|  |
| --- |
| BINF 630- Bioinformatics Methods |
| Assignment-2 |
|  |
| **Momina Tariq** |
| **5/5/2016** |

|  |
| --- |
|  |

**Select one of the protein sequences from the list. Predict secondary and three-dimensional structure for the selected sequence using an advanced secondary structure prediction algorithm and a homology modeling algorithm, respectively. Analyze the quality of your model using one of the structure validation or verification tools and identify structural classification of your model.**

**PROTEIN CHOSEN:** Q1N7D2

**TOOL CHOSEN:** Uniprot

**PROTEIN NAME:** Peptidase

**ORGANISM:** Sphingomonas sp. (strain SKA58)

**QUERY SEQUENCE:**

>tr|Q1N7D2|Q1N7D2\_SPHSS Peptidase OS=Sphingomonas sp. (strain SKA58) GN=SKA58\_11088 PE=4 SV=1

MCMTVRTMLLALGLSATPLLPAAAQAQAIAPDRLESSVRTLASDLFQGRAPGTIGEERTI

GYLVGRFEALGLEPGGPDGQWVQTVPLLHTQLGTPTQLAVTQGSATTPWTFGENVYVSTL

QPKDNVSIDKAPMVFVGYGVSAPERGWDDFKGVDLKGKVAVFLVNDPDFEAVKGEDAVGK

FGGKTMTYYGRWTYKFEEAARRGAIGALIVHDTPGAGYGWNVVKSPGGENYDLVRPADRL

TSLQLQGWIAGEAAKSLFARAGQDLAALRKKARSAASKPVELKGATFTASFPVTQDVVQS

ANVLARIPGAKRPDETIMYGAHWDAYGKGEPDAQGRIYRAGANDDALGIAGMFEIARAFK

AGPAPDRSILFAAWTAEERGLLGSEHYALNPVYPLDKTVANLTIDILQTAGKAKDVILVG

KGQNMLEDDLARFAAWQGRVVTQESLPERGLFYRADHFSMAKRGVPVLLMMGIAGASDLV

DGGRAAGQAWVDAYTGNCYHQACDAVDDSWNLDGAAQDIDLMLDIGRDLATSDRWPEWKP

GSEFKAIRDKSAGVRK

* **SECONDARY STRUCTURE PREDICTION:**

**-SERVER CHOSEN**: JPred (Advance secondary structure Algorithm)

**-PARAMETERS CHOSEN:**

Type of Input= Single Sequence (Raw/Fasta)

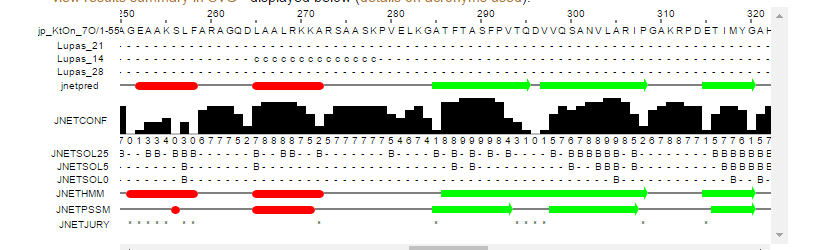


Figure 1 JPred result

**RESULT INTEPRETATION:** *Jpred* is a web server that takes a protein sequence or multiple alignments of protein sequences and from these predicts secondary structure using a neural network called Jnet. The prediction is the definition of each residue into alpha helix, beta sheet or random coil secondary structures.

The green arrows predict the beta strands, whereas the alpha helices are marked by red tubes. The output shows the character ‘B’ in prediction. Other characters such as ‘E’, ‘H’, and ‘\_’ can also appear. They represent extended (E), helical (H) and other (-) types of secondary structure respectively. In the solvent accessibility predictions they represent buried (B) and exposed (-) for each of the 0%, 5% and 25% solvent accessibility cut-offs.

* **3D STRUCTURE PREDICTION:**

**-SERVER CHOSEN**: SWISS-MODEL (homology-modeling server)

**-PARAMETERS CHOSEN:** Default

When a model was built using the query sequence, a total of 287 templates were found to match the target sequence. This list was filtered by a heuristic down to 50. The top templates are:

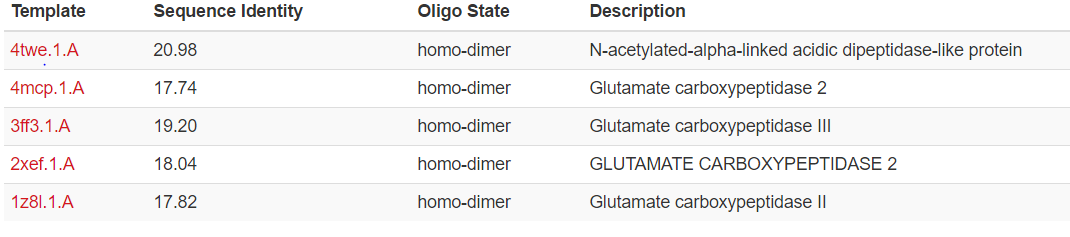


Figure 2 Top templates

The template with the highest sequence identity was chosen, which is 4twe.1.A. This template shows a sequence identity of 20.98, which means that its 20.98% identical to the query sequnce. Accoridng to the description it’s a N-acetylated-alpha-linked acidic dipeptidase-like protein.

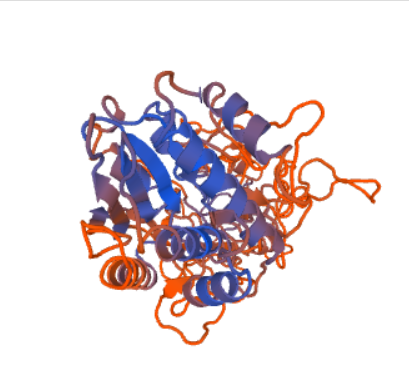


Figure 3 3D structure of N-acetylated-alpha-linked acidic dipeptidase-like protein acquired from SWISS- Model

**Model Results**

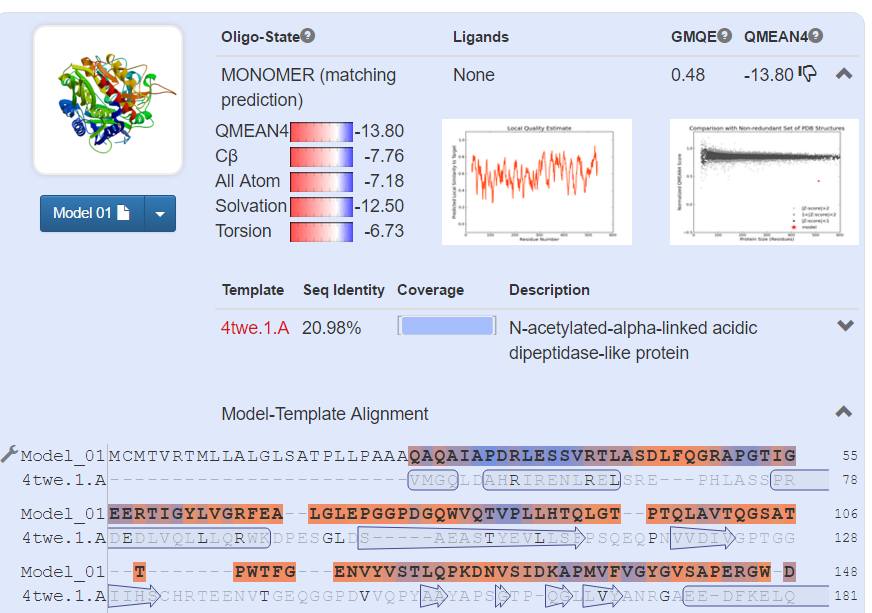
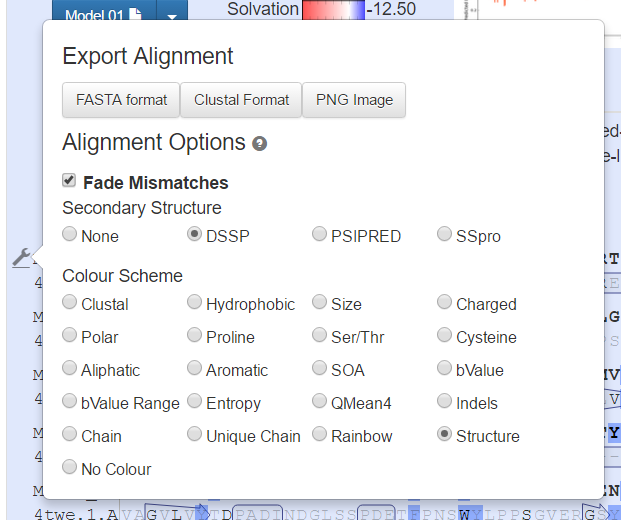




Figure 4 Swiss Model Result



By clicking on the tool symbol on the left most side, one can select the color scheme and the alogorthim to use for secondary structure prediction. Current Swiss- Model results are shown using DSSP algorithm and color scheme is set to be ‘structure”

* **ANALYSIS OF THE QUALITY OF PROTEIN MODEL:**

**-Analysis of the quality of my protein model was done using:** *Procheck*. Procheck was accessed through PDBsum.

The aim of PROCHECK is to assess how normal, or conversely how unusual, the geometry of the residues in a given protein structure is, as compared with stereochemical parameters derived from well-refined, high-resolution structures.



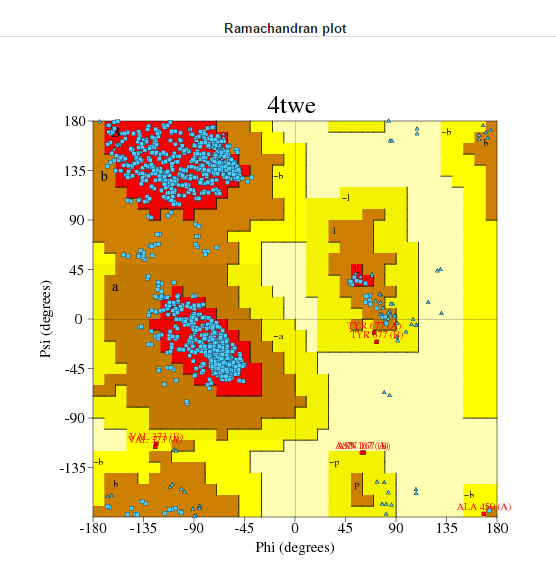
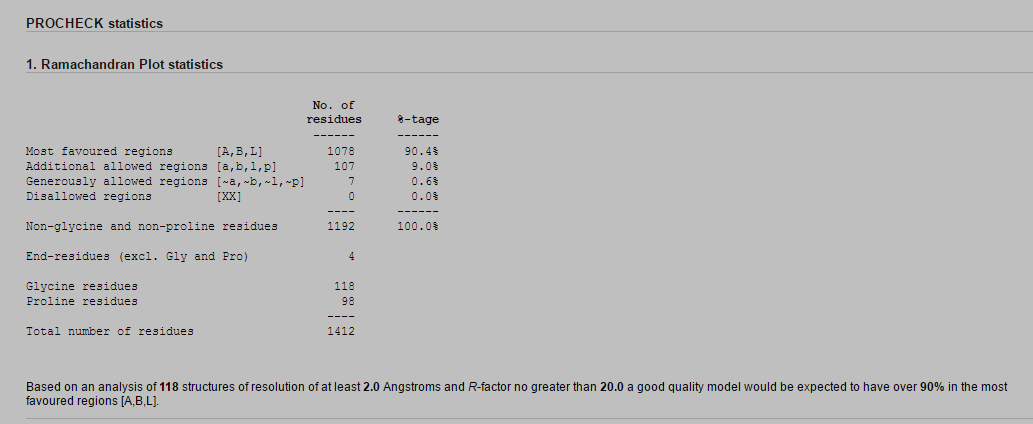


Figure 5 Ramachandran Plot for 4twe

Ramachandran plot is used for the validation of protein. Since few side chains are streically hindered, they come into Vander wall radii and cannot exist in nature.

There are different regions on the Ramachandran plot. The red color depicts the core region in which the most favorable phi and psi angles are possible. The yellow region shows the less possible confirmations and this is termed as the allowed region. Generous region is the green colored region in which a little lesser confirmations than the allowed region are possible. The grey color shows the non-allowed region where conformations are not allowed.

Ramachandran plot gives the most possible regions for the secondary structures. Thus, with the possible no.of phi psi angles a three-dimensional structure can be predicted.



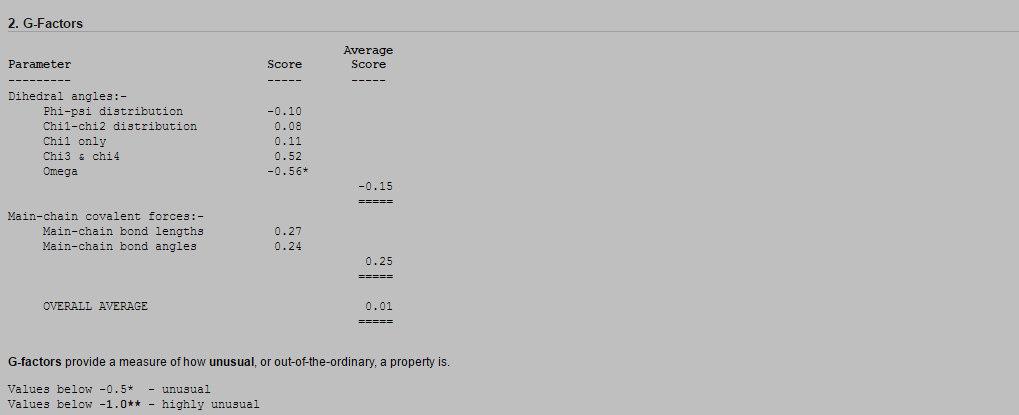


Figure 6: PROCHECK Statistics

* **STRUCTURAL CLASSIFICATION OF PROTEIN MODEL:**

**-DATABASE USED:** CATH

**-PARAMETERS CHOSEN:** Search was done using protein sequence in FASTA format.

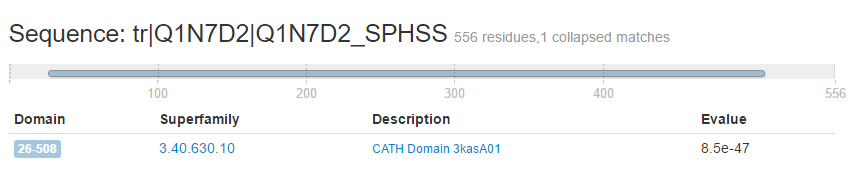


Figure 7 The following figure highlight regions of my query protein sequence that match structural domains in CATH.

The following figure highlights regions of my query protein sequence that match structural domains in CATH. The e-value is 8.5e-47. The Expect **value** (**E**) is a parameter that describes the number of hits one **can** "expect" to see by chance when searching a database of a particular size. It decreases exponentially as the Score (S) of the match increases. Essentially, the **E value** describes the random background noise. So, in this case the e-value is not that great because the lower the E-value, or the closer it is to zero, the more "significant" the match is.

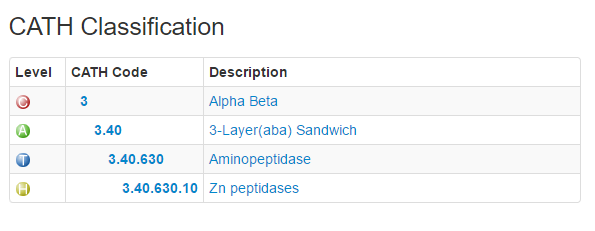


Figure 8 CATH Classification Table

According to the CATH classification table, the query protein belongs to the class Alpha-Beta. It belongs to the superfamily of Zn peptidases. The table below shows the superfamily summary of this protein.

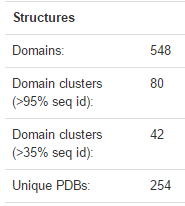
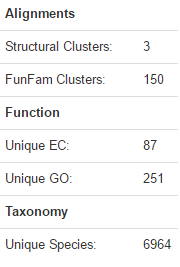


Figure 9 A general summary of information for this superfamily.

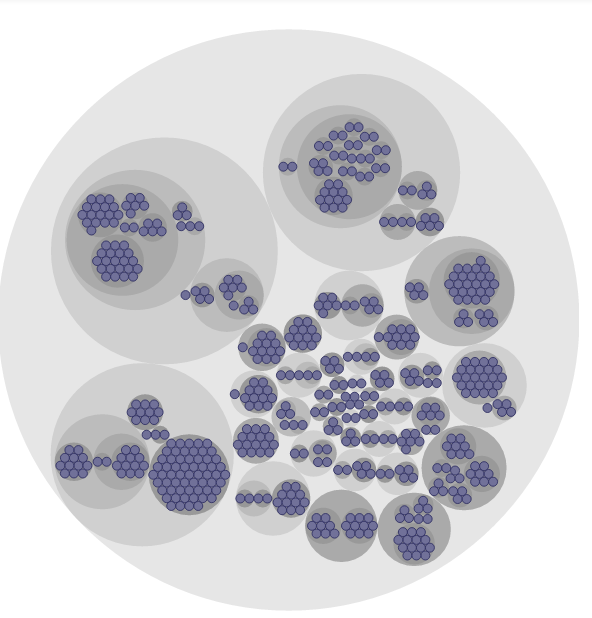


Figure 10 CATH Domains

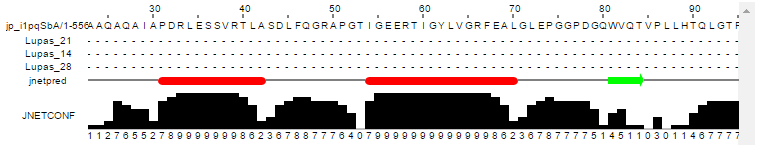
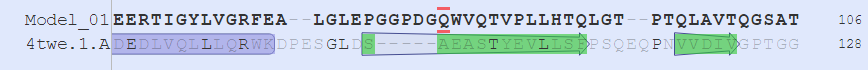
The following diagram provides an overview of the CATH structural domains within this superfamily. Domains have been grouped into S35, S60, S95, S100 clusters which reflect increasingly strict sequence identity cutoffs. For example, all domains grouped into the same S35 cluster are guaranteed to share at least 35% sequence identity.



* **COMPARISON OF SECONDARY STRUCTURES WITH THE SECONDARY STRUCTURES IN 3D MODEL:**

Secondary structure prediction is a set of techniques in [bioinformatics](https://en.wikipedia.org/wiki/Bioinformatics) that aim to predict the local [secondary structures](https://en.wikipedia.org/wiki/Secondary_structure) of [proteins](https://en.wikipedia.org/wiki/Protein) based only on knowledge of their [amino acid](https://en.wikipedia.org/wiki/Amino_acid) sequence. For proteins, a prediction consists of assigning regions of the amino acid sequence as likely [alpha helices](https://en.wikipedia.org/wiki/Alpha_helix), [beta strands](https://en.wikipedia.org/wiki/Beta_sheet) (often noted as "extended" conformations), or [turns](https://en.wikipedia.org/wiki/Turn_(biochemistry)). The success of a prediction is determined by comparing it to the results of the [DSSP](https://en.wikipedia.org/wiki/DSSP_(protein)) algorithm (or similar e.g. [STRIDE](https://en.wikipedia.org/wiki/STRIDE_(protein))) applied to the [crystal structure](https://en.wikipedia.org/wiki/X-ray_crystallography) of the protein. Specialized algorithms have been developed for the detection of specific well-defined patterns such as [transmembrane helices](https://en.wikipedia.org/wiki/Transmembrane_helix" \o "Transmembrane helix) and [coiled coils](https://en.wikipedia.org/wiki/Coiled_coil) in proteins.

Using the DSSP algorithm, it was found that my protein is 35% helical and has 16% of beta sheets.

**GLN 80;**

I took a portion of a protein to compare between its predicted secondary structure and 3D structure. The results seem to match together because an alpha helix that is predicted in the secondary structure is also a beta strand in its 3D structure.

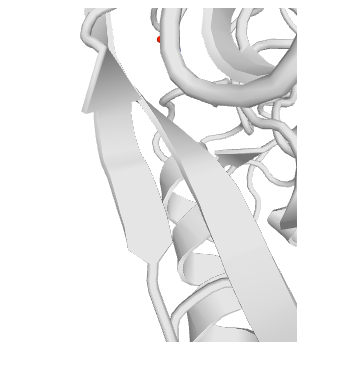
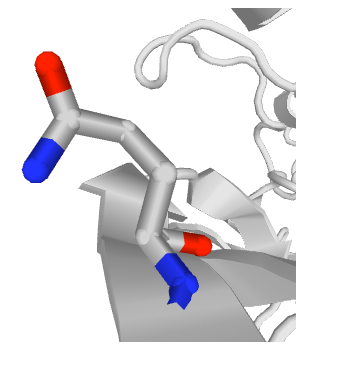


Figure 11 Beta strand at position 80-91

Figure 12 Zoomed in view of Beta strand

However, the beta strand in Jpred terminates at an early position as compared to what is predicted by Swiss- model. The beta- strand is extended from position 80-91.