Multiple Sequence Alignment
BIFX-550

S. *Ravi*chandran, Ph.D.
Biology,
Hood College, Frederick, MD 21701

- Online
  - Clustalw
  - NCBI
- Software
  - Jalview
  - http://www.jalview.org/Web_Installers/install.htm

Typos/issues with the Book

- Clustalw site mentioned in the book doesn't work
- Pfam-B DB no longer exists
- EBI, NCBI, TCOFFEE, ENSEMBL can be used for creating MSA
- Software (Mac and Windows versions)
  - http://www.jalview.org/download
- Any Questions

Discussion Questions (Chapter 4)

- [4.1] Why doesn’t anyone offer “Basic Global Alignment Search Tool” (BGAST) to complement BLAST?

- [4-2] Should you consider a significant expect value to be 1, 0.05, or 10−5? Does this depend on the particular search you are doing?

Discussion question

- [4-3] Why is it that database programs such as BLAST must make a trade-off between sensitivity and selectivity? How does the BLASTP algorithm address this issue?
  - What is sensitivity?
  - What is selectivity?

http://www.people.virginia.edu/~wrp/papers/wrp_protsci04.pdf

- Two characteristics of a search algorithm are important when searching a database: sensitivity and selectivity. A more sensitive algorithm will find a larger percentage of the total number of true positives, or homologs in the database, at a given threshold of statistical significance or false positives. A more selective algorithm will find a smaller number of false positives, or nonhomologs that receive high similarity scores to the query, at a given threshold of coverage. Generally there is a trade-off between these two characteristics, such that improving the performance of one degrades the performance of the other.

Scaled or Bit score

- “It can be used to compare alignments that may or may not have used the same scoring matrix.”

Aim

- How to create MSA of proteins?
- Databases that use MSA
  - PFAM
- MSA of genomic DNA
  - Regions from different species

Why MSA?

- How is my protein/gene related to other proteins/genes?
- What is different from pairwise alignments
  - Two sequences may not align well, but in presence of other sequences may align well to explain the relationship between them
- Protein functions are assigned based on homology (alignments) than biochemical assays (mostly)

Why MSA?

- Often used for
  - Secondary Structure perdition
  - Phylogenetic analysis
  - Generating Position-specific Scoring matrices for use with highly sensitive search such as RPS BLAST, PSI-BLAST, Delta-BLAST etc.
- Mutagenesis experiments

Considerations during MSA

- Goal is to create alignments with many aligning characters (sequence letters) as possible
- Keeping the function in mind
  - What can replace an amino acid without affecting the function of the protein?
- 3D structure can often guide or modify/correct the MSA

DNA vs Protein in MSA

- For MSA
  - Protein MSA is more informative on protein side
    - 20 aa vs 4 nt ; less prone for inaccuracies

- Exceptions
  - Regulatory elements use nucleotide

Definition

- “MSA is a collection of 3 or more protein (or nucleic acid) sequences that are partially or completely aligned”

- Each column
  - Homologous resides (evolutionary sense using structural information)
  - Same position in 3D structure

MSA

Not Easy to create a multiple sequence alignment

Easy Case

GADPH
Glyceraldehyde 3-phosphate dehydrogenase

Protein Acc.	Organism
NP_002037.2	H.sapiens
XP_508955.1	P.troglodytes
XP_001105471.1	M.mulatta
NP_001003142.1	C.lupus
XP_003435697.1	C.lupus
XP_003434435.1	C.lupus
NP_001029206.1	B.taurus
NP_032110.1	M.musculus
XP_001476757.1	M.musculus
NP_058704.1	R.norvegicus
NP_989636.1	G.gallus
NP_001108586.1	D.rerio
NP_001259584.1	D.melanogaster
NP_525108.2	D.melanogaster
XP_318655.2	A.gambiae
NP_496237.1	C.elegans
NP_496192.1	C.elegans
NP_508534.3	C.elegans
NP_508535.1	C.elegans
NP_012483.3	S.cerevisiae
NP_011708.3	S.cerevisiae
NP_012542.1	S.cerevisiae
XP_456022.1	K.lactis
NP_596154.1	S.pombe
NP_595236.1	S.pombe
XP_003717853.1	M.oryzae
XP_956977.1	N.crassa
NP_001060897.1	O.sativa
NP_001004949.1	X.tropicalis

```
NP_002037.2          1   MGKVKVGVNGFGRIGRLVTRAAFNSGKVDIVAINDPFIDLNYMVYMFQYD   50
XP_508955.1          1   MGKVKVGVNGFGRIGRLVTRAAFNSGKVDIVAINDPFIDLNYMVYMFQYD   50
XP_001105471.1       1   MGKVKVGVNGFGRIGRLVTRAAFNSGKVDIVAINDPFIDLNYMVYMFQYD   50
NP_001003142.1       1   --MVKVGVNGFGRIGRLVTRAAFNSGKVDIVAINDPFIDLNYMVYMFQYD   48
XP_003435697.1       1   --MVKVGVNGFGHIGRLVTRAAFNSGKVDIVAINDPFIDLNYMVYMFQYD   48
XP_003434435.1       1   --MVKVGVNGFGRIGRLVTRAAFNSGKVDIVTINDPFIDLNYMVYMFQYD   48
NP_001029206.1       1   --MVKVGVNGFGRIGRLVTRAAFNSGKVDIVAINDPFIDLHYMVYMFQYD   48
NP_032110.1          1   --MVKVGVNGFGRIGRLVTRAAICSGKVEIVAINDPFIDLNYMVYMFQYD   48
XP_001476757.1       1   --MVKVGVNGFGRIGRLVTRAAICSGKVEIVAINDPFIDLNYMVYMFQYD   48
NP_058704.1          1   --MVKVGVNGFGRIGRLVTRAAFSCDKVDIVAINDPFIDLNYMVYMFQYD   48
NP_989636.1          1   --MVKVGVNGFGRIGRLVTRAAVLSGKVQVVAINDPFIDLNYMVYMFKYD   48
NP_001108586.1       1   --MVKVGINGFGRIGRLVTRAAFLTKKVEIVAINDPFIDLDYMVYMFQYD   48
NP_001259584.1       1   --MSKIGINGFGRIGRLVLRAAIDKG-ANVVAVNDPFIDVNYMVYLFKFD   47
NP_525108.2          1   --MSKIGINGFGRIGRLVLRAAIDKG-ASVVAVNDPFIDVNYMVYLFKFD   47
XP_318655.2          1   --MSKIGINGFGRIGRLVLRAAITKG-ASVVAINDPFIGVDYMVYLFKYD   47
NP_496237.1          1   MSKANVGINGFGRIGRLVLRAAVEKDTVQVVAVNDPFITIDYMVYLFKYD   50
NP_496192.1          1   MSKANVGINGFGRIGRLVLRAAVEKDTVQVVAVNDPFITIDYMVYLFKYD   50
NP_508534.3          1   MTKPSVGINGFGRIGRLVLRAAVEKDSVNVVAVNDPFISIDYMVYLFQYD   50
NP_508535.1          1   MPKPSVGINGFGRIGRLVLRAAVEKDSVNVVAVNDPFISIDYMVYLFQYD   50
NP_012483.3          1   --MIRIAINGFGRIGRLVLRLALQRKDIEVVAVNDPFISNDYAAYMVKYD   48
NP_011708.3          1   --MVRVAINGFGRIGRLVMRIALSRPNVEVVALNDPFITNDYAAYMFKYD   48
NP_012542.1          1   --MVRVAINGFGRIGRLVMRIALQRKNVEVVALNDPFISNDYSAYMFKYD   48
XP_456022.1          1   --MVKVAINGFGRIGRLVLRIALQRKALEVVAVNDPFISVDYAAYMFKYD   48
NP_596154.1          1   MAIPKVGINGFGRIGRIVLRNALVAKTIQVVAINDPFIDLEYMAYMFKYD   50
NP_595236.1          1   MAIPKVGINGFGRIGRIVLRNAILTGKIQVVAVNDPFIDLDYMAYMFKYD   50
XP_003717853.1       1   --MVKCGINGFGRIGRIVFRNAIEHPDCEIVAVNDPFIEPKYAKYMLEYD   48
XP_956977.1          1   -MVVKVGINGFGRIGRIVFRNAIEHDDIHIVAVNDPFIEPKYAAYMLRYD   49
NP_001060897.1       1   MGKIKIGINGFGRIGRLVARVALQSEDVELVAVNDPFITTDYMTYMFKYD   50
NP_001004949.1       1   ----------------------------------------MAYMFKYD    8
```

Not So Easy Case

```
NP_005203.2          50   YVPNSYPYYGTNLYQRRPAIA-INNPYVPRTYYANPAV-VRPHAQIPQRQ   97
XP_001143538.1       50   YVPNSYPYYGTNLYQRRPAIA-INNPYVPRTYYANPAV-VRPHAQIPQRQ   97
XP_001095721.1       50   YVPNSYPYYGTNLYQHRPAIA-INNPYVPRTYYANPAV-VRPHAQIPQRQ   97
XP_005628327.1       51   YVLNSFSHYEPNYYPHRPAEP-INHQYVPYPFYAKPAVAVRTHAQIPQWQ   99
NP_776719.1          51   YVLSRYPSYGLNYYQQKPVAL-INNQFLPYPYYAKPAA-VRSPAQILQWQ   98
NP_031812.2          51   SVLN-FNQYEPNYYHYRPSLPATASPYMYYPLVVRLLL-LRSPAPISKWQ   98
NP_113750.1          50   SVLN-RNHYEPIYYHYRTSVP--VSPYAYFPVGLKLLL-LRSPAQILKWQ   95

NP_005203.2          98   YLPN--------SHPPTVVRRPNLHPSFIAIPPKKIQDKIIIPTINTIAT   139
XP_001143538.1       98   YLPN--------SHPPTVVHRPNLHPSFIAIPPKKIQDKIIIPTINTIAT   139
XP_001095721.1       98   YLPN--------SHLPTVVRRPNLHPSFIAIPPKKIQDKIIIPTINTITT   139
XP_005628327.1      100   VLPN--------AYPPTMMHRPQLHPSFIAIPPKKIQDKTSIPTINTIAT   141
NP_776719.1          99   VLSNTVPAKSCQAQPTTMARHPHPHLSFMAIPPKKNQDKTEIPTINTIAS   148
NP_031812.2          99   SMPN--------FPQSAGVPYAIPNPSFLAMPTNENQDNTAIPTIDPITP   140
NP_113750.1          96   PMPN--------FPQPVGVPHPIPNPSFLAIPTNEKHDNTAIPASNTIAP   137
```

Protein Acc.	Organism
NP_005203.2	H.sapiens
XP_001143538.1	P.troglodytes
XP_001095721.1	M.mulatta
XP_005628327.1	C.lupus
NP_776719.1	B.taurus
NP_031812.2	M.musculus
NP_113750.1	R.norvegicus

MSA outline

- No one correct way to align sequences
  - Protein sequence evolve time and more rapidly than their 3D folds
- Conditions for MSA
  - Choose homologous sequences (reasonable E-value cutoff)
  - A software
  - Maximize the total score of the a series of pairwise alignments
    - With appropriate gap-opening and extension penalties
  - Analyze/inspect the output
    - Remove large gaps contributing sequences
    - Rerun if necessary

How to carry out MSA

- Collect Homologous protein sequences
  - Orthologous/paralogous
- Choose appropriate software that uses appropriate scoring functions
- Choose appropriate parameters
  - Gap penalties
- Interpret the output and re-run if needed

How can residues/residue features help us to get a meaningful alignment

- Cysteines involved in S-S conserved
- TM domains
- Secondary structural elements should guide the alignment
- Loops and coils
- Order of secondary structural elements

Selecting Sequences

- Use reasonable high quality sequences
- What alignment method? Global or Local
  - Alignment time is a bottle-neck
  - Keep in mind that most alignment methods are ineffective when the # of sequences increases
- Is there a guideline for # of sequences?
  - 10-15 sequences (minimum but can be lower)
  - 50-100 sequences (upper limit)

Selecting Sequences

- Do you need to have the sequences of same length?
  - Yes (roughly of same length)
  - How can we achieve this?
    - We can trim the sequences
    - Use PSI-BLAST and other techniques to get some leads on how to accomplish this step.
- Use visualization software to guide you during this step.
  - Jalview or DiscoveryStudio sequence viewer

Selecting Sequences

- Use prealigned DBs and 3D structure based alignments for guidance.

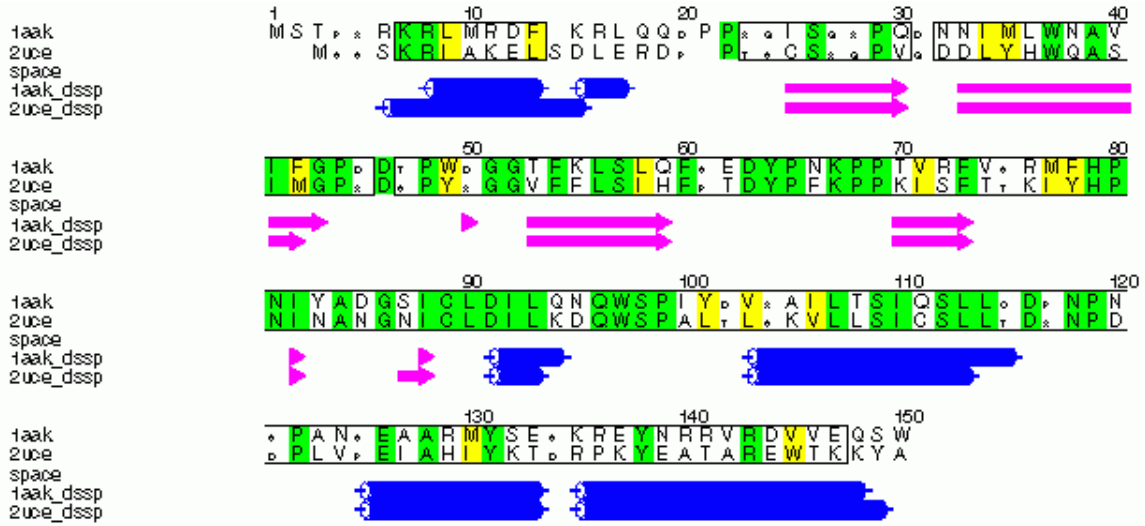
Comparing Homologous enzymes

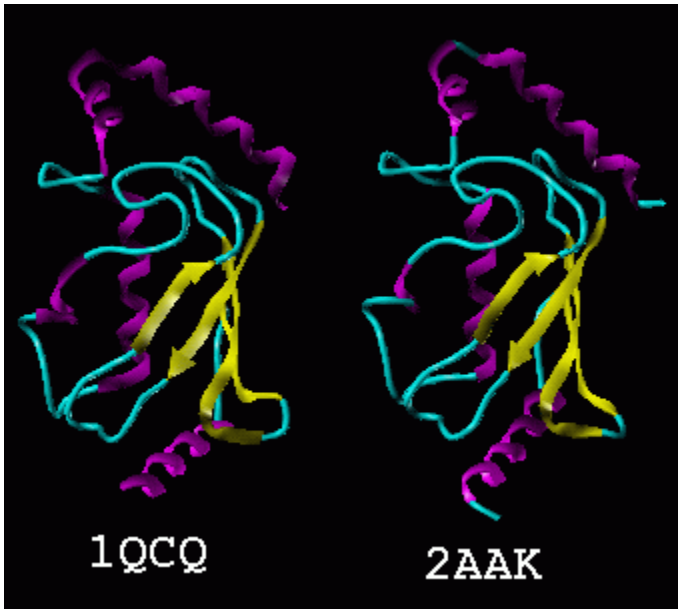
Family: Ubiquitin Conjugating enzyme

1QCQ: Arabidopsis Thaliana 2AAK: Baker's Yeast

Sequence Identity 43%

Topological Similarity





Using PredictProtein

Russell *et al,* JMB, **269,** 423-439 **1997**

Pairwise alignments doesn't work when comparing distant homologs

MSA will have more information to properly align the misaligned region

Where do we use MSA?

- Impact Assessment
  - SIFT, Polyphen2 etc.
  - How can we predict which ones are bad?
- MSA is a very sensitive method to detect distant homologs
- MSA can lead to critical residues in a family

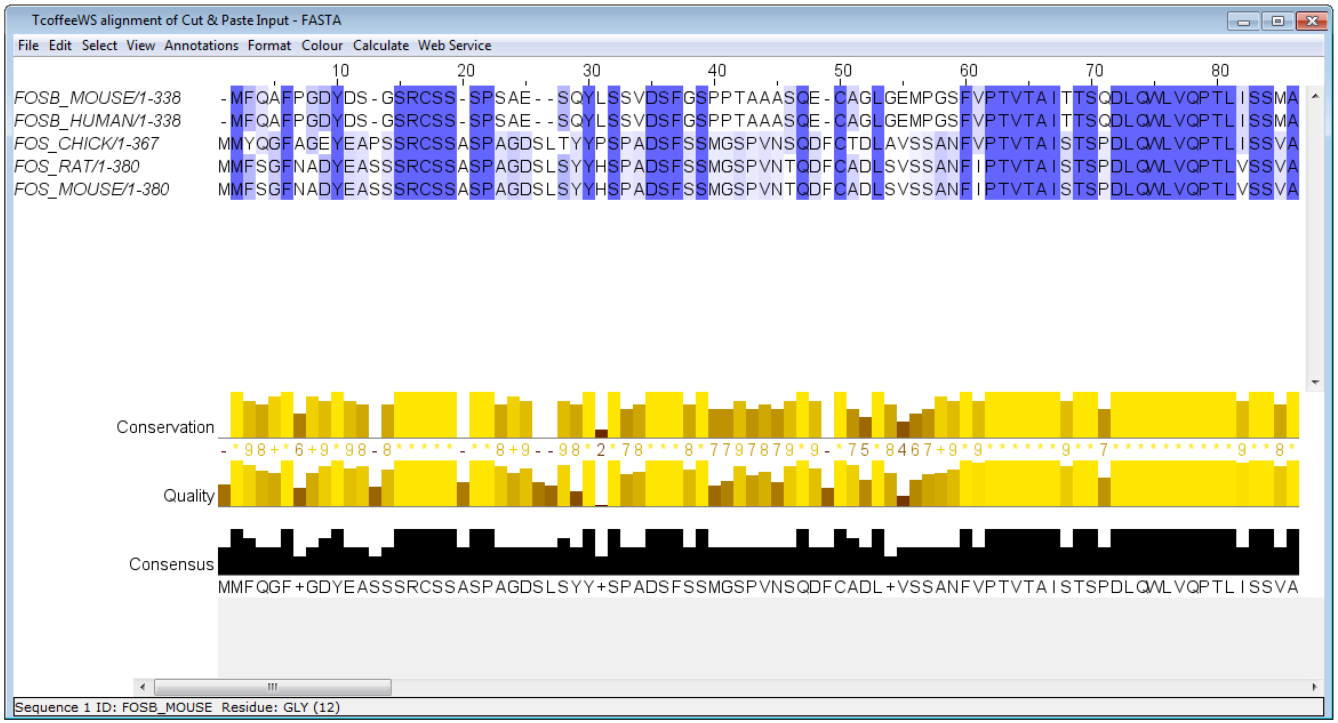
Where do we use MSA?

- Phylogenetic tree based algorithm's (upcoming class) begin with MSA
  - Optimal alignment is critical
- Regulatory regions of a gene
  - Transcription factor binding site
  - Conserved non-coding regions

Software for Viewing the MSA

- Jalview
- Demo
  - Sequences from next page

Conservation Color
81-100% Dark Blue
61-80% Medium Blue
41-60% Light Blue
<= 40% White



```
>FOSB_MOUSE Protein fosB
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT
DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY
TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL

>FOSB_HUMAN Protein fosB
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSSGGASGS
GGPSTSGTTSGPGPARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT
DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
LPGSAPAKEDGFSWLLPPPPPPPLPFQTSQDAPPNLTASLFTHSEVQVLGDPFPVVNPSY
TSSFVLTCPEVSAFAGAQRTSGSDQPSDPLNSPSLLAL

>FOS_CHICK Proto-oncogene protein c-fos
MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANF
VPTVTAISTSPDLQWLVQPTLISSVAPSQNRGHPYGVPAPAPPAAYSRPAVLKAPGGRGQ
SIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRNRRRELTDTLQAETDQLEEEKSALQA
EIANLLKEKEKLEFILAAHRPACKMPEELRFSEELAAATALDLGAPSPAAAEEAFALPLM
TEAPPAVPPKEPSGSGLELKAEPFDELLFSAGPREASRSVPDMDLPGASSFYASDWEPLG
AGSGGELEPLCTPVVTCTPCPSTYTSTFVFTYPEADAFPSCAAAHRKGSSSNEPSSDSLS
SPTLLAL

>FOS_RAT Proto-oncogene protein c-fos
MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF
IPTVTAISTSPDLQWLVQPTLVSSVAPSQTRAPHPYGLPTPSTGAYARAGVVKTMSGGRA
QSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRNRRRELTDTLQAETDQLEDEKSALQ
TEIANLLKEKEKLEFILAAHRPACKIPNDLGFPEEMSVTSLDLTGGLPEATTPESEEAFT
LPLLNDPEPKPSLEPVKNISNMELKAEPFDDFLFPASSRPSGSETARSVPDVDLSGSFYA
ADWEPLHSSSLGMGPMVTELEPLCTPVVTCTPSCTTYTSSFVFTYPEADSFPSCAAAHRK
GSSSNEPSSDSLSSPTLLAL

>FOS_MOUSE Proto-oncogene protein c-fos
MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF
IPTVTAISTSPDLQWLVQPTLVSSVAPSQTRAPHPYGLPTQSAGAYARAGMVKTVSGGRA
QSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRNRRRELTDTLQAETDQLEDEKSALQ
TEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEMSVASLDLTGGLPEASTPESEEAFT
LPLLNDPEPKPSLEPVKSISNVELKAEPFDDFLFPASSRPSGSETSRSVPDVDLSGSFYA
ADWEPLHSNSLGMGPMVTELEPLCTPVVTCTPGCTTYTSSFVFTYPEADSFPSCAAAHRK
GSSSNEPSSDSLSSPTLLAL
```

Another Example to try in Clustal Omega

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

Demo using ClustalW

Accessing MSA methods

1. Exact methods
   - Goal: Maximize the alignment score
   - Pros
     - Multi-dimensional version of Dynamic Programming
     - Exact result
   - Cons
     - $O(2^N L^N)$ (N: # of seqs; L = average sequence length
     - Not possible beyond 2 or 3 sequences

Accessing MSA methods

2. Progressive Sequence Alignment

- Fitch and Yasunobu (1975)
- Progressive: Strategy is to start with 2 (closest) and build the alignment by including the rest.
- Pros:
  - Can handle 100s or even thousands of seqs.
- Cons:
  - Final alignment depends on the order in which the alignments were joined.

Examples of Progressive Alignment Methods (Software)

- ClustalW
- Clustal Omega
- ProbCons
- Muscle
- T-COFFEE

- For the next exercise, use
  - http://www.ebi.ac.uk/Tools/msa/
- Clustal Omega
- Inputs are shown below

```
>beta_globin 2hhbB NP_000509.1 [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
>myoglobin 2MM1 NP_005359.1 [Homo sapiens]
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVL
TALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFR
KDMASNYKELGFQG
>neuroglobin 1OJ6A NP_067080.1 [Homo sapiens]
MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVML
VIDAAVTNVEDLSSLEEYLASLGRKHRAVGVKLSSFSTVGESLLYMLEKCLGPAFTPATRAAWSQLYGAV
VQAMSRGWDGE
>soybean_globin 1FSL leghemoglobin P02238 LGBA_SOYBN [Glycine max]
MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLANGVDPTNPKLTGHAEKLFALV
RDSAGQLKASGTVVADAALGSVHAQKAVTDPQFVVVKEALLKTIKAAVGDKWSDELSRAWEVAYDELAAA
IKKA
>rice_globin 1D8U rice Non-Symbiotic Plant Hemoglobin NP_001049476.1 [Oryza sativa (japonica)
MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLRNSDVPLEKNPK
LKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDAHFEVVKFALLDTIKEEVPADMWS
PAMKSAWSEAYDHLVAAIKQEMKPAE
```

Progressive Alignment

- Steps
  - All possible pairs of sequences aligned to get scores for each alignment
  - Similarity scores are used to construct a guide tree
  - Multiple alignment is built in a systematic manner using DP starting with the closest pair

Steps of Progressive Alignment

- Stage 1
  - Needleman and Wunch Global alignment is carried out
    - In this case 10 alignments were made

http://embossgui.sourceforge.net/demo/

Clustalw2

Five distantly related proteins

(a) Stage 1: series of pairwise alignments

SeqA	Name	Length	SeqB	Name	Length	Score
1	beta_globin	147	2	myoglobin	154	25.17
1	beta_globin	147	3	neuroglobin	151	15.65
1	beta_globin	147	4	soybean_globin	144	13.19
1	beta_globin	147	5	rice_globin	166	21.09
2	myoglobin	154	3	neuroglobin	151	16.56
2	myoglobin	154	4	soybean_globin	144	8.33
2	myoglobin	154	5	rice_globin	166	12.99
3	neuroglobin	151	4	soybean_globin	144	17.36
3	neuroglobin	151	5	rice_globin	166	18.54
4	soybean_globin	144	5	rice_globin	166	43.06

1

Compare the scores of distantly to closely related proteins

(a) Stage 1: series of pairwise alignments (closely related globin proteins)

SeqA	Name	Length	SeqB	Name	Length	Score
1	human_NP_000509	147	2	Pan_troglodytes_XP_508242	147	100.0
1	human_NP_000509	147	3	Canis_familiaris_XP_537902	147	89.8
1	human_NP_000509	147	4	Mus_musculus_NP_058652	147	80.27
1	human_NP_000509	147	5	Gallus_gallus_XP_444648	147	69.39
2	Pan_troglodytes_XP_508242	147	3	Canis_familiaris_XP_537902	147	89.8
2	Pan_troglodytes_XP_508242	147	4	Mus_musculus_NP_058652	147	80.27
2	Pan_troglodytes_XP_508242	147	5	Gallus_gallus_XP_444648	147	69.39
3	Canis_familiaris_XP_537902	147	4	Mus_musculus_NP_058652	147	78.91
3	Canis_familiaris_XP_537902	147	5	Gallus_gallus_XP_444648	147	71.43
4	Mus_musculus_NP_058652	147	5	Gallus_gallus_XP_444648	147	66.67

1: beta_globin	100.00	25.34	20.98	17.65	16.78
2: myoglobin	25.34	100.00	15.65	12.41	14.29
3: neuroglobin	20.98	15.65	100.00	18.12	19.59
4: soybean_globin	17.65	12.41	18.12	100.00	43.06
5: rice_globin	16.78	14.29	19.59	43.06	100.00

Stage 2

- Similarity scores between sequences are often converted into distance matrix
  - Introduced by Feng and Doolittle
- (N*(N-1))/2 is the number of alignments needed to computer the scores and matrices. This is why these approaches are slow

How to measure distances between sequences?

1) Count the number of mismatches in a pairwise alignment.
2) Feng and Doolittle
    1) Convert similarity scores to distance scores

$$D = -\ln S_{eff} \text{ (2 sequences i and j)}$$

$$S_{eff} = \frac{S_{real(i,j)} - S_{rand(i,j)}}{S_{iden(i,j)} - S_{rand(i,j)}} 100$$

$D = 0$; when $S_{eff} = 1$

$D = \infty$; when $S_{eff} = 0$

S_rand(i,j) = mean score from 1000 random shuffling(s) of sequences;
S_iden(ij) mean of (S_ii + S_jj)/2

Distance matrix to Guide Tree (upcoming classes)

- Two approaches
- UPGMA
  - Unweighted Pair-Group Method of Arithmetic averages
- Neighbor joining method
- Two things are important in trees
  - Topology (the branching order)
  - Branching length (drawn to reflect the evolutionary distance)

Finally Guide Tree is used as a guidance to create MSA

Aligned sequences (high score) are placed at the nodes of a tree, then progressively other sequences or pairs are added. Sometimes profiles are created and added to alignments or added to profiles

Compare the boxed sequence stretch with the next alignment (page 35)

5 distantly related proteins

Red residues indicate alpha helices using 3D structures

Tree

- Common tree output format is Newick

http://embossgui.sourceforge.net/demo/

Clustalw2

Five distantly related proteins

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

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

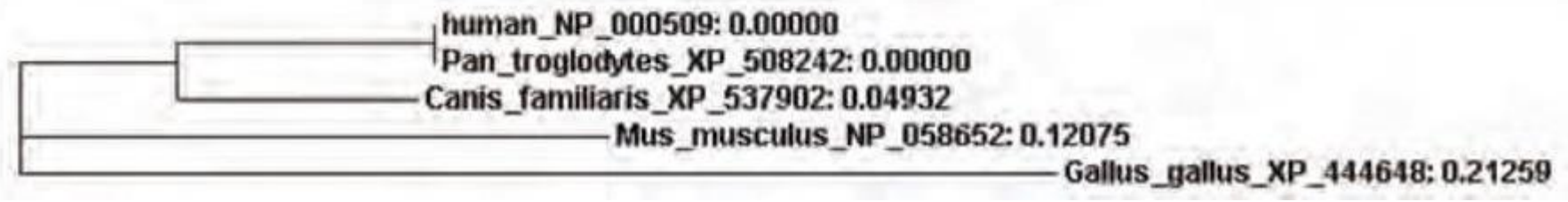
Soybean and rice are extant sequences; appear at the terminal nodes of the tree

Second example

Note the chicken has the lowest score in the alignments. This defines where the chicken finally is placed on the tree

(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);
```



Add other information

- Guide trees are not true phylogenetic trees
  - Phylogenetic trees are usually based on a model that includes multiple substitution on the aligned position
- Dynamic Programming can be used to align the closest pair identified in Guide Tree
- But they are used in the III stage in ClustalW to create MSA

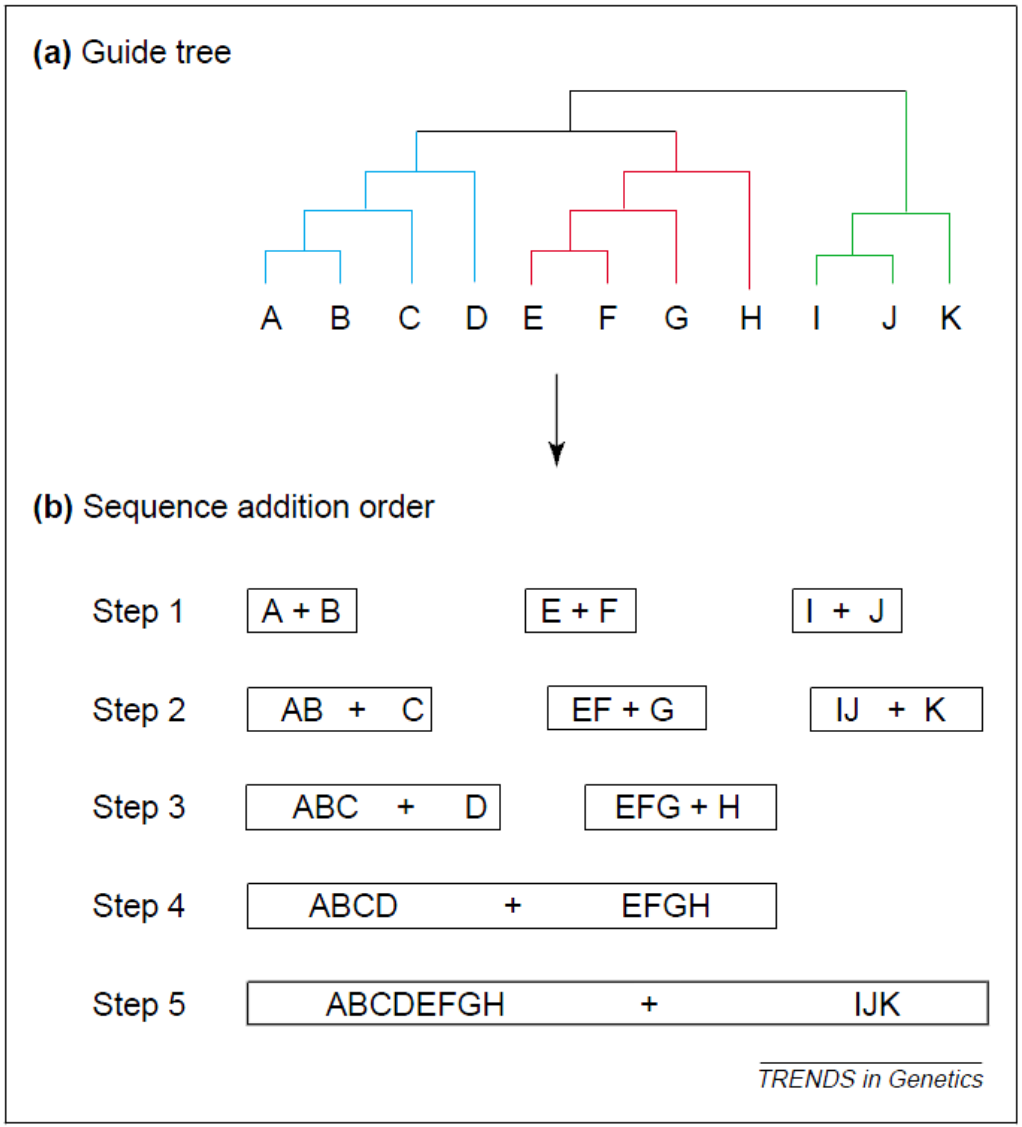
Stage 3

- MSA is created using guide tree
  - Select two most similar sequences
    - Appear at the terminal node of the tree
  - Next dissimilar sequence is either added to the first pair or a new pair with the sequences 3 and 4 (for ex.) and the pairs are then added together
  - This will be done until there are no sequences left

MSA of 5 closely related globins

Acidic/Basic: Blue/Red

Compare the boxed sequence stretch with the previous alignment; page 29



Review TRENDS in Genetics Vol.19 No.6 June 2003

Phylogeny for the faint of heart: a tutorial

Sandra L. Baldauf

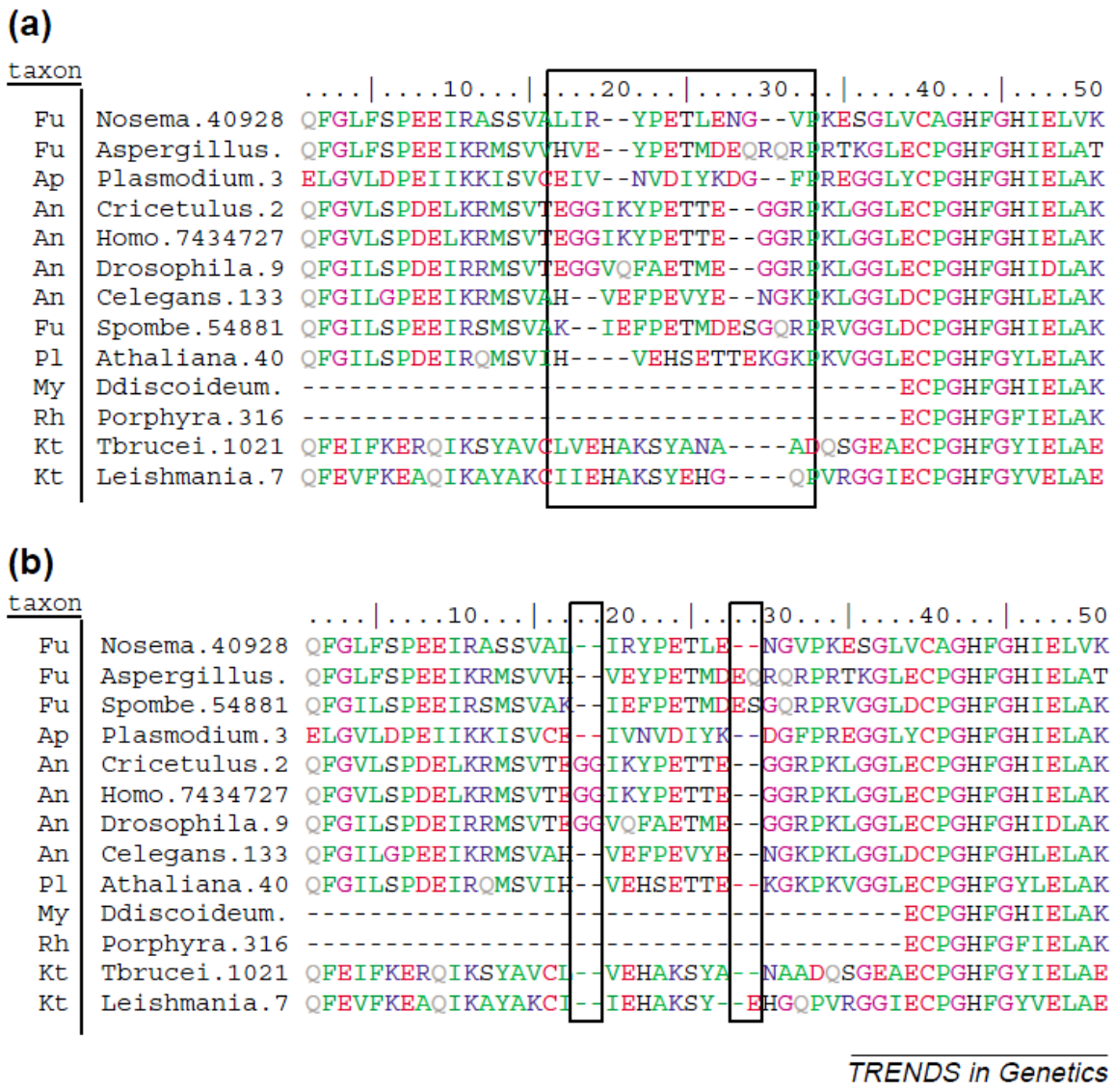
Department of Biology, University of York, Box 373, York, UK YO10 5YW

Figs 4a and b from the above paper

Feng-Doolittle Approach

- "Once a gap, always a gap"
  - 2 closely related sequence information should be weighed more. So, a gap shown in the first alignments most probably a true one!
- ClustalW
  - Once a gap always a gap policy
  - In addition, closely related sequences are downweighted (reducing their over-dominant impact on the alignment)

More on once a gap always a gap



Review TRENDS in Genetics Vol.19 No.6 June 2003

Phylogeny for the faint of heart: a tutorial

Sandra L. Baldauf

Department of Biology, University of York, Box 373, York, UK YO10 5YW

Figs 5a and b from the above paper

Failure of once a gap always a gap

Manual adjustment yields meaningful alignment

Once a gap always a gap

- In most genes, INDELs do not happen with ease
  - STOP codons, out of frame mutations yielding bad folding etc.
- So, gap has to be penalized for insertion more than extension

Other improvements

- Loytyonoja and Goldman addresses how to deal with deletions and how to reduce the domination of closely related sequences
  - Refer to page 214 of the book

Alignment is important for Trees

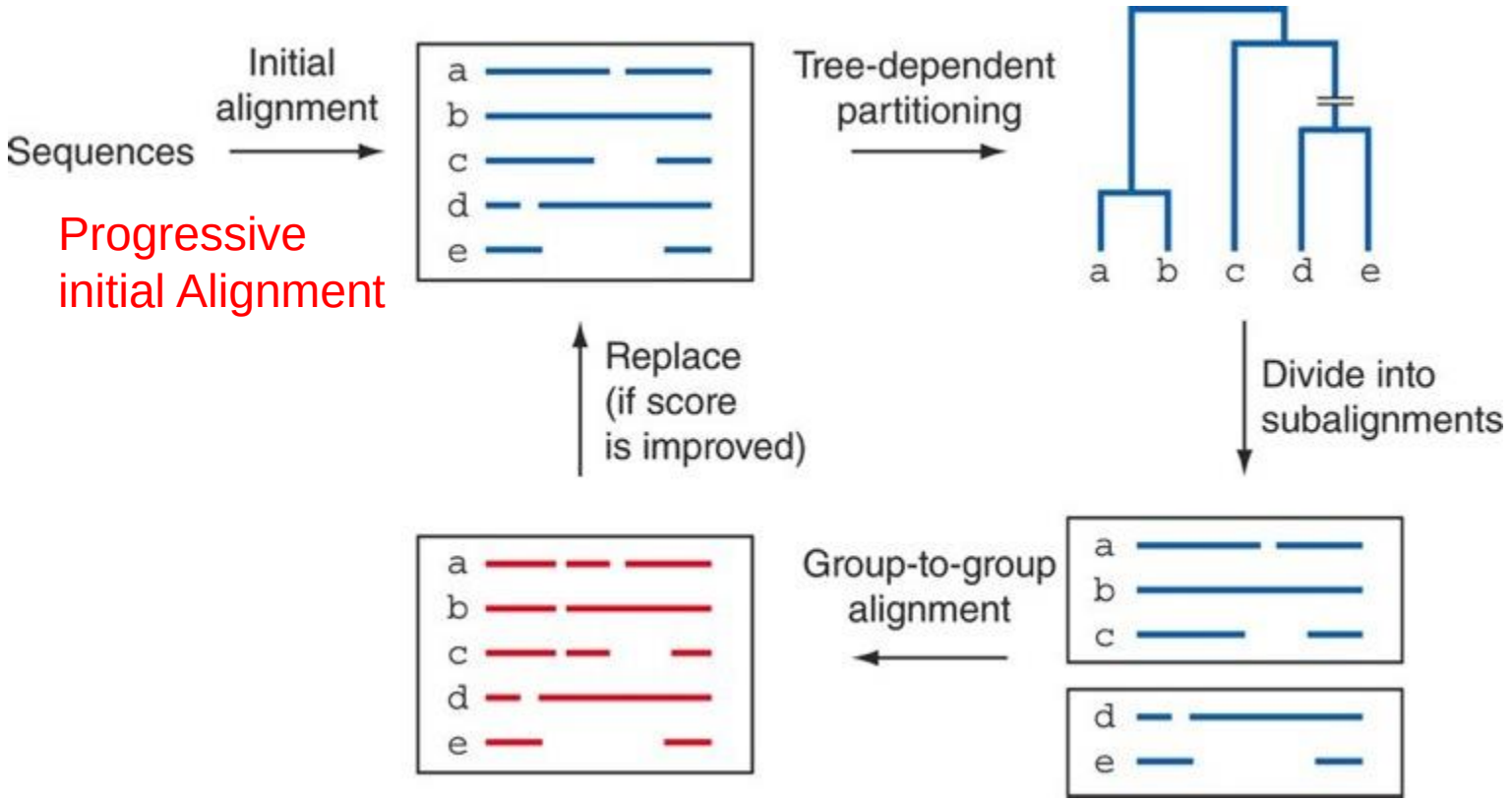
- An approach is to look at the alignment and delete all residues around gaps (and surrounding residues)
  - Why?
    - If there is misalignment then there is no real information
    - If there is a huge insertion and the rest of the alignment is all correct. The group that shares the extra (let us say 9-NT ins), then you have 9 extra characters for the OTUs that have it.

- In reality, a gap is a single evolutionary event (regardless of size)
  - The gap size is not important because the opening penalty is much bigger than extension penalty

Iterative Approaches

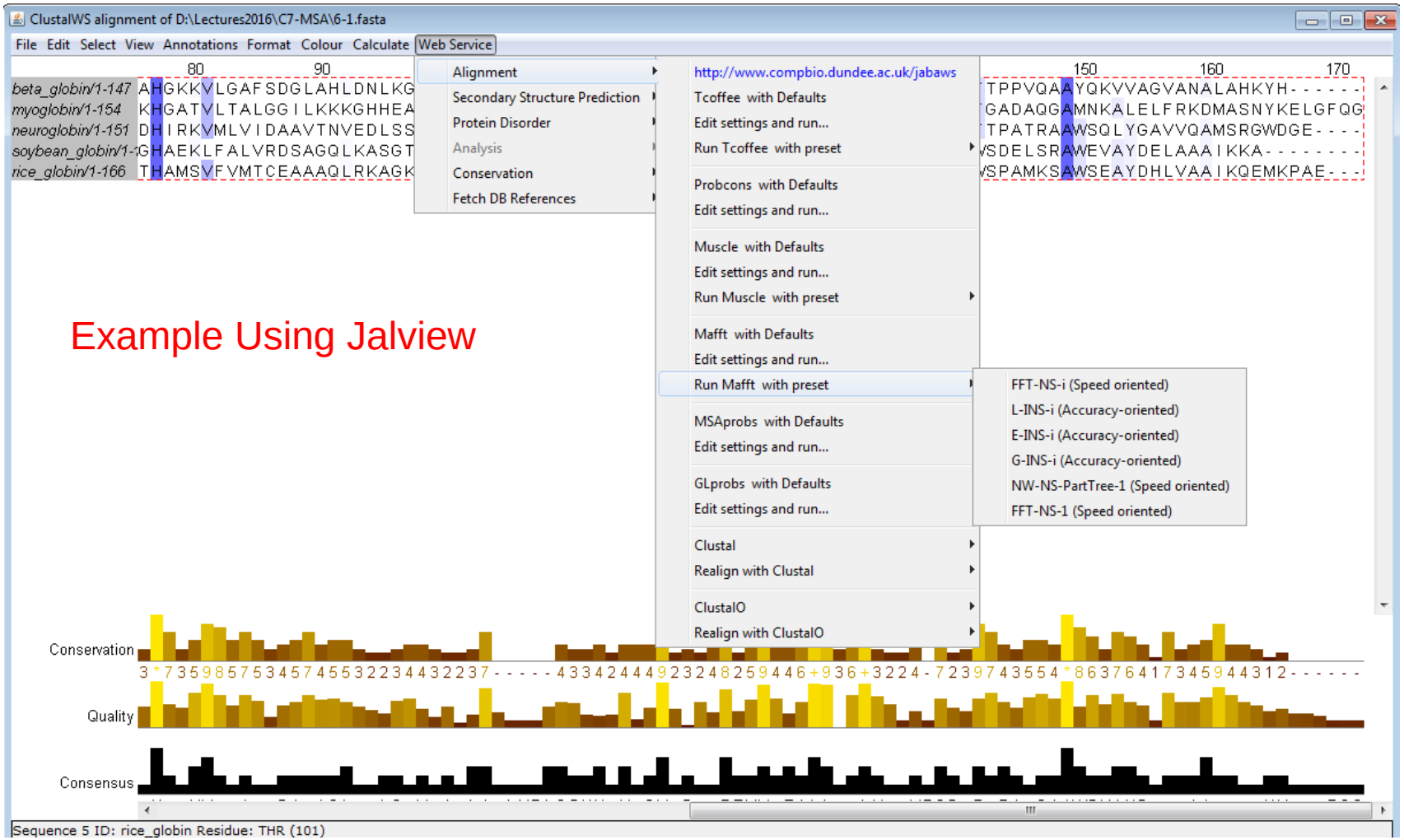
- Progressive methods have a limitation that once an error occurs it progresses and cannot be corrected
- Iterative approaches offer a solution to that approach
- So, what are Iterative approaches?

Iterative Refinement by MAFFT



Examples of Iterative Approach

- MAFFT
  - Highly accurate based on benchmarking studies
  - As implemented in Jalview seems to be very accurate and easy to use
  - L-ins-i
    - Seems to be most accurate; uses local alignment to improve the MSA


ClustalWS alignment of D:\Lectures2016\C7-MSA\6-1.fasta
File Edit Select View Annotations Format Colour Calculate Web Service
Alignment
Secondary Structure Prediction
Protein Disorder
Analysis
Conservation
Fetch DB References
http://www.compbio.dundee.ac.uk/jabaws
Tcoffee with Defaults
Edit settings and run...
Run Tcoffee with preset
Probcons with Defaults
Edit settings and run...
Muscle with Defaults
Edit settings and run...
Run Muscle with preset
Mafft with Defaults
Edit settings and run...
Run Mafft with preset
MSAprobs with Defaults
Edit settings and run...
GLprobs with Defaults
Edit settings and run...
Clustal
Realign with Clustal
ClustalO
Realign with ClustalO
FFT-NS-i (Speed oriented)
L-INS-i (Accuracy-oriented)
E-INS-i (Accuracy-oriented)
G-INS-i (Accuracy-oriented)
NW-NS-PartTree-1 (Speed oriented)
FFT-NS-1 (Speed oriented)
Example Using Jalview
Conservation
Quality
Consensus
Sequence 5 ID: rice_globin Residue: THR (101)

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

All MSAs are not same

We will compare MUSCLE with MAFFT

(book compares more than two methods)

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

Figure 6.7 from Pevsner Bioinformatics and Functional Genomics (III Ed) Copyright Figure

Look at the Alignment of 1, 2 and 3 positions

3 position is a conserved Histidine
Critical for binding of Oxygen is not aligned in Muscle method

Evaluate how the programs handle the alignment

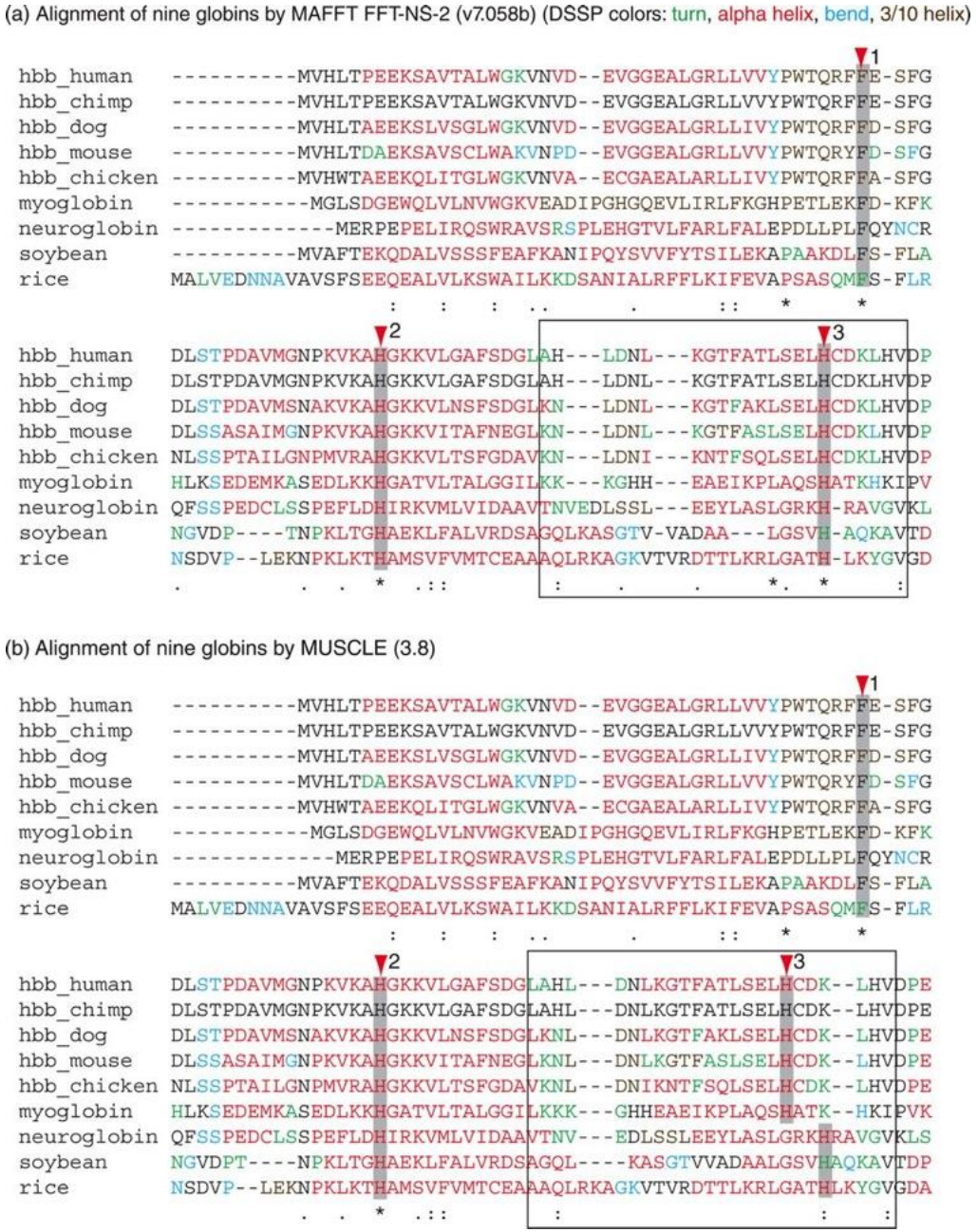
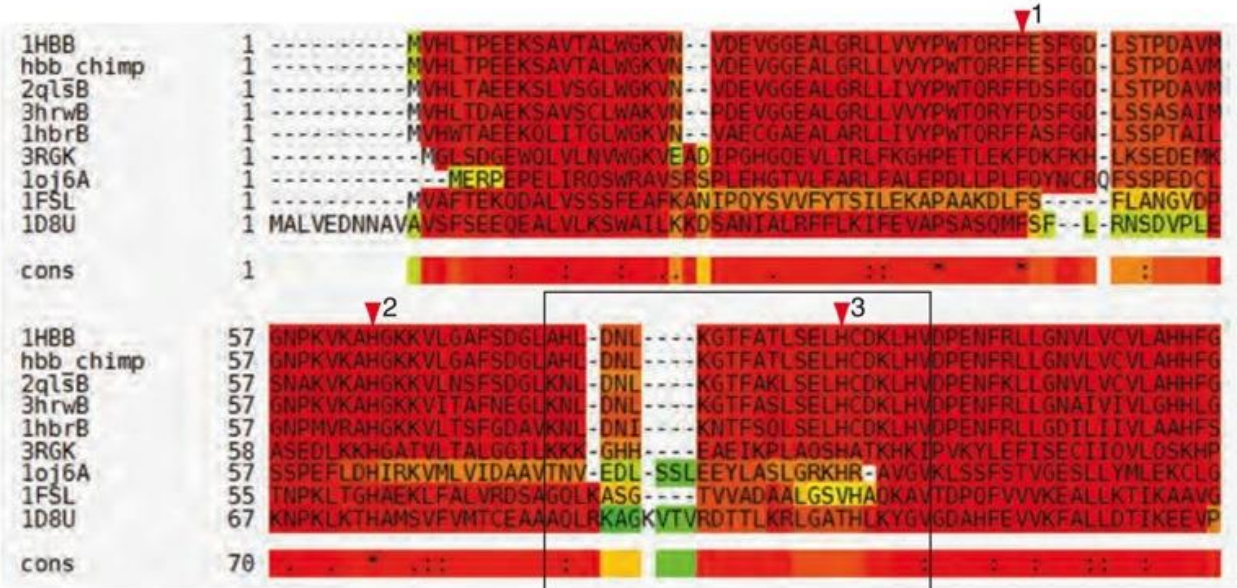


Figure 6.7 from Pevsner Bioinformatics and Functional Genomics (III Ed) Copyright Figure

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

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

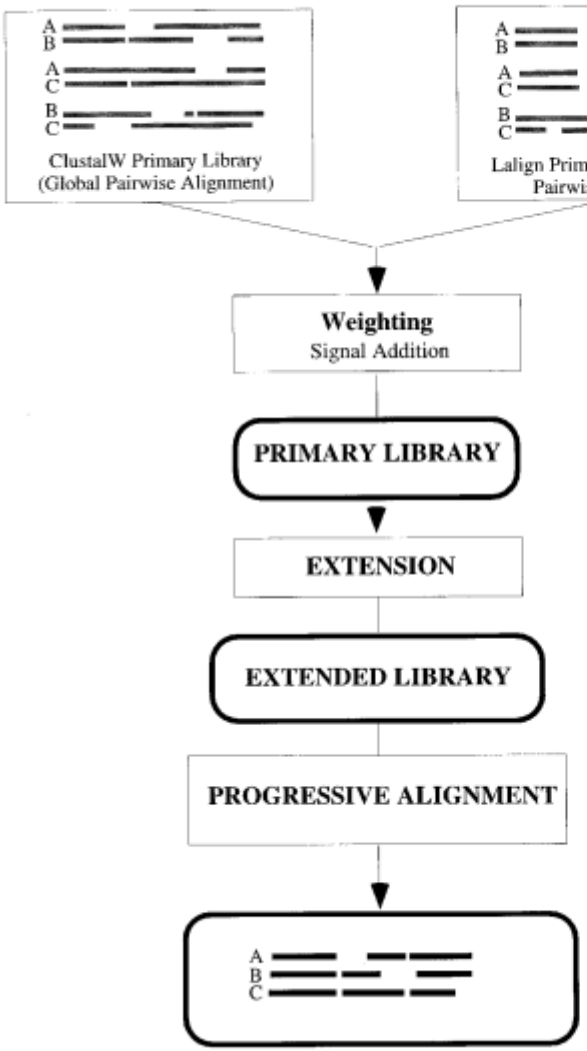


Consistency-based Approaches

$$CS = \sum_{i=1}^{M} \frac{C_i}{M}$$

- T-COFFEE
  - Tree-based Consistency Objective Function For Alignment Evaluation
  - Steps
    - Computes a library of pairwise alignments
      - Global alignments
      - Local alignments (10 highest scores are selected)
      - Each pair of aligned is assigned a weight
      - Primary/Extended Library
      - Progressive alignment (Library vs Query sequences)

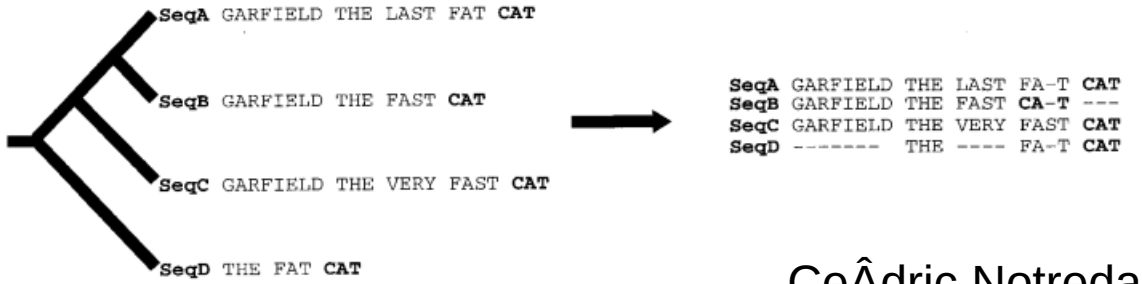
Objective scoring function: sum-of-pairs scores; Assign 1 if the all the residues in a column are aligned to the reference and 0 if not (crude method)



CeÂdric Notredame, Desmond G. Higgin
and Jaap Heringa, JMB, 302,205,2000

Figure 1. Layout of the T-Coffee strategy; the main steps required to compute a multiple sequence alignment using the T-Coffee method. Square blocks designate procedures while rounded blocks indicate data structures.

a)Regular Progressive Alignment Strategy



Note the word CAT is misaligned

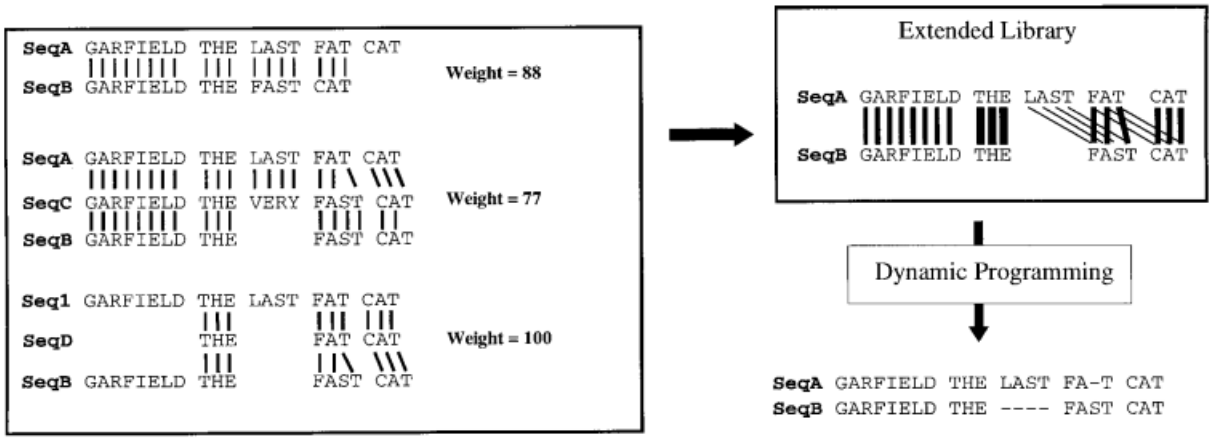
CeÂdric Notredame, Desmond G. Higgin and Jaap Heringa, JMB, 302,205,2000

b)Primary Library

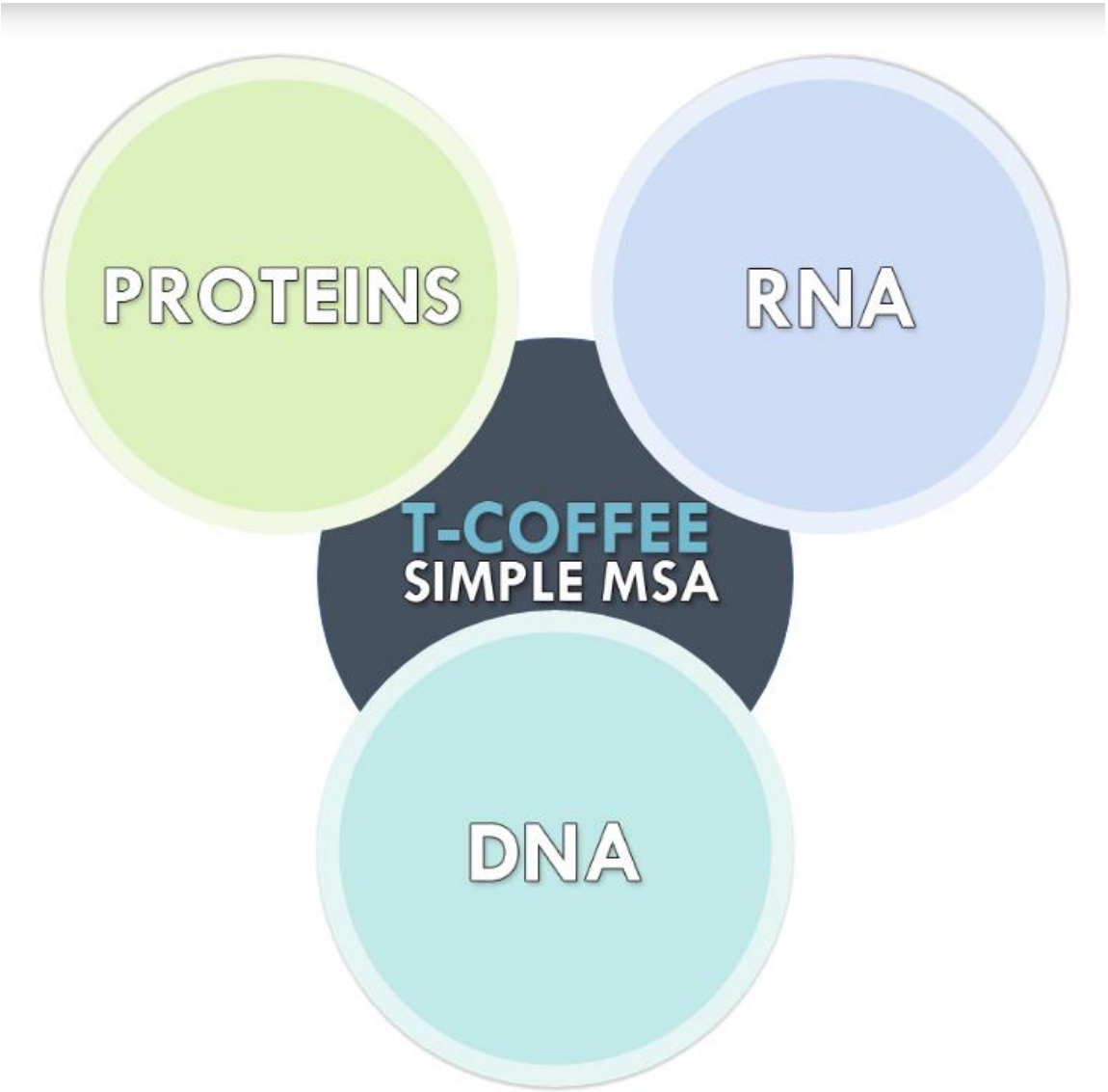
```
SeqA GARFIELD THE LAST FAT CAT   Prim. Weight = 88
SeqB GARFIELD THE FAST CAT ---

SeqA GARFIELD THE LAST FA-T CAT  Prim. Weight = 77
SeqC GARFIELD THE VERY FAST CAT

SeqA GARFIELD THE LAST FAT CAT   Prim. Weight =100
SeqD ------- THE ---- FAT CAT

SeqB GARFIELD THE ---- FAST CAT  Prim Weight = 100
SeqC GARFIELD THE VERY FAST CAT

SeqB GARFIELD THE FAST CAT       Prim. Weight = 100
SeqD ------- THE FA-T CAT

SeqC GARFIELD THE VERY FAST CAT  Prim. Weight = 100
SeqD ------- THE ---- FA-T CAT
```

c)Extended Library for seq1 and seq2

A and B are aligned through C and D



http://tcoffee.crg.cat/



Structure-based approach

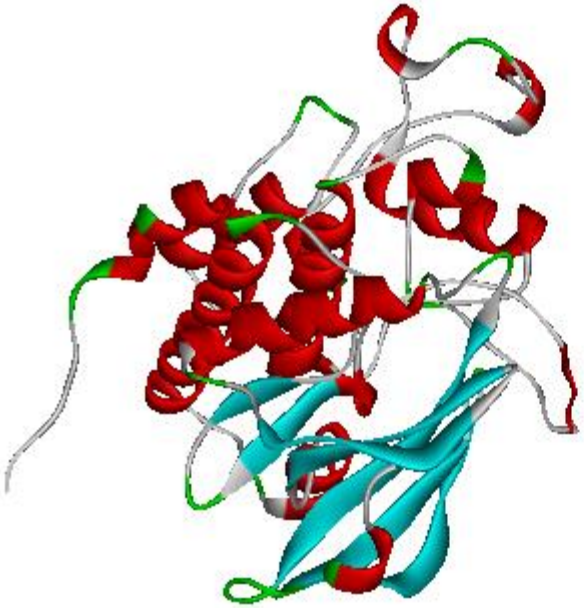
- Structure evolves slowly than sequences, so adding a structure to the sequence alignment is more helpful
- PRALINE and T-COFFEE (Expresso modules) are examples of this approach
- You can do this via T-COFFEE server
  - Add sequences;
  - Add PDB for some
  - Choose algorithm and submit

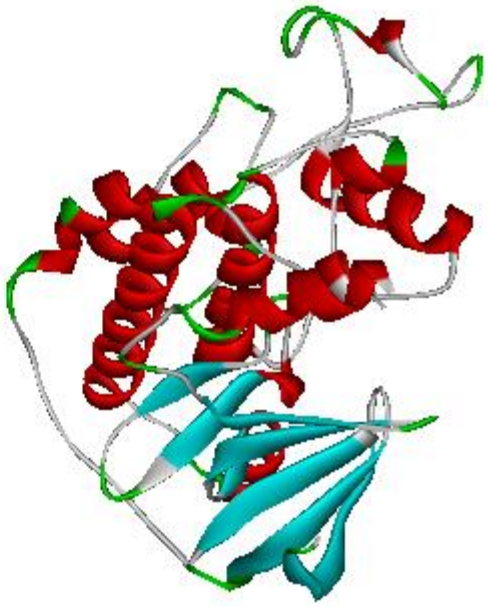
Cyclin Dependent
Kinase (Alpha + Beta)

42% identity
RMSD 1.8 Ang

Human CDK6

Human CDK2





1blx

1ckp

Structural Overlay done using DeepView

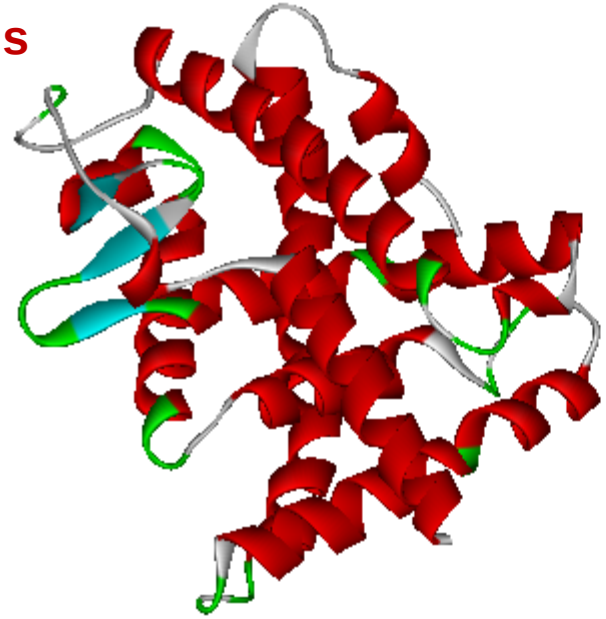
Nayeem et al, Protein Science, 15:808 (2006)

Nuclear Hormone Receptor (All Alpha)

Core fold is conserved

Loop regions differ

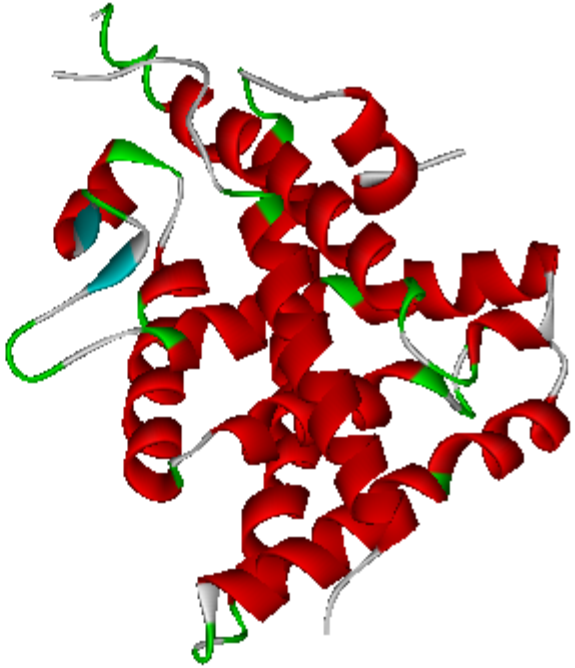
Human Nuclear receptor subfamily 1 group B member 3



2LBD

38% identity
RMSD 1.3 Ang

Human Nuclear receptor NER



1P8D

Structural Overlay done using DeepView

Nayeem et al, Protein Science, 15:808 (2006)

```
1CYD_C|PDBID|CHAIN|SEQUENCE      MKLNFSGLRALVTGAGKGIGRDTVKALHASGAKVVAVTRTNSDLVSL--- 47
1A27__|PDBID|CHAIN|SEQUENCE      -----ARTVVLITGCSSGIGLHLAVRLASDPSQSFKVYATLRDLKTQGRL 45
                                       :   .*:**...***  .  .  * :. ::  .  *  *  ** :

1CYD_C|PDBID|CHAIN|SEQUENCE      -----AKECP--GIEPVCVDLGDWDATEKALG--GIGPVDLLVNNAALVI 88
1A27__|PDBID|CHAIN|SEQUENCE      WEAARALACPPGSLETLQLDVRDSKSVAAARERVTEGRVDVLVCNAGLGL 95
                                      *  **   .:*.: :*: * .:.  *      * **:** **.* :

1CYD_C|PDBID|CHAIN|SEQUENCE      MQPFLEVTKEAFDRSFSVNLRSVFQVSQMVARDMINRGVPGSIVNVSSMV 138
1A27__|PDBID|CHAIN|SEQUENCE      LGPLEALGEDAVASVLDVNVVGTVRMLQAFLPDMKRRGS-GRVLVTGSVG 144
                                 : *:  : ::*.   :.**: ...:: * .  ** .**  * ::  ..*:

1CYD_C|PDBID|CHAIN|SEQUENCE      AHVTFPNLITYSSTKGAMTMLTKAMAMELGPHKIRVNSVNPTVVLTDMGK 188
1A27__|PDBID|CHAIN|SEQUENCE      GLMGLPFNDVYCASKFALEGLCESLAVLLLPFGVHLSLIECGPVHTAFME 194
                                 . : :*   .*.::* *:  * :::*: * *. :::. ::   * * : :

1CYD_C|PDBID|CHAIN|SEQUENCE      KVSADPEFARKLKERHPLRKFAEVEDVVNSIL----------FLLSDR 226
1A27__|PDBID|CHAIN|SEQUENCE      KVLGSPEEVLDRTDIHTFHRFYQYLAHSKQVFREAAQNPEEVAEVFLTAL 244
                                 ** ..**  .  . .: *.:::* :     :.::           .:*:

1CYD_C|PDBID|CHAIN|SEQUENCE      SASTSGGGILVDAGYLAS------------------------- 244
1A27__|PDBID|CHAIN|SEQUENCE      RAPKPTLRYFTTERFLPLLRMRLDDPSGSNYVTAMHREVFGDVPA 289
                                  *...     :.   :*.
```

ClustalW 1.83
www.ebi.ac.uk

19% identity
RMSD
1.7 Ang

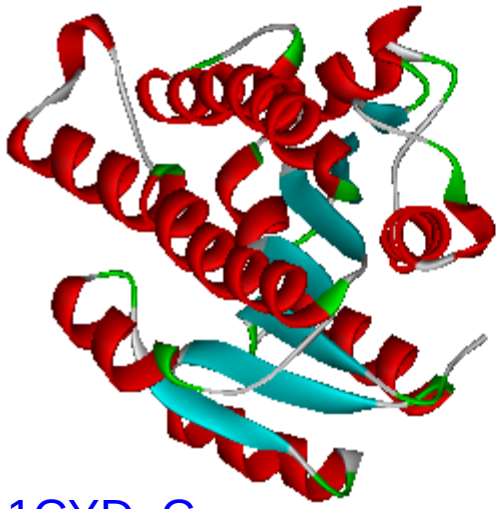
Human Estradiol 17-beta-dehydrogenase 1

Mouse Carbonyl reductase [NADPH] 2

Nayeem et al, Protein Science, 15:808 (2006)

Short Chain Dehydrogenase (Alpha + Beta)

Structural Overlay done using DeepView



1CYD_C

1A27

Structure-based approach

- One can compare sequence alignment
- Repeat them with structure information
- One can compare how well are we doing
- Are the residues important for function are aligned?

Run T-COFFEE

```
>hbb_human
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL
DNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
>hbb_chimp
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL
DNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
>hbb_dog
MVHLTAEEKSLVSGLWGKVNVDEVGGEALGRLLIVYPWTQRFFDSFGDLSTPDAVMSNAKVKAHGKKVLNSFSDGLKNL
DNLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVVAGVANALAHKYH
>hbb_mouse
MVHLTDAEKSAVSCLWAKVNPDEVGGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNPKVKAHGKKVITAFNEGLKNL
DNLKGTFASLSELHCDKLHVDPENFRLLGNAIVIVLGHHLGKDFTPAAQAAFQKVVAGVATALAHKYH
>hbb_chicken
MVHWTAEEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLTSFGDAVKNL
DNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH
>myoglobin
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVLTALGGILKK
KGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG
>neuroglobin
MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNV
EDLSSLEEYLASLGRKHRAVGVKLSSFSTVGESLLYMLEKCLGPAFTPATRAAWSQLYGAVVQAMSRGWDGE
>soybean
MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLANGVDPTNPKLTGHAEKLFALVRDSAGQLKA
SGTVVADAALGSVHAQKAVTDPQFVVVKEALLKTIKAAVGDKWSDELSRAWEVAYDELAAAIKKA
>rice
MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLRNSDVPLEKNPKLKTHAMSVF
VMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDAHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAI
KQEMKPAE
```

iRMSD-APDB from T-COFFEE server

- http://tcoffee.crg.cat/apps/tcoffee/do:irmsd

- Repeat the calculations using sequence/profile based methods and when we compare with structure based methods, do we learn anything new?

Methods for calculating trees

- Distance-based methods:
  - a.k.a clustering or algorithmic methods
    - UPGMA, neighbor-joining etc
- Discrete data methods
  - Tree searching methods
    - Parsimony, maximum likelihood, Bayesian methods

Methods for calculating trees

- Distance-Matrix Methods (we briefly looked at it before)
  - Distance could be % sequence difference is computed for all pairwise combinations of OTUs
  - Assemble distances into a tree

Methods for calculating trees

- Discrete methods
  - Examine each column separately
  - Then build trees that accommodates all the information
    - More information than Distance based methods
    - But, Distance based methods are faster
  - Also computes a hypothesis for each column, you can use that to comment about evolution
    - Catalytic sites or regulatory sites etc.

Benchmarking MSA

- How do we know, how we are doing?
- There are gold standard (true positive) sets available for us to compare
- Good estimators of Bench-mark
  - Solvability: sequences should not be too easy > 50 % similarity or too difficult
  - Scalability: The time to finish the problems
  - Free access to the results
  - Evolution: Should be able to modify and test

Benchmarking MSA

- Good estimators of Bench-mark
  - Evolution: Should be able to expand over time and include newer cases
- Benchmarks available
  - BALIbase
  - And others
- One good way to choose the sequences is based on the availability of structural information

Benchmarking MSA

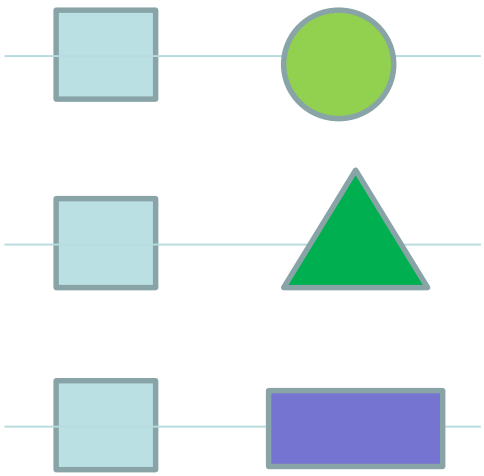
- But not all sequences have 3D structures
- Non-core cannot be aligned well and the structures cannot offer any help there

Metrics commonly used to evaluate the performance

- Sum-of-pairs scores
  - Counting the number of pairs between the target and the reference alignment, divided by the total number of pairs in the reference
    - Cons: works best for global alignment; Not evolutionary basis

Databases for MSA

Sequence → Function



Multiple Seq Alignment (MSA)

DNA Binding Function to the domains

Why Sequence Alignment is Important?

Consensus Sequence (Majority is taken as the consensus; skip Variation)
Search DB

Not very useful (binary etc.)

Pattern (regular expression)
GHx(2)[IV] (no probabilities)

PSSM (uses frequencies for each position)

Profiles (extension of PSSM; with gaps)

Profiles

		-	A	G	G	C	T	A	T	C	A	C	C	T	G
		T	A	G	-	C	T	A	C	C	A	-	-	-	G
		C	A	G	-	C	T	A	C	C	A	-	-	-	G
		C	A	G	-	C	T	A	T	C	A	C	-	G	G
		C	A	G	-	C	T	A	T	C	A	C	-	G	G
A		0	1	0	0	0	0	1	0	0	0.8	0	0	0	0
C		0.6	0	0	0	1	0	0	0.4	1	0	0.6	0.2	0	0
G		0	0	1	0.2	0	0	0	0	0	0.2	0	0	0.4	1
G		0.2	0	0	0	0	1	0	0.6	0	0	0	0	0.2	0
-		0.2	0	0	0.8	0	0	0	0	0	0	0.4	0.8	0.4	0

Profile based alignment

```
S1    G T C  T    G   A
S2    G T C  A    G   C
S2,4  G T C  A/T  GA  A/C (profile)
```

- 4 sequences
- 1,2,3,4
- 1-2, 1-3, 1-4, 2-3, 2-4, 3-4
- Find the high scoring pair and create a profile
- Use the profile to align to the next pair
  - Use DP for this step
- …..

Aligning Sequences

- Aligning sequence vs Sequence
  - We know how to do this
- Can we align sequence vs Profile
  - Yes
- Can we align Profile vs Profiles
  - Yes

Andy talk

Patterns

[FY]-x-C-x(2)-{VA}-x-H(3)

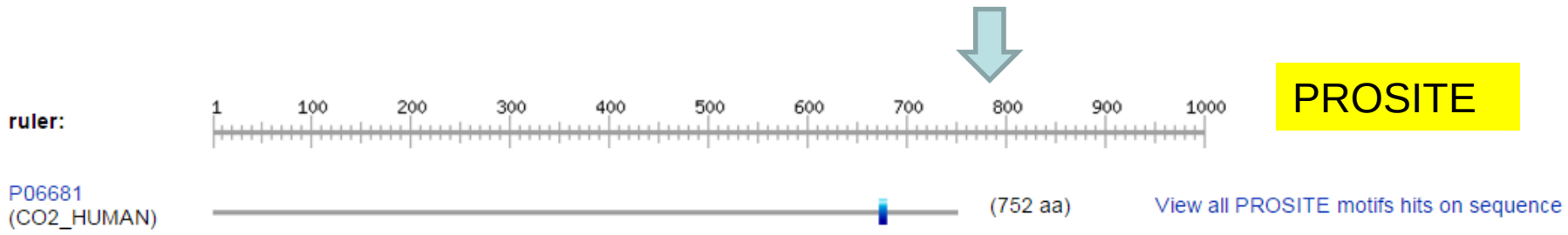
[] OR ex [FY] F or Y
x Any
X(2) any two residues
{} NOT example {VA} not V or A
H(3) 3 HIS residues

- PFAM

```
GHEGV
GHKKV
GHIKV
GH**V
```

```
GHEGV
GHKKI
GHIKV
GHAMI
GH**[I/V]
```

P-x(2)-G-E-S-G(2)-[AS] ⟹ http://prosite.expasy.org/scanprosite/



Complement C2 (EC 3.4.21.43) (C3/C5 convertase) [Cleaved into: Complement C2b fragment; Complement C2a fragment]. *Homo sapiens (Human)*

USERPAT1 : Hits on PDB 3D structures: [2I6Q-A, 2I6S-A, 2ODP-A, 2ODQ-A]
Pattern: **P-x(2)-G-E-S-G(2)-[AS]**
Approximate number of expected random matches [Ref: PMID 11535175] in ~ 100'000 sequences (50'000'000 residues): 0.44

674 - 682: PckGESGGA

NCBI Tool

```
  1 mgplmvlfcl lflypglads apscpqnvni sggtftlshg wapgslltys cpqglypspa
 61 srlckssgqw qtpgatrsls kavckpvrcp apvsfengiy tprlgsypvg gnvsfecedg
121 filrgspvrq crpngmwdge tavcdngagh cpnpgislga vrtgfrfghg dkvryrcssn
181 lvltgssere cqgngvwsgt epicrqpysy dfpedvapal gtsfshmlga tnptqktkes
241 lgrkiqiqrs ghlnlyllld csqsvsendf lifkesaslm vdrifsfein vsvaiitfas
301 epkvlmsvln dnsrdmtevi sslenanykd hengtgtnty aalnsvylmm nnqmrllgme
361 tmawqeirha iilltdgksn mggspktavd hireilninq krndyldiya igvgkldvdw
421 relnelgskk dgerhafilq dtkalhqvfe hmldvskltd ticgvgnmsa nasdqertpw
481 hvtikpksqe tcrgalisdq wvltaahcfr dgndhslwrv nvgdpksqwg kefliekavi
541 spgfdvfakk nqgilefygd diallklaqk vkmstharpi clpctmeanl alrrpqgstc
601 rdhenellnk qsvpahfval ngsklninlk mgvewtscae vvsqektmfp nltdvrevvt
661 dqflcsgtqe despckgesg gavflerrfr ffqvglvswg lynpclgsad knsrkraprs
721 kvppprdfhi nlfrmqpwlr qhlgdvlnfl pl
```

PFAM

- Precomputed MSA
  - Of protein domains and Conserved Domains that are important for function
- Proteins
  - Contain many domains
    - Many of them are repeated
  - The domain combinations are responsible for function
  - Unknown proteins, if identified, can assign function

PFAM

Book is outdated on this Database; there are no more PFAM-A and PFAM-B

- PFAM-A
- Experts at EBI manually collect members of a known family, align and create a seed alignment (done via HMMER software)
- In the next step, HMMER using seed alignment finds other members that are belonging to the family.
- New members are then used to generate a “full alignment” for the protein family

PFAM DB Summary

- Consists of 2 alignments
  - Pfam-A and full-alignments (larger)
- Pfam-A
  - Seed alignments; small high confidence number of representative family members
  - Stable ID (accession numbers)
  - Expert curated

PFAM

- Full alignment
  - Additional sequences are aligned to Pfam-A (ex PF00042; globins)
  - Not annotated/accession numbers

Query:
beta globin 2hhbB NP_000509.1 [Homo sapiens]

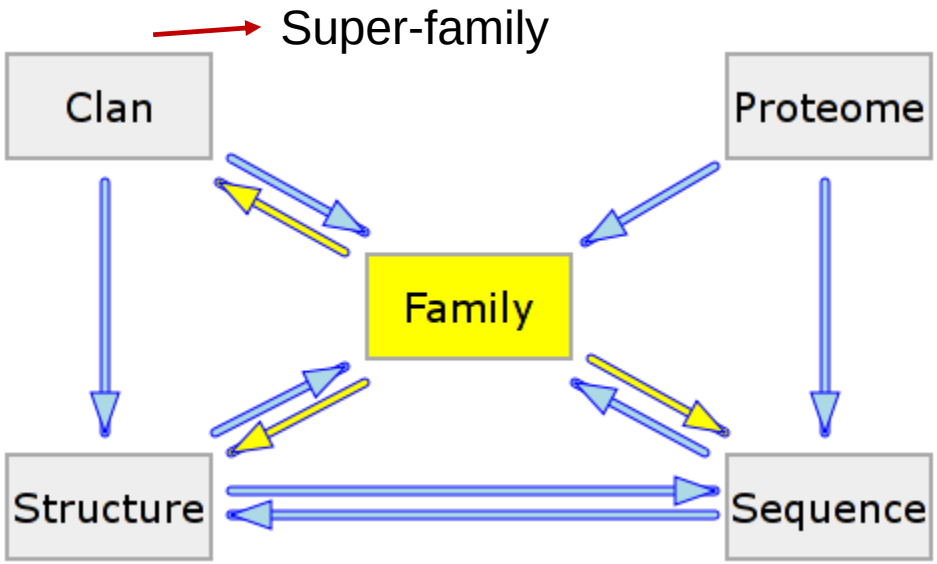
Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value
				Start	End	Start	End	From	To			
Globin	Globin	Domain	CL0090	8	112	8	112	1	110	110	106.6	8.5e-31

```
#HMM      dkalvkaswekvkanaeelgaeilkrlFkaypdtkklFkkfgdlssaedlksspkfkahakkvlaalgeavknldndealkaalkklaarHaergkvdpanFklfgeall
#MATCH    +k++v a+w+kv  n +e+g e+l rl+ +yp+t+++F++fgdls+ ++++++pk+kah+kkvl+a+++ +++l  d +lk+++++l++ H++++vdp+nF+l+g++l+
#PP       6899********..99**************************************************..7.**************************9986
#SEQ      EKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL--D-NLKGTFATLSELHCDKLHVDPENFRLLGNVLV
```

Clan:
Bac_globin; Globin; Phycobilsome and Protoglobin

PFAM

*"a **clan** as a collection of families that have arisen from a single evolutionary origin.*



Identifying distant family members

Globin-like

Add annotation

The globin fold is an evolutionary conserved six helical fold that is found in bacteria and eukaryotes.

This clan contains **4** families and the total number of domains in the clan is **6147**. The clan was built by RD Finn.

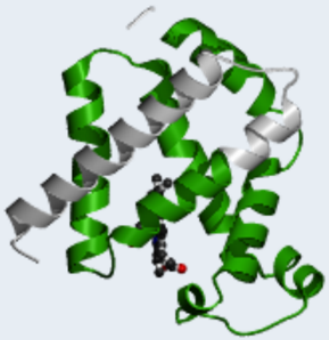
Members

This clan contains the following 4 member families:

Bac_globin Globin Phycobilisome Protoglobin

External database links

CATH:	1.10.490.10
SCOP:	46458

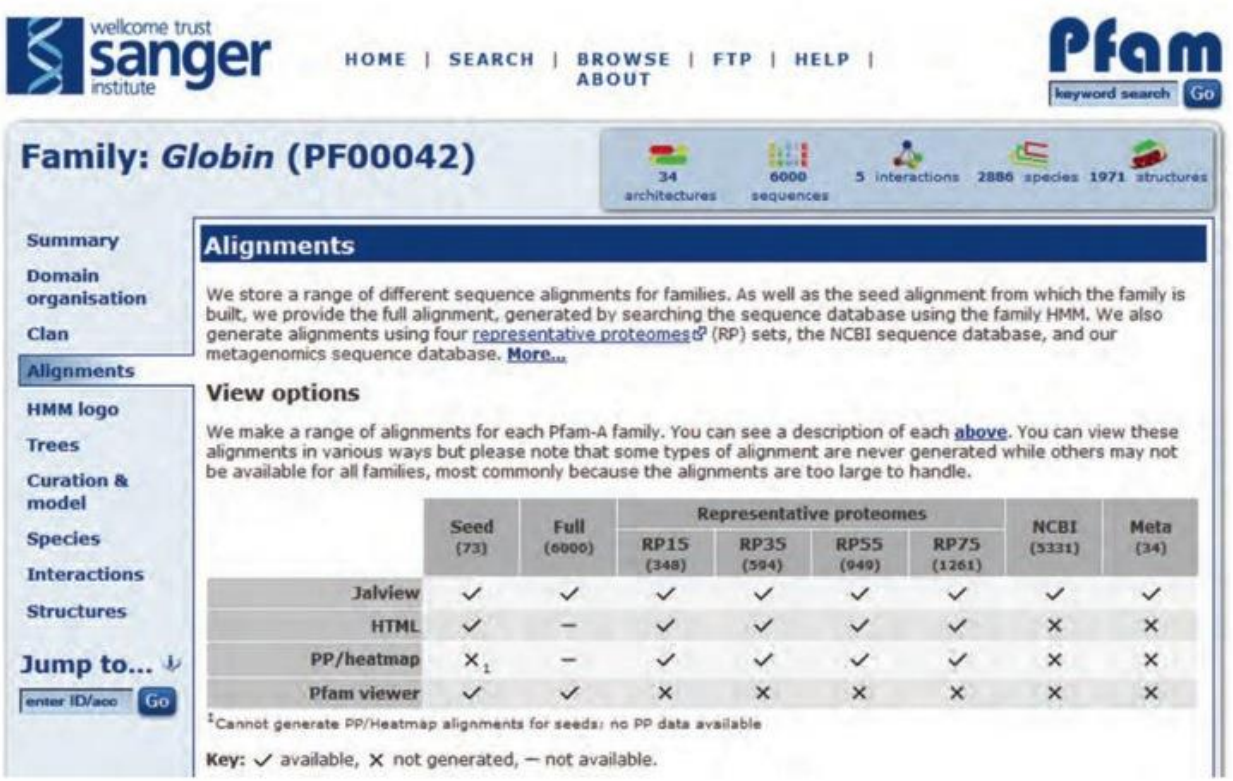


Example structure

PDB entry 1V5H: Crystal Structure of Human Cytoglobin (Ferric Form)
View a different structure:
1V5H

Mycobacterium leprae (a globin from the leprosy-causing bacterium)

(a) Pfam alignments



(b) Pfam seed alignment

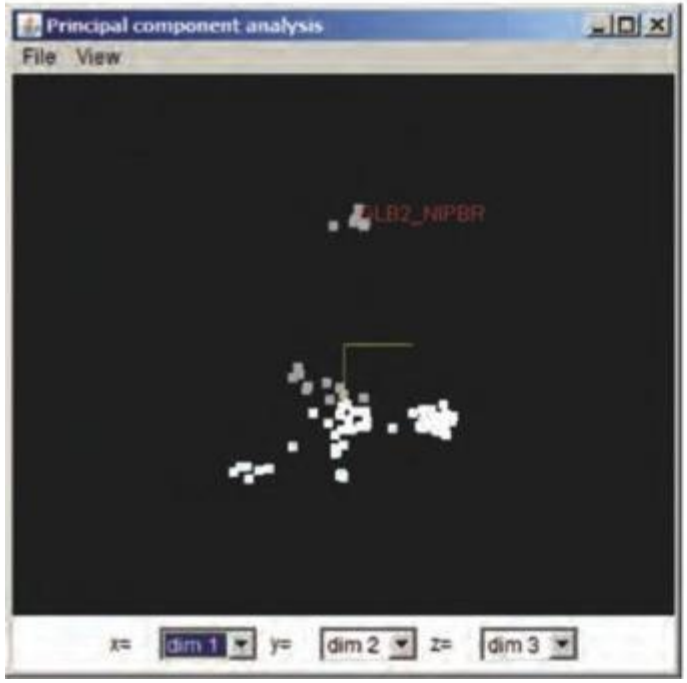
PFAM → PF00044 → Download and analyze
Or use
Jalview to carry out the analysis

Each point is one sequence;

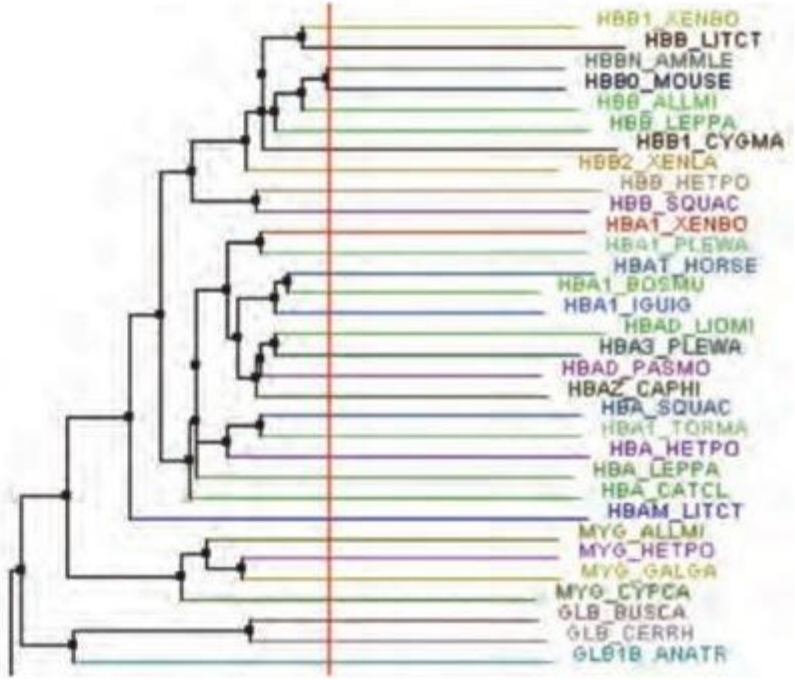
Here what we are doing is reducing the dimensionality to view the set of 23 proteins in a smaller dimensional space.

Visualization of Pfam seed alignment of globins using JalView

(a) Principal components analysis (PCA)

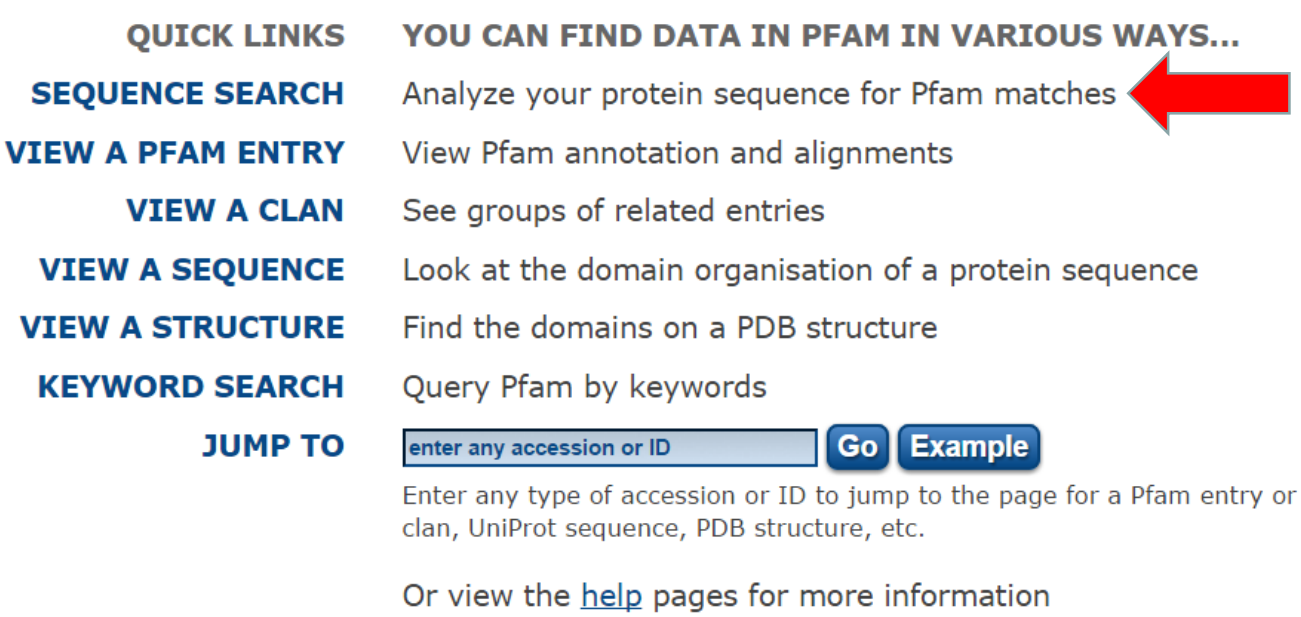


(b) Neighbor-joining tree



Example using PFAM

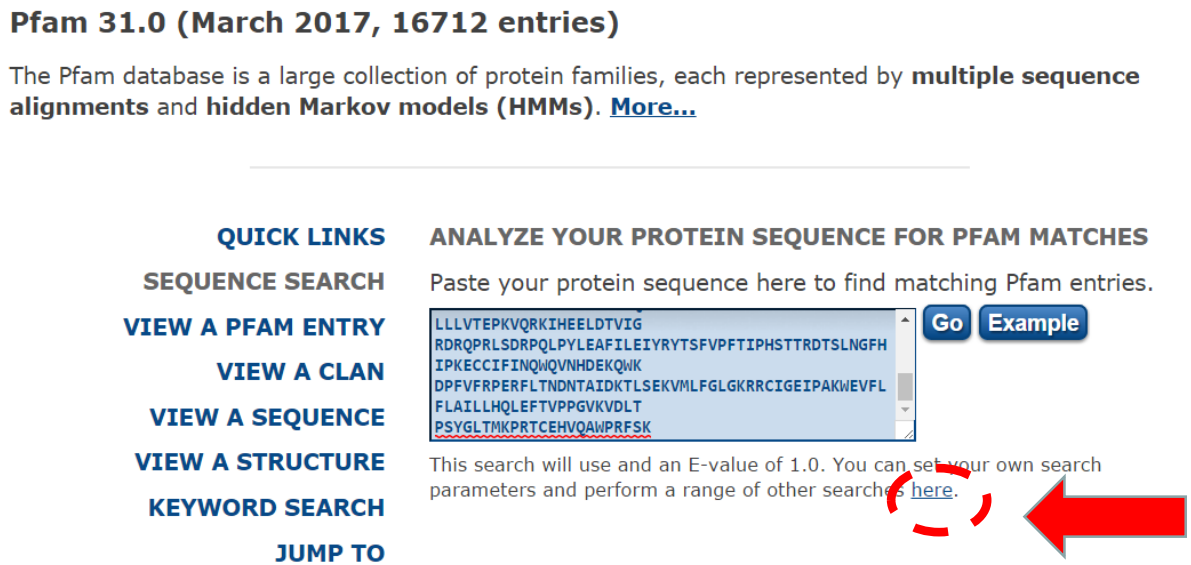
http://pfam.xfam.org/



Copy the sequence from next page

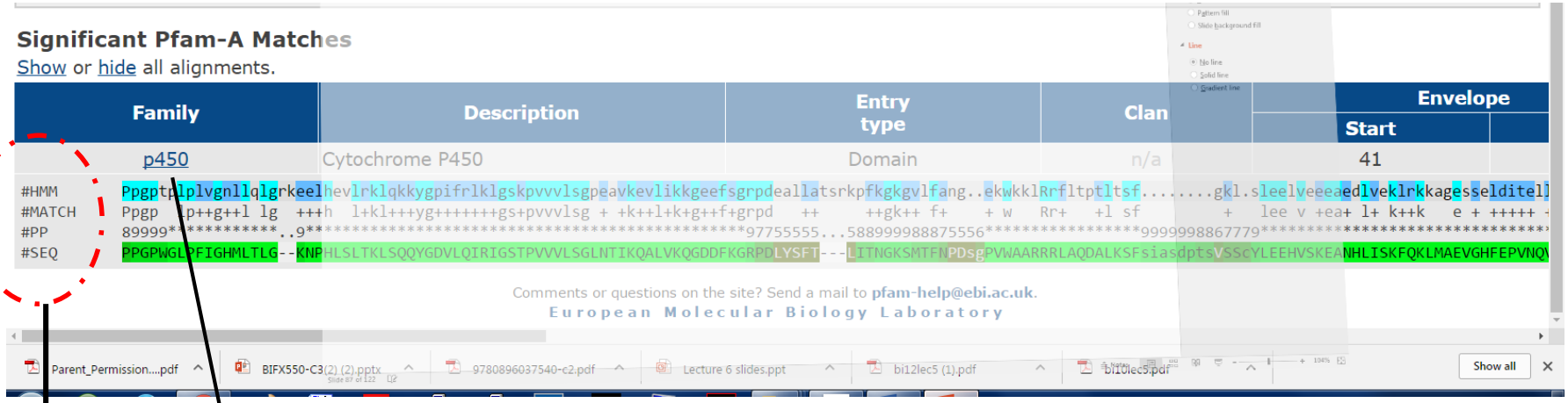
>Pfam_Query_sequence
MAFSQYISLAPELLLATAIFCLVFWVLRGTRTQVPKGLKSPPG
PWGLPFIGHMLTLGKNPHLSLTKLSQQ
YGDVLQIRIGSTPVVVLSGLNTIKQALVKQGDDFKGRPDLYSF
TLITNGKSMTFNPDSGPVWAARRRLAQ
DALKSFSIASDPTSVSSCYLEEHVSKEANHLISKFQKLMAEVG
HFEPVNQVVESVANVIGAMCFGKNFPR
KSEEMLNLVKSSKDFVENVTSGNAVDFFPVLRYLPNPALKRF
KNFNDNFVLSLQKTVQEHYQDFNKNSIQ
DITGALFKHSENYKDNGGLIPQEKIVNIVNDIFGAGFETVTTAIF
WSILLLVTEPKVQRKIHEELDTVIG
RDRQPRLSDRPQLPYLEAFILEIYRYTSFVPFTIPHSTTRDTSL
NGFHIPKECCIFINQWQVNHDEKQWK
DPFVFRPERFLTNDNTAIDKTLSEKVMLFGLGKRRCIGEIPAK
WEVFLFLAILLHQLEFTVPPGVKVDLT
PSYGLTMKPRTCEHVQAWPRFSK

Example using PFAM

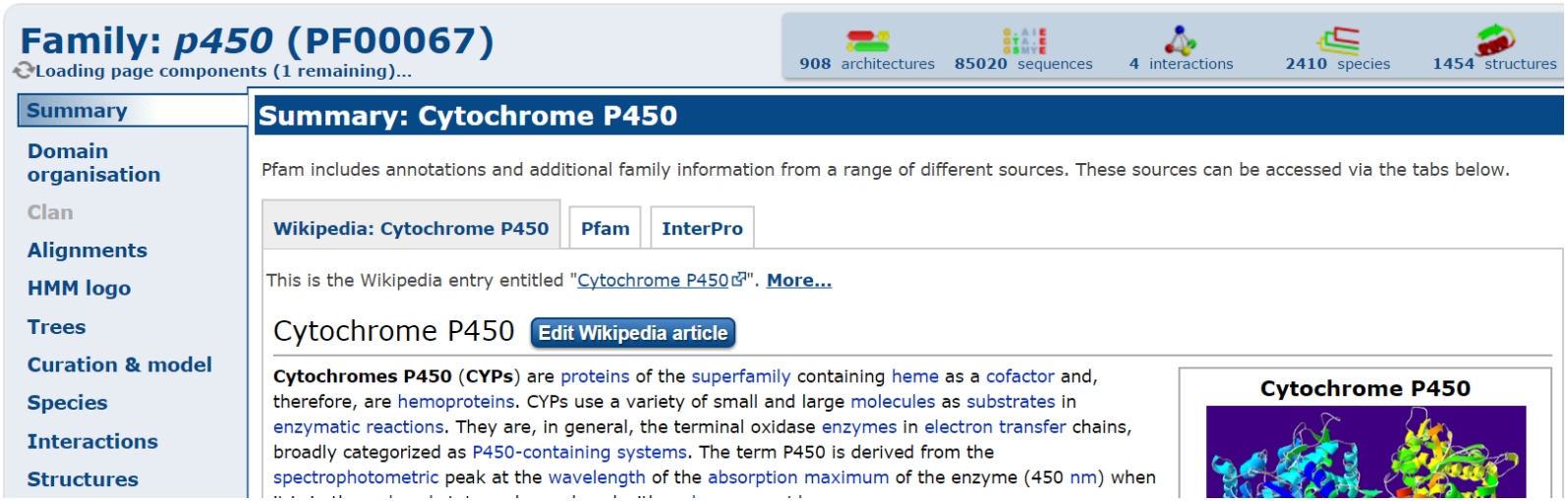


Click here for additional options

After exploring, go back to the previous screen and submit the sequence with default options



#Seq is your query; #PP is the posterior Probability, Higher the better;
Top line HMM is the model sequence that represents the family



SMART

- Smart Modular Architecture Research Tool
- http://smart.embl-heidelberg.de/
- DB
  - Protein families involved in cellular signaling
  - Extracellular domains
  - Chromatin function
- Uses profile HMM like PFAM
  - Also uses HMMER software

SMART

- Used in two modes
- Normal Mode
  - DB searches against Swiss-Prot, SP-TrEMBL and Stable Ensembl proteomes
- Genomic Mode
  - DB searches against Organisms from Ensembl and SwissProt including Eukaryotes, Bacteria and Archea

Conserved Domain Db (CDD)

- Identify conserved domains in protiens
- Gets guidance from 3D information to identify domain boundaries
- Source data
  - Pfam A
  - SMART
  - COG (orthologous prokaryotic protein families)
  - TIGRFAM
  - PRK (Protein clusters of related RefSeq entries)

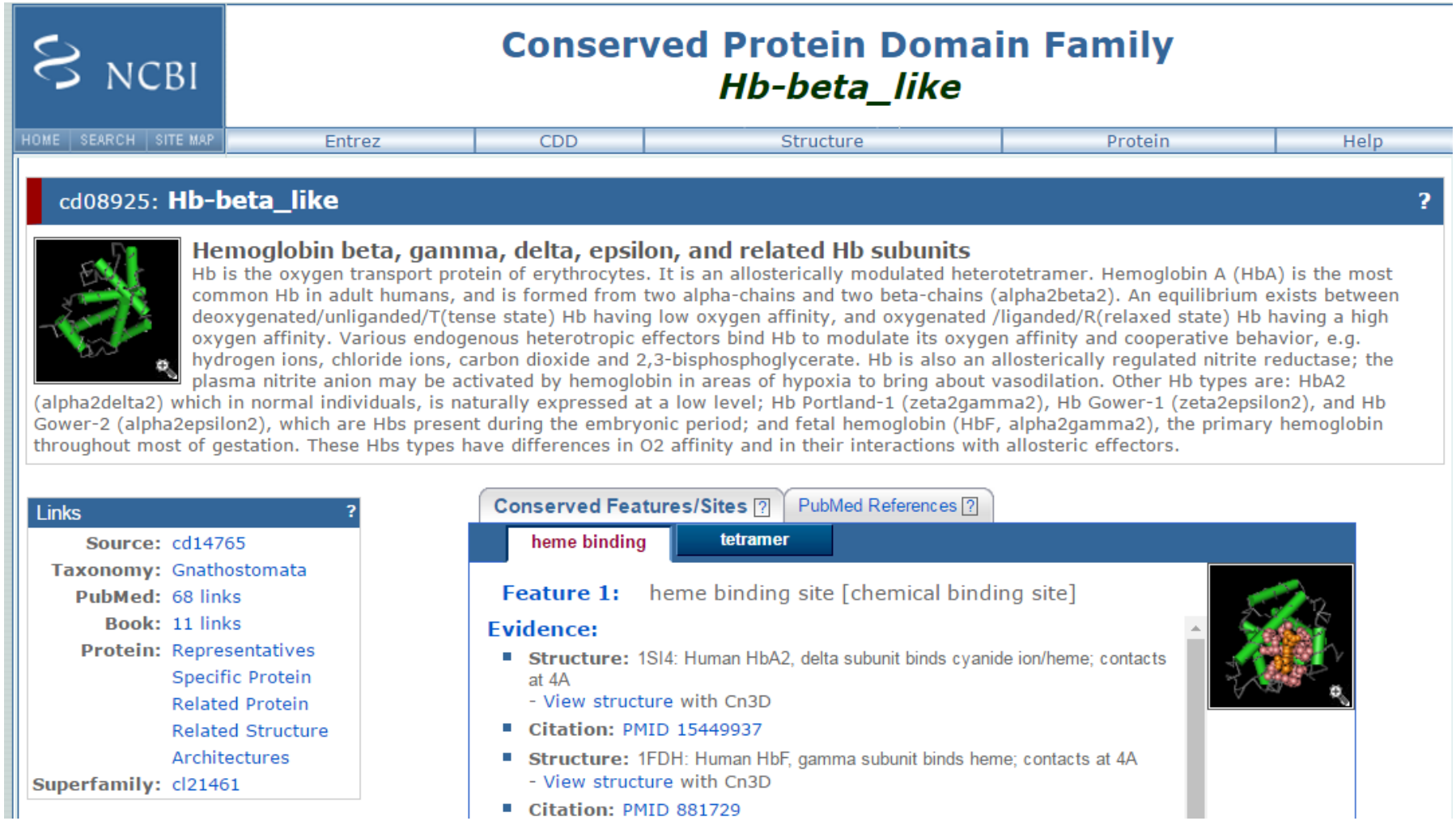
CDD (Conserved Domain Database)

- NCBI tool
- http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
- Search using sequence or text queries
- Uses Reverse position specific BLAST (RPS-BLAST for searching)
  - Query is searched against PSSM libraries
    - Many PSSMs

CDD

- Searches against profiles generated from prealigned sequences
- CDD
  - CD search is carried out using RPS-BLAST; Variant of PSI BLAST
- What is it used for?
  - Identify conserved domains in query sequence

Many ways to get to CDD; Blast, or from Gene→Protein → identify conserved Domains



Delta Blast

- Most sensitive protein-protein search tool in NCBI
- constructs PSSM using the results of CDD search
- And use that to search further sequence DB

DELTA-BLAST

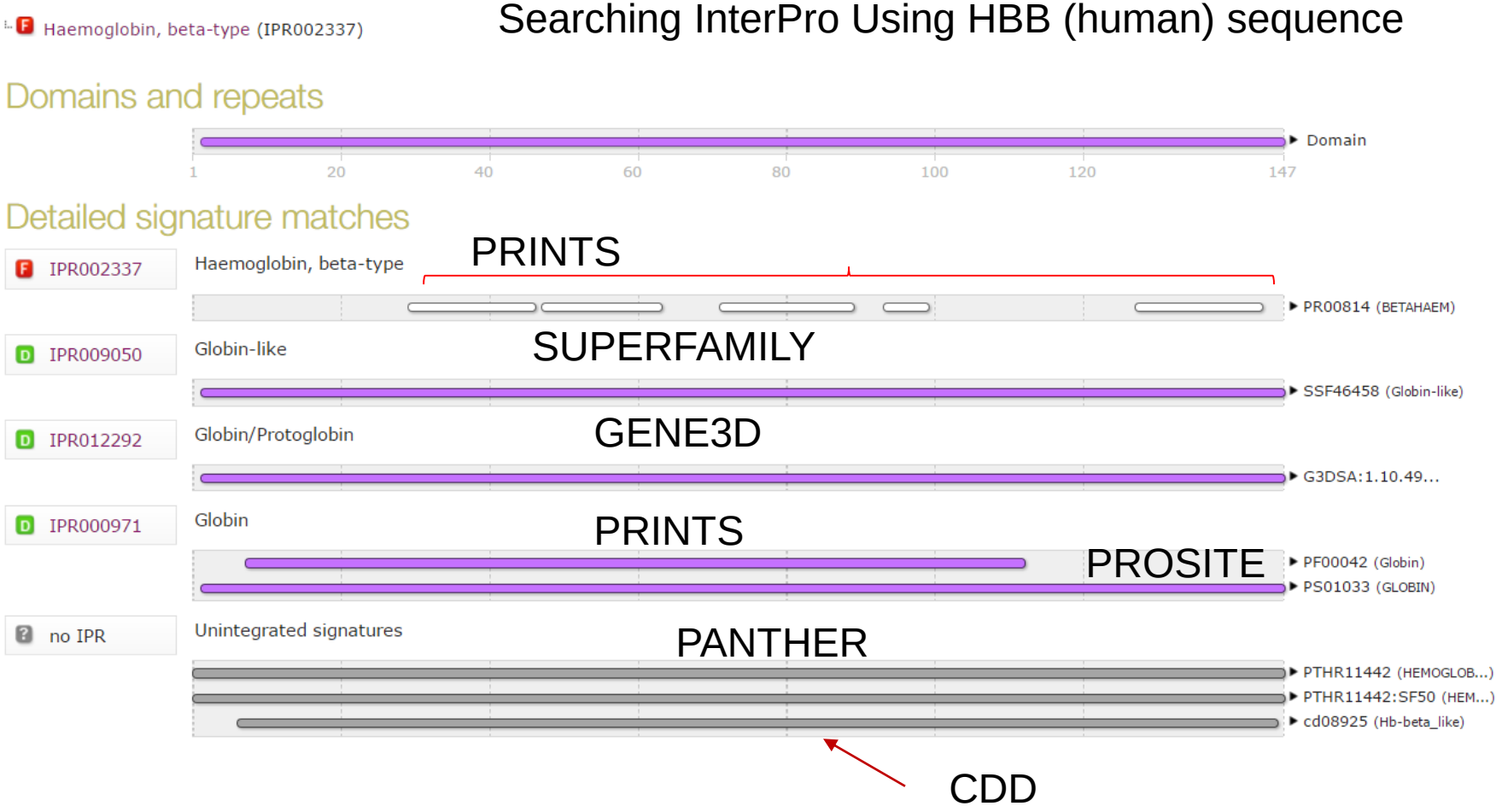
- Domain enhanced lookup time accelerated BLAST
- Most accurate tool than usual BLASTP
- How is it different
- Query is searched using pre-created PSSMs
  - Basically this first step is a RPS-BLAST
  - Use the resulting PSSM to search against protein Database

Integerated MSA Resources

- Interpro and iProClass
- These allow to explore the features using several DBs in parallel
  - PROSITE, PRINTS, ProDom, Pfam, TIGRFAMs with cross ref to BLOCKS
- www.ebi.ac.uk/interpro

Gleaned from other DBs

Searching InterPro Using HBB (human) sequence



EMBL-EBI

InterPro
Protein sequence analysis & classification

Member database information

Signature database	Version	Signatures*	Integrated signatures**
CATH-Gene3D	3.5.0	2626	1725
CDD	3.14	11273	318
HAMAP	201605.11	2087	2080
PANTHER	10.0	95118	5358
Pfam	29.0	16295	15699
PIRSF	3.01	3285	3223
PRINTS	42.0	2106	2001
ProDom	2006.1	1894	1130
PROSITE patterns	20.119	1309	1289
PROSITE profiles	20.119	1136	1108
SMART	7.1	1312	1265
SUPERFAMILY	1.75	2019	1417
TIGRFAMs	15.0	4488	4452

* Some signatures may not have matches to UniProtKB proteins.

** Not all signatures of a member database may be integrated at the time of an InterPro release

Manual vs Automated

- Pfam
  - Curated manually
  - Sean Eddy and Colleagues;
- PROSITE, BLOCAS and PRINTS
  - Manually annotated
- ProDom DOMO
  - Automated

Manual vs Automated

- Is Manual better?
  - Yes, but time consuming
- PSI-BLAST & DELTA-BLAST
  - If error occurs in first cycle, then it easily propagates

Multiple Sequence Alignment Genomic Regions

- Motivation
  - Genomes are sequenced in a phase that is faster than ever
  - Species sequence comparison is very informative for phylogenetic analysis
  - Identify DNA regions that are under
    - positive selection
      - Hence rapidly changing in a given lineage
    - Negative selection

Differences in MSA of DNA

- Few sequences with lengths 1-10s of millions
  - Protein sequence lengths are much smaller
- Far diverged comparison like Fish vs Human (islands of good conservation followed by low conservation)
- Repetitive elements in DNA (transposons and long and short interspersed nuclear elements

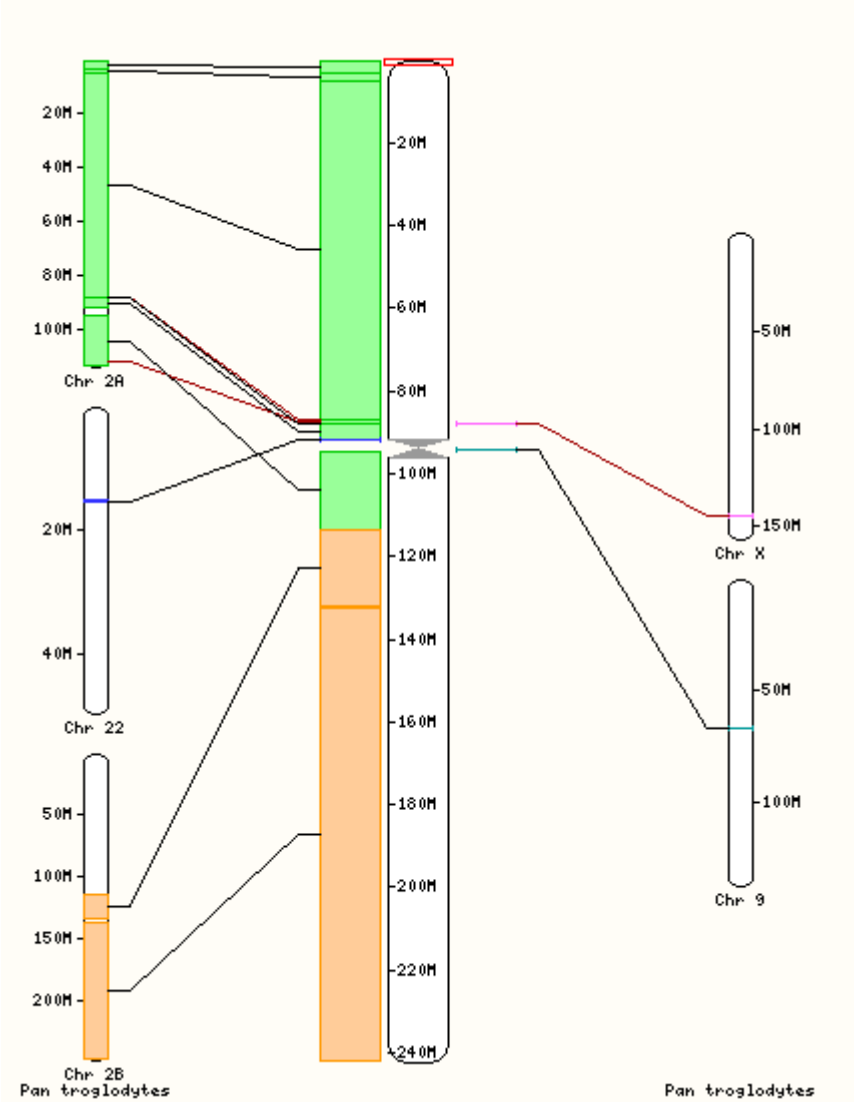
Differences in MSA of DNA

- Chromosomal duplications, delectations and translocations
  - Issues for alignment algorithms
- No reliable benchmarks for genomic alignments compared to proteins that are based on 3D structures

Synteny between Human Chromosome 2 and Chimpanzee

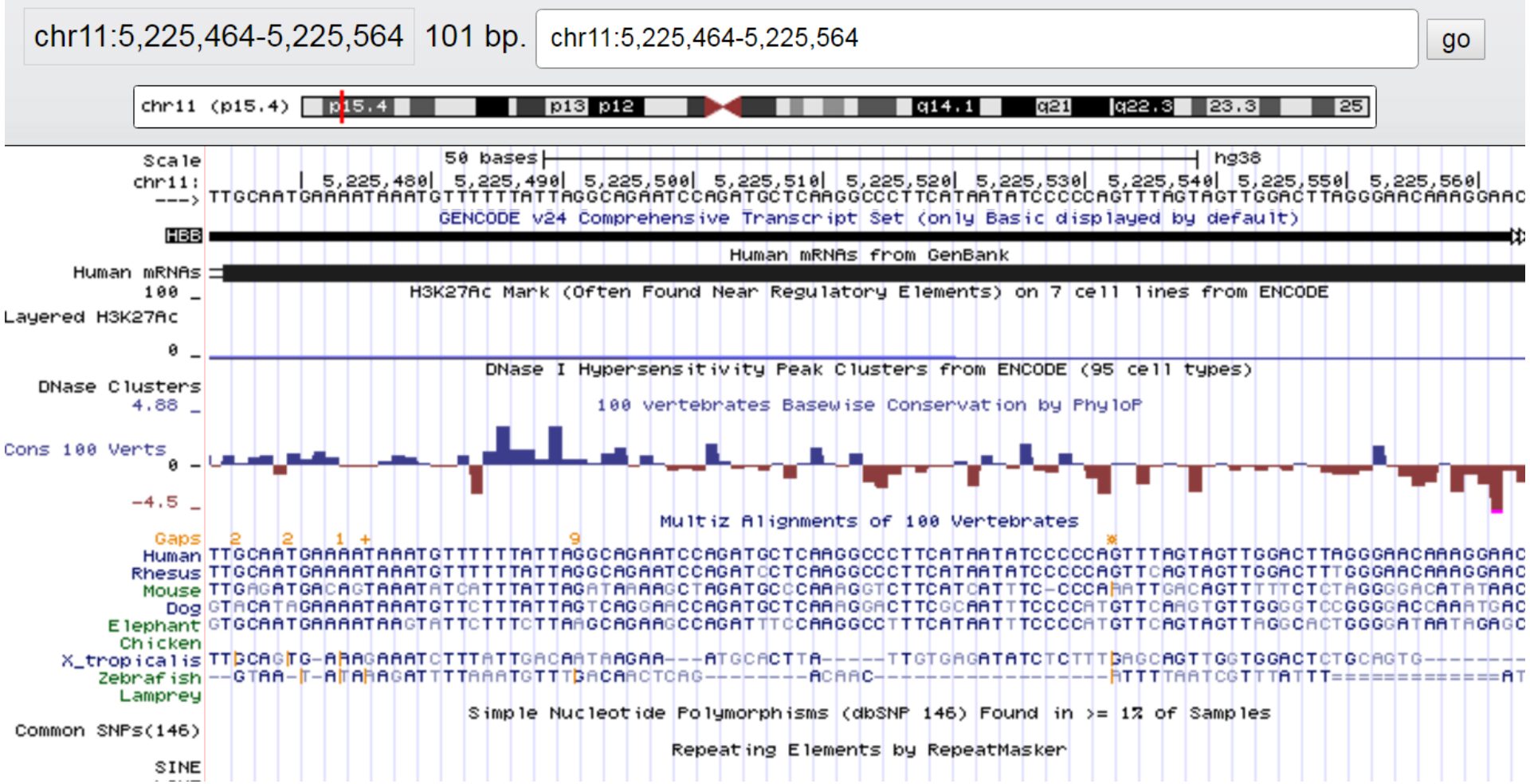
You can see that there is a significant similarity in two different chromosomes of Chmp

2A and 2B



DNA Alignments in UCSC

- Go to UCSC
- Type HBB
- Pick the first entry
- Hide unwanted things
  - Right click and hide
- Zoom only 100 residues
- You will be shown the multiple sequence alignment using Multiz



PhyloP conservation scores are shown as bars

"*sites predicted to be conserved are assigned positive scores (and shown in blue), while sites predicted to be fast-evolving are assigned negative scores (and shown in red).* "

What is shown on the page?

- phyloP
  - Conservation scores estimated at individual columns
- Pros and Cons of PhyloP and other conservation scores can be seen here

https://genome.ucsc.edu/cgi-bin/hgTables?db=hg38&hgta_group=compGeno&hgta_track=cons100way&hgta_table=phyloP100way&hgta_doSchema=describe+table+schema

Galaxy via UCSC

- Start with UCSC
- Select HBB region
- Select BED output format and send to Galaxy (coding exons only)
- In Galaxy tools menu choose “Fetch alignments” and then “Extract Pairwise MAF Blocks” to convert the genomic interval to the

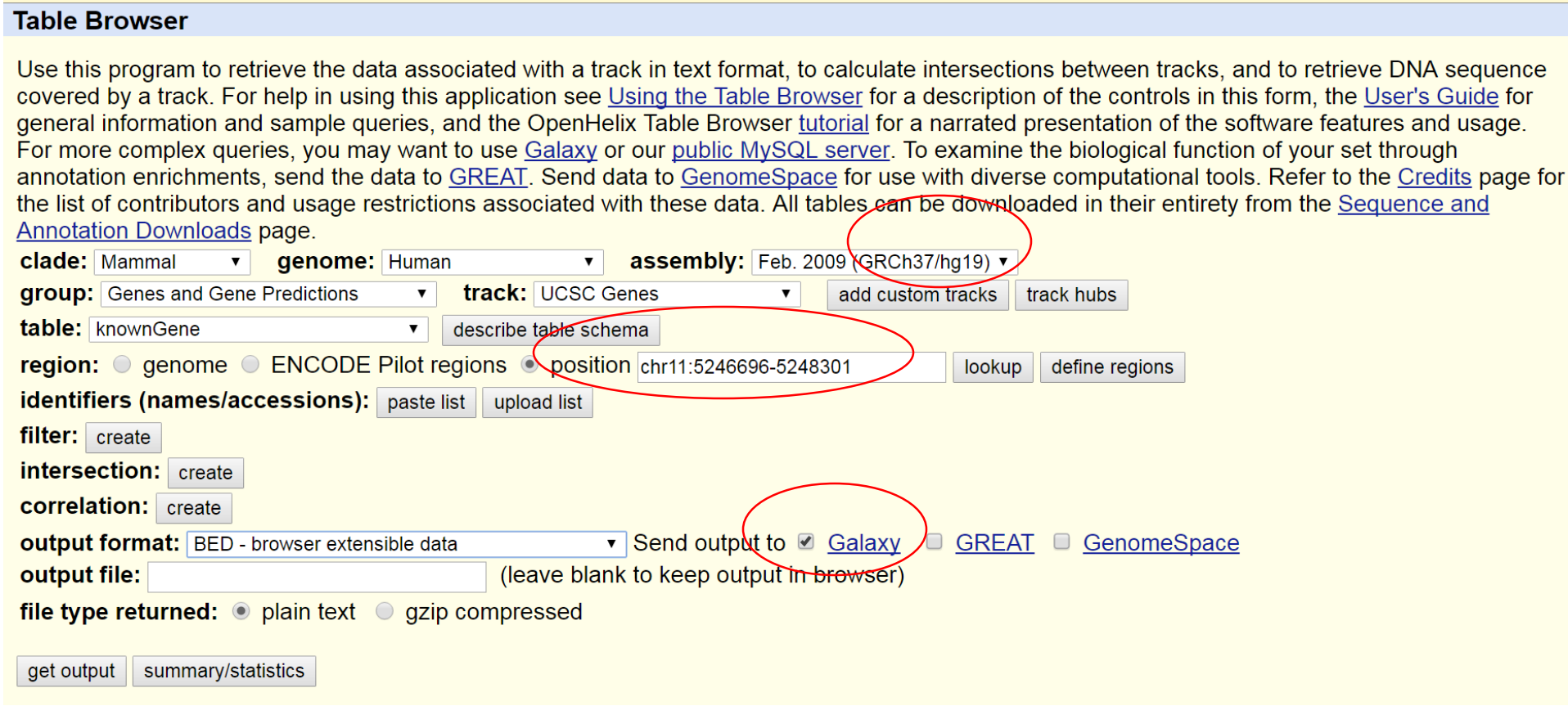
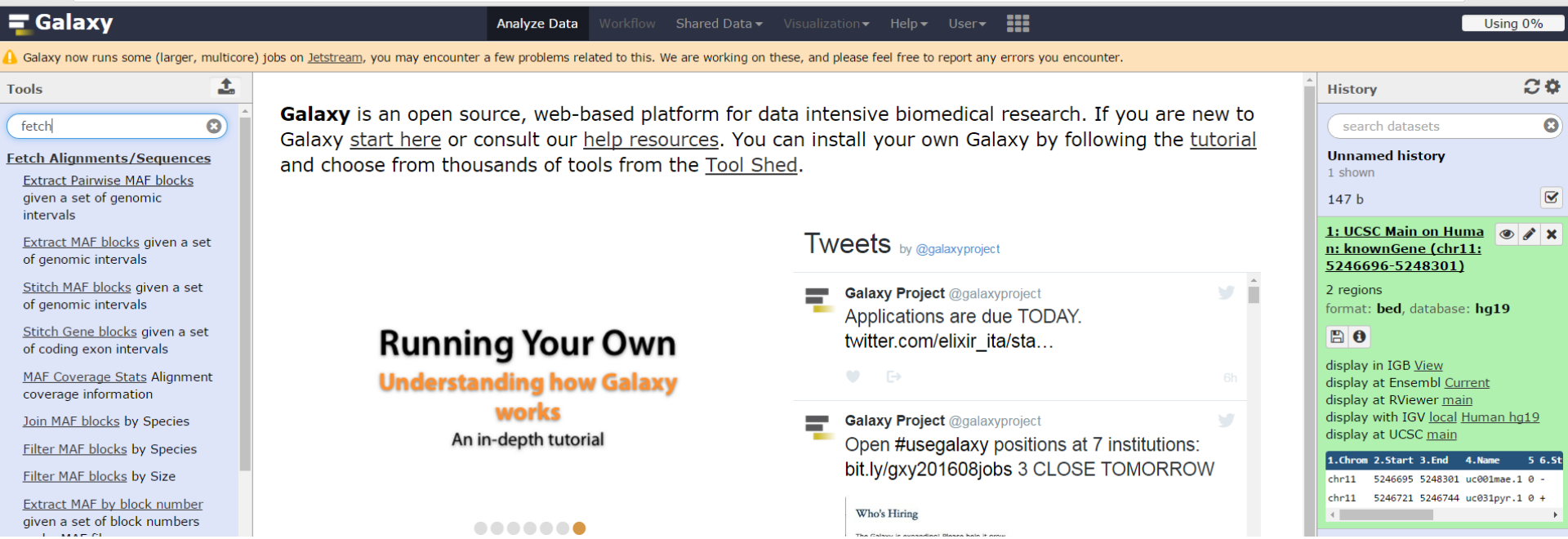
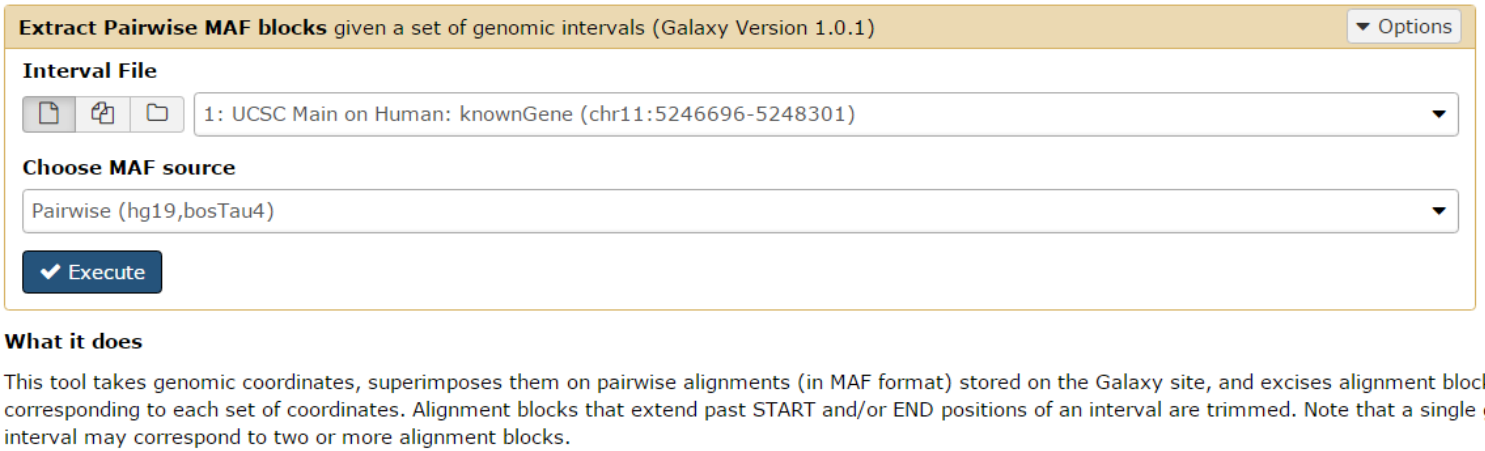


Table Browser
Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.
clade: Mammal
genome: Human
assembly: Feb. 2009 (GRCh37/hg19)
group: Genes and Gene Predictions
track: UCSC Genes
add custom tracks
track hubs
table: knownGene
describe table schema
region: genome ENCODE Pilot regions position chr11:5246696-5248301
lookup
define regions
identifiers (names/accessions):
paste list
upload list
filter: create
intersection: create
correlation: create
output format: BED - browser extensible data
Send output to Galaxy GREAT GenomeSpace
output file:
(leave blank to keep output in browser)
file type returned: plain text gzip compressed
get output
summary/statistics

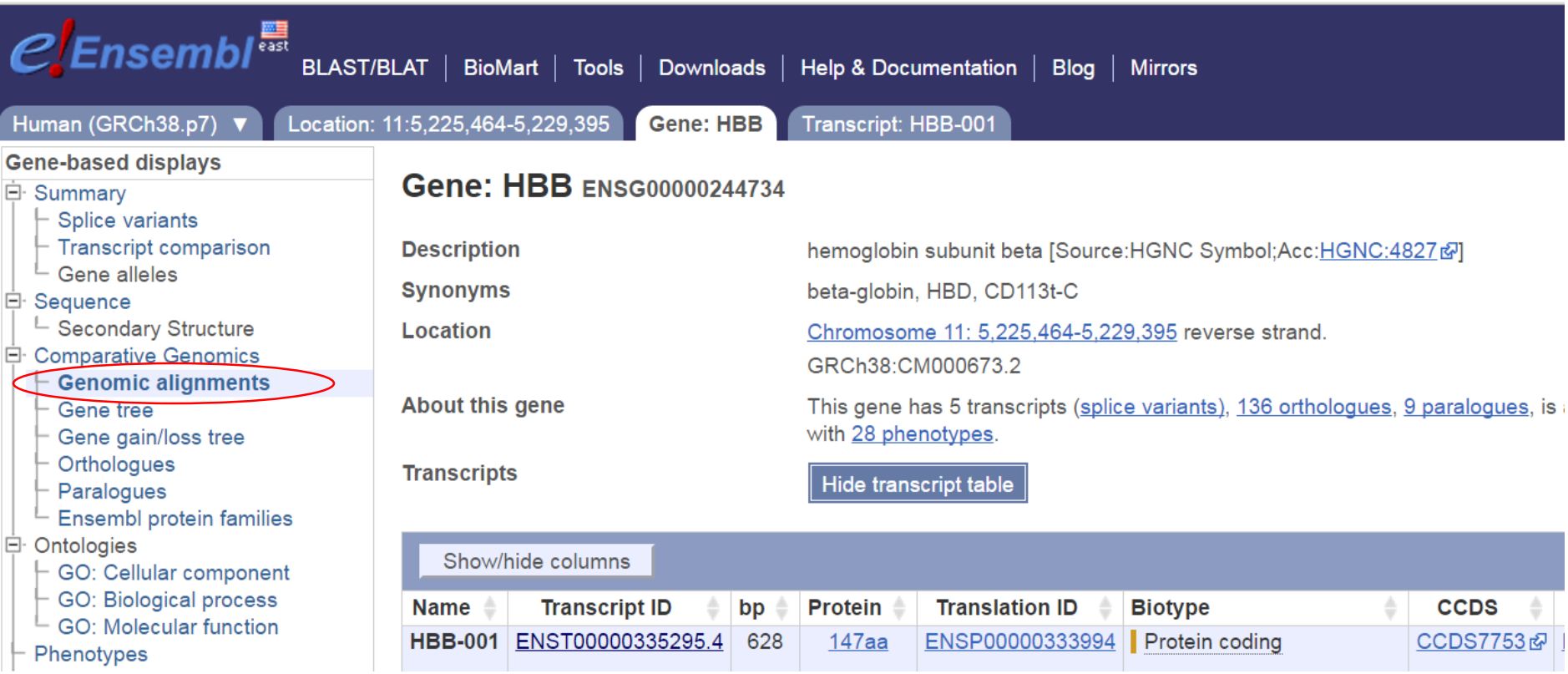


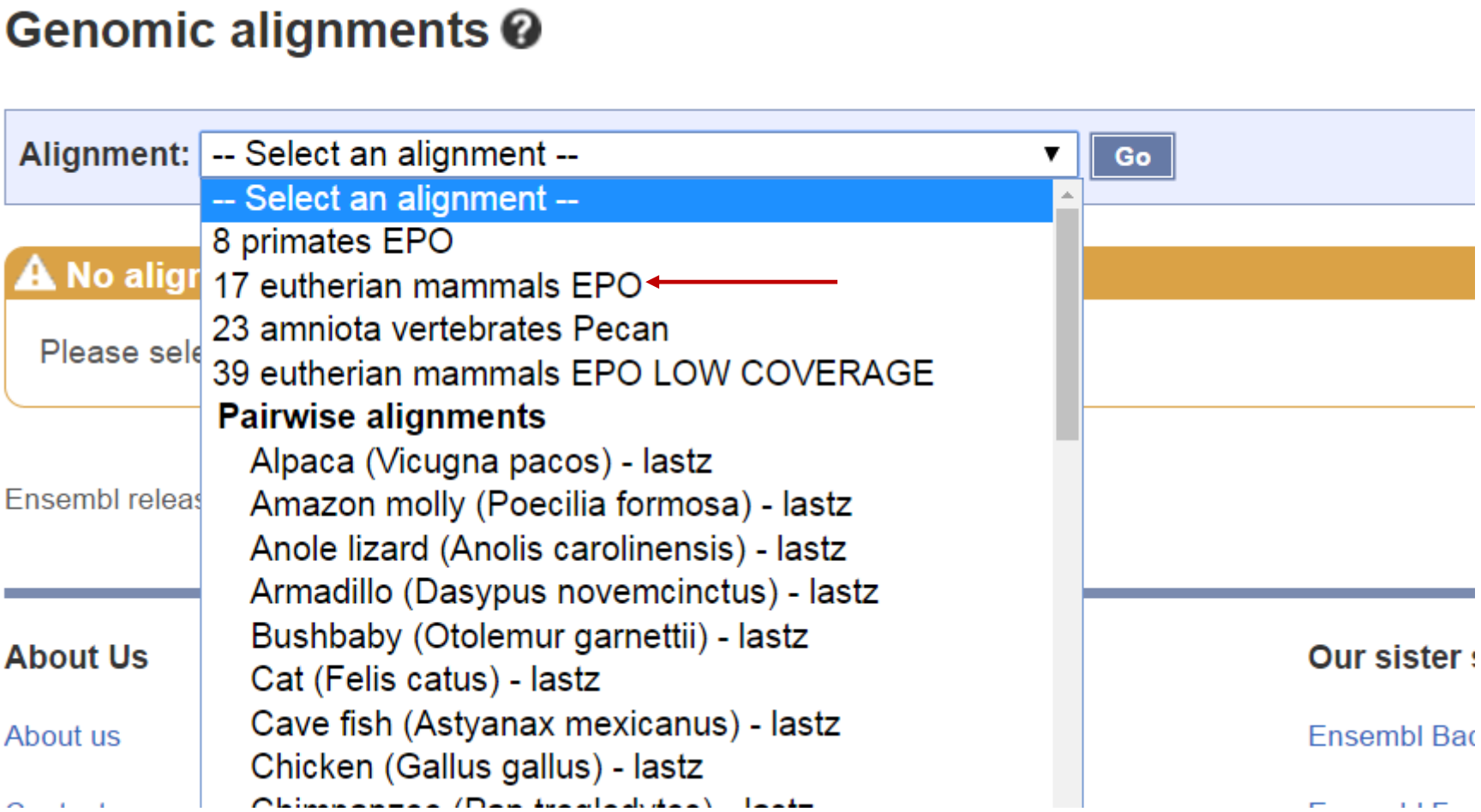
A browser session will open up with Galaxy



Choose Cow Genome for comparison

Ensembl


e!Ensembl east
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Human (GRCh38.p7) ▼
Location: 11:5,225,464-5,229,395
Gene: HBB
Transcript: HBB-001
Gene-based displays
Summary
Splice variants
Transcript comparison
Gene alleles
Sequence
Secondary Structure
Comparative Genomics
Genomic alignments
Gene tree
Gene gain/loss tree
Orthologues
Paralogues
Ensembl protein families
Ontologies
GO: Cellular component
GO: Biological process
GO: Molecular function
Phenotypes
Gene: HBB ENSG00000244734
Description
hemoglobin subunit beta [Source:HGNC Symbol;Acc:HGNC:4827]
Synonyms
beta-globin, HBD, CD113t-C
Location
Chromosome 11: 5,225,464-5,229,395 reverse strand.
GRCh38:CM000673.2
About this gene
This gene has 5 transcripts (splice variants), 136 orthologues, 9 paralogues, is
with 28 phenotypes.
Transcripts
Hide transcript table
Show/hide columns
Name
Transcript ID
bp
Protein
Translation ID
Biotype
CCDS
HBB-001
ENST00000335295.4
628
147aa
ENSP00000333994
Protein coding
CCDS7753


Genomic alignments
Alignment:
-- Select an alignment --
Go
-- Select an alignment --
8 primates EPO
17 eutherian mammals EPO
23 amniota vertebrates Pecan
39 eutherian mammals EPO LOW COVERAGE
Pairwise alignments
Alpaca (Vicugna pacos) - lastz
Amazon molly (Poecilia formosa) - lastz
Anole lizard (Anolis carolinensis) - lastz
Armadillo (Dasypus novemcinctus) - lastz
Bushbaby (Otolemur garnettii) - lastz
Cat (Felis catus) - lastz
Cave fish (Astyanax mexicanus) - lastz
Chicken (Gallus gallus) - lastz
No align
Please sele
Ensembl releas
About Us
About us
Our sister
Ensembl Bac

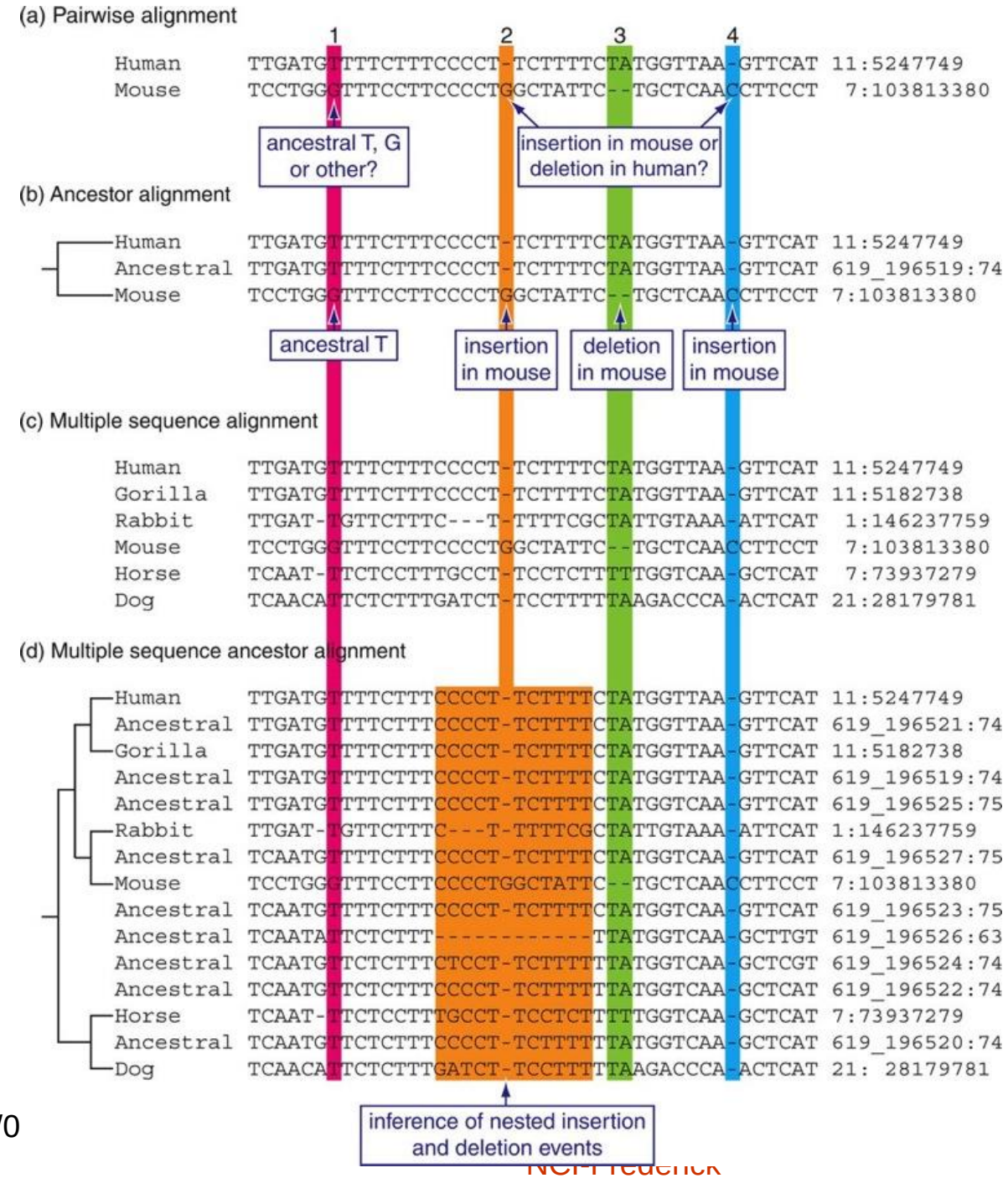
Conserved regions are marked as blue shade

```
Human       ACTTAT--TTGCCTGGTATGCCTGGGCTTTTGATGGTCTTAGTATAGCTTGCAGCCTTGTCCCTGCAGGGTATTATGGGTAATAGAAAGAAAAGTCTGCGTTACACTCTAGTCACACTAA
Chimpanzee  ACTTATTTTTGCCTGGTATGCCTGGGCTTTTGATGGTCTTAGTATAGCTTGCACCCTTGTCCCTGCAGGATATTATGGGTCATAGAAAGAAAAGTCTGCATTACACTCTAGTCACACTAA
Gorilla     ACTTATTTTTGCCTGGTATGCCTGGGCTTTTGATGGTCTTAGTATAGCTTGCAGCCTTGTCCCTGCAGGATATTATGGGTAATAGAAAGAAAAGTCTGCATTACACTCTAGTCACACTAA

Human       GTAACTACCATTGGAAAAGCAACCCCTGCCTTGAAGCCAGGATGATGGTATCTGCAGCAGTTGCCAACACAAGAGAAGGATCCATAGTTCATCATTTAAAAAAGAAAACAAAATAGAAAA
Chimpanzee  GTAACTACCATTGGAAAAGCAACCCCTGCCTTGAAGCCAGGATGATGGTATCTGCAGCAGTTGCCAACACAAGAGAAGGATCCATAGTTCATCATTTAAAAAAGAAAACAAAATAGAAAA
Gorilla     GTAACTACCATTGGAAAAGCAACCCCTGCCTTGAAGCCAGGATGATGGTATCTGCAGCAGTTGCCAACACAAGAGAAGGATCCATAGTTCATCATTTAAAAAAGAAAACAAAATAGAAAA

Human       AGGAAAACTATTTCTGAGCATAAGAAGTTGTAGGGTAAGTCTTTAAGAAGGTGACAATTTCTGCCAATCAGGATTTCAAAGCTCTTGCTTTGACAATTTTGGTCTTTCAGAATACTATAA
Chimpanzee  AGGAAAACTATTTCTGAGCATAAGAAGTTGTAGGGTAAGTCTTTAAGAAGGTGACAATTTCTGCCAATCAGGATTTCAAAGCTGTTGCTTTGACAATTTTGGTCTTTCAGAATACTATAA
Gorilla     AGGAAAACTATTTCTGAGCATAAGAAGTTGTAGGGTAAGTCTTTAAGAAGGTGACAATTTCTGCCAATCAGGATTTCAAAGCTGTTGCTTTGACAATTTTGGTCTTTCAGAATACTATAA

Human       ATATAACCTATATTATAATTTCATAAAGTCTGTGCATTTTCTTTGACCCAGGATATTTGCAAAAGACATATTCAAACTTCCGCAGAACACTTTATTTCACATATACATGCCTCTTATATC
Chimpanzee  ATATAACCTATATTATAATTTCATAAAGTCTGTGGATTTTCTTTGACCCGGGATATTTGCAAAAGACATATTCAAACTTCCGCAGAACACTTTATTTCACATATACATGCCTCTTATATC
Gorilla     ATATAACCTATATTATAATTTCATAAAGTCTGTGGATTTTCTTTGACCCGGGATATTTGCAAAAGACATATTCAAACTTCCACAGAACACTTTATTTCACATATACATGCCTCTTATACC

Human       AGGGATGTGAAACAGGGTCTTGAAAACTGTCTAAATCTAAAACAATGCTAATGCAGGTTTAAATTTAATAAAATAAAATCCAAAATCTAACAGCCAAGTCAAATCTGCATGTTTTAACAT
Chimpanzee  AGGGATGTGAAACAGGGTCTTGAAAACTGTCTAAATCTAAAACAATGCTAATGCAGGTTTAAATTTGATAAAATAAAATCCAAAATCTAAAAGCCAAGTCAAATCTGTATGTTTTAACAT
Gorilla     AGGGATGTCAAACAGGGTCTTGAAAACTGCCTAAATCTAAAGCAATGCTAATGCAGGTTTAAATTTAATAAAATAAAATCCAAAATCTAAAAGTCAAGTCAAATCTGTATGTTTTAACAT

Human       TTAAAATATTTTAAAGACGTCTTTTCCCAGGATTCAACATGTGAAATCTTTTCTCAGGGATACACGTGTGCCTAGATCCTCATTGCTTTAGTTTTTTACAGAGGAATGAATATAAAAAGA
Chimpanzee  TTAAAATATTTTAAAGACGTCTTTTCCCAGGATTCAACACGTGAAATCTTTTCTCAGGGATACACGTGTGCCTAGATCCTCATTGCTTTAGTTTTTTACAGAGGAATGAGCATACAAAGA
Gorilla     TTAAAATATTTTAAAGACGTCTTTTCCCAGGATTCAACATGTGAAATCTTTTCTCAGGGATACAAGTGTGCCTAGATCCTCGTTGCTTTAGTTTTTTACAGAGGAATGAATATAAAAAGA

Human       AAATACTTAAATTTTATCCCTCTTACCTCTATAATCATACATAGGCATAATTTTTTAACCTAGGCTCCAGATAGCCATAGAAGAACCAAACACTTTCTGCGTGTGTGAGAATAATCAGAG
Chimpanzee  AAATACTTAAATTTTATCTCTCTTACCTCTATAATCATACACAGGCATAATTTTTTAACCTAGGCTCCAGATAGCCATAGAAGAACCAAACACTTTCTGCGTGTGTGAGAATAATCAGAG
Gorilla     AAATACTTAAATTTTCTCTCTCTTACGTCTATAATCGTACATAGGCATAATTTTTTAACCTAGGCTCCAGATAGCCATACAAGAACCAAACACTTTCTGCGTGTGTGAGAATAATCAGAG

Human       TGAGATTTTTTCACAAGTACCTGATGAGGGTTGAGACAGGTAGAAAAAGTGAGAGATCTCTATTTATTTAGCAATAATAGAGAAAGCATTTAAGAGAATAAAGCAATGGAAATAAGAAAT
Chimpanzee  TGAGATTTTTTCACAAGTACCTGATGAGGGTTGAGACAGGTAGAAAAAGTGAGAGATCTCTATTTATTTAGCAATAATAGAGAAAGTATTTAAGAGAATAAAGCAATGGAAATAAGAAAT
Gorilla     TGAGATTTTTTCACAAGTACCTGATGAGGGTTGAGACAGGTAGAAAAAGTGAGAGATCTTTATTTATTTAGCAATAATAGAGAAAGTATTTAAGAGAATACAGCAATGGAAATAAGAAAT
```

Red Core exons identified by Ensembl

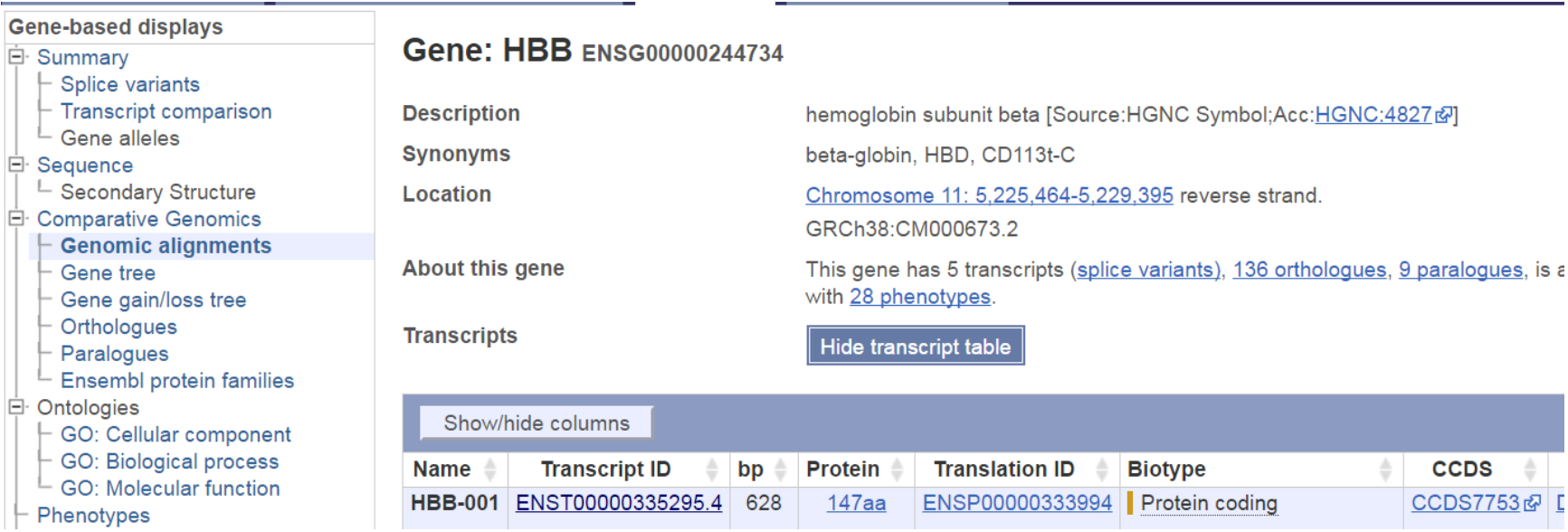
EPO Pipeline

- Enredo, Pecan, Ortheus is a three step pipeline for whole-genome multiple alignments
- LastZ and BlastZ are used to align genome sequences at the DNA level.
- PECAN is the MSA algorithm
  - Phylogenetic tree based alignment
  - Heuristic

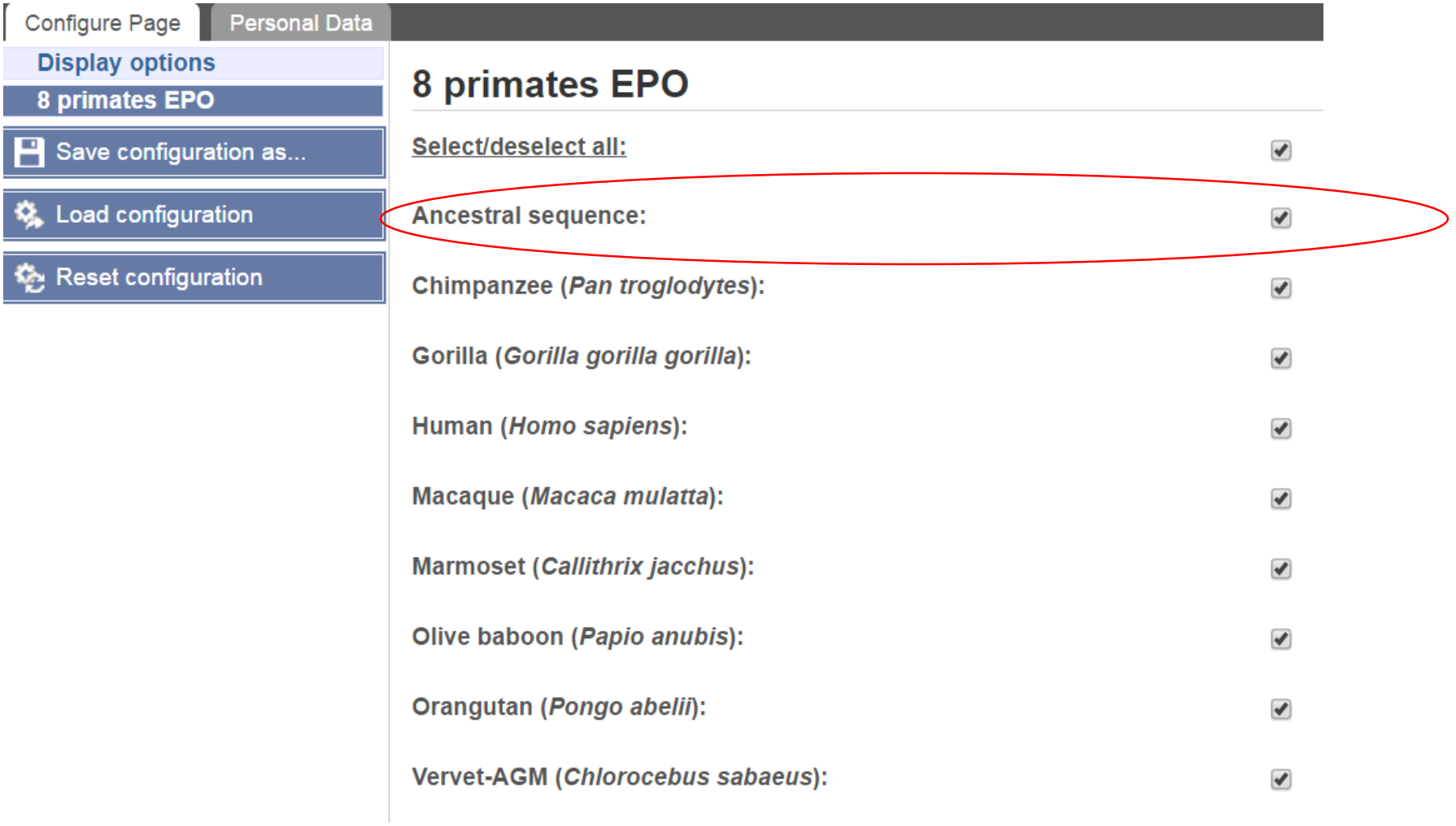


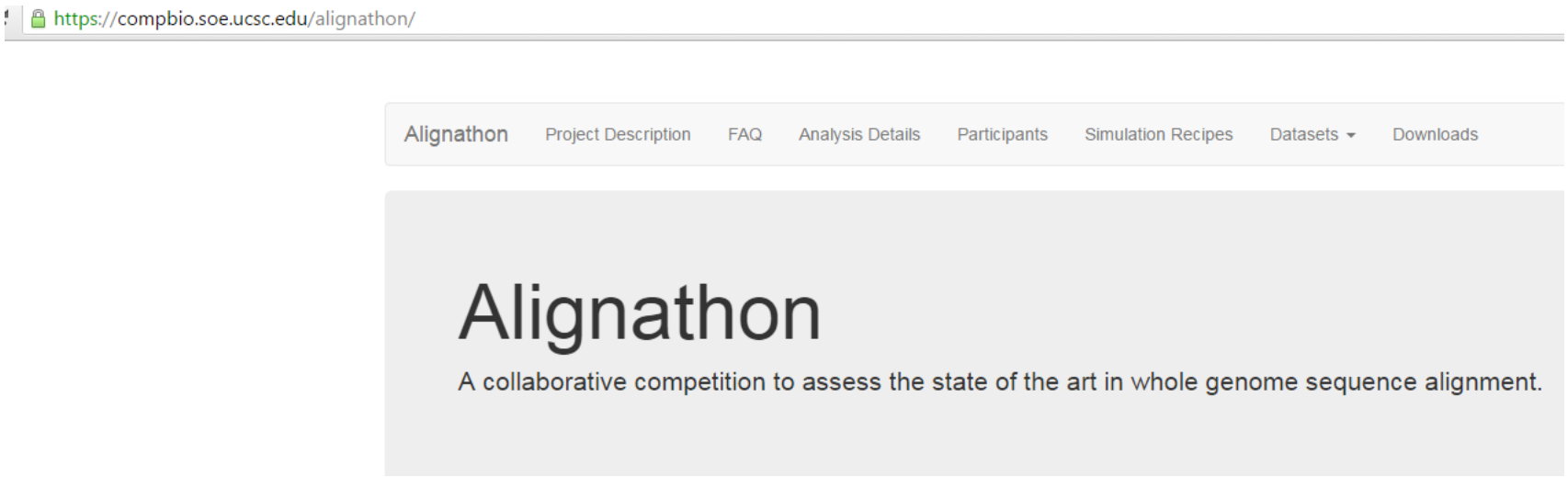
Ensembl EPO is a phylogeny based prediction pipeline

How to view Ancestral sequence in Ensembl?



Select Configure


Configure Page
Personal Data
Display options
8 primates EPO
Save configuration as...
Load configuration
Reset configuration
8 primates EPO
Select/deselect all:
Ancestral sequence:
Chimpanzee (Pan troglodytes):
Gorilla (Gorilla gorilla gorilla):
Human (Homo sapiens):
Macaque (Macaca mulatta):
Marmoset (Callithrix jacchus):
Olive baboon (Papio anubis):
Orangutan (Pongo abelii):
Vervet-AGM (Chlorocebus sabaeus):



https://compbio.soe.ucsc.edu/alignathon/index.html

After the talk

- Problems/Computer Lab
  - 6-1, 6-2, 6-3, 6-4 and 6-8

Thanks

ravichandran@hood.edu