

Chapter 6: Multiple Sequence Alignment

Learning objectives

- Explain the three main stages by which ClustalW performs multiple sequence alignment (MSA);
- Describe several alternative programs for MSA (such as MUSCLE, ProbCons, and TCoffee);
- Explain how they work, and contrast them with ClustalW;
- Explain the significance of performing benchmarking studies and describe several of their basic conclusions for MSA;
- Explain the issues surrounding MSA of genomic regions

Outline: multiple sequence alignment (MSA)

Introduction; definition of MSA; typical uses

- Five main approaches to multiple sequence alignment

 - Exact approaches

 - Progressive sequence alignment

 - Iterative approaches

 - Consistency-based approaches

 - Structure-based methods

Benchmarking studies: approaches, findings, challenges

Databases of Multiple Sequence Alignments

 - Pfam: Protein Family Database of Profile HMMs

 - SMART

 - Conserved Domain Database

 - Integrated multiple sequence alignment resources

 - MSA database curation: manual versus automated

Multiple sequence alignments of genomic regions

 - UCSC, Galaxy, Ensembl, alignathon

Perspective

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

CLUSTAL W (1.83) multiple sequence alignment

```

beta globin  -----MVHLTPEEKSAVTALWGKVNVDD--EVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin   -----MGLSDGEWQLVLNVWVGKVEADIPGHGQEVLIQLFKGHPETLEKFDKFK- 48
neuroglobin -----MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALPDLLPLFQYNCR 47
soybean      -----MVAFTTEKQDALVSSSFEEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice         MALVEDNNAVAVSFSSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLR- 59
              :   :   :   :   . . .   .   ::   *   *
              ▼
beta globin  DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLS-----ELHCDKLHVDPE 102
myoglobin   HLKSEDEMKAISED LKKHGATVLTALGGILKKKGHHAEIKPLA-----QSHATKHKIPVK 103
neuroglobin QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLA---LGRKHRAVGVKLS 104
soybean      --NGVDPT--NPKLTGHA EKLFALVRDSAGQLKASGTVVADAA---LGSVHAQKAVTDP 101
rice         --NSDVPLEKNPKLKTHAMSVFVMTCEAAQAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117
              .   . . . *   .::   :   :   :
              ▼
beta globin  NFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH----- 147
myoglobin   YLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin SFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151
soybean      QFVVVKEALLKTIKAAV-GDKWSDLSRAWEVAYDELA AAIKKA----- 144
rice         HFEVVKFALLDTIKEEVPADMWS PAMKSAWSEAYDHLVAAIKQEMKPAE--- 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 globinMVHLTPEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFES.FG	▼
myoglobinMGLSDGEWQLVLNVWGKVEADIPGHGQEVLIIRLFKGH PETLEKFDK.FK	
neuroglobinMERPEPELIQSWRAVSRSPLEHGTVL FARLFALEPDLLPLFQYNCR	
soybeanMVAFT EKQDALVSSSFEAFKANI PQYSVVFYTSILEKAPAAKDLS..FL	
rice	MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS..FL	
Consistency	000000000014265438257934573463364343624453686433*35344*50063	
▽		
beta globin	DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL DNLKGT FATLSEL..HCDKLH....VDP	▼
myoglobin	HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQS..HATKHK....IPV	
neuroglobin	QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHRAVG....VKL	
soybean	A.NGVDP..TNPKLTGHA EKLFALVRDSAGQL.KASGTVVADAA....LGSVHAQKAVTD	
rice	R.NSDVPLEKNPKLKTHAMSVFVMTCEAAAQL.RKAGKVTVRDTTLKRLGATHLKYGVGD	
Consistency	3166354224776653*4368635424454451335634333542003335440000922	
▽		
beta globin	ENFRLLGNVLVCVLAHHF.GKEFTPPVQAAYQKV VAGVANALAHKYH.....	
myoglobin	KYLEFISECIIQVLQSKH.PGDFGADAQGAMNKALELFRKDMASNYKELGFQG	
neuroglobin	SSFSTVGESLLYMLEKCL.GPAFTPATRAAWSQLYGAVVQAMSRGWD..GE..	
soybean	PQFVVVKEALLKTIKAAV.GDKWSELSRAWEVAYDELA AAIKKA.....	
rice	AHFEVVKFALLDTIKEEVPADMWS PAMKSAWSEAYDHLVAAIKQEMKPAE...	
Consistency	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-FG
myoglobin   -----MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIIRLFKGHPETLEKFDK-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--LEKNPKLKTHAMSFVMTCEAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA

```

. . * .:: : :

```

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

```

: : :: : : * . . :

Probcons

(c)

PROBCONS

beta globin	M-----VHLT PEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
myoglobin	M-----GLS DGEWQLVLNVWGKVEADIPGHGQEVLIIRLFKGH PETLEKFDK-FK
neuroglobin	M-----ERPE PELIRQSWRAVSRS PLEHGT VLFA R LFAL EPDLLPLFQYNCR
soybean	M-----VAFTE EKQDALVSSSF EAFKAN IPQYSVVFY TSILEK APAAKD L SF-LA
rice	MALVEDNNAVAVSFS EEQEALVLKSWAILKKDSANIALRFFLKI FEVAPSASQMFSF-LR
	* * : : : : :: * * .

beta globin	DLST PDAVMGNPKVKAHGKKVLGAFSDGLAHLD--- NLK---GTF FATLSELHCD KLHVDP
myoglobin	HLKSEDEMKA SEDLKKHGATVLTALGGI--- LKKKGHE---AEI KPLAQSHAT KHKIPV
neuroglobin	QFSSPEDCLSS PEFLDHIRKVMLVIDAAVTN VEDLSSLE---EY LASLGRKHRAV -GVKL
soybean	NGVDP---- T NPKLTGHA EKL FALVRDSAGQLKAS GT VV---- AD AALGSVHAQK-A VT D
rice	NSDVP--LEKN PKLKTHAMSVFVMTCEAAAQLRKAGKV TVRDT TLKRLGATHLKY -GVGD
	. : . . . * . . . :: . * . * :

beta globin	ENFRLLGNVLVCVLAHHF -GKEFT PPVQAAYQKVVAGVANALAHK-----YH
myoglobin	KYLEFISECIIQVLQSKH -PGDFG ADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin	SSFSTVGESLLYMLEKCL-GPAFT PATRAAWSQLYGAVVQAMSRG--- W-DGE
soybean	PQFVVVKEALLKTIKAAV -GDKWS DELSRAWEVAYDELA AAIK-----KA
rice	AHFEVVKFALLDTIKEE VPADMWS PAMKSAWSEAYDHLVAAIKQE--- MKPAE
	: : :: : : * . . :

TCoffee

(d)

CLUSTAL FORMAT for T-COFFEE Version_5.13

```

beta globin  -----MVHLTPEEKSAVTALWGKVNV--EVGGEALGRLLVVYPWTQRFFESFG
myoglobin   -----MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIIRLFKGHPETLEKFD-KFK
neuroglobin -----MERPEPELIHQSWRAVSRSPLEHGTVLFARLFALPDLLPLFQYNCR
soybean      -----MVAFTTEKQDALVSSSFEEAFKANIPQYSVVFYTSILEKAPAAKDLS-FLA
rice         MALVEDNNAVAVSFS EEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS-FLR
              :   :   :   :   . . .   .   : :   *   * .

              ▽
beta globin  DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNL---KGTF---ATLSELHCDKLHVDP
myoglobin   HLKSEDEMKA SEDLKKHGATVLTAL---GGILKKKGHHEAE---IKPLAQSHATKHKIEV
neuroglobin QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDL---SSLEEYLASLGRKH-RAVGVKL
soybean     NGVDP----TNPKLTGHA EKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDLP
rice        NSDVP--LEKNPKLKTHAMSVFVMTCEAAQRLRKAGKVTVRD TTKRLGATHLKYGVGDA
              .   . . . * . : :           :           * . *

beta globin  ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH-----
myoglobin   KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDG----E
soybean     Q-FVVVKEALLKTIKAAV-GDKWSD ELSRAW EVAYDELA AAIKKA-----
rice        H-FEVVKFALLDTIKEEVPADMWS PAMKSAWSEAYDHLVAAIKQE---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

Outline: multiple sequence alignment (MSA)

Introduction; definition of MSA; typical uses

Five main approaches to multiple sequence alignment

- Exact approaches

- Progressive sequence alignment

- Iterative approaches

- Consistency-based approaches

- Structure-based methods

Benchmarking studies: approaches, findings, challenges

Databases of Multiple Sequence Alignments

- Pfam: Protein Family Database of Profile HMMs

- SMART

- Conserved Domain Database

- Integrated multiple sequence alignment resources

- MSA database curation: manual versus automated

Multiple sequence alignments of genomic regions

- UCSC, Galaxy, Ensembl, alignathon

Perspective

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.

Use ClustalW to do a progressive MSA

STEP 1 - Enter your input sequences

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

```
>beta_globin 2hhbB NP_000509.1 [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL
DNLKGTFTATLSELHCD
KLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
>myoglobin 2MM1 NP_005359.1 [Homo sapiens]
MGLSDGEWQLVLNVWGKVEADIPGHGQEVURLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKGH
FEAFIKPLAQSHAT
KHKIPVKYLEFISECIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQGG
>neuroglobin 1OJ6A NP_067080.1 [Homo sapiens]
MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVMVLVIDAAVTNVEDL
SSLEEYLA SLGRKHR
```

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

10

0.1

STEP 3 - Set your Multiple Sequence Alignment Options

Protein Weight Matrix

GAP OPEN

GAP EXTENSION

GAP DISTANCES

NO END GAPS

BLOSUM

10

0.20

5

no

ITERATION

NUMITER

CLUSTERING

none

1

NJ

Output Options

FORMAT

ORDER

Clustal w/ numbers

input

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)

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

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)

(a) Stage 1: series of pairwise alignments

ClustalW 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

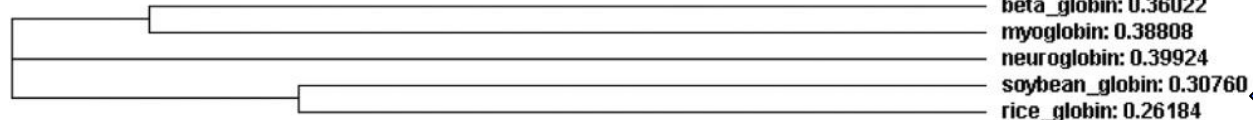


best
score

(highest percent
pairwise identity)

(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, $(n-1)(n) / 2$

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      -----MVHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin       -----MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIIRLFKGHPETLEKFDKFK- 48
neuroglobin     -----MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALPDLLPLFQYNCR 47
soybean_globin  -----MVAFTTEKQDALVSSSFSAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice_globin     MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKI FEVAPSASQMFSLR- 59
                  :   :   :   :   ..   .   ::   *   *.

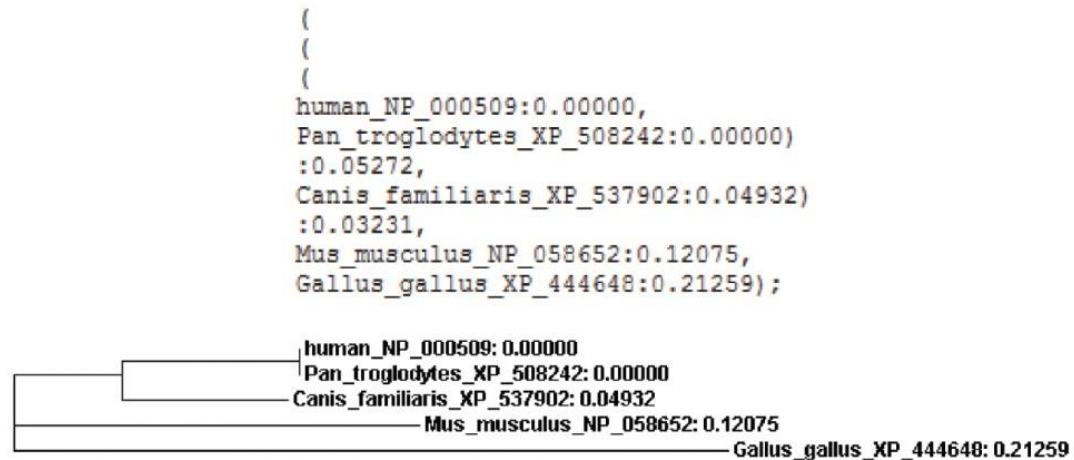
beta_globin      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL DNLKGT FAT-----LSELHCDKLHVDP 101
myoglobin       HLKSEDEMKA SEDLKKHGATVLTALGGILKKKGHHEAEIKP-----LAQSHATKHKIPV 102
neuroglobin     QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEY---LASLGRKHRAVGVKLS 104
soybean_globin  --NGVDPT--NPKLTGHA EKLFALVRDSAGQLKASGTVVAD---AALGSVHAQKAVTDP 101
rice_globin     --NSDVPLEKNPKLKT HAMS VFVMTCEAAAQLRKAGKVTVRD TTKRLGATHLKYGVGDA 117
                  .   .   ..   *   .::   :   *   *

beta_globin      ENFRL LGNVLVCVLAHHFGKEFTPPVQAAYQKV VAGVANALAHKYH----- 147
myoglobin       KYLEFI SECI IQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin     SFSTVGESLLYMLEKCLG-PAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151
soybean_globin  QFVVVKEALLKTIKAAVG-DKWSDELSRAWEVAYDELA AAIKKA----- 144
rice_globin     HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166
                  .   :   :   :   *   .   .   :
```


(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)



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

```

MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS 50
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS 50
MVHLTAEEKSLVSGLWGKVNVDEVGGEALGRLLIVYPWTQRFFDSFGDLS 50
MVHLTDAEKSAVSC LWAKVNPDEVGGEALGRLLVVYPWTQRYFDSFGDLS 50
MVHWTAEKQLITGLWGKVNAECGAELARLLIVYPWTQRFFASFGNLS 50
***  *   **  .  ::  **  ***   *  *  .  ***  .  ***  :  *****  :  *  ***  :  **

```

human_NP_000509
Pan_troglodytes_XP_508242
Canis_familiaris_XP_537902
Mus_musculus_NP_058652
Gallus_gallus_XP_444648

```

TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLEVD 100
TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLEVD 100
TPDAVMSNAKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLEVD 100
SASAIMGNPKVKAHGKKVITAFNEGLKNLDNLKGTFASLSELHCDKLEVD 100
SPTAILGNPMVRAHGKKVLITFSGDAVKNLDNIKNTFSQLSELHCDKLEVD 100
:.  *::.*  *::*****:  :*:::  :  :***:*  .**  :  *****

```

human_NP_000509
Pan_troglodytes_XP_508242
Canis_familiaris_XP_537902
Mus_musculus_NP_058652
Gallus_gallus_XP_444648

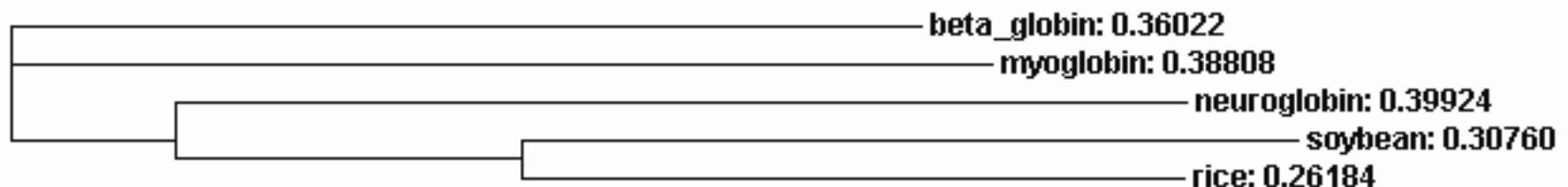
```

PENFRLLGNVLVCVLAHHFGKEFTPVQAAYQKVVAGVANALAHKYH 147
PENFRLLGNVLVCVLAHHFGKEFTPVQAAYQKVVAGVANALAHKYH 147
PENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALAHKYH 147
PENFRLLGNAIVIVLGHHLGKDFTPAAQAAFQKVVAGVATALAHKYH 147
PENFRLLGDILITVLAHAFSGKDFTEPCQAANQKLVRVVAHALARKYH 147
****:***:  ::  **  *::*::***  ****:*::*  **  ***:***

```


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

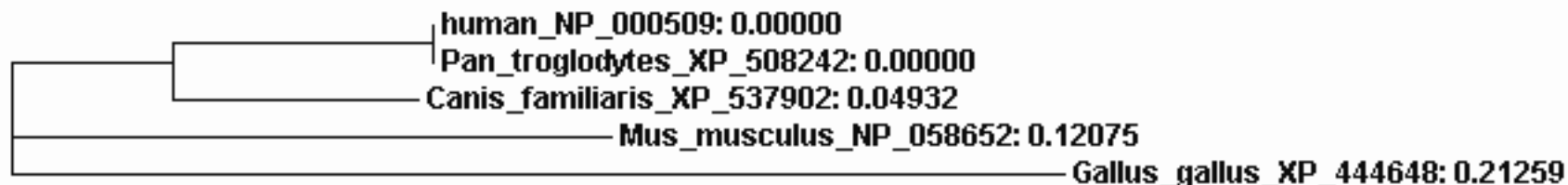
```
(  
  beta_globin:0.36022,  
  myoglobin:0.38808,  
  (  
    neuroglobin:0.39924,  
    (  
      soybean:0.30760,  
      rice:0.26184)  
    :0.13652)  
  :0.06560);
```



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

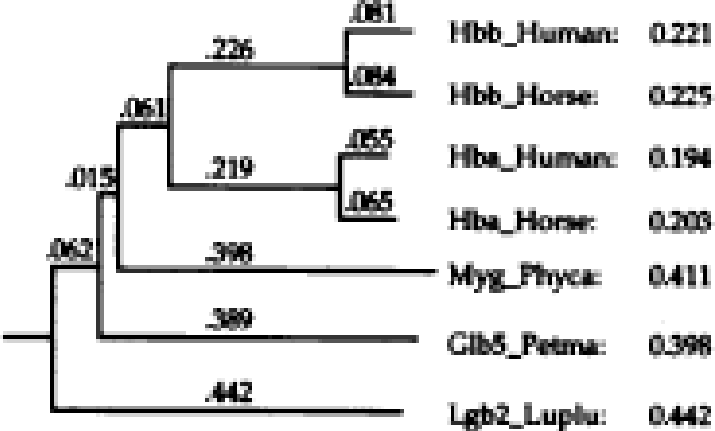
- Individual weights are assigned to sequences; very closely related sequences are given less weight, while distantly related sequences are given more weight
- Scoring matrices are varied dependent on the presence of conserved or divergent sequences, e.g.:

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

- Residue-specific gap penalties are applied

Rooted neighbor-joining
tree (guide tree) and
sequence weights

Progressive alignment: Align following the guide tree



```

-----VELTPEEAAVFAADQNT-----PSTVOGALGALLVVFSTQVFSTQDLST
-----VQLDDEEAAVFLALMDKV-----KEVVOGALGALLVVFSTQVFSTQDLST
-----VLSADKTVFLAANKYDAGAGETGAALEKMFLEFETKTVFPEFLS-----
-----VLSADKTVFLAANKYDAGAGETGAALEKMFLEFETKTVFPEFLS-----
-----VLSGKQGLVFLVMAKFAADVAGMOGDLIRLSTSESTLAFDFFQFLST
FIVDTGAVAPLAAEKTKIRANAFSTSTSTGVDILVFSTSTGAGFFPEFLSTT
-----GALTGGGALVFAAKH-----KAFKFGKTFILVLEDAKAFDFFPEFLSTSE
      *           *           *           *

```

PDAVWMPFVYKARCKEYVLAFFEDDCHIELD---HLCOTTFATLSEKCHDCLAVDPOHFFL
 PDAVWMPFVYKARCKEYVLAFFEDDCHIELD---HLCOTTFATLSEKCHDCLAVDPOHFFL
 ---HGLAOTFKRCKEYVADALFMAVAVD---DMFHALSALSOLMAKCLAVDPVHFFL
 ---HGLAOTFKARCKEYVADALFLAVCHIELD---DLFOLALHLSOLMAKCLAVDPVHFFL
 KAKHMAHMLKCKOVTVYLTALQALFLKFD---HMLAKLPLAQSHAFLKIKIPKYLEF
 ADQLKALDVWMAKRIINAVIDAVAMCOT---HMLAKLPLAQSHAFLKIKIPKYLEF
 VF---GMLHGLAARCKEYVLAFFEDDCHIELD---HLCOTTFATLSEKCHDCLAVDPOHFFL

LQWFLVGVLAARFQKSTFFVQAATQKVVAGVANDLRKTH-----
 LQWFLVGVLAARFQKSTFELQASTQKVVAGVAKALAKETH-----
 LQWCLLVTLAARLPAETTPAVHAILQWFLAVSVTVLSEKTR-----
 LQWCLLVTLAVLPHSTTPAVHAILQWFLAVSVTVLSEKTR-----
 ISEATINVLSEHPQDQGAQAQAMHIALELFKDIAATKELQYQ-----
 LAAVIADTVAAQ-----DAQFELAMNICILLASAT-----
 VQEAILETIEKVGAAKWEELNAYTLATKELATVIEKMDAA-----

Outline: multiple sequence alignment (MSA)

Introduction; definition of MSA; typical uses

Five main approaches to multiple sequence alignment

- Exact approaches

- Progressive sequence alignment

- Iterative approaches

- Consistency-based approaches

- Structure-based methods

Benchmarking studies: approaches, findings, challenges

Databases of Multiple Sequence Alignments

- Pfam: Protein Family Database of Profile HMMs

- SMART

- Conserved Domain Database

- Integrated multiple sequence alignment resources

- MSA database curation: manual versus automated

Multiple sequence alignments of genomic regions

- UCSC, Galaxy, Ensembl, alignathon

Perspective

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

Iterative approaches: MAFFT

MAFFT version 6

Multiple alignment program for amino acid or nucleotide sequences

[Download version](#)

[Mac OS X](#)

[Windows](#)

[Linux](#)

[Source](#)

[Usage](#)

Online version

Alignment

[Phylogeny](#)

[Merits and limitations](#)

[Algorithms](#)

[Tips](#)

[Aligning large data](#)

[Mafft-homologs](#)

[Benchmarks](#)

[Feedback](#)



Contact address
has changed!!

kkato@
kuicr.kyoto-u.ac.jp



kato@
bioreg.kyushu-u.ac.jp

Multiple sequence alignment and NJ / UPGMA phylogeny

Input:

Paste protein or DNA sequences in fasta format. [Example](#)

```
>gi|55743122|ref|NP_006735.2| retinol-binding protein 4, plasma precursor
MKVWVALLLLAALGSGRAERDCRVSSFRVKNFDFKARFSGTWYAMAKKDPEGLFLQDNIAEFSVDETGQ
MSATAKGRVRLNNWVDCADMVGTFTDTEPAKFMMKYWGVASFQKGNDDHWIVDTDYDTYAVQYSCRL
LNLDTGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL

>gi|12843160|dbj|BAB25881.1| unnamed protein product [Mus musculus]
MEWVWVALLAALGGGSAERDCRVSSFRVKNFDFKARFSGTWYAMAKKDPEGLFLQDNIAEFSVDEKGH
MSATAKGRVRLNNWVDCADMVGTFTDTEPAKFMMKYWGVASFQKGNDDHWIIDTDYDTYAVQYSCRL
QNLDTGTCADSYSFVFSRDPNGLSPETRLRVRQRQEELCLERQYRWIEHNGYCDGRSRNSL

>gi|4502163|ref|NP_001638.1| apolipoprotein D precursor [Homo sapiens]
MVMLLLLSALAGLFGAAEGQAFHLGKCPNPFVQENFDVNYLGRWYEIEKIPTTFENGRCIQANYSLME
NGKIKVLNQELRADGTQVNIQEGEATFVNLTEPAKLEVKFSWFMPSPAFYNYLATDYENYALVYSTCIIQL
FHVDFANILARNPNLPPETVDSLKNILTSNNIDVKKMTVTDQVNCPLKS
```

or upload a file:

[Browse...](#)

[Use structural alignment\(s\)](#)

Output order:

- ☐ Same as input
- ☒ Aligned

Notify when finished (optional; recommended when submitting large data):

Email address:

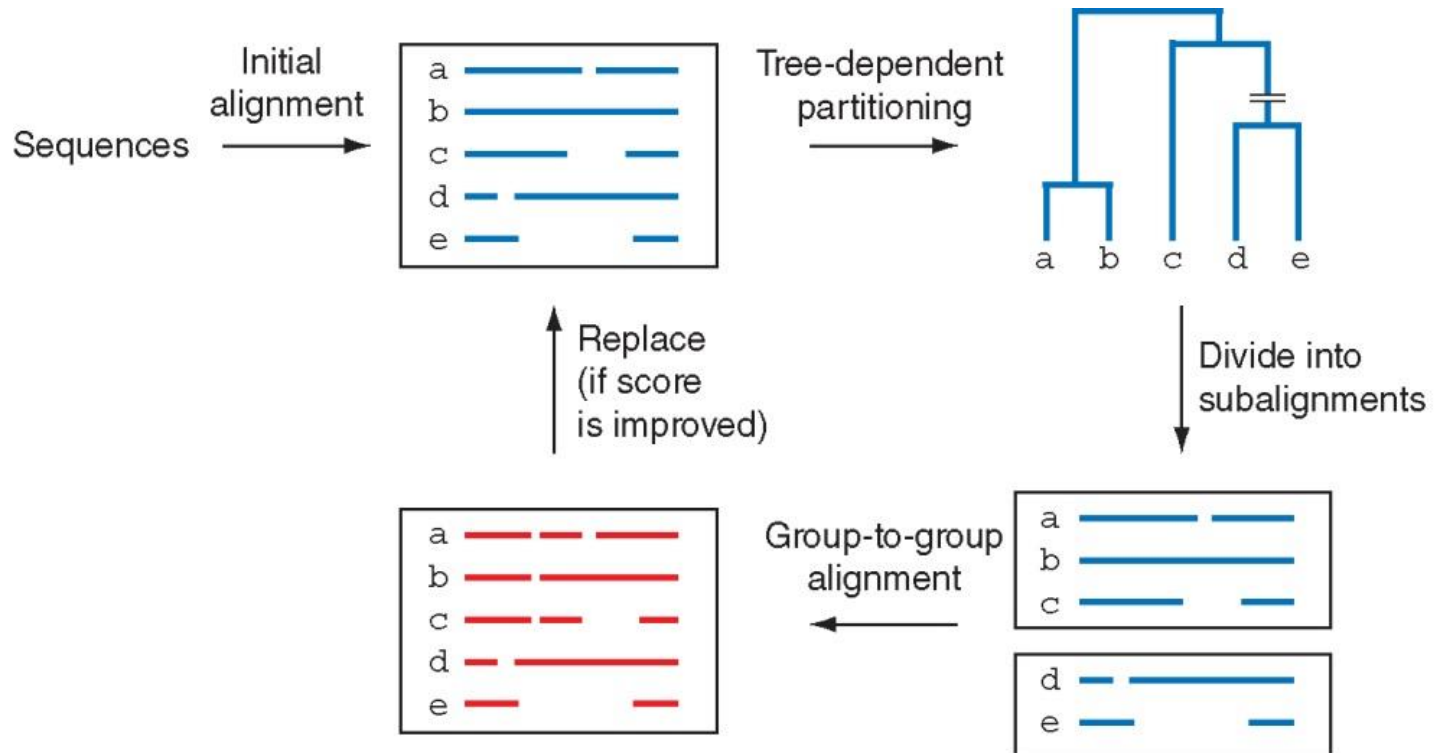
[Submit](#)

[Reset](#)

[Advanced settings](#)

Has about 1000
advanced settings!

Iterative method of MAFFT



MAFFT

(a) Alignment of nine globins by MAFFT FFT-NS-2 (v7.058b) (DSSP colors: **turn**, **alpha helix**, **bend**, 3/10 helix)

```

hbb_human      -----MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVPWTQRFFE-SFG
hbb_chimp      -----MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVPWTQRFFE-SFG
hbb_dog        -----MVHLTAEEKSLVSGLWGKVNVD--EVGGEALGRLLIVPWTQRFFD-SFG
hbb_mouse      -----MVHLTDAEKSAVSCLWAKVNPD--EVGGEALGRLLVVPWTQRFYFD-SFG
hbb_chicken    -----MVHWTAEEKQLITGLWGKVNVA--ECGAEALARLLIVPWTQRFFA-SFG
myoglobin      -----MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPTLEKFD-KFK
neuroglobin    -----MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean        -----MVAFTEKQDALVSSSFEAFKANIPQSVVFYTSILEKAPAAKDLS-FLA
rice           MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS-FLR
               :   :   :   .   .   .   :   :   *   *
               ▼2               ▼3
hbb_human      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAH---LDNL---KGTFATLSELHCDKLHVDP
hbb_chimp      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAH---LDNL---KGTFATLSELHCDKLHVDP
hbb_dog        DLSTPDAVMSNAKVKAHGKKVLNSFSDGLKN---LDNL---KGTFAKLSELHCDKLHVDP
hbb_mouse      DLSSASAIMGNPKVKAHGKKVITAFNEGLKN---LDNL---KGTFASLSELHCDKLHVDP
hbb_chicken    NLSSPTAILGNPMVRAHGKKVLTSFGDAVKN---LDNI---KNTFSQLSELHCDKLHVDP
myoglobin      HLKSEDEMKASEDLKKHGATVLTALGGILKK---KGHH---EAEIKPLAQSHATKHKIPV
neuroglobin    QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSL---EEYLASLGRKH-RAVGVKL
soybean        NGVDP---TNPKLTGHAEKLFALVRDSAQLKASGTV-VADAA---LGSVH-AQKAVTD
rice           NSDVP---LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATH-LKYGVGD
               .   .   .   *   .   :   :   .   .   .   *   *   :

```

(b) Alignment of nine globins by MUSCLE (3.8)

MUSCLE

```

hbb_human      -----MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVPWTQRFFE-SFG
hbb_chimp      -----MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVPWTQRFFE-SFG
hbb_dog        -----MVHLTAEEKSLVSGLWGKVNVD--EVGGEALGRLLIVPWTQRFFD-SFG
hbb_mouse      -----MVHLTDAEKSAVSCLWAKVNPD--EVGGEALGRLLVVPWTQRFYFD-SFG
hbb_chicken    -----MVHWTAEEKQLITGLWGKVNVA--ECGAEALARLLIVPWTQRFFA-SFG
myoglobin      -----MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPTLEKFD-KFK
neuroglobin    -----MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean        -----MVAFTEKQDALVSSSFEAFKANIPQSVVFYTSILEKAPAAKDLS-FLA
rice           MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS-FLR
               :   :   :   .   .   .   :   :   *   *
               ▼2               ▼3
hbb_human      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL---DNLKGTFATLSELHCDK--LHVDPE
hbb_chimp      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL---DNLKGTFATLSELHCDK--LHVDPE
hbb_dog        DLSTPDAVMSNAKVKAHGKKVLNSFSDGLKNL---DNLKGTFAKLSELHCDK--LHVDPE
hbb_mouse      DLSSASAIMGNPKVKAHGKKVITAFNEGLKNL---DNLKGTFASLSELHCDK--LHVDPE
hbb_chicken    NLSSPTAILGNPMVRAHGKKVLTSFGDAVKNL---DNIKNTFSQLSELHCDK--LHVDPE
myoglobin      HLKSEDEMKASEDLKKHGATVLTALGGILKKK---GHHEAEIKPLAQSHATK--HKIPVK
neuroglobin    QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNV---EDLSSLEEYLASLGRKHRAVGVKLS
soybean        NGVDPT---NPKLTGHAEKLFALVRDSAQL---KASGTVVADAALGSVHAQKAVTDP
rice           NSDVP---LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
               .   .   *   .   :   :   :   :   :   :

```


ProbCons

(c) Alignment of nine globins by ProbCons (version 1.12)

```

hbb_human      M-----VHLTPEEKSAVTALWGKNVD--EVGGEALGRLLVVYPWTORFFES-FG
hbb_chimp      M-----VHLTPEEKSAVTALWGKNVD--EVGGEALGRLLVVYPWTORFFES-FG
hbb_dog        M-----VHLTAEKSLVSGLWGKNVD--EVGGEALGRLLIYYPWTORFFDS-FG
hbb_mouse      M-----VHLTDAEKSAVSLWAKVNDP--EVGGEALGRLLVVYPWTORFYDS-FG
hbb_chicken    M-----VHWTAEKQLITGLWGKNVA--ECGAEALARLLIYYPWTORFFAS-FG
myoglobin      M-----GLSDGEWQLVLNVWGKVEADIPGHGQEVLRIRLFKHPETLEKFDK-FK
neuroglobin    M-----ERPEPELIROSWRAVSRSPLHGTVLFARLFALEPDLLPLFQYNCR
soybean        M-----VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice           MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKI FEVAPSASQMFSLR
*               *       :       :       :       .       :       *       *

hbb_human      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNL-----KGTFFATLSELHCDKLHVP
hbb_chimp      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNL-----KGTFFATLSELHCDKLHVP
hbb_dog        DLSTPDAVMSNAKVKAHGKKVLNSFSDGLKNDNL-----KGTFAKLSELHCDKLHVP
hbb_mouse      DLSSASAIMGNPKVKAHGKKVITAFNEGLKNDNL-----KGTFFASLSELHCDKLHVP
hbb_chicken    NLSSPTAILGNPMVRAHGKKVLTSFGDAVKNDNI-----KNTFSQLSELHCDKLHVP
myoglobin      HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHH-----EAEIKPLAQSHATKHKIPV
neuroglobin    QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSL-----EYLASLGRKHRAV-GVKL
soybean        NGVDP---TNPKLTGHAELKLFALVRDSAGQLKASGTV---V-ADAALGSVHAQK-AVTD
rice           NSDVP--LEKNPKLKTAMSVFVMTCEAAQRLKAGKVTVRDRTLKRLGATHLKY-GVGD
.               .       .       *       .       .       :       .       .       *       *       :

```

(d) Alignment of nine globins by T-COFFEE (Expresso version_10.00)

T-COFFEE

```

1HBB      1 -----MVHLTPEEKSAVTALWGKNV--VDEVGGEALGRLLVVYPWTORFFESFGD--LSTPDAVM
hbb_chimp 1 -----MVHLTPEEKSAVTALWGKNV--VDEVGGEALGRLLVVYPWTORFFESFGD--LSTPDAVM
2q15B     1 -----MVHLTAEKSLVSGLWGKNV--VDEVGGEALGRLLIYYPWTORFFDSFGD--LSTPDAVM
3hrwB     1 -----MVHLTDAEKSAVSLWAKVN--PDEVGGEALGRLLVVYPWTORFYDSFGD--LSSASAIM
1hbrB     1 -----MVHWTAEKQLITGLWGKNV--VAECGAEALARLLIYYPWTORFFASFGN--LSSPTAIL
3RGK      1 -----MGLSDGEWQLVLNVWGKVEADIPGHGQEVLRIRLFKHPETLEKFDKFKH--LKSEDEMK
10j6A     1 -----MERPEPELIROSWRAVSRSPLHGTVLFARLFALEPDLLPLFQYNCRQFSSPEDCL
1FSL      1 -----MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS---FLANGVDP
1D8U      1 MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKI FEVAPSASQMFSL--L-RNSDVPLE

cons      1      :       :       :       .       .       *       *       :       :

hbb_human  57  GNPVKVKAHGKKVLGAFSDGLAHL--DNL-----KGTFFATLSELHCDKLHVPENFRLLGNVLCVLAHHFG
hbb_chimp  57  GNPVKVKAHGKKVLGAFSDGLAHL--DNL-----KGTFFATLSELHCDKLHVPENFRLLGNVLCVLAHHFG
2q15B     57  SNAKVKAHGKKVLNSFSDGLKNL--DNL-----KGTFAKLSELHCDKLHVPENFKLLGNVLCVLAHHFG
3hrwB     57  GNPVKVKAHGKKVITAFNEGLKNL--DNL-----KGTFFASLSELHCDKLHVPENFRLLGNVLCVLAHHFG
1hbrB     57  GNPVRAHGKKVLTSFGDAVKNL--DNI-----KNTFSQLSELHCDKLHVPENFRLLGDILIIVLAHFS
3RGK      58  ASEDLKKHGATVLTALGGILKKK--GHH-----EAEIKPLAQSHATKHKIPVKYLEFISECIIIVLOSHP
10j6A     57  SSPEFLDHIRKVMLVIDAAVTNV--EDL--SSL--EYLASLGRKHRAV--AVGKLSFSSTVGESLLYMEKCLG
1FSL      55  TNPKLTGHAELKLFALVRDSAGQLKASG---TVVADALGSVHAQKAVTDPOFVVVKEALLKTIKAAVG
1D8U      67  KNPKLKTAMSVFVMTCEAAQRLKAGKVTVRDRTLKRLGATHLKYGVGDAHFEVVKFALLDTIKEEVP

cons      70  .       .       *       .       .       :       :       :       :       :       :

```


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

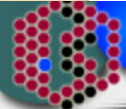
- Make profiles of each half of the tree

- Re-align the profiles

- Accept/reject the new alignment

Access to MUSLCE at EBI

<http://www.ebi.ac.uk/muscle/>

**EMBL-EBI**
European Bioinformatics Institute

Get N

EBI HomeAbout EBIGroupsServicesToolboxDatabasesDownloadsSubmissions


SEQUENCE ANALYSIS

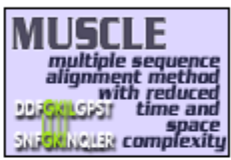
- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- Muscle Help
- Jalview Help
- Guide Tree
- Alignment
- Colours

- Similar Applications
 - ▶ ClustalW
 - ▶ T-Coffee

MUSCLE Submission Form

MUSCLE stands for **M**ultiple **S**equences **C**omparison by **L**og-**E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than CLUSTALW or T-Coffee, depending on the chosen options.

 [Download Software](#)



EMAIL	RESULTS	ALIGNMENT TITLE	OUTPUT FORMAT	OUTPUT TREE
<input type="text"/>	<input type="text" value="interactive"/>	<input type="text" value="Sequence"/>	<input type="text" value="fasta"/>	<input type="text" value="none"/>

Enter or Paste a set of Sequences in any supported format:

Help

Upload a file:

Browse...

Run

Reset

Outline: multiple sequence alignment (MSA)

Introduction; definition of MSA; typical uses

Five main approaches to multiple sequence alignment

- Exact approaches

- Progressive sequence alignment

- Iterative approaches

- Consistency-based approaches

- Structure-based methods

Benchmarking studies: approaches, findings, challenges

Databases of Multiple Sequence Alignments

- Pfam: Protein Family Database of Profile HMMs

- SMART

- Conserved Domain Database

- Integrated multiple sequence alignment resources

- MSA database curation: manual versus automated

Multiple sequence alignments of genomic regions

- UCSC, Galaxy, Ensembl, alignathon

Perspective

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.

ProbCons output for the same alignment: consistency iteration helps

(c)

PROBCONS

```

beta globin  M-----VHLTPEEKSAVTALWGKVNVDD--EVGGEALGRLLVVYPWTQRFFES-FG
myoglobin   M-----GLSDGEWQLVLNVWGKVEADIPGHGQEVLIIRLFKGHPETLEKFDK-FK
neuroglobin M-----ERPEPELIHQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean      M-----VAFTEKQDALVSSSFEEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice         MALVEDNNAVAVSFS EEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
*           * : : : : : . . . : : * *

```

```

beta globin  DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD---NLK---GTFATLSELHCDKLHVDP
myoglobin   HLKSEDEMKA SEDLKKHGATVLTALGGI---LKKKGHHE---AEIKPLAQSHATKHKIPV
neuroglobin QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLE---EYLASLGRKHRV-GVKL
soybean      NGVDP----TNPKLTGHA EKLFALVRDSAGQLKASGTVV---ADAAALGSVHAQK-AVTD
rice         NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKY-GVGD
.           : . . . * . . : : . . . * . * :

```

```

beta globin  ENFRLLG NVLVCVLAH HF-GKEFTPPVQAAYQKVVAGVANALAHK-----YH
myoglobin   KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRG---W-DGE
soybean      PQFVVVKEALLKTIKAAV-GDKWSELSRAW EVAYDELA AAIK-----KA
rice         AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQE---MKPAE
:           : : : : : : * . . :

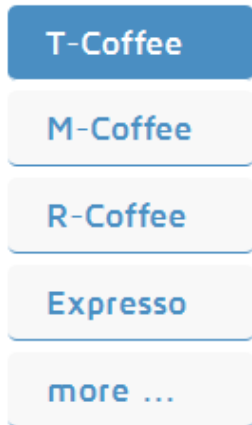
```




A collection of tools for Computing, Evaluating and Manipulating Multiple Alignments of DNA, RNA, Protein Sequences and Structures

T-Coffee Server

Quick links to the most popular T-Coffee modes:



Other T-Coffee links

[Documentation](#)

[Downloads](#)

[Support & discussion group](#)

Access to T-Coffee:
<http://tcoffee.org>

- Make a MSA
- MSA w. structural data
- Compare MSA methods
- Make an RNA MSA
- Combine MSA methods
- Consistency-based
- Structure-based

APDB ClustalW output:

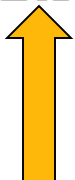
TCoffee can incorporate structural information into a MSA

```
T-COFFEE, Version 4.71(Thu Nov 16 15:08:43 2006)
Cedric Notredame
CPU TIME:0 sec.
# APDB Evaluation: Color Range Blue-[0 % -- 100 %]-Red
# Sequence Score: APDB
# Local Score: APDB

SCORE=47
*
  BAD  AVG  GOOD
*
2hbbB  : 224
1V5HA  : 213
2MM1   : 219
1OJ6A  : 194
1FSL   : 157

2hbbB  -----MVHLTPEEKSAVTALWG--KVVNDEVGGEALGRLLVVYP
1V5HA   MEKVPGEMEIERERSEELSEAERKAVQAMWRLYANCEDVGVAILLVRFFVNFP
2MM1    -----MGLSDGEWQLVLNVWGKVEADIPGHGQEVLLIRLFKGHP
1OJ6A   -----MERPEPELIQSWRAVSRSPLEHGTVLFARLFALEP
1FSL    -----MVAFTEKQDALVSSSFEAFKANIPQYSVVFFYTSILEKAP

          :   :   :   :   .   .   ::  *
```



Protein Data Bank accession numbers

Outline: multiple sequence alignment (MSA)

Introduction; definition of MSA; typical uses

Five main approaches to multiple sequence alignment

- Exact approaches

- Progressive sequence alignment

- Iterative approaches

- Consistency-based approaches

- Structure-based methods

Benchmarking studies: approaches, findings, challenges

Databases of Multiple Sequence Alignments

- Pfam: Protein Family Database of Profile HMMs

- SMART

- Conserved Domain Database

- Integrated multiple sequence alignment resources

- MSA database curation: manual versus automated

Multiple sequence alignments of genomic regions

- UCSC, Galaxy, Ensembl, alignathon

Perspective

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 Espresso 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.

Name hiv-1 protease

Number of sequences 4
Alignment Length 106
Longest Sequence 104
Shortest Sequence 98
Average Percent Identity 49
Maximum Percent Identity 86
Minimum Percent Identity 35

Sequence Name SWISSPROT Accession
1fmb P32542
7upjB P03366
pol_sivcz P17283
POL_SIVMK P05897

Family 1fmb 7upjB pol_sivcz POL_SIVMK

1fmb 1 vTYNLEKRPTTIVLINDTPLNVLLDTGADTSVLTTahynrlkyrgrk.YQ
7upjB 1 pQFSLWKRPVVTAYIEGQPVEVLLDTGADDSIVAG.....iel.gnn.YS
pol_sivcz 1 pQITLWQRPLIPVKVEGQLCEALLDTGADDTVIER.....iqlggl..WK
POL_SIVMK 1 pQFSLWRRPVVTAHIEGQPVEVLLDTGADDSIVTG.....iel.gph.YT

1fmb 50 GTGIGGVGGNVETFS.TPVTIKKKGRHIKTRMLVADIPVTILGRDILQDL
7upjB 44 PKIVGGIGGFINTLEYKNVEIEVLNKKVRATIMTGDTPINIFGRNILTAL
pol_sivcz 44 PKMIGGIGGF IKVKQFDNVHIEIEGRKVVGTVLVGPTPVNIIGRNILTQ
POL_SIVMK 44 PKIVGGIGGFINTKEYKNVEIEVLGKRIKRTIMTGDTPINIFGRNLLTAL

1fmb 99 GAKLV1
7upjB 94 GMSLN1
pol_sivcz 94 GCTLV.
POL_SIVMK 94 GMSLN1

Key

alpha helix RED
beta strand GREEN
core blocks UNDERSCORE

BaliBase: comparison of multiple sequence alignment algorithms

Multiple sequence alignment: methods

Benchmarking tests suggest that ProbCons, a consistency-based/progressive algorithm, performs the best on the BAliBASE set, although MUSCLE, a progressive alignment package, is an extremely fast and accurate program.

ClustalW has been the most popular program. It has a nice interface (especially with ClustalX) and is easy to use. But several programs perform better. There is no one single best program to use, and your answers will certainly differ (especially if you align divergent protein or DNA sequences)

Outline: multiple sequence alignment (MSA)

Introduction; definition of MSA; typical uses

Five main approaches to multiple sequence alignment

- Exact approaches

- Progressive sequence alignment

- Iterative approaches

- Consistency-based approaches

- Structure-based methods

Benchmarking studies: approaches, findings, challenges

Databases of Multiple Sequence Alignments

- Pfam: Protein Family Database of Profile HMMs

- SMART

- Conserved Domain Database

- Integrated multiple sequence alignment resources

- MSA database curation: manual versus automated

Multiple sequence alignments of genomic regions

- UCSC, Galaxy, Ensembl, alignathon

Perspective



Pfam
keyword search

34 architectures 6000 sequences 5 interactions 2886 species 1971 structures

enter ID/acc

We make a range of alignments for each Pfam-A family. You can see a description of each [above](#). You can view these alignments in various ways but please note that some types of alignment are never generated while others may not be available for all families, most commonly because the alignments are too large to handle.

	Seed (73)	Full (6000)	Representative proteomes				NCBI (5331)	Meta (34)
			RP15 (348)	RP35 (594)	RP55 (949)	RP75 (1261)		
Jalview	✓	✓	✓	✓	✓	✓	✓	✓
HTML	✓	—	✓	✓	✓	✓	✗	✗
PP/heatmap	X ₁	—	✓	✓	✓	✓	✗	✗
Pfam viewer	✓	✓	✗	✗	✗	✗	✗	✗

¹Cannot generate PP/Heatmap alignments for seeds; no PP data available

Key: ✓ available, ✗ not generated, — not available.

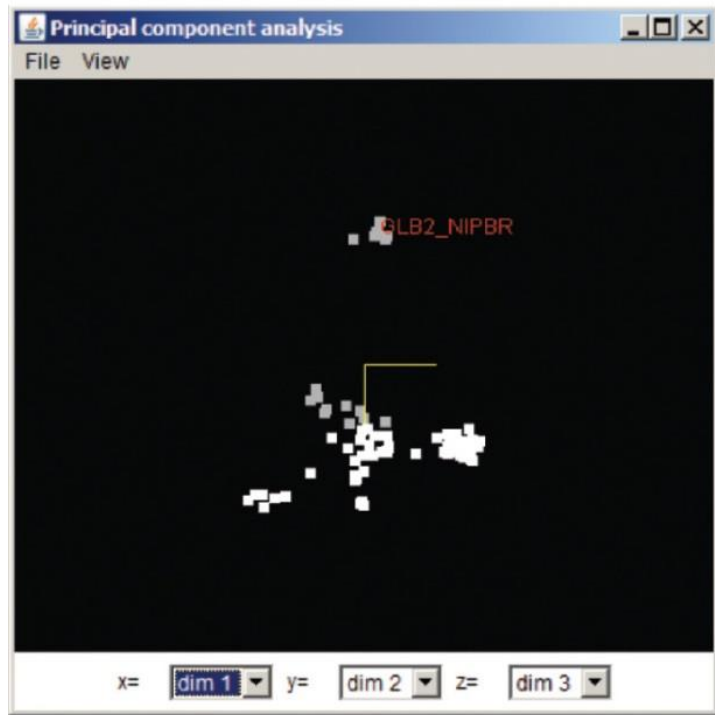
Seed sequence alignment for PF00042

Q20638 CAEL/74-184
Q19601 CAEL/105-215
Q18311 CAEL/32-140
GLB4_LUMTE/11-120
GLB4_LUMTE/11-120 (SS)
GLB3_TYLHE/8-117
GLB4_TYLHE/8-117
GLB1_TYLHE/7-110
GLB2_TYLHE/9-115
GLB2_LUMTE/8-114
GLB2_LUMTE/8-114 (SS)
GLE_TUBTU/6-112
GLB3_LAMSP/7-113

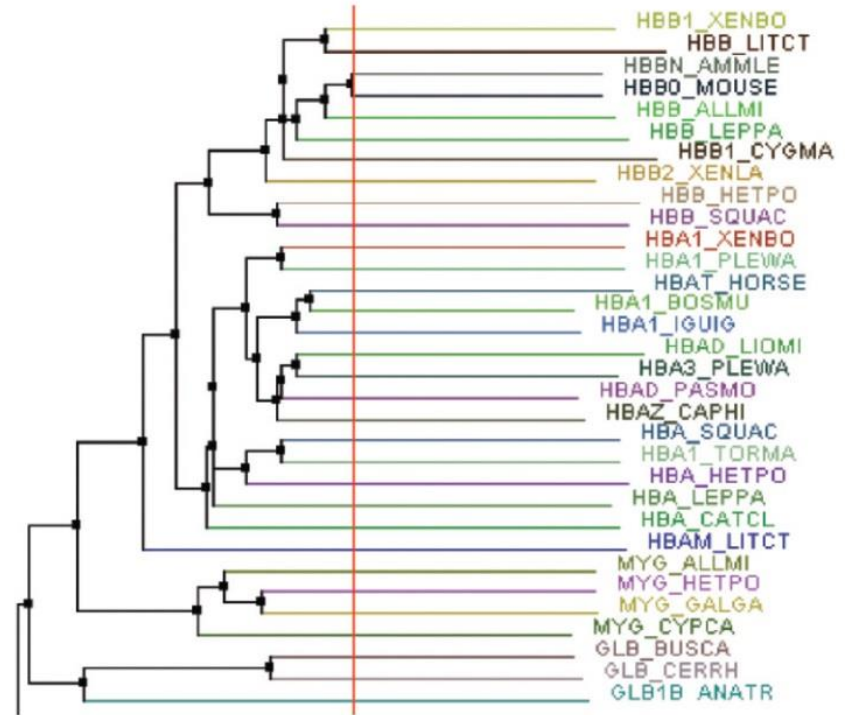
EKELLRRITWSD.EFD.....NLYELGSAITYCYIFDHNPNCKQLFP.F.ISKYQGEDEWKESKEFRSQALKFVQTLAQVVK
 ERILLEQSVNRK.TRK.....TGADHIGSKIFFMVLTAQPDIAKIFG.L.EKIPTRGLKYDPRFRQHALVVTKLDFVIR
 TKKLVIQENFR.VLA.....QCPELFTEIHWKSAITRSTSIKALFG.I.AE.N.ESEPMQNAFLGLSSTIQAFFYKLIID
 DRREIRHIWDD.VWSSS.FIDRRVAIVRAVFDLKFHYTSIKALFERVKIDF.....ESEKSHLVRVANGLLDLIN
 HHHHHHHHHHHS.-S.-S.SCHHHHHHHHHHHHHHHHHHSGGGGGGGGCCCTTIST.....TSSHHHHHHHHHHHHHHHHHFT
 DRHEVLDNWKG.IWSAE.FIGRRVAIGQAIQFELFALDPNAKGVFGVNV.D.K....PSEADKKAHVIRVINGLDLAVN
 DRREVQALNRS.IWSAE.DIGRRITLIGRLLFEELFEIDGATKGLFKRVNVDDT.....HSPPEFAHVLRVNGLDITLIG
 QRIKVKQQAQ.VYSV...GESRIDFALDVFNFNFRINPDRS.LFNRVNNDV.....YSPPEFAHMRVVFAGFDILIS
 QRLKVKQQAQ.AYGV...GHERVELGIALWKSMAQDNDARDLFRKHGVEDV.....HSPAFEAHMRVVFAGFDLRVIS
 EQLVKSENGR.AYGS...GHDREAFSQAIWRATFAQVPESRSLEKRVHGGDT.....SHPAFIAHAERVLGGLDIATIS
 HHHHHHHHHHHS.-S.-S.HHHHHHHHHHHHHHHHHHGGGGGGGGGGGTTT-T.....TSHHHHHHHHHHHHHHHHHHC
 QRFKVKHQNAE.AFGT...SHHRLDFGLKLNWISIFRDAPEIRGLFKRVGDG.N.....AYSSEFEAHAERVLGGLDMTIS
 ORLKVKRONAE.AYGS...GNDREFFGHITHTVKDAPASDLRKLFRVGNID.....HTPEFAHAHTRVLGGLDMTIS

Pfam alignment retrieved in the JalView Java viewer

(a) Principal components analysis (PCA)



(b) Neighbor-joining tree



Databases on which Interpro (release 51.0) is based

Database	Contents (entries)
PANTHER 9.0	60,000
Pfam 27.0	14,800
PIRSF 3.01	3,300
PRINTS 42.0	2,000
ProDom 2006.1	1,900
PROSITE 20.105 patterns	1,300
PROSITE 20.105 profiles	1,100
SMART 6.2	1,000
TIGRFAMs 15.0	4,500
CATH-Gene3D 3.5.0	2,600
SUPERFAMILY 1.75	2,000
UniProtKB 2015_04	47,300,000
UniProtKB/Swiss-Prot 2015_04	531,000
UniProtKB/TrEMBL 2015_04	46,715,000
GO Classification	27,000

http://www.ebi.ac.uk/interpro/release_notes.html

Outline: multiple sequence alignment (MSA)

Introduction; definition of MSA; typical uses

Five main approaches to multiple sequence alignment

- Exact approaches

- Progressive sequence alignment

- Iterative approaches

- Consistency-based approaches

- Structure-based methods

Benchmarking studies: approaches, findings, challenges

Databases of Multiple Sequence Alignments

- Pfam: Protein Family Database of Profile HMMs

- SMART

- Conserved Domain Database

- Integrated multiple sequence alignment resources

- MSA database curation: manual versus automated

Multiple sequence alignments of genomic regions

- UCSC, Galaxy, Ensembl, alignathon

Perspective

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.

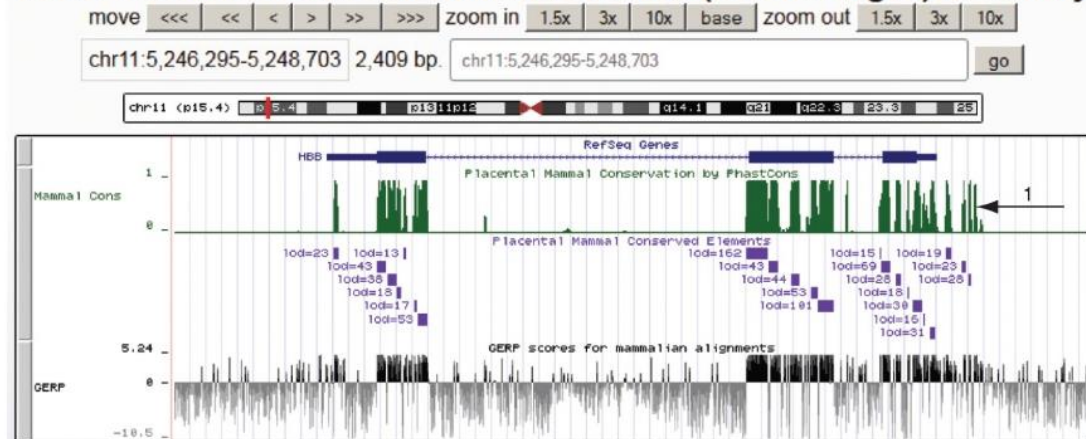
Alignment of divergent sequences often reveals islands of conservation (providing “anchors” for alignment).

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.

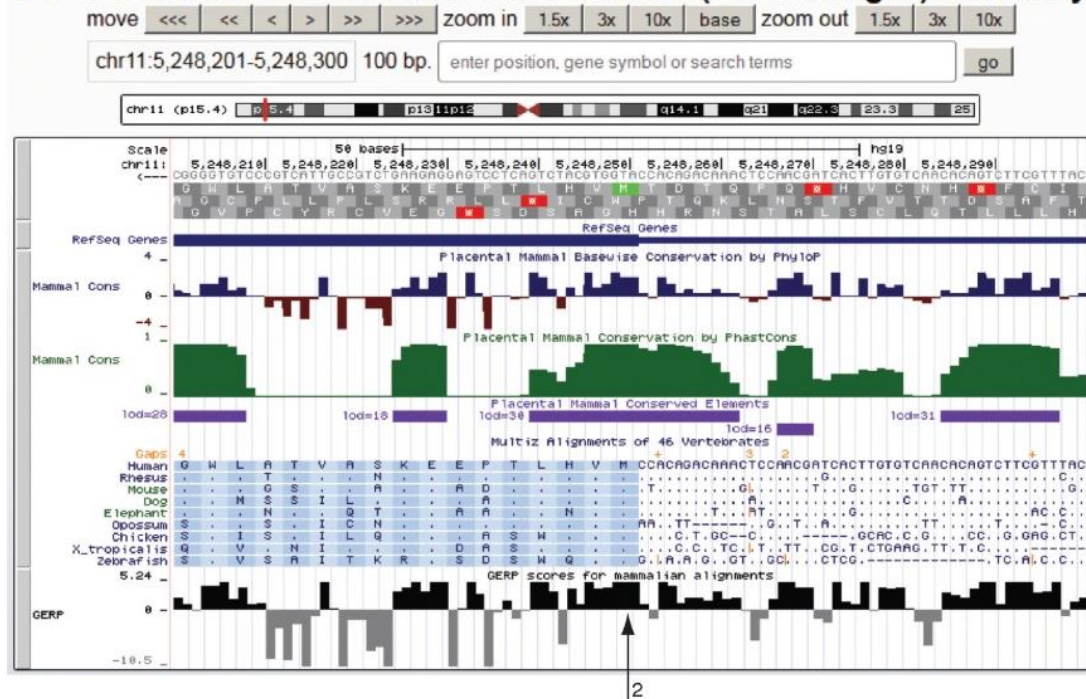
(a) *HBB* gene (zoomed out 1.5x to 2,409 base pairs)

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly



(b) View of *HBB* gene (100 base pairs)

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly



Analyzing multiple sequence alignments at Ensembl

(a) Ensembl entry for *HBB*

Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | More ▾

Human (GRCh37) ▾ Location: 11:5,246,694-5,250,625 ▾ Gene: *HBB* ▾

Gene: *HBB* ENSG00000244734

Description hemoglobin, beta [Source:HGNC Symbol;Acc:4827]
Location [Chromosome 11: 5,246,694-5,250,625](#) reverse strand.
INSDC coordinates chromosome:GRCh37:CM000673.1:5246694:5250625:1
Transcripts This gene has 4 transcripts (splice variants) [Hide transcript table](#)

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CDS incomplete	CCDS
<i>HBB-001</i>	ENST00000335295	754	ENSP00000333994	147	Protein coding	-	CCDS7753
<i>HBB-004</i>	ENST00000380315	502	ENSP00000369671	90	Protein coding	3'	-
<i>HBB-002</i>	ENST00000485743	680	No protein product	-	Retained intron	-	-
<i>HBB-003</i>	ENST00000475226	319	No protein product	-	Retained intron	-	-

Genomic alignments ⓘ

Alignment: -- Select an alignment -- [Go](#)

Go to a graph

Key

- 6 primates EPO
- 13 eutherian mammals EPO
- 20 amniota vertebrates Pecan
- 36 eutherian mammals EPO LOW COVERAGE

Features

Human > chr11

Pairwise alignments

Alpaca (Vicugna pacos) - blastz
 Anole lizard (Anolis carolinensis) - translated blat
 Armadillo (Dasypus novemcinctus) - blastz
 Bushbaby (Oryzomys flavescens) - lastz
 Cat (Felis catus) - lastz
 Chicken (Gallus gallus) - lastz
 Chicken (Gallus gallus) - translated blat
 Chimpanzee (Pan troglodytes) - lastz
 Chinese softshell turtle (Pelodiscus sinensis) - lastz
 Ciona intestinalis - translated blat
 Ciona savignyi - translated blat
 Cod (Gadus morhua) - translated blat
 Coelacanth (Latimeria chalumnae) - translated blat

TTTGAACCAATGATAAACCACTCCCATAGATGAGTGCATGA:
 ACTTAAGAAAGATTAAAGACTGGAGTAAGGAAATGGACTCTGTG:
 GGGCTGGAATAAAAGTAGAATAGACCTGCACCTGCTGTGCATCCAT:
 CTGATTAGATTGAAAGTAGAGGCTCTGACCATACCAATTTGCAC:
 TGTCCCTGCAGGGTATTATGGGTAAAGAAAGAAAGTCTCGTTAC:
 CAGTTGCCAACAAGAGAAGGATCCATAGTTTCATCATTTAAAAAG:
 TTCTGCCAATCAGGATTTCAAAGCTCTTGTCTTGACATTTTGGTC:
 TGCATAAGACATATTCAAACTTCGCGACAAACATTTATTTACATAT:
 TTTAAATTTAATAAAATAAATCCAAATCTAACAGCAAGTCAAAAT:
 GATACACGTGTGCTAGATCCTCAATTGCTTTAGTTTTTACAGAGG:

Analyzing multiple sequence alignments at Ensembl

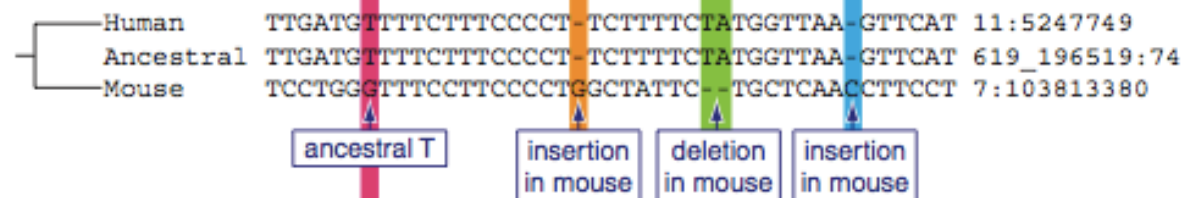
(b) Ensembl multiple sequence alignment (Enredo/Pecan/Ortheus software)

Homo sapiens	11: 5246983	TTCATACCTCTT-ATCTTCCTCCACAGCTCCTGGGCAACGTGCTGG
Gorilla gorilla gorilla	11: 5181973	TTCATACCTCTT-GTCTTCCTCCACAGCTCCTGGGCAATGTGCTGG
Pongo abelii	11: 65239065	TTCATACCTCTT-GTCTCCCTCCACAGCTCCTGGGCAATGTGCTGG
Oryctolagus cuniculus	1:146237264	TTCATGCCTTCT--TCTCTTTCCTACAGCTCCTGGGCAACGTGCTGG
Mus musculus	7:103812810	TTGATGGTTCTT--CCATCTTCCACAGCTCCTGGGCAATATGATCG
Bos taurus	15: 49339417	CCCTTGCTTAATG-TCTTTTCCACACAGCTCCTGGGCAACGTGCTAG
Bos taurus	15: 49074455	CCCTTGCTTAATG-TCTTTTCCACACAGCTCCTGGGCAACGTGCTGG
Sus scrofa	9: 5633260	CCCTTCCTTTTTA-TCTCTCTCCACAGCTCCTGGGCAACGTGATAG
Equus caballus	7: 73936736	CCCCCTCTTT-TT-TCTCTTCCACACAGCTCCTGGGCAACGTGCTGG
Canis lupus familiaris	21: 28179266	CACATGCCTCTTG-TCT--TCCCCACAGCTGCTGGGCAACGTGTTGG

(a) Pairwise alignment



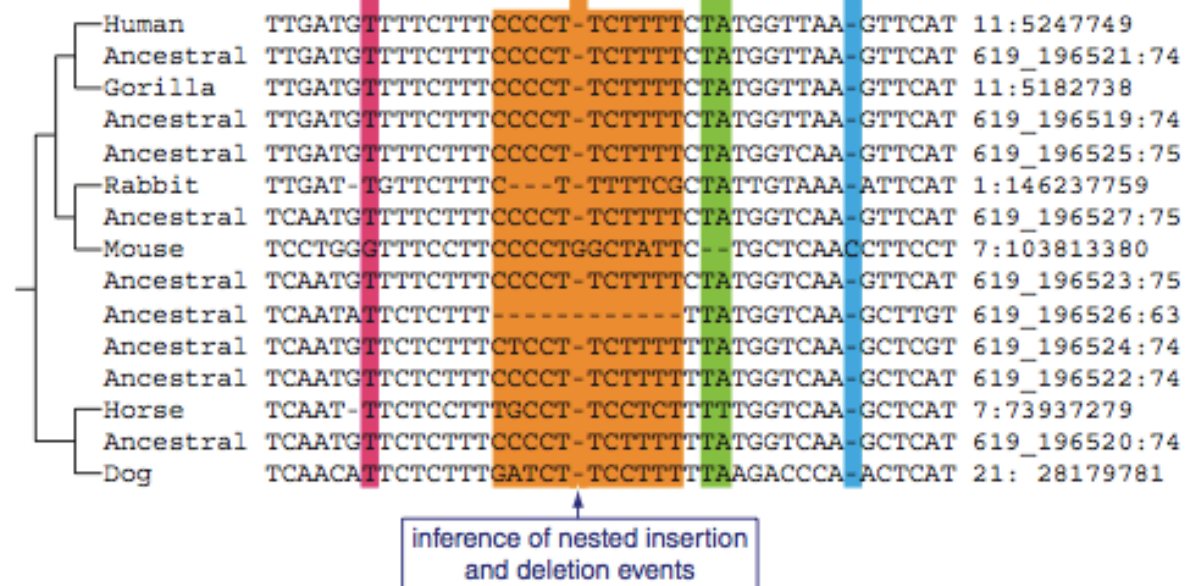
(b) Ancestor alignment



(c) Multiple sequence alignment



(d) Multiple sequence ancestor alignment



Outline: multiple sequence alignment (MSA)

Introduction; definition of MSA; typical uses

Five main approaches to multiple sequence alignment

- Exact approaches

- Progressive sequence alignment

- Iterative approaches

- Consistency-based approaches

- Structure-based methods

Benchmarking studies: approaches, findings, challenges

Databases of Multiple Sequence Alignments

- Pfam: Protein Family Database of Profile HMMs

- SMART

- Conserved Domain Database

- Integrated multiple sequence alignment resources

- MSA database curation: manual versus automated

Multiple sequence alignments of genomic regions

- UCSC, Galaxy, Ensembl, alignathon

Perspective

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.