Multiple sequence alignment

Monday, December 6, 2010

Bioinformatics
J. Pevsner
pevsner@kennedykrieger.org

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
- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

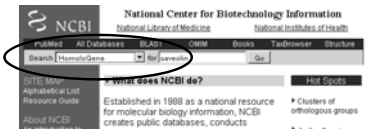
[3] MEGA to make a multiple sequence alignment

[4] Multiple alignment of genomic DNA

Multiple sequence alignment: definition

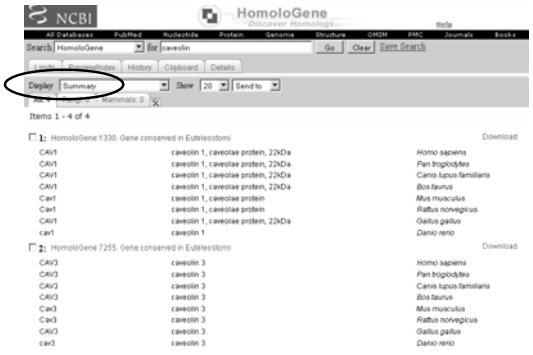
- a collection of three or more protein (or nucleic acid) sequences that are partially or completely aligned
- homologous residues are aligned in columns across the length of the sequences
- residues are homologous in an evolutionary sense
- residues are homologous in a structural sense

Example: someone is interested in caveolin



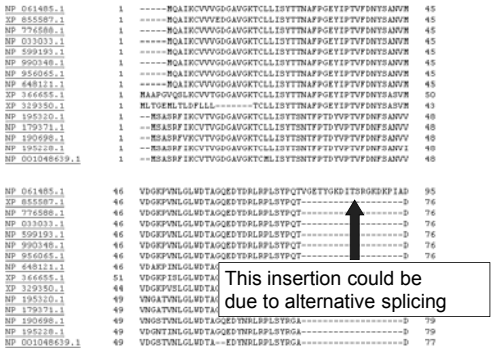
Step 1: at NCBI change the pulldown menu to HomoloGene and enter caveolin in the search box

Step 2: inspect the results. We'll take the first set of caveolins. Change the Display to Multiple alignment.

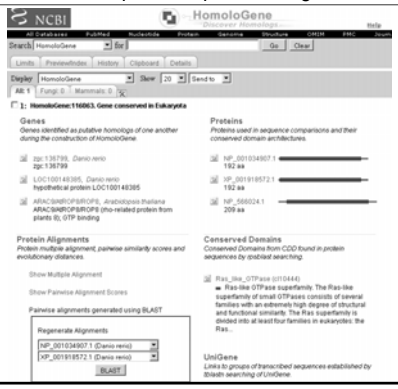


Step 3: inspect the multiple alignment. Note that these eight proteins align nicely, although gaps must be included.

Here's another multiple alignment, Rac:



HomoloGene includes groups of eukaryotic proteins. The site includes links to the proteins, pairwise alignments, and more



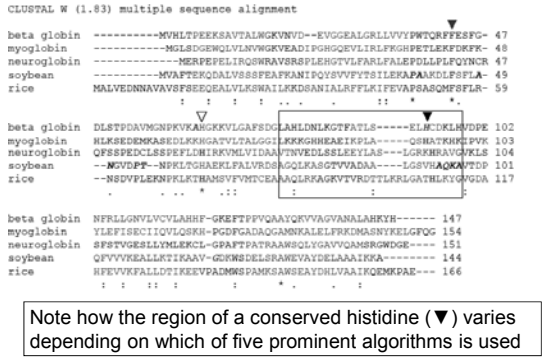
Example: 5 alignments of 5 globins

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

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

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

ClustalW



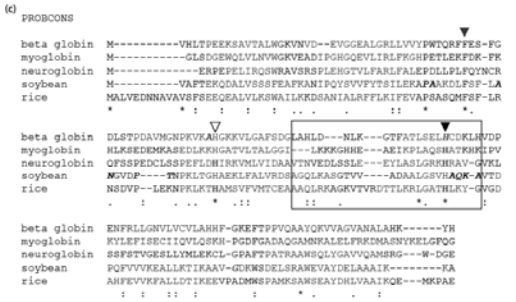
Note how the region of a conserved histidine (▼) varies depending on which of five prominent algorithms is used

Praline

MUSCLE

Page 194

Probcons



Page 195

TCoffee

Page 195

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

Page 180

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)

Page 181

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)
- Regulatory regions of genes may have consensus sequences identifiable by MSA

Page 181

Multiple sequence alignment: outline

[1] Introduction to MSA
- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

[3] MEGA to make a multiple sequence alignment

[4] Multiple alignment of genomic DNA

[5] Introduction to molecular evolution and phylogeny

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

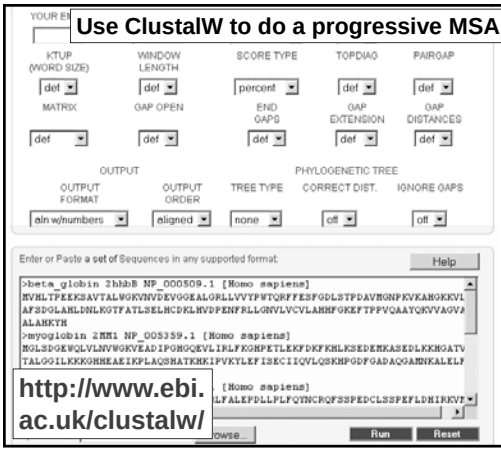
Example of MSA using ClustalW: two data sets

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



http://www.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

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

Page 185

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

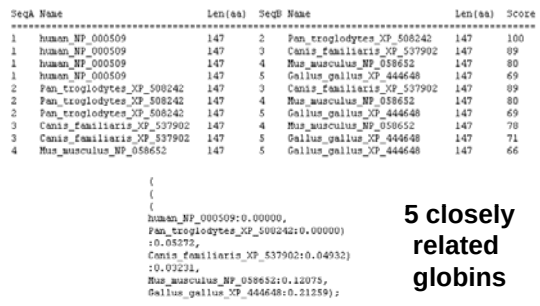
```
(
beta_globin:0.36022,
myoglobin:0.38808,
(
neuroglobin:0.39924,
(
soybean:0.30760,
rice:0.26184)
:0.13652)
:0.06560);
```

Page 186

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 lecture)
- ClustalW provides a syntax to describe the tree

Page 187



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

Page 188

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.”

Page 188

Clustal W alignment of 5 distantly related globins

```
CLUSTAL W (1.83) multiple sequence alignment


beta_globin         ---------------MVHLTPEEKSAVTALWG--KVNVDEVGGEALGRLLVVYPWTQRFF 43
cytoglobin          MEKVPGEMEIERRERSEELSEAERKAVQAMWARLYANCEDVGVAILVRFFVNFPSAKQYF 60
myoglobin           ----------------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKF 44
neuroglobin         ----------------MERPEPELLRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLF 42
leghemoglobin       ----------------MGFTEKQEALVNSSWELFKQNPS-YSVLFYTILEKAPAAKGMF 43
                                     :  :   :   *    .   .      ::    *     *

beta_globin         ES-PGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLA---HLDNLKGTFATLSELHCDKLHV 99
cytoglobin          SQ-FKHMEDPLEMERSPQLRKHACRVMGALNTVVENLHDPDKVSSVLALVGKAHALKHKV 119
myoglobin           DK-FKHLKSEDEMKASEDLKKHGATVLTALGGILK---KKGHHEAEIKPLAQSHATKHKI 100
neuroglobin         QYNCRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEVLASLGRGHDAVC V 101
leghemoglobin       S----FLKDSAEVVDSPKLQAHAEKVFGMVHDSAIQLRASGEVVLGDATLGAIHIQKGVV 99
                    .    :.      . .. *  *: .       .       :. *    :

beta_globin         DPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH---------- 147
cytoglobin          EPVYFKILSGVILEVVAEEFASDFPPETQRAWAKLRGLIYSHVTAAYKEVGWVQQVPNAT 179
myoglobin           PVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG------ 154
neuroglobin         KLSSFSTVGESLLYMLEKCLGPAFTPATRAAWSQLYGAVVQAMSRGWDGE---------- 151
leghemoglobin       DP-HFVVVKEALLETIKEASGEKWSEELSTAWEVAYEGLASAIKKAMN----------- 146
                        :  :   :: :      :      *      . . :   .

beta_globin         ----------
cytoglobin          TPPATLPSSGP 190
myoglobin           ----------
neuroglobin         ----------
leghemoglobin       ----------
```

Fig. 6.3
Page 187

Clustal W alignment of 5 closely related globins

```
CLUSTAL W (1.83) multiple sequence alignment

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

* asterisks indicate identity in a column

Fig. 6.5
Page 189

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

Page 189

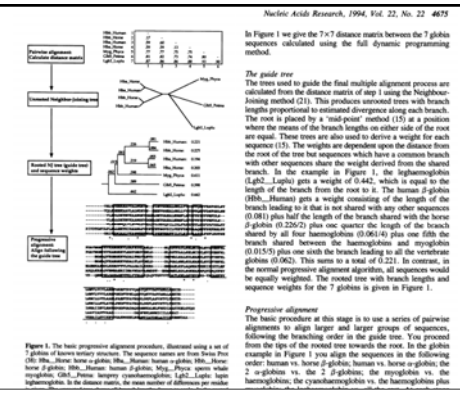
Additional features of ClustalW improve its ability to generate accurate MSAs

- 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

Page 190



Nucleic Acids Research, 1994, Vol. 22, No. 22 **4675**

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

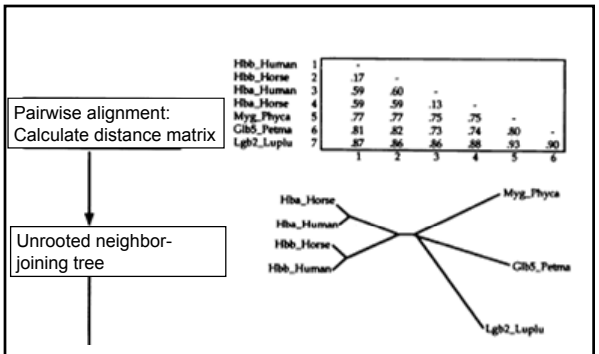
The trees used to guide the final multiple alignment process are calculated from the distance matrix of step 1 using the Neighbour-Joining method (21). This produces unrooted trees with branch lengths proportional to estimated divergence along each branch. The root is placed by a 'mid-point' method (15) 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.

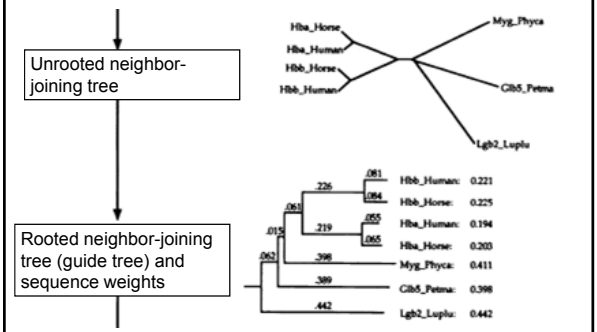
Progressive alignment

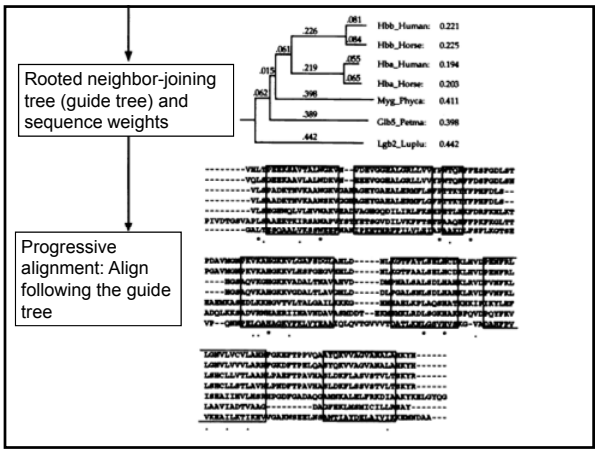
The basic procedure at this stage is to use a series of pairwise alignments to align larger and larger groups of sequences, following the branching order in the guide tree. You proceed from the tips of the rooted tree towards the root. In the globin example in Figure 1 you align the sequences in the 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

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 (36): 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

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







Multiple sequence alignment: outline

[1] Introduction to MSA
- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

[3] MEGA to make a multiple sequence alignment

[4] Multiple alignment of genomic DNA

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

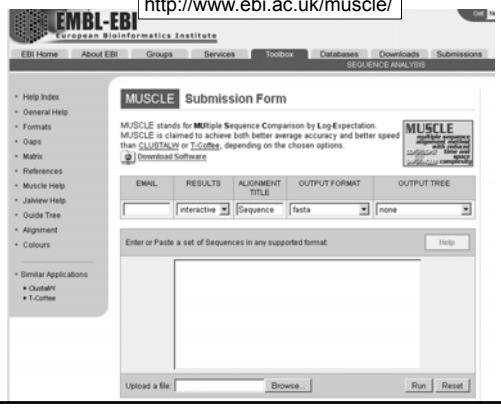
MUSCLE: next-generation progressive MSA

[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

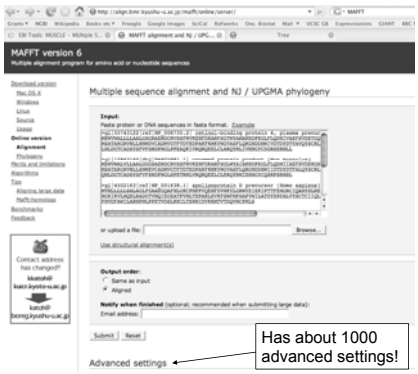
Access to MUSLCE at EBI
http://www.ebi.ac.uk/muscle/



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!

Multiple sequence alignment: outline

[1] Introduction to MSA
- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

[3] MEGA to make a multiple sequence alignment

[4] Multiple alignment of genomic DNA

Multiple sequence alignment: consistency

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

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

Examples: T-COFFEE, Prrp, DiAlign, ProbCons

ProbCons—consistency-based approach

Combines iterative and progressive approaches with a unique probabilistic model.

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

Progressive alignment hierarchically along guide tree

Post-processing and iterative refinement (a little like MUSCLE)

ProbCons uses an HMM to make alignments

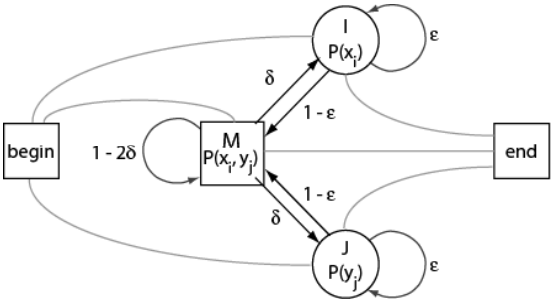


Fig. 5.12
Page 158

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.

Page 193

ProbCons output for the same alignment: consistency iteration helps

Multiple sequence alignment: outline

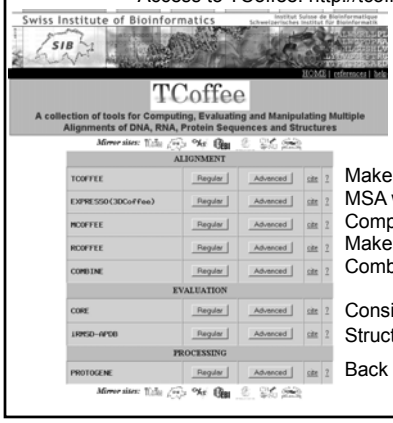
[1] Introduction to MSA
- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

[3] MEGA to make a multiple sequence alignment

[4] Multiple alignment of genomic DNA

Access to TCoffee: http://tcoffee.org



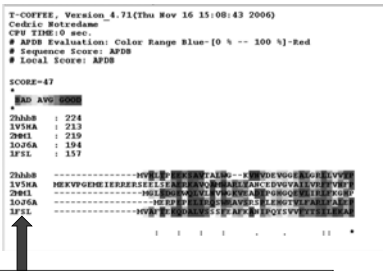
Make a MSA
MSA w. structural data
Compare MSA methods
Make an RNA MSA
Combine MSA methods

Consistency-based
Structure-based

Back translate protein MSA

Page 194

APDB ClustalW output:
TCoffee can incorporate structural information into a MSA



Protein Data Bank accession numbers

Multiple sequence alignment: outline

[1] Introduction to MSA
- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

[3] MEGA to make a multiple sequence alignment

[4] Multiple alignment of genomic DNA

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.

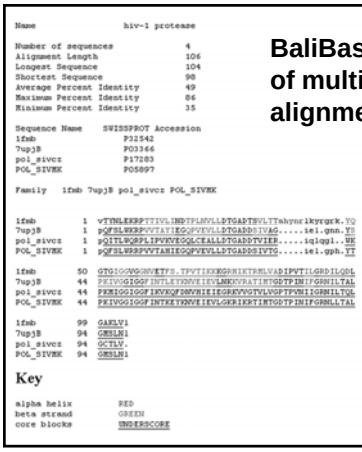
Strategy for assessment of alternative multiple sequence alignment algorithms

[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



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 is 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)

Multiple sequence alignment: review questions

[1] Explain how ClustalW works.

[2] Name two alternative methods. How do they work? What is the evidence that they are better than ClustalW?

[3] What does it mean to do a benchmarking study, and why is it important?

[4] Why is it important that various programs that make MSAs often give different answers?

Multiple sequence alignment: outline

[1] Introduction to MSA
- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

[3] MEGA to make a multiple sequence alignment

[4] Multiple alignment of genomic DNA

Multiple sequence alignment to profile HMMs

► Hidden Markov models (HMMs) are "states" that describe the probability of having a particular amino acid residue at arranged in a column of a multiple sequence alignment

► HMMs are probabilistic models

► HMMs may give more sensitive alignments than traditional techniques such as progressive alignment

Structure of a hidden Markov model (HMM)

delete state

insert state

main state

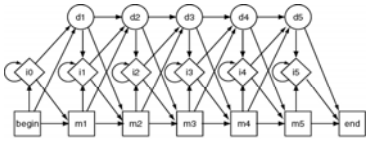
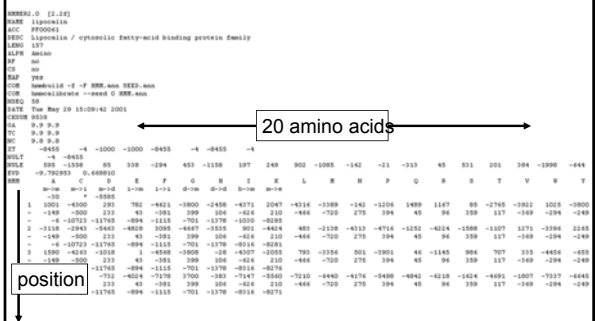


Fig. 5.12

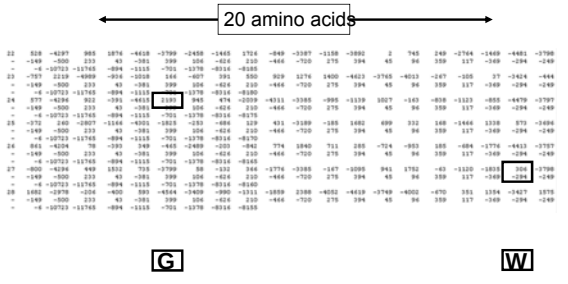
PFAM (protein family) database is a leading resource for the analysis of protein families

http://pfam.sanger.ac.uk/

PFAM HMM for lipocalins: resembles a position-specific scoring matrix



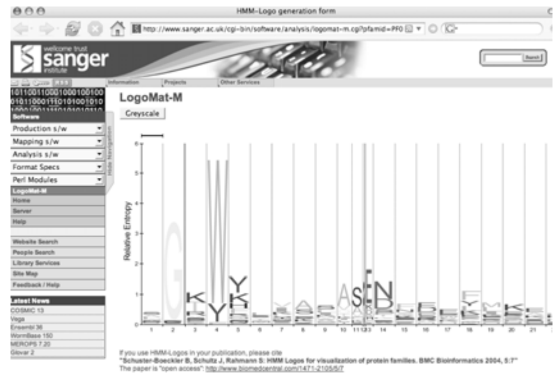
PFAM HMM for lipocalins: GXW motif



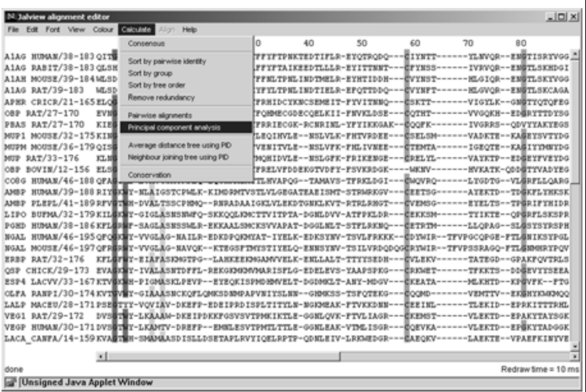
PFAM GCG MSF format

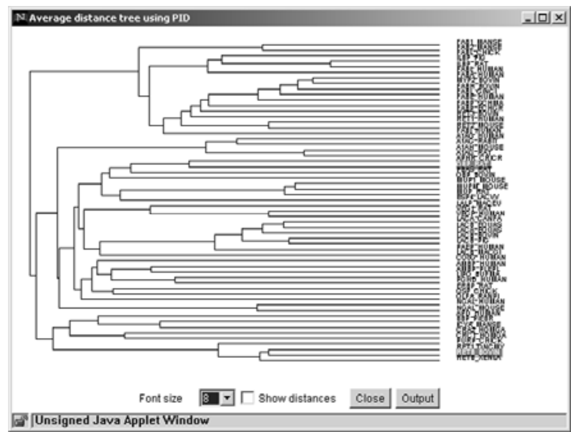
```
                    1                                                   50
A1AG_HUMAN_38-183   QITGKWF.YI ASAFRNEEYN .KSVQEIQAT FFYFTPNKTE DTIFLR.EYQ
A1AG_RABIT_38-183   QLSHKWF.FT ASAFRNPKYK .QLVQHTQAA FFYFTAIKEE DTLLLR.EYI
A1AH_MOUSE_39-184   WLSDKWF.FI GAAVLNPDYR .QEIQKTQMV FFNLTPNLIN DTMELR.EYH
A1AG_RAT_39-183     WLSDKWF.YM GAAFRDPVFK .QAVQTIQTE YFYLTPNLIN DTIELR.EFQ
APHR_CRICR_21-165   ELQGKWY.TI VIAADNLEKI .EEGGPLRFY FRHIDCYKNC SEMEIT.FYV
OBP_RAT_27-170      EVNGDWR.TL YIVADNVEKV .AEGGSLRAY FQHMECGDEC QELKII.FNV
PBAS_RAT_27-170     KIEGNWR.TV YLAASSVEKI .NEGSPLRTY FRRIECGK.R CNRINL.YFY
MUP1_MOUSE_32-175   KINGEWH.TI ILASDKREKI .EDNGNFRLF LEQIHVLE.. NSLVLK.FHT
MUPM_MOUSE_36-179   QISGYWF.SI AEASYEREKI .EEHGSMRAF VENITVLE.. NSLVFK.FHL
MUP_RAT_33-176      KLNGDWF.SI VVASNKREKI .EENGSMRVF MQHIDVLE.. NSLGFK.FRI
OBP_BOVIN_12-156    ELSGPWR.TV YIGSTNPEKI .QENGPFRTY FRELVFDDEK GTVDFY.FSV
CO8G_HUMAN_46-188   QFAGTWL.LV AVGSACRFLQ .EQGHRAEAT TLHVAPQG.. TAMAVS.TFR
AMBP_HUMAN_39-188   RIYGKWY.NL AIGSTCPWLK .KIMDRMTVS TLVLGEGATE AEISMT.STR
AMBP_PLEPL_41-189   RFVGTWH.DV ALTSSCPHMQ ..RNRADAAI GKLVLEKDTG NKLKVT.RTR
LIPO_BUFMA_32-179   KILGKWY.GI GLASNSNWFQ .SKKQQLKMC TTVITPTA.D GNLDVV.ATF
PGHD_HUMAN_38-186   KFLGRWF.SA GLASNSSWLR .EKKAALSMC KSVVAPAT.D GGLNLT.STF
NGAL_HUMAN_46-195   QFQGKWY.VV GLAG.NAILR .EDKDPQKMY AT.IYELK.E DKSYNV.TSV
NGAL_MOUSE_46-197   QFRGRWY.VV GLAG.NAVQK ..KTEGSFTM YSTIYELQ.E NNSYNV.TSI
ERBP_RAT_32-176     KFLGFWY.EI AFASKMGTPG ..LAHKEEKM GAMVVELK.E NLLALT.TTY
QSP_CHICK_29-173    EVAGKWY.IV ALASNTDFFL .REKGKMKMV MARISFLG.E DELEVS.YAA
ESP4_LACVV_33-167   KTVGKWH.PI GMASKLPEVP ..EYEQKISP MDHMVELT.D GDMKLT.ANY
OLFA_RANPI_30-174   KVTGVWY.GI AAASNCKQFL QMKSDNMPAP VNIYSLNN.. GHMKSS.TSF
LALP_MACEU_28-171   PSEGTYY.VQ VIAV.DKEFP .EDEIPRDIS PLTITYLN.N GKMEAK.FTV
VEG1_RAT_29-172     DVSGTWY.LK AAAW.DKEIP DKKFGSVSVT PMKIKTLE.G GNLQVK.FTV
```

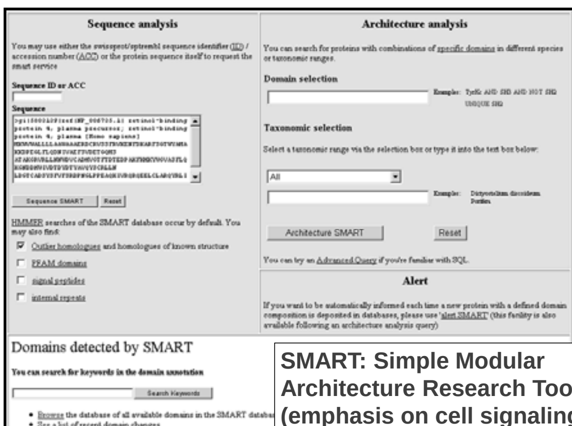
Pfam (protein family) database



PFAM JalView viewer





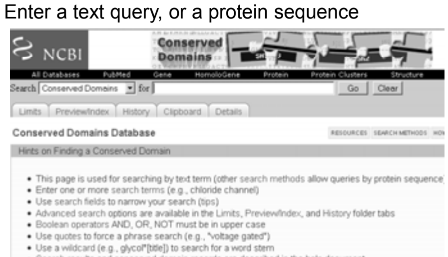


SMART: Simple Modular Architecture Research Tool (emphasis on cell signaling)

Page 199

CDD: Conserved domain database (at NCBI): CDD = Pfam + SMART

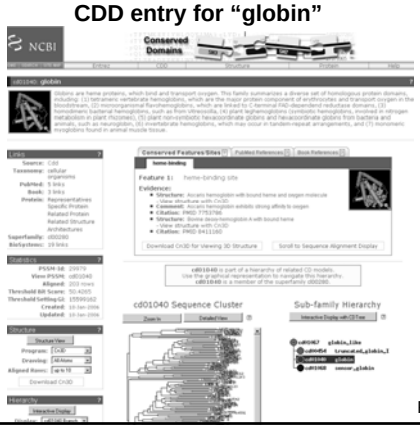
[1] Go to NCBI → Domains & Structure (left sidebar)
[2] Click CDD
[3] Enter a text query, or a protein sequence



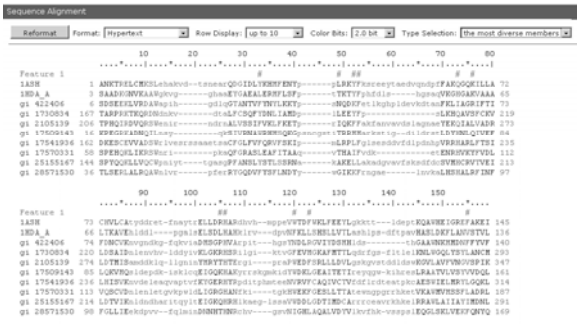
CDD entry for “globin”

CDD = PFAM + SMART

CDD entry for “globin”



CDD entry for “globin”



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

Multiple sequence alignment: outline

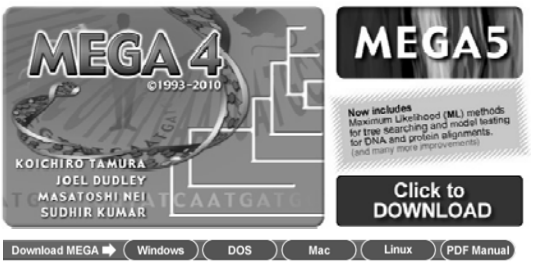
[1] Introduction to MSA
- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[2] Hidden Markov models (HMMs), Pfam and CDD

[3] MEGA to make a multiple sequence alignment

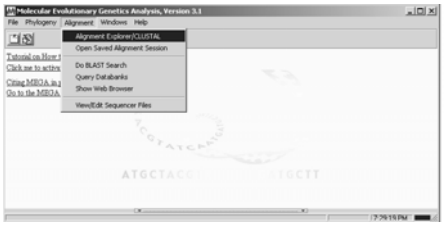
[4] Multiple alignment of genomic DNA

MEGA version 4: Molecular Evolutionary Genetics Analysis

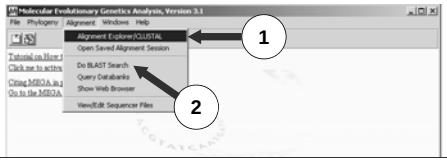


Download from www.megasoftware.net

MEGA version 4: Molecular Evolutionary Genetics Analysis

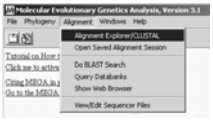


MEGA version 4: Molecular Evolutionary Genetics Analysis

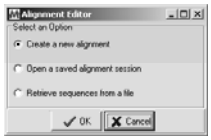


Two ways to create a multiple sequence alignment
1. Open the Alignment Explorer, paste in a FASTA MSA
2. Select a DNA query, do a BLAST search

Once your sequences are in MEGA, you can run ClustalW then make trees and do phylogenetic analyses



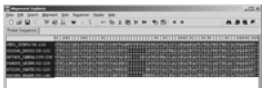
[1] Open the Alignment Explorer



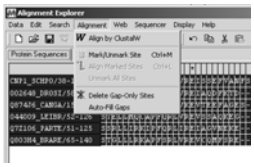
[2] Select "Create a new alignment"

[3] Click yes (for DNA) or no (for protein)

[4] Find, select, and copy a multiple sequence alignment (e.g. from Pfam; choose FASTA with dashes for gaps)



[5] Paste it into MEGA



[6] If needed, run ClustalW to align the sequences

[7] Save (Ctrl+S) as .mas then exit and save as .meg

Multiple sequence alignment: outline

[1] Introduction to MSA
- Exact methods
- Progressive (ClustalW)
- Iterative (MUSCLE)
- Consistency (ProbCons)
- Structure-based (Expresso)
- Conclusions: benchmarking studies

[3] Hidden Markov models (HMMs), Pfam and CDD

[4] MEGA to make a multiple sequence alignment

[5] Multiple alignment of genomic DNA

Multiple sequence alignment of genomic DNA

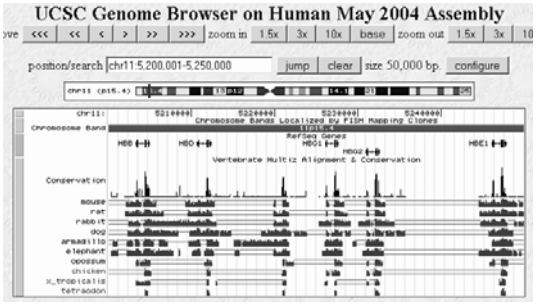
There are typically few sequences (up to several dozen), each having up to millions of base pairs. Adding more species improves accuracy.

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

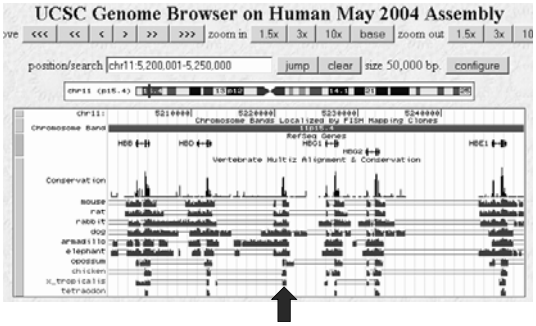
Chromosomes are subject to inversions, duplications, deletions, and translocations (often involving millions of base pairs). E.g. human chromosome 2 is derived from the fusion of two acrocentric chromosomes.

There are no benchmark datasets available.

Multiple alignment of genomic DNA at UCSC
50,000 base pairs (at http://genome.ucsc.edu)

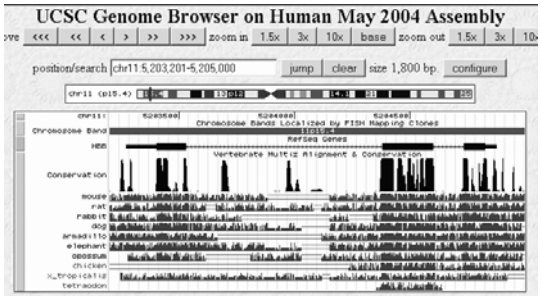


Note conserved regions: exons and regulatory sites (scale: 50,000 base pairs)

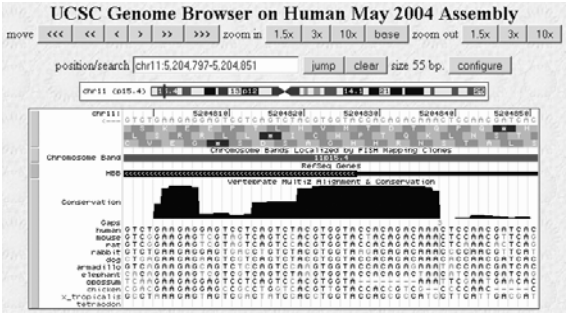


regulatory

Multiple alignment of beta globin gene
scale: 1,800 base pairs



Multiple alignment of beta globin gene
scale: 55 base pairs



This week: please download MEGA software and paste in a set of protein sequences. We'll use MEGA next week to make phylogenetic trees.

Download from www.megasoftware.net