BINF 630-Bioinformatics Methods

Homework Assignment 1

Momina Tariq

4/7/2016

1. Write a regular expression that describes the alignment in the box.

```
QIQAAKIWAAKPYVDESRISIWGWSYGGF
QIAAAKHWAQKDYIDEDRLAIWGWSYGGY
QIQAAKAWGKKPYVDKTRMAIWGWSYGGF
QIEATRQFSKMGFVDDKRIAIWGWSYGGY
QIEAARQFLKMGFVDSKRVAIWGWSYGGY
QVFAAKELLKNRWADKDHIGIWGWSYGGF
QVFAAKEVLKNRWADKDHIGIWGXSYGGF
QVFAAKELLKNRWADKDHIGIWGWSYGGF
QVFAAKELLKNRWADKDHIGIWGWSYGGF
VGSASVSMMPRLPRLPQLLDQPGSSSGGY
FIAAAEYLKAEGYTRTDRLAIRGGSNGGL
FQCAAEYLIKEGYTSPKRLTINGGSNGGL
FQCAAEYLIKEGYTTSKRLTINGGSNGGL
FIAAGEYLQKNGYTSKDYMALSGRSNGGL
YLDACDALLKLGYGSPSLCYAMGGSAGGM
FIAAAKHLIDQNYTSPTKMAARGGSAGGL
QITAVRKFIEMGFIDEKRIAIWGWSYGGY
QLTAVRKFIEMGFIDEERIAIWGWSYGGY
```

REGULAR EXPRESSION: [QVFY]-X(2)-A-X(8)-[YFWP]-X(7)-[IQLA]-X-G-X-S-[YSNA]-G(2)-[FYLM]

Note: 'X' was chosen for a column with amino acids more than 4.

2. Find 5 protein sequences from different organisms that contain the pattern described by the regular expression from Q1. List the IDs, names, and function of found proteins

Tool Chosen: Scan Prosite
Parameters: Default

LIST OF PROTEINS:

i. **ID:** DPP4_MOUSE
Name: Dipeptidyl peptidase 4
Organism: Mus musculus (Mouse)
Function: Cell surface glycoprotein receptor involved in the costimulatory signal essential for T-cell receptor (TCR)-mediated T-cell activation.

ii. **ID:** DAP2_YEAST
Name: Dipeptidyl aminopeptidase B
Organism: Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)
Function: Catalysis of the hydrolysis of N-terminal dipeptides from a polypeptide chain.

iii. **ID:** DDAPB_ARTBC
Name: Probable dipeptidyl-aminopeptidase B
Organism: *Arthroderma benhamiae (strain ATCC MYA-4681 / CBS 112371) (Trichophyton mentagrophytes)*

Function: Type IV dipeptidyl-peptidase which removes N-terminal dipeptides sequentially from polypeptides having unsubstituted N-termini provided that the penultimate residue is proline.

iv. **ID:** DAPB_ASPTN
Name: Probable dipeptidyl-aminopeptidase B
Organism: *Aspergillus terreus (strain NIH 2624 / FGSC A1156)*
Function: Type IV dipeptidyl-peptidase which removes N-terminal dipeptides sequentially from polypeptides having unsubstituted N-termini provided that the penultimate residue is proline.

v. **ID:** SEPR_HUMAN
Name: Prolyl endopeptidase FAP
Organism: Homo sapiens (Human)
Function: Cell surface glycoprotein serine protease that participates in extracellular matrix degradation and involved in many cellular processes including tissue remodeling, fibrosis, wound healing, inflammation and tumor growth.

3. Find 2 protein structures from different organisms that contain the pattern described by the regular expression from Q1. List the IDs of found protein structures.

Database chosen: PDB
Motif Query: [QVFY]X{2}AX{8}[YFWP]X{7}[IQLA]XGXS[YSNA]G{2}[FYLM]
Parameters: Default

LIST OF PROTEIN STRUCTURES:

i. **ID:** 4L72
Name: Dipeptidyl peptidase 4
Organism: Homo sapiens
Function: Cell surface glycoprotein receptor involved in the costimulatory signal essential for T-cell receptor (TCR)-mediated T-cell activation

ii. **ID:** 4BCB
Name: Prolyl endopeptidase
Organism: Sus scrofa (Pig)
Function: Cleaves peptide bonds on the C-terminal side of prolyl residues within peptides that are up to approximately 30 amino acids long.

4. Build a multiple sequence alignment for all protein sequences from Q2 and Q3.

Tool Chosen: T-Coffee
Parameters: Default

```
T-COFFEE, Version_11.00.d625267 (2016-01-11 15:25:41 - Revision d625267 - Build 507)
Cedric Notredame
SCORE=818
*
 BAD AVG GOOD
*

sp|P28843|DPP4_     NTOWITWSPEGHKLAYVWKNDIYVKVEPHL-----PSH-RITSTGEENVIYNGITDWVYEEEV
sp|P18962|DAP2_     EVALAIWSPNSNDIAYVODNNIYIYSAISK-----KTIRAVTNDGSSF-LFNGKPDWVYEEEV
sp|D4AQT0|DAPB_     RIOLATWSPTSDAVAFTRDNNLYIRNLTSK-----SVK-AITTDGGTN-LFYGIPDWVYEEEV
sp|Q0CXB1|DAPB_     RVOLASWSPNSDAVVFVRDNNMFLRKLSSD-----KVV-PITKDGGKD-LFYGVPDWVYEEEV
sp|Q12884|SEPR_     PIOYLCWSPVGSKLAYVYONNIYLKORPGD-----PPF-OITFNGRENKIFNGIPDWVYEEEM
4L72_A|PDBID|CH     NTOWVTWSPVGHKLAYVWNNDIYVKIEPNL-----PSY-RITWTGKEDIIYNGITDWVYEEEV
4BCB_A|PDBID|CH     TVALRGYA-------FSEDGEYFAYGLSASGSDWVTIK-FMKVDGAKE-L----PD-VLERV-

cons                      ::       :  ..: :                 :.  *    :    .* * *.


sp|P28843|DPP4_     FGAYSALWWSPNNTFLAYAOF---------------------------ND---TGVPLIEYSF
sp|P18962|DAP2_     FEDDKAAWWSPTGDYLAFLKI---------------------------DE---SEVGEFIIPY
sp|D4AQT0|DAPB_     FEGNCATWWSLDGKYISYLRT---------------------------NE---TLVPEFPIDF
sp|Q0CXB1|DAPB_     LSGNSATWWSNDAKYVAFLRT---------------------------NE---STVPEYPVOY
sp|Q12884|SEPR_     LATKYALWWSPNGKFLAYAEF---------------------------ND---TDIPVIAYSY
4L72_A|PDBID|CH     FSAYSALWWSPNGTFLAYAOF---------------------------ND---TEVPLIEYSF
4BCB_A|PDBID|CH     --KFSCMAWTHDGKGMFYNAYPQQDGKSDGTETSTNLHQKLYYHVLGTDQSEDILCAEFPDEP

cons                     .  *:     : :                              ::


sp|P28843|DPP4_     YS-DE--------SLOYPKTVWIPYPKAGAVNPTVKFFIVNIDSLSSSSSAAPIOIPAPASVA
sp|P18962|DAP2_     YVODE--------KDIYPEMRSIKYPKSGTPNPHAELWVYSMKDG------TSFHPRISGNKK
sp|D4AQT0|DAPB_     YLSSPPGYSPKPNEESYPYVOOIKYPKAGAPNPTVNLOFYDVERE------ESFSVDVKDTLK
sp|Q0CXB1|DAPB_     FLSRPSGKKPLPGLEDYPDVROIKYPKAGAPNPVVNLOFYNVEKN------EVFSVEVPDDFA
sp|Q12884|SEPR_     YG-DE----------OYPRTINIPYPKAGAKNPVVRIFIIDTTYPAYVG---POEVPVPAMIA
4L72_A|PDBID|CH     YS-DE--------SLOYPKTVRVPYPKAGAVNPTVKFFVVNTDSLSSVTNATSIOITAPASML
4BCB_A|PDBID|CH     --------KWMGGAELSDDGRYVLLSIREGCDPVNRLWYCDLQQE------SNGITGILKWVK

cons                                      :  .     :*  .:   .


sp|P28843|DPP4_     RGDHYLCDVVW-ATEERISLOWLRRIONYSVMAICD-YDKINLTWNCPSEOQHVEMSTTGWVG
sp|P18962|DAP2_     DGSLLITEVTW-VGNGNVLVKTTDRSSDILTVFLID-TIA--KTSN-VVRN---ESSNGGWWE
sp|D4AQT0|DAPB_     DDDRLIVEVIP-GSKGKVLVRETNRESYIVKVAVID-ANK--REGK-IVRSDNIDEIDGGWVE
sp|Q0CXB1|DAPB_     DDDRIIIEVLW-AAESNVLVRATNRESDVLKIFLID-TES--RTGK-MVRLEDIVGLDGGWVE
sp|Q12884|SEPR_     SSDYYFSWLTW-VTDERVCLOWLKRVONVSVLSICD-FREDWOTWDCPKTOEHIEESRTGWAG
4L72_A|PDBID|CH     IGDHYLCDVTW-ATOERISLOWLRRIONYSVMDICD-YDESSGRWNCLVARQHIEMSTTGWVG
4BCB_A|PDBID|CH     LIDNFEGEYDYVTNEGTVFTFKTNRHSPNYRLINIDFTDPEESKWKVLVPE-H-EKDVLEWVA

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

5. Identify the conserved regions in the alignment from Q4 and explore their biological significance.

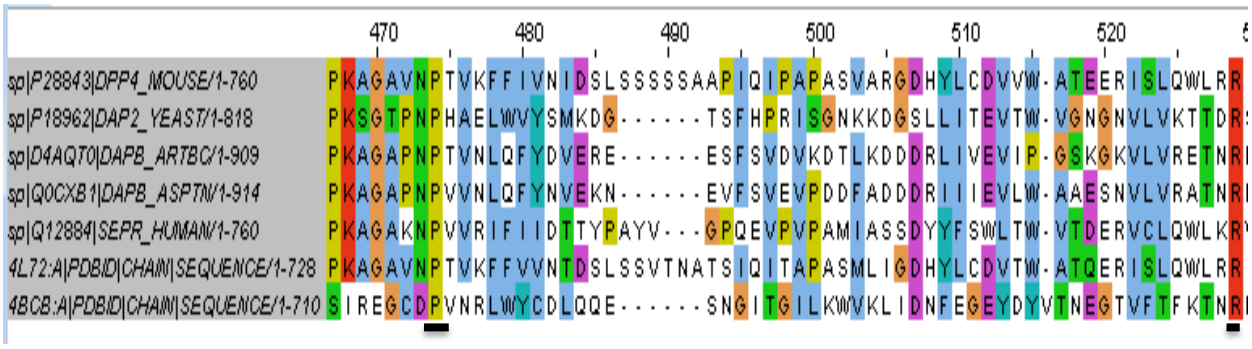
Tool Chosen: Jalview

Parameters: Web service was set to alignment with T-coffee with defaults.

This is an automatically calculated quantitative alignment annotation which measures the number of conserved physico-chemical properties conserved for each column of the alignment.

The most conserved columns in each group have the most intense colors, and the least conserved are the palest.

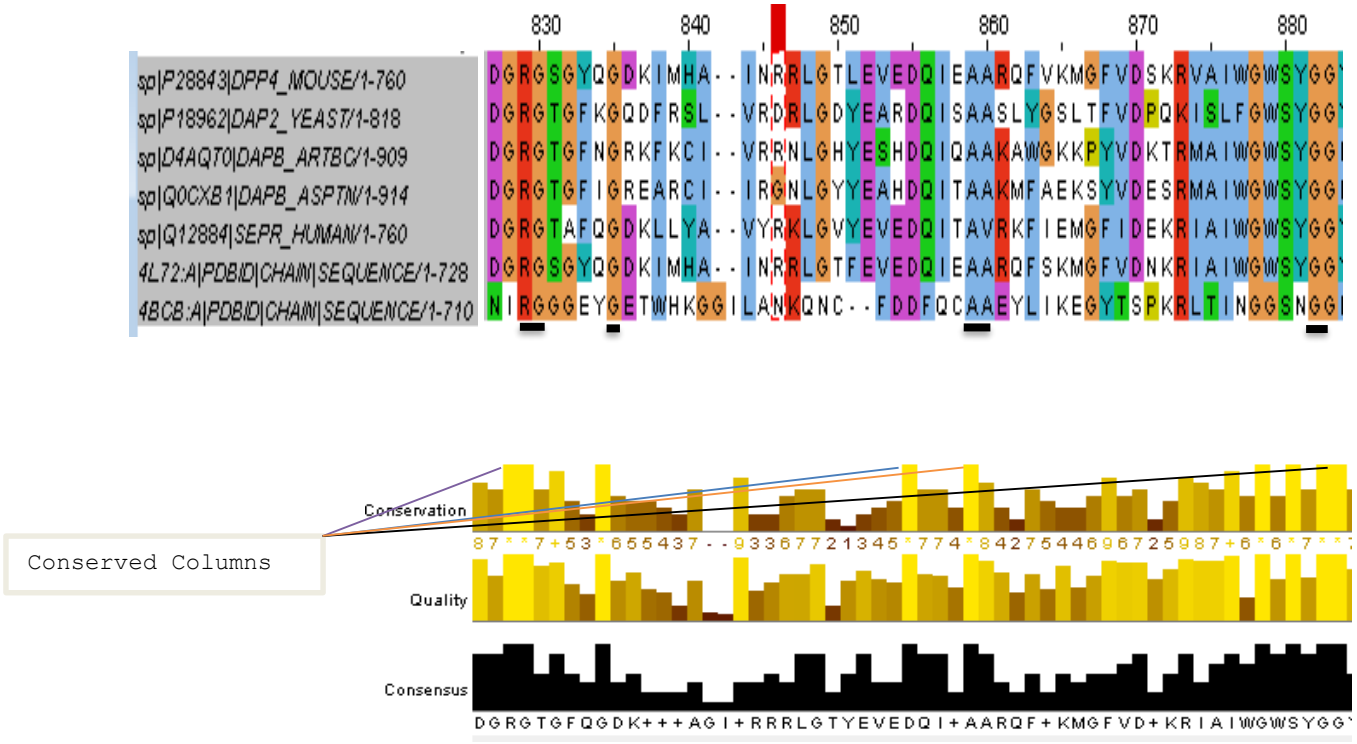
Conservation is visualized on the alignment or a sequence group as a histogram giving the score for each column. Conserved columns are indicated by '*'.



Conserved Columns

BIOLOGICAL SIGNIFICANCE: The figure above shows that the 7 sequences have conserved proline (P) and arginine (R) regions marked by '*'. Proline is a small, non-polar aliphatic amino acid. Proteins with proline rich region play a purely structural role. These regions are also involved in complex multiple protein association phenomenon.

A number of proteins containing arginine-rich motifs (ARMs) are known to bind RNA and are involved in regulating RNA processing in viruses and cells.



BIOLOGICAL SIGNIFICANCE: In this figure the sequences contains conserves regions of glycine (G), alanine (A) and arginine (R).

Glycine is a non- polar aliphatic amino acid. It is an essential component of important biological molecules, a key substance in many metabolic reactions. Glycine-rich regions are proposed to be involved in protein-protein interactions in some mammalian protein families. Glycine rich proteins exert important roles in very diverse processes such as signal transduction, stress response, transcriptional regulation and development.

Alanine rich regions in bacteria are shown to be important for translational stability and translocation.

6. Evaluate statistical parameters of the regular expression from Q1 based on similar expressions in the Prosite database.

A regular expression is qualitative; it either does match or does not. There is no threshold above for we consider the match as statistically significant. However, it is possible to evaluate the accuracy of PROSITE patterns thanks to the statistics on the number of hits obtained while scanning the SWISSPROT database.

A normalised score of 8.5 is typically defined as the default cut-off value in PROSITE profiles. However, in some cases this threshold is not suitable or appropriate. There are profiles producing statistically significant matches to members of structurally related protein families, for which a higher cut-off value is indicated. Contrariwise, for short structural repeats it is sometimes necessary to choose a lower cut-off value to reach suitable sensitivity. Such decisions are always based on the match lists for SWISSPROT presented as quality control information in each entry.

Usually, a second low cut-off level with a normalised threshold score of 6.5 is defined for weak matches, which must be interpreted with caution. Nevertheless, they can be very useful for gene discovery and the detection of remote homologues.

Also, every new protein entering SWISS-PROT is checked for the occurrence of PROSITE patterns and profiles and a match status for the relevant PROSITE entries is evaluated. At every new PROSITE release, these SWISSPROT match statuses are used to establish statistics for most motifs. These statistics allow the user to assess the ability of a motif to detect all or most of the sequences it is designed to describe (sensitivity) as well as its ability to give as few false positive results as possible (specificity). In addition, this process allows motifs to be permanently improved to give a better fit to the increasing number of proteins in SWISS-PROT.

- Number of hits found by the pattern **[QVFY]-X(2)-A-X(8)-[YFWP]-X(7)-[IQLA]-X-G-X-S-[YSNA]-G(2)-[FYLM]=** 47
- Approximate number of expected random matches: 2.182971e-02

REFERENCES:

- Bucher, P., Karplus, K., Moeri, N. and Hofmann, K. (1996), 'A flexible motif search technique based on generalized profiles', Comput. Chem., Vol. 20(1), pp. 3–23. 7.
- Gribskov, M., McLachlan, A. D. and Eisenberg, D. (1987), 'Profile analysis: detection of distantly related proteins', Proc. Natl Acad. Sci. USA, Vol. 84(13), pp. 4355–4358. 8.
- Gribskov, M., Lu¨thy, R. and Eisenberg, D. (1990), 'Profile analysis', Methods Enzymol., Vol. 183, pp. 146–159. 9.
- Luthy, R., Xenarios, I. and Bucher, P. (1994), 'Improving the sensitivity of the sequence profile method', Protein Sci., Vol. 3(1), pp. 139–146. 10.
- Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994), 'Improved sensitivity of profile searches through the use of sequence weights and gap excision', Comput. Appl. Biosci., Vol. 10(1), pp. 19–29.
- Crowley, P. J., L. J. Brady, D. A. Piacentini, and A. S. Bleiweis. 1993. Identification of a salivary agglutinin-binding domain within cell surface adhesin P1 of *Streptococcus mutans*. Infect. Immun. 61:1547-1552.