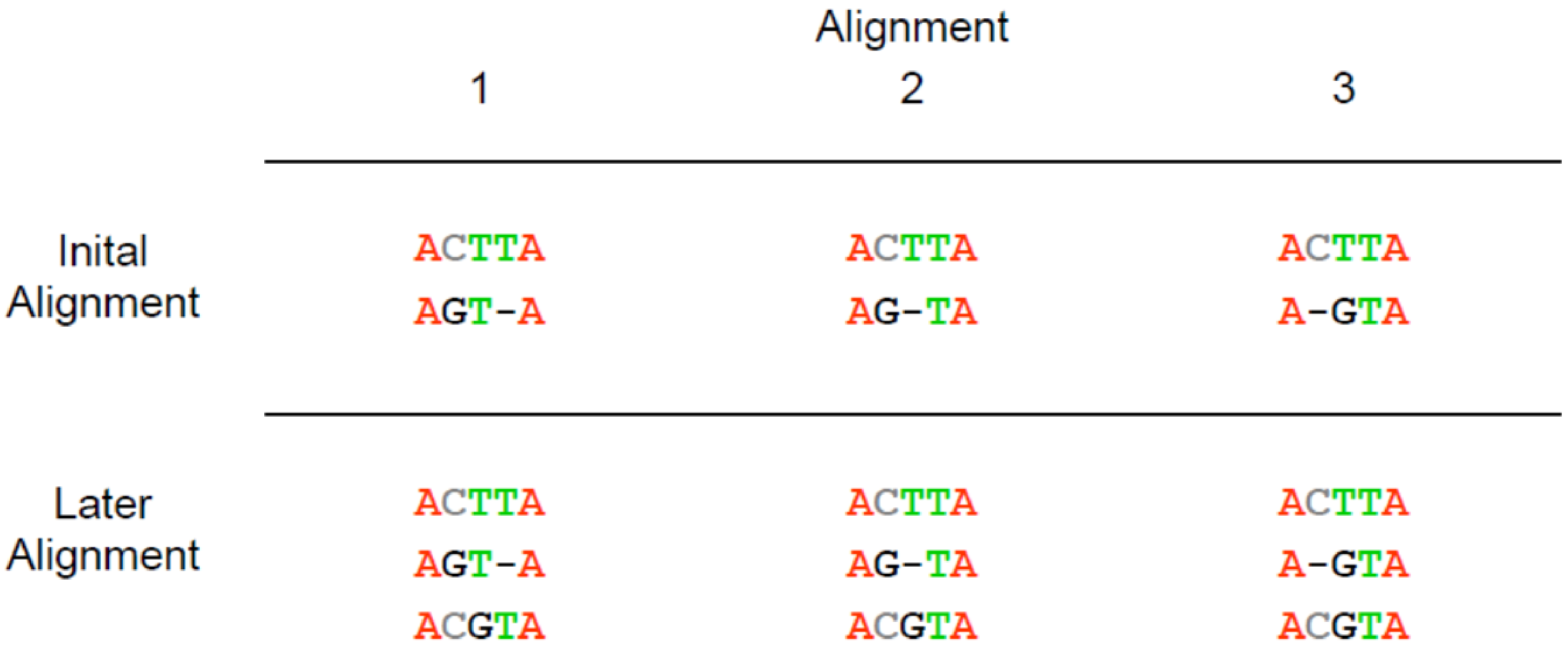
- ClustalW is a "greedy" algorithm
  - makes the best immediate solution (local choice) in hopes of finding the best overall (global) solution
  - choices are made regardless of later consequences
  - early mistakes get propagated throughout the rest of the alignment

	Alignment		
	1	2	3
Inital Alignment	ACTTA AGT-A	ACTTA AG-TA	ACTTA A-GTA

new sequence ACGTA

Potential problems with ClustalW

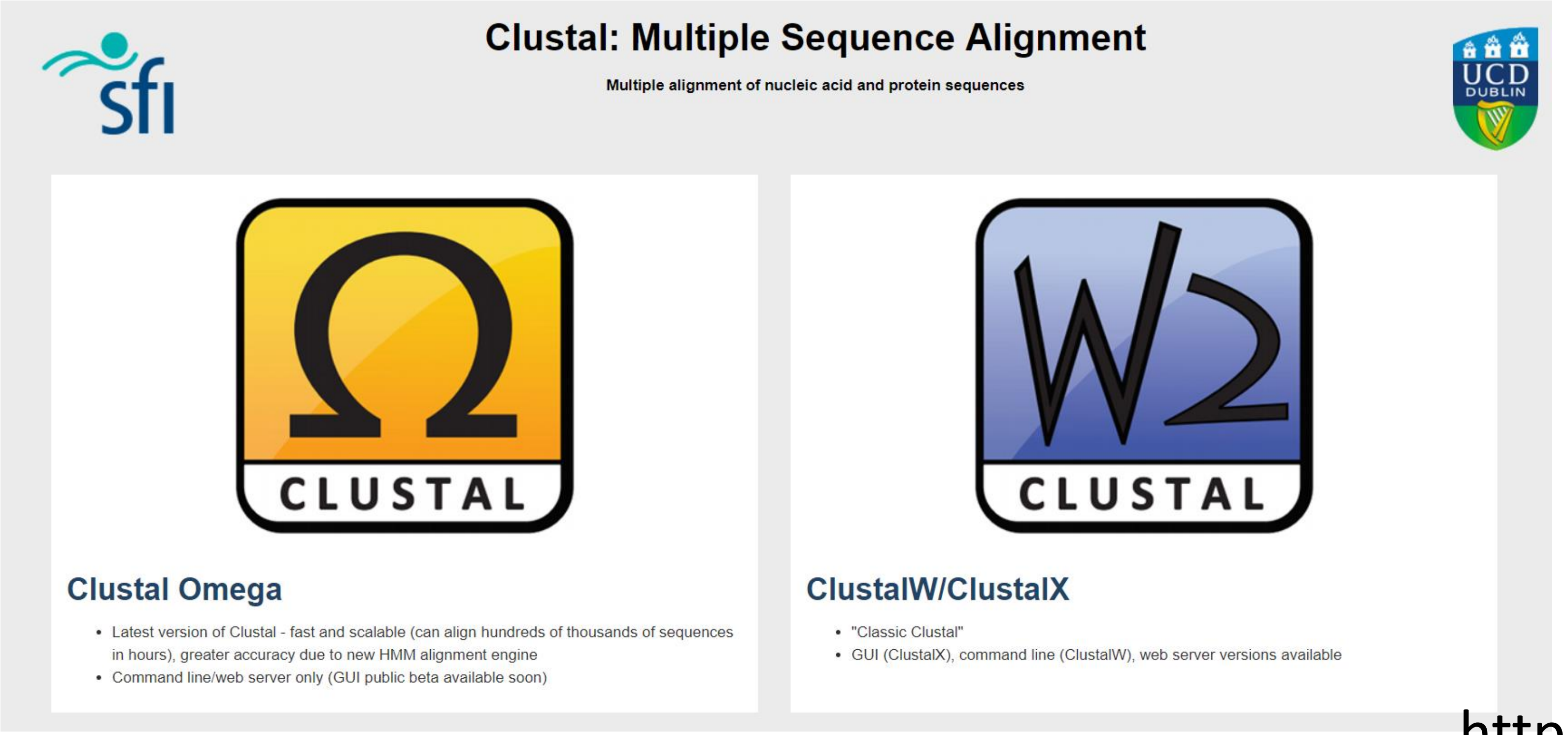
- ClustalW is a “greedy” algorithm
  - makes the best immediate solution (local choice) in hopes of finding the best overall (global) solution
  - choices are made regardless of later consequences
  - early mistakes get propagated throughout the rest of the alignment



		Alignment	
	1	2	3
Inital Alignment	ACTTA AGT-A	ACTTA AG-TA	ACTTA A-GTA
Later Alignment	ACTTA AGT-A ACGTA	ACTTA AG-TA ACGTA	ACTTA A-GTA ACGTA

Clustal Omega

- Profile HMMs to model groups of sequences whereas Clustal W uses sequence profiles to store information about groups of sequences



Clustal: Multiple Sequence Alignment

Multiple alignment of nucleic acid and protein sequences

Clustal Omega

- Latest version of Clustal - fast and scalable (can align hundreds of thousands of sequences in hours), greater accuracy due to new HMM alignment engine
- Command line/web server only (GUI public beta available soon)

ClustalW/ClustalX

- "Classic Clustal"
- GUI (ClustalX), command line (ClustalW), web server versions available

http://www.clustal.org/

Iterative alignment

- Progressive alignment:
  - The order of selection of sequences can influence the alignment
  - Once there is a gap, always a gap

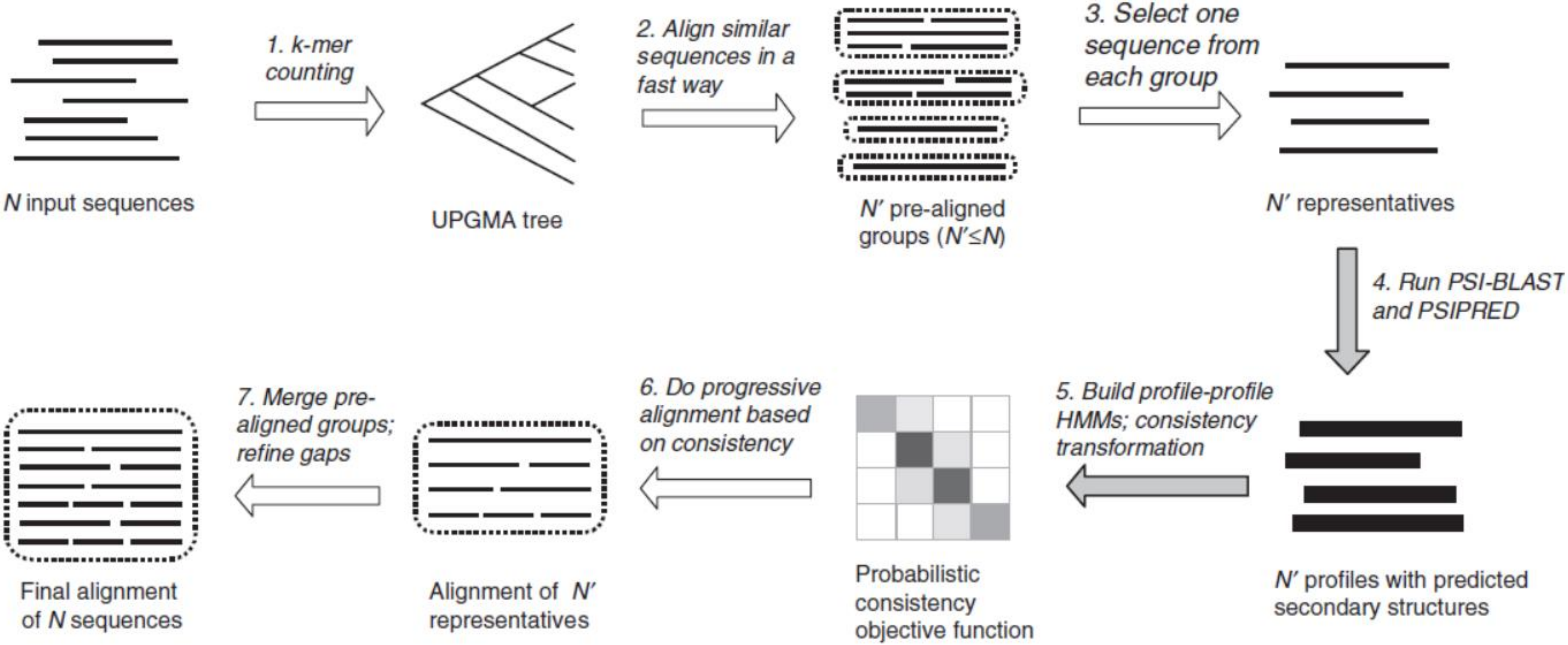
- How to avoid committing to a non-optimal pairwise decision?
  - Revisit alignments
  - This is the focus of iterative alignments

Iterative alignment

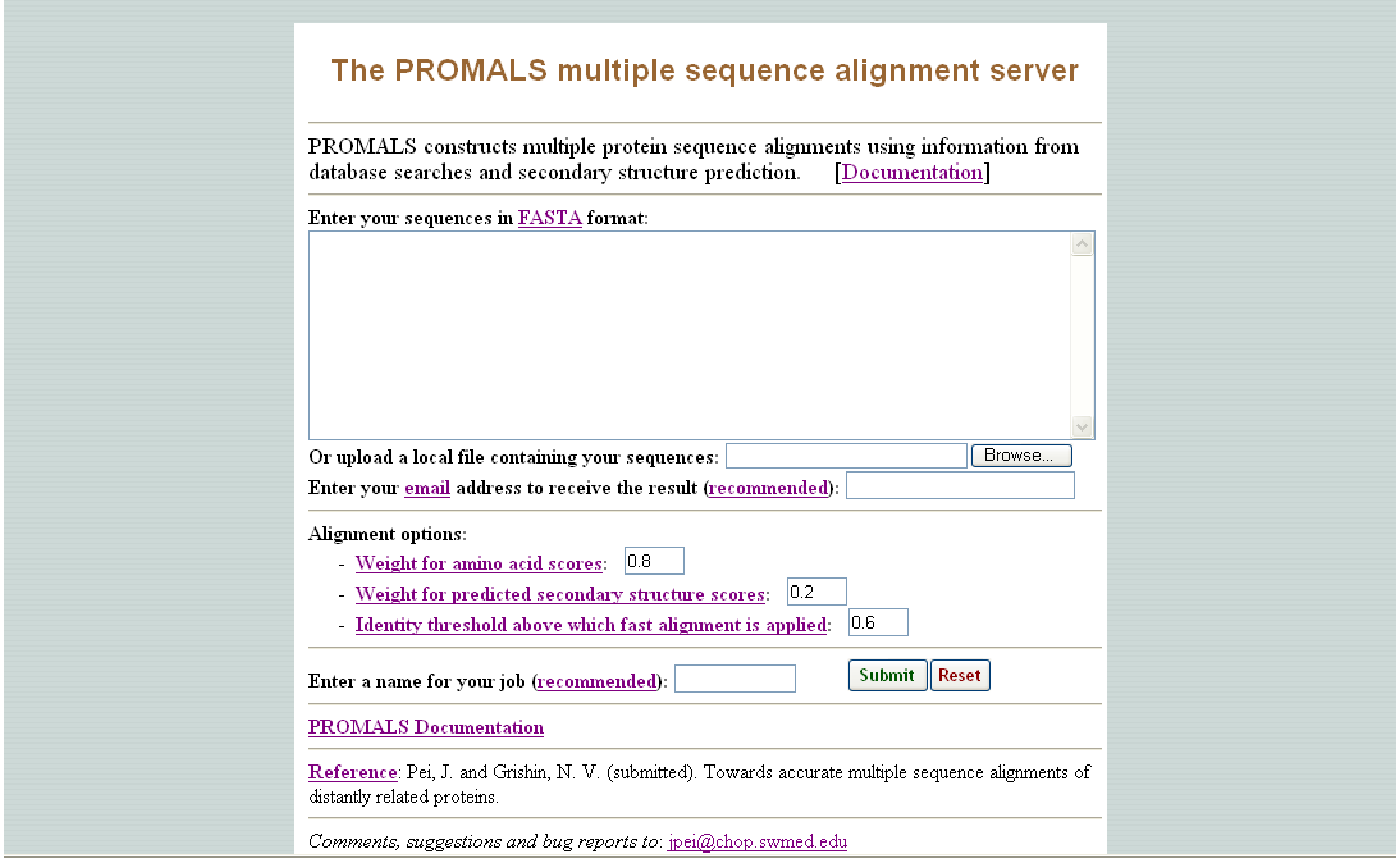
- Basic iterative refinement algorithm
    - Remove a sequence from the current multiple alignment
    - Realign the removed sequence back to the multiple alignment
    - Repeat until removal and realignment of any sequence does not improve the alignment score

- MUSCLE (multiple sequence alignment by log-expectation)

Profile-HMM method: Promals


N input sequences
1. k-mer counting
UPGMA tree
2. Align similar sequences in a fast way
N′ pre-aligned groups (N′≤N)
3. Select one sequence from each group
N′ representatives
4. Run PSI-BLAST and PSIPRED
N′ profiles with predicted secondary structures
5. Build profile-profile HMMs; consistency transformation
Probabilistic consistency objective function
6. Do progressive alignment based on consistency
Alignment of N′ representatives
7. Merge pre-aligned groups; refine gaps
Final alignment of N sequences

http://prodata.swmed.edu/promals



The PROMALS multiple sequence alignment server

PROMALS constructs multiple protein sequence alignments using information from database searches and secondary structure prediction. [Documentation]

Enter your sequences in FASTA format:

Or upload a local file containing your sequences:

Enter your email address to receive the result (recommended):

Alignment options:

- Weight for amino acid scores: 0.8
- Weight for predicted secondary structure scores: 0.2
- Identity threshold above which fast alignment is applied: 0.6

Enter a name for your job (recommended):

PROMALS Documentation

Reference: Pei, J. and Grishin, N. V. (submitted). Towards accurate multiple sequence alignments of distantly related proteins.

Comments, suggestions and bug reports to: jpei@chop.swmed.edu

History of MSA

1975 Sankoff

Formulated multiple alignment problem and gave DP solution

1988 Carrillo-Lipman

Branch and Bound approach for MSA

1990 Feng-Doolittle

Progressive alignment

1994 Thompson-Higgins-Gibson-ClustalW

Most popular multiple alignment program

1998 DIALIGN (*Segment-based multiple alignment)*

2000 T-coffee *(consensus-based)*

2004 MUSCLE

Acceptable result

2005 Kalign

Fast and acceptable result, gappy

2005 ProbCons (uses Bayesian consistency)

2006 M-Coffee (consensus meta-approach)

2006 Expresso (3D-Coffee; use structural template)

2007 PROMALS (profile-profile alignment)

Slow but most accurate

2009 FastTree

Fast and working with large datasets; ok

Recommendations

- Many dozens of MSA programs have been introduced in recent years. None is optimal. Each offers unique strengths and weaknesses.

- MSA algorithms assume that sequences are homologous
  - MSA programs will align anything and all sequences, even if they are not homologous.

- Ideally sequences with one domain or sequences with same domain architecture

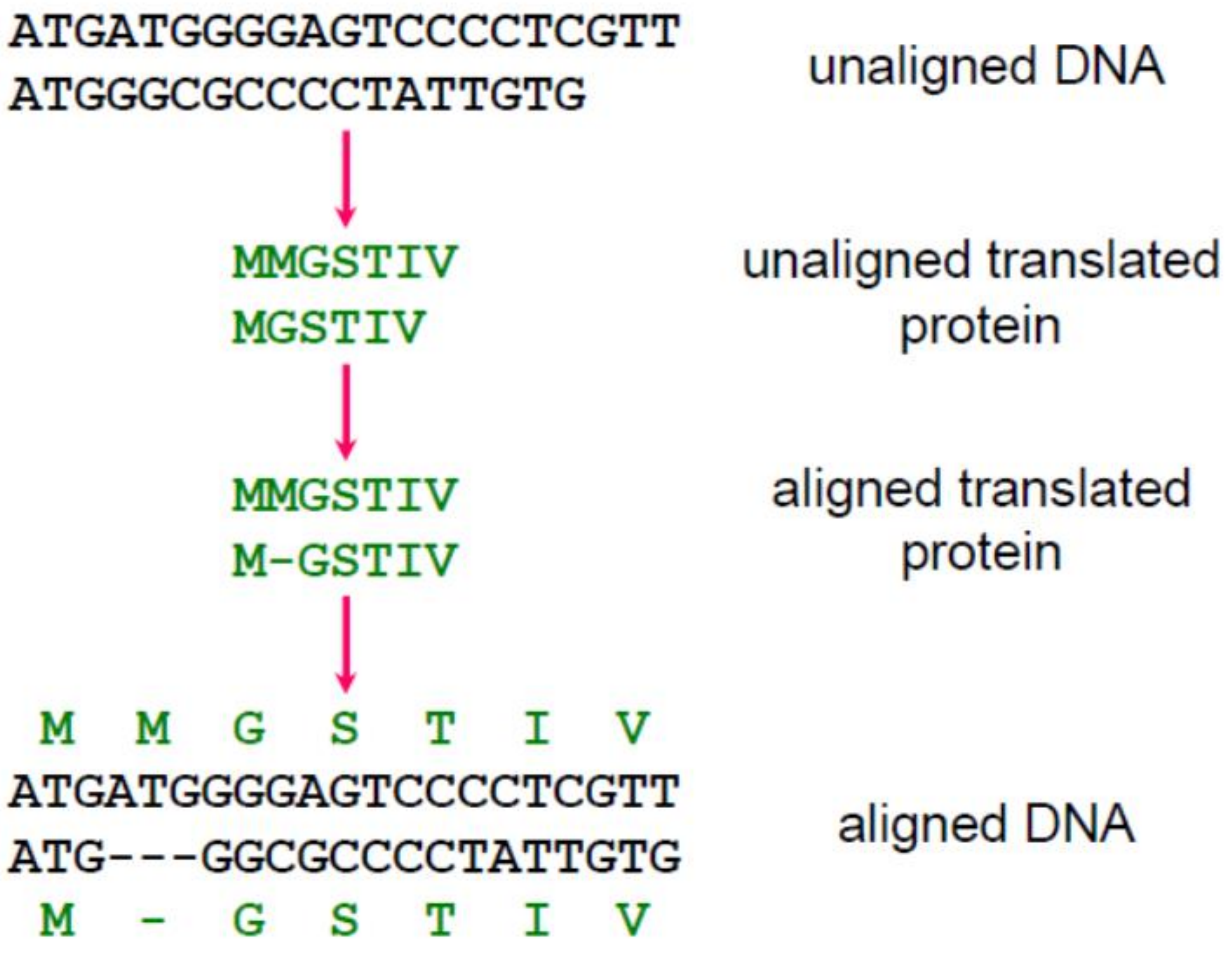
- Proteins are easier to align than DNA

- If it looks wrong it probably is wrong!

- Manual alignment is needed

Proteins are easier to align than DNA

- Therefore, if your DNA sequences are too divergent try aligning their amino acid translation, and then translating the sequence back to DNA



```
ATGATGGGAGTCCCCTCGTT
ATGGGCGCCCCTATTGTG
```

unaligned DNA

```
MMGSTIV
MGSTIV
```

unaligned translated protein

```
MMGSTIV
M-GSTIV
```

aligned translated protein

```
 M  M  G  S  T  I  V
ATGATGGGAGTCCCCTCGTT
ATG---GGCGCCCCTATTGTG
 M  -  G  S  T  I  V
```

aligned DNA

Multiple sequence alignment editors

- BioEdit - MS-Windows
- Genedoc - MS-Windows
- EditSeq/MegAlign - Lasergene - Mac or MS-Windows
- DNA Strider - Macintosh
- Seq-Al - Macintosh
- ASAD - Excel - Macintosh or MS-Windows
- SeqPup - Mac. MS-Windows, X-Windows

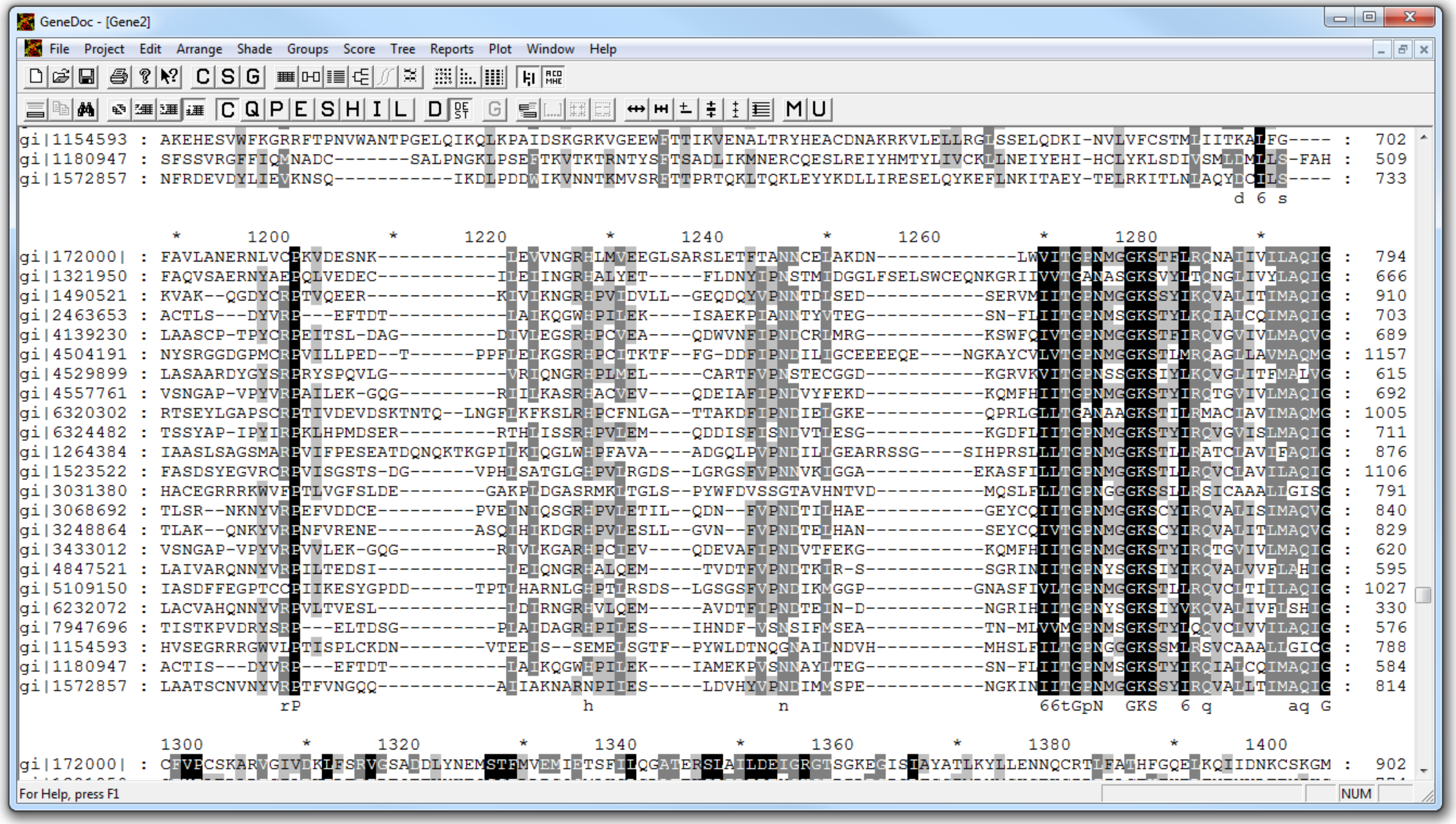
MSA-Visualization and improvement

- GeneDoc (Windows)
- Download: http://genedoc.software.informer.com/download/
  - Arranging and Editing
    - GeneDoc's Grab and Drag arrangement mode allows you to move residues around like beads on a string
  - Shading Alignments
  - Reports: Stats, Score, Composition
  - Exporting and Copying Figures

http://www.nrbsc.org/old/gfx/genedoc/gdpaf.htm

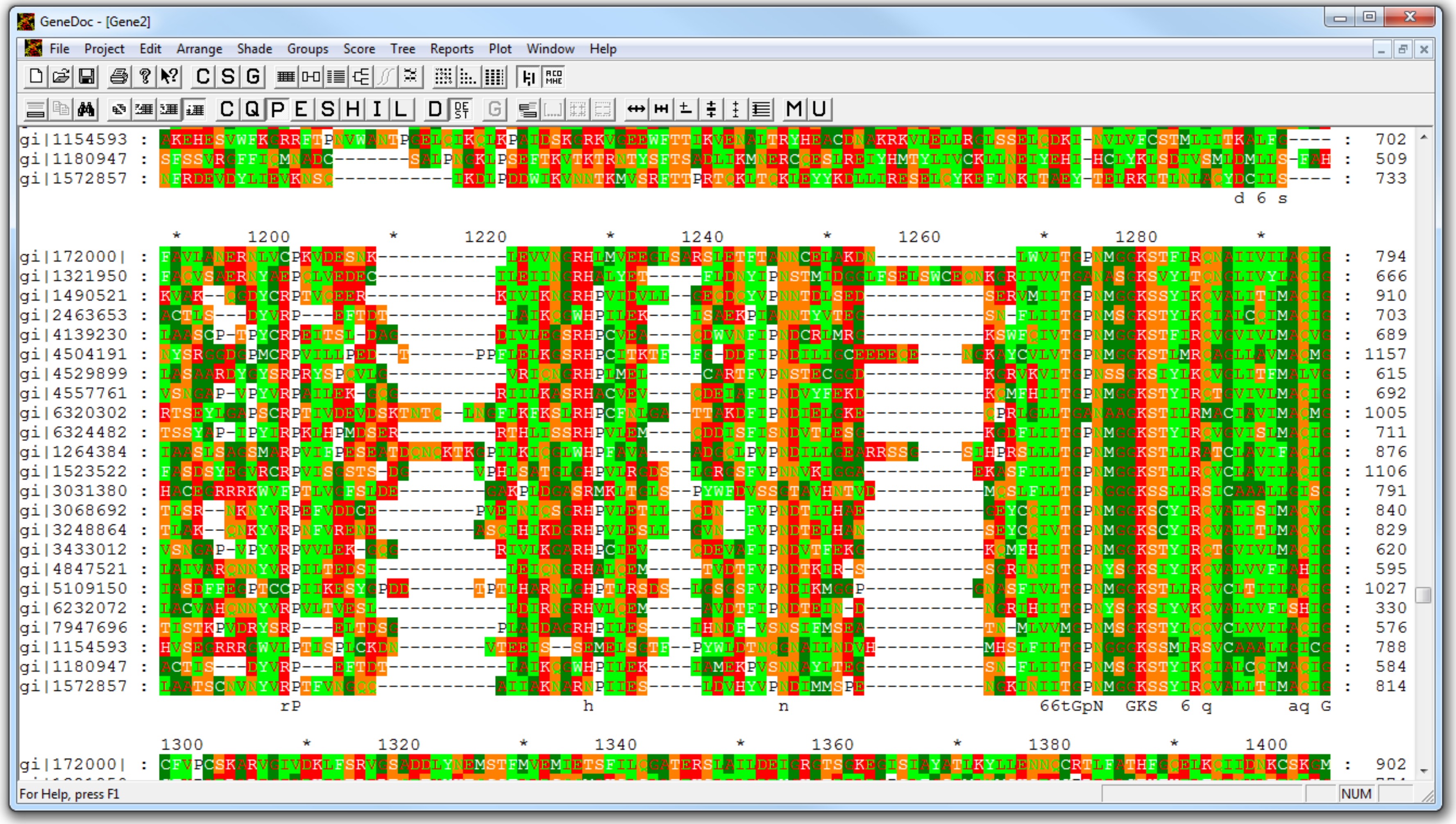
GeneDoc: Conservation Mode

- GeneDoc (Windows)


GeneDoc - [Gene2]
File Project Edit Arrange Shade Groups Score Tree Reports Plot Window Help
C S G
C Q P E S H I L D
M U
For Help, press F1
NUM

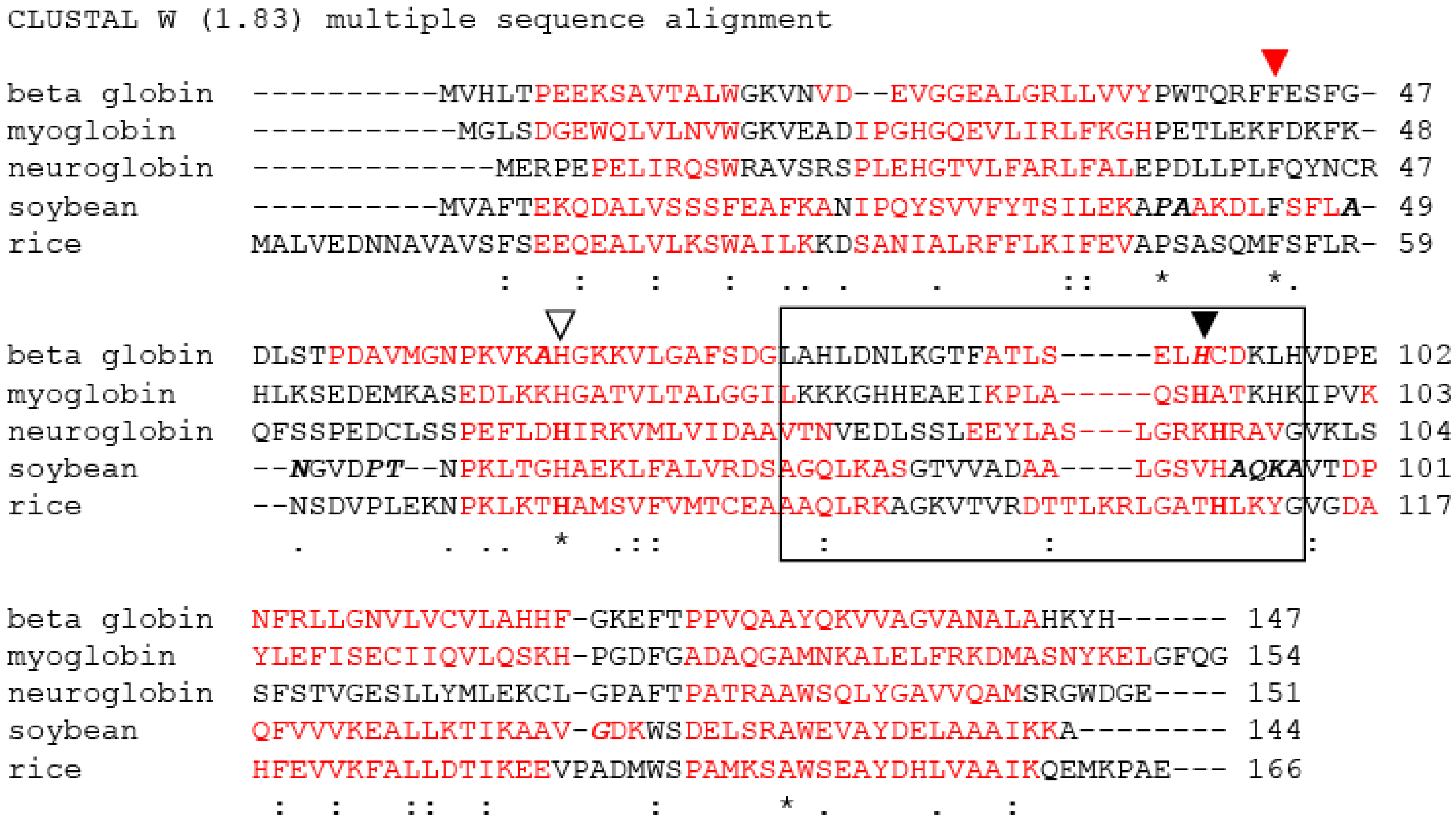
```
gi|1154593 : AKEHESVWFKGRRFTPNVWANTPGELQIKQLKPAIDSKGRKVGEEWFTTIKVENALTRYHEACDNAKRKVLELLRGLSSELQDKI-NVLVFCSTMLIITKALFG---- :  702
gi|1180947 : SFSSVRGFFIQMNADC-------SALPNGKLPSEFTKVTKTRNTYSFTSADLIKMNERCQESLREIYHMTYLIVCKLLNEIYEHI-HCLYKLSDIVSMLDMILS-FAH :  509
gi|1572857 : NFRDEVDYLIEVKNSQ-----------IKDLPDDWIKVNNTKMVSRFTTPRTQKLTQKLEYYKDLLIRESELQYKEFLNKITAEY-TELRKITLNLAQYDCILS---- :  733
                                                                                                             d 6 s

                  *        1200         *        1220         *        1240         *        1260         *        1280         *
gi|172000| : FAVLANERNLVCPKVDESNK------------LEVVNGRHLMVEEGLSARSLETFTANNCELAKDN-------------LWVITGPNMGGKSTFLRQNALIVILAQIG :  794
gi|1321950 : FAQVSAERNYAEPQLVEDEC-----------ILEIINGRHALYET-----FLDNYIPNSTMIDGGLFSELSWCEQNKGRIIVVTGANASGKSVYLTQNGLIVYLAQIG :  666
gi|1490521 : KVAK--QGDYCRPTVQEER------------KIVIKNGRHPVIDVLL--GEQDQYVPNNTDLSED-----------SERVMIITGPNMGGKSSYIKQVALITIMAQIG :  910
gi|2463653 : ACTLS---DYVRP---EFTDT-----------LAIKQGWHPILEK----ISAEKPIANNTYVTEG-----------SN-FLIITGPNMSGKSTYLKQIALCQIMAQIG :  703
gi|4139230 : LAASCP-TPYCRPEITSL-DAG---------DIVLEGSRHPCVEA----QDWVNFIPNDCRLMRG-----------KSWFQIVTGPNMGGKSTFIRQVGVIVLMAQVG :  689
gi|4504191 : NYSRGGDGPMCRPVILLPED--T-------PPFLELKGSRHPCITKTF--FG-DDFIPNDILIGCEEEEQE----NGKAYCVLVTGPNMGGKSTLMRQAGLLAVMAQMG : 1157
gi|4529899 : LASAARDYGYSRPRYSPQVLG-----------VRIQNGRHPLMEL-----CARTFVPNSTECGGD-----------KGRVKVITGPNSSGKSIYLKQVGLITFMALVG :  615
gi|4557761 : VSNGAP-VPYVRPAILEK-GQG----------RIILKASRHACVEV----QDEIAFIPNDVYFEKD-----------KQMFHIITGPNMGGKSTYIRQTGVIVLMAQIG :  692
gi|6320302 : RTSEYLGAPSCRPTIVDEVDSKTNTQ--LNGFLKFKSLRHPCFNLGA--TTAKDFIPNDIELGKE-----------QPRLGLLTGANAAGKSTILRMACIAVIMAQMG : 1005
gi|6324482 : TSSYAP-IPYIRPKLHPMDSER---------RTHLISSRHPVLEM----QDDISFISNDVTLESG-----------KGDFLIITGPNMGGKSTYIRQVGVISLMAQIG :  711
gi|1264384 : IAASLSAGSMARPVIFPESEATDQNQKTKGPILKIQGLWHPFAVA----ADGQLPVPNDILLGEARRSSG----SIHPRSLLLTGPNMGGKSTLLRATCLAVIFAQLG :  876
gi|1523522 : FASDSYEGVRCRPVISGSTS-DG------VPHLSATGLGHPVLRGDS--LGRGSFVPNNVKIGGA----------EKASFILLTGPNMGGKSTLLRQVCLAVILAQIG : 1106
gi|3031380 : HACEGRRRKWVFPTLVGFSLDE--------GAKPLDGASRMKLTGLS--PYWFDVSSGTAVHNTVD----------MQSLFLLTGPNGGGKSSLLRSICAAALLGISG :  791
gi|3068692 : TLSR--NKNYVRPEFVDDCE---------PVEINIQSGRHPVLETIL--QDN--FVPNDTILHAE-----------GEYCQIITGPNMGGKSCYIRQVALISIMAQVG :  840
gi|3248864 : TLAK--QNKYVRPNFVRENE---------ASQIHIKDGRHPVLESLL--GVN--FVPNDTELHAN-----------SEYCQIVTGPNMGGKSCYIRQVALITLMAQVG :  829
gi|3433012 : VSNGAP-VPYVRPVVLEK-GQG----------RIVLKGARHPCIEV----QDEVAFIPNDVTFEKG-----------KQMFHIITGPNMGGKSTYIRQTGVIVLMAQIG :  620
gi|4847521 : LAIVARQNNYVRPILTEDSI------------LEIQNGRHALQEM-----TVDTFVPNDTKIR-S-----------SGRINIITGPNYSGKSIYIKQVALVVFLAHIG :  595
gi|5109150 : IASDFFEGPTCCPIIKESYGPDD------TPTLHARNLGHPTLRSDS--LGSGSFVPNDIKMGGP----------GNASFIVLTGPNMGGKSTLLRQVCLTIILAQIG : 1027
gi|6232072 : LACVAHQNNYVRPVLTVESL------------LDIRNGRHVLQEM-----AVDTFIPNDTEIN-D-----------NGRIHIITGPNYSGKSIYVKQVALIVFLSHIG :  330
gi|7947696 : TISTKPVDRYSRP---ELTDSG---------PLAIDAGRHPILES----IHNDF-VSNSIFMSEA-----------TN-MLVVMGPNMSGKSTYLQQVCLVVILAQIG :  576
gi|1154593 : HVSEGRRRGWVLPTISPLCKDN--------VTEEIS--SEMELSGTF--PYWLDTNQGNAILNDVH----------MHSLFILTGPNGGGKSSMLRSVCAAALLGICG :  788
gi|1180947 : ACTIS---DYVRP---EFTDT-----------LAIKQGWHPILEK----IAMEKPVSNNAYLTEG-----------SN-FLIITGPNMSGKSTYIKQIALCQIMAQIG :  584
gi|1572857 : LAATSCNVNYVRPTFVNGQQ------------AIIAKNARNPIIES-----LDVHYVPNDIMMSPE-----------NGKINIITGPNMGGKSSYIRQVALLTIMAQIG :  814
                         rP                        h                  n                 66tGpN  GKS  6 q      aq G

             1300         *        1320         *        1340         *        1360         *        1380         *        1400
gi|172000| : CFVPCSKARVGIVDKLFSRVGSADDLYNEMSTFMVEMIETSFILQGATERSLAILDEIGRGTSGKEGISIAYATLKYLLENNQCRTLFATHFGQELKQIIDNKCSKGM :  902
```

GeneDoc: Property Mode


GeneDoc - [Gene2]
File Project Edit Arrange Shade Groups Score Tree Reports Plot Window Help
C S G
C Q P E S H I L D G M U
For Help, press F1
NUM

```
gi|1154593 : AKEHESVWFKGRRFTPNVWANTPGELQIKQLKPAIDSKGRKVGEEWFTTIKVENALTRYHEACDNAKRKVLELLRGLSSELQDKI-NVLVFCSTMLIITKALFG---- :  702
gi|1180947 : SFSSVRGFFIQMNADC-------SALPNGKLPSEFTKVTKTRNTYSFTSADLIKMNERCQESIREIYHMTYLIVCKLLNEIYEHI-HCLYKLSDIVSMLDMLLS-FAH :  509
gi|1572857 : NFRDEVDYLIEVKNSQ-----------IKDLPDDWIKVNNTKMVSRFTTPRTQKLTQKLEYYKDLLIRESELQYKEFLNKITAEY-TELRKITLNLAQYDCILS---- :  733
                                                                                                               d 6 s

              *      1200         *      1220         *      1240         *      1260         *      1280         *
gi|172000| : FAVLANERNLVCPKVDESNK------------LEVVNGRHLMVEEGLSARSLETFTANNCELAKDN-------------LWVITGPNMGGKSTFLRQNAIIVILAQIG :  794
gi|1321950 : FAQVSAERNYAEPQLVEDEC-----------ILEIINGRHALYET-----FLDNYIPNSTMIDGGLFSELSWCEQNKGRIIVVTGANASGKSVYLTQNGLIVYLAQIG :  666
gi|1490521 : KVAK--QGDYCRPTVQEER------------KIVIKNGRHPVIDVLL--GEQDQYVPNNTDLSED-----------SERVMIITGPNMGGKSSYIKQVALITIMAQIG :  910
gi|2463653 : ACTLS---DYVRP---EFTDT-----------LAIKQGWHPILEK----ISAEKPIANNTYVTEG-----------SN-FLIITGPNMSGKSTYLKQIALCQIMAQIG :  703
gi|4139230 : LAASCP-TPYCRPEITSL-DAG---------DIVLEGSRHPCVEA----QDWVNFIPNDCRLMRG-----------KSWFQIVTGPNMGGKSTFIRQVGVIVLMAQVG :  689
gi|4504191 : NYSRGGDGPMCRPVILLPED--T------PPFLELKGSRHPCITKTF--FG-DDFIPNDILIGCEEEEQE----NGKAYCVLVTGPNMGGKSTLMRQAGLLAVMAQMG : 1157
gi|4529899 : LASAARDYGYSRPRYSPQVLG-----------VRIQNGRHPLMEL-----CARTFVPNSTECGGD-----------KGRVKVITGPNSSGKSIYLKQVGLITFMALVG :  615
gi|4557761 : VSNGAP-VPYVRPAILEK-GQG---------RIILKASRHACVEV----QDEIAFIPNDVYFEKD-----------KQMFHIITGPNMGGKSTYIRQTGVIVLMAQIG :  692
gi|6320302 : RTSEYLGAPSCRPTIVDEVDSKTNTQ--LNGFLKFKSLRHPCFNLGA--TTAKDFIPNDIELGKE-----------QPRLGLLTGANAAGKSTILRMACIAVIMAQMG : 1005
gi|6324482 : TSSYAP-IPYIRPKLHPMDSER---------RTHLISSRHPVLEM----QDDISFISNDVTLESG-----------KGDFLIITGPNMGGKSTYIRQVGVISLMAQIG :  711
gi|1264384 : IAASLSAGSMARPVIFPESEATDQNQKTKGPILKIQGLWHPFAVA----ADGQLPVPNDILLGEARRSSG----SIHPRSLLLTGPNMGGKSTLIRATCLAVIFAQLG :  876
gi|1523522 : FASDSYEGVRCRPVISGSTS-DG------VPHLSATGLGHPVLRGDS--LGRGSFVPNNVKIGGA----------EKASFILLTGPNMGGKSTLLRQVCLAVILAQIG : 1106
gi|3031380 : HACEGRRRKWVFPTLVGFSLDE--------GAKPLDGASRMKLTGLS--PYWFDVSSGTAVHNTVD----------MQSLFLLTGPNGGGKSSLIRSICAAALLGISG :  791
gi|3068692 : TLSR--NKNYVRPEFVDDCE---------PVEINIQSGRHPVLETIL--QDN--FVPNDTILHAE-----------GEYCQIITGPNMGGKSCYIRQVALISIMAQVG :  840
gi|3248864 : TLAK--QNKYVRPNFVRENE---------ASQIHIKDGRHPVLESLL--GVN--FVPNDTELHAN-----------SEYCQIVTGPNMGGKSCYIRQVALITLMAQVG :  829
gi|3433012 : VSNGAP-VPYVRPVVLEK-GQG---------RIVLKGARHPCIEV----QDEVAFIPNDVTFEKG-----------KQMFHIITGPNMGGKSTYIRQTGVIVLMAQIG :  620
gi|4847521 : LAIVARQNNYVRPILTEDSI------------LEIQNGRHALQEM-----TVDTFVPNDTKIR-S-----------SGRINIITGPNYSGKSIYIKQVALVVFLAHIG :  595
gi|5109150 : IASDFFEGPTCCPIIKESYGPDD------TPTLHARNLGHPTIRSDS--LGSGSFVPNDIKMGGP----------GNASFIVLTGPNMGGKSTLLRQVCLTIILAQIG : 1027
gi|6232072 : LACVAHQNNYVRPVLTVESL------------LDIRNGRHVLQEM-----AVDTFIPNDTEIN-D-----------NGRIHIITGPNYSGKSIYVKQVALIVFLSHIG :  330
gi|7947696 : TISTKPVDRYSRP---ELTDSG---------PLAIDAGRHPILES----IHNDF-VSNSIFMSEA-----------TN-MLVVMGPNMSGKSTYLQQVCLVVILAQIG :  576
gi|1154593 : HVSEGRRRGWVLPTISPLCKDN--------VTEEIS--SEMELSGTF--PYWLDTNQGNAILNDVH----------MHSLFILTGPNGGGKSSMLRSVCAAALLGICG :  788
gi|1180947 : ACTIS---DYVRP---EFTDT-----------LAIKQGWHPILEK----IAMEKPVSNNAYLTEG-----------SN-FLIITGPNMSGKSTYIKQIALCQIMAQIG :  584
gi|1572857 : LAATSCNVNYVRPTFVNGQQ-----------AIIAKNARNPIIES-----LDVHYVPNDIMMSPE-----------NGKINIITGPNMGGKSSYIRQVALLTIMAQIG :  814
                        rP                          h                 n                       66tGpN  GKS  6 q       aq G

             1300         *      1320         *      1340         *      1360         *      1380         *      1400
gi|172000| : CFVPCSKARVGIVDKLFSRVGSADDLYNEMSTFMVEMIETSFILQGATERSLAILDEIGRGTSGKEGISIAYATLKYLLENNQCRTLFATHFGQELKQIIDNKCSKGM :  902
```

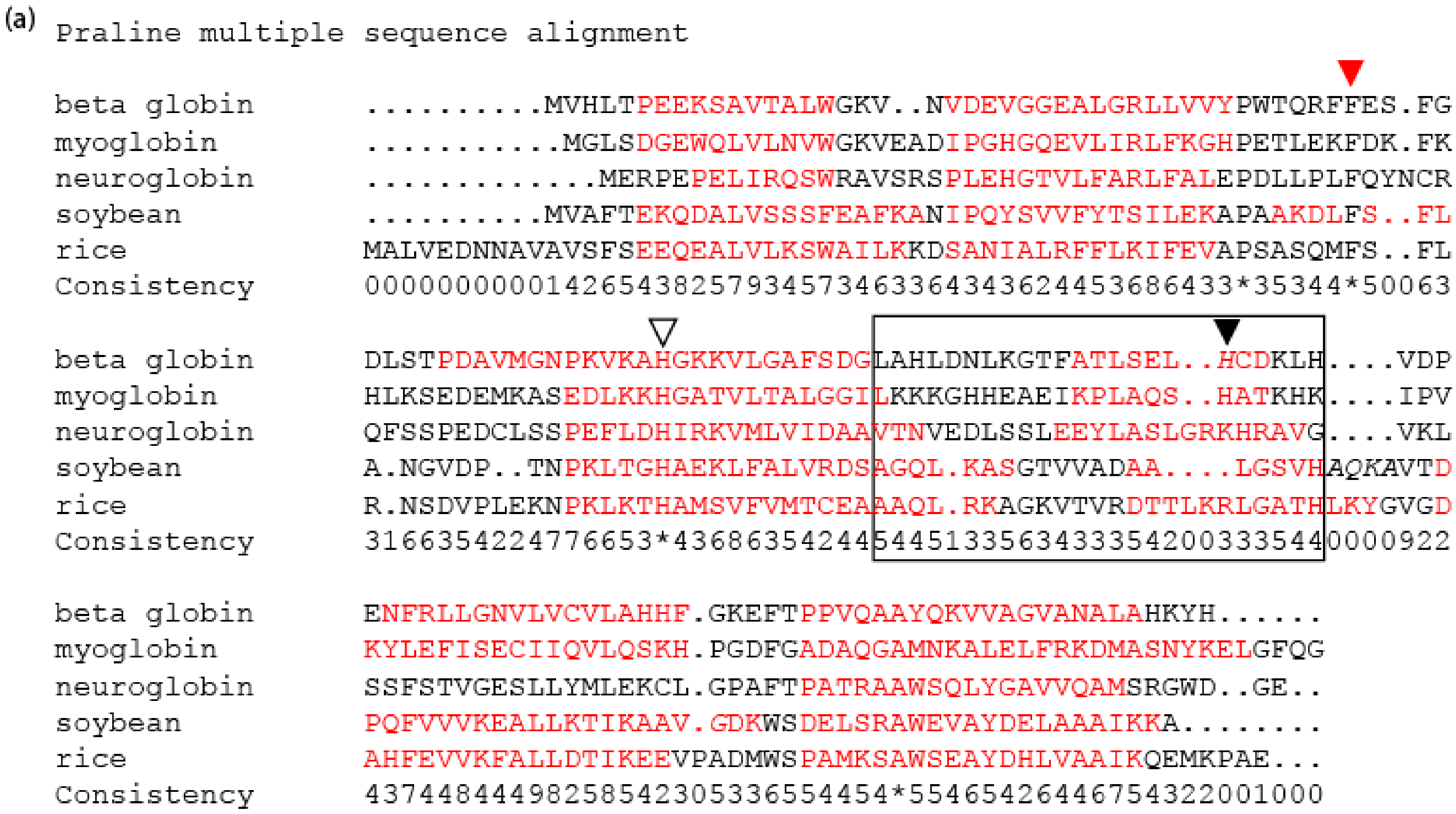
ClustalW



```
CLUSTAL W (1.83) multiple sequence alignment
                                                                   ▼
beta globin  ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin    -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFK- 48
neuroglobin  -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR 47
soybean      ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice         MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLR- 59
                          :   :   :   :  .. .    .      ::   *     *.
                             ▽                                 ▼
beta globin  DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLS-----ELHCDKLHVDPE 102
myoglobin    HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLA-----QSHATKHKIPVK 103
neuroglobin  QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLAS---LGRKHRAVGVKLS 104
soybean      --NGVDPT--NPKLTGHAEKLFALVRDSAGQLKASGTVVADAA----LGSVHAQKAVTDP 101
rice         --NSDVPLEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117
               .       . ..   *  .::        :           :             :

beta globin  NFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH------ 147
myoglobin    YLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin  SFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151
soybean      QFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA-------- 144
rice         HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166
              :  :   ::  :        :      * .     .   :
```

Note how the region of a conserved histidine (▼) varies depending on which of five prominent algorithms is used

Praline

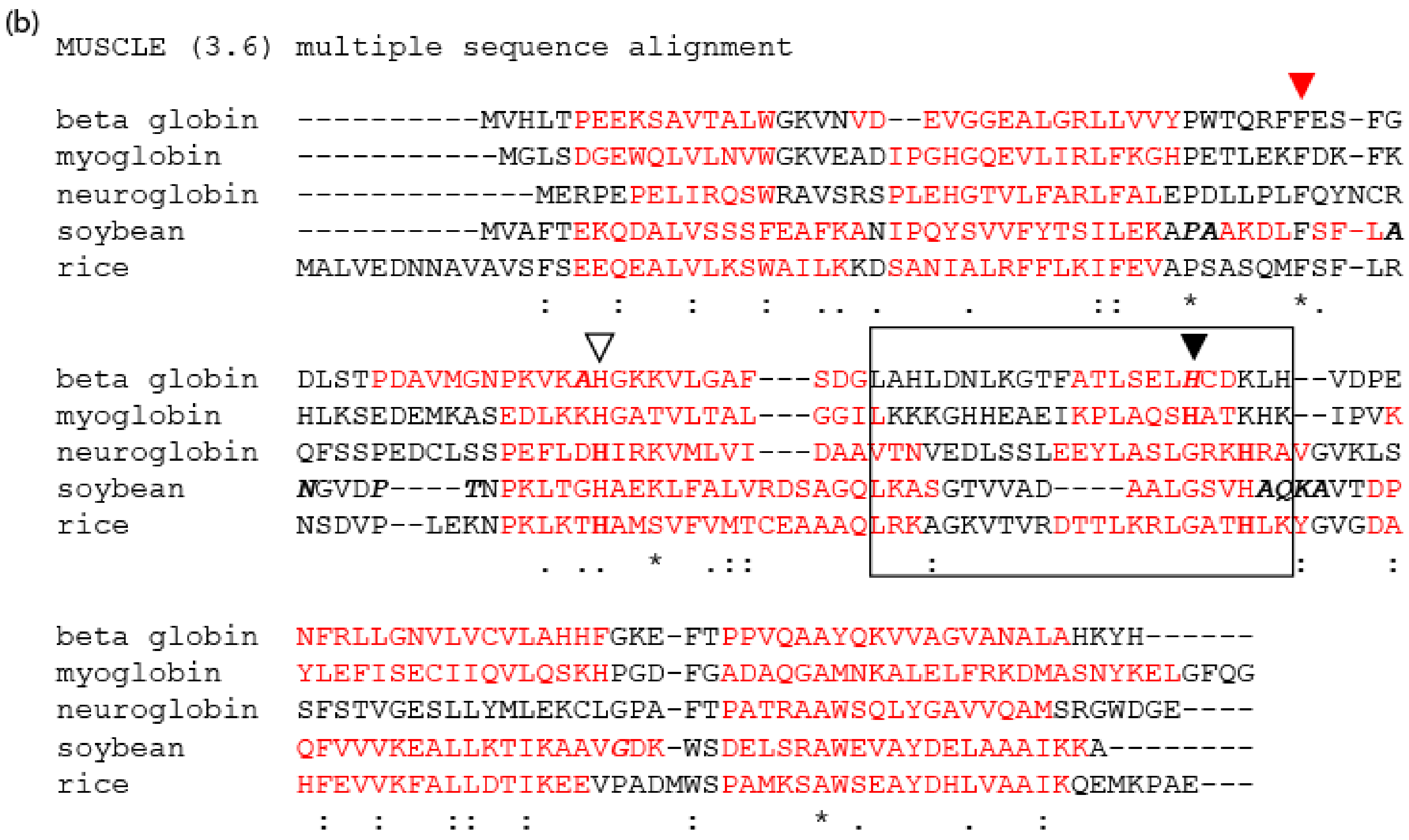


(a)

```
Praline multiple sequence alignment

beta globin     ..........MVHLTPEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFES.FG
myoglobin       ...........MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK.FK
neuroglobin     .............MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean         ..........MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS..FL
rice            MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS..FL
Consistency     000000000014265438257934573463364343624453686433*35344*50063

beta globin     DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSEL..HCDKLH....VDP
myoglobin       HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQS..HATKHK....IPV
neuroglobin     QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHRAVG....VKL
soybean         A.NGVDP..TNPKLTGHAEKLFALVRDSAGQL.KASGTVVADAA....LGSVHAQKAVTD
rice            R.NSDVPLEKNPKLKTHAMSVFVMTCEAAAQL.RKAGKVTVRDTTLKRLGATHLKYGVGD
Consistency     316635422477665 3*43686354244544513356343335420033354400000922

beta globin     ENFRLLGNVLVCVLAHHF.GKEFTPPVQAAYQKVVAGVANALAHKYH......
myoglobin       KYLEFISECIIQVLQSKH.PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin     SSFSTVGESLLYMLEKCL.GPAFTPATRAAWSQLYGAVVQAMSRGWD..GE..
soybean         PQFVVVKEALLKTIKAAV.GDKWSDELSRAWEVAYDELAAAIKKA........
rice            AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE...
Consistency     43744844498258542305336554454*554654264467543220010 00
```

Note also the changing pattern of gaps within the boxed region in these five different alignments.

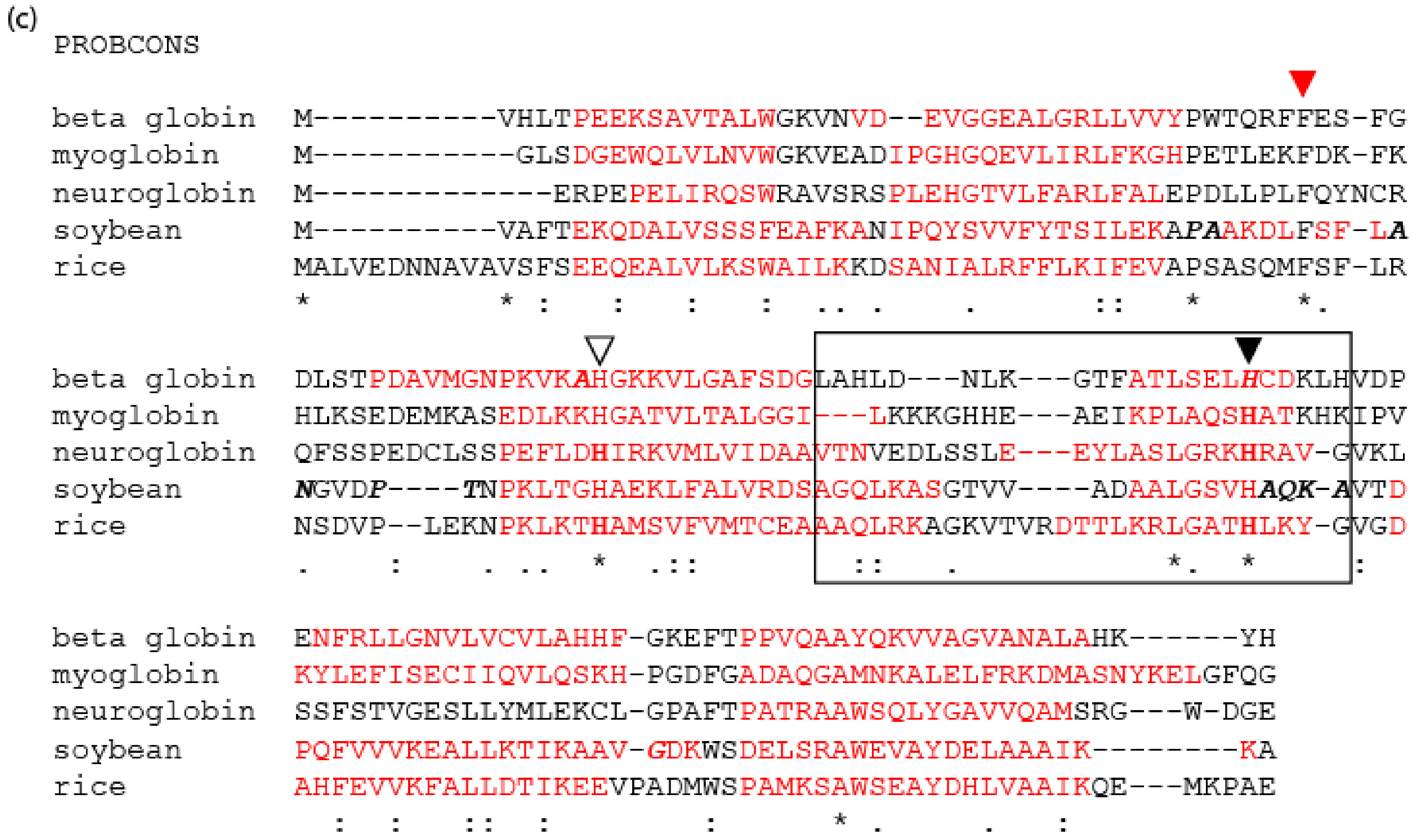
MUSCLE



(b)

```
MUSCLE (3.6) multiple sequence alignment

                                                                   ▼
beta globin  ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
myoglobin    -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK
neuroglobin  -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean      ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice         MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
                          :   :   :   :  .. .    .      ::   *     *.

                             ▽                               ▼
beta globin  DLSTPDAVMGNPKVKAHGKKVLGAF---SDGLAHLDNLKGTFATLSELHCDKLH--VDPE
myoglobin    HLKSEDEMKASEDLKKHGATVLTAL---GGILKKKGHHEAEIKPLAQSHATKHK--IPVK
neuroglobin  QFSSPEDCLSSPEFLDHIRKVMLVI---DAAVTNVEDLSSLEEYLASLGRKHRAVGVKLS
soybean      NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDP
rice         NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
                          . ..  *  .::         :                   :    :

beta globin  NFRLLGNVLVCVLAHHFGKE-FTPPVQAAYQKVVAGVANALAHKYH------
myoglobin    YLEFISECIIQVLQSKHPGD-FGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin  SFSTVGESLLYMLEKCLGPA-FTPATRAAWSQLYGAVVQAMSRGWDGE----
soybean      QFVVVKEALLKTIKAAVGDK-WSDELSRAWEVAYDELAAAIKKA--------
rice         HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE---
              :  :   ::  :        :      * .     .   :
```

Probcons

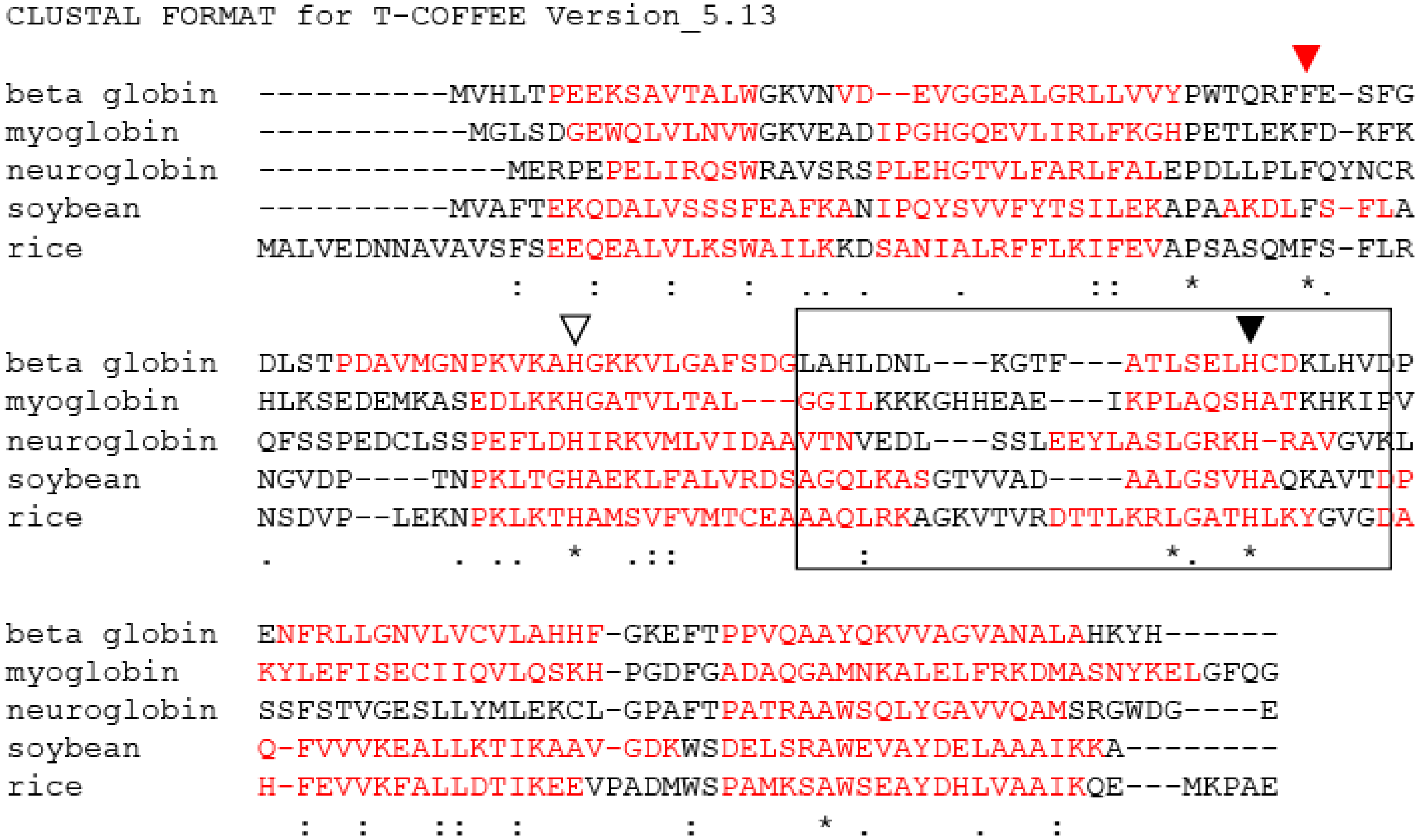


(c)

```
PROBCONS

beta globin    M----------VHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
myoglobin      M-----------GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK
neuroglobin    M-------------ERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean        M----------VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice           MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
               *          * :   :   :   :  .. .    .      ::   *     *.

beta globin    DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD---NLK---GTFATLSELHCDKLHVDP
myoglobin      HLKSEDEMKASEDLKKHGATVLTALGGI---LKKKGHHE---AEIKPLAQSHATKHKIPV
neuroglobin    QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLE---EYLASLGRKHRAV-GVKL
soybean        NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVV----ADAALGSVHAQK-AVTD
rice           NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKY-GVGD
               .    :    . ..  *  .::        ::   .           *.  *     :

beta globin    ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHK------YH
myoglobin      KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin    SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRG---W-DGE
soybean        PQFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIK--------KA
rice           AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQE---MKPAE
                 :  :   ::  :        :      * .     .   :
```

TCoffee

(d)



```
CLUSTAL FORMAT for T-COFFEE Version_5.13

beta globin  ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFE-SFG
myoglobin    -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFD-KFK
neuroglobin  -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean      ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS-FLA
rice         MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS-FLR
                          :   :   :   :  .. .    .      ::   *     *.

beta globin  DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNL---KGTF---ATLSELHCDKLHVDP
myoglobin    HLKSEDEMKASEDLKKHGATVLTAL---GGILKKKGHHEAE---IKPLAQSHATKHKIPV
neuroglobin  QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDL---SSLEEYLASLGRKH-RAVGVKL
soybean      NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDP
rice         NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
             .         . ..  *  .::         :               *.  *

beta globin  ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH------
myoglobin    KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin  SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDG----E
soybean      Q-FVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA--------
rice         H-FEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQE---MKPAE
               :  :   ::  :        :      * .     .   :
```