Chapter 6:
Multiple Sequence Alignment

Multiple sequence alignment: definition

- a collection of three or more protein (or nucleic acid) sequences that are partially or completely aligned
- homologous residues are aligned in columns across the length of the sequences
- residues are homologous in an evolutionary sense
- residues are homologous in a structural sense

Example: 5 alignments of 5 globins

Let's look at a multiple sequence alignment (MSA) of five globins proteins. We'll use five prominent MSA programs: ClustalW, Praline, MUSCLE (used at HomoloGene), ProbCons, and TCoffee. Each program offers unique strengths.

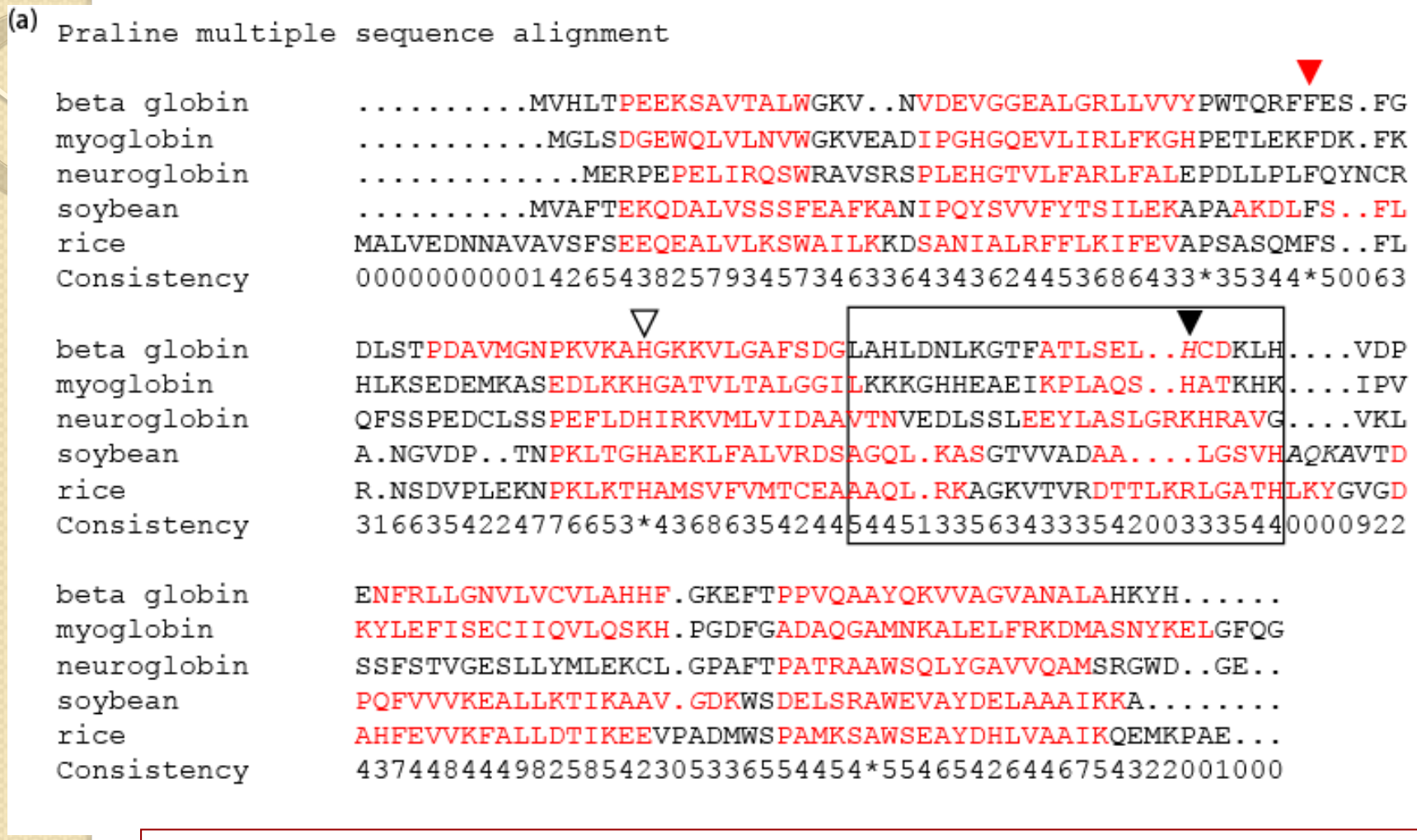
We'll focus on a histidine (H) residue that has a critical role in binding oxygen in globins, and should be aligned. But often it's not aligned, and all five programs give different answers.

Our conclusion will be that there is no single best approach to MSA. Dozens of new programs have been introduced in recent years.

ClustalW

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

Praline

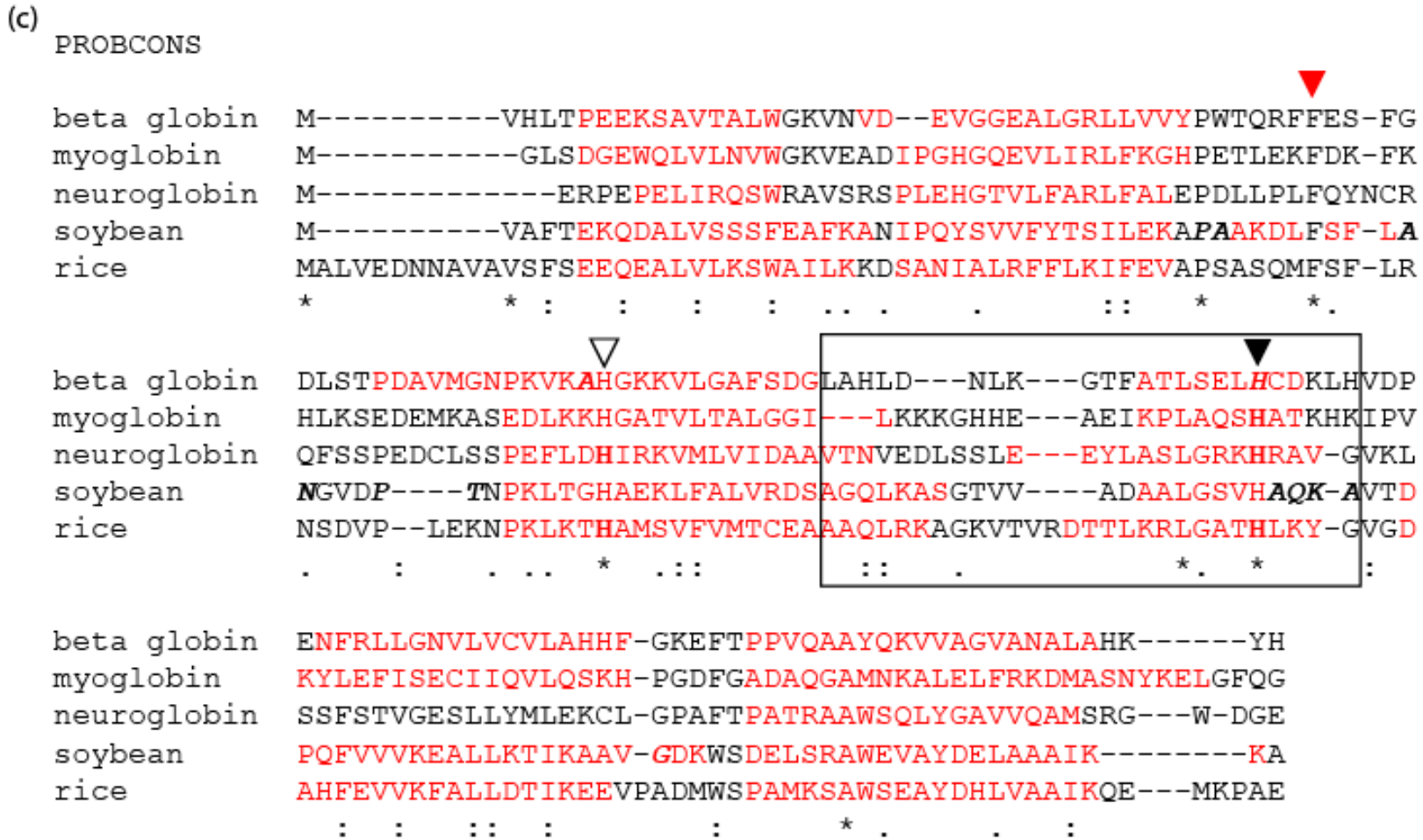


Note also the changing pattern of gaps within the boxed region in these five different alignments.

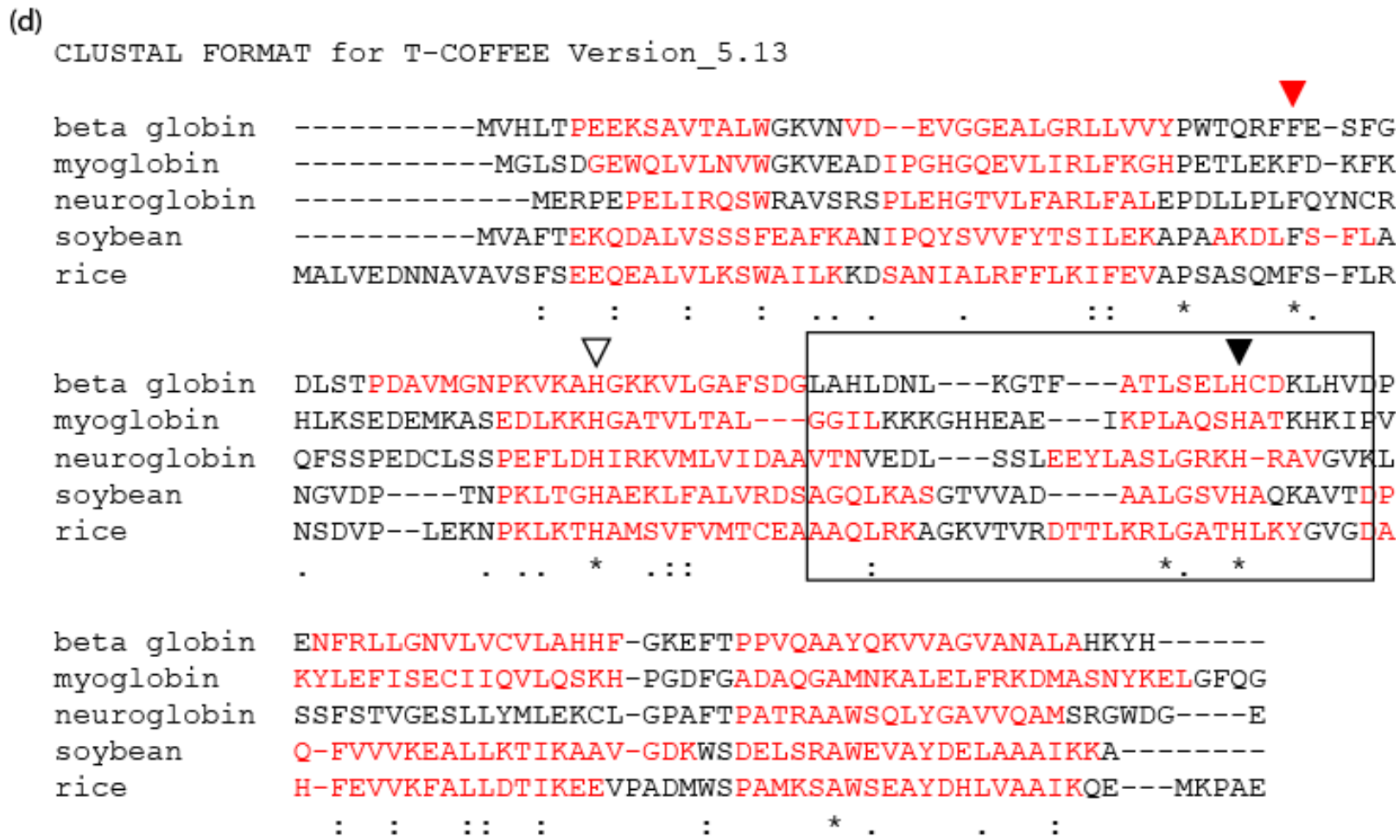
MUSCLE


(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---IKPLAQSHATKHKIP
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

Multiple sequence alignment: properties

- not necessarily one "correct" alignment of a protein family
- protein sequences evolve...
- ...the corresponding three-dimensional structures of proteins also evolve
- may be impossible to identify amino acid residues that align properly (structurally) throughout a multiple sequence alignment
- for two proteins sharing 30% amino acid identity, about 50% of the individual amino acids are superposable in the two structures

Multiple sequence alignment: features

- some aligned residues, such as cysteines that form disulfide bridges, may be highly conserved
- there may be conserved motifs such as a transmembrane domain
- there may be conserved secondary structure features
- there may be regions with consistent patterns of insertions or deletions (indels)

Multiple sequence alignment: uses

- MSA is more sensitive than pairwise alignment to detect homologs
- BLAST output can take the form of a MSA, and can reveal conserved residues or motifs
- A single query can be searched against a database of MSAs (e.g. PFAM)
- Regulatory regions of genes may have consensus sequences identifiable by MSA

Multiple sequence alignment: exact methods

Exact methods of multiple alignment use dynamic programming and are guaranteed to find optimal solutions. But they are not feasible for more than a few sequences.

Multiple sequence alignment: methods

Progressive methods: use a guide tree (related to a phylogenetic tree) to determine how to combine pairwise alignments one by one to create a multiple alignment.

Examples: CLUSTALW, MUSCLE

Multiple sequence alignment: methods

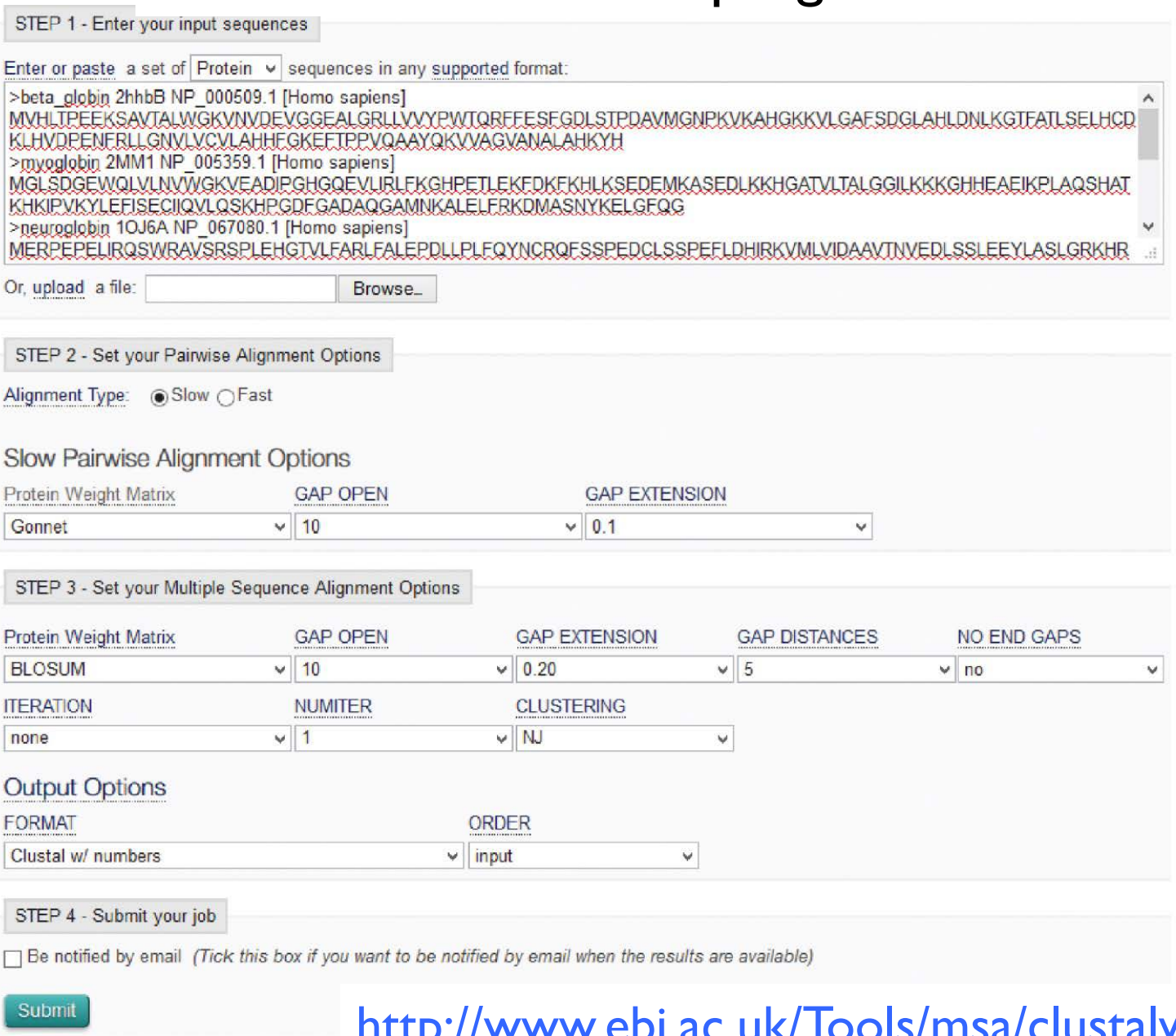
Example of MSA using ClustalW: two data sets

Five distantly related globins (human to plant)

Five closely related beta globins

Obtain your sequences in the FASTA format!
You can save them in a Word document or text editor.
Visit www.bioinfbook.org for web documents 6-3 and 6-4

Use ClustalW to do a progressive MSA



http://www.ebi.ac.uk/Tools/msa/clustalw2/

B&FG 3e
Fig. 6.1
Page 209

ClustalW stage 1: series of pairwise alignments

(a) Stage 1: series of pairwise alignments

SeqA	Name	Length	SeqB	Name	Length	Score
1	beta_globin	147	2	myoglobin	154	25.17
1	beta_globin	147	3	neuroglobin	151	15.65
1	beta_globin	147	4	soybean_globin	144	13.19
1	beta_globin	147	5	rice_globin	166	21.09
2	myoglobin	154	3	neuroglobin	151	16.56
2	myoglobin	154	4	soybean_globin	144	8.33
2	myoglobin	154	5	rice_globin	166	12.99
3	neuroglobin	151	4	soybean_globin	144	17.36
3	neuroglobin	151	5	rice_globin	166	18.54
4	soybean_globin	144	5	rice_globin	166	43.06

1

best score (highest percent pairwise identity)

ClustalW stage 1: series of pairwise alignments

(a) Stage 1: series of pairwise alignments

SeqA	Name	Length	SeqB	Name	Length	Score
1	beta_globin	147	2	myoglobin	154	25.17
1	beta_globin	147	3	neuroglobin	151	15.65
1	beta_globin	147	4	soybean_globin	144	13.19
1	beta_globin	147	5	rice_globin	166	21.09
2	myoglobin	154	3	neuroglobin	151	16.56
2	myoglobin	154	4	soybean_globin	144	8.33
2	myoglobin	154	5	rice_globin	166	12.99
3	neuroglobin	151	4	soybean_globin	144	17.36
3	neuroglobin	151	5	rice_globin	166	18.54
4	soybean_globin	144	5	rice_globin	166	43.06

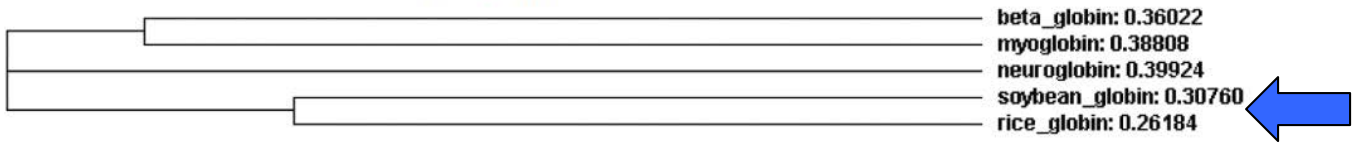
1

best score (highest percent pairwise identity)

ClustalW stage 2: create a guide tree

(b) Stage 2: create a guide tree (calculated from a distance matrix)

```
(
(
beta_globin:0.36022,
myoglobin:0.38808)
:0.06560,
neuroglobin:0.39924,
(
soybean_globin:0.30760,
rice_globin:0.26184)
:0.13652);
```



Note that the two proteins with the highest percent pairwise identity (soybean and rice globin) also have the shortest connecting branch lengths in the tree

Feng-Doolittle MSA occurs in 3 stages

[1] Do a set of global pairwise alignments (Needleman and Wunsch's dynamic programming algorithm)

[2] Create a guide tree

[3] Progressively align the sequences

Progressive MSA stage 1 of 3: generate global pairwise alignments

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	beta_globin	147	2	myoglobin	154	25
1	beta_globin	147	3	neuroglobin	151	15
1	beta_globin	147	4	soybean	144	13
1	beta_globin	147	5	rice	166	21
2	myoglobin	154	3	neuroglobin	151	16
2	myoglobin	154	4	soybean	144	8
2	myoglobin	154	5	rice	166	12
3	neuroglobin	151	4	soybean	144	17
3	neuroglobin	151	5	rice	166	18
4	soybean	144	5	rice	166	43

best score

Number of pairwise alignments needed

For n sequences, $(n\text{-}1)(n) / 2$

For 5 sequences, $(4)(5) / 2 = 10$

For 200 sequences, $(199)(200) / 2 = 19{,}900$

Feng-Doolittle stage 2: guide tree

- Convert similarity scores to distance scores
- A tree shows the distance between objects
- Use UPGMA (defined in the phylogeny chapter)
- ClustalW provides a syntax to describe the tree

ClustalW alignment of five distantly related beta globin orthologs

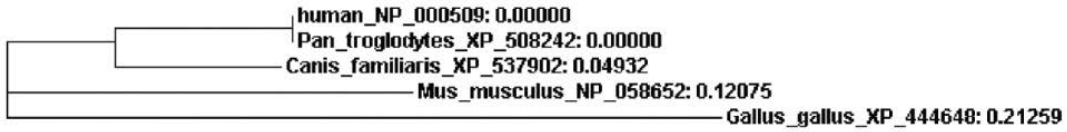
B&FG 3e
Fig. 6.3
Page 211

(a) Stage 1: series of pairwise alignments (closely related globin proteins)

SeqA	Name	Length	SeqB	Name	Length	Score
1	human_NP_000509	147	2	Pan_troglodytes_XP_508242	147	100.0
1	human_NP_000509	147	3	Canis_familiaris_XP_537902	147	89.8
1	human_NP_000509	147	4	Mus_musculus_NP_058652	147	80.27
1	human_NP_000509	147	5	Gallus_gallus_XP_444648	147	69.39
2	Pan_troglodytes_XP_508242	147	3	Canis_familiaris_XP_537902	147	89.8
2	Pan_troglodytes_XP_508242	147	4	Mus_musculus_NP_058652	147	80.27
2	Pan_troglodytes_XP_508242	147	5	Gallus_gallus_XP_444648	147	69.39
3	Canis_familiaris_XP_537902	147	4	Mus_musculus_NP_058652	147	78.91
3	Canis_familiaris_XP_537902	147	5	Gallus_gallus_XP_444648	147	71.43
4	Mus_musculus_NP_058652	147	5	Gallus_gallus_XP_444648	147	66.67

(b) Stage 2: create a guide tree (calculated from a distance matrix)

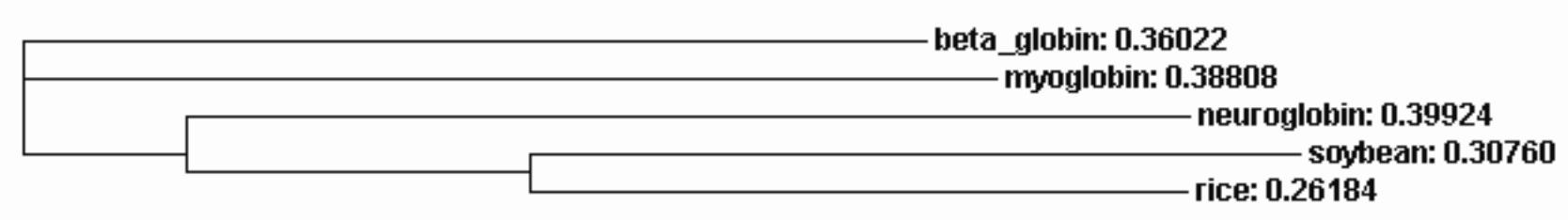
```
(
(
(
human_NP_000509:0.00000,
Pan_troglodytes_XP_508242:0.00000)
:0.05272,
Canis_familiaris_XP_537902:0.04932)
:0.03231,
Mus_musculus_NP_058652:0.12075,
Gallus_gallus_XP_444648:0.21259);
```



ClustalW alignment of five closely related beta globin orthologs

Progressive MSA stage 2 of 3: generate a guide tree calculated from the distance matrix (5 distantly related globins)

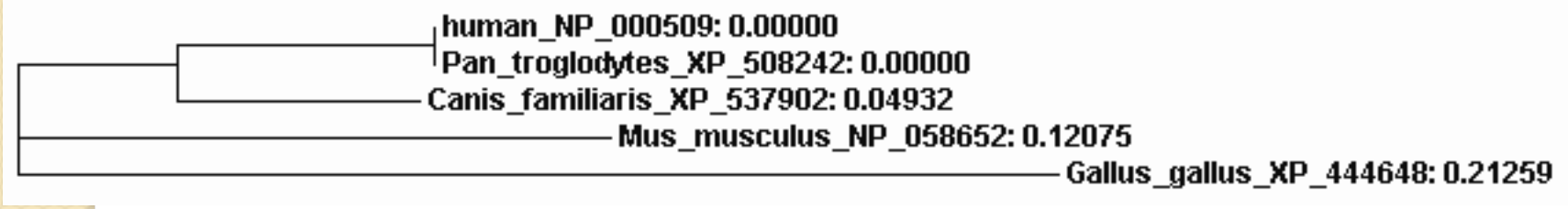
```
(
beta_globin:0.36022,
myoglobin:0.38808,
(
neuroglobin:0.39924,
(
soybean:0.30760,
rice:0.26184)
:0.13652)
:0.06560);
```



SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	human_NP_000509	147	2	Pan_troglodytes_XP_508242	147	100
1	human_NP_000509	147	3	Canis_familiaris_XP_537902	147	89
1	human_NP_000509	147	4	Mus_musculus_NP_058652	147	80
1	human_NP_000509	147	5	Gallus_gallus_XP_444648	147	69
2	Pan_troglodytes_XP_508242	147	3	Canis_familiaris_XP_537902	147	89
2	Pan_troglodytes_XP_508242	147	4	Mus_musculus_NP_058652	147	80
2	Pan_troglodytes_XP_508242	147	5	Gallus_gallus_XP_444648	147	69
3	Canis_familiaris_XP_537902	147	4	Mus_musculus_NP_058652	147	78
3	Canis_familiaris_XP_537902	147	5	Gallus_gallus_XP_444648	147	71
4	Mus_musculus_NP_058652	147	5	Gallus_gallus_XP_444648	147	66

```
(
(
(
human_NP_000509:0.00000,
Pan_troglodytes_XP_508242:0.00000)
:0.05272,
Canis_familiaris_XP_537902:0.04932)
:0.03231,
Mus_musculus_NP_058652:0.12075,
Gallus_gallus_XP_444648:0.21259);
```

5 closely related globins



Feng-Doolittle stage 3: progressive alignment

- Make a MSA based on the order in the guide tree
- Start with the two most closely related sequences
- Then add the next closest sequence
- Continue until all sequences are added to the MSA
- Rule: "once a gap, always a gap."

Why "once a gap, always a gap"?

- There are many possible ways to make a MSA
- Where gaps are added is a critical question
- Gaps are often added to the first two (closest) sequences
- To change the initial gap choices later on would be to give more weight to distantly related sequences
- To maintain the initial gap choices is to trust that those gaps are most believable

Additional features of ClustalW improve its ability to generate accurate MSAs

- Individual weights are assigned to sequences; very closely related sequences are given less weight, while distantly related sequences are given more weight

- Scoring matrices are varied dependent on the presence of conserved or divergent sequences, e.g.:

PAM20	80-100% id
PAM60	60-80% id
PAM120	40-60% id
PAM350	0-40% id

- Residue-specific gap penalties are applied

Iterative approaches: MAFFT

- Uses Fast Fourier Transform to speed up profile alignment
- Uses fast two-stage method for building alignments using k-mer frequencies
- Offers many different scoring and aligning techniques
- One of the more accurate programs available
- Available as standalone or web interface
- Many output formats, including interactive phylogenetic trees
- We will skip the details

Multiple sequence alignment methods

Iterative methods: compute a sub-optimal solution and keep modifying that intelligently using dynamic programming or other methods until the solution converges.

Examples: MUSCLE, IterAlign, Praline, MAFFT

MUSCLE: next-generation progressive MSA

[1] Build a draft progressive alignment

- Determine pairwise similarity through k-mer counting (not by alignment)
- Compute distance (triangular distance) matrix
- Construct tree using UPGMA
- Construct draft progressive alignment following tree

MUSCLE: next-generation progressive MSA

[2] Improve the progressive alignment

Compute pairwise identity through current MSA

Construct new tree with Kimura distance measures

Compare new and old trees: if improved, repeat this step, if not improved, then we're done

MUSCLE: next-generation progressive MSA

[3] Refinement of the MSA

Split tree in half by deleting one edge

Make profiles of each half of the tree

Re-align the profiles

Accept/reject the new alignment

Multiple sequence alignment: consistency

Consistency-based algorithms: generally use a database of both local high-scoring alignments and long-range global alignments to create a final alignment

These are very powerful, very fast, and very accurate methods

Examples: T-COFFEE, Prrp, DiAlign, ProbCons

ProbCons—consistency-based approach

Combines iterative and progressive approaches with a unique probabilistic model.

Uses Hidden Markov Models to calculate probability matrices for matching residues, uses this to construct a guide tree

Progressive alignment hierarchically along guide tree

Post-processing and iterative refinement (a little like MUSCLE)

ProbCons—consistency-based approach

Sequence x x_i

Sequence y y_j

Sequence z z_k

If x_i aligns with z_k

and z_k aligns with y_j

then x_i should align with y_j

ProbCons incorporates evidence from multiple sequences to guide the creation of a pairwise alignment.

Multiple sequence alignment: methods

How do we know which program to use?

There are benchmarking multiple alignment datasets that have been aligned painstakingly by hand, by structural similarity, or by extremely time- and memory-intensive automated exact algorithms.

Some programs have interfaces that are more user-friendly than others. And most programs are excellent so it depends on your preference.

If your proteins have 3D structures, use these to help you judge your alignments. For example, try Expresso at http://www.tcoffee.org.

Strategy for assessment of alternative multiple sequence alignment algorithms

[1] Create or obtain a database of protein sequences for which the 3D structure is known. Thus we can define "true" homologs using structural criteria.

[2] Try making multiple sequence alignments with many different sets of proteins (very related, very distant, few gaps, many gaps, insertions, outliers).

[3] Compare the answers.

Multiple sequence alignment of genomic DNA

There are typically few sequences (up to several dozen), each having up to millions of base pairs. Adding more species improves accuracy.

Alignment of divergent sequences often reveals islands of conservation (providing “anchors” for alignment).

Chromosomes are subject to inversions, duplications, deletions, and translocations (often involving millions of base pairs). E.g. human chromosome 2 is derived from the fusion of two acrocentric chromosomes.

There are no benchmark datasets available.

Perspective: multiple sequence alignment (MSA)

- Many dozens of MSA programs have been introduced in recent years. None is optimal. Each offers unique strengths and weaknesses.
- Key methods include consistency-, iterative-, and structure-based multiple alignment.
- Alignment of genomic DNA presents specialized challenges and different sets of tools. MSA are readily available through genome browsers such as Ensembl, UCSC, and NCBI.