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)

<table>
<thead>
<tr>
<th>Introduction; definition of MSA; typical uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Five main approaches to multiple sequence alignment</td>
</tr>
<tr>
<td>Exact approaches</td>
</tr>
<tr>
<td>Progressive sequence alignment</td>
</tr>
<tr>
<td>Iterative approaches</td>
</tr>
<tr>
<td>Consistency-based approaches</td>
</tr>
<tr>
<td>Structure-based methods</td>
</tr>
</tbody>
</table>

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

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

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

Note how the region of a conserved histidine (▼) varies depending on which of five prominent algorithms is used.
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  ---------MVHLP EEKSAVTALWGKVNVD-- EVGGRALGLVLYPVWTQRFFES-FG
myoglobin     ---------MGLSDG EWQLVLNVWGKVEADIPHGQEVLRIFKGLPETFKEFDK-FK
neuroglobin   ---------MERPEPELIRQS WRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean       ---------MVAFT EKQDALVSSSEFAFKAIPQYSVVFYTSILEKAAPAA KDLFSF-LA
rice          MALVEDNNAVAVSFSEEQEALV LKSWAILKDSANIALRFFLKL IFEVAPSASQMF S-F-LR

\[\downarrow\]

beta globin  DLSTPDAVMGNPKVKAHGKKVLGAF---SDGLAHL DNLKGFTATLSHELCDKLH---VDPE
myoglobin     HLKSEDEM KASEDLKKHGATVTLA---GGILKKGKHEAEIKPLAQSHATKHK---IPV
neuroglobin   QFSSPEDCLSSPEFLDHIRKVML VI---DAATVNEVELSSLEELASLGKHRAVGVKLS
soybean       NGVDP---TNPKLT GHAEKLFALVRDSAGQLKASGTVVAD---AALGSVHAQKAVTD P
rice          NSDVP---LEKNPKL KTHAMSVFMVTCEAAAQLRKAGKVTVRDTLKRLGATHLKYGVDA


beta globin  NFRLLLGNVLVCVLAHHFGE-FTPPVQAAYQKVVA GVANALAHKYH-----
myoglobin     YLEFISECIIQVLQS KHPGD-FGADAQ GAMN KALELF KDMAS NYELGFQG
neuroglobin   SFSTVG ESLLYMLEK C LPGA-FTPATRAAWSQLYGAVVQAMSRGWDEGE-----
soybean       QFVV VKE AllK TIAAVGD K-WSDELS RAEV EAYDE LA AA I KKA-----
rice          HFEVVKF ALLDTI EEVPADMWS PAMKSAWSEAYDHLVAAIKQEMKPAE-----

Probcons

(c) PROBCONS

beta globin  M---------VHLTPEEKSAVTALWGBKVNVD--EVGGGAEALRLWVYPTQRTFFES-FG
myoglobin    M---------GLSDEGQLNVNLWGKVEADIPGHGQEVRLFLKHPETLEKFDK-FK
neuroglobin  M---------ERPEPELIRQSWRAVSRSPELEHTVLFARLFALEPDDLPLFQYNCR
soybean      M---------VAFTKQDALVSSSFEEAFKANIPQYSVVFYTSILEKAPAADLFSF-LA
rice         MALVEDNNAAVAVSFSEEQEVALKSWAILKKDSANIALRFFLKIIFEVAPSASQMFSF-LR

beta globin  DLSTPDAVMGNPKVAHGKVLGAFSDGLAHLD---NLK---GTFATLSELHCDKLHVDP
myoglobin    HLKSEDEMKAISEDLKKHGATVLTALGGI-LKKGHHE---AEIKPLAQSHATKHKIPV
neuroglobin  QFSSPEDCSSPEFLDHIRVKMLVIDAAAAVTNEVDLSSLE---EYLASLGRKHARV-GVKL
soybean      NGVDP-----TNPKLTGHAEKLFALVIRDSAGQLKASGTVV----ADAALGVSVAQAQ-AVTD
rice         NSDVP--LEKNPKLKTHAMSVFVMTCEAAAAQLRKAEGKTVDTRTLKRLGATHLKY-GVGD

beta globin  ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVLVAGVANALAHK-------YH
myoglobin    KYLEFISECIIQVLSQKH-PGDFFGADQAMKNKALELFRKDMASNYKELGFQG
neuroglobin  SSFSTVGESELLYMLEKCL-GPAFTPATRAAWSQLYGAQAVQAMSRG---W-DGE
soybean      PQFVVVKEALLKTIAAV-GDKWSDELRAEWEVAYDELAAAIAK-------KA
rice         AHFEVVKFALLDTIKEEVPADMSAMPAMSAYEYDHLVAAIQKE---MKPAE
TCoffee

CLUSTAL FORMAT for T-COFFEE Version 5.13

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta globin</td>
<td>MVHLTPEEKSAVTALWGGKVNVD---EVGGEDALGRLLVYPWTQRFFESFG</td>
</tr>
<tr>
<td>myoglobin</td>
<td>MGLSDGWQVLNVKVEADIPHQGEVQLIRFLFKHPELUKEFKD-KFK</td>
</tr>
<tr>
<td>neuroglobin</td>
<td>MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR</td>
</tr>
<tr>
<td>soybean</td>
<td>MAVFTEKQDALVSSSFEAFKANIPQYSYVFYTSILEKAPAANDLSFS-FLA</td>
</tr>
<tr>
<td>rice</td>
<td>MALVEDNNAVAVSFSEEQEAIVLKSWAILKKSANIALRFLKIFEVAPSASQMFSLR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta globin</td>
<td>DLSIPDAVNM酿酒KHKDLKVLGAFSDGLAHLDNL---KGTF---ATLSELHCZLKLHVPD</td>
</tr>
<tr>
<td>myoglobin</td>
<td>HSKSDEMKAESDLKKGATVTLAL---GGILKKKKGHEAE---IKPLAQSHATKHKIPV</td>
</tr>
<tr>
<td>neuroglobin</td>
<td>QFSSPEDECLSSPEFLEDHIRKVMLVIDAVAATVNEDL---SSLLEYLASLGRKH-RAVGVHL</td>
</tr>
<tr>
<td>soybean</td>
<td>NGDVP---TNPKLTGHAEKFLAVSAGKLQASGTVVAD---AALGSAQAQKAVHTDP</td>
</tr>
<tr>
<td>rice</td>
<td>NSDVP---LEKNPKLTHAMSVMFVMTCEAAAQLRKAQKVTVRDTTLKRLGATHLKYGVGLA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta globin</td>
<td>ENFRLGREENVLVCVLHAF-H-GKEFPPVCQAAYKVVAGVANALAHKYH---</td>
</tr>
<tr>
<td>myoglobin</td>
<td>KYLEFISECIIQVLQSKH-PGDFGADQAGMNKALEILFRKDMASNYKELGFQG</td>
</tr>
<tr>
<td>neuroglobin</td>
<td>SFFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRRWGDG---E</td>
</tr>
<tr>
<td>soybean</td>
<td>Q-FVVKEALLKTIKAAV-GDKWSDELRAWEVAYDELAAAIAKKA---</td>
</tr>
<tr>
<td>rice</td>
<td>H-FEVVFALLDITKEEVPADMWSAMKSAWSEAYDHLVAIAKQE---MKPAE</td>
</tr>
</tbody>
</table>
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

http://www.ebi.ac.uk/Tools/msa/clustalw2
(a) Stage 1: series of pairwise alignments

<table>
<thead>
<tr>
<th>SeqA</th>
<th>Name</th>
<th>Length</th>
<th>SeqB</th>
<th>Name</th>
<th>Length</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>beta_globin</td>
<td>147</td>
<td>2</td>
<td>myoglobin</td>
<td>154</td>
<td>25.17</td>
</tr>
<tr>
<td>1</td>
<td>beta_globin</td>
<td>147</td>
<td>3</td>
<td>neuroglobin</td>
<td>151</td>
<td>15.65</td>
</tr>
<tr>
<td>1</td>
<td>beta_globin</td>
<td>147</td>
<td>4</td>
<td>soybean_globin</td>
<td>144</td>
<td>13.19</td>
</tr>
<tr>
<td>1</td>
<td>beta_globin</td>
<td>147</td>
<td>5</td>
<td>rice_globin</td>
<td>166</td>
<td>21.09</td>
</tr>
<tr>
<td>2</td>
<td>myoglobin</td>
<td>154</td>
<td>3</td>
<td>neuroglobin</td>
<td>151</td>
<td>16.56</td>
</tr>
<tr>
<td>2</td>
<td>myoglobin</td>
<td>154</td>
<td>4</td>
<td>soybean_globin</td>
<td>144</td>
<td>8.33</td>
</tr>
<tr>
<td>2</td>
<td>myoglobin</td>
<td>154</td>
<td>5</td>
<td>rice_globin</td>
<td>166</td>
<td>12.99</td>
</tr>
<tr>
<td>3</td>
<td>neuroglobin</td>
<td>151</td>
<td>4</td>
<td>soybean_globin</td>
<td>144</td>
<td>17.36</td>
</tr>
<tr>
<td>3</td>
<td>neuroglobin</td>
<td>151</td>
<td>5</td>
<td>rice_globin</td>
<td>166</td>
<td>18.54</td>
</tr>
<tr>
<td>4</td>
<td>soybean_globin</td>
<td>144</td>
<td>5</td>
<td>rice_globin</td>
<td>166</td>
<td>43.06</td>
</tr>
</tbody>
</table>

ClustalW stage 1: series of pairwise alignments

1. best score (highest percent pairwise identity)
ClustalW stage 1: series of pairwise alignments

<table>
<thead>
<tr>
<th>SeqA</th>
<th>Name</th>
<th>Length</th>
<th>SeqB</th>
<th>Name</th>
<th>Length</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>beta_globin</td>
<td>147</td>
<td>2</td>
<td>myoglobin</td>
<td>154</td>
<td>25.17</td>
</tr>
<tr>
<td>1</td>
<td>beta_globin</td>
<td>147</td>
<td>3</td>
<td>neuroglobin</td>
<td>151</td>
<td>15.65</td>
</tr>
<tr>
<td>1</td>
<td>beta_globin</td>
<td>147</td>
<td>4</td>
<td>soybean_globin</td>
<td>144</td>
<td>13.19</td>
</tr>
<tr>
<td>1</td>
<td>beta_globin</td>
<td>147</td>
<td>5</td>
<td>rice_globin</td>
<td>166</td>
<td>21.09</td>
</tr>
<tr>
<td>2</td>
<td>myoglobin</td>
<td>154</td>
<td>3</td>
<td>neuroglobin</td>
<td>151</td>
<td>16.56</td>
</tr>
<tr>
<td>2</td>
<td>myoglobin</td>
<td>154</td>
<td>4</td>
<td>soybean_globin</td>
<td>144</td>
<td>8.33</td>
</tr>
<tr>
<td>2</td>
<td>myoglobin</td>
<td>154</td>
<td>5</td>
<td>rice_globin</td>
<td>166</td>
<td>12.99</td>
</tr>
<tr>
<td>3</td>
<td>neuroglobin</td>
<td>151</td>
<td>4</td>
<td>soybean_globin</td>
<td>144</td>
<td>17.36</td>
</tr>
<tr>
<td>3</td>
<td>neuroglobin</td>
<td>151</td>
<td>5</td>
<td>rice_globin</td>
<td>166</td>
<td>18.54</td>
</tr>
<tr>
<td>4</td>
<td>soybean_globin</td>
<td>144</td>
<td>5</td>
<td>rice_globin</td>
<td>166</td>
<td>43.06</td>
</tr>
</tbody>
</table>

ClustalW stage 2: create a guide tree

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

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

<table>
<thead>
<tr>
<th>SeqA Name</th>
<th>Len(aa)</th>
<th>SeqB Name</th>
<th>Len(aa)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta_globin</td>
<td>147</td>
<td>myoglobin</td>
<td>154</td>
<td>25</td>
</tr>
<tr>
<td>beta_globin</td>
<td>147</td>
<td>neuroglobin</td>
<td>151</td>
<td>15</td>
</tr>
<tr>
<td>beta_globin</td>
<td>147</td>
<td>soybean</td>
<td>144</td>
<td>13</td>
</tr>
<tr>
<td>beta_globin</td>
<td>147</td>
<td>rice</td>
<td>166</td>
<td>21</td>
</tr>
<tr>
<td>myoglobin</td>
<td>154</td>
<td>neuroglobin</td>
<td>151</td>
<td>16</td>
</tr>
<tr>
<td>myoglobin</td>
<td>154</td>
<td>soybean</td>
<td>144</td>
<td>8</td>
</tr>
<tr>
<td>myoglobin</td>
<td>154</td>
<td>rice</td>
<td>166</td>
<td>12</td>
</tr>
<tr>
<td>neuroglobin</td>
<td>151</td>
<td>soybean</td>
<td>144</td>
<td>17</td>
</tr>
<tr>
<td>neuroglobin</td>
<td>151</td>
<td>rice</td>
<td>166</td>
<td>18</td>
</tr>
<tr>
<td>soybean</td>
<td>144</td>
<td>rice</td>
<td>166</td>
<td>43</td>
</tr>
</tbody>
</table>
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
(a) Stage 1: series of pairwise alignments (closely related globin proteins)

<table>
<thead>
<tr>
<th>SeqA</th>
<th>Name</th>
<th>Length</th>
<th>SeqB</th>
<th>Name</th>
<th>Length</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>human_NP_000509</td>
<td>147</td>
<td>2</td>
<td>Pan_troglodytes_XP_508242</td>
<td>147</td>
<td>100.0</td>
</tr>
<tr>
<td>1</td>
<td>human_NP_000509</td>
<td>147</td>
<td>3</td>
<td>Canis_familiaris_XP_537902</td>
<td>147</td>
<td>89.8</td>
</tr>
<tr>
<td>1</td>
<td>human_NP_000509</td>
<td>147</td>
<td>4</td>
<td>Mus_musculus_NP_058652</td>
<td>147</td>
<td>80.27</td>
</tr>
<tr>
<td>1</td>
<td>human_NP_000509</td>
<td>147</td>
<td>5</td>
<td>Gallus_gallus_XP_444648</td>
<td>147</td>
<td>69.39</td>
</tr>
<tr>
<td>2</td>
<td>Pan_troglodytes_XP_508242</td>
<td>147</td>
<td>3</td>
<td>Canis_familiaris_XP_537902</td>
<td>147</td>
<td>89.8</td>
</tr>
<tr>
<td>2</td>
<td>Pan_troglodytes_XP_508242</td>
<td>147</td>
<td>4</td>
<td>Mus_musculus_NP_058652</td>
<td>147</td>
<td>80.27</td>
</tr>
<tr>
<td>2</td>
<td>Pan_troglodytes_XP_508242</td>
<td>147</td>
<td>5</td>
<td>Gallus_gallus_XP_444648</td>
<td>147</td>
<td>69.39</td>
</tr>
<tr>
<td>3</td>
<td>Canis_familiaris_XP_537902</td>
<td>147</td>
<td>4</td>
<td>Mus_musculus_NP_058652</td>
<td>147</td>
<td>78.91</td>
</tr>
<tr>
<td>3</td>
<td>Canis_familiaris_XP_537902</td>
<td>147</td>
<td>5</td>
<td>Gallus_gallus_XP_444648</td>
<td>147</td>
<td>71.43</td>
</tr>
<tr>
<td>4</td>
<td>Mus_musculus_NP_058652</td>
<td>147</td>
<td>5</td>
<td>Gallus_gallus_XP_444648</td>
<td>147</td>
<td>86.67</td>
</tr>
</tbody>
</table>

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

```
{
    (human_NP_000509:0.00000, Pan_troglodytes_XP_508242:0.00000) :0.05272, Canis_familiaris_XP_537902:0.04932) :0.03231, Mus_musculus_NP_058652:0.12075, Gallus_gallus_XP_444648:0.21259);
```
ClustalW alignment of five closely related beta globin orthologs
Progressive MSA stage 2 of 3: generate a guide tree calculated from the distance matrix (5 distantly related globins)
5 closely related globins

<table>
<thead>
<tr>
<th>SeqA Name</th>
<th>Len(aa)</th>
<th>SeqB Name</th>
<th>Len(aa)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>human_NP_000509</td>
<td>147</td>
<td>2 Pan troglodytes XP_508242</td>
<td>147</td>
<td>100</td>
</tr>
<tr>
<td>human_NP_000509</td>
<td>147</td>
<td>3 Canis familiaris XP_537902</td>
<td>147</td>
<td>89</td>
</tr>
<tr>
<td>human_NP_000509</td>
<td>147</td>
<td>4 Mus musculus NP_058652</td>
<td>147</td>
<td>80</td>
</tr>
<tr>
<td>human_NP_000509</td>
<td>147</td>
<td>5 Gallus gallus XP_444648</td>
<td>147</td>
<td>69</td>
</tr>
<tr>
<td>Pan troglodytes XP_508242</td>
<td>147</td>
<td>3 Canis familiaris XP_537902</td>
<td>147</td>
<td>89</td>
</tr>
<tr>
<td>Pan troglodytes XP_508242</td>
<td>147</td>
<td>4 Mus musculus NP_058652</td>
<td>147</td>
<td>80</td>
</tr>
<tr>
<td>Pan troglodytes XP_508242</td>
<td>147</td>
<td>5 Gallus gallus XP_444648</td>
<td>147</td>
<td>69</td>
</tr>
<tr>
<td>Canis familiaris XP_537902</td>
<td>147</td>
<td>4 Mus musculus NP_058652</td>
<td>147</td>
<td>78</td>
</tr>
<tr>
<td>Canis familiaris XP_537902</td>
<td>147</td>
<td>5 Gallus gallus XP_444648</td>
<td>147</td>
<td>71</td>
</tr>
<tr>
<td>Mus musculus NP_058652</td>
<td>147</td>
<td>5 Gallus gallus XP_444648</td>
<td>147</td>
<td>66</td>
</tr>
</tbody>
</table>

```cpp
{
{
  human_NP_000509: 0.000000,
  Pan troglodytes XP_508242: 0.000000
  : 0.05272,
  Canis familiaris XP_537902: 0.04932
  : 0.03231,
  Mus musculus NP_058652: 0.12075,
  Gallus gallus XP_444648: 0.21259);
```
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
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

Multiple sequence alignment and NJ / UPGMA phylogeny

Input:
Paste protein or DNA sequences in fasta format. Example

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

(b) Alignment of nine globins by MUSCLE (3.8)
(c) Alignment of nine globins by ProbCons (version 1.12)

**ProbCons**

<table>
<thead>
<tr>
<th>Protein</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>hbb_human</td>
<td>M---------VHLTPEAKSAVTALVWGNVVD---EVGGAEALGRLLLVYVPTQORFFES---FG</td>
</tr>
<tr>
<td>hbb_chimp</td>
<td>M---------VHLTPEAKSAVTALVWGNVVD---EVGGAEALGRLLLVYVPTQORFFES---FG</td>
</tr>
<tr>
<td>hbb_dog</td>
<td>M---------VHLTDAEKLAVSCMLAIVND---EVGGAEALGRLLLVYVPTQORFFES---FG</td>
</tr>
<tr>
<td>hbb_mouse</td>
<td>M---------VHLTDAEKLAVSCMLAIVND---EVGGAEALGRLLLVYVPTQORFFES---FG</td>
</tr>
<tr>
<td>hbb_chicken</td>
<td>M---------VHWTAEKLAVSCMLAIVND---EVGGAEALGRLLLVYVPTQORFFES---FG</td>
</tr>
<tr>
<td>myoglobin</td>
<td>M---------VHWTAEKLAVSCMLAIVND---EVGGAEALGRLLLVYVPTQORFFES---FG</td>
</tr>
<tr>
<td>neuroglobin</td>
<td>M---------GLSDGKWLVNKGKBEADPQHPQVBLNLRLFGPHEKLFDK---FK</td>
</tr>
<tr>
<td>soybean</td>
<td>M---------GLSDGKWLVNKGKBEADPQHPQVBLNLRLFGPHEKLFDK---FK</td>
</tr>
<tr>
<td>rice</td>
<td>MALVEDNNAVAFSPQEEALVLKSWAILKDSANLARPFLKTPAPVAPAMQPSR---LR</td>
</tr>
</tbody>
</table>

(d) Alignment of nine globins by T-COFFEE (Expresso version 10.00)

**T-COFFEE**

<table>
<thead>
<tr>
<th>Protein</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1HBB</td>
<td>MVHTLPEAKSAVTALVWGNVVD---EVGGAEALGRLLLVYVPTQORFFES---FG</td>
</tr>
<tr>
<td>hbb_chimp</td>
<td>MVHTLPEAKSAVTALVWGNVVD---EVGGAEALGRLLLVYVPTQORFFES---FG</td>
</tr>
<tr>
<td>2qLSB</td>
<td>MVHTLPEAKSAVTALVWGNVVD---EVGGAEALGRLLLVYVPTQORFFES---FG</td>
</tr>
<tr>
<td>3hrWB</td>
<td>MVHTLPEAKSAVTALVWGNVVD---EVGGAEALGRLLLVYVPTQORFFES---FG</td>
</tr>
<tr>
<td>1hrB</td>
<td>MVHTLPEAKSAVTALVWGNVVD---EVGGAEALGRLLLVYVPTQORFFES---FG</td>
</tr>
<tr>
<td>3RKG</td>
<td>MVHTLPEAKSAVTALVWGNVVD---EVGGAEALGRLLLVYVPTQORFFES---FG</td>
</tr>
<tr>
<td>1oj6A</td>
<td>MVHTLPEAKSAVTALVWGNVVD---EVGGAEALGRLLLVYVPTQORFFES---FG</td>
</tr>
<tr>
<td>4FSL</td>
<td>MVHTLPEAKSAVTALVWGNVVD---EVGGAEALGRLLLVYVPTQORFFES---FG</td>
</tr>
<tr>
<td>1D8U</td>
<td>MVHTLPEAKSAVTALVWGNVVD---EVGGAEALGRLLLVYVPTQORFFES---FG</td>
</tr>
<tr>
<td>cons</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1HBB</td>
<td>GNPVKRHKHKLVGLVFA伪造内容</td>
</tr>
<tr>
<td>hbb_chimp</td>
<td>GNPVKRHKHKLVGLVFA伪造内容</td>
</tr>
<tr>
<td>2qLSB</td>
<td>GNPVKRHKHKLVGLVFA伪造内容</td>
</tr>
<tr>
<td>3hrWB</td>
<td>GNPVKRHKHKLVGLVFA伪造内容</td>
</tr>
<tr>
<td>1hrB</td>
<td>GNPVKRHKHKLVGLVFA伪造内容</td>
</tr>
<tr>
<td>3RKG</td>
<td>GNPVKRHKHKLVGLVFA伪造内容</td>
</tr>
<tr>
<td>1oj6A</td>
<td>GNPVKRHKHKLVGLVFA伪造内容</td>
</tr>
<tr>
<td>4FSL</td>
<td>GNPVKRHKHKLVGLVFA伪造内容</td>
</tr>
<tr>
<td>1D8U</td>
<td>GNPVKRHKHKLVGLVFA伪造内容</td>
</tr>
<tr>
<td>cons</td>
<td>70</td>
</tr>
</tbody>
</table>
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/

MUSCLE Submission Form

MUSCLE stands for **M**Ultiple **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](image)

<table>
<thead>
<tr>
<th>EMAIL</th>
<th>RESULTS</th>
<th>ALIGNMENT TITLE</th>
<th>OUTPUT FORMAT</th>
<th>OUTPUT TREE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>interactive</td>
<td>Sequence</td>
<td>fasta</td>
<td>none</td>
</tr>
</tbody>
</table>

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

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, Prrip, 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

<table>
<thead>
<tr>
<th>Sequence x</th>
<th>$x_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence y</td>
<td>$y_j$</td>
</tr>
<tr>
<td>Sequence z</td>
<td>$z_k$</td>
</tr>
</tbody>
</table>

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.
<table>
<thead>
<tr>
<th>Organism</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta globin</td>
<td>M--------VHLTPPEKSATLWVGKVNVD--EYGGGEALGRLLVVYPWTQRFFES-FG</td>
</tr>
<tr>
<td>myoglobin</td>
<td>M---------GLSDGEWQLVLYNVGVKEADIPGHQEVOLVEFLKGFPETLEKFDK-FK</td>
</tr>
<tr>
<td>neuroglobin</td>
<td>M---------ERPEPHELIRQSWRAVSRSPLEHGTVLFARLFALFEDDLLLFQYNCR</td>
</tr>
<tr>
<td>soybean</td>
<td>M---------VAFTKDQALVSSSSFEAKAFKANIPQYSVVFYSILEKAPAAKLDFSF-LA</td>
</tr>
<tr>
<td>rice</td>
<td>MALVEDNNAVASFSEEQEOALVKSWAILKKSANIALRFFLKIFEVAPSAASQMFSLR</td>
</tr>
</tbody>
</table>

**Consistency iteration helps**
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:

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

Access to TCoffee: http://tcoffee.org
APDB ClustalW output:
TCoffee can incorporate structural information into a MSA
Outline: multiple sequence alignment (MSA)

Introduction; definition of MSA; typical uses
Five main approaches to multiple sequence alignment
   1. Exact approaches
   2. Progressive sequence alignment
   3. Iterative approaches
   4. Consistency-based approaches
   5. Structure-based methods

Benchmarking studies: approaches, findings, challenges

Databases of Multiple Sequence Alignments
   1. Pfam: Protein Family Database of Profile HMMs
   2. SMART
   3. Conserved Domain Database

Integrated multiple sequence alignment resources

MSA database curation: manual versus automated

Multiple sequence alignments of genomic regions
   1. 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 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).

BaliBase: comparison of multiple sequence alignment algorithms

<table>
<thead>
<tr>
<th>Name</th>
<th>hiv-1 protease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sequences</td>
<td>4</td>
</tr>
<tr>
<td>Alignment Length</td>
<td>106</td>
</tr>
<tr>
<td>Longest Sequence</td>
<td>104</td>
</tr>
<tr>
<td>Shortest Sequence</td>
<td>98</td>
</tr>
<tr>
<td>Average Percent Identity</td>
<td>49</td>
</tr>
<tr>
<td>Maximum Percent Identity</td>
<td>86</td>
</tr>
<tr>
<td>Minimum Percent Identity</td>
<td>35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>SWISSPROT Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>1fmb</td>
<td>P32542</td>
</tr>
<tr>
<td>7upjB</td>
<td>P03366</td>
</tr>
<tr>
<td>pol_sivcz</td>
<td>P17283</td>
</tr>
<tr>
<td>POL_SIVMK</td>
<td>P05897</td>
</tr>
</tbody>
</table>

Family 1fmb 7upjB pol_sivcz POL_SIVMK

1fmb
---
1
vTYNELEKRPTTIVLINDTPLNVLLDTGADTSVLT
Tahynr lkyrgk.YQ

7upjB
---
1
pQFSLWKRPVVTAYIEGQPVEVLDDTGADDSIVAG....iel.gnn.QS

pol_sivcz
---
1
pQITLWQRLIPVKEGQLCEALLDTGADTVIER....iqqlqgl..UK

POL_SIVMK
---
1
pQFSLWRPPVVTAHIEGQPVEVLDDTGADDSIVTG....iel.gph.YT

1fmb
---
50
GTGQGCGGNVEQFS.TPVTIKKKGHRHFKRMLVADIPVTLGRDILQDL

7upjB
---
44
PKIVGGGFGINTLEYKNVEIEVLKVRATINTGDPINIFGRNILTAL

pol_sivcz
---
44
PKHMGIGGFGIKVKQFDNQHIEIEGRKVVGTVLGVGTPVNIIGRNILTQL

POL_SIVMK
---
44
PKIVGGGFGINTKEYKNVEIEVLGKRKRTIMTGDPINIFGRNLLLTAL

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)
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
(a) Pfam alignments

(b) Pfam seed alignment
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

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

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)

(b) View of HBB gene (100 base pairs)
Analyzing multiple sequence alignments at Ensembl

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>11</td>
<td>5246983   TTCATAACCTCTT-ATTTTCCTCCTCCACAGCTCCTGCGGAAACAGTATGCTGG</td>
</tr>
<tr>
<td>Gorilla gorilla</td>
<td>11</td>
<td>5181973   TTCATAAATCTTTTGCTCCCTCCTCCACAGCTCCTGCGGAAACAGTATGCTGG</td>
</tr>
<tr>
<td>Pongo abelii</td>
<td>11</td>
<td>65239065  TTCATAACCTCTT-GTCTCCCTCCACAGCTCCTGCGGAAACAGTATGCTGG</td>
</tr>
<tr>
<td>Oryctolagus cuniculus</td>
<td>1:146237264</td>
<td>TTCATGGCTTTGCTCCTTTTACAGCTCCTGCGGAAACAGTATGCTGG</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>7:103812810</td>
<td>TTGATGGTTTCTT-GGATCCCTCCTCCACAGCTCCTGCGGAAACAGTATGCTGG</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>15: 49339417</td>
<td>CCTTGGCTTATGCTTTTTACAGCTCCTGCGGAAACAGTATGCTGG</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>15: 49074455</td>
<td>CCTTGGCTTATGCTTTTTACAGCTCCTGCGGAAACAGTATGCTGG</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>9:   5633260</td>
<td>CCCCCTTTTTTA-TCTCTCTCCACAGCTCCTGCGGAAACAGTATGCTGG</td>
</tr>
<tr>
<td>Equus caballus</td>
<td>7:   73936736</td>
<td>CCCCCTTTTT-TT-TCTCTCTCCACAGCTCCTGCGGAAACAGTATGCTGG</td>
</tr>
<tr>
<td>Canis lupus familiaris</td>
<td>21: 28179266</td>
<td>CACATGGCTTGT-TCT--TCCCCACAGCTGCTGCGGAAACAGTATGCTGG</td>
</tr>
</tbody>
</table>
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
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.