Sandhya Tiwari
MOL 217 - Applied Bioinformatics
2 May 2012

ASSIGNMENT 2 - FINAL REPORT

In this assignment, we were given three sequences by collaborators for analysis and modelling. For an initial assessment of these sequences and their homologues, the sequences were put through BLASTP (protein-protein BLAST)[cite] against the NR (All non-redundant GenBank CDS translations +PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects) protein databases.

Sequence no.	Top hit	Identity (%)
1	Triosephosphate Isomerase, putative [Ricinus communis]	100
2	Endochitinase A [Serratia liquefaciens]	100
3	Methylaspartate Ammonia-Lyase [Eubacterium infirmum F0142]	100

As the collaborator had pointed out, they all possess the alpha-beta barrel fold. More specifically, the TIM-barrel fold, which was confirmed by the Conserved Domains information provided in the BLASTP output.

Triosephosphate isomerase (TIM) is an enzyme that catalyzes the reversible interconversion of triose phosphate isomers; dihydroxyacetone phosphate and D-glyceraldehyde 3-phosphate. This plays an important role in glycolysis and is essential for efficient energy production and is found in nearly every organism which performs this reaction.

Endochitinase A is a variant of endochitinases which catalyse the hydrolysis of nonterminal (1->4)-beta linkages of N-acetyl-D-glucosamine (GlcNAc) polymers of chitin and chitodextrins. They typically cleave randomly within the chitin chain. The catalytic motif is defined by "D-X-X-D-X-D-X-E". This enzyme is a member of the Glycoside Hydrolase Family 18 which also contains catalytically inactive members from plants that function as inhibitors or lectins.

Methylaspartate Ammonia-Lyase (with cofactor magnesium) cleaves the carbon-nitrogen bond in L-threo-3-methylaspartate to produce mesaconate and ammonia. It is known to participate in c5-branched dibasic acid metabolism and nitrogen metabolism.

Given that we now know the possible identity and/or functions of these protein sequences, it was important to find if structures were available of these sequences or of proteins closely related to them. The BLASTP was then repeated against the PDB structures database.

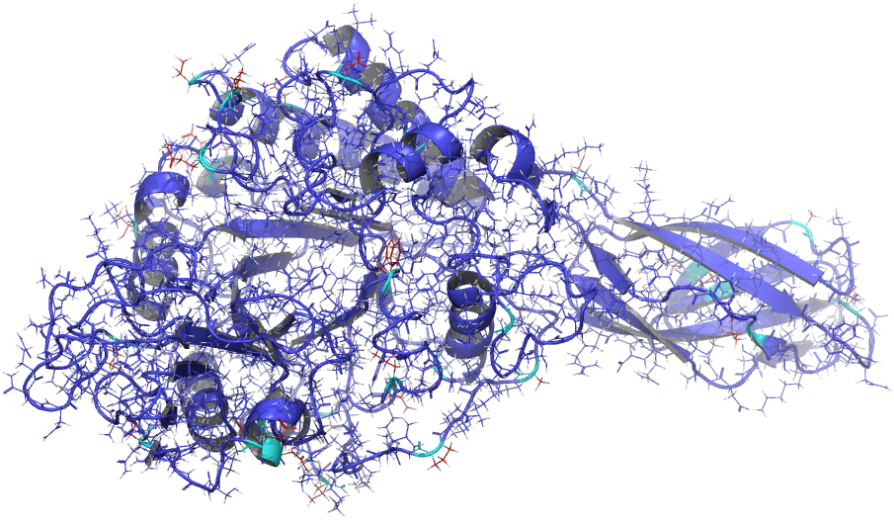
Sequence no.	Top PDB structure hit	Identity (%)	Query coverage (%)
1	1SW3 - Triosephosphate Isomerase [Gallus gallus]	65	77
2	1RD6 - Chitinase A mutant W167A [Serratia Marcescens]	94	99

Sequence no.	Top PDB structure hit	Identity (%)	Query coverage (%)
3	1KCZ - Beta-Methylaspartase [Clostridium tetanomorphum]	67	98

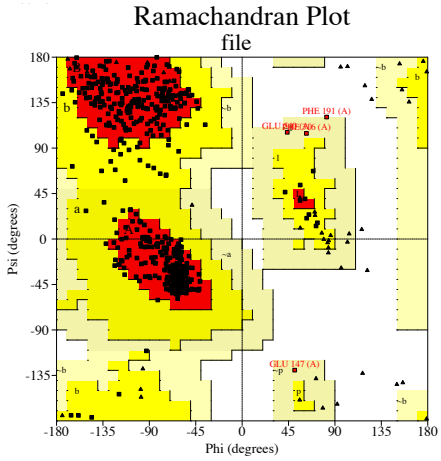
From the above output summary, we find that the sequences do not have a structure of their own, but there are homologues which do. The caveats are that a) the hits cover different lengths of the sequences (query coverage), which means only parts of the sequences may be modelled based on this value and b) the identity also varies, which has a direct impact on the validity on any kinds of models made.

For sequence 2, which is the Endochitinase A, it makes sense to proceed with the homology modelling approach due to the high identity and coverage of the template 1RD6. For this, we used the Prime homology modelling application in Maestro (Schrodinger Suite 2012). Even there are many available structures with the same percentage of coverage and identity, given the high values, it seems valid to just use 1RD6 as the template for the alignment and the sequence. SSPro was used to predict the secondary structure of the sequence, which was also used for this alignment.

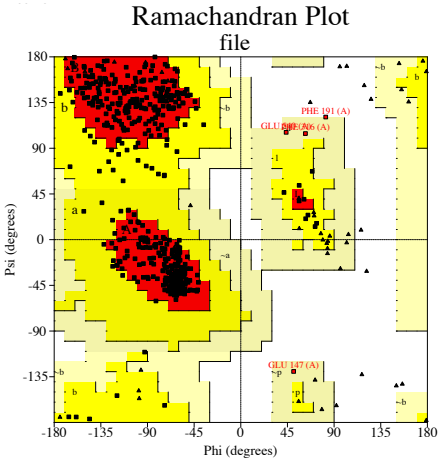
The model built shows a colour scheme where the blue atoms are those derived directly from the template, cyan atoms have backbone atoms from the template, and predicted side chains, and red atoms have been predicted or modeled. In general, the model does not show much deviation from the template, which leaves the need for further refinement to a minimum.



To validate this model, I ran it through PROCHECK (customised output webpage here: http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=q723&code=133702).

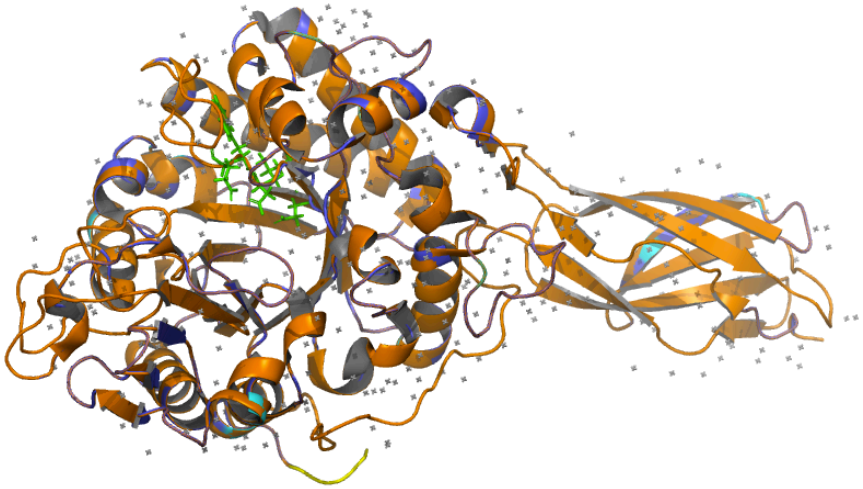


Ramachandran Plot for Sequence 2 model.



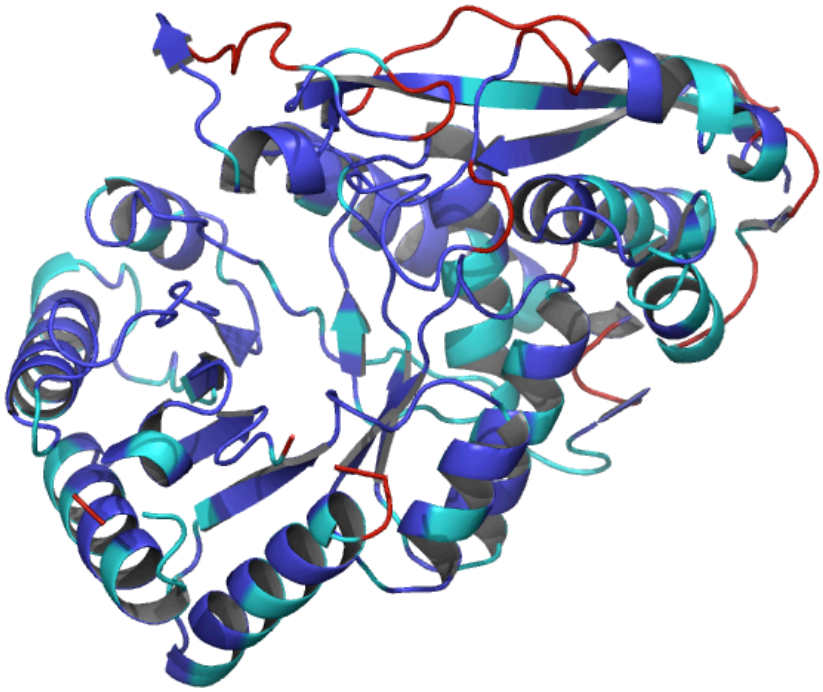
Ramachandran Plot for template 1RD6.

Looking at the Ramachandran plot of both the template and the model, we do not find significant differences between them. According to the statistics generated, the model fares just as well with having residues in most favoured regions [A,B,L] which 87.2%, 0.1% higher than the template, even though it is below the 90% threshold recommended based on analysis done on 118 structures of 2.0 Å resolution and better.



Superimposition of the template (orange) on to the model for sequence 2. The green sticks signify the catalytic motif of chitinase which is conserved in the sequences of both proteins. The grey dots symbolise water from the template structure, and the stretch of loop coloured yellow was ignored in the model-building, as the template sequence is longer than that of the target.

A different strategy was used to model sequence 3, which has high query coverage, but a much lower sequence identity. We attempted to create a chimeric model from two structures, 1KCZ (mentioned above) and 1KKR (a methyaspartate ammonia lyase structure from Citrobacter amalonaticus). This method produced the model below, where the predicted parts mostly lie in the loop regions, and there is a broken section within the TIM barrel, in one of the loops. This structure would have to be refined further to be built properly, or perhaps a single structure model would have been best.



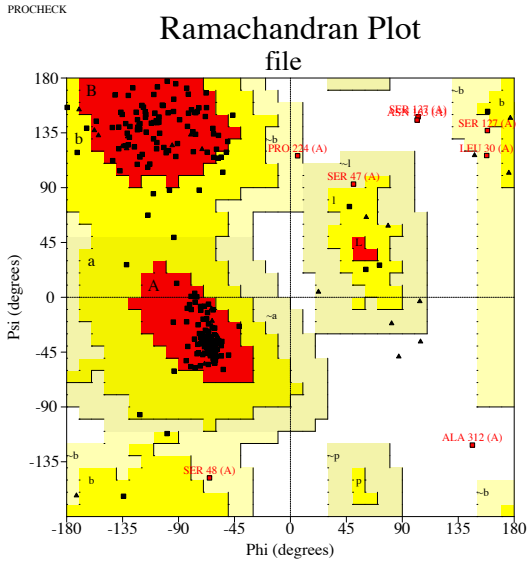
Finally, for sequence 1, which was the most different to the top BLAST hit than the other two sequences, I decided to use I-TASSER. I-TASSER server is a service where 3D models are built based on multiple-threading alignments by LOMETS and iterative TASSER assembly simulations. Potential functions are also sought by comparing the models to protein function databases. 5 models were generated but of them all, I picked the one with the highest C-score of -1.18 (confidence score calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations). Most C-scores lie within -5 and 2, where a higher value signifies higher confidence.

The image below is that of the best model of sequence 1. The top ten templates used by I-TASSER are describe in a table below.

Rank	PDB Hit	Identity 1	Identity 2	Coverage	Norm. Z-score
1	1mo0A	0.56	0.46	0.82	3.20
2	1b9bA	0.46	0.39	0.79	3.06
3	3m9yA	0.41	0.34	0.79	5.94
4	3th6A	0.59	0.46	0.78	4.80
5	1mo0A	0.56	0.46	0.82	4.73
6	2dp3A	0.50	0.40	0.80	2.54
7	3m9yA	0.41	0.34	0.79	6.05
8	1mo0A	0.56	0.46	0.81	3.70
9	1ssdA	0.63	0.49	0.79	3.12
10	2yc6A	0.48	0.40	0.80	2.98

Identity 1 is the percentage sequence identity of the templates in the threading aligned region with the query sequence, Identity 2 is the percentage sequence identity of the whole template chains with query sequence, Coverage represents the coverage of the threading alignment that is equal to the number of aligned residues divided by the length of query protein and lastly, Norm. Z-score is the normalized Z-score of the threading alignments. Alignment with a Normalized Z-score >1 mean a good alignment and vice versa. The PDB hit 1mo0 chain A features thrice (highlighted in yellow). This structure was found to be that of TIM from C. elegans. It also ranks top as a PDB identified to be a structural analogue of this model. 1SW3, the structure we identified earlier, is second on this list.

To validate this model, I ran it through PROCHECK (customised output webpage here:http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=r390&code=152236)



Aside from the 7 residues in the unfavourable regions, the statistic for the residues in most favourable regions [A,B,L] is 90.1%, which is within the range for a good model.