Multiple sequence alignment
Multiple sequence alignment: today’s goals

• to define what a multiple sequence alignment is and how it is generated; to describe profile HMMs

• to introduce databases of multiple sequence alignments

• to introduce ways you can make your own multiple sequence alignments

• to show how a multiple sequence alignment provides the basis for phylogenetic trees
Multiple sequence alignment: outline

[1] Introduction to MSA

  1) Exact
  2) Progressive (ClustalW)
  3) Iterative (MUSCLE, MAFFT)
  4) Consistency (ProbCons)
  5) Structure-based (Expresso, PRALINE)

Conclusions from benchmarking

[3] Databases of MSAs (hidden Markov models)

[4] Multiple alignment of genomic regions

[5] MEGA to make a multiple sequence alignment
Multiple sequence alignment: definition

• a collection of three or more protein (or nucleic acid) sequences that are partially or completely aligned

• homologous residues are aligned in columns across the length of the sequences

• residues are homologous in an evolutionary sense, and, in a structural sense
Multiple sequence alignments

• The challenge of alignment is to establish the site-wise conservation obscured by evolution

• Sequences can act as intermediates between highly dissimilar sequences and can connect these fairly distantly related sequences into an alignment

• The resulting alignment will better reflect evolutionary forces
Multiple sequence alignment: properties

Generally align proteins:

- nucleotides less well-conserved

- nucleotide sequences are less informative (fewer characters) -> harder to align with high confidence

How do you know if you have the “correct” alignment of a protein family? Is there one “correct” alignment?

- for two proteins sharing 30% amino acid identity, about 50% of the individual amino acids are superposable in the two structures
Proportion of structurally superposable residues in pairwise alignments as a function of sequence identity

After Chothia & Lesk (1986)
Multiple sequence alignment: outline

[1] Introduction to MSA

   1) Exact
   2) Progressive (ClustalW)
   3) Iterative (MUSCLE, MAFFT)
   4) Consistency (ProbCons)
   5) Structure-based (Expresso, PRALINE)
   Conclusions from benchmarking

[3] Databases of MSAs (hidden Markov models)

[4] Multiple alignment of genomic regions

[5] MEGA to make a multiple sequence alignment
CLUSTAL W (1.83) multiple sequence alignment

beta globin
-----------MVHLPKEKSAVNTALWGVKNV--EVPGEALGRLLVYPTQRFFESFG- 47
myoglobin
-----------MGDSDEWQLVLNVGVNDKIPDHGQPEVLIRLFLKHPETLEFKDFK- 48
neuroglobin
-----------MERPEPELRQSWRAVNSFPLEHGTGTFARLFALPDLPFLQYNCR 47
soybean
-----------MVAFTREQODALVSSSSFEEFKANIPQYSVVFYTSILEKAPAKDLFSFLA- 49
rice
MALVEDNNAVAVSFSSEQEALVKSAILKDSANIALRFFLKIFEVAPSASQMSFLR- 59

beta globin
DLSTPDAVMGNPKVKAHGKVLGAFSDGLAHLDNLKGTFA---ELHCDKLVHVDPE 102
myoglobin
HLKSEDEMKADELKKGATVLTALGGILKKKHIGHEAEIKPLA---QSHATKHKPVK 103
neuroglobin
QFSSEPDCLEPSPELDHIRKVMVLDAAVNTVEDLSSLEELYLS---LGRKHRAVGKLS 104
soybean
--NGVDP--NPKLTGHAELKFAVSDLSSAGQLKASGTVVADA----LGSVHAKQAVTDP 101
rice
--NADVPAEKNPKLTKHAMSVMVFYMTCEAAQLRKGKVTDDTXTLRLGATHLYGVD 117

beta globin
NFRLLGNVLCVLCVLAHKHF--GKEFTPPVQAAYQKVYAGVANALAHKYH----- 147
myoglobin
YLFISFECIIQVLSKHPGDFGADAOQAMKALEELFRKDSMYKELGFQG 154
neuroglobin
SFSTVGESLLYMLEKCL-PATEPAFTPATRAAWSQLYGAVVQAMSRGWDE---- 151
soybean
QFVVVKEALLKTIKAAV--GDWSDELRSRAWEVAYDELAAAIIKA-------- 144
rice
HFEVVVFALLDTIKEEVPADMWSPAMSWSSEAYDLVAAIKQEMKPAE--- 166
Praline

(a) Praline multiple sequence alignment

<table>
<thead>
<tr>
<th></th>
<th>beta globin</th>
<th>myoglobin</th>
<th>neuroglobin</th>
<th>soybean</th>
<th>rice</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta globin</td>
<td>MVHTPEEKSATCIVGK..NVDEVGGEALGRLLVYPWTQRFFES.FG</td>
<td>MGLSDGEWQLVLNVKWVVEADIPGHQGEVILRKFLGHPETLEKFDK.FK</td>
<td>MERPELIRQSWMRSPLERHTGVLVARLFALAFPDDLPLFQYNCR</td>
<td>MVAFTKEQDALVSSEFAFKANIPQYSVVFYTSILEKAPAAPKDLFS..FL</td>
<td>MALVEDNNNAVHSFSESEQKLVLSWAILKKDSANIALRRFFLMKIEVAPSASQMFS..FL</td>
</tr>
<tr>
<td>myoglobin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>neuroglobin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>soybean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistency</td>
<td>00000000000142654382579345734633643463642453686433<em>35344</em>50063</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[beta globin] DLSTPDAVMGNPKVAKAVGHKVLGAFSGDGLAHLDNLKGTAFLSL..HCDKLH....VDP
[myoglobin] HLKSEDNEMKASEDLKKGATVLTALGGILKKKHegaEIKPLAQ..HATKHK....IPV
[neuroglobin] QFSSPEDCLSSPEFILDHVRKVMVIAAAVNTVEALSSLEELALSILGRKHRAVO..VKL
[soybean] A.NGVD..TNPKLTGHAEKLFALVRDAGQL.KASVTGVDAA..LGVSHEAQKAVTD
[rice] R.NSDVPKLEKNPKTHAMSVFVMTCAAAQL.RKAGKVTVRDTTTLKLGATHLKYGVGD

Consistency 3166354224776653*4368635424454451335634333542003335440009022

[beta globin] ENFRLLLGNVLVVLCAHLHNF..GKEFPDPPVQAAYQKVGVAVANALAHKYYH....
[myoglobin] KYLEFISECIIQVLQSKH..PGDFGADAQGMKNKALELFRKDMSNYKELGFOQG
[neuroglobin] SSFSTVGESLLYMLEKCL.GPAFTPATARAWSQYGAVQAMPAGWD..GE..
[soybean] PQFVVVKEALLKTIKAV.GDKWSSDELRSRAWEYVAYDELAAAIIKAK....
[rice] AHFEVVKFALLTDIKEEVPADMSPAMKSAWSEAYDHLLVAIAIQEMKPAE...

Consistency 43744844498258542305336554454*5546542644646754322001000
MUSCLE

(b) MUSCLE (3.6) multiple sequence alignment

beta globin
----------MVHLPSSAVTALW GKVNVD--EVGGEALGRLLVVVYPWTQRFFES-FG
myoglobin
----------MGLSDGEWLVLNW GKVEADIPGHGQEVLRFLFKGHPETLEKFDK-FK
neuroglobin
----------MERPEPELIRQS WRASVRSPLEHGTVLFARLFALAPDLPLLFLQYNCR
soybean
----------MVAETFEKQDALVSSSFEAKANIPQYSVVFYTSILEKAPA AKDLFSF-LA
rice
MALVEDNNAVAVSFSEQEALVKLKSWAILKKSANIALRFLKIFEVAPSASQMFSF-LR

\n
beta globin
DLSTPDAVMGNPKVKAHGKKVLGAF---SDGLAHLDNLKGTATLSELHCDKLH--VDPE
myoglobin
HLKSEDEMKASEDLKHKHGATVLTA--GGILKKGKHHEAEIKPLAQSHATKHK--IPVK
neuroglobin
QFSSPEDCLSSPEFLDHIRKVMLVI---DAAVTVEDLSLEELASLGRKHRAVGVKLS
soybean
NGVDP----TNPKLTGHAELKFALVRDSAGQLKASGTVD----AALGSVHAQKAVTDP
rice
NSDVP---LEKNPKLKTAMSVFVMTCFAAQLRKKAGKVTRDDTLKRLGATHLKYGVDA

beta globin
NFRLLGNNVLCVCLAHHFGE-FTPPVQAAYQKVVGAVANALAHKYH------
myoglobin
YLEFISECIIQVLQSKHPGD-FGADAQGAMNKENLELFKDMASNKYLELGFGQG
neuroglobin
SFSTVGESLLYMLEKCLGPA-FTPATRAAWSQLYGAVVQAMSRGWDE----
soybean
QFVVVKEALLKTIKAAVGDK-WSDELSRAEVEAYDELAAAIIKKA---------
rice
HFEVVKFAALDITKEEVPADMWSVAMKSAWSEAYDLVAAIKQEMKPAE----
<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta globin</td>
<td>M----------------------VHLTPEEKSAVTALWGKVNVD--EYGGEALGRLLVVYPWTQRFFES-FG</td>
</tr>
<tr>
<td>myoglobin</td>
<td>M----------------------GSLDGEWQLVNWVGBVEADIPGHQEQVFLIRLFKHFETLEKFDK-FK</td>
</tr>
<tr>
<td>neuroglobin</td>
<td>M----------------------ERPEPELRQSWRAPPSSPSEHGTVLFAFLFALPEPDLPLFQYPNCR</td>
</tr>
<tr>
<td>soybean</td>
<td>M----------------------VAFTKODAVSSSFEAFKANNIPQYSSVFYTSILEKAPAADLSF-AL</td>
</tr>
<tr>
<td>rice</td>
<td>MALVEDANNAVSSEEEQAEALVLYKSWAILKKDSANIALRFLKLKIFEVAPSASQMFSL-LR</td>
</tr>
<tr>
<td>beta globin</td>
<td>DLSTPDAVMGNPVKAHGKhKGVLGAFSDGLAHLD---NLK---GTFATLSELHCDKLHVDP</td>
</tr>
<tr>
<td>myoglobin</td>
<td>HLKSEDEMKASENDLKKHAGTVALTAGG1---LKKKHGHE---AEIKPLAQSHATKHKIPV</td>
</tr>
<tr>
<td>neuroglobin</td>
<td>QFSSPDCLSSPEFLDHIKIVMLVDAATVIVEDLSSE---EYLASLGRKRAV-GVKL</td>
</tr>
<tr>
<td>soybean</td>
<td>NGVDP------TPKLTHGAELKLFALVDSAGQLKASGTVV------ADAALGVHFAQK-AVTD</td>
</tr>
<tr>
<td>rice</td>
<td>NSDVP---LEKNPKLTTHAMSVMFVTCEAIAQLRKAGKVTVDRTTTLRGLATHLKY-GVGD</td>
</tr>
<tr>
<td>beta globin</td>
<td>ENFRLLGNVLVCVLAHHF--GKEFTPPVQAAYQKVAGVANALAHK------YH</td>
</tr>
<tr>
<td>myoglobin</td>
<td>KYLEFISECIIQVLQSKH-PGDFGADAOQGAMKNKALELFRKDMASYKELGFQG</td>
</tr>
<tr>
<td>neuroglobin</td>
<td>SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRG---W-DGE</td>
</tr>
<tr>
<td>soybean</td>
<td>PQFVVVEKAllekTIKAAV-GDKWSDELSRAEVEVAYDELAIAK-----------------KA</td>
</tr>
<tr>
<td>rice</td>
<td>AHFEVVVKFALLDTIKEEVPAADMWSPAMKSAWSEAYDHLVAAIKAQ---MKPAE</td>
</tr>
</tbody>
</table>
TCooffee

CLUSTAL FORMAT for T-COFFEE Version 5.13

beta globin
myoglobin
neuroglobin
soybean
rice

beta globin
myoglobin
neuroglobin
soybean
rice

beta globin
myoglobin
neuroglobin
soybean
rice
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 (Words) identity, about 50% of the individual amino acids are superposable in the two structures (Meaning)
Multiple sequence alignment: features

- some aligned residues, such as cysteine 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

- Population data can be analyzed in a MSA (PopSet)

- A single query can be searched against a database of MSAs (e.g. PFAM, Blocks, CDD, etc.)

- Regulatory regions of genes may have consensus sequences identifiable by MSA
Multiple sequence alignment: outline

[1] Introduction to MSA

   1) Exact
   2) Progressive (ClustalW)
   3) Iterative (MUSCLE, MAFFT)
   4) Consistency (ProbCons)
   5) Structure-based (Expresso, PRALINE)
Conclusions from benchmarking

[3] Databases of MSAs (hidden Markov models)

[4] Multiple alignment of genomic regions

[5] MEGA to make a multiple sequence alignment
Multiple sequence alignment: methods

Exact methods: dynamic programming
Instead of the 2-D dynamic programming matrix in the Needleman-Wunsch technique, think about a 3-D, 4-D or higher order matrix.

Exact methods give optimal alignments but are not feasible in time or space for more than ~10 sequences. Still an extremely active research field. Useful not only in bioinformatics, but also in language theory.
Multiple sequence alignment: methods

Progressive methods: use a guide tree (a little like a phylogenetic tree but NOT a phylogenetic tree) to determine how to combine pairwise alignments one by one to create a multiple alignment.

Making multiple alignments using trees was a very popular subject in the 1980s. Fitch and Yasunobu (1974) may have first proposed the idea, but Hogeweg and Hesper (1984) and many others worked on the topic.

Feng and Doolittle (1987) made one important contribution that got their names attached to this alignment method.

Examples: ClustalW, MUSCLE
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: IterAlign, Praline, MAFFT
Multiple sequence alignment: methods

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

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

Examples: T-COFFEE, Prrp, DiAlign, ProbCons
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.
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 is the most popular program. It has a nice interface (especially with CLUSTALX) and is easy to use.

BUT IT IS NOT THE ONLY CHOICE! AND NOT THE BEST CHOICE!
Multiple sequence alignment: outline

[1] Introduction to MSA

   1) Exact
   2) Progressive (ClustalW)
   3) Iterative (MUSCLE, MAFFT)
   4) Consistency (ProbCons)
   5) Structure-based (Expresso, PRALINE)

Conclusions from Benchmarking

[3] Databases of MSAs (hidden Markov models)

[4] Multiple alignment of genomic regions

[5] MEGA to make a multiple sequence alignment
Multiple sequence alignment: methods

Example of MSA using ClustalW: two data sets

Five distantly related lipocalins (human to *E. coli*)

Five closely related RBPs

When you do this, obtain the sequences of interest in the FASTA format!
(You can save them in a Word document)
Get sequences from Entrez Protein
You can display sequences from Entrez Protein in the fasta format

>gi|55743122|ref|NP_006735.2| retinol-binding protein 4, plasma precursor
MKWWALLLAALGSGAERDCRVSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAESAIFSVDEGTQGMSATAKGRVRLNNWVDVCADMVGTFTDTDEDPAKFKMKYWGVASFLQKGNDDHIVDTDYAVYQYRCRLLNLGDTCADSYSFVSRDPNGLPPEAQKIVQRQEEELCLARQYRILVHGNCYCDGRSERNL

>gi|12843160|dbj|BAB25881.1| unnamed protein product [Mus musculus]
MEWWALLLLALGCGGSAERDCRVSFRVKENFDKARFSGLWYAIAKKDPEGLFLQDNIIAESFVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTDEDPAKFKMKYWGVASFLQRGNNDDHWIIDTDYDTFALQYSCRLQNLGDTCADSYSFVSRDPNGLPETRRRLVRQREELCLERQYRWRIEHNGYCQSPRSRNSL

>gi|4502163|ref|NP_001638.1| apolipoprotein D precursor [Homo sapiens]
MVMLLLLLSLAAGLFGAEGQAFHGLGKCPNNPQPQENFDVNYKLGRYETEKIPTTFENGRCIQANYSLME NGKIKVNLQ RELADGTVNQIEGEGATPVNLTEPAKLEKVFWSWMPASAPYWI LATDYEN YALVY SCTCCIQLFHVDFAWILARNPNPLLPTETVDSLKNILTSSNIDVKKMTVTDQVNPCKLS

>gi|127528|sp|P11590|MUP4_MOUSE Major urinary protein 4 precursor (MUP 4)
MKLLLLGLTLVCHAEETSKGQNLNEKINGEWFSLILASDKREKIEEHSRMVRFVHEHVLENSLAF KFHTVIDEGCESEIFLVADKTEKAGYESVMDGFNFTFTILKTDYDYNYIMFHLINEKDGKTQMLEMYGRKA DLNSDIKEFKVCLCEEHGIITKENIIDLTKTNRLKARE

>gi|732003|sp|P39281|BLC_ECOLI Outer membrane lipoprotein bhc precursor
MRLPLAATAAATATTAFLVACSSPTPPRGTGTVNNDNFDAKRYLGTWEIAREFDHHRFERGKLTATYSLRDDDGCLNVINKYNPDRGMWQQSEGKAYFTGAPTRAALKVSFFGFYGGYVIALDREYRHALVCGPDRDYLWI LSRTPTISDEVKQEMLAVATREGFDVSKFIFWVQPZS
When you get a DNA sequence from Entrez Nucleotide, you can click CDS to select only the coding sequence.

This is very useful for phylogeny studies.
[1] Enter a query at NCBI such as globin
[2] Click on HomoloGene (left side)
[3] Choose a HomoloGene family, and view in the fasta format
Use ClustalW to do a progressive MSA

http://www2.ebi.ac.uk/clustalw/
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

Sequence format is Pearson
Sequence 1: gi|5803139|ref|NP_006735.1| 199 aa
Sequence 2: gi|6174963|sp|Q00724|RETB_MOUS 201 aa
Sequence 3: gi|132407|sp|P04916|RETB_RAT 201 aa
Sequence 4: gi|89271|pir|A39486 201 aa
Sequence 5: gi|132403|sp|P18902|RETB_BOVIN 183 aa
Progressive MSA stage 1 of 3: generate global pairwise alignments

Number of pairwise alignments needed: For $n$ sequences, $C = \frac{(n-1)(n)}{2}$, $n=5$, $C = \frac{4*5}{2} = 10$

Start of Pairwise alignments

Aligning...

1. Sequences (1:2) Aligned. Score: 84
2. Sequences (1:3) Aligned. Score: 84
7. Sequences (2:5) Aligned. Score: 85

five closely related lipocalins

Best score
CLUSTAL W (1.81) Multiple Sequence Alignments

Sequence format is Pearson
Sequence 1: gi|5803039|ref|NP_006735.1| 199 aa
Sequence 2: gi|12843160|dbj|BAB25881.1| 201 aa
Sequence 3: gi|4502163|ref|NP_001638.1| 189 aa
Sequence 4: gi|127528|sp|P11590|MUP4_MOUSE 178 aa
Sequence 5: gi|732003|sp|P39281|BLC_ECOLI 177 aa

Start of Pairwise alignments
Aligning...

<table>
<thead>
<tr>
<th>Sequences</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>(4:5)</td>
<td>7</td>
</tr>
<tr>
<td>(3:4)</td>
<td>9</td>
</tr>
<tr>
<td>(2:3)</td>
<td>17</td>
</tr>
<tr>
<td>(2:4)</td>
<td>9</td>
</tr>
<tr>
<td>(3:5)</td>
<td>27</td>
</tr>
<tr>
<td>(2:5)</td>
<td>19</td>
</tr>
<tr>
<td>(1:2)</td>
<td>84</td>
</tr>
<tr>
<td>(1:3)</td>
<td>14</td>
</tr>
<tr>
<td>(1:4)</td>
<td>8</td>
</tr>
<tr>
<td>(1:5)</td>
<td>12</td>
</tr>
</tbody>
</table>

Guide tree file created: [/net/nfs0/vol1/production/w3nobody/tmp/838554.269763-180145.dnd]

Start of Multiple Alignment
There are 4 groups
Aligning...

<table>
<thead>
<tr>
<th>Group</th>
<th>Sequences</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4080</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Delayed</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Delayed</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Delayed</td>
</tr>
</tbody>
</table>

Sequence: 3 Score: 1544
Sequence: 5 Score: 1408
Sequence: 4 Score: 1239
Alignment Score 1459

CLUSTAL-Alignment file created  [/net/nfs0/vol1/production/w3nobody/tmp/838554.269763-180145.aln]

five distantly related lipocalins

best score
Feng-Doolittle stage 2: guide tree

1. Convert similarity scores to distance scores
2. A tree shows the distance between objects
3. Use UPGMA (defined in the phylogeny lecture)
4. ClustalW provides a syntax to describe the tree

A guide tree is not a phylogenetic tree
Progressive MSA stage 2 of 3: generate guide tree

( (gi|5803139|ref|NP_006735.1|:0.04284, (gi|6174963|sp|Q00724|RETB_MOUS:0.00075, gi|132407|sp|P04916|RETB_RAT:0.00423) :0.10542) :0.01900, gi|89271|pir||A39486:0.01924, gi|132403|sp|P18902|RETB_BOVIN:0.01902);

five closely related lipocalins
Progressive MSA stage 2 of 3: generate a guide tree calculated from the distance matrix

five distantly related lipocalins
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.”
Clustal W alignment of 5 closely related lipocalins

CLUSTAL W (1.82) multiple sequence alignment

| gi|89271|pir||A39486 | MEWVWALVLLAALGSAQAERDCRVSSFRVKENFDKARFSGTWYAMAKKDP 50
| gi|132403|sp|P18902|RETB_BOVIN | ------------------
| gi|5803139|ref|NP_006735.1| | ERDCRVSSFRVKENFDKARFAGTWYAMAKKDP 32
| gi|6174963|sp|Q00724|RETB_MOUS | MKWVWALLLLAAW--AAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDP 48
| gi|132407|sp|P04916|RETB_RAT | MEWVWALVLLAALGGGSAERDCRVSSFRVKENFDKARFSGLWYAIAMAKKDP 50
| gi|5803139|ref|NP_006735.1| | MEWVWALVLLAALGGGSAERDCRVSSFRVKENFDKARFSGLWYAIAMAKKDP 50
| gi|6174963|sp|Q00724|RETB_MOUS | ********************:* ***:*****
| gi|132407|sp|P04916|RETB_RAT |

| gi|89271|pir||A39486 | EGLFLQDNIVAEFSVDENGHMSATAKGRVRLNNNDVCDMVGTFDTDED 100
| gi|132403|sp|P18902|RETB_BOVIN | EGLFLQDNIVAEFSVDENGHMSATAKGRVRLNNNDVCDMVGTFDTDED 82
| gi|5803139|ref|NP_006735.1| | EGLFLQDNIVAEFSVDENGHMSATAKGRVRLNNNDVCDMVGTFDTDED 98
| gi|6174963|sp|Q00724|RETB_MOUS | EGLFLQDNIVAEFSVDENGHMSATAKGRVRLNNNDVCDMVGTFDTDED 100
| gi|132407|sp|P04916|RETB_RAT | EGLFLQDNIVAEFSVDENGHMSATAKGRVRLNNNDVCDMVGTFDTDED 100
| gi|5803139|ref|NP_006735.1| | ********************:* ***:*****
| gi|6174963|sp|Q00724|RETB_MOUS |
| gi|132407|sp|P04916|RETB_RAT |

| gi|89271|pir||A39486 | PAKFKMKYWGVASFLQKGNDDHWIITDTDYTAYAQYSCRLQNLNGTCADS 150
| gi|132403|sp|P18902|RETB_BOVIN | PAKFKMKYWGVASFLQKGNDDHWIITDTDYTEFVAQYSCRLNNDGTCADS 132
| gi|5803139|ref|NP_006735.1| | PAKFKMKYWGVASFLQKGNDDHWIITDTDYAVQYSCRLNLNGTCADS 148
| gi|6174963|sp|Q00724|RETB_MOUS | PAKFKMKYWGVASFLQKGNDDHWIITDTDYTFLAQYSCRLQNLNGTCADS 150
| gi|132407|sp|P04916|RETB_RAT | PAKFKMKYWGVASFLQKGNDDHWIITDTDYTFLAQYSCRLQNLNGTCADS 150
| gi|5803139|ref|NP_006735.1| | ********************:* ***:*****
| gi|6174963|sp|Q00724|RETB_MOUS |
| gi|132407|sp|P04916|RETB_RAT |

* asterisks indicate identity in a column
Progressive MSA stage 3 of 3: progressively align the sequences following the branch order of the tree

Distantly related lipocalins
Progressive MSA stage 3 of 3: CLUSTALX output

Note that you can download CLUSTALX locally, rather than using a web-based program!
Progressive MSA stage 3 of 3:
Why following the branch order of the tree?
Order matters

Adapted from C. Notredame, Pharmacogenomics 2002
Progressive MSA stage 3 of 3: Why following the branch order of the tree? Order matters
Why “once a gap, always a gap”?

- There are many possible ways to make a MSA
- Where gaps are added is a critical question, and the main task
- 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 in the closer sequences are more 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.

- Similar to the pairwise alignment, scoring matrices are varied dependent on the divergent of the sequences:
  
<table>
<thead>
<tr>
<th>Scoring Matrix</th>
<th>Identity Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM20</td>
<td>80-100% id</td>
</tr>
<tr>
<td>PAM60</td>
<td>60-80% id</td>
</tr>
<tr>
<td>PAM120</td>
<td>40-60% id</td>
</tr>
<tr>
<td>PAM350</td>
<td>0-40% id</td>
</tr>
</tbody>
</table>

- Residue-specific gap penalties are applied.
Multiple sequence alignment: outline

[1] Introduction to MSA

   1) Exact
   2) Progressive (ClustalW)
   3) Iterative (MUSCLE, MAFFT)
   4) Consistency (ProbCons)
   5) Structure-based (Expresso, PRALINE)

Conclusions from Benchmarking

[3] Databases of MSAs (hidden Markov models)

[4] Multiple alignment of genomic regions

[5] MEGA to make a multiple sequence alignment
Multiple sequence comparison by log-expectation. More information.

Paste sequences in FASTA format:

```
>gi|127528|sp|P11590|MUP4_MOUSE Major urinary protein 4 precursor (MUP 4)
MKLLLCGLTLLCIAHAEATSKQNLMWKEINGEWFILLASDKEKIEEHGSMRFVEHIHLHENSLAF
KFHTVIDGECEISEIFLVADKTEKAGEYSVMYDFNFTILKTYYDNYIMFHLINEKDGKTFQMLMELYGRKA
DLNLDKEFVKLCCEHHGIIKIEIDLTITKMRKCLRE

>gi|732003|sp|P39281|BLC_ECOLI Outer membrane lipoprotein blc precursor
MRLLPLVAAATAAFVLVAVCSSPTPRGVTVVNNFDKARYLGTYEXITARFDHFRFEGLEKVTATYSLRDG
GLNVINKGNYPDGKNQSEQKAYTFGAPTRAALKVSGGPYYGYNVIALREYRHALVCGBPDRLYLI
LSRTPTISDEVKQEMLAVATREGFDVSKFIWQQPS
```

(Please, no more than 200 sequences on our server.)

OR

Upload FASTA file:

Email address: swheelan@jhmi.edu (required)

Confirm email address: swheelan@jhmi.edu (required)

Email subject line: MUSCLE results
http://www.ebi.ac.uk/muscle/

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**

<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:

![Sequence Entry Field]
SeaView is a graphical multiple sequence alignment editor available at http://pbil.univ-lyon1.fr/software/seaview.html
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
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 most accurate programs available
- Available as standalone or web interface
- Many output formats, including interactive phylogenetic trees
Iterative approaches: MAFFT

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

Has about 1000 advanced settings!
### Iterative approaches: MAFFT

#### MAFFT-\_L-INS\_j Result

**CLUSTAL format alignment by MAFFT (v6.808a)**

<table>
<thead>
<tr>
<th></th>
<th>human</th>
<th>Dog</th>
<th>Mouse</th>
<th>RAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Dog</td>
<td>MQVLAAGRR----LPSVFHPGRYESTPSRPG--YKAGGPAAPADRLPWLYARPAPPRGL</td>
<td>RAAPVGGLPMTWEWVVALVLAALGSARAESDCRVSNFQVKKNFDKRFAVTWYAMAKKD</td>
<td>TRLGLRLRCEWVVALVLAALGGSERCRVSSFVRKENFDKARFSVGLWYAIAKKD</td>
<td>MPAFPPPSRPPPAPPAPPRGLSRVVTKARPPPGAIKVGGKPLASVASR-----ARERGGQAC</td>
</tr>
<tr>
<td>Mouse</td>
<td>------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>RAT</td>
<td>------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>--------------</td>
</tr>
</tbody>
</table>

**Alignment details:**

<table>
<thead>
<tr>
<th>human</th>
<th>Dog</th>
<th>Mouse</th>
<th>RAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>----------MKWWAVLPNLLAALGSGERDRCRVSSFVRKENFDKARFSVGTYAMAKKD</td>
<td>RAAPVGGLPMTWEWVVALVLAALGSARAESDCRVSNFQVKKNFDKRFAVTWYAMAKKD</td>
<td>TRLGLRLRCEWVVALVLAALGGSERCRVSSFVRKENFDKARFSVGLWYAIAKKD</td>
</tr>
<tr>
<td>Dog</td>
<td>PEGLFLQDNIVAEFSVDETGQMSATAKGRVRLNNWDVCADMVGTFTDTEPPKFKMKYW</td>
<td>PEGLFLQDNIVAEFSVDETGQMSATAKGRVRLNNWDVCADMVGTFTDTEPPKFKMKYW</td>
<td>PEGLFLQDNIVAEFSVDETGQMSATAKGRVRLNNWDVCADMVGTFTDTEPPKFKMKYW</td>
</tr>
<tr>
<td>Mouse</td>
<td>PEGLFLQDNIVAEFSVDETGQMSATAKGRVRLNNWDVCADMVGTFTDTEPPKFKMKYW</td>
<td>PEGLFLQDNIVAEFSVDETGQMSATAKGRVRLNNWDVCADMVGTFTDTEPPKFKMKYW</td>
<td>PEGLFLQDNIVAEFSVDETGQMSATAKGRVRLNNWDVCADMVGTFTDTEPPKFKMKYW</td>
</tr>
<tr>
<td>RAT</td>
<td>PEGLFLQDNIVAEFSVDETGQMSATAKGRVRLNNWDVCADMVGTFTDTEPPKFKMKYW</td>
<td>PEGLFLQDNIVAEFSVDETGQMSATAKGRVRLNNWDVCADMVGTFTDTEPPKFKMKYW</td>
<td>PEGLFLQDNIVAEFSVDETGQMSATAKGRVRLNNWDVCADMVGTFTDTEPPKFKMKYW</td>
</tr>
</tbody>
</table>

**Alignment details:**

<table>
<thead>
<tr>
<th>human</th>
<th>Dog</th>
<th>Mouse</th>
<th>RAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>GVASFLQKGNNDDHWIVTDYDTTYAVQYSCRLNLNGTDGCADSYSFVFSRDPNGLFPEAQKIK</td>
<td>GVASFLQKGNNDDHWIVTDYDTTYAVQYSCRLNLNGTDGCADSYSFVFSRDPNGLFPEAQKIK</td>
<td>GVASFLQKGNNDDHWIVTDYDTTYAVQYSCRLNLNGTDGCADSYSFVFSRDPNGLFPEAQKIK</td>
</tr>
<tr>
<td>Dog</td>
<td>GVASFLQKGNNDDHWIVTDYDTTYAVQYSCRLNLNGTDGCADSYSFVFSRDPNGLFPEAQKIK</td>
<td>GVASFLQKGNNDDHWIVTDYDTTYAVQYSCRLNLNGTDGCADSYSFVFSRDPNGLFPEAQKIK</td>
<td>GVASFLQKGNNDDHWIVTDYDTTYAVQYSCRLNLNGTDGCADSYSFVFSRDPNGLFPEAQKIK</td>
</tr>
<tr>
<td>Mouse</td>
<td>GVASFLQKGNNDDHWIVTDYDTTYAVQYSCRLNLNGTDGCADSYSFVFSRDPNGLFPEAQKIK</td>
<td>GVASFLQKGNNDDHWIVTDYDTTYAVQYSCRLNLNGTDGCADSYSFVFSRDPNGLFPEAQKIK</td>
<td>GVASFLQKGNNDDHWIVTDYDTTYAVQYSCRLNLNGTDGCADSYSFVFSRDPNGLFPEAQKIK</td>
</tr>
<tr>
<td>RAT</td>
<td>GVASFLQKGNNDDHWIVTDYDTTYAVQYSCRLNLNGTDGCADSYSFVFSRDPNGLFPEAQKIK</td>
<td>GVASFLQKGNNDDHWIVTDYDTTYAVQYSCRLNLNGTDGCADSYSFVFSRDPNGLFPEAQKIK</td>
<td>GVASFLQKGNNDDHWIVTDYDTTYAVQYSCRLNLNGTDGCADSYSFVFSRDPNGLFPEAQKIK</td>
</tr>
</tbody>
</table>

**Alignment details:**

<table>
<thead>
<tr>
<th>human</th>
<th>Dog</th>
<th>Mouse</th>
<th>RAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>VRQRQEELCLARQYLIVHNHYCDGRSERNL</td>
<td>VRQRQEELCLARQYLIVHNHYCDGRSERNL</td>
<td>VRQRQEELCLARQYLIVHNHYCDGRSERNL</td>
</tr>
<tr>
<td>Dog</td>
<td>VRQRQEELCLARQYLIVHNHYCDGRSERNL</td>
<td>VRQRQEELCLARQYLIVHNHYCDGRSERNL</td>
<td>VRQRQEELCLARQYLIVHNHYCDGRSERNL</td>
</tr>
<tr>
<td>Mouse</td>
<td>VRQRQEELCLARQYLIVHNHYCDGRSERNL</td>
<td>VRQRQEELCLARQYLIVHNHYCDGRSERNL</td>
<td>VRQRQEELCLARQYLIVHNHYCDGRSERNL</td>
</tr>
<tr>
<td>RAT</td>
<td>VRQRQEELCLARQYLIVHNHYCDGRSERNL</td>
<td>VRQRQEELCLARQYLIVHNHYCDGRSERNL</td>
<td>VRQRQEELCLARQYLIVHNHYCDGRSERNL</td>
</tr>
</tbody>
</table>
Iterative approaches: MAFFT

MAFFT-L-INS-i Result

CLUSTAL (-like) formatted alignment by MAFFT (v6.500a)

| gi  | 55743122 | ref  | MKWVWALLLAAALGS----GRAERD-----CRVSSFR-----VKENFDKARFSGTWWYAMAKK |
| gi  | 12843160 | dbj  | MEWKWALVLLAALGG-----GSAERD-----CRVSSFR-----VKENFDKARFSGLWYIAIKK |
| gi  | 4502163  | ref  | MVML----LLLALAGLFGAAEGQAFHLGK----PNPP-----VQENFDVNYKLGRWYEIEKI |
| gi  | 732003   | sp   | MRLL----PLV--------AAATAAFLVACSSPTPRGVTVVNNFDAKYRGTYWIERAF |
| gi  | 127528   | sp   | MKLL-----LCGLTTLVCIHAEEATS--------------KQGNLNVEKINGEWFSSILA |
| gi  | 55743122 | ref  | DPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLNNWDV--------------CADMVGTFDTE |
| gi  | 12843160 | dbj  | DPEGLFLQDNIIAEFSVDETGQMSATAKGRVRLNNWEV--------------CADMVGTFDTE |
| gi  | 4502163  | ref  | PTTFENG-RCIQANYSLMEMGKIKVLNQELRAD-GTVNQ--------IEGEATPVNLT |
| gi  | 732003   | sp   | DHRFERGKEVATYSLRDDDGLNIVNKPGDPRQMWQ--------SEGKAYFTG-- |
| gi  | 127528   | sp   | SDKRE---------KIEEHGSRMVFVEHIHVLENSLAFKFHTVIDECSEIFLVADKE |
| gi  | 55743122 | ref  | DPAKFKMKYWGVASFLQKGNNDDHIVTDYD-TYAVQYSC--------------RLLNLGDGCADTSYSFV |
| gi  | 12843160 | dbj  | DPAKFKMKYWGVASFLQQRNNDHIIIDTDYD-TFALQYSC--------------RLQNLGDGCADTSYSFV |
| gi  | 4502163  | ref  | EPAKLEVFKFSWFMP--------APYWILATDYE-NYALVYSCCTICIQLFHVD--------FAWI |
| gi  | 732003   | sp   | APTRAALKVSFFGPFY--------GGYNVIALDREYRHALVCGP--------DRD----YLWI |
| gi  | 127528   | sp   | KAGEYSVMYDFNTFT--------ILKTDYD-NYIMFH--------LNEKDGKTQ-LMEL |
| gi  | 55743122 | ref  | FSRDPNLGPEEA-QKIVRQRQEELCLARQYRILVHNGYCDGRSERNLL |
| gi  | 12843160 | dbj  | FSRDPNLGSPET-RRLVRQRQEELCLERQYRUIEHWNCYSQRRSRNLS |
| gi  | 4502163  | ref  | LARNPN-LPPETVDSLKNILTSNNDVQMTVTDQVNCPLS-------- |
| gi  | 732003   | sp   | LSRTPT-ISDEVKQEMLAVATREGFDVSKFIWVQPG--------S-------- |
| gi  | 127528   | sp   | YGRKAD-LNSDIKEKVFKLCEEHGIKENIIDLTKNRCLKARE-------- |
Iterative approaches: MAFFT  JalView
Multiple sequence alignment: outline

[1] Introduction to MSA

   1) Exact
   2) Progressive (ClustalW)
   3) Iterative (MUSCLE, MAFFT)
   4) Consistency (ProbCons)
   5) Structure-based (Expresso, PRALINE)
      Conclusions from benchmarking

[3] Databases of MSAs (hidden Markov models)

[4] Multiple alignment of genomic regions

[5] MEGA to make a multiple sequence alignment
http://probcons.stanford.edu/

ProbCons—consistency-based approach

PROBCONS is an efficient protein multiple sequence alignment program, which has demonstrated a statistically significant improvement in accuracy compared to several leading alignment tools.

**Basic Parameters**

- E-mail address: swheelan@jhmi.edu
- E-mail address (again): swheelan@jhmi.edu
- Input sequence file: /Users/sarahwheelan/Desktop/lipocali

**Additional Options**

- Consistency reps: 2
- Iterative refinement reps: 100
- Pre-training reps: 0
- Output format: MFA, CLUSTALW

**Compute Alignment**

Run PROBCONS! Clear values

Comments to Chuong Do (huongdo@cs.stanford.edu)
ProbCons: consistency-based approach

Combines iterative and progressive approaches with a unique probabilistic model.

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

Progressive alignment hierarchically along guide tree.

Post-processing and iterative refinement (a little like MUSCLE).
ProbCons—consistency-based approach

Sequence x \( x_i \)
Sequence y \( y_j \)
Sequence z \( z_k \)

If \( x_i \) aligns with \( z_k \)
and \( z_k \) aligns with \( y_j \)
then \( x_i \) should align with \( y_j \)

ProbCons incorporates evidence from multiple sequences to guide the creation of a pairwise alignment.
ProbCons output for the same alignment: how consistency iteration helps.
Multiple sequence alignment: outline

[1] Introduction to MSA

   1) Exact
   2) Progressive (ClustalW)
   3) Iterative (MUSCLE, MAFFT)
   4) Consistency (ProbCons)
   5) Structure-based (Expresso, PRALINE)

Conclusions from benchmarking

[3] Databases of MSAs (hidden Markov models)

[4] Multiple alignment of genomic regions

[5] MEGA to make a multiple sequence alignment
EXPRESSO (3DCoffee)  http://tcoffee.org

<table>
<thead>
<tr>
<th>ALIGNMENT</th>
<th>TCOFFEE</th>
<th>EXPRESSO(3DCoffee)</th>
<th>MCOFFEE</th>
<th>RCOFFEE</th>
<th>COMBINE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Regular</td>
<td>Regular</td>
<td>Regular</td>
<td>Regular</td>
<td>Regular</td>
</tr>
<tr>
<td></td>
<td>Advanced</td>
<td>Advanced</td>
<td>Advanced</td>
<td>Advanced</td>
<td>Advanced</td>
</tr>
<tr>
<td></td>
<td>site</td>
<td>site</td>
<td>site</td>
<td>site</td>
<td>site</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EVALUATION</th>
<th>CORE</th>
<th>iRMSD-AFDB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Regular</td>
<td>Regular</td>
</tr>
<tr>
<td></td>
<td>Advanced</td>
<td>Advanced</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PROCESSING</th>
<th>PROTOGENE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Regular</td>
<td>Advanced</td>
</tr>
</tbody>
</table>

Make an MSA
MSA using structural data
Compare MSA methods
Make an RNA MSA
Combine MSA methods
Consistency-based
Structure-based
Back translate protein MSA

http://tcoffee.org
PRALINE Input: pure iterative approach

Centre for Integrative Bioinformatics VU
vrije Universiteit amsterdam

PRALINE multiple sequence alignment

Paste in your sequences in FASTA format (MAX 500 sequences, length 2000):

Or Upload a FASTA file (MAX 500 sequences, length 2000):

Enter a name for your job

Options

Exchange weights matrix:

Associated gap penalties:

Progressive alignment strategy:

Structural features:
**Praline output: pure iterative approach**

<table>
<thead>
<tr>
<th>Human</th>
<th>Dog</th>
<th>Rat</th>
<th>Mouse</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>KENFDKARFS</td>
<td>GTWYAMAKKD</td>
<td>PEGLFLQDNIVAESFVDE</td>
<td>QMSATAKGRV</td>
<td>7 * * * * * * 5 * * 7 * * * * * * 9 * * * * * * 4 * 5 * * * * * *</td>
</tr>
</tbody>
</table>
| RLLNNWVDVCA | DMVGFTETDTE | DPAFKKMKYW | VASFLQ | 7 * * * * | 80%
| RLLNNWVDVCA | DMVGFTETDTE | DPAFKKMKYW | VASFLQ | 7 * * * * | 80%
| RLLSNWVEVCA | DMVGFTETDTE | DPAFKKMKYW | VASFLQ | 7 * * * * | 80%
| RLLSNWVEVCA | DMVGFTETDTE | DPAFKKMKYW | VASFLQ | 7 * * * * | 80%
| RLYAVQYSCR  | LLNLDGTCAD | SYSFVSRSRP | NGLPEAQKI | 4 * * * * * * * * 45 * 67 7 8 * * * * * * |
| RLYAVQYSCR  | LLNLDGTCAD | SYSFVSRSRP | NGLPEAQKI | 4 * * * * * * * * 45 * 67 7 8 * * * * * * |
| RFLAQYSCR   | LLNLDGTCAD | SYSFVSRSRP | NGLPEAQKI | 4 * * * * * * * * 45 * 67 7 8 * * * * * * |
| RFLAQYSCR   | LLNLDGTCAD | SYSFVSRSRP | NGLPEAQKI | 4 * * * * * * * * 45 * 67 7 8 * * * * * * |
| ARQYRLIVHN  | GYCDBGRSERNLL | ERQYRELHN | GYCQSPRPSRN | 5 * * * 4 4 * ** 6 5 6 6 5 * |
Praline output: pure iterative approach

Boxes highlight a region that is difficult to align.
<table>
<thead>
<tr>
<th>Database</th>
<th>Accession</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAFFT</td>
<td>gi</td>
<td>55743122</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>gi</td>
<td>12843160</td>
</tr>
<tr>
<td>MAFFT</td>
<td>gi</td>
<td>4502163</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>gi</td>
<td>127528</td>
</tr>
<tr>
<td>ProbCons</td>
<td>gi</td>
<td>55743122</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>gi</td>
<td>12843160</td>
</tr>
<tr>
<td>ProbCons</td>
<td>gi</td>
<td>4502163</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>gi</td>
<td>127528</td>
</tr>
<tr>
<td>ProbCons</td>
<td>gi</td>
<td>732003</td>
</tr>
</tbody>
</table>

**Alignment Output**

<table>
<thead>
<tr>
<th>Database</th>
<th>Accession</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAFFT</td>
<td>gi</td>
<td>55743122</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>gi</td>
<td>12843160</td>
</tr>
<tr>
<td>MAFFT</td>
<td>gi</td>
<td>4502163</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>gi</td>
<td>127528</td>
</tr>
<tr>
<td>ProbCons</td>
<td>gi</td>
<td>55743122</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>gi</td>
<td>12843160</td>
</tr>
<tr>
<td>ProbCons</td>
<td>gi</td>
<td>4502163</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>gi</td>
<td>127528</td>
</tr>
<tr>
<td>ProbCons</td>
<td>gi</td>
<td>732003</td>
</tr>
</tbody>
</table>

**Consensus Output**

<table>
<thead>
<tr>
<th>Database</th>
<th>Accession</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAFFT</td>
<td>gi</td>
<td>55743122</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>gi</td>
<td>12843160</td>
</tr>
<tr>
<td>MAFFT</td>
<td>gi</td>
<td>4502163</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>gi</td>
<td>127528</td>
</tr>
</tbody>
</table>

**Consensus Output**

<table>
<thead>
<tr>
<th>Database</th>
<th>Accession</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAFFT</td>
<td>gi</td>
<td>55743122</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>gi</td>
<td>12843160</td>
</tr>
<tr>
<td>MAFFT</td>
<td>gi</td>
<td>4502163</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>gi</td>
<td>127528</td>
</tr>
</tbody>
</table>
Multiple sequence alignment: outline

[1] Introduction to MSA

   1) Exact
   2) Progressive (ClustalW)
   3) Iterative (MUSCLE, MAFFT)
   4) Consistency (ProbCons)
   5) Structure-based (Expresso, PRALINE)

Conclusions from benchmarking

[3] Databases of MSAs (hidden Markov models)

[4] Multiple alignment of genomic regions

[5] MEGA to make a multiple sequence alignment
Multiple sequence alignment: Conclusions from benchmarking

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, everyone’s old favorite, continues to be a decent program and is included in almost every MSA paper you will see. It has withstood the test of time. Plus, it has a nice interface (especially with CLUSTALX) and is easy to use. But it might be time to move on.
### Multiple sequence alignment algorithms

<table>
<thead>
<tr>
<th>Progressive</th>
<th>Local</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PIMA</td>
<td>CLUSTAL PileUp</td>
</tr>
<tr>
<td>Iterative</td>
<td>DIALIGN Mafft</td>
<td>SAGA Muscle</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>1fmb</th>
<th>1fmb</th>
</tr>
</thead>
<tbody>
<tr>
<td>vTNLEKRPTTIVLLNDPLNLLDDTGGDTSVLTTahynrlkrygrk.VQ</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7upjB</th>
<th>7upjB</th>
</tr>
</thead>
<tbody>
<tr>
<td>pQFLSFLKPVRVTAYIEGQPVEVLLDTGADDTSIVAG..iel.gnn.YS</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pol_sivcz</th>
<th>pol_sivcz</th>
</tr>
</thead>
<tbody>
<tr>
<td>pQITLWQRPLIPVKVEGQLCEALDLGADDTVIER..ielqgl..UK</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>POL_SIVMK</th>
<th>POL_SIVMK</th>
</tr>
</thead>
<tbody>
<tr>
<td>pQFLSFLWRRPVVTAHIEGQPVEVLLDTGADDSTIVG..iel.gph.YT</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1fmb</th>
<th>1fmb</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTGIGGIVGVGRVES.SPVTIHHKGRHDKMLGLVADIPVTLRGLDILQDL</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7upjB</th>
<th>7upjB</th>
</tr>
</thead>
<tbody>
<tr>
<td>PKIVGGGIGFINTLEFGKVEIEVLTNKNKVRATINTGTDTPINIFGRNILTAL</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pol_sivcz</th>
<th>pol_sivcz</th>
</tr>
</thead>
<tbody>
<tr>
<td>PKMIGGGIGFIVKQFVNQHIEIGKVRVGTGVLGPTVPNIFGRNILTQL</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>POL_SIVMK</th>
<th>POL_SIVMK</th>
</tr>
</thead>
<tbody>
<tr>
<td>PKIIVGGGIGFINTKEKVEIEVLTNRKIKRTINTGTDTPINIFGRNLLTAL</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1fmb</th>
<th>1fmb</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAKLV1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7upjB</th>
<th>7upjB</th>
</tr>
</thead>
<tbody>
<tr>
<td>GHSLN1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pol_sivcz</th>
<th>pol_sivcz</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCTLV.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>POL_SIVMK</th>
<th>POL_SIVMK</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMSLN1</td>
<td></td>
</tr>
</tbody>
</table>

Key

- **alpha helix**: RED
- **beta strand**: GREEN
- **core blocks**: UNDERSCORE
Conclusions: assessment of alternative multiple sequence alignment algorithms

[1] As percent identity among proteins drops, performance (accuracy) declines also. This is especially severe for proteins < 25% identity.

Proteins <25% identity: 65% of residues align well

Proteins <40% identity: 80% of residues align well

[2] “Orphan” sequences are highly divergent members of a family. Surprisingly, orphans do not disrupt alignments. Also surprisingly, global alignment algorithms outperform local.
Conclusions: assessment of alternative multiple sequence alignment algorithms

[3] Separate multiple sequence alignments can be combined (e.g. RBPs and lactoglobulins).

Iterative algorithms (MUSCLE, MAFFT, PRRP, SAGA) outperform progressive alignments (Clustal)

[4] When proteins have large N-terminal or C-terminal extensions, local alignment algorithms are superior. PileUp (global) is an exception.
A new major shakeup in the alignment world . . .


VERY simple experiment: take a bunch of sequences, align them using popular programs, then reverse them, and align them again, e.g.

...ADDSYP -> PYSDDA...
...ADYSYP -> PYSYDA...

The alignments should be the same, right????

NO! the agreement is pathetic, from 8-50% using the most generous measures.
A new major shakeup in the alignment world . . .

What does this mean? Why did this happen?

Most likely, our gap insertion algorithms are biased toward left-to-right reading order. Clearly, this is not good, because evolution probably does not have a similar bias. This is currently referred to as the HoT (Heads or Tails) problem.

How to fix it? In the only two MSA papers released since this bombshell, the authors mentioned the issue and noted that every study should also reverse all of their alignments, although they themselves did not do so.
Two kinds of multiple sequence alignment resources

[1] Databases of multiple sequence alignments

Text-based searches:
CDD, Pfam (profile HMMs), PROSITE

Database searches with a query sequence:
BLAST, CDD, PFAM

[2] Multiple sequence alignment by custom input

Muscle, Clustal W, Clustal X
Multiple sequence alignment programs

AMAS
CINEMA
ClustalW
ClustalX
DIALIGN
HMMT
Match-Box
MultAlin
MSA
Muslc
PileUp
SAGA
T-COFFEE
Databases of multiple sequence alignments

- BLOCKS
- CDD
- DOMO (Gapped MSA)
- INTERPRO
- iProClass
- MetaFAM
- Pfam
- PRINTS
- PRODOM (PSI-BLAST)
- PROSITE
- SMART

HMM
Databases of multiple sequence alignments

- BLOCKS
- CDD
- DOMO (Gapped MSA)
- INTERPRO
- iProClass
- MetaFAM
- Pfam
- PRINTS
- PRODOM (PSI-BLAST)
- PROSITE
- SMART

Integrative resources
PFAM (protein family) database:
http://pfam.sanger.ac.uk/

Pfam 22.0 (July 2007, 9318 families)

The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). More...

USING PFAM

SEQUENCE SEARCH
Analyze your protein sequence for Pfam matches

VIEW A PFAM FAMILY
View Pfam family annotation and alignments

VIEW A CLAN
See groups of related families

VIEW A SEQUENCE
Look at the domain organisation of a UniProt sequence

VIEW A STRUCTURE
Find the domains on a PDB structure

KEYWORD SEARCH
Query Pfam by keywords

Or view the help pages for more information
PFAM entry: lipocalin

Accession number: PF00061
Definition: Lipocalin / cytosolic fatty-acid binding protein family
Author: Eddy SR
Alignment method of seed: ClustalW
Source of seed members: Prosite and HMM_iterative_training
Gathering cutoffs: 9.9 9.9
Trusted cutoffs: 9.90 9.90
Noise cutoffs: 9.80 9.80
HMM build command line: hmmbuild -f HMM SEED
HMM build command line: hmmscan --seed 0 HMM
Database Reference: PROSITE: PD0C00187
Database Reference: PROSITE: PD0C00188
Database Reference: PRINTS: PR00178;
Database Reference: PRINTS: PR00179;
Database Reference: SCOP: 1hms; sf: [SCOP-USA] [CATH-PDBSUM]
Database reference: PFAMB: PB020384;
Database reference: PFAMB: PB023754;
Database Reference: INTERPRO: IP0000566;
Database Reference: PDB: 1pmp A: 3; 131;
Database Reference: PDB: 1pmp B: 3; 131;
Database Reference: PDB: 1pmp C: 3; 131;
Database Reference: PDB: 2a2u C: 14; 157;
Database Reference: PDB: 2a2u D: 14; 157;
Database Reference: PDB: 1ew3 A: 32; 176;
Comment: Lipocalins are transporters for small hydrophobic molecules, such as lipids, steroid hormones, bilins, and retinoids.
Comment: Alignment subsumes both the lipocalin and fatty acid binding protein signatures from PROSITE. This is supported on structural and functional grounds.
Comment: Structure is an eight-stranded beta barrel.
Number of members: 274

Retrieve a Pfam alignment for lipocalin

Which alignment: [Seed alignment] [What format: Plain text]

Use a mime-capable external viewer like behv
Retrieve alignment

Visualize domain structure of proteins in a Pfam lipocalin alignment

Which alignment: [Seed alignment] [Retrieve domain structures]

Retrieve Pfam profile HMM for lipocalin

Retrieve HMM
PFAM HMM for lipocalins

<table>
<thead>
<tr>
<th>Position</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>C</td>
</tr>
<tr>
<td>3</td>
<td>D</td>
</tr>
<tr>
<td>4</td>
<td>E</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
</tr>
<tr>
<td>6</td>
<td>G</td>
</tr>
<tr>
<td>7</td>
<td>H</td>
</tr>
<tr>
<td>8</td>
<td>I</td>
</tr>
<tr>
<td>9</td>
<td>K</td>
</tr>
<tr>
<td>10</td>
<td>L</td>
</tr>
<tr>
<td>11</td>
<td>M</td>
</tr>
<tr>
<td>12</td>
<td>N</td>
</tr>
<tr>
<td>13</td>
<td>P</td>
</tr>
<tr>
<td>14</td>
<td>Q</td>
</tr>
<tr>
<td>15</td>
<td>R</td>
</tr>
<tr>
<td>16</td>
<td>S</td>
</tr>
<tr>
<td>17</td>
<td>T</td>
</tr>
<tr>
<td>18</td>
<td>V</td>
</tr>
<tr>
<td>19</td>
<td>W</td>
</tr>
<tr>
<td>20</td>
<td>Y</td>
</tr>
</tbody>
</table>

20 amino acids
PFAM HMM for lipocalins: GXW motif

| HMM | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | U | Y |
| m->m | -30 | +5585 | 985 | 1876 | -4618 | -3799 | -2458 | -1465 | 1726 | -849 | -3387 | -1158 | -3892 | 2 | 745 | 249 | -2764 | -1469 | -4481 | -3798 |
| m->i | -149 | -500 | 233 | 43 | -381 | 399 | 106 | -626 | 210 | -466 | -720 | 275 | 394 | 45 | 96 | 359 | 117 | -369 | -294 | -249 |
| i->i | -149 | -500 | 233 | 43 | -381 | 399 | 106 | -626 | 210 | -466 | -720 | 275 | 394 | 45 | 96 | 359 | 117 | -369 | -294 | -249 | 20 amino acids |

HMM: Hidden Markov Model
PFAM GCG MSF format
Pfam (protein family) database

If you use HMM-Logos in your publication, please cite
Distinct domain architectures for Globin

1746 proteins with Globin architecture
Q27126  URECA[ urechis caupo (innkeeper worm) (spoonworm)] f-i hemoglobin

183 proteins with Globin, FAD_binding_6, NAD_binding_1 architecture
Q6XX21  CRYNV[ cryptococcus neoformans var. grubii (filobasidiella neoformans var. grubii)] flavohemoglobin

31 proteins with Globin, Globin architecture
Q683R3  BIOGL[ biophthora glabrata (bloodfluke planorb)] hemoglobin (fragment)

4 proteins with Globin, Globin, Globin, Globin, Globin, Globin, Globin, Globin, Globin, Globin architecture
Q9NG75  9CRUS[ paratema zietziana] hemoglobin p polymer precursor
SMART: Simple Modular Architecture Research Tool (emphasis on cell signaling)
Domains within the query sequence of 199 residues

Mouse over domain / undefined region to see the limits; click on it to go to further annotation; right-click to save whole protein as PNG image

Transmembrane segments as predicted by the TMBHMM program ( ), coiled coil regions determined by the Coils2 program ( ) and Segments of low compositional complexity, determined by the SPCE program ( ) Hits only found by BLAST are indicated by (highlight) for hits in the schnipsel database and (red) for hits against PDB.

Architecture analysis

Display all proteins with similar domain organisation.
Display all proteins with similar domain composition.

The SMART diagram above represents a summary of the results shown below. Domains with scores less significant than established cutoffs are not shown in the diagram. Features are also not shown when two or more occupy the same piece of sequence; the priority for display is given by SMART > PFAM > PROSPERO repeats > Signal peptide > Transmembrane > Coiled coil > Low complexity. In either case, features not shown in the above diagram are marked 'hidden'

Confidently predicted domains, repeats, motifs and features:

<table>
<thead>
<tr>
<th>name</th>
<th>begin</th>
<th>end</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>low complexity</td>
<td>5</td>
<td>16</td>
<td>-</td>
</tr>
</tbody>
</table>

Summary of BLAST results. Note that the probabilities are not directly comparable to those in the table above.

<table>
<thead>
<tr>
<th>name</th>
<th>seq</th>
<th>begin</th>
<th>end</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB:1RBP</td>
<td>lrbp</td>
<td>17</td>
<td>198</td>
<td>1.00e-108</td>
</tr>
</tbody>
</table>
Databases of multiple sequence alignments

- BLOCKS
- CDD
- DOMO
- INTERPRO
- iProClass
- MetaFAM
- PFAM
- PRINTS
- PRODOM
- PROSITE
- SMART

Conserved Domain Database (CDD) at NCBI = PFAM + SMART
CDD: Conserved domain database

[1] Go to NCBI → Structure
[2] Click CDD
[3] Enter a text query, or a protein sequence
CDD: Conserved domain database

Proteins often contain several modules or domains, each with a distinct evolutionary origin and function. NCBI’s Conserved Domain Database is a collection of multiple sequence alignments for ancient domains and full-length proteins. The CD-Search service may be used to identify the conserved domains present in a protein query sequence:

Submit Query

Search Database | CDD v2.05 - 11399 PSSMs

Enter a **Protein** query as Accession, GI, or Sequence in FASTA format:

lipocalin

Read about the FASTA format description. Click [here](#) for advanced options.
CDD = PFAM + SMART
CDD uses RPS-BLAST: reverse position-specific

Purpose: to find conserved domains in the query sequence

Query = your favorite protein

Database = set of many position-specific scoring matrices (PSSMs), i.e. a set of MSAs

CDD is related to PSI-BLAST, but distinct

CDD searches against profiles generated from pre-selected alignments
MSA databases: manual vs. automated curation

Manual curation:
- Pfam
- PROSITE
- BLOCKS
- PRINTS

Automated curation:
- DOMO
- PRODOM
- MetaFam

Advantage:
- fewer alignment errors
- more comprehensive
Multiple sequence alignment: outline

[1] Introduction to MSA

   1) Exact
   2) Progressive (ClustalW)
   3) Iterative (MUSCLE, MAFFT)
   4) Consistency (ProbCons)
   5) Structure-based (Expresso, PRALINE)
   Conclusions from benchmarking

[3] Databases of MSAs (hidden Markov models)

[4] Multiple alignment of genomic regions

[5] MEGA to make a multiple sequence alignment
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.