B.Sc. (Hons.) Biotechnology
Core Course 13:
Basics of Bioinformatics and Biostatistics (BIOT 3013)

Unit 5:
Multiple sequence alignment

Dr. Satarudra Prakash Singh
Department of Biotechnology
Mahatma Gandhi Central University, Motihari

Multiple sequence alignment (MSA): Why?

- Pair-wise alignment can concluded that there is probably a functional relationship between the two sequences.
- If it is known that there is a functional similarity amongst a number of sequences, then we can use MSA to find out where the similarity comes from.
- It extract biologically important information (widely dispersed sequence similarities) that can give biologist hints about the evolutionary history of certain sequences.

Why we do multiple alignments?

- Active site residues are under evolutionary pressure to maintain their functional integrity and undergo very fewer mutations than less functionally important amino acids.
- MSA is used to study closely related genes or proteins in order to find the evolutionary relationships between genes.
- It identify shared patterns among functionally or structurally related genes.
- It is used characterize protein families and determine the consensus sequence of several aligned sequences.

Example of multiple alignment

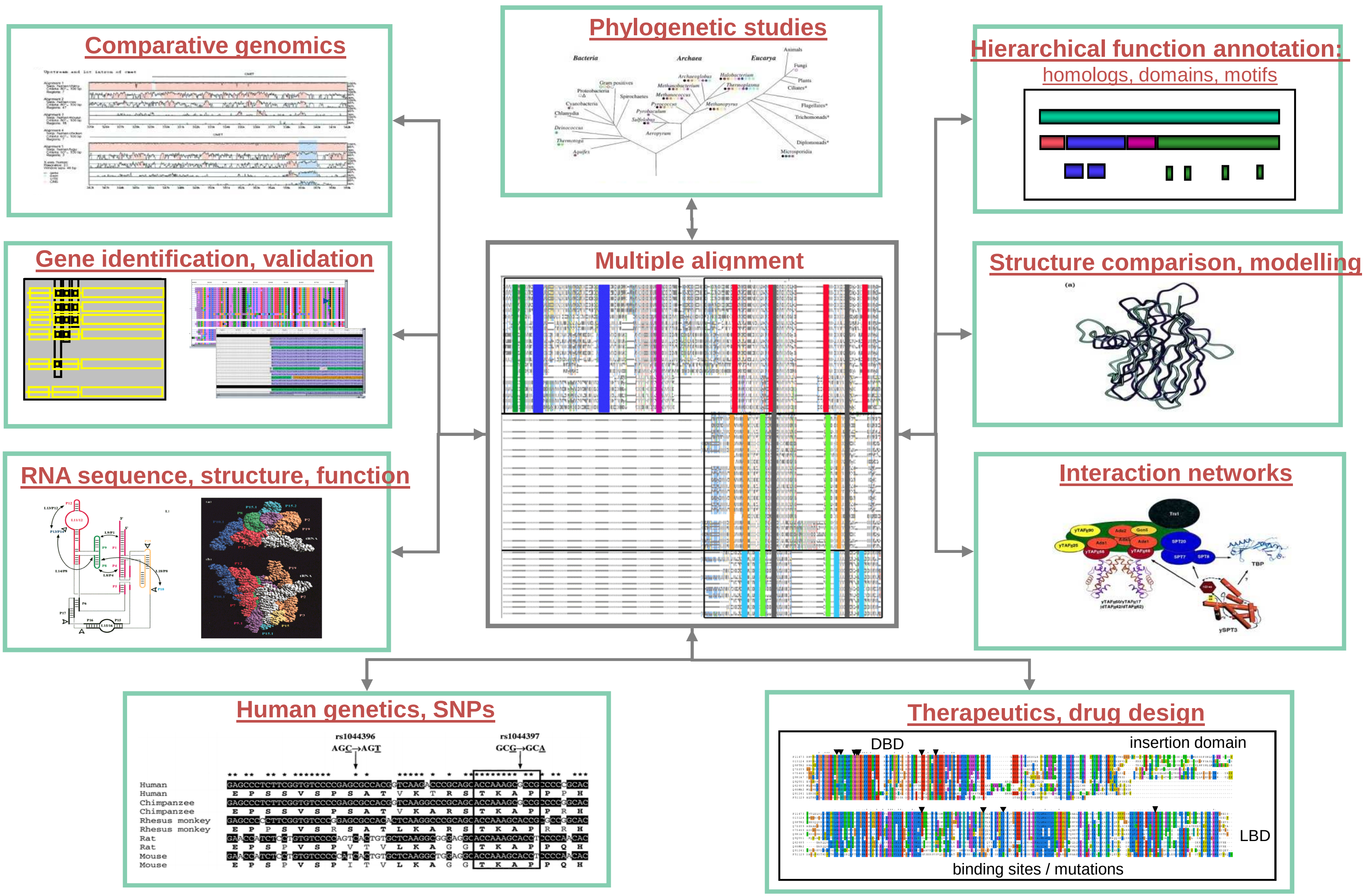
Example: part of an alignment of SH2 domains from 14 sequences

```
                          *      *      :        * **:*                          *  :                :  :  ::
lnk_rat      -----------YPWFHGPISRVRAAQLVQLQGPDAHGVFLVRQSESRR-GEYVLTFNLQ-----GRAKHLRLVLTERGQCRVQH--LHFPSVVDML
crk1_mouse   -----------SAWYMGPVTRQEAQTRLQGQR---HGMFLVRDSSTCP-GDYVLSVSEN-----SRVSHYIINSLPNRRFKIGD--QEFDHLPALL
nck_human    -------------WYYGKVTRHQAEMALNERGH--EGDFLIRDSESSP-NDFSVSLKAQ-----GKNKHFKVQLK-ETVYCIGQ--RKFSTMEELV
ht16_hydat   -------------WYHGKITREVAVQVLLRKGGR-DGFFLIRDCGNAP-EDYVLSMMFR-----SQILHFQINCLGDNKFSIDNG-PIFQGLDMLI
pip5_human   -----------KPWYYDSLSRGEAEDMLMRIPR--DGAFLIRKREGS--DSYAITFRAR-----GKVKHCRINRDG-RHFVLGTS-AYFESLVELV
fer_human    -------------WYHGAIPRIEAQELLKK-----QGDFLVRESHGKP-GEYVLSVYSD-----GQRRHFIIQYV-DNMYRFEG--TGFSNIPQLI
1ab2         -----------EEWFHGVLPREEVVRLLNN-----DGDFLVRETIRNEESQIVLSVCW------NGHKHFIVQTTGEGNFRFEG--PPFASIQELI
1mil         -----------HSWYHGPVSRNAAEYLLSSGI---NGSFLVRESESSP-GQRSISLRYE-----GRVYHYRINTASDGKLYVSSE-SRFNTLAELV
1blj         -----------EPWFHGKLSRREAEALLQL-----NGDFLVRESTTTP-GQYVLTGLQS-----GQPKHLLLVDP-EGVVRTKD--HRFESVSHLI
1shd         GSVAPVETLEVEKWFFRTISRKDAERQLLAPMNK-AGSFLIRESESNK-GAFSLSVKDITTQ-GEVVKHYKIRSLDNGGYYISPR-ITFPTLQALV
1lkkA        -------SIQAEEWYFGKITRRESERLLLNAENP-RGTFLVRESEA-----YCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSR-TQFNSLQQLV
1csy         --------LEPEPWFFKNLSRKDAERQLLAPGNT-HGSFLIRESESTA-GSFSLSVRDFDQNQGEVVKHYKIRNLDNGGFYISPR-ITFPGLHELV
1bfi         -------SHEKMPWFHGKISREESEQIVLIGSKT-NGKFLIRARDNN--GSYALCLLHE-----GKVLHYRIDKDKTGKLSIPEG-KKFDTLWQLV
1gri         -------HHDEKTWNVGSSNRNKAENLLRGKR---DGTFLVRESSKQ--GCYACSVVVD-----GEVKHCVINKTATG-YGFAEPYNLYSSLKELV
             -------EMKPHPWFFGKIPRAKAEEMLSKQRH--DGAFLIRESESAP-GDFSLSVKFG-----NDVQHFKVLRDGAGKYFLWV--VKFNSLNELV
```

* conserved identical residues

: conserved similar residues

Central role of multiple alignments

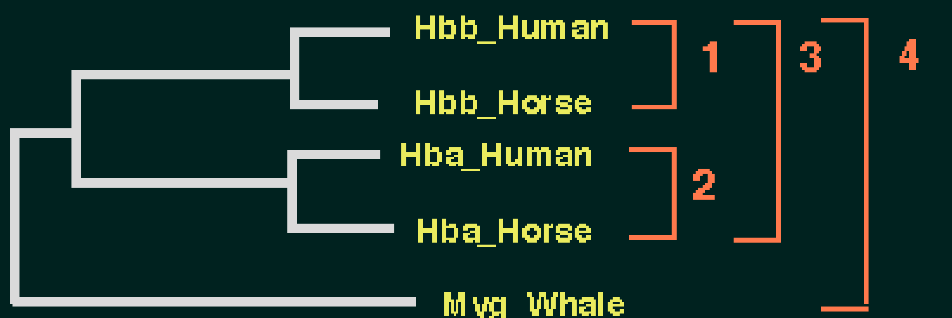


Multiple Alignment Method

- Compare all sequences pair-wise.
- Perform cluster analysis on the pair-wise data to generate a hierarchy for alignment.
- This may be in the form of a binary tree or a simple ordering.
- Build the multiple alignment by first aligning the most similar pair of sequences, then the next most similar pair and so on.
- Once an alignment of two sequences has been made, then this is fixed.

Overview of ClustalW Procedure

Hbb_Human	1	-				
Hbb_Horse	2	.17	-			
Hba_Human	3	.59	.60	-		
Hba_Horse	4	.59	.59	.13	-	
Myg_Whale	5	.77	.77	.75	.75	-



alpha-helices

```
1  PEEKSAVTALWGKVN--VDEVGG
2  GEEKAAVLALWDKVN--EEEVGG
3  PADKTNVKAAWGKVGAHAGEYGA
4  AADKTNVKAAWSKVGGHAGEYGA
5  EHEWQLVLHVWAKVEADVAGHGQ
```

1 2 3 4

CLUSTAL W

Quick pairwise alignment: calculate distance matrix

↓

Neighbor-joining tree (guide tree)

↓

Progressive alignment following guide tree

Most popular MSA tool: Clustalw (http://www.ebi.ac.uk/clustalw)

- Both progressive global and local alignments can be done in ClustalW.

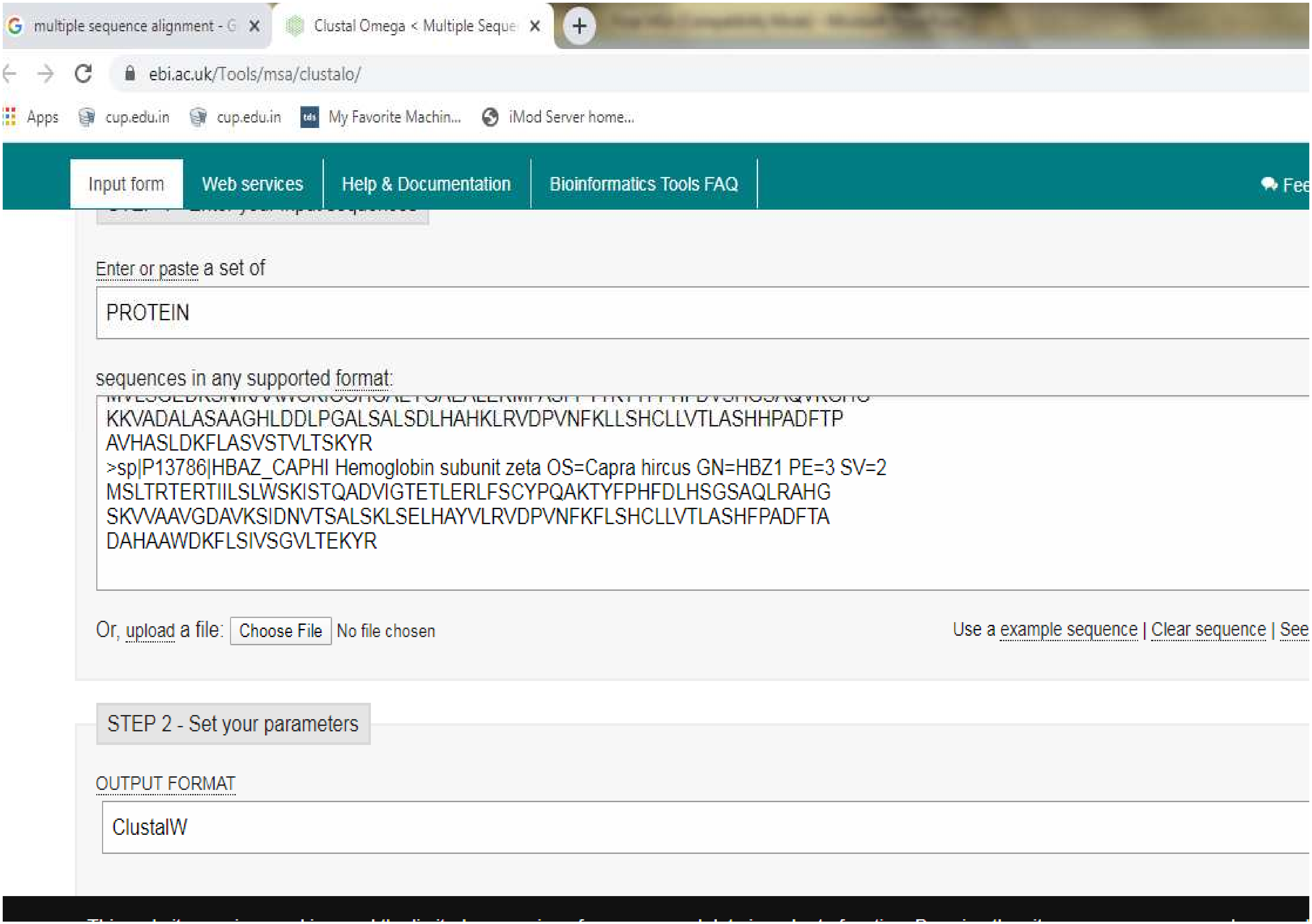
- The user has the option to control parameters to make the best alignments (e.g., word size, matrix, gap open, extension, etc.).

Clustalw/ClustalO

- It also provides two phylogenetic trees, a cladogram (equal length of branched tree showing common ancestry) or a phylogram (unequal length of branched tree showing evolutionary distances).

- Alignment can be further edited using the Jalview program (http://www.ebi.ac.uk/jalview).

- The main challenges for MSA is to handle growing data set sizes of nucleic acid and proteins.


multiple sequence alignment - G
Clustal Omega < Multiple Seque
ebi.ac.uk/Tools/msa/clustalo/
Apps
cup.edu.in
cup.edu.in
My Favorite Machin...
iMod Server home...
Input form
Web services
Help & Documentation
Bioinformatics Tools FAQ
Enter or paste a set of
PROTEIN
sequences in any supported format:
KKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTP
AVHASLDKFLASVSTVLTSKYR
>sp|P13786|HBAZ_CAPHI Hemoglobin subunit zeta OS=Capra hircus GN=HBZ1 PE=3 SV=2
MSLTRTERTIILSLWSKISTQADVIGTETLERLFSCYPQAKTYFPHFDLHSGSAQLRAHG
SKVVAAVGDAVKSIDNVTSALSKLSELHAYVLRVDPVNFKFLSHCLLVTLASHFPADFTA
DAHAAWDKFLSIVSGVLTEKYR
Or, upload a file: Choose File No file chosen
Use a example sequence | Clear sequence | See
STEP 2 - Set your parameters
OUTPUT FORMAT
ClustalW

Results for job clustalo-I20200418-095050-0522-88062448-p2m


```
CLUSTAL O(1.2.4) multiple sequence alignment


sp|P69905|HBA_HUMAN      MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
sp|P01942|HBA_MOUSE      MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG
sp|P13786|HBAZ_CAPHI     MSLTRTERTIILSLWSKISTQADVIGTETLERLFSCYPQAKTYFPHFDLHSGSAQLRAHG
                         * *:  ::: : : *.*:. :.   *:*:***:* .:* :********:  ****::.**

sp|P69905|HBA_HUMAN      KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
sp|P01942|HBA_MOUSE      KKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTP
sp|P13786|HBAZ_CAPHI     SKVVAAVGDAVKSIDNVTSALSKLSELHAYVLRVDPVNFKFLSHCLLVTLASHFPADFTA
                         .**. *: .*.  :*:: .*** **:***: *********:**********:* **:**

sp|P69905|HBA_HUMAN      AVHASLDKFLASVSTVLTSKYR
sp|P01942|HBA_MOUSE      AVHASLDKFLASVSTVLTSKYR
sp|P13786|HBAZ_CAPHI     DAHAAWDKFLSIVSGVLTEKYR
                          .**: ****: ** ***.***
```

PLEASE NOTE: Showing colors on large alignments is slow.

Scoring of MSA :Entropy score

- Define frequencies for the occurrence of each letter in each column of multiple alignment
  - $p_A = 1$, $p_T=p_G=p_C=0$ (1st column)
  - $p_A = 0.75$, $p_T = 0.25$, $p_G=p_C=0$ (2nd column)
  - $p_A = 0.50$, $p_T = 0.25$, $p_C=0.25$ $p_G=0$ (3rd column)
- Compute entropy of each column

$$- \sum_{X=A,T,G,C} p_X \log p_X$$

AAA
AAA
AAT
ATC

Entropy: Example

$$entropy\begin{pmatrix} A \\ A \\ A \\ A \end{pmatrix} = 0$$

Best case

Worst case

$$entropy\begin{pmatrix} A \\ T \\ G \\ C \end{pmatrix} = -\sum \frac{1}{4}\log\frac{1}{4} = -4(\frac{1}{4} * -2) = 2$$

Entropy of an Alignment: Example

column entropy:

$$-(\ p_A \log p_A + p_C \log p_C + p_G \log p_G + p_T \log p_T)$$

A	A	A
A	C	C
A	C	G
A	C	T

- Column 1 = $-[1*\log(1) + 0*\log 0 + 0*\log 0 + 0*\log 0]$
  $= 0$
- Column 2 = $-[(^1/_4)*\log(^1/_4) + (^3/_4)*\log(^3/_4) + 0*\log 0 + 0*\log 0]$
  $= -[\ (^1/_4)*(-2) + (^3/_4)*(-.415)\ ] = +0.811$
- Column 3 = $-[(^1/_4)*\log(^1/_4)+(^1/_4)*\log(^1/_4)+(^1/_4)*\log(^1/_4) +(^1/_4)*\log(^1/_4)]$
  $= 4* -[(^1/_4)*(-2)] = +2.0$
- Alignment Entropy = 0 + 0.811 + 2.0 = +2.811

References

- https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/download/lectures/PCB_Lect05_Multip_Align.pdf
- https://www.genome.jp/tools-bin/clustalw
- https://www.ebi.ac.uk/seqdb/confluence/display/JDSAT/Clustal+Omega+Help+and+Documentation
- https://academic.oup.com/bib/article/17/6/1009/2606431

Thank you.

Email: sprakashsingh@mgcub.ac.in