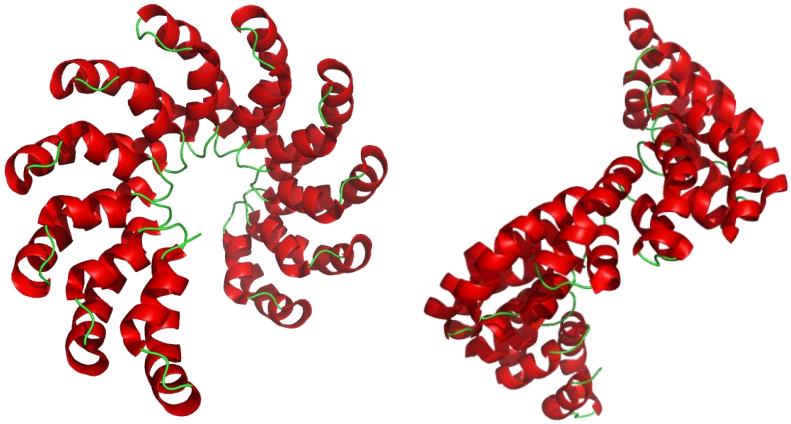
WEB-BASED SKILL

2018-2019



Topological and ribbon representations of repeat or solenoid proteins. The picture shows the ribbon diagram of the crystal structure of the TAL effector DNA-binding domain of dHax3 in two orientations. (PDB entry: 3V6P).

Name: →

Student ID Number: →

WEB-BASED SKILL

ON-LINE ANALYSIS OF PROTEIN STRUCTURE AND FUNCTION.

For this Web-Based Skills you should consult the lecture notes concerning protein structure and the Chapter in the accompanying text book on "Exploring Evolution and Bioinformatics".

These websites will also be used in this work and the Workshops:

1). http://www.uniprot.org/ (**UniProt**)

2). http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/ (**PDBSum**)

3). http://www.uniprot.org/align/ (**UniProt - Sequence Alignment**)

These may assist later:

4). https://www.ebi.ac.uk/pdbe/ (**Protein Data Bank - EBI**)

5). http://pymol.org/educational/ (**PyMol – Molecular Graphics Package**)

NOTE Symbols like this (1) next to the text refer to the **HELP POWERPOINT** that accompanies this Web-Based Skills 8 (in **PDF** format).

Proteins are the macromolecules that carry out numerous functions in the body and their sequences are unique to their roles. You should recall from the lectures that this is because their sequence dictates their three dimensional structure. You should also remember that some proteins start out as having a longer sequence than they eventually "use" as their native conformation. This can be for a number of reasons as given in the lecture notes.

In this Web-Based Skill study and the accompanying Workshops we are going to introduce you to a number of specialised sites (given above) that handle information about proteins and other macromolecules. Combined these sites provide an enormous amount of data and as you have been told, this has given rise to specialist research that utilises this data; bioinformatics.

You have already looked at DNA sequence comparisons in an earlier Web-Based Skill in this series. That was working with the DNA, but the information in this molecule is first transcribed into RNA and then translated into the protein sequence and hence into the molecules that the body ultimately uses.

In the lectures we considered homology and in particular we discussed the difference between orthologs and paralogs. Today you will explore something of these differences further.

PREPARATION REFLECTION
(Indicate four relevant issues you have acquired by undertaking this preparation).

- → I am looking forward to making use of bioinformatic databases to acquire information about proteins and nucleic acids.

- → I am eager to identify protein sequences and study it for functional attributes and homology with other proteins.

- → I would like to know what other information is available apart from structural and functional characteristics.

- → I would like to find out how to identify the effects of mutations on protein structure and functional properties.

Identify IN YOUR OWN WORDS* what is meant by the following terms:

- **HOMOLOGY**

→ Homology refers to a certain amount of similarity between two protein sequences such that the important functional amino acid residues between the two proteins are conserved in both the protein sequences, to impart same or similar functions.

- **ORTHOLOG**

→ Orthologs refer to proteins performing the same function in two different species. For example, a protein that performs a specific function in a mouse and another protein that performs the same function in a rat are orthologs.

- **PARALOG**

→ Paralogs are two proteins within the same species that perform similar functions. For example, haemoglobin and myoglobin are two proteins found in a number of species that perform similar functions in different locations. These arise due to gene duplication events giving rise to genes encoding similar proteins within the same organism.

*__NOT__ from Wikipedia, or the internet!! Take the material from the lectures and accompanying text book then actually put this information into your **OWN WORDS**

The characteristics of the residues within a protein work to create the Tertiary structure of that protein. As a result some residues that are far from each other in the sequence become close together in their relative three-dimensional positions and in turn these can be used to create the function of a protein. This means that there are some "significant" residues in a protein (for example a small glycine residue in Collagen every third position) that if mutated by mistake can cause the protein to mis-fold, or maybe lose catalytic activity and hence the protein might dysfunction. These are the residues that have a greater importance overall and which therefore have to be conserved between different species. A failure to do this could ultimately mean that the organism might not survive. The homologous regions in a protein show where these

important residues could be found. This is easier to see when comparing sequences between species that are NOT very close in evolutionary terms. Our protein sequences are very close to those of our primate cousins the Gorilla and Chimpanzee for example, but we are more distant from a Chicken (most of us!!), so the relative homology between proteins from these two species will more accurately show regions important to preserve structure or function or both.

1 Go to site **1)** above (**UniProt**) and in the query box (at the top – labelled UniProtKB) type in **P61626**. This is the "Accession Code" for one of the proteins you will be working with. Find the name of this protein (Bold letters beside "Protein") and also locate and paste in the box below the sequence for this protein in "**FASTA**" format. (Note that there are boxes on the left side under "**Display**" and one of them is "**Sequence**". Click on that and it will take you to that section of the data file on this protein. There you will see the FASTA link to put the sequence in the box below in that format).

2

Identify what this protein is (the name beside "Protein") and paste in its sequence here in FASTA* Format:

- **NAME**

→ Lysozyme C

- **SEQUENCE (in FASTA format)**

```
>sp|P61626|LYSC_HUMAN Lysozyme C OS=Homo sapiens OX=9606 GN=LYZ PE=1 SV=1
MKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRA
TNYNAGDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRD
PQGIRAWVAWRNRCQNRDVRQYVQGCGV
```

*** LOOK <u>IN</u> THE <u>SEQUENCE</u> SECTION ➔ <u>FASTA</u>** (**blue** background)

From the lectures you know that Protein structures can be determined in various ways; from a solution structure using NMR Spectroscopy, and from the crystal structure, using Crystallography. The information for almost all of these determined structures is stored in the Protein Data Bank, the PDB (site **4)** above). In particular the atomic coordinates are stored there. Other websites (for example site **2)** above) are available which summarise this information, and combine it with data from other sources to enable more research to be undertaken on any given protein or structure.

You have been provided with a **PDB code** (four alpha numeric characters) for your own personal protein structure.

3 Go to site **2)** above (**PDBSum**) and in the query box put in that code.

4 In the MIDDLE OF THE PAGE for your specific protein there is a **boxed area** which contains information about the sequence of the protein (it will have a heading probably "Protein Chain") with a **link** inside to the **UniProt page** of your protein identified through its accession code (similar to the one given above). Click that link to get to the UniProt page for your protein.

Again in the box below name your protein and provide its sequence in FASTA format.

Identify what this protein is and paste in its sequence here in FASTA Format:

- **NAME**

→ DNA binding protein

- **SEQUENCE (in FASTA format)**

```
>3v6p:A
GKQALETVQRLLPVLCQAHGLTPQQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQALLPVL
CQAHGLTPEQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPQQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPQQVVAIA
SNGGGKQALETVQRLLPVLCQAHGLTPQQVVAIASNSGGKQALETVQRLLPVLCQAHGLTPQQVVAIASNGGGKQALETVQRL
LPVLCQAHGLTPQQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPEQV
VAIASHDGGKQALETVQRLLPVLCQAHGLTPQQVVAIASNG
```

5

Go to site **3**) above. Paste the two sequences that you have in FASTA format into the box, one below the other. This is for a **BLAST** (you have met this before) alignment of the two sequences. Run the alignment. Use "Print Screen" then "Crop" this to show this alignment result here.

ALIGNMENT RESULT BELOW HERE

Alignment

How to print an alignment in color

```
P61626 LYSC_HUMAN     1   -------MK----------------------------ALIVL-----------GL-----      9
3v6p:A                1   GKQALETVQRLLPVLCQAHGLTPQQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQV      60
                              ::                                ** .:          **

P61626 LYSC_HUMAN    10   -----------------VLLSVTVQGKVF-------------------------ERCEL      26
3v6p:A               61   VAIASHDGGKQALETVQALLPVLCQAHGLTPEQVVAIASNGGGKQALETVQRLLPVLCQA      120
                                        .** *  *.: :                              *:

P61626 LYSC_HUMAN    27   AR----TLKRLGMDG--YRGI----SLANWMCLAKWESGYNTRATNYNAGDRSTDYGIFQ      76
3v6p:A              121   HGLTPQQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPQQVVAIASNGGGKQALETVQR      180
                                :  :. :*   :.:     *   :* *:  :  :. *   *.*.:.:   : :

P61626 LYSC_HUMAN    77   INSRYWC-NDGKTPGAVNACH-----------------LSCSA--LLQDNI---------      107
3v6p:A              181   L-LPVLCQAHGLTPQQVVAIASNSGGKQALETVQRLLPVLCQAHGLTPQQVVAIASNGGG      239
                          :     *  .* **  * *                    : *.*  *  :::

P61626 LYSC_HUMAN   108   ---------ADAVACAKRVVRDPQGIRAWVAW--------------RNRCQNRDV-----      139
3v6p:A              240   KQALETVQRLLPVLCQAHGL-TPQQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQV      298
                                  * *  : :  ** : * .:                  ** :.:

P61626 LYSC_HUMAN   140   --------------------RQY---------------------------------VQG      145
3v6p:A              299   VAIASNGGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQA      358
                                                                                  *.

P61626 LYSC_HUMAN   146   CGV------------                                            148
3v6p:A              359   HGLTPQQVVAIASNG                                           373
                           *:
```

The aligned sequences will identify those residues conserved across species that are likely to be either important for protein structure or function, or both. Identical aligned residues are marked with an asterisk (*) while those "conservatively replaced" are marked with a colon (:) and those semi-conservatively replaced are marked with a simple full stop (.)

6 Identify the first **FIVE** residues that are **FULLY CONSERVED** (*) starting on the **SECOND PAIRED LINE** of the aligned sequences. Give them as their full name, three letter code and their position **for both proteins**. For example if it was an arginine then you could put the pair as:

Arginine Arg 42 [*For the **P61626** Protein*]
Arginine Arg 42 [*For **your** Protein*] **CAREFUL**: NOTE THERE MAY BE **GAPS**

FIVE RESIDUES:

- →Leucine Leu 11 [For the P61626 Protein]
  Leucine Leu 79 [For the 3V6P Protein]
- →Leucine Leu 12 [For the P61626 Protein]
  Leucine Leu 80 [For the 3V6P Protein]
- →Valine Val 14 [For the P61626 Protein]
  Valine Val 82 [For the 3V6P Protein]
- →Glutamine Gln 17 [For the P61626 Protein]
  Glutamine Gln 85 [For the 3V6P Protein]
- →Cysteine Cys 24 [For the P61626 Protein]
  Cysteine Cys 118 [For the 3V6P Protein]

In the Workshop that is linked to this Web-Based Skill you will explore the links between sequence and structure further.

In the box below give THREE comments on what you have learnt from this exercise.

THREE COMMENTS ON WHAT YOU HAVE LEARNT:

- → Through this exercise, I learnt how to use various important bioinformatic sites such as UniProt and PDBSum, to find out protein sequences, align different protein sequences and identify conserved residues between the sequences.

- → I realized that there may be differences in the positions of conserved residues between different proteins, but the combination of conserved residues imparting functional attributes may still be the same.

- → I was unable to find the UniProt link to my protein on the PDBSum page as mentioned in the instructions, and so I learnt how to use PDBSum to find out the FASTA sequence of the protein.

PERSONAL REFLECTION (Impact of this exercise on my graduate attributes)
→

This exercise has taught me how to identify protein sequences and align different protein sequences to find out regions of conserved amino acids. This exercise has a lot of implications for my graduate studies as it will help me study the relationships and the extent of homology between different proteins. Using the Align tool, I will be able to identify conserved functional groups across species. Also, I have acquired familiarity with important bioinformatic sites, using which I will be able to access and study sequential and structural information about proteins and nucleic acids.