Chapter VI: Multiple Sequence Alignment

Presentations use info from:

Jonathan Pevsner, Ph.D.
http://bioinfbook.org
pevsner@kennedykrieger.org
Bioinformatics and Functional Genomics
(3rd edition, ©2015 John Wiley & Sons, Ltd.)
You may use this PowerPoint for teaching purposes

dr. Stanislav Kolenčík
stanislav.kolencik@famnit.upr.si

What will you learn?

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); MAFFT

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

When we consider a protein (or gene), one of the most fundamental questions is <u>what other proteins are related to it</u>.

Biological sequences often occur in families. These families may consist of related genes within an organism (paralogs), sequences within a population (e.g., polymorphic variants), or genes in other species (orthologs).

Pfam: Multiple sequence alignments and HMM-profiles of protein domains

Erik L. L. Sonnhammer, Sean R. Eddy, Ewan Birney, Alex Bateman, Richard Durbin

Nucleic Acids Research, Volume 26, Issue 1, 1 January 1998, Pages 320–322,
https://doi.org/10.1093/nar/26.1.320
Published: 01 January 1998 **Article history** ▾

PDF Split View Cite Permissions Share ▾

Abstract

Pfam contains multiple alignments and hidden Markov model based profiles (HMM-profiles) of complete protein domains. The definition of domain boundaries, family members and alignment is done semi-automatically based on expert knowledge, sequence similarity, other protein family databases and the ability of HMM-profiles to correctly identify and align the members. Release 2.0 of Pfam contains 527 manually verified families which are available for browsing and on-line searching via the World Wide Web in the UK at http://www.sanger.ac.uk/Pfam/ and in the US at http://genome.wustl.edu/Pfam/ Pfam 2.0 matches one or more domains in 50% of Swissprot-34 sequences, and 25% of a large sample of predicted proteins from the *Caenorhabditis elegans* genome.

Issue Section: Articles

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, they are presumably derived from a common ancestor.

- residues are homologous in a structural sense; aligned residues tend to occupy corresponding positions in the three-dimensional structure of each aligned protein.

- Compared to pairwise alignments, **multiple sequence alignments** are very powerful because two sequences that may not align well to each other can be aligned via their relationship to a third sequence, thereby integrating information in a way not possible using only pairwise alignments.

- We can therefore define members of a gene or protein family and identify conserved regions.

- **The overwhelming majority of proteins have been identified through the sequencing of genomic DNA or complementary DNA**.

- **The function of most proteins is therefore assigned on the basis of homology to other known proteins rather than on the basis of results from biochemical or cell biological (functional) assays.**

Five algorithmic approaches:

- (1) exact methods;
- (2) progressive alignment (e.g., ClustalW);
- (3) iterative approaches (e.g., PRALINE, IterAlign, MUSCLE);
- (4) consistency-based methods (e.g., MAFFT, ProbCons);
- (5) structure-based methods that include information about one or more known three-dimensional protein structures to facilitate creation of a multiple sequence alignment (e.g., Expresso).

Example: 5 alignments of 5 globins

MSA programs: e.g., ClustalW, Praline, MUSCLE (used at HomoloGene), ProbCons, and TCoffee. Each program offers unique strengths.

Multiple sequence alignment (MSA) of five globins proteins…

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

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   ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin     -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFK- 48
neuroglobin   -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR 47
soybean       ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice          MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLR- 59
                            :   :   :   :  .. .    .      ::   *     *.

beta globin   DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLS-----ELHCDKLHVDPE 102
myoglobin     HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLA-----QSHATKHKIPVK 103
neuroglobin   QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLAS---LGRKHRAVGVKLS 104
soybean       --NGVDPT--NPKLTGHAEKLFALVRDSAGQLKASGTVVADAA----LGSVHAQKAVTDP 101
rice          --NSDVPLEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117
                .       . ..  *  .::        :           :              :

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

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

Praline

(a) Praline multiple sequence alignment

```
beta globin      ..........MVHLTPEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFES.FG
myoglobin        ...........MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK.FK
neuroglobin      .............MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean          ..........MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS..FL
rice             MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS..FL
Consistency      0000000000142654382579345734633643436244536864 33*35344*50063

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

beta globin      ENFRLLGNVLVCVLAHHF.GKEFTPPVQAAYQKVVAGVANALAHKYH......
myoglobin        KYLEFISECIIQVLQSKH.PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin      SSFSTVGESLLYMLEKCL.GPAFTPATRAAWSQLYGAVVQAMSRGWD..GE..
soybean          PQFVVVKEALLKTIKAAV.GDKWSDELSRAWEVAYDELAAAIKKA........
rice             AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE...
Consistency      437448444982585423053365544 54*5546542644675432200 1000
```

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

MUSCLE

(b)

```
MUSCLE (3.6) multiple sequence alignment

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

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

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

Probcons

(c)

```
PROBCONS

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

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

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

TCoffee

(d)

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

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

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

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

Multiple sequence alignment: properties

- not necessarily one "correct" alignment of a protein family;
- **protein sequences evolve**...;
- ...the corresponding 3-D 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 [Profiles (such as those described for DELTA-BLAST and hidden Markov models) depend on accurate multiple sequence alignments].

- BLAST output can take the form of a MSA, and can reveal conserved residues or motifs.

- Algorithms that predict whether variants are harmful often rely on DNA and/or protein multiple sequence alignments to assess cross-species conservation. Deleterious variants tend to occur at more conserved positions.

- 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

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]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCD
KLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
>myoglobin 2MM1 NP_005359.1 [Homo sapiens]
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHAT
KHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG
>neuroglobin 1OJ6A NP_067080.1 [Homo sapiens]
MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHR
```

Or, upload a file: Browse...

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)*

Submit

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

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]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCD
KLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
>myoglobin 2MM1 NP_005359.1 [Homo sapiens]
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHAT
KHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG
>neuroglobin 1OJ6A NP_067080.1 [Homo sapiens]
MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHR
```

Or upload a file: Browse

EMBL-EBI | Services | Research | Training | Industry | About us | EMBL-EBI Hinxton

ClustalW2

Input form | Web services | Help & Documentation | Bioinformatics Tools FAQ | Feedback | Share

Tools > Multiple Sequence Alignment > ClustalW2

ClustalW2 is a general purpose DNA or protein multiple sequence alignment program for **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Please Note

The ClustalW2 services have been retired. To access similar services, please visit the Multiple Sequence Alignment tools page. For protein alignments we recommend Clustal Omega. For DNA alignments we recommend trying MUSCLE or MAFFT. If you have any questions/concerns please contact us via the feedback link above.

FORMAT: Clustal w/ numbers

ORDER: 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)*

Submit

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)

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

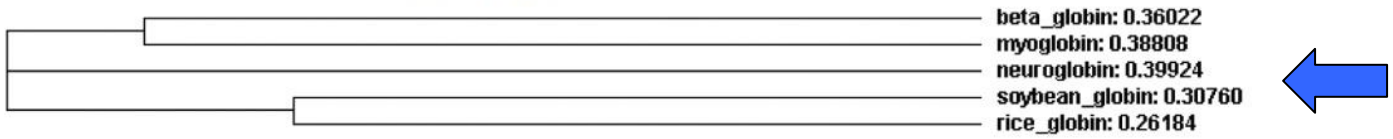
best score

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

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 guide tree is calculated from the distance matrix with the unweighted pair group method (UPGMA) or neighbor-joining method (NJ).

ClustalW alignment of five distantly related beta globin orthologs

```
beta_globin     ----------MVHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin       -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFK- 48
neuroglobin     -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR 47
soybean_globin  ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA- 49
rice_globin     MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLR- 59
                             :   :   :   :  ..      .      ::   *     *.

beta_globin     DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAT------LSELHCDKLHVDP 101
myoglobin       HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKP------LAQSHATKHKIPV 102
neuroglobin     QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEY---LASLGRKHRAVGVKLS 104
soybean_globin  --NGVDPT--NPKLTGHAEKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDP 101
rice_globin     --NSDVPLEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117
                  .       . ..  *  .::        :                 *.  *

beta_globin     ENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------ 147
myoglobin       KYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin     SFSTVGESLLYMLEKCLG-PAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151
soybean_globin  QFVVVKEALLKTIKAAVG-DKWSDELSRAWEVAYDELAAAIKKA-------- 144
rice_globin     HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166
                    .    :  :        :      * .     .   :
```

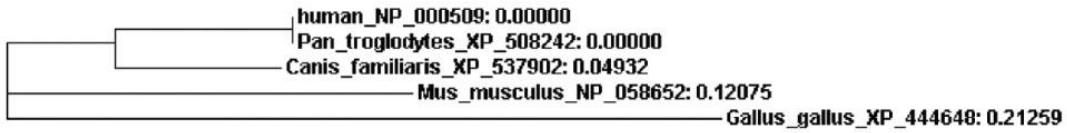
The region of the second histidine is prone to misalignment, and we will explore how other programs treat this region.

(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);
```



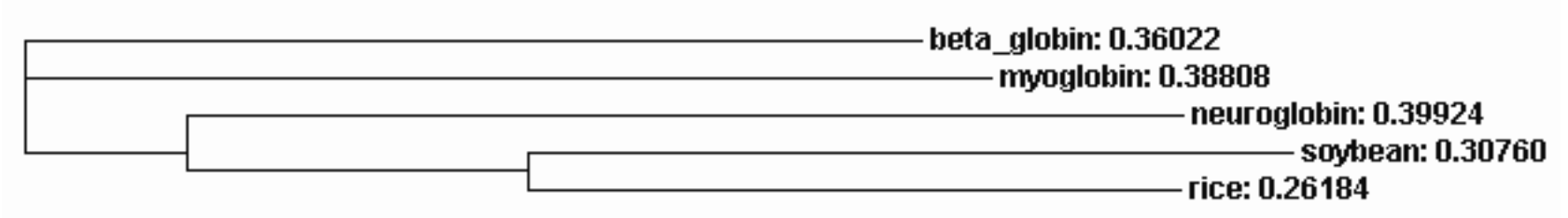
ClustalW alignment of five closely related beta globin orthologs

```
                                                                    ▼
human_NP_000509                 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS 50
Pan_troglodytes_XP_508242       MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS 50
Canis_familiaris_XP_537902      MVHLTAEEKSLVSGLWGKVNVDEVGGEALGRLLIVYPWTQRFFDSFGDLS 50
Mus_musculus_NP_058652          MVHLTDAEKSAVSCLWAKVNPDEVGGEALGRLLVVYPWTQRYFDSFGDLS 50
Gallus_gallus_XP_444648         MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLS 50
                                *** *  **. :: **.***  * *.***.***:*******:* ***:**

                                          ▼                         ▼
human_NP_000509                 TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD 100
Pan_troglodytes_XP_508242       TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD 100
Canis_familiaris_XP_537902      TPDAVMSNAKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVD 100
Mus_musculus_NP_058652          SASAIMGNPKVKAHGKKVITAFNEGLKNLDNLKGTFASLSELHCDKLHVD 100
Gallus_gallus_XP_444648         SPTAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVD 100
                                :. *::.*. *:******: :*.:.: :***:*.**: ***********

human_NP_000509                 PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147
Pan_troglodytes_XP_508242       PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH 147
Canis_familiaris_XP_537902      PENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALAHKYH 147
Mus_musculus_NP_058652          PENFRLLGNAIVIVLGHHLGKDFTPAAQAAFQKVVAGVATALAHKYH 147
Gallus_gallus_XP_444648         PENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH 147
                                ****:***: :: **. *:.*:***  ***:**:*  ** ***:***
```

Progressive MSA stage 2 of 3:

generate a guide tree calculated from
the distance matrix (5 distantly related globins)

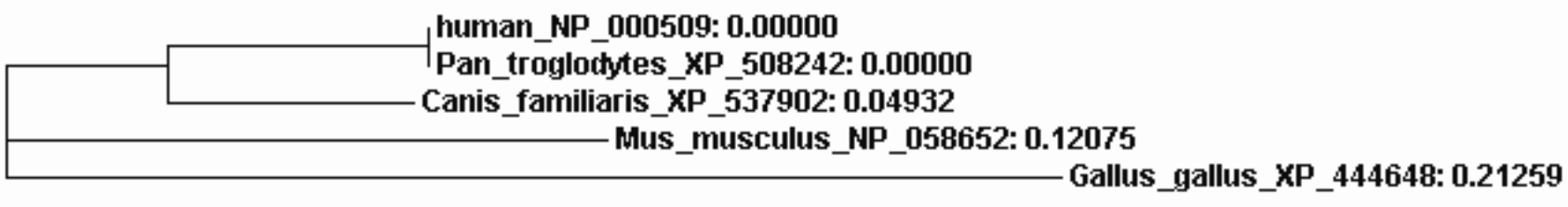
```
(
beta_globin:0.36022,
myoglobin:0.38808,
(
neuroglobin:0.39924,
(
soybean:0.30760,
rice:0.26184)
:0.13652)
:0.06560);
```



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.”

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

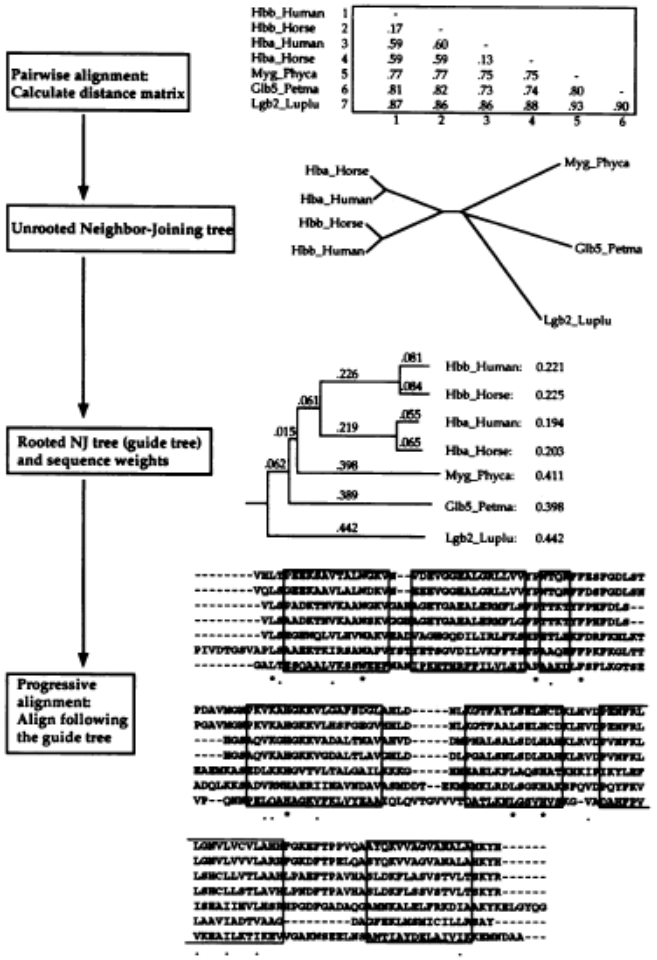


Figure 1. The basic progressive alignment procedure, illustrated using a set of 7 globins of known tertiary structure. The sequence names are from Swiss Prot (38): Hba__Horse: horse α-globin; Hba__Human: human α-globin; Hbb__Horse: horse β-globin; Hbb__Human: human β-globin; Myg__Phyca: sperm whale myoglobin; Glb5__Petma: lamprey cyanohaemoglobin; Lgb2__Luplu: lupin leghaemoglobin. In the distance matrix, the mean number of differences per residue

In Figure 1 we give the 7×7 distance matrix between the 7 globin sequences calculated using the full dynamic programming method.

The guide tree

The trees used to guide the final multiple alignment process are calculated from the distance matrix of step 1 using the Neighbour-Joining method (21). This produces unrooted trees with branch lengths proportional to estimated divergence along each branch. The root is placed by a 'mid-point' method (15) at a position where the means of the branch lengths on either side of the root are equal. These trees are also used to derive a weight for each sequence (15). The weights are dependent upon the distance from the root of the tree but sequences which have a common branch with other sequences share the weight derived from the shared branch. In the example in Figure 1, the leghaemoglobin (Lgb2__Luplu) gets a weight of 0.442, which is equal to the length of the branch from the root to it. The human β-globin (Hbb__Human) gets a weight consisting of the length of the branch leading to it that is not shared with any other sequences (0.081) plus half the length of the branch shared with the horse β-globin (0.226/2) plus one quarter the length of the branch shared by all four haemoglobins (0.061/4) plus one fifth the branch shared between the haemoglobins and myoglobin (0.015/5) plus one sixth the branch leading to all the vertebrate globins (0.062). This sums to a total of 0.221. In contrast, in the normal progressive alignment algorithm, all sequences would be equally weighted. The rooted tree with branch lengths and sequence weights for the 7 globins is given in Figure 1.

Progressive alignment

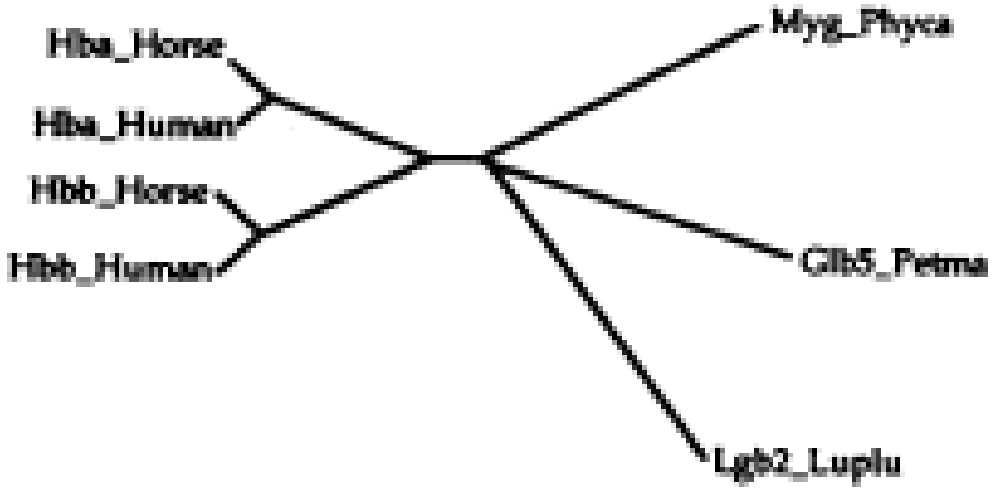
The basic procedure at this stage is to use a series of pairwise alignments to align larger and larger groups of sequences, following the branching order in the guide tree. You proceed from the tips of the rooted tree towards the root. In the globin example in Figure 1 you align the sequences in the following order: human vs. horse β-globin; human vs. horse α-globin; the 2 α-globins vs. the 2 β-globins; the myoglobin vs. the haemoglobins; the cyanohaemoglobin vs. the haemoglobins plus

See Thompson et al. (1994) for an explanation of the three stages of progressive alignment implemented in ClustalW

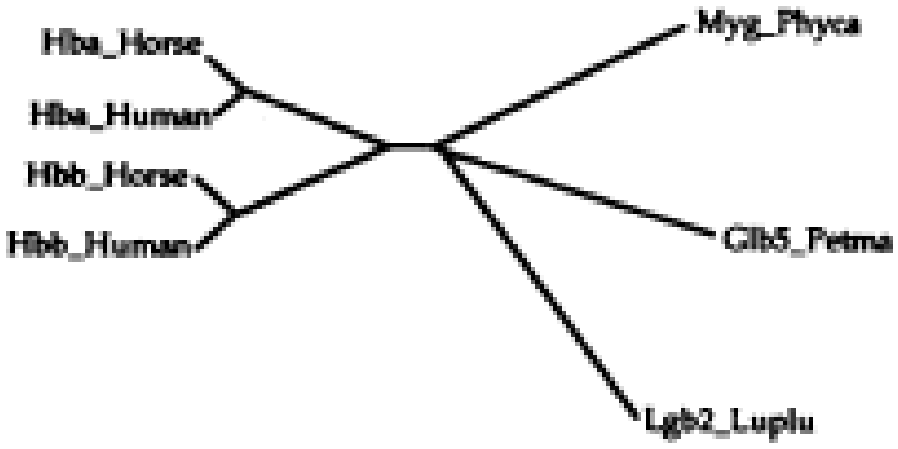
Pairwise alignment:
Calculate distance matrix

Unrooted neighbor-joining tree

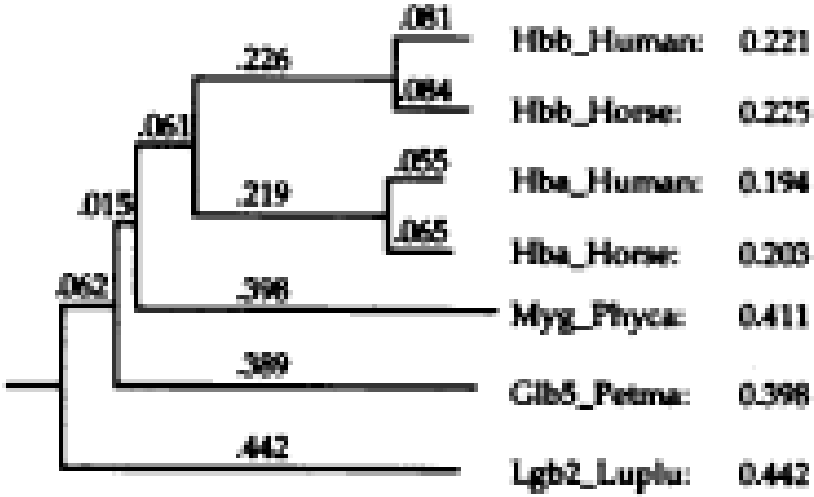
		1	2	3	4	5	6
Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90



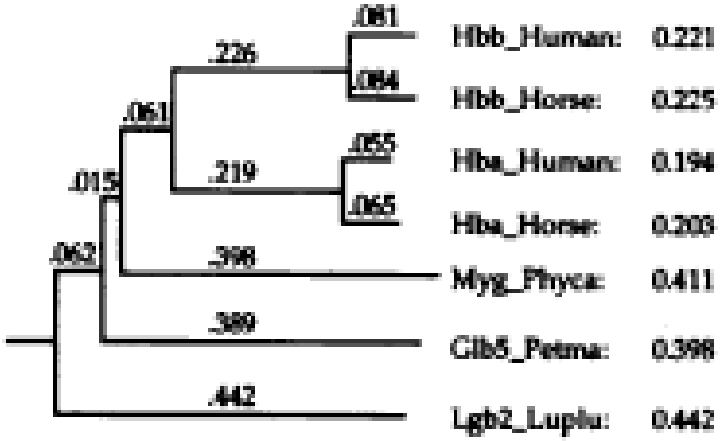
Unrooted neighbor-joining tree



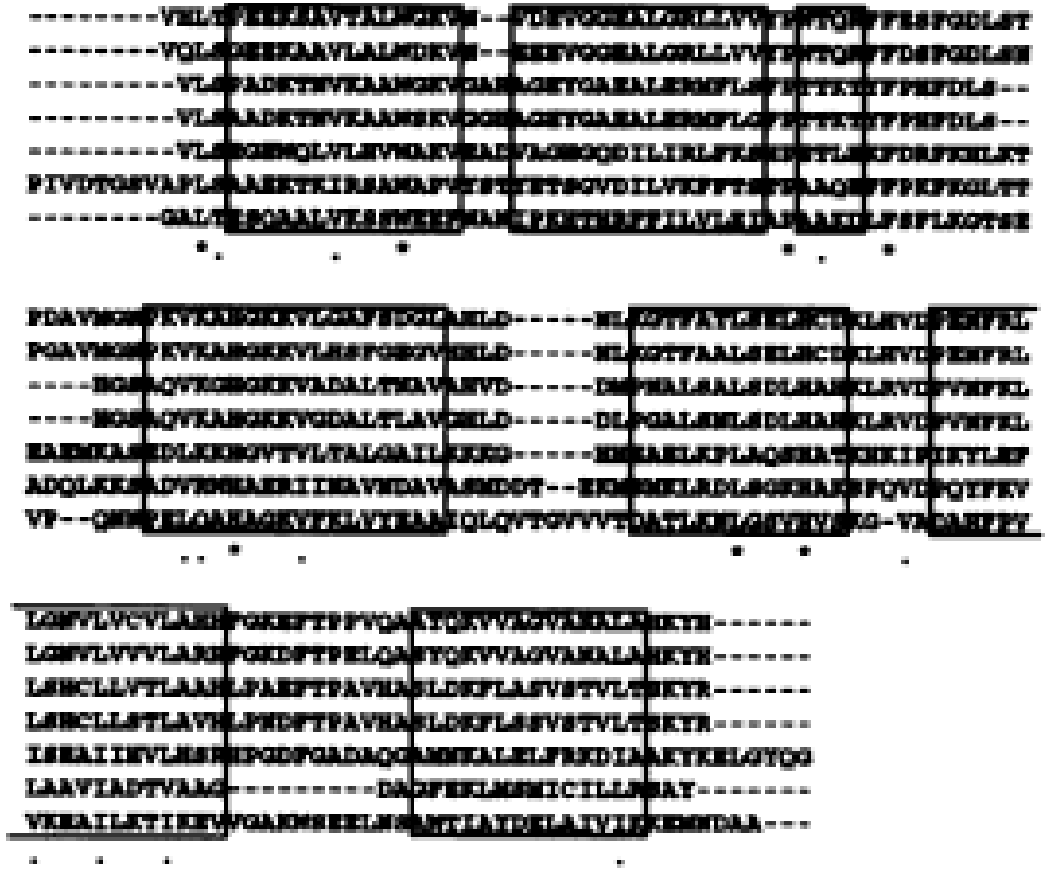
Rooted neighbor-joining tree (guide tree) and sequence weights



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



Progressive alignment: Align following the guide tree

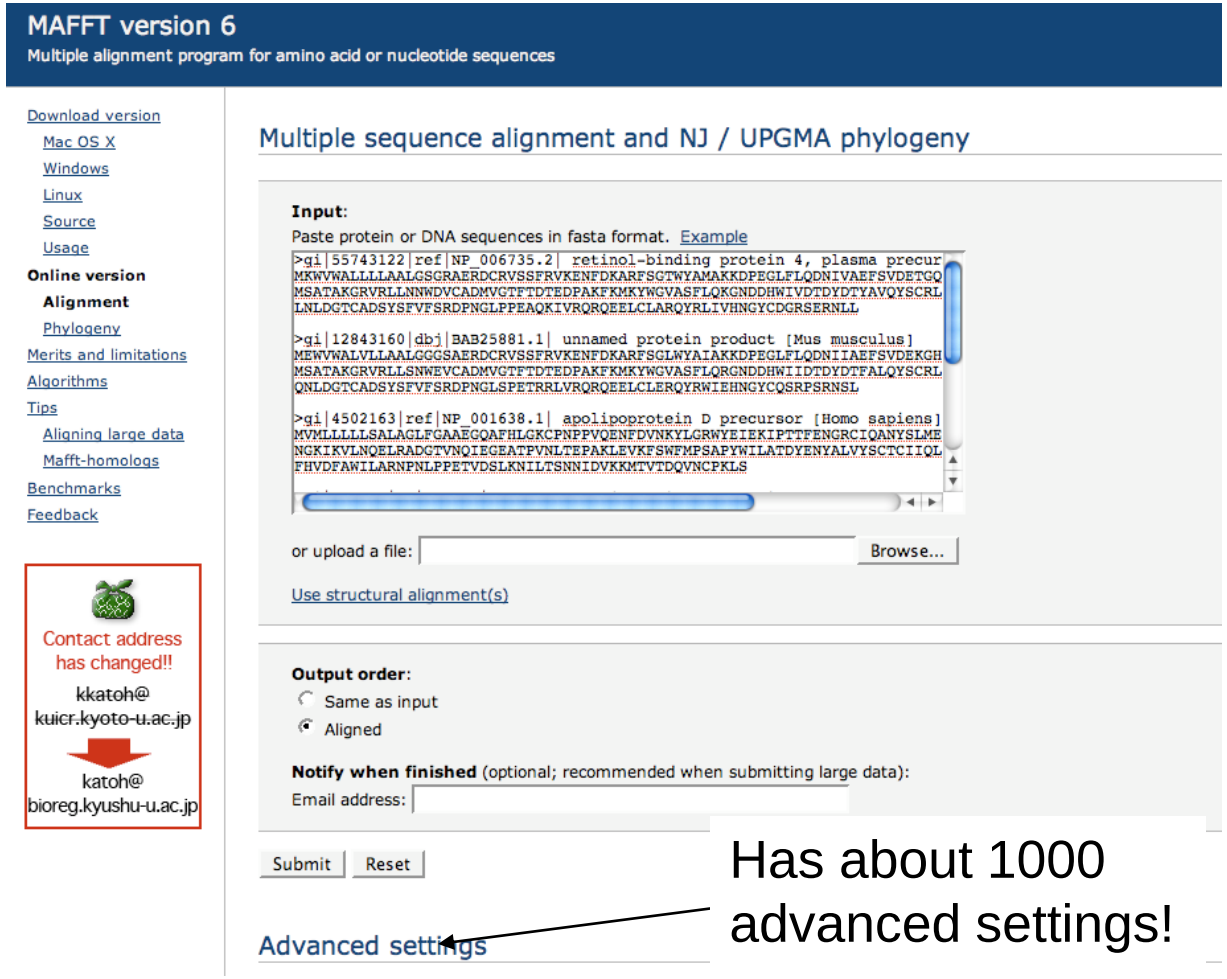


Iterative approaches: MAFFT

- 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

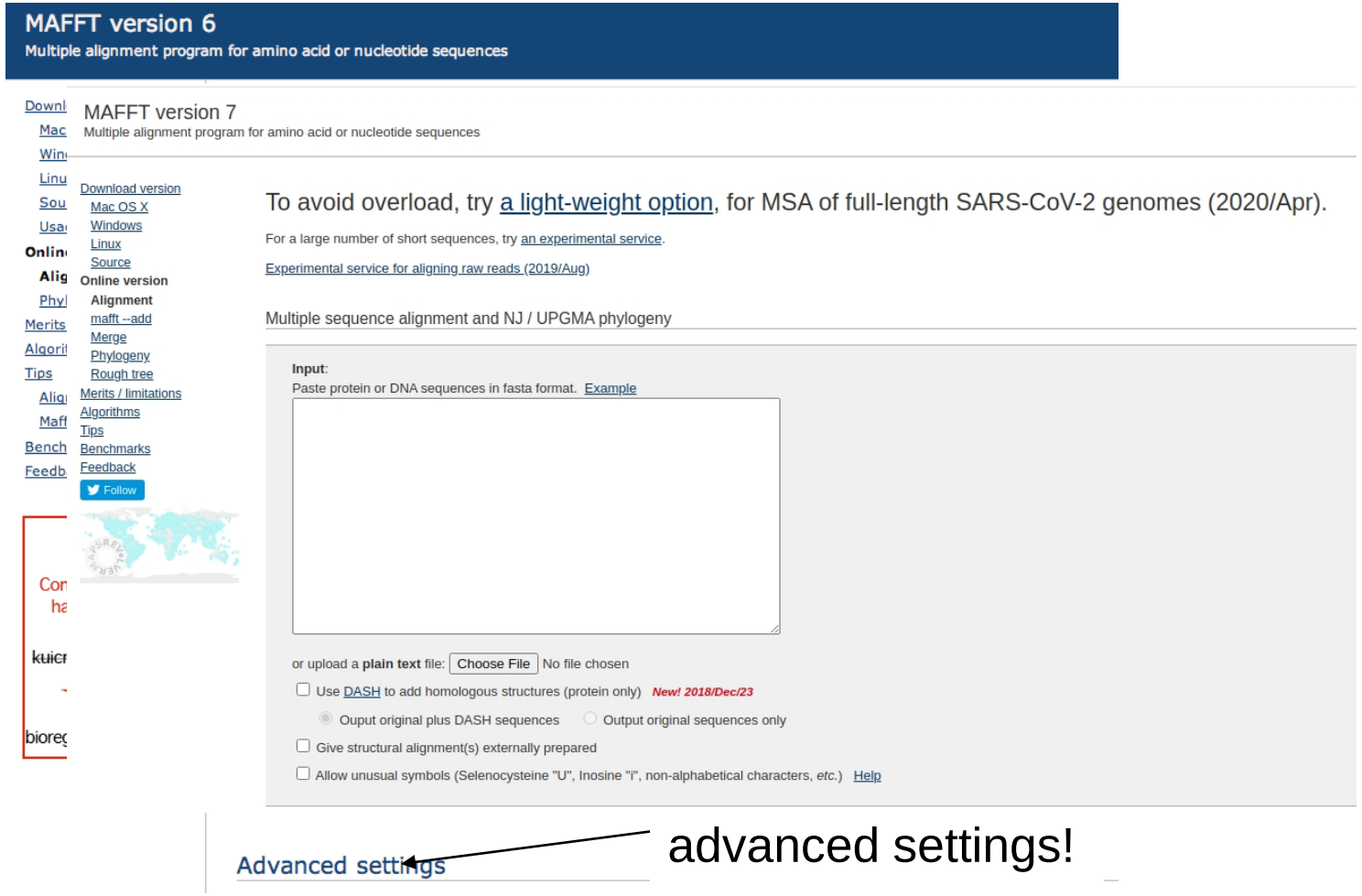
- One of the more accurate programs available
- Available as standalone or web interface
- Many output formats, including interactive phylogenetic trees

Iterative approaches: MAFFT



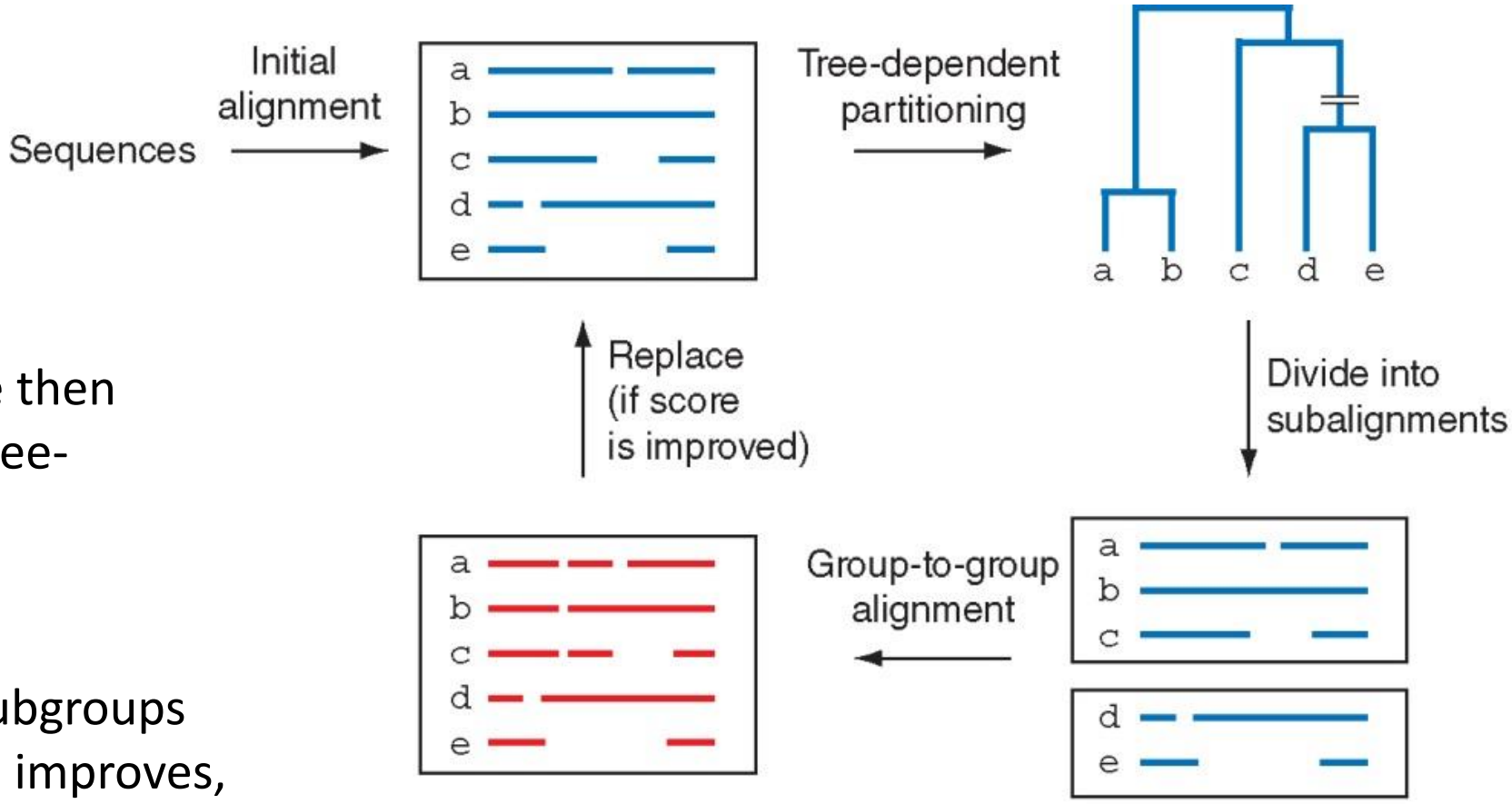
Has about 1000 advanced settings!

Iterative approaches: MAFFT



advanced settings!

Iterative method of MAFFT



A progressive alignment is made then divided into subalignments by tree-dependent partitioning.

Partitions are re-aligned, then subgroups are aligned. If an objective score improves, this new alignment replaces the initial one and the process may be repeated.

MAFFT

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

```
                                                                                1
hbb_human    ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFE-SFG
hbb_chimp    ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFE-SFG
hbb_dog      ----------MVHLTAEEKSLVSGLWGKVNVD--EVGGEALGRLLIVYPWTQRFFD-SFG
hbb_mouse    ----------MVHLTDAEKSAVSCLWAKVNPD--EVGGEALGRLLVVYPWTQRYFD-SFG
hbb_chicken  ----------MVHWTAEEKQLITGLWGKVNVA--ECGAEALARLLIVYPWTQRFFA-SFG
myoglobin    -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFD-KFK
neuroglobin -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean      ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS-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----TNPKLTGHAEKLFALVRDSAGQLKASGTV-VADAA---LGSVH-AQKAVTD
rice         NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATH-LKYGVGD
             .         .  .  *  .::        :    .     .     *.  *     :
```

MUSCLE

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

```
                                                                                1
hbb_human    ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFE-SFG
hbb_chimp    ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFE-SFG
hbb_dog      ----------MVHLTAEEKSLVSGLWGKVNVD--EVGGEALGRLLIVYPWTQRFFD-SFG
hbb_mouse    ----------MVHLTDAEKSAVSCLWAKVNPD--EVGGEALGRLLVVYPWTQRYFD-SFG
hbb_chicken  ----------MVHWTAEEKQLITGLWGKVNVA--ECGAEALARLLIVYPWTQRFFA-SFG
myoglobin    -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFD-KFK
neuroglobin -------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean      ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS-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----NPKLTGHAEKLFALVRDSAGQL----KASGTVVADAALGSVHAQKAVTDP
rice         NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
                       .  .  *  .::        :                    :    :
```

ProbCons

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

```
                                                                                                  1
hbb_human      M----------VHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
hbb_chimp      M----------VHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
hbb_dog        M----------VHLTAEEKSLVSGLWGKVNVD--EVGGEALGRLLIVYPWTQRFFDS-FG
hbb_mouse      M----------VHLTDAEKSAVSCLWAKVNPD--EVGGEALGRLLVVYPWTQRYFDS-FG
hbb_chicken    M----------VHWTAEEKQLITGLWGKVNVA--ECGAEALARLLIVYPWTQRFFAS-FG
myoglobin      M-----------GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK
neuroglobin    M-------------ERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean        M----------VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice           MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
               *          *     :   :   :  ..      .      ::   *     *
                                 2                                  3
hbb_human      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNL------KGTFATLSELHCDKLHVDP
hbb_chimp      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNL------KGTFATLSELHCDKLHVDP
hbb_dog        DLSTPDAVMSNAKVKAHGKKVLNSFSDGLKNLDNL------KGTFAKLSELHCDKLHVDP
hbb_mouse      DLSSASAIMGNPKVKAHGKKVITAFNEGLKNLDNL------KGTFASLSELHCDKLHVDP
hbb_chicken    NLSSPTAILGNPMVRAHGKKVLTSFGDAVKNLDNI------KNTFSQLSELHCDKLHVDP
myoglobin      HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHH------EAEIKPLAQSHATKHKIPV
neuroglobin    QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSL---EEYLASLGRKHRAV-GVKL
soybean        NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTV---V-ADAALGSVHAQK-AVTD
rice           NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKY-GVGD
               .         .  .  *  .::        :    ..:         *.  *     :
```

T-COFFEE

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

```
                                                                                                  1
1HBB           1 ----------MVHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTORFFESFGD-LSTPDAVM
hbb_chimp      1 ----------MVHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTORFFESFGD-LSTPDAVM
2qlsB          1 ----------MVHLTAEEKSLVSGLWGKVN--VDEVGGEALGRLLIVYPWTORFFDSFGD-LSTPDAVM
3hrwB          1 ----------MVHLTDAEKSAVSCLWAKVN--PDEVGGEALGRLLVVYPWTORYFDSFGD-LSSASAIM
1hbrB          1 ----------MVHWTAEEKQLITGLWGKVN--VAECGAEALARLLIVYPWTORFFASFGN-LSSPTAIL
3RGK           1 -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKH-LKSEDEMK
1oj6A          1 -------------MERPEPELIROSWRAVSRSPLEHGTVLFARLFALEPDLLPLFOYNCRQFSSPEDCL
1FSL           1 ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS-----FLANGVDP
1D8U           1 MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF--L-RNSDVPLE

cons           1           :   :   :  ..     .      ::   *     *      :

                           2                                  3
1HBB          57 GNPKVKAHGKKVLGAFSDGLAHL-DNL----KGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
hbb_chimp     57 GNPKVKAHGKKVLGAFSDGLAHL-DNL----KGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
2qlsB         57 SNAKVKAHGKKVLNSFSDGLKNL-DNL----KGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHHFG
3hrwB         57 GNPKVKAHGKKVITAFNEGLKNL-DNL----KGTFASLSELHCDKLHVDPENFRLLGNAIVIVLGHHLG
1hbrB         57 GNPMVRAHGKKVLTSFGDAVKNL-DNI----KNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFS
3RGK          58 ASEDLKKHGATVLTALGGILKKK-GHH----EAEIKPLAOSHATKHKIPVKYLEFISECIIOVLOSKHP
1oj6A         57 SSPEFLDHIRKVMLVIDAAVTNV-EDL-SSLEEYLASLGRKHR-AVGVKLSSFSTVGESLLYMLEKCLG
1FSL          55 TNPKLTGHAEKLFALVRDSAGQLKASG----TVVADAALGSVHAQKAVTDPQFVVVKEALLKTIKAAVG
1D8U          67 KNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDAHFEVVKFALLDTIKEEVP

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

Access to MUSLCE at EBI
http://www.ebi.ac.uk/muscle/

EMBL-EBI
European Bioinformatics Institute

EBI Home | About EBI | Groups | Services | Toolbox | Databases | Downloads | Submissions

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 **MU**ltiple **S**equence **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

MUSCLE multiple sequence alignment method with reduced time and space complexity

EMAIL	RESULTS	ALIGNMENT TITLE	OUTPUT FORMAT	OUTPUT TREE
	interactive	Sequence	fasta	none

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

Upload a file: Browse... Run Reset

Multiple sequence alignment: consistency

In progressive alignments using the Feng–Doolittle approach, pairwise alignment scores are generated and used to build a tree. Consistency-based methods adopt a different approach (example in the next slide).

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

Examples: T-COFFEE, Prrp, DiAlign, ProbCons
+ Clustal Omega.

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 output for the same alignment: consistency iteration helps

(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

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:

- T-Coffee
- M-Coffee
- R-Coffee
- Expresso
- more ...

Other T-Coffee links

- Documentation
- Downloads
- Support & discussion group

Access to TCoffee:
http://tcoffee.org

https://www.ebi.ac.uk/Tools/msa/tcoffee/

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

https://www.ebi.ac.uk/Tools/msa/

Multiple Sequence Alignment

Feedback | Share

Tools > Multiple Sequence Alignment

Multiple Sequence Alignment (MSA) is generally the alignment of three or more biological sequences (protein or nucleic acid) of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences studied.

By contrast, **Pairwise Sequence Alignment** tools are used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences.

Clustal Omega

New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large alignments.

Launch Clustal Omega

EMBOSS Cons

EMBOSS Cons creates a consensus sequence from a protein or nucleotide multiple alignment.

Launch EMBOSS Cons

Kalign

Very fast MSA tool that concentrates on local regions. Suitable for large alignments.

Launch Kalign

MAFFT

MSA tool that uses Fast Fourier Transforms. Suitable for medium-large alignments.

Launch MAFFT

MUSCLE

Accurate MSA tool, especially good with proteins. Suitable for medium alignments.

Launch MUSCLE

MView

Transform a Sequence Similarity Search result into a Multiple Sequence Alignment or reformat a Multiple Sequence Alignment using the MView program.

Launch MView

T-Coffee

Consistency-based MSA tool that attempts to mitigate the pitfalls of progressive alignment methods. Suitable for small alignments.

Launch T-Coffee

WebPRANK

The EBI has a new phylogeny-aware multiple sequence alignment program which makes use of evolutionary information to help place insertions and deletions. Try it out at WebPRANK.

https://www.ebi.ac.uk/Tools/msa/

Multiple Sequence Alignment

Feedback | Share

Tools > Multiple Sequence Alignment

Multiple Seq
homology car

By contrast, P
between two

Clustal Ome
New MSA
Launch

EMBOSS C
EMBOSS
Launch

Kalign
Very fast
Launch

MAFFT
MSA tool
Launch

MUSCLE
Accurate
Launch

MView
Transform
Launch

T-Coffee
Consister
Launch

WebPRANK
The EBI
Try it out

Clustal Omega

Input form | Web services | Help & Documentation | Bioinformatics Tools FAQ | Feedback | Share

Tools > Multiple Sequence Alignment > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

Or, upload a file: Choose File No file chosen

Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

The default settings will fulfill the needs of most users.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job

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

Submit

Structure-Based Methods

APDB: a web server to evaluate the accuracy of sequence alignments using structural information

FREE

Fabrice Armougom, Olivier Poirot, Sébastien Moretti, Desmond G. Higgins, Phillip Bucher, Vladimir Keduas, Cedric Notredame Author Notes

Bioinformatics, Volume 22, Issue 19, 1 October 2006, Pages 2439–2440,
https://doi.org/10.1093/bioinformatics/btl404
Published: 01 October 2006 **Article history** ▾

PDF Split View Cite Permissions Share ▾

Abstract

Summary: The APDB webserver uses structural information to evaluate the alignment of sequences with known structures. It returns a score correlated to the overall alignment accuracy as well as a local evaluation. Any sequence alignment can be analyzed with APDB provided it includes at least two proteins with known structures. Sequences without a known structure are simply ignored and do not contribute to the scoring procedure.

Availability: APDB is part of the T-Coffee suite of tools for alignment analysis, it is available on Author Webpage. A standalone version of the package is also available as a freeware open source from the same address.

Contact:cedric.notredame@europe.com

Issue Section: Sequence analysis

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
*
2hhbB    : 224
1V5HA    : 213
2MM1     : 219
1OJ6A    : 194
1FSL     : 157

2hhbB    ---------------MVHLTPEEKSAVTALWG--KVNVDEVGGEALGRLLVVYP
1V5HA   MEKVPGEMEIERRERSEELSEAERKAVQAMWARLYANCEDVGVAILVRFFVNFP
2MM1     ----------------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHP
1OJ6A    ------------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEP
1FSL     ---------------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAP

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

Protein Data Bank accession numbers

Programs that enable you to **incorporate structural information** include **PRALINE** (Simossis and Heringa, 2005) and the T-COFFEE module Expresso (Armougom et al., 2006b).

When you use the Expresso program at the T-COFFEE website, you submit a series of sequences (typically in the FASTA format).
Each sequence is automatically searched by BLAST against the Protein Data Bank (PDB) database, and matches (sharing >60% amino acid identity) are used to provide a template to guide the creation of the multiple sequence 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.

BaliBase: comparison of multiple sequence alignment algorithms

```
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.....iqlqgl..WK
POL_SIVMK     1  pQFSLWRRPVVTAHIEGQPVEVLLDTGADDSIVTG.....iel.gph.YT

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

1fmb         99  GAKLVl
7upjB        94  GMSLNl
pol_sivcz    94  GCTLV.
POL_SIVMK    94  GMSLNl
```

Key

```
alpha helix         RED
beta strand         GREEN
core blocks         UNDERSCORE
```

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)

(a) Pfam alignments

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

keyword search Go

Family: *Globin* (PF00042)

34 architectures | 6000 sequences | 5 interactions | 2886 species | 1971 structures

- Summary
- Domain organisation
- Clan
- Alignments
- HMM logo
- Trees
- Curation & model
- Species
- Interactions
- Structures

Jump to... enter ID/acc Go

Alignments

We store a range of different sequence alignments for families. As well as the seed alignment from which the family is built, we provide the full alignment, generated by searching the sequence database using the family HMM. We also generate alignments using four representative proteomes (RP) sets, the NCBI sequence database, and our metagenomics sequence database. More...

View options

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	×[1]	–	✓	✓	✓	✓	×	×
Pfam viewer	✓	✓	×	×	×	×	×	×

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

Key: ✓ available, × not generated, – not available.

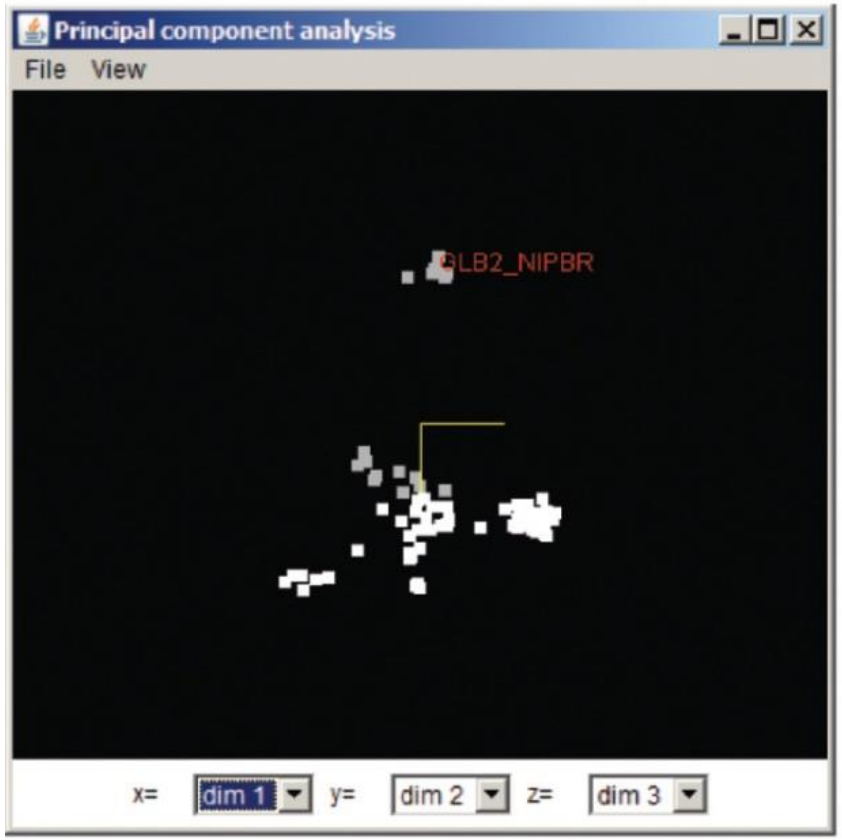
(b) Pfam seed alignment

Seed sequence alignment for PF00042

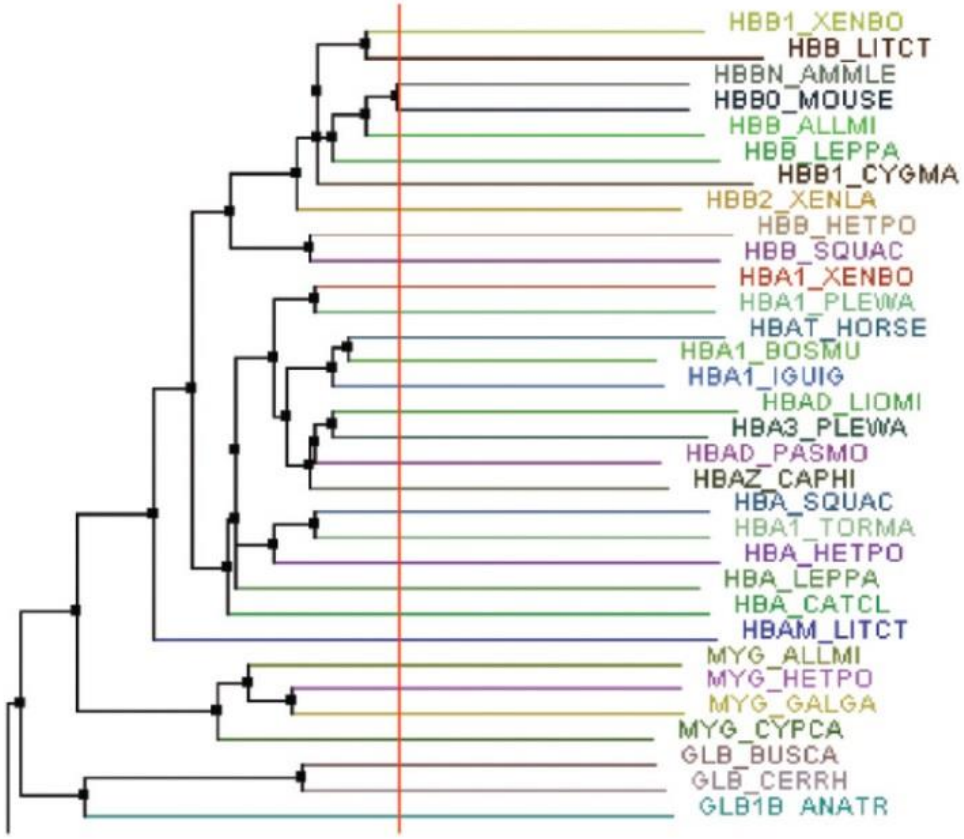
```
Q20638_CAEEL/74-184        EKELLRRIWSD.EFD......NLYELGSAIYCYIFDHNPNCKQLFP.F.ISKYQGDEWKESKEFRSQALKFVQILAQVVK
Q19601_CAEEL/105-215       ERILLEQSWRK.TRK.....IGADHIGSKIFFMVLTAQPDIKAIFG.L..EKIPTGRLKYDPRFRQHALVYTKILDFVIR
Q18311_CAEEL/32-140        TKKLVIQEWPR.VLA......QCPELFTEIWHKSATRSTSIKLAFG.I.AE.N..ESPMQNAAFLGLSSTIQAFFYKLII
GLB4_LUMTE/11-120          DRREIRHIWDD.VWSSS.FTDRRVAIVRAVFDDLFKHYPTSKALFERVKIDEP......ESGEFKSHLVRVANGLDLLIN
GLB4_LUMTE/11-120 (SS)     HHHHHHHHHHH.HS--S.SCHHHHHHHHHHHHHHHHSGGGGGGGCCCTTTST......TSSHHHHHHHHHHHHHHHHT
GLB3_TYLHE/8-117           DRHEVLDNWKG.IWSAE.FTGRRVAIGQAIFQELFALDPNAKGVFGRVNVD.K.....PSEADWKAHVIRVINGLDLAVN
GLB4_TYLHE/8-117           DRREVQALWRS.IWSAE.DTGRRTLIGRLLFEELFEIDGATKGLFKRVNVDDT......HSPEEFAHVLRVVNGLDTLIG
GLB1_TYLHE/7-110           QRIKVKQQWAQ.VYSV...GESRTDFAIDVFNNFFRTNPDRS.LFNRVNGDNV......YSPEFKAHMVRVFAGFDILIS
GLB2_TYLHE/9-115           QRLKVKQQWAK.AYGV...GHERVELGIALWKSMFAQDNDARDLFKRVHGEDV......HSPAFEAHMARVFNGLDRVIS
GLB2_LUMTE/8-114           EGLKVKSEWGR.AYGS...GHDREAFSQAIWRATFAQVPESRSLFKRVHGDDT......SHPAFIAHAERVLGGLDIAIS
GLB2_LUMTE/8-114 (SS)      HHHHHHHHHHH.HH-S...HHHHHHHHHHHHHHHHH-GGGGGGGGGGTTT-T......TSHHHHHHHHHHHHHHHHHHC
GLB_TUBTU/6-112            QRFKVKHQWAE.AFGT...SHHRLDFGLKLWNSIFRDAPEIRGLFKRVDGD.N.....AYSAEFEAHAERVLGGLDMTIS
GLB3_LAMSP/7-113           QRLKVKRQWAE.AYGS...GNDREEFGHFIWTHVFKDAPSARDLFKRVRGDNI......HTPAFRAHATRVLGGLDMCIA
```

Pfam alignment retrieved in the JalView Java viewer

(a) Principal components analysis (PCA)



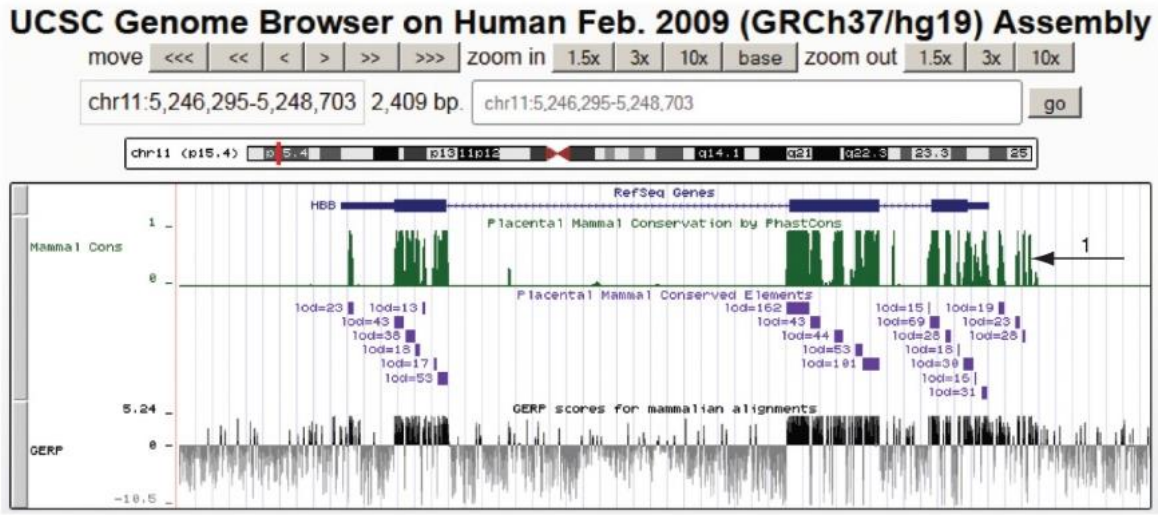
(b) Neighbor-joining tree



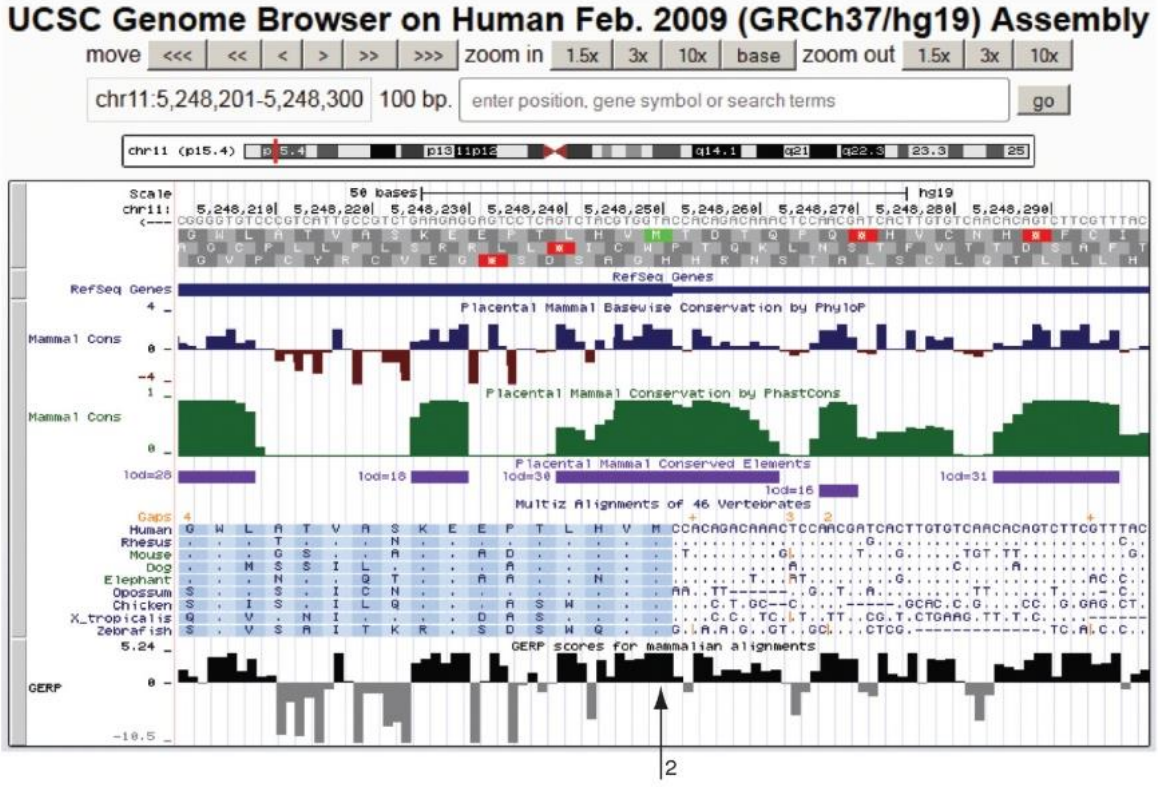
Multiple sequence alignment of genomic DNA

- 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)

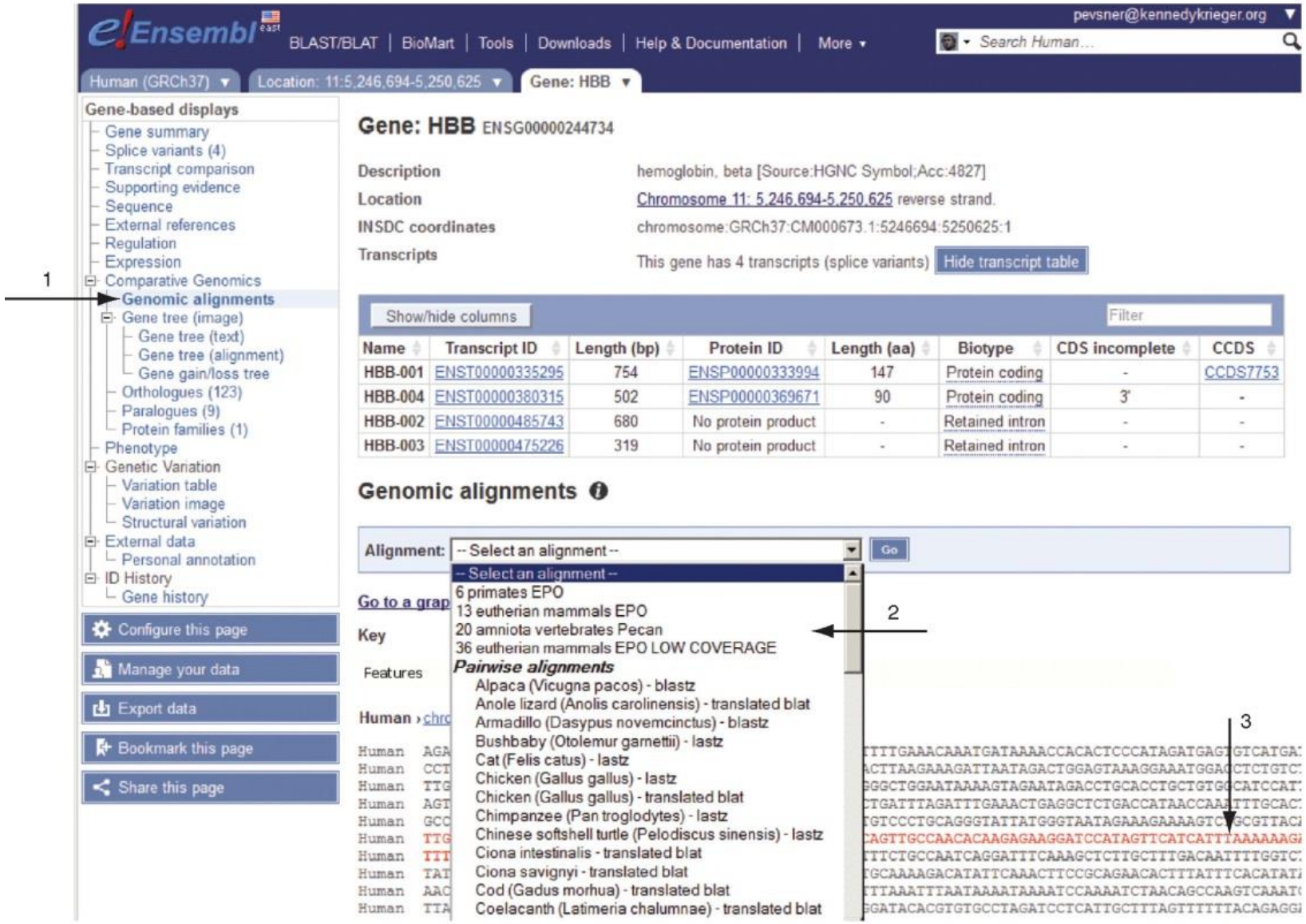


(b) View of *HBB* gene (100 base pairs)



Analyzing multiple sequence alignments at Ensembl

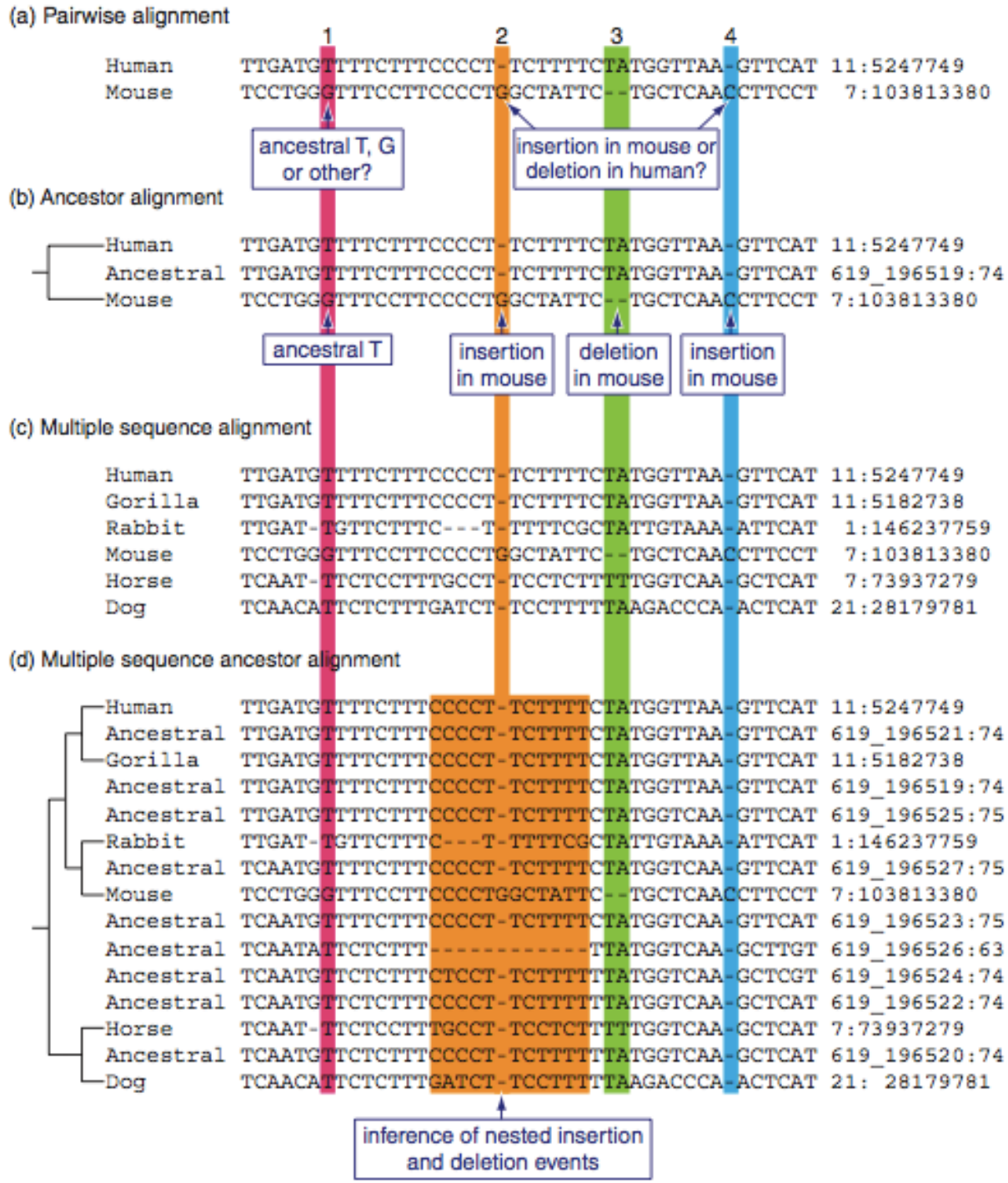
(a) Ensembl entry for *HBB*



Analyzing multiple sequence alignments at Ensembl

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

```
Homo sapiens             11:  5246983 TTCATACCTCTT-ATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGG
Gorilla gorilla gorilla  11:  5181973 TTCATACCTCTT-GTCTTCCTCCCACAGCTCCTGGGCAATGTGCTGG
Pongo abelii             11: 65239065 TTCATACCTCTT-GTCTCCCTCCCACAGCTCCTGGGCAATGTGCTGG
Oryctolagus cuniculus     1:146237264 TTCATGCCTTCT--TCTCTTTCCTACAGCTCCTGGGCAACGTGCTGG
Mus musculus              7:103812810 TTGATGGTTCTT--CCATCTTCCCACAGCTCCTGGGCAATATGATCG
Bos taurus               15: 49339417 CCCTTGCTTAATG-TCTTTTCCACACAGCTCCTGGGCAACGTGCTAG
Bos taurus               15: 49074455 CCCTTGCTTAATG-TCTTTTCCACACAGCTCCTGGGCAACGTGCTGG
Sus scrofa                9:  5633260 CCCTTCCTTTTTA-TCTCTCTCCCACAGCTCCTGGGCAACGTGATAG
Equus caballus            7: 73936736 CCCCCTCTTT-TT-TCTCTTCCCCACAGCTCCTGGGCAACGTGCTGG
Canis lupus familiaris   21: 28179266 CACATGCCTCTTG-TCT--TCCCCACAGCTGCTGGGCAACGTGTTGG
```


(a) Pairwise alignment
1
2
3
4
Human TTGATGTTTTCTTTCCCCT-TCTTTTCTATGGTTAA-GTTCAT 11:5247749
Mouse TCCTGGGTTTCCTTCCCCTGGCTATTC--TGCTCAACCTTCCT 7:103813380
ancestral T, G or other?
insertion in mouse or deletion in human?
(b) Ancestor alignment
Human TTGATGTTTTCTTTCCCCT-TCTTTTCTATGGTTAA-GTTCAT 11:5247749
Ancestral TTGATGTTTTCTTTCCCCT-TCTTTTCTATGGTTAA-GTTCAT 619_196519:74
Mouse TCCTGGGTTTCCTTCCCCTGGCTATTC--TGCTCAACCTTCCT 7:103813380
ancestral T
insertion in mouse
deletion in mouse
insertion in mouse
(c) Multiple sequence alignment
Human TTGATGTTTTCTTTCCCCT-TCTTTTCTATGGTTAA-GTTCAT 11:5247749
Gorilla TTGATGTTTTCTTTCCCCT-TCTTTTCTATGGTTAA-GTTCAT 11:5182738
Rabbit TTGAT-TGTTCTTTC---T-TTTTCGCTATTGTAAA-ATTCAT 1:146237759
Mouse TCCTGGGTTTCCTTCCCCTGGCTATTC--TGCTCAACCTTCCT 7:103813380
Horse TCAAT-TTCTCCTTTGCCT-TCCTCTTTTTGGTCAA-GCTCAT 7:73937279
Dog TCAACATTCTCTTTGATCT-TCCTTTTTAAGACCCA-ACTCAT 21:28179781
(d) Multiple sequence ancestor alignment
Human TTGATGTTTTCTTTCCCCT-TCTTTTCTATGGTTAA-GTTCAT 11:5247749
Ancestral TTGATGTTTTCTTTCCCCT-TCTTTTCTATGGTTAA-GTTCAT 619_196521:74
Gorilla TTGATGTTTTCTTTCCCCT-TCTTTTCTATGGTTAA-GTTCAT 11:5182738
Ancestral TTGATGTTTTCTTTCCCCT-TCTTTTCTATGGTTAA-GTTCAT 619_196519:74
Ancestral TTGATGTTTTCTTTCCCCT-TCTTTTCTATGGTCAA-GTTCAT 619_196525:75
Rabbit TTGAT-TGTTCTTTC---T-TTTTCGCTATTGTAAA-ATTCAT 1:146237759
Ancestral TCAATGTTTTCTTTCCCCT-TCTTTTCTATGGTCAA-GTTCAT 619_196527:75
Mouse TCCTGGGTTTCCTTCCCCTGGCTATTC--TGCTCAACCTTCCT 7:103813380
Ancestral TCAATGTTTTCTTTCCCCT-TCTTTTCTATGGTCAA-GTTCAT 619_196523:75
Ancestral TCAATATTCTCTTT----------TTATGGTCAA-GCTTGT 619_196526:63
Ancestral TCAATGTTCTCTTTCTCCT-TCTTTTTTATGGTCAA-GCTCGT 619_196524:74
Ancestral TCAATGTTCTCTTTCCCCT-TCTTTTTTATGGTCAA-GCTCAT 619_196522:74
Horse TCAAT-TTCTCCTTTGCCT-TCCTCTTTTTGGTCAA-GCTCAT 7:73937279
Ancestral TCAATGTTCTCTTTCCCCT-TCTTTTTTATGGTCAA-GCTCAT 619_196520:74
Dog TCAACATTCTCTTTGATCT-TCCTTTTTAAGACCCA-ACTCAT 21: 28179781
inference of nested insertion and deletion events

Perspective: multiple sequence alignment (MSA)

01

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

02

Key methods include consistency-, iterative-, and structure-based multiple alignment.

03

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.