NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

School of Science

Information Technologies in Medicine and Biology

Direction: *Bioinformatics*

Algorithms in Structural Bioinformatics

Postgraduate Student: *Begetis Nikolaos* Professor: *Chrysina Evangelia*

Deadline Date: *30/03/2013*

Assignment 1

1

As we learned in class, a TIM barrel is a conserved protein fold consisting of eight α-helices and eight parallel β-strands. These helices and strands, alternate along the peptide backbone. TIM barrels are common protein folds, and one of the most intriguing features among members of this class of proteins is that although they all exhibit the same tertiary fold there is very little sequence similarity between them.

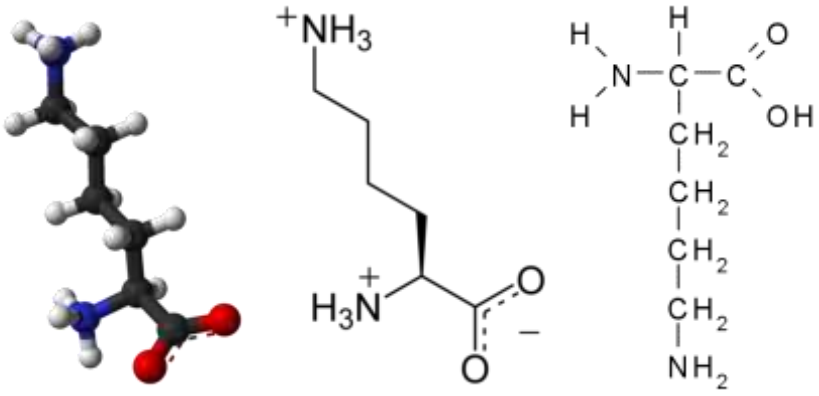
In this assignment's first task we were asked to select an amino-acid that is detected in a TIM barrel protein fold and describe its name in all ways and the atoms of its strains. To do that, we decided to look for the Lysine (Lys) amino-acid in the Chitinase protein enzyme.

To start with a brief definition, Lysine is an amino-acid with the chemical formula:

$$HO_2CCH(NH_2)(CH_2)_4NH_2$$

Lysine's codons are AAA and AAG, and it is a base, as are arginine and histidine. It participates in hydrogen bonding and as a general base in catalysis. It can be considered to be somewhat amphipathic as the part of the side chain nearest to the backbone is long, carbon-containing and hydrophobic, whereas the end of the side chain is positively charged. For this reason, one can find lysines where part of the side chain is buried and only the charged portion is on the outside of the protein. However, this is by no means always the case and generally lysines prefer to be on the outside of proteins. Lysines are also frequently involved in salt-bridges.

Lysines are quite frequent in protein active or binding sites. Lysine contains a positively-charged amino group on its side chain that is sometimes involved in forming hydrogen bonds with negatively-charged non-protein atoms.



Figures 1-3

Looking at Lysine's structure, the (NH_3^+) amino group is attached to the fifth carbon beginning from the α-carbon, which is attached to the carboxyl (C=OOH) group.

In the first of the above figures one can see, a ball-and-stick model of the L-lysine cation, (S)-2,6-diammoniohexanoate, [C6H15N2O2]+, from the crystal structure of L-lysine monohydrochloride dihydrate.

In respect to Lysine's chemical structure shown in the third figure the $H_2N-(CH_2)_4$ chain consists the side chain whereas the rest consist the main chain of this amino-acid.

The molecular formula of Lysine is $C_6H_{14}N_2O_2$ and as we can infer from the figure below Lysine (K) is a Hydrophobic, Polar, Positive Charged amino-acid.

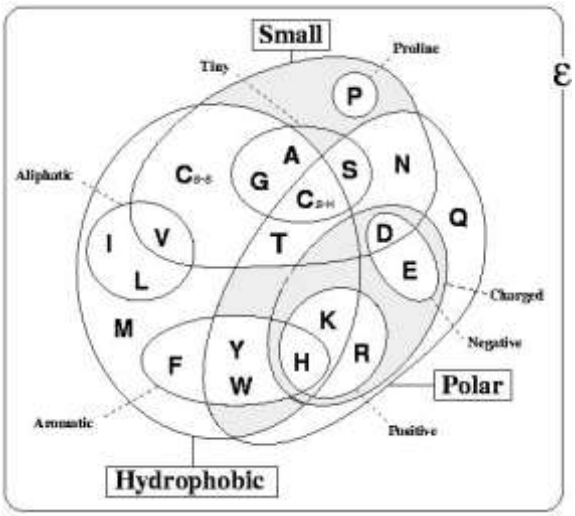


Figure 4

In respect to human symptoms, Lysine found reduces the feeling of helplessness, counteracts fatigue, helps in the absorption of calcium, need to proper growth. Fights cold sores and herpes simplex I and other viruses and helps lower triglycerides.

To become more specific and before we move on to the next task we leave aside all the bibliographic references and we visit the PDB Protein Data Bank of EMBL-EBI in Europe.

There we search for Chitinase 1ctn and from the results given we can infer that all the above bibliographic references were correct. For instance in the figure's below (Fig. 5) label there was referred the following text:

- "Default element colours are as defined by PyMOL e.g. ***carbon***, nitrogen, oxygen, sulphur, phosphorus, selenium."
- "***Positively*** and negatively-charged atoms of standard residues in proteins and polynucleotides are blue and red respectively."
- "Similarly, electro-positive and electro-negative ***polar*** atoms are skyblue and salmon respectively."

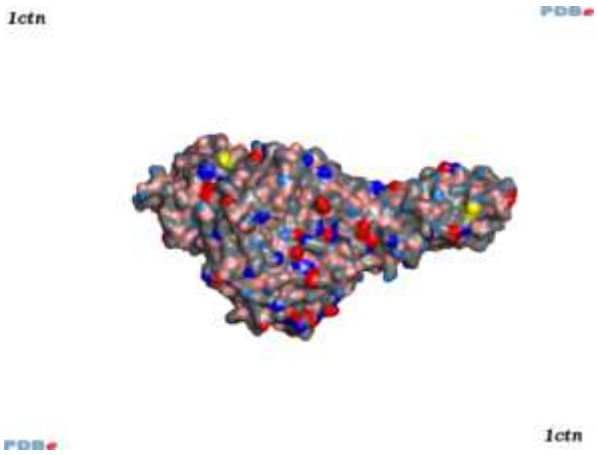


Figure 5

We highlighted and emphasized the words respective to the Lysine amino-acid.

2

In respect to what was indicated above, we chose the Chitinase protein enzyme to capture the Primary, Secondary, Tertiary and Quaternary structures of this protein. Using the EMBL-EBI PDB Protein Data Bank web tool we selected each one of the structures as they are listed in the left menu of the web tool. The results are provided below:

Primary Structure:

In order to get the primary structure we selected in the header menu the Primary Structure the checkbox of the fasta string format. The result is shown below (Lys in highlighted in red):

```
>1ctn_A; molId:1; molType:protein; unp:P07254; molName:CHITINASE A;
AAPGKPTIAWGNTKFAIVEVDQAATAYNNLVKVKNAADVSVSWNLWNGDTGTTAKILLNGKEAWSGPSTGSSGTA
NFKVNKGGRYQMQVALCNADGCTASDATEIVVADTDGSHLAPLKEPLLEKNKPYKQNSGKVVGSYFVEWGVYGRN
FTVDKIPAQNLTHLLYGFIPICGGNGINDSLKEIEGSFQALQRSCQGREDFKISIHDPFAALQKAQKGVTAWDDP
YKGNFGQLMALKQAHPDLKILPSIGGWTLSDPFFFMGDKVKRDRFVGSVKEFLQTWKFFDGVDIDWEFPGGKGAN
PNLGSPQDGETYVLLMKELRAMLDQLSTETGRKYELTSAISAGKDKIDKVAYNVAQNSMDHIFLMSYDFYGAFDL
KNLGHQTALNAPAWKPDTAYTTVNGVNALLAQGVKPGKIVVGTAMYGRGWTGVNGYQNNIPFTGTATGPVKGTWE
NGIVDYRQIAGQFMSGEWQYTYDATAEAPYVFKPSTGDLITFDDARSVQAKGKYVLDKQLGGLFSWEIDADNGDI
LNSMNASLGNSAGVQ
```

*The results are also provided with the **1ctn.ent** file given with in the deliverable zipped file*

Secondary Structure:

As it is shown in Figure 6 taken from the PDB web tool, by selecting in the left bar the Secondary Structure and by the header bar the Secondary Structure checkbox the secondary structure consists of 25 strands (137 amino-acids, 25.4%) and 19 helices (228 amino-acids, 42.3%).

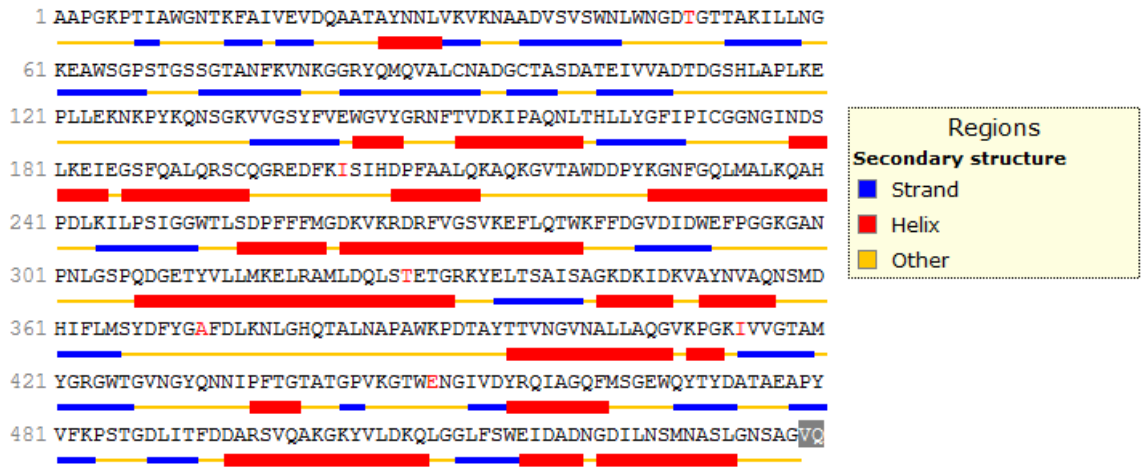


Figure 6

Tertiary and Quaternary Structures:

Now, having installed the CCP4 tool proposed in class we can depict the above secondary structure to by following the steps of Figure 7 to what in shown in Figure 8.

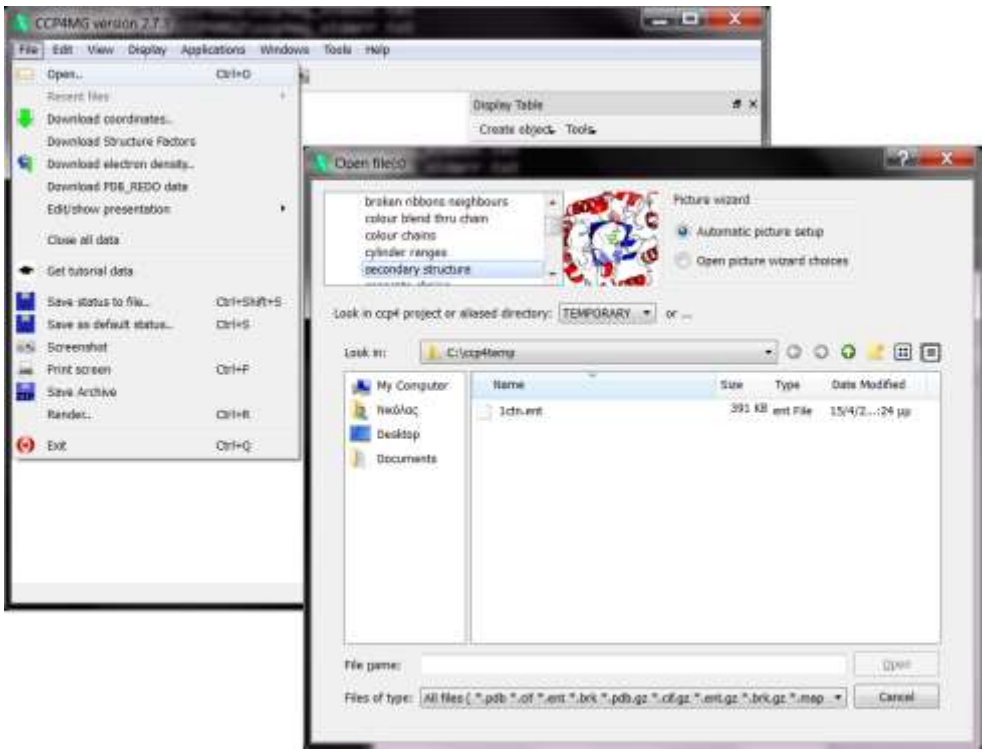


Figure 7

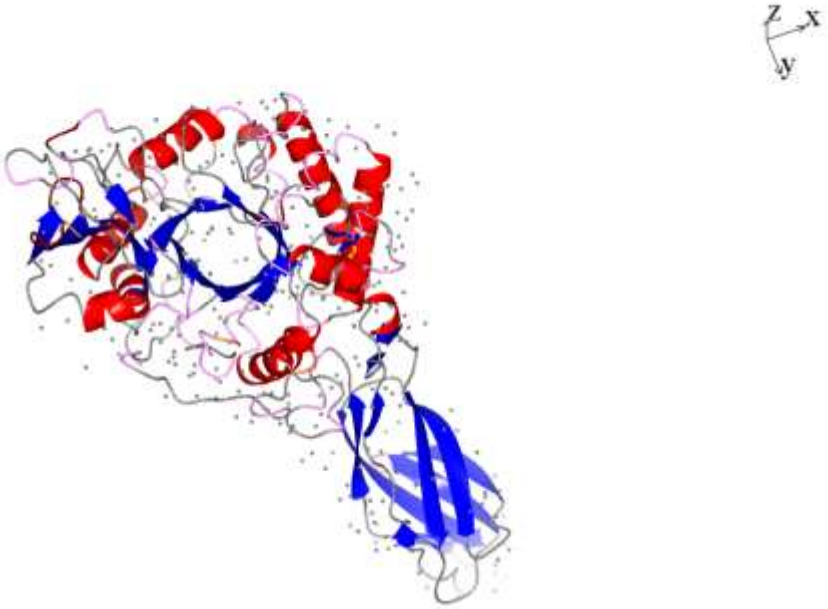


Figure 8

In Figure 8 we selected to show the quaternary filtered by the secondary structure types' format (helices have different color than strands etc.) in this angle in order to be easier to be observed the TIM-barrel fold in the middle.

Figure 9 shows the same structure but with more details. To achieve this depicting we selected in the right sidebar of CCP4 all the three domains to show all their atoms with the format of the secondary structure (Figure 10).

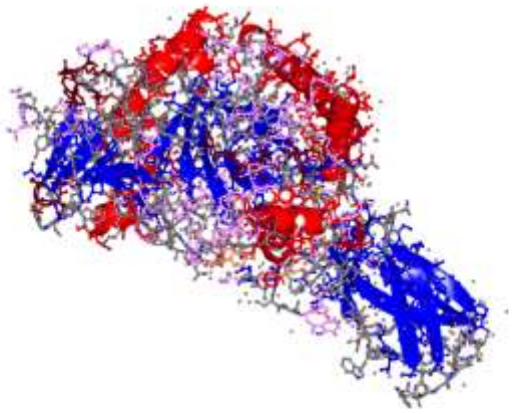


Figure 9

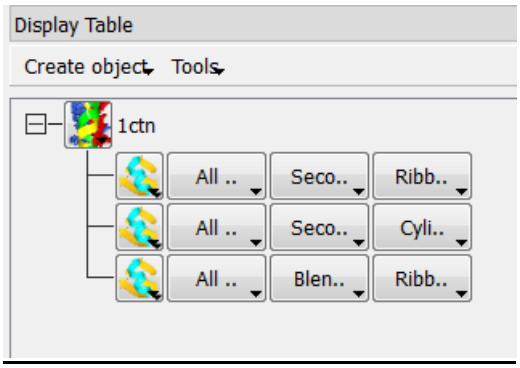


Figure 10

3

As a third task we were asked to find how many structures have a TIM Barrel Fold according to the Protein Data Bank PDB.

The procedure we followed was to recur to the EMBL-EBI PDB and to RCSB PDB and hit in the “search” form and the “browse” form respectively the search key words “TIM barrel fold” and “TIM barrel”, respectively. There we were given as results what in shown in the following two figures (Figure 11 and 12)

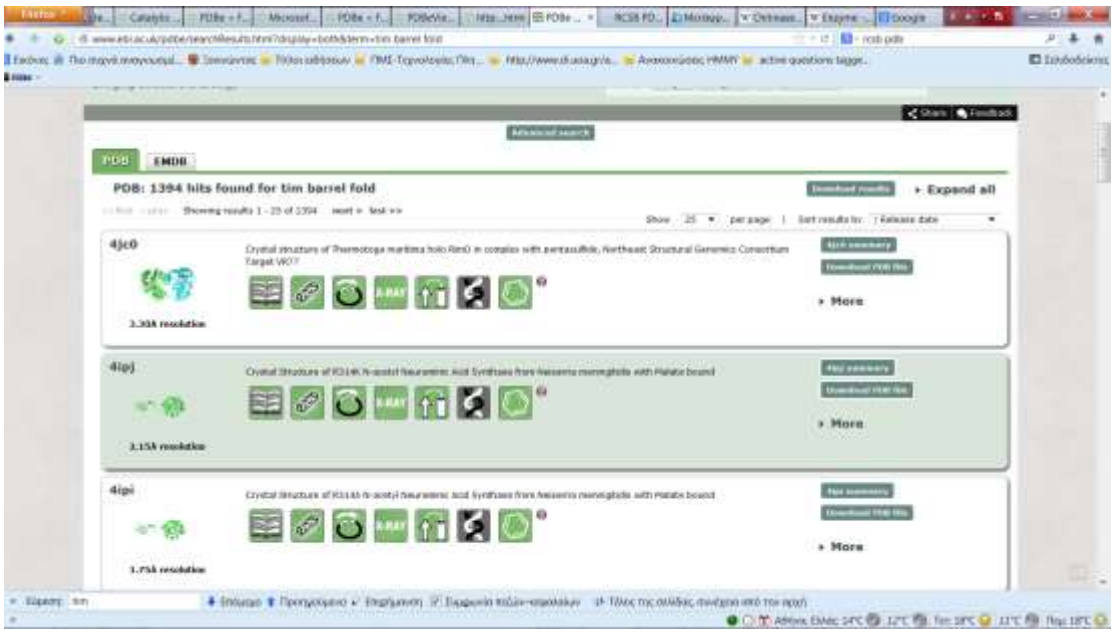


Figure 11

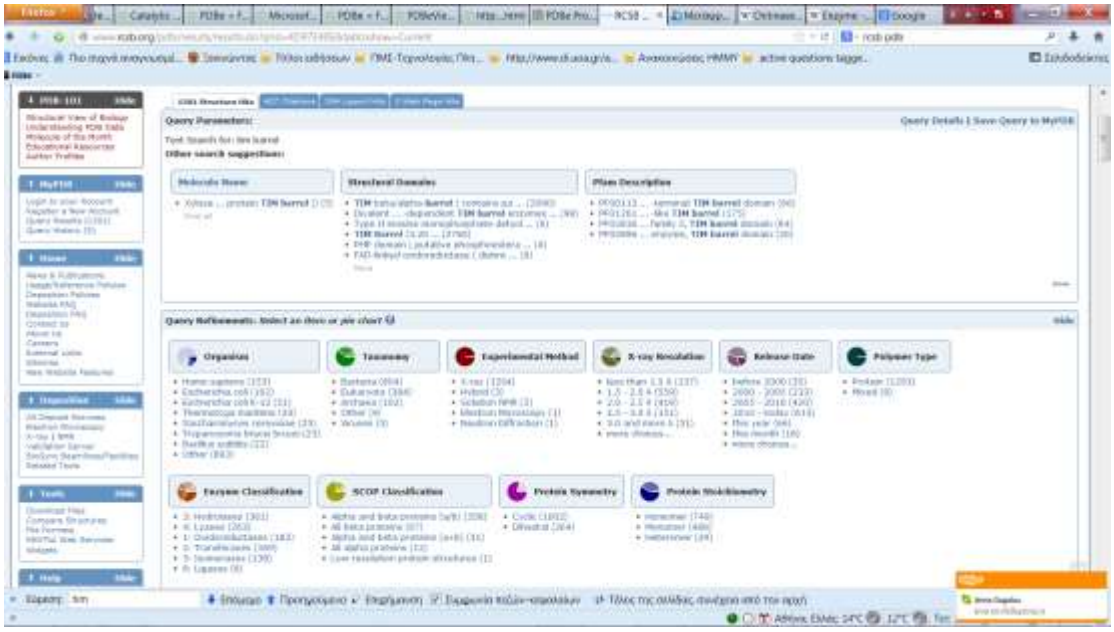


Figure 12

As we can see in the above two figures in the EMBL-EBI PDB data bank there were found 1394 hits for TIM barrel fold, while in RCSB PDB were found 1301 hits. This difference in numbers is caused due to the updates each databank does and due to which structures does every protein data bank accept as valid.

Some more figures showing some statistical details on the above TIM barrel folds are shown below (Figures 13-16).

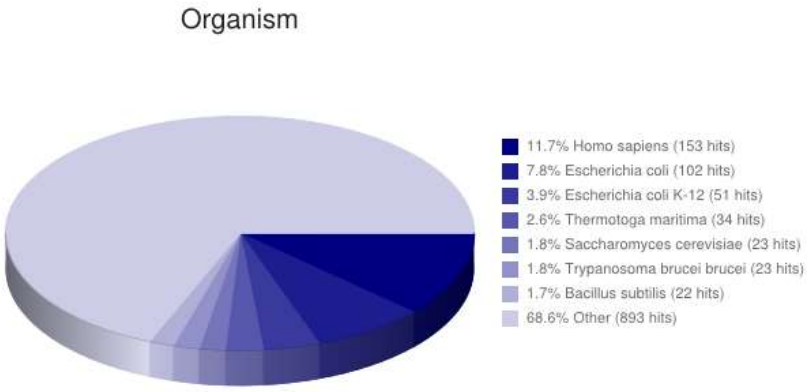


Figure 13

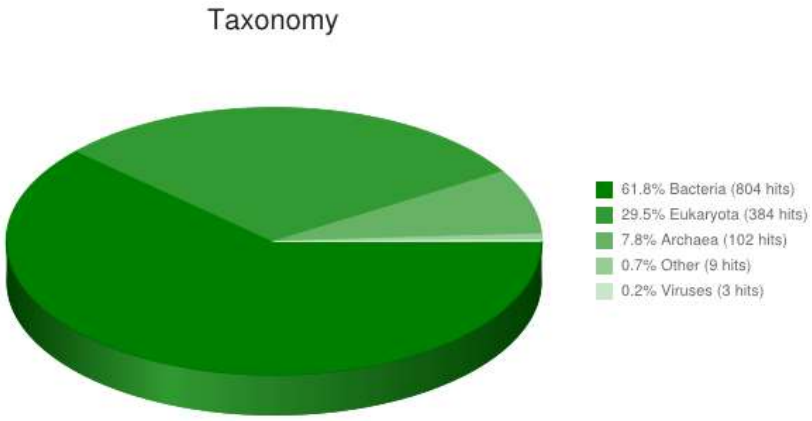


Figure 14

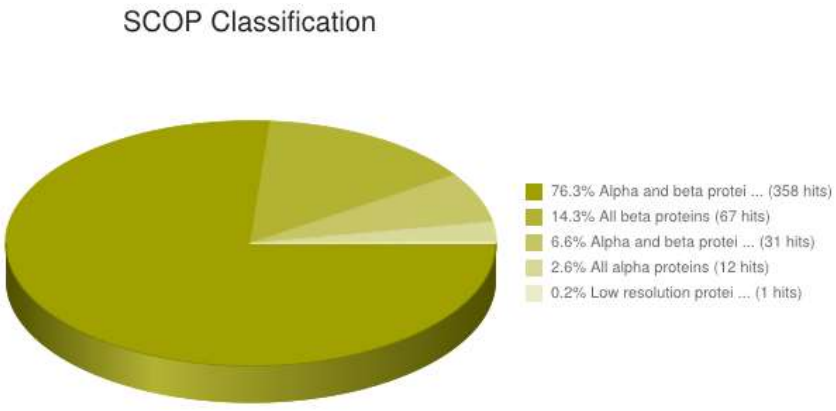


Figure 15

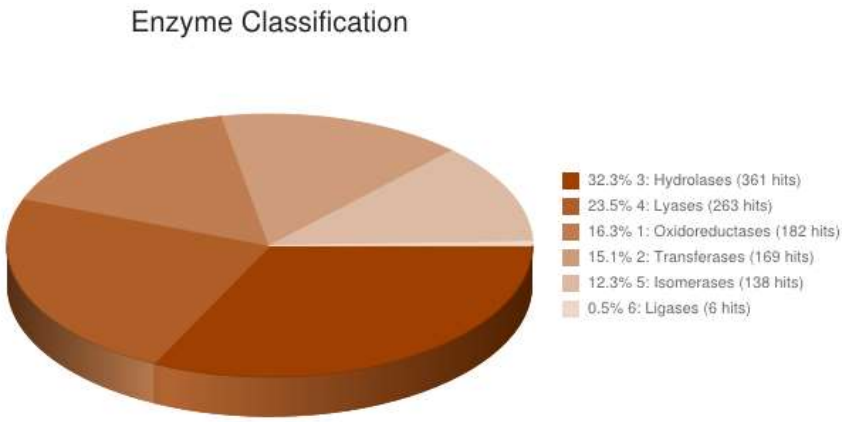


Figure 16

4

As a fourth task we were asked to read the PDBFold tool tutorial and apply thePDB fold for the Chitinase enzyme with PDB code: 1ctn (the same we selected in purpose above so that we have something integrated).

So, to begin with, we started with a pairwise alignment. We selected in the left menu bar in PDBe the button for “Structure similarity”. Then, in the next page we launched PDBeFold tool which started up the webservice, where we selected the **pairwise** alignment, for the chitinase enzyme with code 1ctn as requested (Figure 17).

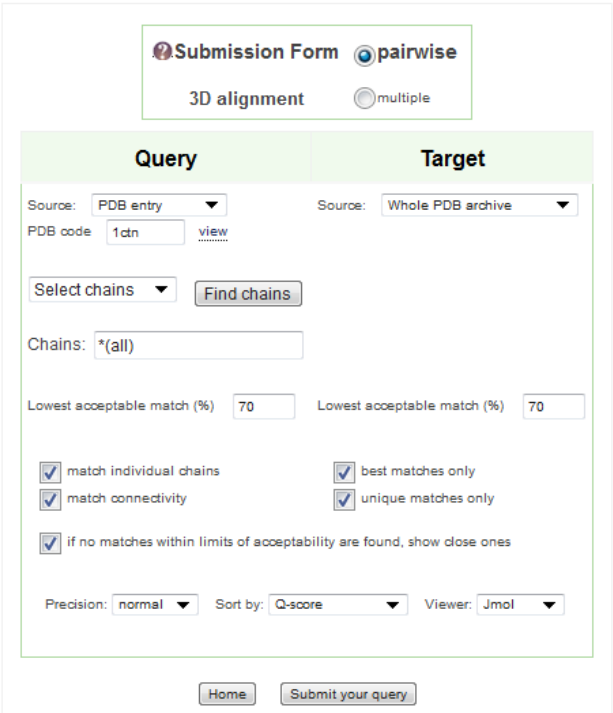


Figure 17

All the rest of input selections were kept with the default values. Some of the results sorted in descending order are shown in Figure 18. Results can also be found in file reslist_pair.dat packaged in the deliverable.

Figure 18

Now, given the above alignment and excluding the first in list item because it represents the 1ctn crystal structure of chitinase itself, we select the second protein in list with PDB code 2wly:2 found in chitinase's chain A, too. This protein has an 98% percent of Ca matches (Q) and an 99% of sequence matches. A part of these statistics and all the rest of features are presented in Figure 19 (all statistics can be found in file "pairmatch_with_2wly.dat"), while the superpose of the two proteins in 3D superposed view is shown in Figure 20, where one can observe the few alterations in the 3D surface as it being a little blurred.

Figure 19

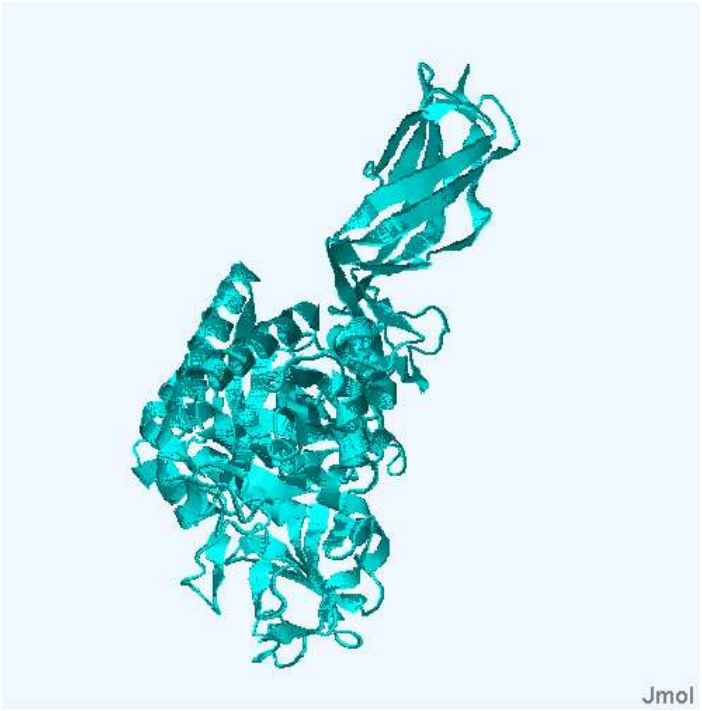


Figure 20

Following the same steps for the last of the 36 proteins in the above pairwise alignment with PDB code 3as2:A found in chitinase's chain A, too. This protein has an 85% percent of Ca matches (Q) and an 54% of sequence matches. A part of these statistics and all the rest of features are presented in Figure 21 (all statistics can be found in file "pairmatch_with_3as2.dat"), while the superpose of the two proteins in 3D superposed view is shown in Figure 22. Now, it is more clear for one to observe the alterations between the two in the 3D surface. With the grey color one can see the 1ctn chitinase and with the turquoise color the 3as2 chitinase surface superposed on the 1ctn.

Figure 21

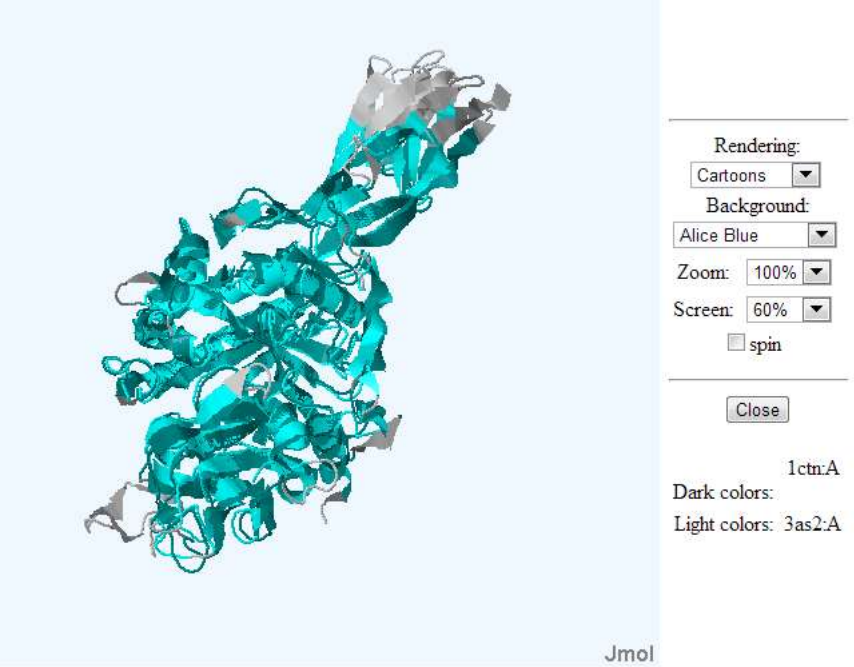


Figure 22

Having shown the pairwise alignment, we continue with the multiple alignment. We follow exactly the same procedure. Then, in the next page we launched PDBeFold tool which started up the webservice, where we selected the **multiple** alignment, for the chitinase enzyme with code 1ctn as requested (Figure 23), we also selected the most of the proteins we found in the previous pairwise alignment as the sequences we wanted 1ctn to be aligned with. As it was expected the PDBeFold tool lagged for a few minutes.

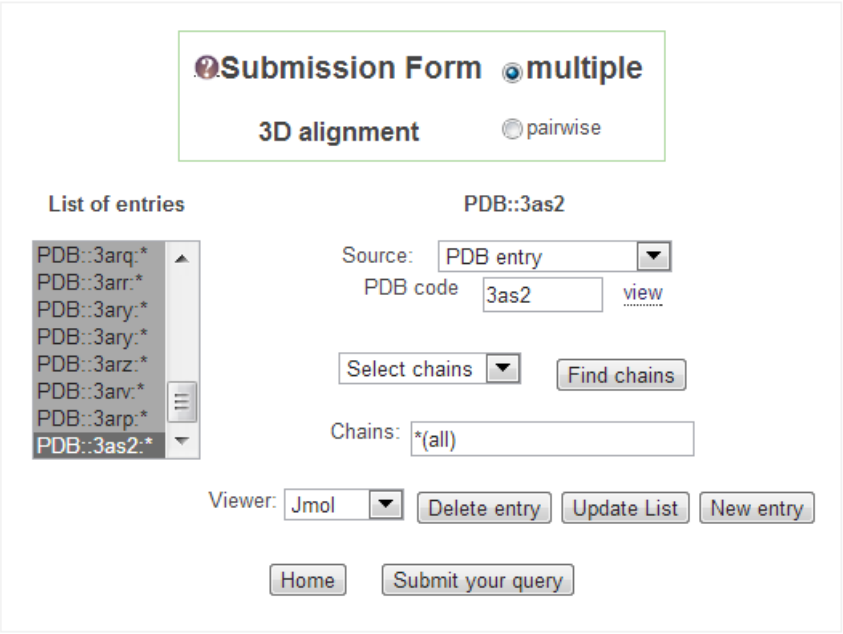
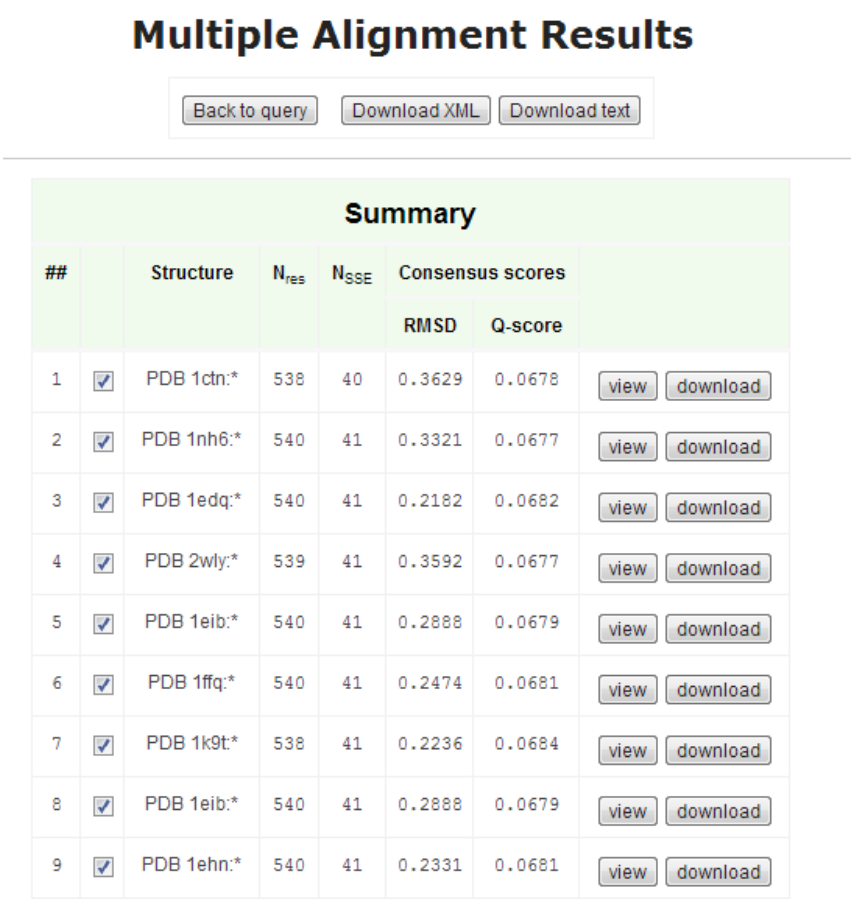


Figure 23

All the rest of input selections were kept with the default values. Some of the results sorted in descending order are shown in Figure 24. Results can also be found in file reslist_mult.dat packaged in the deliverable. Moreover, the secondary structure alignment in shown in Figure 25.



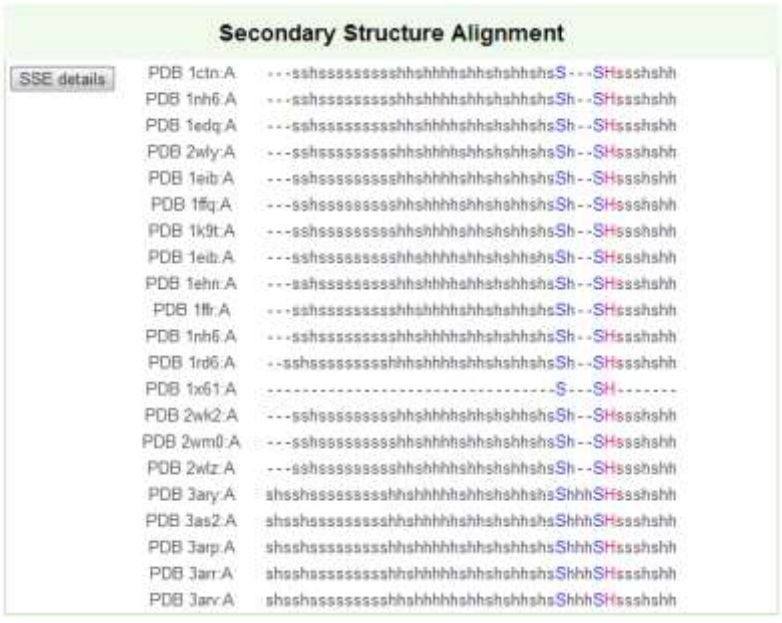
Multiple Alignment Results

Back to query | Download XML | Download text

Summary

##		Structure	N_{res}	N_{SSE}	Consensus scores		
					RMSD	Q-score	
1	☑	PDB 1ctn:*	538	40	0.3629	0.0678	view download
2	☑	PDB 1nh6:*	540	41	0.3321	0.0677	view download
3	☑	PDB 1edq:*	540	41	0.2182	0.0682	view download
4	☑	PDB 2wly:*	539	41	0.3592	0.0677	view download
5	☑	PDB 1eib:*	540	41	0.2888	0.0679	view download
6	☑	PDB 1ffq:*	540	41	0.2474	0.0681	view download
7	☑	PDB 1k9t:*	538	41	0.2236	0.0684	view download
8	☑	PDB 1eib:*	540	41	0.2888	0.0679	view download
9	☑	PDB 1ehn:*	540	41	0.2331	0.0681	view download

Figure 24



Secondary Structure Alignment

SSE details

PDB 1ctn:A ---sshsssssssssshhshhhhhshshhshsS---SHssshshh
PDB 1nh6:A ---sshsssssssssshhshhhhhshshhshsSh--SHssshshh
PDB 1edq:A ---sshsssssssssshhshhhhhshshhshsSh--SHssshshh
PDB 2wly:A ---sshsssssssssshhshhhhhshshhshsSh--SHssshshh
PDB 1eib:A ---sshsssssssssshhshhhhhshshhshsSh--SHssshshh
PDB 1ffq:A ---sshsssssssssshhshhhhhshshhshsSh--SHssshshh
PDB 1k9t:A ---sshsssssssssshhshhhhhshshhshsSh--SHssshshh
PDB 1eib:A ---sshsssssssssshhshhhhhshshhshsSh--SHssshshh
PDB 1ehn:A ---sshsssssssssshhshhhhhshshhshsSh--SHssshshh
PDB 1ffr:A ---sshsssssssssshhshhhhhshshhshsSh--SHssshshh
PDB 1nh6:A ---sshsssssssssshhshhhhhshshhshsSh--SHssshshh
PDB 1rd6:A --sshsssssssssshhshhhhhshshhshsSh--SHssshshh
PDB 1x61:A ----------S---SH-------
PDB 2wk2:A ---sshsssssssssshhshhhhhshshhshsSh--SHssshshh
PDB 2wm0:A ---sshsssssssssshhshhhhhshshhshsSh--SHssshshh
PDB 2wtz:A ---sshsssssssssshhshhhhhshshhshsSh--SHssshshh
PDB 3ary:A shsshsssssssssshhshhhhhshshhshsShhhSHssshshh
PDB 3as2:A shsshsssssssssshhshhhhhshshhshsShhhSHssshshh
PDB 3arp:A shsshsssssssssshhshhhhhshshhshsShhhSHssshshh
PDB 3arr:A shsshsssssssssshhshhhhhshshhshsShhhSHssshshh
PDB 3arv:A shsshsssssssssshhshhhhhshshhshsShhhSHssshshh

Figure 25

Finally, Figure 26 shows the superposition in 3D of all the proteins we tested in multiple alignment tool. There, one can observe all the alterations in the 3D surface of each protein from all the rest of proteins.

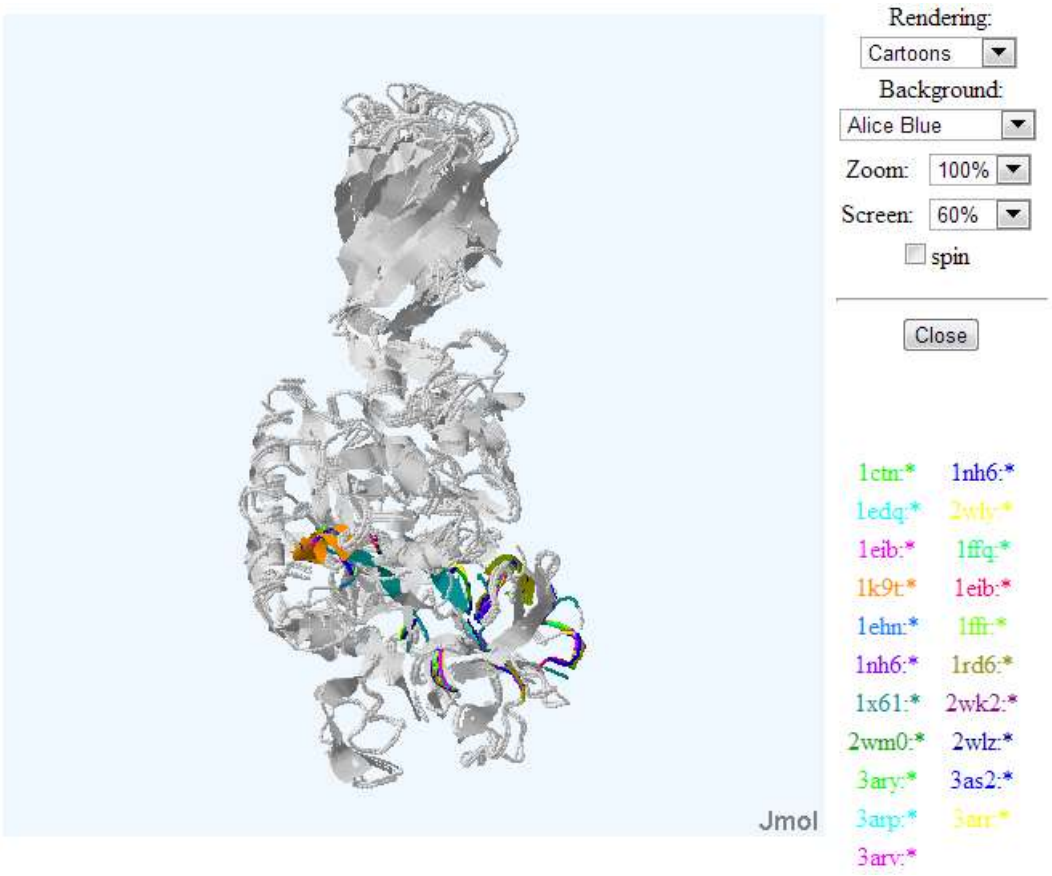


Figure 26

5

For the fifth and last task of this exercise we were asked to propose methods of administrating the above information we gathered in the previous task´s application.

So, to begin with, there exist a lot of tools that make a multiple alignment in Protein´s primary structure. Therefore, sometimes, we need to check also the secondary, tertiary and quaternary structures using the RMSD to find the secondary identity.

We were glad to find that sometimes when we test two or more protein alignments in spite of having a less than 50% sequence identity, they appear to share a high level of secondary structure. It can easily be observed in the above provided figures. This tells us that proteins may share the same phylogenecy us despite having the most of the amino-acids altered the conserve their structure, which means that they conserve the amino-acid properties as shown in Figure 4 in the beginning of this report.

So to conclude and answer this task's question, we learned from the above experiment that whether a multiple alignment in primary structure is not too good, we must not exclude it

from the final results for our experiments. It is quite probable that it has a good alignment in the secondary structure and so it has conserved the proteins properties.

To sum up and dive a little deeper, we can also infer which areas alter mostly in the proteins secondary structure, and so we can check explicitly the regions in the primary structure where the alterations of the secondary structure occur. An example for this case in shown in the above Figure 26, where all the proteins we tested seem to have been altered in the same secondary structure areas.