

Multiple Sequence Alignment

BIFX-550

S. Ravichandran, 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/w
rp_protsci04.pdf](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 prediction
 - 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 residues (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	NP_002037.2	1	MGKVKVGVNGFGRIGRLVTRAFFNSGKVDIVAINDPFIDLNLYMVMYMFQYD
XP_003435697.1	C.lupus	XP_508955.1	1	MGKVKVGVNGFGRIGRLVTRAFFNSGKVDIVAINDPFIDLNLYMVMYMFQYD
XP_003434435.1	C.lupus	XP_001105471.1	1	MGKVKVGVNGFGRIGRLVTRAFFNSGKVDIVAINDPFIDLNLYMVMYMFQYD
NP_001029206.1	B.taurus	NP_001003142.1	1	--MVKVGVNGFGRIGRLVTRAFFNSGKVDIVAINDPFIDLNLYMVMYMFQYD
NP_032110.1	M.musculus	NP_001003142.1	1	--MVKVGVNGFGRIGRLVTRAFFNSGKVDIVAINDPFIDLNLYMVMYMFQYD
XP_001476757.1	M.musculus	XP_003435697.1	1	--MVKVGVNGFGRIGRLVTRAFFNSGKVDIVAINDPFIDLNLYMVMYMFQYD
NP_058704.1	R.norvegicus	XP_003434435.1	1	--MVKVGVNGFGRIGRLVTRAFFNSGKVDIVTINDPFIDLNLYMVMYMFQYD
NP_989636.1	G.gallus	NP_001029206.1	1	--MVKVGVNGFGRIGRLVTRAFFNSGKVDIVAINDPFIDLHYMVMYMFQYD
NP_001108586.1	D.rerio	NP_032110.1	1	--MVKVGVNGFGRIGRLVTRAFFNSGKVEIVAINDPFIDLNLYMVMYMFQYD
NP_001259584.1	D.melanogaster	NP_001476757.1	1	--MVKVGVNGFGRIGRLVTRAFFNSGKVEIVAINDPFIDLNLYMVMYMFQYD
NP_525108.2	D.melanogaster	NP_318655.2	1	--MVKVGVNGFGRIGRLVTRAFFNSCDKVDIVAINDPFIDLNLYMVMYMFQYD
XP_318655.2	A.gambiae	NP_058704.1	1	--MVKVGVNGFGRIGRLVTRAFFNSCDKVDIVAINDPFIDLNLYMVMYMFQYD
NP_496237.1	C.elegans	NP_989636.1	1	--MVKVGVNGFGRIGRLVTRAFFNSGKVQVVAINDPFIDLNLYMVMYMFQYD
NP_496192.1	C.elegans	NP_001108586.1	1	--MVKVGINGFGRIGRLVTRAFFLTKKVEIVAINDPFIDLDYMVMYMFQYD
NP_508534.3	C.elegans	NP_001259584.1	1	--MSKIGINGFGRIGRLVLAIDKG-ANVVAVNDPFIDVNLYMVLFKFD
NP_508535.1	C.elegans	NP_525108.2	1	--MSKIGINGFGRIGRLVLAIDKG-ASVVAVNDPFIDVNLYMVLFKFD
NP_012483.3	S.cerevisiae	NP_318655.2	1	--MSKIGINGFGRIGRLVLAIAITKG-ASVVAINDPFIGVDYMVLFKYD
NP_011708.3	S.cerevisiae	NP_496237.1	1	MSKANVGINGFGRIGRLVLAVERKDTVQVVAVNDPFITIDYMVLFKYD
NP_012542.1	S.cerevisiae	NP_496192.1	1	MSKANVGINGFGRIGRLVLAVERKDTVQVVAVNDPFITIDYMVLFKYD
XP_456022.1	K.lactis	NP_496192.1	1	MSKANVGINGFGRIGRLVLAVERKDSVNVVAVNDPFISIDYMVLFKYD
NP_596154.1	S.pombe	NP_508534.3	1	MTKPSVGINGFGRIGRLVLAVERKDSVNVVAVNDPFISIDYMVLFKYD
NP_595236.1	S.pombe	NP_508535.1	1	MPKPSVGINGFGRIGRLVLAVERKDSVNVVAVNDPFISIDYMVLFKYD
XP_003717853.1	M.oryzae	NP_012483.3	1	--MIRIAINGFGRIGRLVLRALQRKDIEVVAVNDPFISNDYAAYMVKYD
XP_956977.1	N.crassa	NP_011708.3	1	--MVRVAINGFGRIGRLVMRIALSRPNVEVVALNDPFITNDYAAYMFKYD
NP_001060897.1	O.sativa	NP_011708.3	1	--MVRVAINGFGRIGRLVMRIALQRKNVEVVALNDPFISNDYSAYMFKYD
NP_001004949.1	X.tropicalis	NP_012542.1	1	--MVKAINGFGRIGRLVLRALQRKALEVVAVNDPFISVDYAAAYMFKYD
		NP_456022.1	1	MAIPKVGINGFGRIGRIVRNALVAKTIQVVAINDPFIDLEYMAYMFKYD
		NP_596154.1	1	MAIPKVGINGFGRIGRIVRNAILTGKIQVVAVNDPFIDLDYMAYMFKYD
		NP_595236.1	1	--MVKGCGINGFGRIGRIVRNAIEHPDCEIVAVNDPFIEPKYAAYMLEYD
		XP_003717853.1	1	-MVKVGINGFGRIGRIVRNAIEHDDIHIVAVNDPFIEPKYAAYMLRYD
		XP_956977.1	1	MGKIKIGINGFGRIGRLVARVALQSEDVELVAVNDPFITTDYMTYMFKYD
		NP_001060897.1	1	-----MAYMFKYD
		NP_001004949.1	1	8

Not So Easy Case

NP_005203.2	50	YVPNSYPYYGTNLYQRRPAIA-INNPyVPrTYYANPAV-VRPQAQIPQRQ	97
XP_001143538.1	50	YVPNSYPYYGTNLYQRRPAIA-INNPyVPrTYYANPAV-VRPQAQIPQRQ	97
XP_001095721.1	50	YVPNSYPYYGTNLYQHRPAIA-INNPyVPrTYYANPAV-VRPQAQIPQRQ	97
XP_005628327.1	51	YVLNSFSHYEPNYYPHRPAEP-INHQYVPPFYAKPAVAVRTHAQIPQWQ	99
NP_776719.1	51	YVLSRYPSYGLNYYQQKPVAL-INNQFLPYPPYYAKPAA-VRSPAQILQWQ	98
NP_031812.2	51	SVLN-FNQYEPNYYHYRPSLPATASPYMYYPLVVRLLL-LRSPAPISKWQ	98
NP_113750.1	50	SVLN-RNHYEPIYYHYRTSVP--VSPYAYFPVGLKLLL-LRSPAQILKWQ	95
NP_005203.2	98	YLPN-----SHPPTVVRRPNLHPSFIAIPPKIQDKIIIPPTINTIAT	139
XP_001143538.1	98	YLPN-----SHPPTVVHRPNLHPSFIAIPPKIQDKIIIPPTINTIAT	139
XP_001095721.1	98	YLPN-----SHLPTVVRRPNLHPSFIAIPPKIQDKIIIPPTINTITT	139
XP_005628327.1	100	VLPN-----AYPPTMMHRPQLHPSFIAIPPKIQDKTSIPTINTIAT	141
NP_776719.1	99	VLSNTVPAKSCQAQPTTMARHPHPLSFMAIPPKKNQDKTEIPTINTIAS	148
NP_031812.2	99	SMPN-----FPQSAGVPYAIIPNPSFLAMPTNENQDNTAIPTIIDPITP	140
NP_113750.1	96	PMPN-----FPQPVGVPHPPIPNNPSFLAIPTNEKHDNTAIPASNTIAP	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.

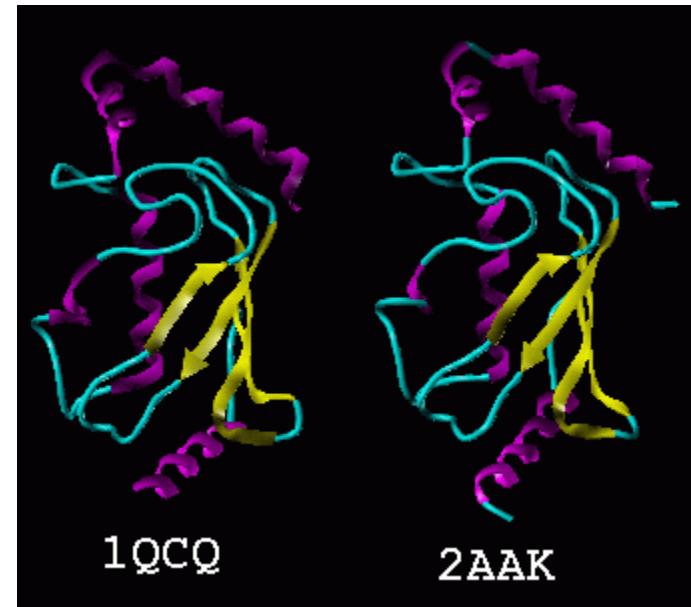
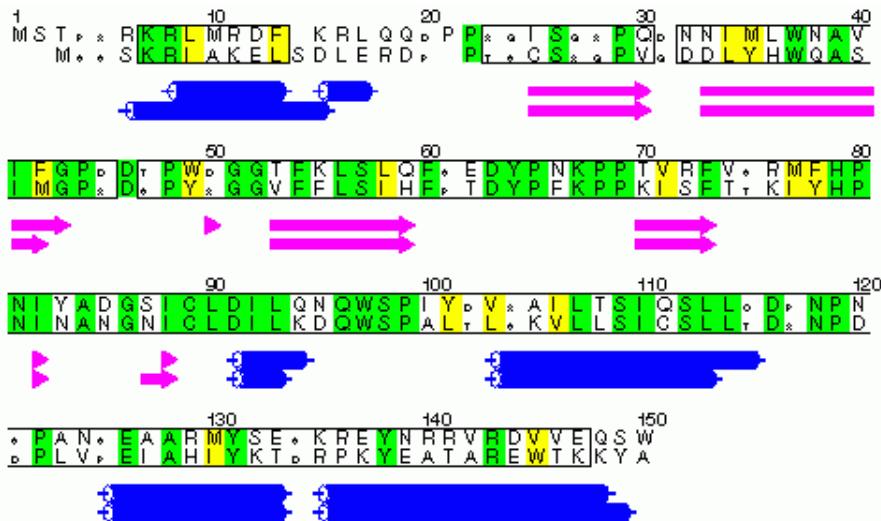
Comparing Homologous enzymes

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

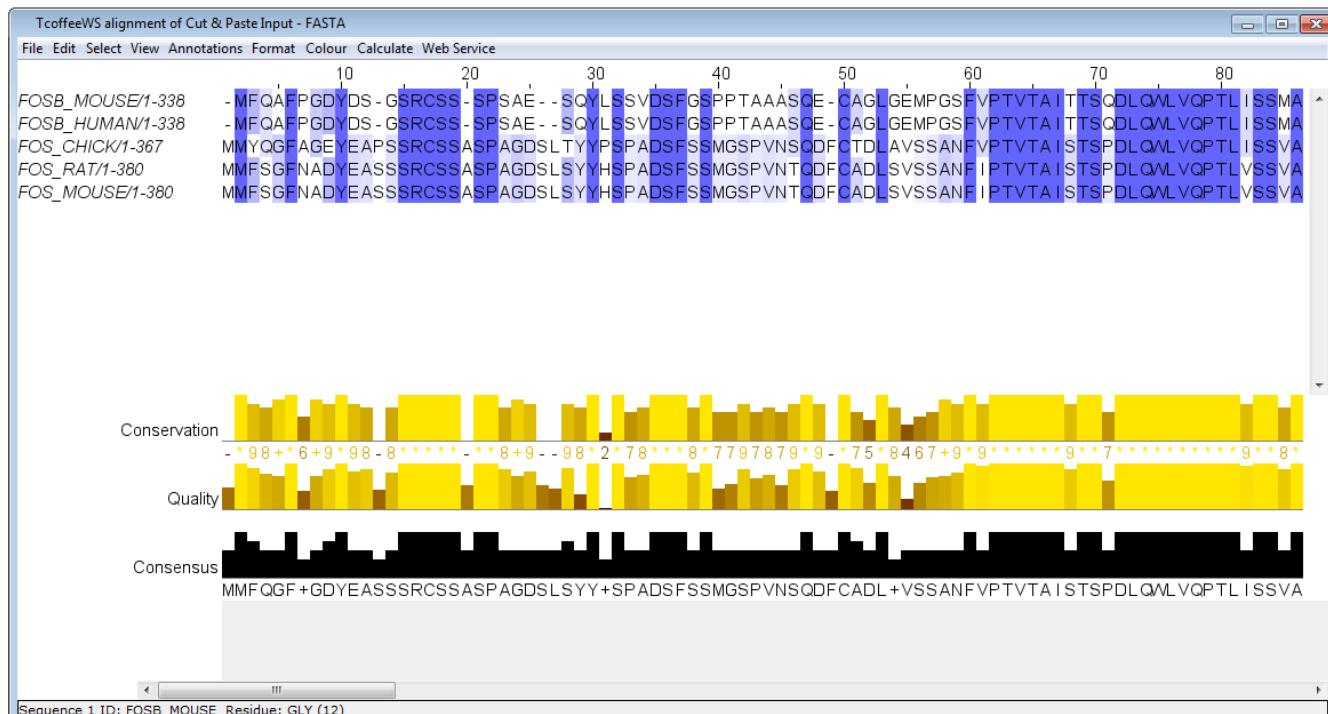
Where do we use MSA?

- 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

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

- Jalview
- Demo
 - Sequences from next page



```
>FOSB_MOUSE Protein fosB
MFQAFPGDYDSSGRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVQPTLISSMAQSQQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
GGPSTTTSGPVSARPARP RRPREETLTPEEEEKRRVRERNKLAACKRNRRRELT
DRLQAETDQLEEEKA ELESEIAELQKEKERLEFVLVAHKPGCKI PYEEGPGPGPLAEVRD
LPGSTS AKE DGF GWLLPPP PPLPFQSS RDAPPN LTASLF THSEV QVL GDP FPV VPSY
TSSFV LTCPEV SAFAGA QRTSG SEQPS DPLNSP SLLAL
```

```
>FOSB_HUMAN Protein fosB
MFQAFPGDYDSSGRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVQPTLISSMAQSQQPLASQPPVVDPYDMPGTSYSTPGMSGYSSGGASGS
GGPSTSGTTS GPGPARPARP RRPREETLTPEEEEKRRVRERNKLAACKRNRRRELT
DRLQAETDQLEEEKA ELESEIAELQKEKERLEFVLVAHKPGCKI PYEEGPGPGPLAEVRD
LPGSAPAKEDGF SWL LPPP PPLPFQTSQD APPN LTASLF THSEV QVL GDP FPV VPSY
TSSFV LTCPEV SAFAGA QRTSG SDQPS DPLNSP SLLAL
```

```
>FOS_CHICK Proto-oncogene protein c-fos
MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANF
VPTVTAISTSPDLQWLVQPTLISSVAPSQNRGH PYGPAPAPPAAYSRAV LKAPGGRGQ
SIGRRGKVEQLSPEEEKRRIRRERNKMAAKCRNRRRELDTLQAETDQLEEEKSALQA
EIANLLKEKEKLEFILA AHRPACKMPEELRFSEELAAATALD LGAPSP AAAEEAFA LPLM
TEAPP AVPPKEPSGSGLELKAE PFDELLFSAGPREASRSVPDM DLPGASSFYASDWEPLG
AGSGGELEPLCTPVVTCTPC STYTSTFVFTYPEADAFPS CAA HRKGSSSNEPSSDLS
SPTLLAL
```

```
>FOS_RAT Proto-oncogene protein c-fos
MMFSGFNADYEASSSRCSSASPAGDSLYYHSPADSFSSMGSPVNTQDFCADLSVSSANF
IPTVTAISTSPDLQWLVQPTLVSSVAPSQTRAPH PYGLPTGSTGAYARAGVVKTMSGGRA
QSIGRRGKVEQLSPEEEKRRIRRERNKMAAKCRNRRRELDTLQAETDQLEDEKSALQ
TEIANLLKEKEKLEFILA AHRPACKIPNDLGFP EEMS VTS LDTGGLPEA TPESEE AFT
LPLLNDPEPKPSLEPVKNISNMELKAEPFDDFLFPASSRPSGSETARSVPDV DLSGSFYA
ADWEPLHSSSLGMGPVTELEPLCTPVVTCTPSCTTYSFVFTYPEADSFPS CAA HRK
GSSSNEPSSDLSPTLLAL
```

```
>FOS_MOUSE Proto-oncogene protein c-fos
MMFSGFNADYEASSSRCSSASPAGDSLYYHSPADSFSSMGSPVNTQDFCADLSVSSANF
IPTVTAISTSPDLQWLVQPTLVSSVAPSQTRAPH PYGLPTQSAGAYARAGMVKTMSGGRA
QSIGRRGKVEQLSPEEEKRRIRRERNKMAAKCRNRRRELDTLQAETDQLEDEKSALQ
TEIANLLKEKEKLEFILA AHRPACKIPDDLGFP EEMS VAS LDTGGLPEA STPESEE AFT
LPLLNDPEPKPSLEPVKSISNVELKAEPFDDFLFPASSRPSGSETSRSPDV DLSGSFYA
ADWEPLHSNSLGMGPVTELEPLCTPVVTCTPGCTTYSFVFTYPEADSFPS CAA HRK
GSSSNEPSSDLSPTLLAL
```

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]
MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLGNVLVCVLAHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
>myoglobin 2MM1 NP_005359.1 [Homo sapiens]
MGLSDGEWQLVNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKAEDLKKHGATVL
TALGGILKKKGHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFR
KDMASNYKELGFQG
>neuroglobin 1OJ6A NP_067080.1 [Homo sapiens]
MERPEPELIQSWRAVSRSPLEHTVLFARLFALEPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVML
VIDAAVTNVEDLSSLEEYLASLGRKHRAVGVKLSSFSTVGESLLYMLEKCLGPAFTPATRAAWSQLYGAV
VQAMSRGWDGE
>soybean_globin 1FSL leghemoglobin P02238 LGBA_SOYBN [Glycine max]
MVAFTEKQDALVSSSFEAFKANIHQYSVVFYTSILEKAPAAKDLFSFLANGVDPTNPKLTGHAEKLFALV
RDSAGQLKASGTVVADAALGSVHAQKAVTDPQFVVVKEALLKTIKAAGDKWSDELSRAWEVAYDELAAA
IKKA
>rice_globin 1D8U rice Non-Symbiotic Plant Hemoglobin NP_001049476.1 [Oryza sativa (japonica)
MALVEDNNNAVASFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLRNSDVPLEKNPK
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

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

CLUSTAL 2.1 multiple sequence alignment

Red residues indicate alpha helices
using 3D structures

beta_globin	-----MVHLT PEEKSAVTALW GKVN--VDEVGGEALGRLL VVY PWTQRFF F EESFG-	47
myoglobin	-----MGLS DGEWQLVLNVW GKVEAD I PGHGQEVLIRLF KGH PETLEK FD KFK-	48
neuroglobin	-----MERPE PELI RQSWRAVSRS PLE HGTVL FARL FALEPDLLPL F QYNCR	47
soybean_globin	-----MVAFT EK QDALVSSS FEAF KAN I PQYSVV FYT SILEK KAPA A KDL F SFLA-	49
rice_globin	MALVEDNNNAVAVSFS EEQE ALVL K SWAIL K KDSANIALRFFLK I FEVAPSASQM F SFLR-	59
	: : : : .. . : : * * .	
beta_globin	DLST PDAVMGNPKVKA HGKKVLGAFSDG LAHLDNLKGTF AT -----LSEL HCDKLHVDP	101
myoglobin	HLKSEDEM KAS EDLKKHGATVLTALGGIL KKKGHHEAEI KP -----LAQSHATKH KI IPV	102
neuroglobin	QFSSPEDCLSS PEFLD HIRKVMLVIDAAVTNVEDLSS EEY---LASLGRK HRAVG V KLS	104
soybean_globin	--NGVDPT--NPKLTG HAEKL FALVRD SAGQLKAS GTVVAD----AALGSV HAQKAV T D P	101
rice_globin	--NSDVPLEKN PKLKT HAMSVFVMTCEAAAQLRK AGKVTVRDTTLKRLGATH HLKYGVGDA	117
	. . * : : : : * . *	
beta_globin	ENFRLLGNVLVCVLAHH F GKEFT PPVQAAYQKV VAGVANAL AHKYH -----	147
myoglobin	KYLEFISECII IQVLQSKH PGDFGADA QGAMNKALEL FRKDMASNY KELGFQG	154
neuroglobin	SFSTVGESLLY MLEKCLG -PAFT PATRAAWSQLY GAVV QAM SRGWDGE-----	151
soybean_globin	QFVVVK EALL KTI KA AVG -DKWS DELSRAWEV A YDELAAAIKK A-----	144
rice_globin	HFEVV KFALLDTIKEE VPADMWS PAMKSAWSEAY DHLVAA IK QEMKPAE---	166

Tree

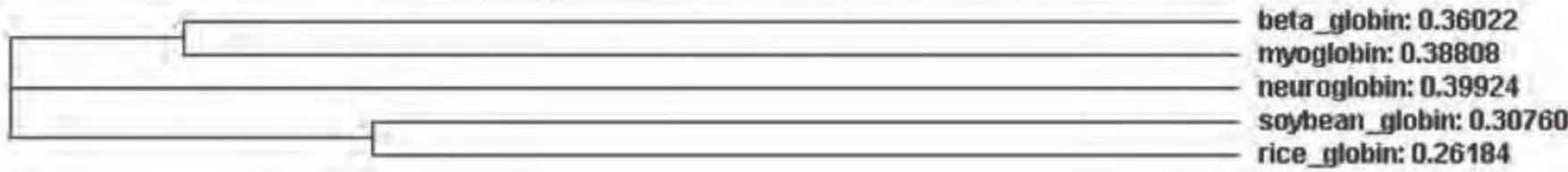
- Common tree output format is Newick

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);
```

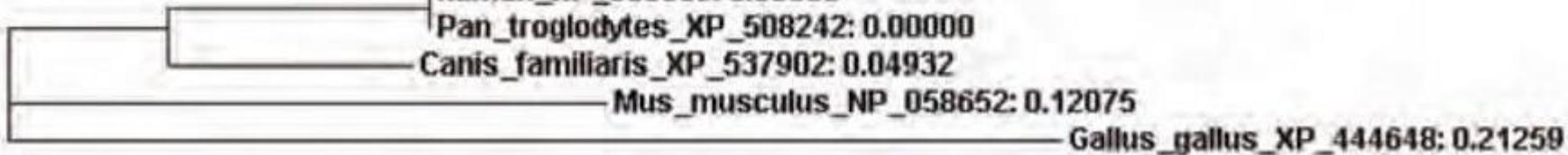


Soybean and rice are extant sequences; appear at the terminal nodes of the tree

Second example

(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);
```



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

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

CLUSTAL 2.1 multiple sequence alignment

F44

human_NP_000509
Pan_troglodytes_XP_508242
Canis_familiaris_XP_537902
Mus_musculus_NP_058652
Gallus_gallus_XP_444648

MVHLTPEEKSAVTALWKGKVNDEVGGEALGRLLVVYPWTQRFESFGDLS 50
MVHLTPEEKSAVTALWKGKVNDEVGGEALGRLLVVYPWTQRFESFGDLS 50
MVHLTAEEKSLVSGLWGKVNDEVGGEALGRLLIVYPWTQRFDFSGDLS 50
MVHLTDAEKSASCLWAKVNDEVGGEALGRLLVVYPWTQRYFDSFGDLS 50
MVHWTAEKKQLITGLWKGKVNVAECGAALARLLIVYPWTQRFASFGNLS 50

human_NP_000509
Pan_troglodytes_XP_508242
Canis_familiaris_XP_537902
Mus_musculus_NP_058652
Gallus_gallus_XP_444648

IPDAVMGNPKVKAHGKKVLGAFSDGLAHLNDNLKGTFAILSELHCDKLHVD 100
IPDAVMGNPKVKAHGKKVLGAFSDGLAHLNDNLKGTFAILSELHCDKLHVD 100
TPDAVMSNAKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVD 100
SASAIMGNPKVKAHGKKVITAFNEGKLNDNLKGTFAISLSELHCDKLHVD 100
SPTAILGNPMVRAHGKKVLTSFGDAVKNLDNIKNTFSQLSELHCDKLHVD 100

human_NP_000509
Pan_troglodytes_XP_508242
Canis_familiaris_XP_537902
Mus_musculus_NP_058652
Gallus_gallus_XP_444648

PENFRLLGNVLVCVLAHHFGKEFTPPVQAAQKVVAGVANALAHKYH 147
PENFRLLGNVLVCVLAHHFGKEFTPPVQAAQKVVAGVANALAHKYH 147
PENFKLLGNVLVCVLAHHFGKEFTPQVQAAQKVVAGVANALAHKYH 147
PENFRLLGNAIVIVLGHHLGKDFTPAAQAAFQKVVAGVATALAHKYH 147
PENFRLLGDILIIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH 147

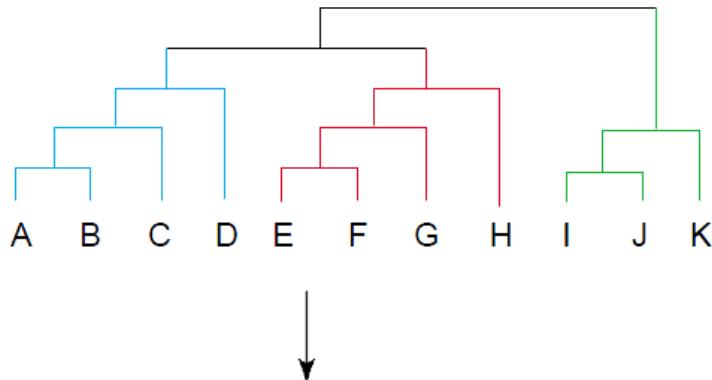
H72

H104

Acidic/Basic: Blue/Red

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

(a) Guide tree



(b) Sequence addition order

Step 1	A + B	E + F	I + J
Step 2	AB + C	EF + G	IJ + K
Step 3	ABC + D	EFG + H	
Step 4	ABCD + EFGH		
Step 5	ABCDEFGH + IJK		

TRENDS in Genetics



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

(a)

taxon

	10....20....30....40....50
Fu	Nosema.40928	QFGLFSPEEIRASSVALIR--YPETLENG--VPKESGLVCAHGFGHIELVK
Fu	Aspergillus.	QFGLFSPEEIKRMSVVHVE--YPETMDEQRQRPRTKGLECPGHFGHIELAT
Ap	Plasmodium.3	ELGVLDPEIIKKISVCEIV--NVDIYKDG--FPREGGLYCPGHFGHIELAK
An	Cricetulus.2	QFGVLSPDELKRMSVTBEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
An	Homo.7434727	QFGVLSPDELKRMSVTBEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
An	Drosophila.9	QFGILSPDEIRRMSVTBEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
An	Celegans.133	QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGLDCPGHFGHIELAK
Fu	Spombe.54881	QFGILSPDEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
P1	Athaliana.40	QFGILSPDEIRQMSVIH---VEHSETTEKGKPKVGGLECPGHFGYLELAK
My	Ddiscoideum.	-----ECPGHFGHIELAK
Rh	Porphyra.316	-----ECPGHFGFIELAK
Kt	Tbrucei.1021	QFEIFKERQIKSYAVCLVEHAKSYANA--ADQSGEAECPGHFGYIELAE
Kt	Leishmania.7	QFEVFKEAQIKAYAKCIIIEHAKSYEHG--QPVRGGIECPGHFGYVELAE

(b)

taxon

	10....20....30....40....50
Fu	Nosema.40928	QFGLFSPEEIRASSVALIR--IRYPETLE--NGVPKESGLVCAHGFGHIELVK
Fu	Aspergillus.	QFGLFSPEEIKRMSVVHVE--VEYPETMDEQRQRPRTKGLECPGHFGHIELAT
Fu	Spombe.54881	QFGILSPDEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
Ap	Plasmodium.3	ELGVLDPEIIKKISVCE--IVNDIYK--DGFPREGGLYCPGHFGHIELAK
An	Cricetulus.2	QFGVLSPDELKRMSVTBEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
An	Homo.7434727	QFGVLSPDELKRMSVTBEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
An	Drosophila.9	QFGILSPDEIRRMSVTBEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
An	Celegans.133	QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGLDCPGHFGHIELAK
P1	Athaliana.40	QFGILSPDEIRQMSVIH---VEHSETTE--KGKPKVGGLECPGHFGYLELAK
My	Ddiscoideum.	-----ECPGHFGHIELAK
Rh	Porphyra.316	-----ECPGHFGFIELAK
Kt	Tbrucei.1021	QFEIFKERQIKSYAVCL--VEHAKSYA--NAADQSGEAECPGHFGYIELAE
Kt	Leishmania.7	QFEVFKEAQIKAYAKCIIIEHAKSY--EHGQPVRGGIECPGHFGYVELAE

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

TRENDS in Genetics

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

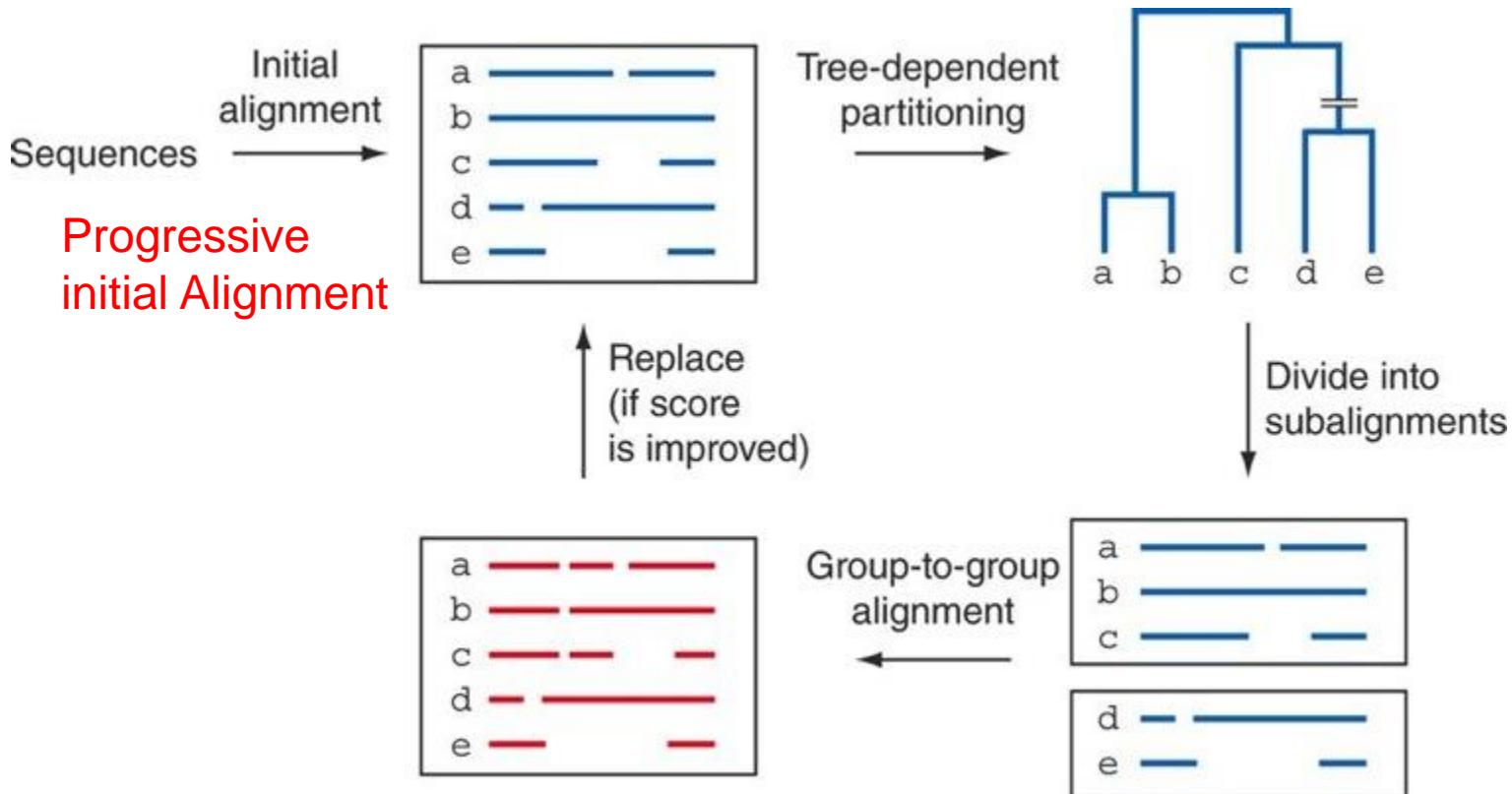
- An approach is to look at the alignment and delete all residues around gaps (and surrounding residues)
 - Why?
 - If there is misalignment then there is no real information
 - If there is a huge insertion and the rest of the alignment is all correct. The group that shares the extra (let us say 9-NT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

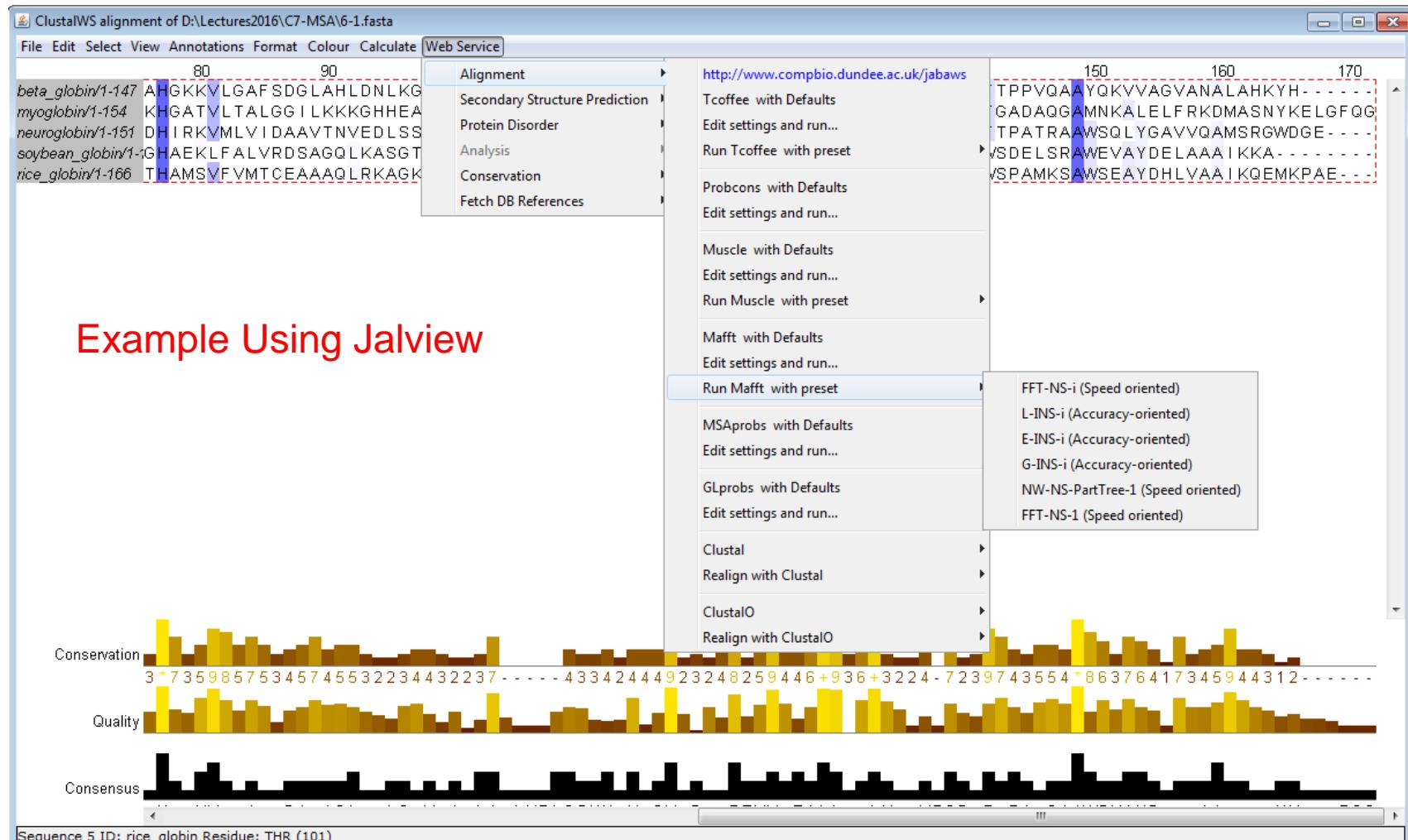
- Progressive methods have a limitation that once an error occurs it progresses and cannot be corrected
- Iterative approaches offer a solution to that approach
- So, what are Iterative approaches?

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



Example Using Jalview

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

All MSAs are not same

We will compare MUSCLE with MAFFT
(book compares more than two methods)

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

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

hbb_human	MVHLTPEEKSAVTALWGKVNVD -- EVGGEALGRLLVVY PWTQRFFE-SFG
hbb_chimp	MVHLTPEEKSAVTALWGKVNVD -- EVGGEALGRLLVVY PWTQRFFE-SFG
hbb_dog	MVHLTAEEKSLVSGLWGKVNVD -- EVGGEALGRLLIIVY PWTQRFFD-SFG
hbb_mouse	MVHLTDAAEKSACVSVCLWAKVNPD -- EVGGEALGRLLVVY PWTQRFYFD-SFG
hbb_chicken	MVHWTAAEKQLITGLWGKVNV -- ECGAEALARLLIVY PWTQRFFA-SFG
myoglobin	MGLSDGEWQLVNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFD-KFK
neuroglobin	MERPEPELIRQSWRAVSRSPLEHGTVLFARLFAEPDLLPLFQYNCR
soybean	MVAFTEKQDALVSSSFEAKFANIPQYSVVFYTISILEKAPAAKDLFS-FLA
rice	MALVEDNNAAVVSFSSEQEALVLKSWAILKKDSANIALRRFLKIFEVAPSASQMES-FLR
	: : : . . . : : * * *
	▼2 ▼3
hbb_human	DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAH -- LDNL -- KGTFA TLSELHCDKLHVDP
hbb_chimp	DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAH -- LDNL -- KGTFA TLSELHCDKLHVDP
hbb_dog	DLSTPDAVMSNAKVKAHGKKVLNSFSDGLKN -- LDNL -- KGTFA KLSL SELHCDKLHVDP
hbb_mouse	DLSSASAISGMNPVKVKAHGKKVITAFNEGLKN -- LDNL -- KGTFA SLSL SELHCDKLHVDP
hbb_chicken	NLSSPTAILGNPMVRAGKKVLTSGFDAVKN -- LDNI -- KNTFS QSLSELHCDKLHVDP
myoglobin	HLKSEDEMKA SEDLKKHGATVLTALGGILKK -- KGHH -- EAEIKPLAQSHATKH KIPV
neuroglobin	QFSSPEDCLSSPFELDHIRKVMLVIDAAVTNVEDLSSL -- EEYLASLGRKH-RAVGVLKL
soybean	NGVDP --- TNPKLTGHAEKLFALVRD SAGQLKAS GTV-VADAA -- LGSVH-AQKAV T D
rice	NSDVP -- LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRD TTLKRLGATH-LKYGVGD

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

hbb_human	- - - - M V H L T P E E K S A V T A L W G K V N V D - - E V G E A L G R L L V V V P W T Q R F E - S F G
hbb_chimp	- - - - M V H L T P E E K S A V T A L W G K V N V D - - E V G E A L G R L L V V V P W T Q R F E - S F G
hbb_dog	- - - - M V H L T A E E K S L V S G L W G K V N V D - - E V G E A L G R L L I V V P W T Q R F E - S F G
hbb_mouse	- - - - M V H L T D A E K S A V S C L W A K V N P D - - E V G E A L G R L L V V V P W T Q R Y F D - S F G
hbb_chicken	- - - - M V H W T A E E K Q L I T G L W G K V N V A - - E C G A E A L A R L L I V V P W T Q R F A - S F G
myoglobin	- - - - M G L S D G E W Q L V L N V W G K V E A D I P G H Q G Q E V L I R L F K G H P E T L E K F D - K F K
neuroglobin	- - - - M E R P E P E L I R Q S W R A V S R S P L E H G T V L F A R L F A E P D L L P L F Q Y N C R
soybean	- - - - M V A F T E K Q D A L V S S S F E A F K A N I P Q Y S V V F Y T S I L E K A P A A K D L F S - F L A
rice	M A L V E D N N A V A V S F S E E Q E A L V L K S W A I L K K D S A N A I L R F F L K I F E V A P S A S Q M F S - F L R

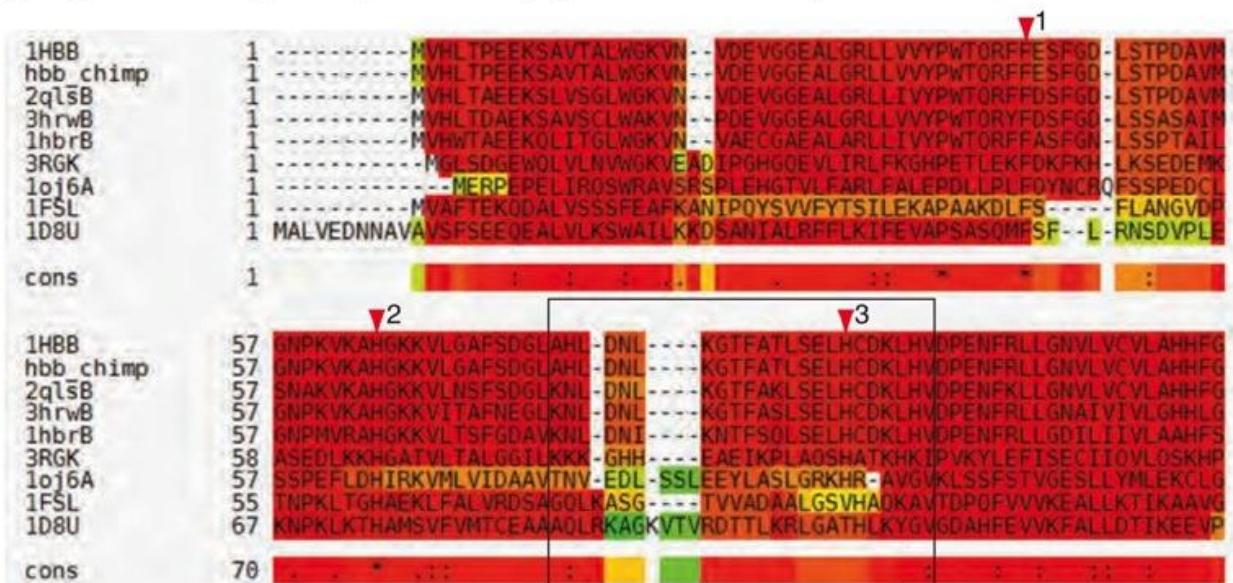
	▼2		▼3
hbb_human	DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL	---	DNLKGTFATLSELHCDK
hbb_chimp	DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL	---	DNLKGTFATLSELHCDK
hbb_dog	DLSTPDAVMSNAVKAHGKKVLNSFSDGLKNL	---	DNLKGTFAKLSSELHCDK
hbb_mouse	DLSSASAICGNPKVKAHGKKVITAFNEGLKNL	---	DNLKGTFFASLSELHCDK
hbb_chicken	NLSSPTAILGNPVMRAHGKKVLTSFGDAVKNL	---	DNIKNTFSQLSELHCDK
myoglobin	HLKSEDEMNASEDLKKHGATVLTALGGILKKK	---	GHHEAEIKPLAQSHATK
neuroglobin	QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNV	---	EDLSSLEEYLASLGRKHRAVGVKLS
soybean	NGVDPT----NPKLTGHAEKLFALVRDSDAGQL	---	KASGTVVADALGSVHQAKAV
rice	NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA	*	TDP

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



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

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



Consistency-based Approaches

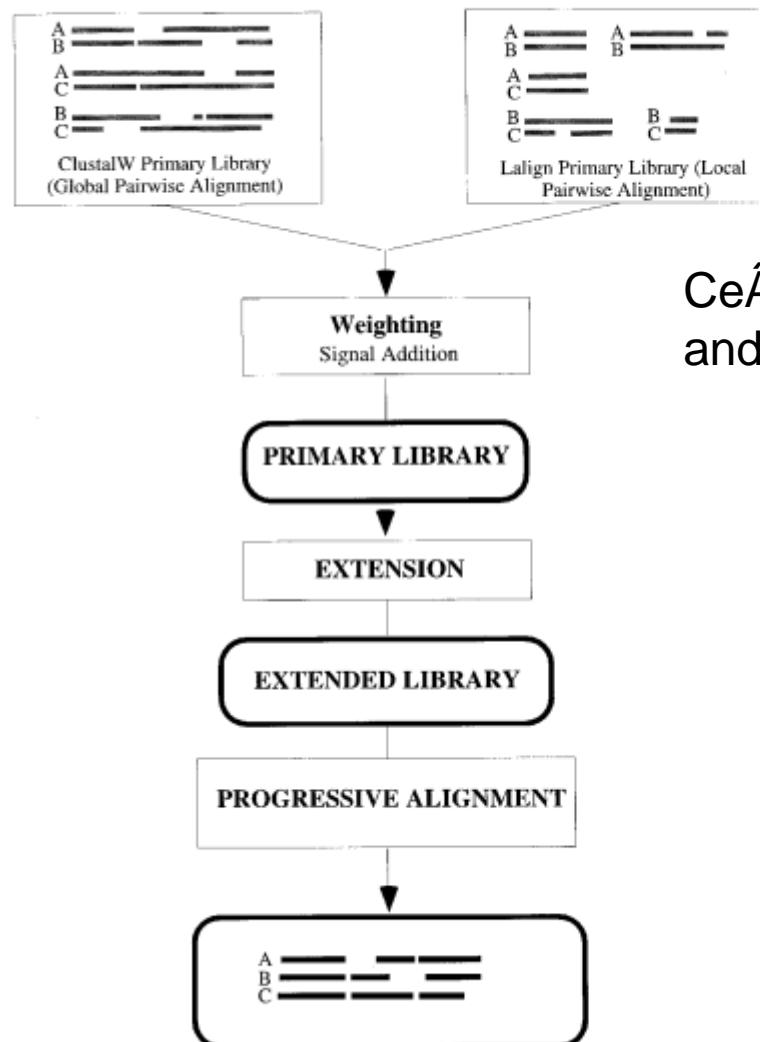
- T-COFFEE

- 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)

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

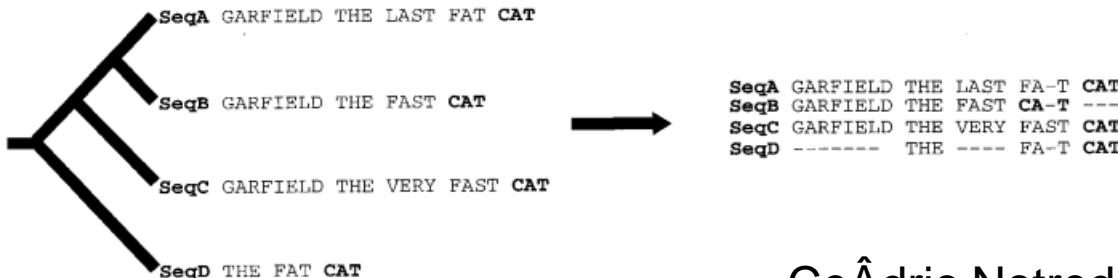
Objective scoring function: sum-of-pairs scores;
Assign 1 if the all the residues in a column are
aligned to the reference and 0 if not (crude method)



CeÂdric Notredame, Desmond G. Higgins
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

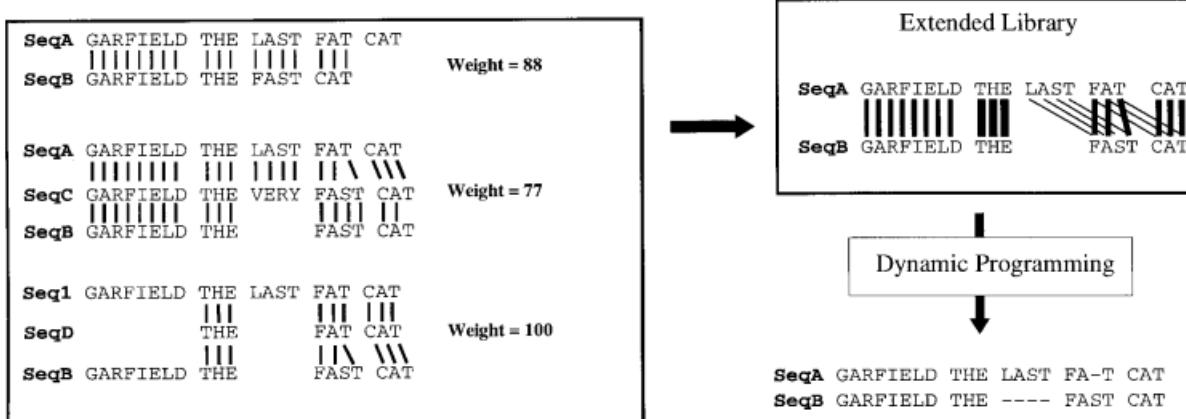


Note the word CAT
is misaligned

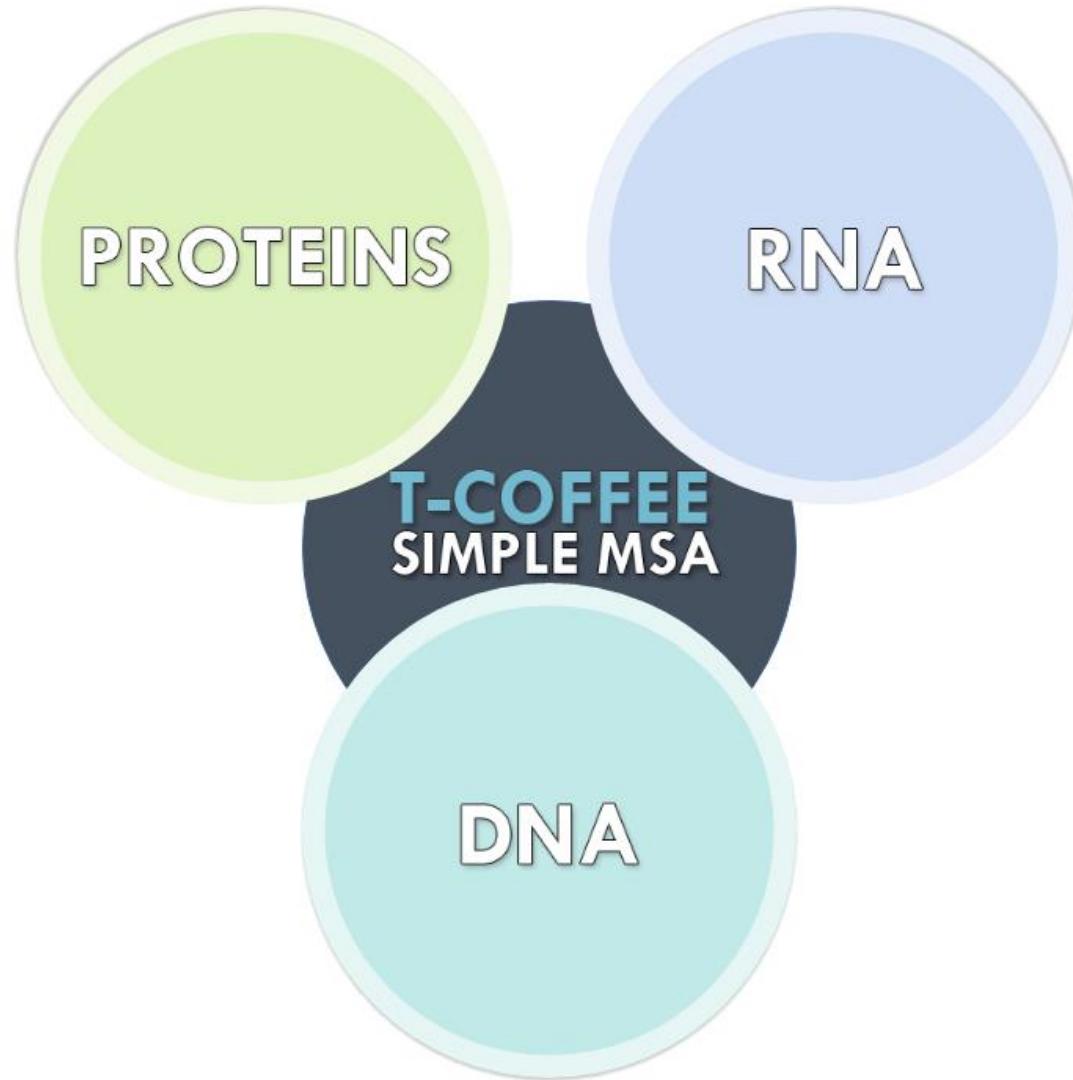
b) Primary Library

SeqA GARFIELD THE LAST FAT CAT	Prim. Weight = 88	SeqB GARFIELD THE ---- FAST CAT	Prim Weight = 100
SeqB GARFIELD THE FAST CAT ---		SeqC GARFIELD THE VERY FAST CAT	
SeqA GARFIELD THE LAST FA-T CAT	Prim. Weight = 77	SeqB GARFIELD THE FAST CAT	Prim. Weight = 100
SeqC GARFIELD THE VERY FAST CAT		SeqD ----- THE FA-T CAT	
SeqA GARFIELD THE LAST FAT CAT	Prim. Weight = 100	SeqC GARFIELD THE VERY FAST CAT	Prim. Weight = 100
SeqD ----- THE ---- FAT CAT		SeqD ----- THE ---- FA-T CAT	

c) Extended Library for seq1 and seq2



A and B
are
aligned
through
C and D

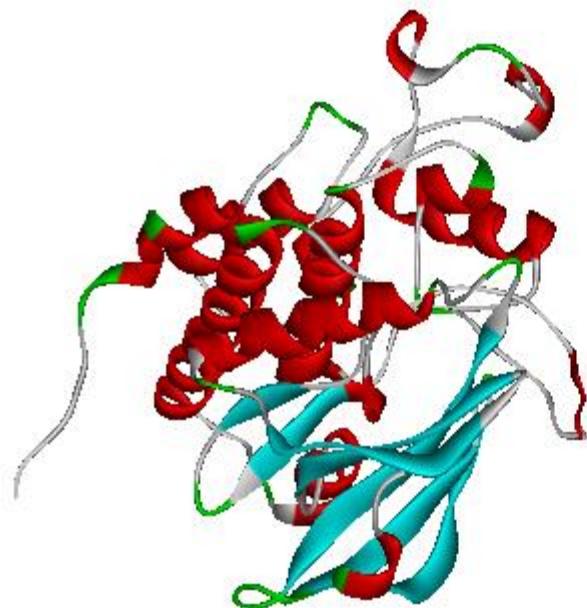


Structure-based approach

- 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

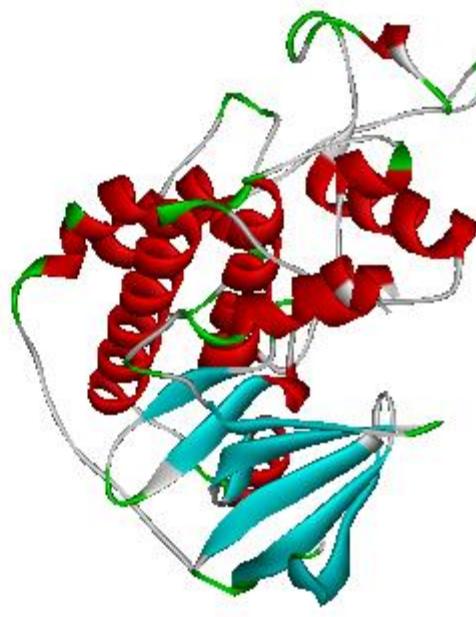
Cyclin Dependent Kinase (Alpha + Beta)

Human CDK6



42% identity
RMSD 1.8 Ang

Human CDK2



Structural Overlay done
using DeepView

1blx

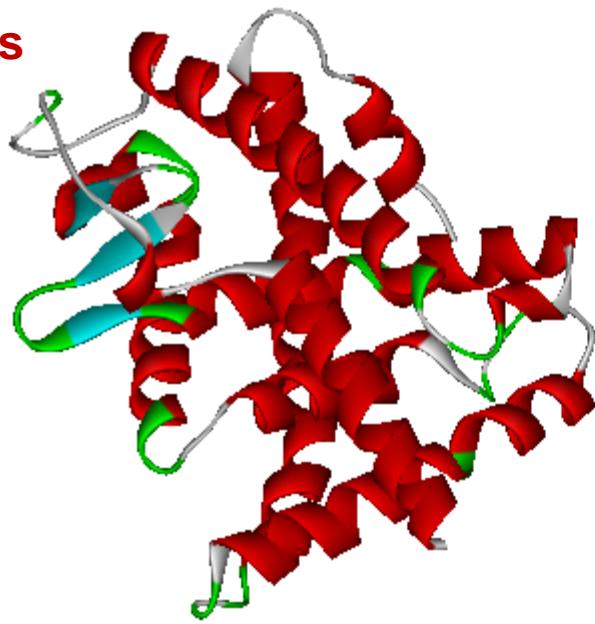
1ckp

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

Nuclear Hormone Receptor (All Alpha)

Core fold is conserved

Loop regions differ

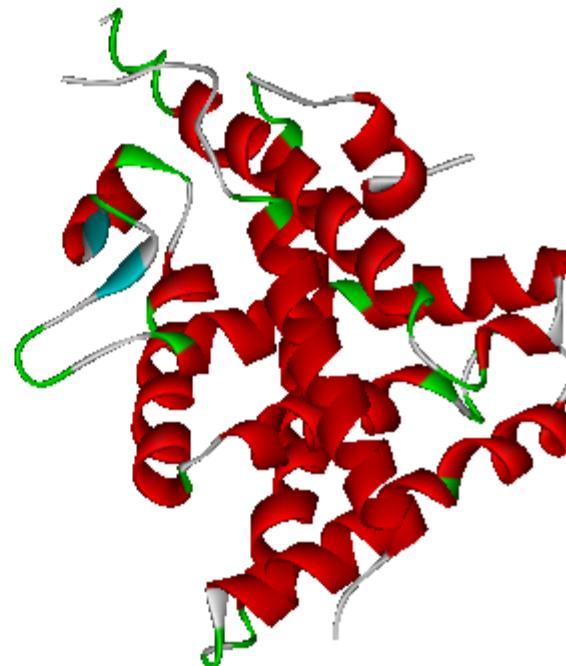


2LBD

Structural Overlay done using DeepView

38% identity
RMSD 1.3 Ang

Human Nuclear receptor NER



1P8D

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

1CYD_C|PDBID|CHAIN|SEQUENCE
 1A27_|PDBID|CHAIN|SEQUENCE

 1CYD_C|PDBID|CHAIN|SEQUENCE
 1A27_|PDBID|CHAIN|SEQUENCE

 1CYD_C|PDBID|CHAIN|SEQUENCE
 1A27_|PDBID|CHAIN|SEQUENCE

 1CYD_C|PDBID|CHAIN|SEQUENCE
 1A27_|PDBID|CHAIN|SEQUENCE

 1CYD_C|PDBID|CHAIN|SEQUENCE
 1A27_|PDBID|CHAIN|SEQUENCE

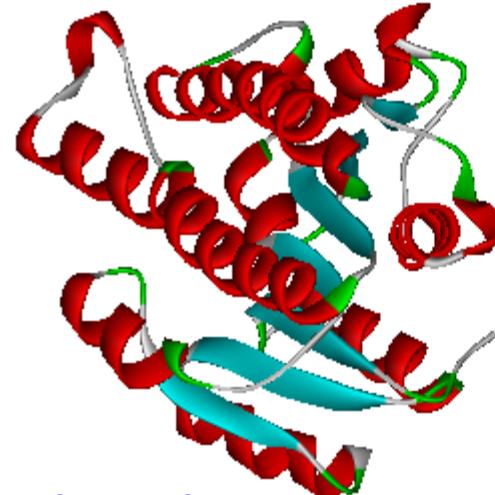
MKLNFSGILRALVTGAGKGIGRDTVKAIHASGAKVVAVTRTNSDLVSI--- 47
 -----ARTVVIITGCSSGIGIHLAVRIASDPSQSFKVYATIRDLKTQGRL 45
 : . * : * . * * . . * . : . * . : * * :
 -----AKECP--GIEPVCVDIGDWDATEKALG--GIGPVDLLVNNAALVI 88
 WEAARALACPPGSIELTLQLDVRDSKSVAARERVTEGRVDVLVCNAGLGL 95
 * * . . * . . * . . * . * * : * * * . * :
 MQPFLEVTKEAFDRSFSVNLRSFQVSQMVARDMINRGVPGSIVNVSSMV 138
 IGPLEALGEDAVASVLDVNVVGTVRMLQAFLPDMKRRGS-GRVLVTGSVG 144
 : * . : : * . . * * : . . : * . * * . * : . * :
 AHVTFPNLIYTSSSTKGAMTMILTKAMAMELGPHKIRVNSVNPTVVI TDMGK 188
 GLMGLPFDVYCASKFALEGICESLA VLLLPFGVHLSLIECGPVHTAFME 194
 . : * . . * . . * : * . : * . . : . : * * : .
 KVSADPEFARKLKERHPLRKFAEVEDVVNSIL-----FLLSDR 226
 KV LGSPEEVLDRTDIHTFHFRYQYLAHSKQVFREAAQNPEEEVAEVFITAL 244
 ** . . * . . . : * . . : . . : . . : . : . :
 SASTSGGGILVDAGYLAS----- 244
 RAPKPTIRYFTTERFLPLLRLDDPSGSNYVTAMHREVFGDVPA 289
 * : . . .

ClustalW 1.83
www.ebi.ac.uk

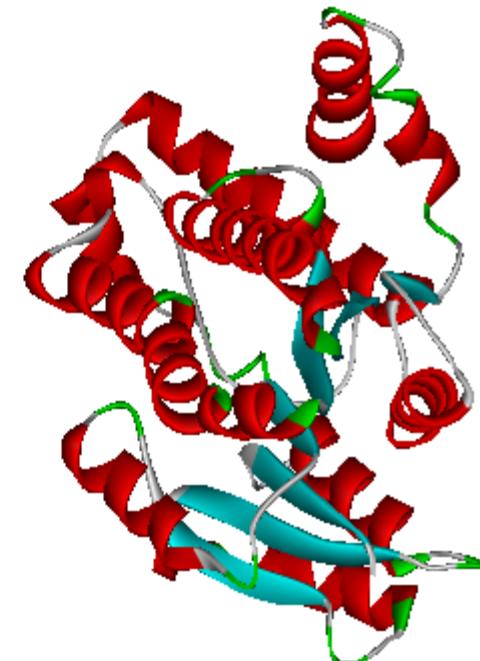
**19% identity
RMSD
1.7 Ang**

Human Estradiol 17-beta-dehydrogenase 1

Mouse Carbonyl reductase [NADPH] 2



1CYD_C



1A27

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

Short Chain Dehydrogenase (Alpha + Beta)

Structural Overlay done
using DeepView

2/29/2020

Dr. S. Ravichandran

72

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
MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL
DNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPQVQAAYQKVAGVANALAHKYH
>hbb_chimp
MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL
DNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPQVQAAYQKVAGVANALAHKYH
>hbb_dog
MVHLTAEEKSLVSGLWGKVNDEVGGEALGRLLIVYPWTQRFFDSFGDLSTPDAVMSNAVKVKAHGKKVLNSFSDGLKNL
DNLKGTFAKLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGKEFTPQVQAAYQKVAGVANALAHKYH
>hbb_mouse
MVHLTDAAEKSASVCLWAKVNPDEVGGEALGRLLIVYPWTQRFFDSFGDLSSASAIMGNPKVKAHGKKVITAFNEGLKNL
DNLKGTFASLSELHCDKLHVDPENFRLLGNAIVIVLGHHLGKDFTPAAQAAFQKVAGVATALAHKYH
>hbb_chicken
MVHWTAAEKQLITGLWGKVNVAECGAEARLLIVYPWTQRFFASFGNLSSPTAILGNPMVRAHGKKVLTSFGDAVKNL
DNIKNTFSQLSELHCDKLHVDPENFRLLGDILIIIVLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH
>myoglobin
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFHKLKSEDEMKAEDLKKHGATVLTALGGILKK
KGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGMNKALELFRKDMASNQYKELGFQG
>neuroglobin
MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNV
EDLSSLEELASLGRKHRAVGVKLSSFSTVGESLLYMLEKCLGPAFTPATRAAWSQLYGAVVQAMSRGWDE
>soybean
MVAFTEKQDALVSSSFEAFKANI PQYSVVFYTSILEKAPAAKDLFSFLANGVDPTNPKLTGHAEKLFALVRDSAGQLKA
SGTVVADAALGSVHAQKAVTDPQFVVVKEALLKTIKAAVGDKWSDELSRAWEVAYDELAAAIKA
>rice
MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLRNSDVPLEKNPKLKTHAMSVF
VMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDAHFEVVKFALLDTIKEEVPADMWS PAMKSAWSEAYDHLVAI
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

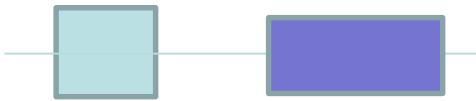
Sequence → Function



Why Sequence Alignment is Important?



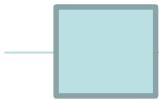
Consensus Sequence (Majority is taken as the consensus; skip Variation)
Search DB



Not very useful (binary etc.)

Multiple Seq Alignment (MSA)

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



PSSM (uses frequencies for each position)

DNA Binding Function to the domains

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

- 4 sequences
 - 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)
- 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	GHEGV
GHKKV	GHKKI
GHIKV	GHIKV
GH**V	GHAMI
	GH** [I/V]

P-x(2)-G-E-S-G(2)-[AS]



<http://prosite.expasy.org/scanprosite/>



PROSITE

ruler:



P06681
(CO2_HUMAN)

(752 aa)

[View all PROSITE motifs hits on sequence](#)

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) - [A S]

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 1flyp glads apscpqnvni sggftftlshg wapgslltys cpqglypspa
61 srlckssgqw qtpgatrsls kavckpvrcp apvsfengiy tprlgssypvg gnvsfecedg
121 filrgspvrq crpngmwdge tavcdngagh cpnpgisla vrtgfrfghg dkvryrcssn
181 lvltgssere cqgngvwsqt epicrqpssy dfpedvapal gtsfshmlga tnptqktkes
241 lgrkiqiqrsl ghl nly lll d csqsvs endf lifkesaslm vdrifsfein vsvaiitfas
301 epkvlmsvln dnsrdmtevi sslenanykd hengtgtnty aalnsvylmm nnqmrl1lgme
361 tmawqeirha iilltdgksn mggspktavd hireilning krndyldiya igvgkldvdw
421 relnelgskk dgerhafilq dtkalhqvfe hmldvskltd ticgvgnmsa nasdqertpw
481 hvtikpkssqe tcrgalisdq wvltaahcfr dgndhslwr vngdpksqwg kefliiekavi
541 spgfdvfakk nqgilefygd diallklaqk vkmstharpi clpctmeanl alrrpqggstc
601 rdhenellnk qsvpahfval ngsklninlk mgvewtscae vvsqektmfp nltdvrevvt
661 dqflcsqtqe des pckgesg gavflerrfr ffqvglvswg lynpclgsad knsrkraps ,
721 kvppprdfhi n1frmqpwl r qhlgdvlnf1 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]

Globin

Clan:

Bac_globin; Globin; Phycobilisome and Protoglobin

Summary

Domain organisation

Alignments

Relationships

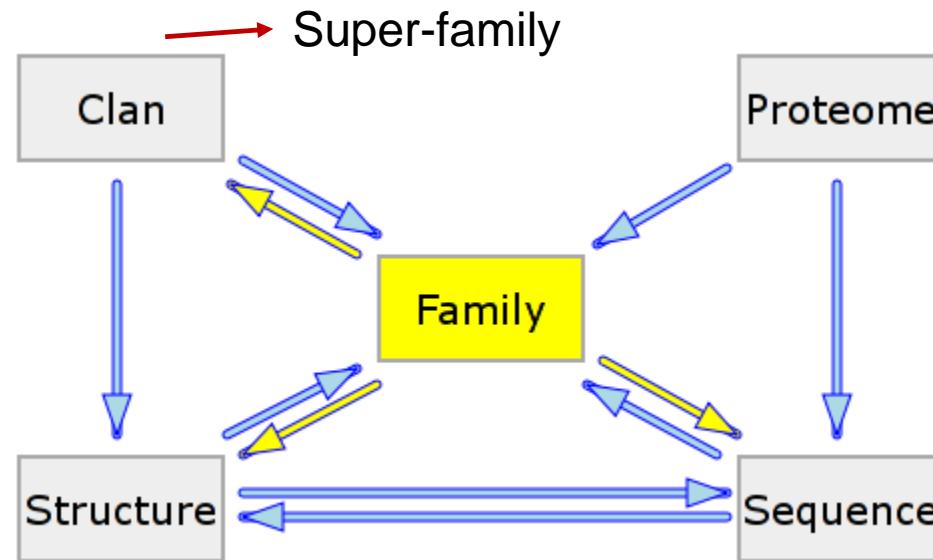
Species

Interactions

Structures

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](#)

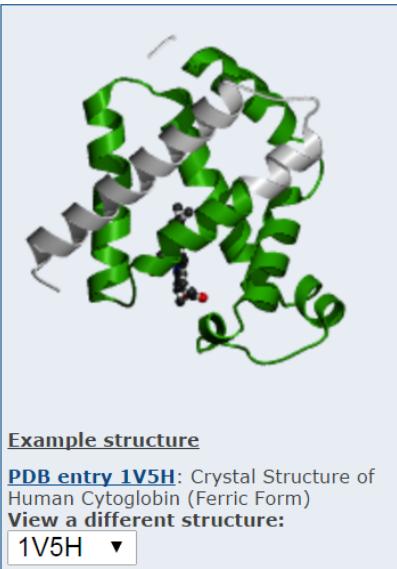
[Phycobilisome](#)

[Protoglobin](#)

External database links

CATH: [1.10.490.10](#)

SCOP: [46458](#)



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

Family: *Globin* (PF00042)

 34 architectures
  6000 sequences
  5 interactions
  2886 species
  1971 structures

Summary

Domain organisation

Clan

All assignments

HMM logo

Trees

model

Species

Interactions

Structures

Jump to... ↴

Alignments

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

| View options

We make a range of alignments for each Pfam-A family. You can see a description of each [above](#). You can view these alignments in various ways but please note that some types of alignment are never generated while others may not be available for all families, most commonly because the alignments are too large to handle.

	Seed (73)	Full (6000)	Representative proteomes				NCBI (3331)	Meta (34)
			RP15 (348)	RP35 (594)	RP55 (949)	RP75 (1261)		
Jalview	✓	✓	✓	✓	✓	✓	✓	✓
HTML	✓	—	✓	✓	✓	✓	✗	✗
PP/heatmap	✗ ₁	—	✓	✓	✓	✓	✗	✗
Pfam viewer	✓	✓	✗	✗	✗	✗	✗	✗

[‡]Cannot generate RR/Heatmap alignments for seeds; no RR data available.

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

(b) Pfam seed alignment

Seed sequence alignment for PF00042

Q20638_CAEEL/74-184
 Q19601_CAEEL/105-215
 Q18311_CAEEL/32-140
 GLB4_LUMTE/11-120
 GLB4_LUMTE/11-120 (SS)
 GLB3_TYLHE/8-117
 GLB4_TYLHE/8-117
 GLB1_TYLHE/7-110
 GLB2_TYLHE/9-115
 GLB2_LUMTE/8-114
 GLB2_LUMTE/8-114 (SS)
 GLB_TUBTU/6-112
 GLB3_JAMSP/7-113

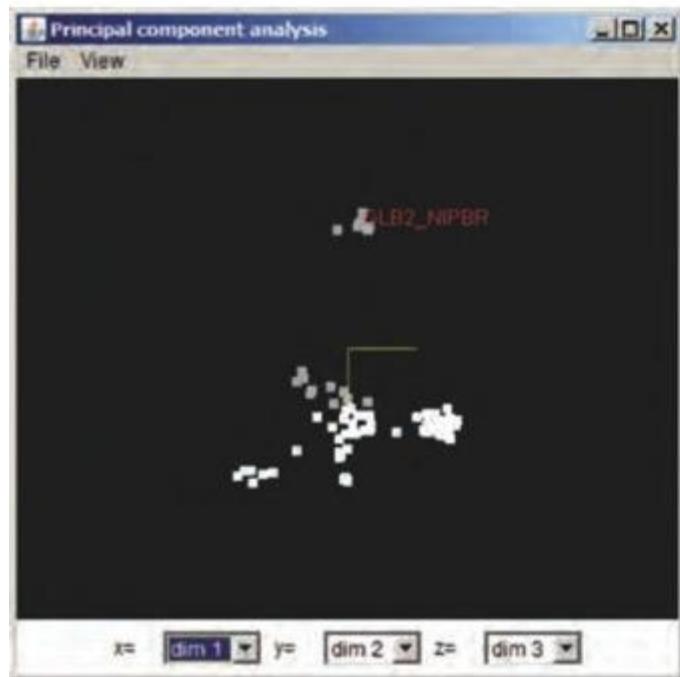
PFAM → PF00044 → Download and analyze
Or use
Jalview to carry out the analysis

Each point is one sequence;

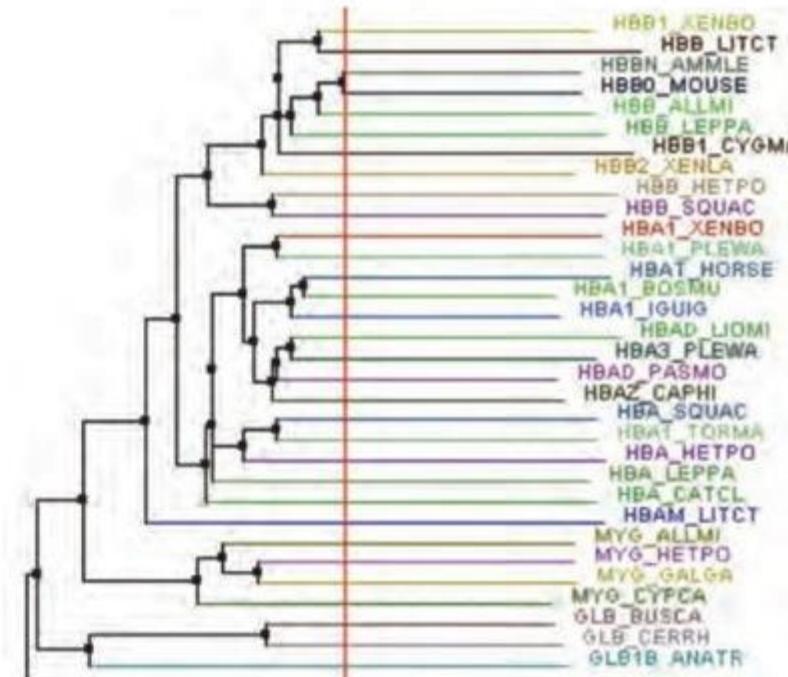
Here what we are doing is reducing the dimensionality to view the set of 23 proteins in a smaller dimensional space.

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/>

QUICK LINKS

[SEQUENCE SEARCH](#)

[VIEW A PFAM ENTRY](#)

[VIEW A CLAN](#)

[VIEW A SEQUENCE](#)

[VIEW A STRUCTURE](#)

[KEYWORD SEARCH](#)

[JUMP TO](#)

YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...

Analyze your protein sequence for Pfam matches



View Pfam annotation and alignments

See groups of related entries

Look at the domain organisation of a protein sequence

Find the domains on a PDB structure

Query Pfam by keywords

[Go](#)

[Example](#)

Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.

Or view the [help](#) pages for more information

Copy the sequence from next page

>Pfam_Query_sequence
MAFSQYISLAPELLLATAIFCLVFWVLRGTRTQVPKGLKSPPG
PWGLPFIGHMLTLGKNPHLSLTKLSQQ
YGDVLQIRIGSTPVVVLSGLNTIKQALVKQGDDFKGRPDLYSF
TLITNGKSMTFPDSDGPVWAARRRLAQ
DALKSFSIASDPTS VSSCYLEEHVSKEANHLISKFQKLMAEVG
HFEPVNQVVESVANVIGAMCFGKNFPR
KSEEMLNLVKSSKDFVENVTSGNAVDFFPVLRYLPNPALKRF
KNFNDNFVLSLQKTVQEHYQDFNKNSIQ
DITGALFKHSENYKDNGGLIPQEKIVNIVNDIFGAGFETVTTAIF
WSILLLVTEPKVQRKIHEELDTVIG
RDRQPRLSDRPQLPYLEAFILEIYRYTSFVPFTIPHSTTRDTSL
NGFHIPKECCIFINQWQVNHDEKQWK
DPFVFRPERFLTNDNTAIDKTLSEKVMLFGLGKRRRCIGEIPAK
WEVFLFLAILLHQLEFTVPPGVKVDLT
PSYGLTMKPRTEHVQAWPRFSK

Example using PFAM

Pfam 31.0 (March 2017, 16712 entries)

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

QUICK LINKS
[SEQUENCE SEARCH](#)
[VIEW A PFAM ENTRY](#)
[VIEW A CLAN](#)
[VIEW A SEQUENCE](#)
[VIEW A STRUCTURE](#)
[KEYWORD SEARCH](#)
[JUMP TO](#)

ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES

Paste your protein sequence here to find matching Pfam entries.

```
LLLVTEPKVQRKIHEELDTVIG  
RDRQPRLSDRPQLPYLEAFILEIYRYTSFVPFTIPHSTTRDTSLNGFH  
IPKECCIFINQWQVNHHDEKQWIK  
DPFVFRPERFLTNNTAIDKTLSEKVMLFGLGKRRCIGEIPAKWEVFL  
FLAILLHQLEFTVPPGVKVVDLT  
PSYGLTMKPRTEHVAWPRFSK
```

[Go](#) [Example](#)

This search will use an E-value of 1.0. You can set your own search parameters and perform a range of other searches [here](#).



Click here for additional options

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

Significant Pfam-A Matches

[Show](#) or [hide](#) all alignments.

Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.

European Molecular Biology Laboratory

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

Family: p450 (PF00067)

>Loading page components (1 remaining)...

908 architectures 85020 sequences 4 interactions 2410 species 1454 structures

Summary

Domain organisation

Clan

Alignments

HMM logo

Trees

Curation & model

Species

Interactions

Structures

Summary: Cytochrome P450

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

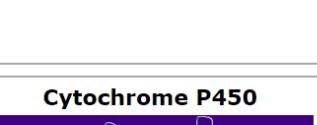
[Wikipedia: Cytochrome P450](#) [Pfam](#) [InterPro](#)

This is the Wikipedia entry entitled "[Cytochrome P450](#)". [More...](#)

Cytochrome P450 [Edit Wikipedia article](#)

Cytochromes P450 (CYPs) are proteins of the superfamily containing heme as a cofactor and, therefore, are hemoproteins. CYPs use a variety of small and large molecules as substrates in enzymatic reactions. They are, in general, the terminal oxidase enzymes in electron transfer chains, broadly categorized as P450-containing systems. The term P450 is derived from the spectrophotometric peak at the wavelength of the absorption maximum of the enzyme (450 nm) when

Cytochrome P450





Schultz et al. (1998) Proc. Natl. Acad. Sci. USA 95, 5857-5864
Ishunin et al. (2014) Nucleic Acids Res. doi: 10.1093/nar/gku240

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

- 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 proteins
- 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

 NCBI

Conserved Protein Domain Family *Hb-beta_like*

HOME SEARCH SITE MAP Entrez CDD Structure Protein Help

cd08925: **Hb-beta_like** ?



Hemoglobin beta, gamma, delta, epsilon, and related Hb subunits

Hb is the oxygen transport protein of erythrocytes. It is an allosterically modulated heterotetramer. Hemoglobin A (HbA) is the most common Hb in adult humans, and is formed from two alpha-chains and two beta-chains (alpha₂beta₂). An equilibrium exists between deoxygenated/unliganded/T(tense state) Hb having low oxygen affinity, and oxygenated /liganded/R(relaxed state) Hb having a high oxygen affinity. Various endogenous heterotropic effectors bind Hb to modulate its oxygen affinity and cooperative behavior, e.g. hydrogen ions, chloride ions, carbon dioxide and 2,3-bisphosphoglycerate. Hb is also an allosterically regulated nitrite reductase; the plasma nitrite anion may be activated by hemoglobin in areas of hypoxia to bring about vasodilation. Other Hb types are: HbA2 (alpha₂delta₂) which in normal individuals, is naturally expressed at a low level; Hb Portland-1 (zeta₂gamma₂), Hb Gower-1 (zeta₂epsilon₂), and Hb Gower-2 (alpha₂epsilon₂), which are Hbs present during the embryonic period; and fetal hemoglobin (HbF, alpha₂gamma₂), the primary hemoglobin throughout most of gestation. These Hb types have differences in O₂ affinity and in their interactions with allosteric effectors.

Links ?

Source: cd14765
Taxonomy: Gnathostomata
PubMed: 68 links
Book: 11 links
Protein: Representatives
Specific Protein
Related Protein
Related Structure
Architectures
Superfamily: cl21461

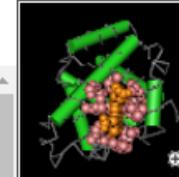
Conserved Features/Sites ? PubMed References ?

heme binding **tetramer**

Feature 1: heme binding site [chemical binding site]

Evidence:

- **Structure:** 1SI4: Human HbA2, delta subunit binds cyanide ion/heme; contacts at 4A
 - [View structure with Cn3D](#)
- **Citation:** PMID 15449937
- **Structure:** 1FDH: Human HbF, gamma subunit binds heme; contacts at 4A
 - [View structure with Cn3D](#)
- **Citation:** PMID 881729



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



Integrated 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

F Haemoglobin, beta-type (IPR002337)

Domains and repeats



Detailed signature matches

F IPR002337

Haemoglobin, beta-type

PRINTS

D IPR009050

Globin-like

SUPERFAMILY

D IPR012292

Globin/Protoglobin

GENE3D

D IPR000971

Globin

PRINTS

PROSITE

? no IPR

Unintegrated signatures

PANTHER

CDD



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

* 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)

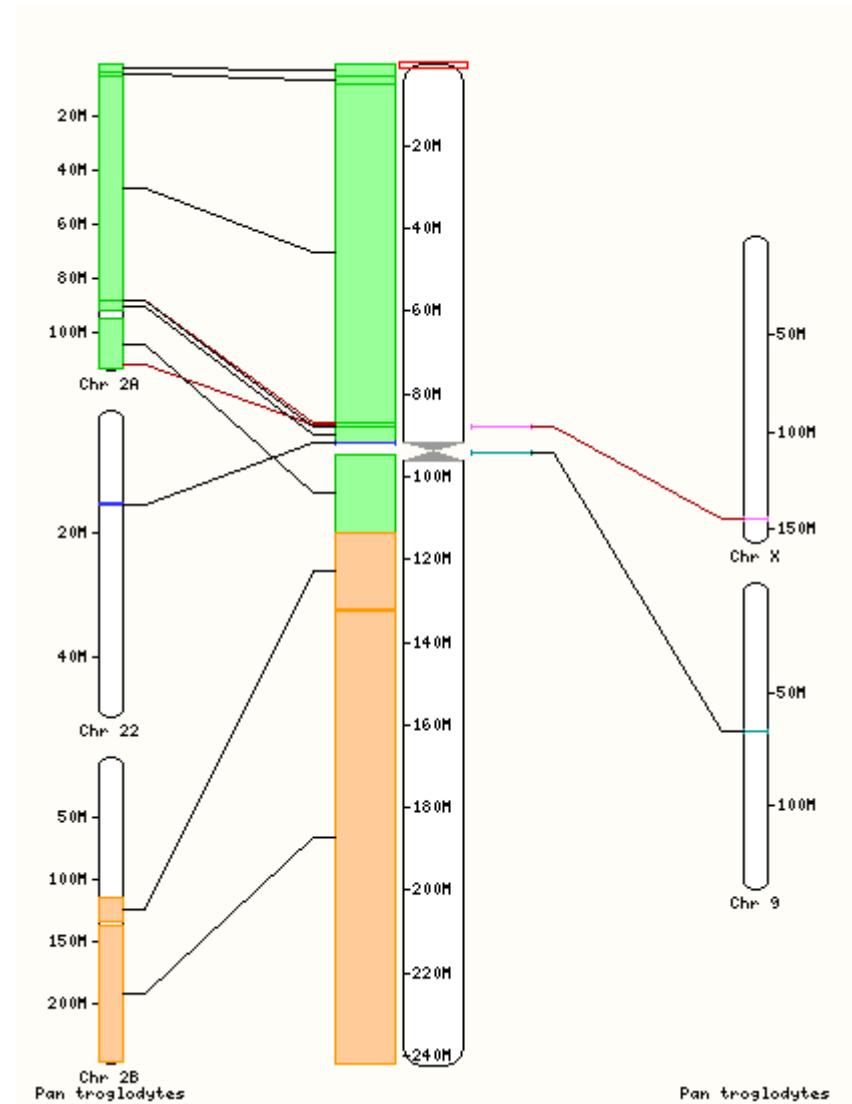
Differences in MSA of DNA

- Chromosomal duplications, deletions 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

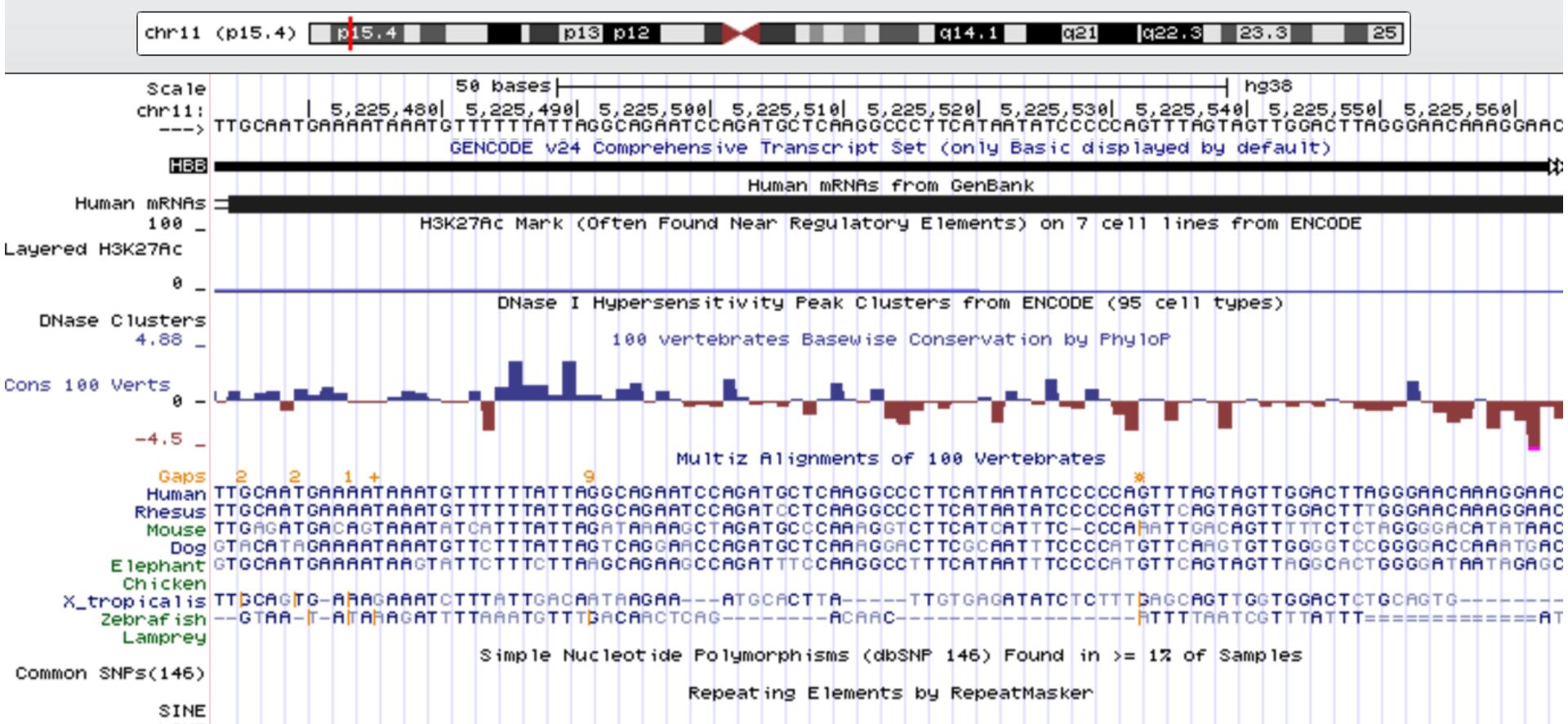
- Go to UCSC
- Type HBB
- Pick the first entry
- Hide unwanted things
 - Right click and hide
- Zoom only 100 residues
- You will be shown the multiple sequence alignment using Multiz

chr11:5,225,464-5,225,564

101 bp.

chr11:5,225,464-5,225,564

go



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](#)

The screenshot shows the Galaxy web interface. On the left, a sidebar titled "Tools" lists various MAF-related tools under the heading "Fetch Alignments/Sequences". A search bar at the top of the sidebar contains the query "fetch". The main content area features a large title "Running Your Own Understanding how Galaxy works" followed by the subtitle "An in-depth tutorial". Below the title is a series of five small circular icons. To the right of the main content is a "Tweets" section showing two tweets from the "Galaxy Project" account (@galaxyproject). The first tweet says "Applications are due TODAY." and includes a link to a Twitter page. The second tweet says "Open #usegalaxy positions at 7 institutions: bit.ly/gxy201608jobs 3 CLOSE TOMORROW" and includes a link to a Twitter page. At the bottom of the tweets is a link "Who's Hiring". On the far right, a "History" panel is open, showing an "Unnamed history" entry for a UCSC Main on Human knownGene (chr11) job with ID 5246696-5248301. The history panel also includes links to view the data in various formats like IGB, Ensembl, RViewer, and IGV, and a preview of the BED file content.

Galaxy now runs some (larger, multicore) jobs on [Jetstream](#), you may encounter a few problems related to this. We are working on these, and please feel free to report any errors you encounter.

Tools

fetch

Fetch Alignments/Sequences

- [Extract Pairwise MAF blocks](#) given a set of genomic intervals
- [Extract MAF blocks](#) given a set of genomic intervals
- [Stitch MAF blocks](#) given a set of genomic intervals
- [Stitch Gene blocks](#) given a set of coding exon intervals
- [MAF Coverage Stats](#) Alignment coverage information
- [Join MAF blocks](#) by Species
- [Filter MAF blocks](#) by Species
- [Filter MAF blocks](#) by Size
- [Extract MAF by block number](#) given a set of block numbers

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#). You can install your own Galaxy by following the [tutorial](#) and choose from thousands of tools from the [Tool Shed](#).

Tweets by @galaxyproject

Galaxy Project @galaxyproject Applications are due TODAY. [twitter.com/elixir_ita/sta...](#) 6h

Galaxy Project @galaxyproject Open #usegalaxy positions at 7 institutions: [bit.ly/gxy201608jobs](#) 3 CLOSE TOMORROW

History

search datasets

Unnamed history

1 shown

147 b

1: UCSC Main on Human knownGene (chr11): 5246696-5248301

2 regions

format: bed, database: hg19

display in IGB [View](#)
display at Ensembl [Current](#)
display at RViewer [main](#)
display with IGV [local Human hg19](#)
display at UCSC [main](#)

1.Chrom	2.Start	3.End	4.Name	5	6..5t
chr11	5246695	5248301	uc001mae.1	0	-
chr11	5246721	5246744	uc031pyr.1	0	+

A browser session will open up with Galaxy

Extract Pairwise MAF blocks given a set of genomic intervals (Galaxy Version 1.0.1) ▼ Options

Interval File
1: UCSC Main on Human: knownGene (chr11:5246696-5248301) ▼

Choose MAF source
Pairwise (hg19,bosTau4) ▼

Execute

What it does

This tool takes genomic coordinates, superimposes them on pairwise alignments (in MAF format) stored on the Galaxy site, and excises alignment blocks corresponding to each set of coordinates. Alignment blocks that extend past START and/or END positions of an interval are trimmed. Note that a single genomic interval may correspond to two or more alignment blocks.

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** (highlighted)
- 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 involved in [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

A No align Please select

Ensembl release

About Us

About us

Our sister

Ensembl Back

...
...

-- 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
- Chimpanzee (*Pan troglodytes*) - lastz



Conserved regions are marked as blue shade

Human	ACTTAT -- TTGCCTGGTATGCCTGGGCTTTGATGGTCTAGTATAGCTGCAGCCTTGCCCTGCAGGTATTATGGTAATAGAAAGAAAAGCTGCCTAACACTCTAGTCACACTAA
Chimpanzee	ACTTATTTGCCTGGTATGCCTGGGCTTTGATGGTCTAGTATAGCTGCAGCCTTGCCCTGCAGGTATTATGGGTATAGAAAGAAAAGCTGCATTACACTCTAGTCACACTAA
Gorilla	ACTTATTTGCCTGGTATGCCTGGGCTTTGATGGTCTAGTATAGCTGCAGCCTTGCCCTGCAGGTATTATGGTAATAGAAAGAAAAGCTGCATTACACTCTAGTCACACTAA
Human	GTAACCTACCATTGAAAAGCAACCCCTGCCTTGAGCCAGGATGATGGTATCTGCAGCAGTTGCCAACACAAGAGAAGGATCCATAGTTCATCATTAAAAAGAAAACAAAATAGAAAA
Chimpanzee	GTAACCTACCATTGAAAAGCAACCCCTGCCTTGAGCCAGGATGATGGTATCTGCAGCAGTTGCCAACACAAGAGAAGGATCCATAGTTCATCATTAAAAAGAAAACAAAATAGAAAA
Gorilla	GTAACCTACCATTGAAAAGCAACCCCTGCCTTGAGCCAGGATGATGGTATCTGCAGCAGTTGCCAACACAAGAGAAGGATCCATAGTTCATCATTAAAAAGAAAACAAAATAGAAAA
Human	AGGAAAACATATTCCTGAGCATAAGAACGTTAGGGTAAGTCTTAAGAAGGTGACAATTCTGCCAATCAGGATTCAAAGCTCTGCTTGACAATTGGTCTTCAGAATACTATAA
Chimpanzee	AGGAAAACATATTCCTGAGCATAAGAACGTTAGGGTAAGTCTTAAGAAGGTGACAATTCTGCCAATCAGGATTCAAAGCTCTGCTTGACAATTGGTCTTCAGAATACTATAA
Gorilla	AGGAAAACATATTCCTGAGCATAAGAACGTTAGGGTAAGTCTTAAGAAGGTGACAATTCTGCCAATCAGGATTCAAAGCTCTGCTTGACAATTGGTCTTCAGAATACTATAA
Human	ATATAACCTATATTATAATTCTAAAGCTGTGCAATTCTTGACCAGGATATTGCAAAAGACATATTCAAACCTCCGAGAACACTTTATTCACATATACTGCCTCTTATAC
Chimpanzee	ATATAACCTATATTATAATTCTAAAGCTGTGGAATTCTTGACCAGGATATTGCAAAAGACATATTCAAACCTCCGAGAACACTTTATTCACATATACTGCCTCTTATAC
Gorilla	ATATAACCTATATTATAATTCTAAAGCTGTGGAATTCTTGACCAGGATATTGCAAAAGACATATTCAAACCTCCAGAACACTTTATTCACATATACTGCCTCTTATAC
Human	AGGGATGTGAAACAGGGTCTTGAAAAGCTGCTAAATCTAAAACATGCTAATGCAGGTTAAATTAAATAAAAATAAACTCAAAGCTGCAATCTGCTTAAACAT
Chimpanzee	AGGGATGTGAAACAGGGTCTTGAAAAGCTGCTAAATCTAAAACATGCTAATGCAGGTTAAATTGATAAAAATAAACTCAAAGCTGCAATCTGCTTAAACAT
Gorilla	AGGGATGTGAAACAGGGTCTTGAAAAGCTGCTAAATCTAAAACATGCTAATGCAGGTTAAATTAAATAAAAATAAACTCAAAGCTGCAATCTGCTTAAACAT
Human	TTAAAATATTAAAGACGTCTTCCCAGGATTCAACATGTGAAATCTTCTCAGGGATACACGTGTGCCTAGATCCTCATTGCTTAGTTTACAGAGGAATGAATATAAAAAGA
Chimpanzee	TTAAAATATTAAAGACGTCTTCCCAGGATTCAACACGTGAAATCTTCTCAGGGATACACGTGTGCCTAGATCCTCATTGCTTAGTTTACAGAGGAATGAGCATACAAAGA
Gorilla	TTAAAATATTAAAGACGTCTTCCCAGGATTCAACATGTGAAATCTTCTCAGGGATACAAGTGTGCCTAGATCCTCGTTGCTTAGTTTACAGAGGAATGAATATAAAAAGA
Human	AAATACTTAAATTTATCCCTTACCTCTATAATCATACATAGGCATAATTCTAACCTAGGCTCCAGATAGCCATAGAAGAACCAAACACTTTCTGCGTGTGAGAATAATCAGAG
Chimpanzee	AAATACTTAAATTTATCTCTTACCTCTATAATCATACACAGGCATAATTCTAACCTAGGCTCCAGATAGCCATAGAAGAACCAAACACTTTCTGCGTGTGAGAATAATCAGAG
Gorilla	AAATACTTAAATTCTCTTACGTCTATAATCGTACATAGGCATAATTCTAACCTAGGCTCCAGATAGCCATACAAGAACCAAACACTTTCTGCGTGTGAGAATAATCAGAG
Human	TGAGATTTTCACAAGTACCTGATGAGGGTTGAGACAGGTAGAAAAGTGGAGAGATCTCTATTATTTAGCAATAATAGAGAAAGCATTAAAGAGAATAAGCAATGGAATAAGAAAAT
Chimpanzee	TGAGATTTTCACAAGTACCTGATGAGGGTTGAGACAGGTAGAAAAGTGGAGAGATCTCTATTATTTAGCAATAATAGAGAAAGTATTAAAGAGAATAAGCAATGGAATAAGAAAAT
Gorilla	TGAGATTTTCACAAGTACCTGATGAGGGTTGAGACAGGTAGAAAAGTGGAGAGATCTTATTATTTAGCAATAATAGAGAAAGTATTAAAGAGAATAACAGCAATGGAATAAGAAAAT

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

(a) Pairwise alignment

Human	TTGATG TTTCTTTCCCCT-TCTTTCTATGGTTAA-GTTCAT	11:5247749
Mouse	TCCGG TTTCTTCCCCTGGCTATTCT-GGCTAACCCTCCT	7:103813380

1 2 3 4

ancestral T, G or other?

insertion in mouse or deletion in human?

(b) Ancestor alignment

Human	TTGATG TTTCTTTCCCCT-TCTTTCTATGGTTAA-GTTCAT	11:5247749
Ancestral	TTGATG TTTCTTTCCCCT-TCTTTCTATGGTTAA-GTTCAT	619_196519:74
Mouse	TCCGG TTTCTTCCCCTGGCTATTCT-GGCTAACCCTCCT	7:103813380

1 2 3 4

ancestral T

insertion in mouse

deletion in mouse

insertion in mouse

(c) Multiple sequence alignment

Human	TTGATG TTTCTTTCCCCT-TCTTTCTATGGTTAA-GTTCAT	11:5247749
Gorilla	TTGATG TTTCTTTCCCCT-TCTTTCTATGGTTAA-GTTCAT	11:5182738
Rabbit	TTGAT- GTTCTTTC---T-TTTTCGCTATTGTAAA-ATTCAT	1:146237759
Mouse	TCCGG TTTCTTCCCCTGGCTATTCT-GGCTAACCCTCCT	7:103813380
Horse	TCAAT- TCTCCTTGCCT-TCCTCTTTGGTCAA-GCTCAT	7:73937279
Dog	TCAACA TCTCTTGTACT-TCCTTTTAAGACCCA-ACTCAT	21:28179781

(d) Multiple sequence ancestor alignment

Human	TTGATG TTTCTTTCCCCT-TCTTTCTATGGTTAA-GTTCAT	11:5247749
Ancestral	TTGATG TTTCTTTCCCCT-TCTTTCTATGGTTAA-GTTCAT	619_196521:74
Gorilla	TTGATG TTTCTTTCCCCT-TCTTTCTATGGTTAA-GTTCAT	11:5182738
Ancestral	TTGATG TTTCTTTCCCCT-TCTTTCTATGGTTAA-GTTCAT	619_196519:74
Ancestral	TTGATG TTTCTTTCCCCT-TCTTTCTATGGTCAA-GTTCAT	619_196525:75
Rabbit	TTGAT- GTTCTTTC---T-TTTTCGCTATTGTAAA-ATTCAT	1:146237759
Ancestral	TCAATG TTTCTTTCCCCT-TCTTTCTATGGTCAA-GTTCAT	619_196527:75
Mouse	TCCGG TTTCTTCCCCTGGCTATTCT-GGCTAACCCTCCT	7:103813380
Ancestral	TCAATG TTTCTTTCCCCT-TCTTTCTATGGTCAA-GTTCAT	619_196523:75
Ancestral	TCAATA TCTCTTT-----TTATGGTCAA-GCTTGT	619_196526:63
Ancestral	TCAATG TCTCTTTCTCCT-TCTTTTTATGGTCAA-GCTCGT	619_196524:74
Ancestral	TCAATG TCTCTTTCCCCT-TCTTTTTATGGTCAA-GCTCAT	619_196522:74
Horse	TCAAT- TCTCCTTGCCT-TCCTCTTTGGTCAA-GCTCAT	7:73937279
Ancestral	TCAATG TCTCTTTCCCCT-TCTTTTTATGGTCAA-GCTCAT	619_196520:74
Dog	TCAACA TCTCTTGTACT-TCCTTTTAAGACCCA-ACTCAT	21: 28179781

inference of nested insertion and deletion events
NOT REACTION

Ensembl EPO is a phylogeny based prediction pipeline

How to view Ancestral sequence in Ensembl?

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 associated 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	View

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*):

Alignathon

Project Description

FAQ

Analysis Details

Participants

Simulation Recipes

Datasets ▾

Downloads

Alignathon

A collaborative competition to assess the state of the art in whole genome sequence alignment.

<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