Chapter 6: Multiple Sequence Alignment

0

• 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

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

CLUSTAL W (1.83) multiple sequence alignment

									•		
beta globin		MVHLT <mark>P</mark>	EEKSA	VTALW	GKVNV	DEV	/GGEALGF	RLLVVY	PWTQRFF	ESFG-	47
myoglobin		MGLS <mark>D</mark>	GEWQL	VLNVW	GKVEA	DIPGH	IGQEVLIE	RLFKGH	PETLEKF	DKFK-	48
neuroglobin		MER	PEPEL	IRQSW	RAVSF	SPLEH	IGTVLFAF	RLFALE	PDLLPLF	QYNCR	47
soybean		MVAFTE	KQDAL	VSSSF	EAFKA	NIPQY	SVVFYTS	SILEKA	PAAKDLF	SFLA-	49
rice	MALVEDNN	AVAVSFSE	EQEAL	VLKSW.	AILKK	DSANI	ALRFFL	<b>KIFEVA</b>	PSASQMF	SFLR-	59
		:	:	: :				::	* *		
		,	$\nabla$						V	٦	
beta globin	DLSTPDAV	MGNPKVKA	HGKKV	LGAFS	DGLAH	ILDNLK	GTFATLS	3	EL <b>H</b> CDKL	HVDPE	102
myoglobin	HLKSEDEM	KASEDLKK	HGATV	LTALG	GILKK	KGHHE	CAEIKPLA	<i></i>	QSHATKHI	KI PVK	103
neuroglobin	QFSSPEDC:	LSSPEFLD	HIRKV	MLVID	AAVTN	VEDLS	SLEEYLA	ASL	GRKHRAV	GVKLS	104
soybean	NGVDPT	NPKLTG	HAEKL	FALVR	DSAGC	LKASC	TVVAD <mark>A</mark>	YT	GSVHAQK	AVTDP	101
rice	NSDVPL	EKNPKLKT	HAMSV	FVMTC	EAAAC	LRKAG	KVTVR <mark>D</mark> I	TLKRL	GATHLKY	GVGDA	117
			* .:	:	:		:	:		:	
										_	

beta globinNFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH------147myoglobinYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG154neuroglobinSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDGE----151soybeanQFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA------144riceHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE---166:::\*:

Note how the region of a conserved histidine  $(\mathbf{\nabla})$  varies depending on which of five prominent algorithms is used

# Praline

(a) Praline multiple sequence alignment

beta globin myoglobin neuroglobin soybean rice Consistency .....MVHLTPEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFES.FG .....MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK.FK .....MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR .....MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS.FL MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS.FL 0000000014265438257934573463364343624453686433\*35344\*50063

beta globin myoglobin neuroglobin soybean rice Consistency DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSEL..*H*CDKLH....VDP HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQS..HATKHK....IPV QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHRAVG....VKL A.NGVDP..TNPKLTGHAEKLFALVRDSAGQL.KASGTVVADAA....LGSVHAQKAVTD R.NSDVPLEKNPKLKTHAMSVFVMTCEAAAQL.RKAGKVTVRDTTLKRLGATHLKYGVGD 3166354224776653\*4368635424454451335634333542003335440000922

beta globin myoglobin neuroglobin soybean rice Consistency ENFRLLGNVLVCVLAHHF.GKEFTPPVQAAYQKVVAGVANALAHKYH..... KYLEFISECIIQVLQSKH.PGDFGADAQGAMNKALELFRKDMASNYKELGFQG SSFSTVGESLLYMLEKCL.GPAFTPATRAAWSQLYGAVVQAMSRGWD..GE.. PQFVVVKEALLKTIKAAV.GDKWSDELSRAWEVAYDELAAAIKKA..... AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE... 43744844498258542305336554454\*55465426446754322001000

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-FGmyoglobin------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FKneuroglobin------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCRsoybean------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LAriceMALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR

:

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

\*.

\*

::

beta globinNFRLLGNVLVCVLAHHFGKE-FTPPVQAAYQKVVAGVANALAHKYH-----myoglobinYLEFISECIIQVLQSKHPGD-FGADAQGAMNKALELFRKDMASNYKELGFQGneuroglobinSFSTVGESLLYMLEKCLGPA-FTPATRAAWSQLYGAVVQAMSRGWDGE----soybeanQFVVVKEALLKTIKAAVGDK-WSDELSRAWEVAYDELAAAIKKA------riceHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE----

::::::: \*.

## Probcons

(c) PROBCONS

rice

beta globin -----VHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG myoglobin -----GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK neuroglobin M------ERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR soybean M-----VAFTEKODALVSSSFEAFKANIPOYSVVFYTSILEKAPAAKDLFSF-LA rice MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASOMFSF-LR \*. : :: : ....

 $\nabla$ beta globin DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD---NLK---GTFATLSELHCDKLHVDP myoglobin HLKSEDEMKASEDLKKHGATVLTALGGI -- LKKKGHHE--- AEIKPLAOSHATKHKIPV neuroglobin OFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLE---EYLASLGRKHRAV-GVKL soybean NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVV----ADAALGSVHAQK-AVTD NSDVP--LEKNPKLKTHAMSVFVMTCEAAAOLRKAGKVTVRDTTLKRLGATHLKY-GVGD \*. \* . : : . . . ::

beta globin ENFRLLGNVLVCVLAHHF-GKEFTPPVOAAYOKVVAGVANALAHK-----YH myoglobin KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG neuroglobin SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRG---W-DGE soybean POFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIK-----KA rice AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKOE---MKPAE

: : : :: : . . :

## TCoffee

CLUSTAL FORMAT for T-COFFEE Version\_5.13

(d)

beta globin------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFE-SFGmyoglobin------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFD-KFKneuroglobin------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCRsoybean------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS-FLAriceMALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS-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

 ...
 \*
 ...

. . . . . .

# 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)

- 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

O

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. 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

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

STEP 1 - Enter your input sequences

Enter or paste a set of Protein v sequences in any supported format:

>beta\_globin 2hhbB NP\_000509.1 [Homo sapiens] MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESEGDLSTPDAVMGNPKVKAHGKKVLGAESDGLAHLDNLKGTFATLSELHCD KLHVDPENERLLGNVLVCVLAHHEGKEETPPVQAAYQKVVAGVANALAHKYH >myoglobin 2MM1 NP\_005359.1 [Homo sapiens] MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKEDKFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHAT KHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELERKDMASNYKELGFQG >neuroglobin 10J6A NP\_067080.1 [Homo sapiens] MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHR Or, upload a file:

STEP 2 - Set your Pairwise Alignment Options

Alignment Type: 
 Slow 
 Fast

#### Slow Pairwise Alignment Options

Protein Weight Matrix	GAP OPEN	GAP EXTENSION
Gonnet v	10 ~	0.1 ~

STEP 3 - Set your Multiple	Sequen	ce Alignment Options							
Protein Weight Matrix		GAP OPEN		GAP EXTENSION		GAP DISTANCES		NO END GAPS	
BLOSUM	v	10	Ý	0.20	v	5	¥	no	v
ITERATION		NUMITER		CLUSTERING					
none	¥	1	¥	NJ	Ý				
Output Options			ORDE	=R					

STEP 4 - Submit your job

Be notified by email (Tick this box if you want to be notified by email when the results are available)

Submit

B&FG 3e

Page 209

Fig. 6.1

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

(a) Stage 1: series of pairwise alignments

## ClustalW stage I: 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

best

score (highest percent pairwise identity)

B&FG 3e Fig. 6.2 Page 210 (a) Stage 1: series of pairwise alignments

### Clusta 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	<mark>15.65</mark>
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

# ClustalW stage 2:

create a guide tree<sup>(b) Stage 2: create a guide tree (calculated from a distance matrix)</sup>

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

( beta\_globin:0.36022, myoglobin:0.38808) :0.06560, neuroglobin:0.39924, ( soybean\_globin:0.30760, rice globin:0.26184)

:0.13652);

#### score (highest percent pairwise identity)

best

B&FG 3e Fig. 6.2 Page 210 
 beta\_globin: 0.36022

 myoglobin: 0.38808

 neuroglobin: 0.39924

 soybean\_globin: 0.30760

 rice\_globin: 0.26184

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
l	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
						N

best

score

Number of pairwise alignments needed

For *n* sequences, (n-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

beta_globin		-MVHL	TPE	EKSAV	/TALWGE	CVN VI	DEVGGEALG	RLLVVY	PWTQR	FFESFG-	47
myoglobin		MGL	SDGI	EWQLV	LNVWGE	(VEAD I F	PGHGQEVLI	RLFKGH	PETLE	KFDKFK-	48
neuroglobin		M	IERPI	<b>PEL</b>	RQSWRA	VSRSPI	LEHGTVLFA	RLFALE	PDLLP	L <mark>F</mark> QYNCR	47
soybean_globin		-MVAF	TEK	DAL	/SSSFE/	FKANI	QYSVVFYI	SILEKA	PAAKD	LFSFLA-	49
rice_globin	MALVEDNNA	VAVSF	SEE	ZEAL	/LKSWA]	LKKDSA	NIALRFFI	KIFEVA	APSASQ	MFSFLR-	59
		:		: :	: :			::	*	*.	
			•								
beta_globin	DLSTPDAVM	GNPKV	KAH	JKKVI	GAFSD	LAHLDI	NLKGTF <mark>AT</mark> -	I	SELHC	DKLHVDP	101
beta_globin myoglobin	DLST <mark>PDAVM</mark> HLKSEDEMK	GNPKV ASEDL	KKH( KAH(	SKKVI SATVI	LGAFSDO LTALGGI	LAHLDI	NLKGTF <mark>AT</mark> - IHEAEI <mark>KP</mark> -	I	,SELHC ⊿AQS <b>H</b> A	DKLHVDP TKHKIPV	101 102
beta_globin myoglobin neuroglobin	DLST <mark>PDAVMO</mark> HLKSEDEMKI QFSSPEDCLS	GNPKV ASEDI SSPEF	KKH(	GKKVI GATVI IRKVN	LGAFSDO LTALGGI 4LVIDA#	LAHLDI LKKKGH VTNVEI	NLKGTF <mark>AT</mark> - HEAEIKP - DLSSL <mark>EEY</mark> -		JSELHC JAQSHA JGRKHR	DKLHVDP TKHKIPV AVGVKLS	101 102 104
beta_globin myoglobin neuroglobin soybean_globin	DLST <mark>PDAVM</mark> HLKSEDEMK QFSSPEDCL NGVDPT-	GNPKV ASEDI SSPEF -NPKI	KAH KKH LDH	GKKVI GATVI IRKVN AEKLI	LGAFSDO LTALGGI 4LVIDA# FALVRDS	LAHLDI LKKKGH VTNVEI AGQLKI	NLKGTF <mark>AT</mark> - HEAEIKP - DLSSL <mark>EEY</mark> - <mark>AS</mark> GTVVAD -	I I AAI	▼ JSEL <b>H</b> C JAQS <b>H</b> A JGRK <b>H</b> R JGSV <b>H</b> A	DKLHVDP TKHKIPV AVGVKLS QKAVFDP	101 102 104 101
beta_globin myoglobin neuroglobin soybean_globin rice_globin	DLST <mark>PDAVM</mark> HLKSEDEMK QFSSPEDCL NGVDPT NSDVPLE	GNPKV ASEDL SSPEF -NPKL KNPKL	KAH KKH LDH TGH	GKKVI GATVI IRKVN AEKLI AMSVI	LGAFSDO LTALGGI 4LVIDA# FALVRDS FVMTCE#	LAHLDI LKKKGH VTNVEI AGQLK <i>I</i> AAQLRI	ILKGTFAT - IHEAEIKP - DLSSL <mark>EEY</mark> - ASGTVVAD - (AGKVTVRL	I LASI AAI )TTLKRI	▼ ISELHC IAQSHA IGRKHR IGSVHA IGSVHA	DKLHVDP TKHKIPV AVGVKLS QKAVIDP KYGVGDA	101 102 104 101 117

beta_globin	ENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH	147
myoglobin	KYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG	154
neuroglobin	SFSTVGESLLYMLEKCLG-PAFTPATRAAWSQLYGAVVQAMSRGWDGE	151
soybean_globin	QFVVVKEALLKTIKAAVG-DKWSDELSRAWEVAYDELAAAIKKA	144
rice_globin	HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE	166
	Next Control of the C	

. : : : \*. . :

B&FG 3e Fig. 6.3 Page 211

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

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

(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);
```

human\_NP\_000509: 0.00000 Pan\_troglodytes\_XP\_508242: 0.00000 Canis\_familiaris\_XP\_537902: 0.04932 Mus\_musculus\_NP\_058652: 0.12075 Gallus\_gallus\_XP\_444648: 0.21259

B&FG 3e Fig. 6.4 Page 212

## ClustalW alignment of five closely related beta globin orthologs

human\_NP\_000509 Pan\_troglodytes\_XP\_508242 Canis\_familiaris\_XP\_537902 Mus\_musculus\_NP\_058652 Gallus\_gallus\_XP\_444648

human\_NP\_000509 Pan\_troglodytes\_XP\_508242 Canis\_familiaris\_XP\_537902 Mus\_musculus\_NP\_058652 Gallus\_gallus\_XP\_444648

human\_NP\_000509 Pan\_troglodytes\_XP\_508242 Canis\_familiaris\_XP\_537902 Mus\_musculus\_NP\_058652 Gallus gallus XP 444648 TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD 100 TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD 100 TPDAVMSNAKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVD 100 SASAIMGNPKVKAHGKKVITAFNEGLKNLDNLKGTFASLSELHCDKLHVD 100 SPTAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVD 100 :. \*::.\*. \*:\*\*\*\*\*: :\*.::

PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147 PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147 PENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALAHKYH 147 PENFRLLGNAIVIVLGHHLGKDFTPAAQAAFQKVVAGVATALAHKYH 147 PENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH 147 \*\*\*\*:\*\*\*: ::: \*\*. \*::\*\*\* \*\*\*:\*\*\*

B&FG 3e Fig. 6.5 Page 213 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);
```

[	beta_globin: 0.36022
	neuroglobin: 0.39924
	sovbean: 0.30760
	rice: 0.26184

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
====			=====			======
1	human_NP_000509	147	2	<pre>Pan_troglodytes_XP_508242</pre>	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"?

0

- 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
PAMI20	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

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

[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

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

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  $y_i$ 

then  $x_i$  should align with  $y_i$ 

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

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

[I] 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.

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.