Indian Journal of Agricultural Sciences **82** (10): 867–74, October 2012

Structure and function prediction of unknown wheat protein using LOMETS and I-TASSER

GEETIKA JETHRA[1], A K MISHRA[2], P S PANDEY[3] and H CHANDRASEKHARAN[4]

Unit of Simulation and Informatics, Indian Agricultural Research Institute, New Delhi 110 012

Received: 19 July 2011; Revised accepted: 27 June 2012

ABSTRACT

Wheat is a vital dietary cereal crop often associated with valuable health effects. A study was carried out to investigate the in-silico analysis of 2D and 3D structure prediction of protein with an unidentified structure and function in *Triticum aestivum*. Primary structure prediction and physicochemical characterization were performed by computing theoretical isoelectric point (pI), molecular weight, and total number of positive and negative residues, extinction coefficient, instability index, aliphatic index and grand average hydropathy (GRAVY). In the present study, a high quality 3D structure and function of wheat protein (CAA35597.1) from NCBI has been predicted for the hypothetical amino acid sequence. For the prediction of secondary and tertiary structure of protein, LOMETS (LOcal MEta-Threading-Server) and I-TASSER (Iterative Threading Assembly Refinement) servers were used respectively. The models were validated using protein structure checking tool PROCHECK. These structures also provide a good foundation for functional analysis. Analysis show that the unknown query protein (CAA35597.1) is similar structurally and functionally to 1tmq_B and 1blu_A of *Tenebrio molitor* (Yellow mealworm) and *Allochromatium vinosum* respectively showing catalytic, binding and inhibitor activity (Molecular) with metabolic and cellular metabolic Activity (Biological).

Key words: I-TASSER, LOMETS, NCBI, Wheat

Wheat belongs to a Poaceae family also known as the grass family. It is one of the largest families. Wheat is grown all over the world for its highly nutritious and useful grain. It is one of the top three most produced crops in the world, including corn and rice. Wheat is the world's most important cereal crop in terms of both areas cultivated and amount of grain produced. Wheat has been cultivated for over 10,000 years and probably originates in the Fertile Crescent, along with other staple crops. It is widely grown throughout the temperate zones and in some tropical/sub-tropical areas at higher elevations. A wide range of wheat products are made by humans, most favourably flour, which is made from the grain itself.

Many molecular and cell biologists face the foremost problem of not having the known structure and function for their protein of interest. The most remarkable effort in the recent era of protein structure determination is development of structure genomics that aims to attain 2D and 3D models of all proteins by an optimized combination of experimental based structure solution and computer-based structure prediction (Terwilliger *et al.* 1998, Brenner *et al.* 2000).

The hurdle that most of the protein structures have not been predicted yet has motivated the structure biologists and bioinformaticians in recent years. Owing to the fact that the percentage of protein sequences in UniProtKB/TrEMBL, with a solved protein structure in the Protein Data Bank (PDB) library2, plunged to 0.6% by the end of the year 2009; this number was 2% in the year 2004 and 1.2% in the year 2007 (Roy *et al.* 2010). Recent advances in computer algorithms for predicting protein structure and function provide biologists with valuable information on their proteins of interest (Stevens *et al.* 2001). Even though the number of experimentally available protein sequences are increasing exponentially over the last few years, there are yet a large number of proteins with unknown fold recognition and without an obvious homology with any other protein that has been resolved.

Computational methods for predicting three-dimensional (3D) protein structures have been historically divided into three categories, based on the availability of template structures in the PDB library, (i) Comparative modeling (ii) Threading, and (iii) Ab-initio modeling.

Comparative modeling: Predicts the three dimensional

[1] Senior Research Fellow (e mail: g_jethra@rediffmail.com); [2] Scientist (SS) (e mail: akmishra@iari.res.in); [3] Principal Scientist (e mail: pspanday@iari.res.in); [4] In-charge and Principal Scientist (e mail: head_usi@iari.res.in)

structure of a given protein sequence (target) based on an alignment to one or more known protein structures (templates). This is done by sequence or sequence profile comparisons (Skolnick *et al.* 2000), and high-resolution models can be generated by merely copying the framework of the template structures or by satisfying the spatial restraints collected from the template structures.

Threading: Homology modeling does not work for sequence difference larger than 20 percent and thus de novo methods are needed for those. The threading methods directly matches the query sequence with the 3D structures of other solved proteins, with the objective of fold recognition similar to the query even when there is no evolutionary relationship between the query and the template protein. The final class of method includes, de novo or Ab-initio methods, predict the structure from sequence alone, without relying on similarity at the fold level between the modeled sequence and any of the known structures because of the non-availability of structurally related proteins in the PDB library. This is the hardest method and the success is limited to only small proteins <120 amino acids (Liwo *et al.* 1999, Simons *et al.* 1997)

Comparative modeling and threading methods use sequence profile and profile–profile alignments for identifying the best templates, but most of the Ab-initio modeling algorithms use evolutionary or knowledge-based information for collecting spatial restraints (Wu *et al.* 2007, Zhang *et al.* 2003) or for identifying local structural building blocks, i e secondary structure (Liwo *et al.* 1999). Recent community-wide critical assessment of protein structure prediction (CASP) experiments (Jauch *et al.* 2007, Battey *et al.* 2007) have revealed the fact that composite approaches in protein structure prediction, which combine various techniques such as threading, Ab-initio modeling and atomic-level structure refinement approaches are significant. I-TASSER (Iterative Threading Assembly Refinement) (Simons *et al.* 1997) is one example of the composite approaches and has been ranked as the best method for the automated protein structure prediction in the last two CASP experiments (Roy *et al.* 2010).

The physio-chemical properties of protein under study were predicted with the help of ProtParam. ProtParam computes various physico-chemical properties that can be deduced from a protein sequence. The parameters computed by ProtParam include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY). Molecular weight and theoretical pI are calculated by Compute pI/Mw. The primary server, used for secondary structure prediction is threading based LOMETS Nine state-of-the-art threading programs are installed and run in a local computer cluster, which ensure the quick generation of initial threading alignments compared with traditional remote-server-based meta services. It generates the consensus models from the top predictions determined by the threading server. A new version of structure modeler, called I-TASSER, is developed (Wu *et al.* 2007) by progressively implementing the TASSER simulations, where template alignments are generated by four simple variants of the profile–profile alignment (PPA) method with different combinations of the Hidden Markov Model (HMM) and PSI-Blast profiles with the Needleman-Wunsch (NW) and Smith-Waterman (SW) alignment algorithms (Kumar *et al.* 2011).

Wheat (*Triticum aestivum*) is one of the principal food crops of the world. It is grown on about one sixth of the total arable land globally. The fast growing population of the world necessitates the achievement of matching increase in the rate of food production. Since new areas that can be brought under cultivation are limited, high yield per unit acreage has to be achieved to meet food requirements of the increasing global population. But there are many limitations also in production of wheat like various diseases, stress conditions (heat, cold drought, salt, etc.), non-availability of quality seed of new wheat cultivars which has to be taken care of. Due to these reasons, improvement of the production of wheat grain is required. This can be achieved by developing various new resistant varieties, with the knowledge of structures and function of various proteins. Many genome sequencing projects are producing linear amino acid sequences, but complete understanding of the biological role of these proteins will require knowledge of their structure and function. Even though experimental structure determination methods resolve this problem by providing high-resolution structural information about a subset of the proteins, computational structure prediction methods will provide valuable information for the large fraction of sequences whose structures will not be determined experimentally.

Protein structure prediction methods have implicit underlying principles that fall into two categories: evolution and folding. Evolution-based methods seek to find conserved sequence patterns, while folding method simulate the physical process of folding. A folding pathway is a time series of protein folding events. Most molecular simulation methods, including molecular dynamics (MD) and Monte Carlo (MC), create a pathway implicitly. Other methods enforce certain characteristics of the folding events during the simulation, including some genetic algorithms, neural nets, and a new rule-based approach (Yuan *et al.* 2003).

The threading and sequence-structure alignment approaches are based on the observation that many protein structures in the PDB are very similar. The key idea of the threading method is to decrease dramatically the number of decoys. This is achieved by constraining conformation. All proteins to a small subset of confirmations obtained by threading through known protein structures that serve as a scaffold for the protein sequence in question and finding the energetically optimal alignment of the sequence to the scaffold

structure. Likewise Ab-initio prediction is the challenging attempt to predict protein structures based only on sequence information and without using templates. It is often divided into two distinct sub-problems: (1) the scoring function that can distinguish between native or native-like structures from non-native ones, and (2) the LOMETS, a meta-threading server, takes method of searching the conformational space (Yuan *et al*. 2003) predictions from different servers that represent a diverse set of state-of-the-art threading algorithms, i e FUGUE (Cozzetto *et al.* 2009), HHSEARCH (Zhang *et al.* 2009), PROSPECT2 (Ekins *et al.* 2007), SAM-T02 (Becker *et al.* 2006), SPARKS2 (Brylinski *et al.* 2008), SP3 (Arakaki *et al.* 2004), PAINT, PPA-I and PPA-II. Models in the LOMET are selected from individual servers purely based on consensus, i e the structure similarity of the considered model with other threading alignments. For the best performance, 30 models are taken from the top predictions of the nine servers sequentially from:

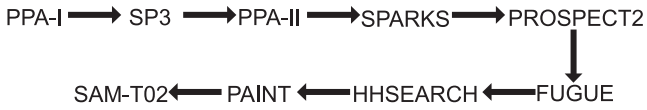


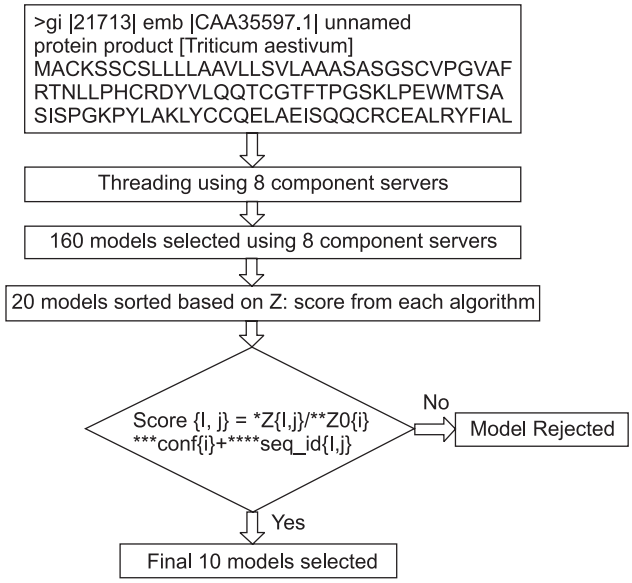
Fig 1 Sequential prediction of nine servers for model selection: LOMETS.

The success of the I-TASSER method in the blind CASP experiments (Kopp *et al.* 2007, Yuan X *et al.* 2003) and the large-scale benchmarking tests (Roy *et al.* 2011) makes it a valuable tool for automated protein structure and function annotation. Global structural comparison of proteins pairs for fold recognition and family assignment can be performed conveniently through structure-based functional assignment (Malmstrom *et al.* 2010, Zhang *et al.* 2006), which in many cases can be directly used to infer function. However, in the recent studies it is gradually recognized that the relationship between structure and function is not always straightforward, as many protein folds/families are known to be functionally promiscuous (Roy *et al.* 2009), and different folds can perform the same function. In a recently developed I-TASSER by Roy A, Kucukural A, Mukherjee S, Hefty P S and Zhang Y, the methodology was extended for annotating the biological function using the predicted protein structures, based on a combination of local and global structural similarities with proteins of known function. Using this method, the biological functions (including ligand-binding sites, Enzyme Commission (EC) numbers and Gene Ontology (GO) terms) of a substantial number of protein targets were correctly identified based on similarities to non-homologous proteins, which otherwise could not have been inferred from sequence or profile-based searches (Altschul *et al.*1997).

MATERIALS AND METHODS

The protein sequence under study (CAA35597.1) was retrieved from NCBI. This protein is present in wheat but only the amino acid sequence is available and was submitted in NCBI on 21 Nov 1989 by Carbonero P Z, E T S Ingenieros Agronomos U P M, Catedra de Bioquimica y Biologia Molecular, E-28040 Madrid, Spain. The fasta sequence is retrieved from (http://www.ncbi.nlm.nih.gov/protein/CAA35597.1). The physio-chemical properties of the desired protein were predicted by using ProtParam.

Further the 2D structure was determined with the help of PSIpred and LOMETS server (steps shown in Flowchart 1).



*z{I,j} is the Z-score of j-th model of i-th server
** {I} is the cutoff of i-th server
***conf{i} is the confidence of i-th server which is defined the average TM-score to native of all predictions in a large-scale benchmark test
****seq_id{I,j} is the sequence identity to query of j-th model of i-th server

Flowchart 1 Sequential Steps of LOMETS for template selection

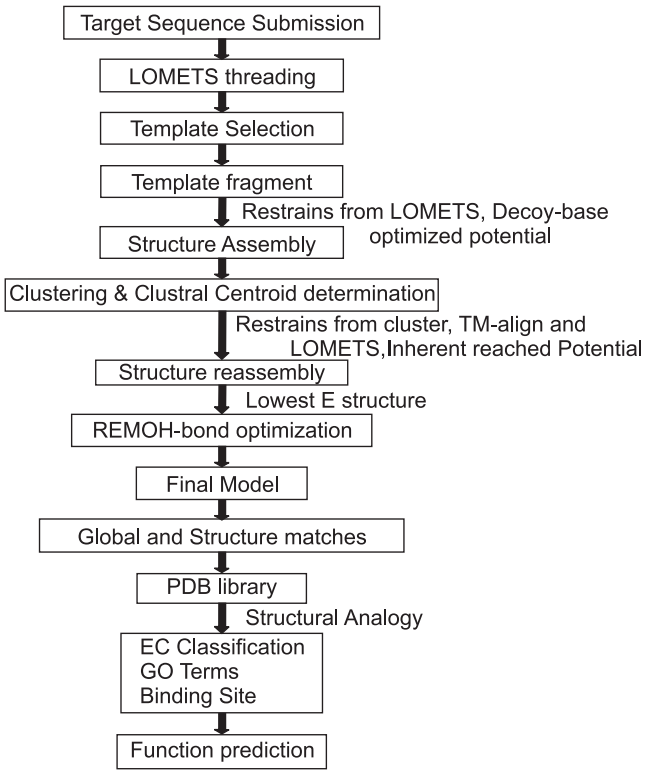
To predict the 3D structure I-TASSER tool (Flowchart 2) was used and for the visualization of the predicted 3D structure RASMOL and Cn3D tools were used. These tools are freely available and can easily be downloaded for offline use. Ligand molecule and Binding site were also determined with the help of I-TASSER (Flowchart 2).

RESULTS AND DISCUSSION

The predicted physio-chemical properties of protein under study were:

Formula: $C_{810}H_{1280}N_{216}O_{231}S_{15,}$ predicted molecular weight 18221.2, Theoretical pI: 7.43, Total number of atoms: 2552, The instability index (II) is computed to be 48.46, Aliphatic index: 95.24 and Grand average of hydropathicity (GRAVY): 0.238

The instability index provides an estimate of the stability of the protein in a test tube. The aliphatic index of a protein



Flowchart 2 A schematic representation of the I-TASSER protocol for protein structure and function prediction (Adopted from Roy *et al.* 2010)

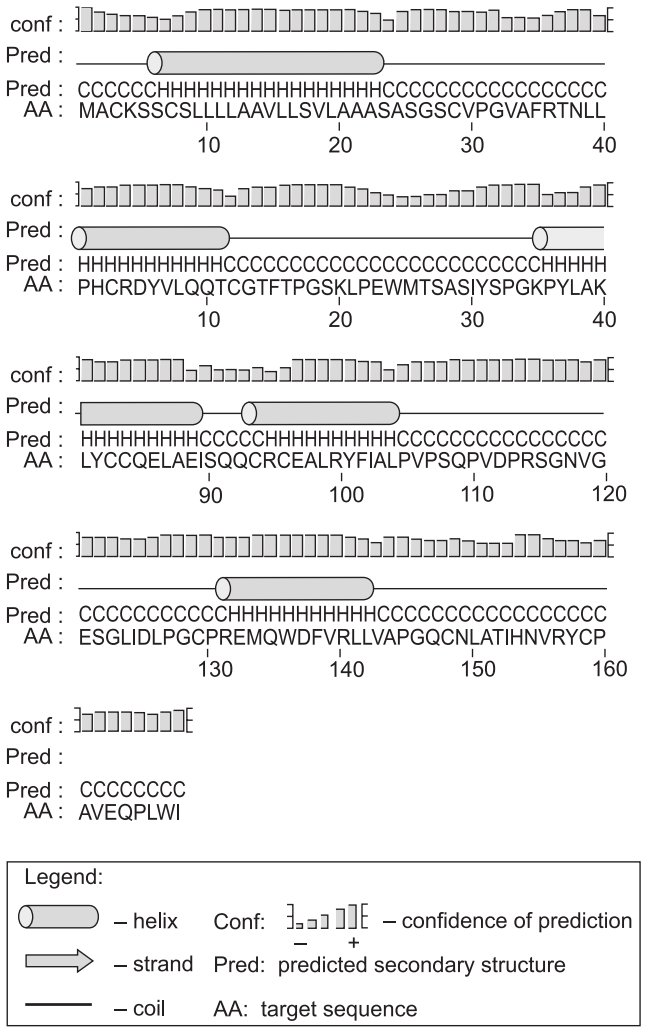


Fig 2 Diagrammatical representation of secondary structure

is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of thermostability of globular proteins.In the present study secondary structure was predicted by PSIPRED. PSIPRED incorporates two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST (Position Specific Iterated BLAST)

2D structure was predicted by PSIPRED (Fig 2) and LOMETS (Local MEta-Threading-Server) after submitting the protein sequence of CAA35597.1, it predicted the secondary structure by threading method. The LOMET server also shows, the top ten model templates predicted by each of the nine individual integrated servers. Out of these, the final top 10 models are selected which have the highest predicted confidence score. The "Confidence Score" column indicates the confidence of the Z-score of the template, the confidence of the particular server and the sequence. This shows identity between the query and the template. The templates whose confidence scores are designated as highest based on the estimated threshold value are considered as the best 10 models for prediction (Table 1). This server also predicts the four spatial restraints, which are the important parameters for the prediction of secondary structure.

These constraints predict parameters like residue number, confidence of contact, distance between the C-alpha pairs and their standard deviation (Table 2), these parameters help predicting the Confidence score of the templates. I-TASSER server predicts the solvent accessibility, i e probable position of the residues in the 3D structure as they are buried or exposed to the surface. This is showed with the help of numerical digits. It most importantly predicts 5 models with highlighted regular secondary structures.

This will help in quickly ascertaining the tertiary structure, class and topology of the query protein from the modeled structure(s). LOMETS servers and I-TASSER based 10 templates are used for predicting the 10 structurally closest structures. The structures are ranked on the basis of TM-score, RMSD value, percentage identity and coverage, but are emphasized on TM-score. Model 1 from the I-TASSER is selected as the best predicted model, wherein a higher score reflects a model of better quality (Fig 3) C-score should be typically in the range [- 5, 2]. In general, models with C-score > - 1.5 have a correct fold (Roy *et al.* 2010) and

Table 1 Shows 10 models with highest Confidence score

Rank	Template	Align_ length	Converage	Z:Score	**Confidence score**	Program	Target-Template alignment	3-D models from threading alignments	Full length models by MODELLER
1	1bluA	117	0.696	16.498	**High**	MUSTER	alignment 1	threading 1	model 1
2	1bea_A	116	0.690	61.570	**High**	HHserch	alignment 2	threading 2	model 2
3	1tmq_B	116	0.690	66.292	**High**	SAM	alignment 3	threading 3	model 3
4	1bea_	116	0.690	27.747	**High**	SP3	alignment 4	threading 4	model 4
5	1bea_	116	0.690	24.260	**High**	SPARKS3	alignment 5	threading 5	model 5
6	1blu_A	117	0.696	31.410	**High**	PPA-I	alignment 6	threading 6	model 6
7	tryp_alpha_ acryl 1bea	116	0.690	19.460	**High**	FUGUE	alignment 7	threading 7	model 7
8	1bluA	117	0.696	7.450	**High**	PROSPECT2	alignment 8	threading 8	model 8
9	1hssa	105	0.625	21.671	**H**igh	SPARKS3	alignment 9	threading 9	model 9
10	1blea_	116	0.690	7.241	**High**	PROSPECT2	alignment 10	threading 10	model 10

Table 2 Spatial restraints

Constraint	Total number of pairs showing constraints
Sidechain contacts	643
C alpha atom contacts	253
Short-range C-alpha atom distances	592
Long-range C-alpha distances	3 348

used for further analysis.

By downloading the PDB file for model 1, Ramachandran plot is developed for the structure validation (Fig 4).The plot developed validates the structure as all the empirically distributed data-points present in the structure are observed to lie in the allowed region. This indicates, the model 1 predicted with the help of I-TASSER have approx. 90% of the residues in the allowed region conformation. Through this we also found the conformation values of φ and ψ angles possible for all amino acid residues present in the protein.

Function of the query protein is based on the functional analogs and the confidence score of the predicted 3D model. Function prediction is based on 3 subsections: Enzyme Commission (EC) numbers, GO terms and ligand-binding sites (Roy *et al.* 2010).

The threshold value of the EC number for the predicted functional analogs should be >1.1 (Roy *et al.* 2010) in other cases, where all the analogs do not cross the threshold value

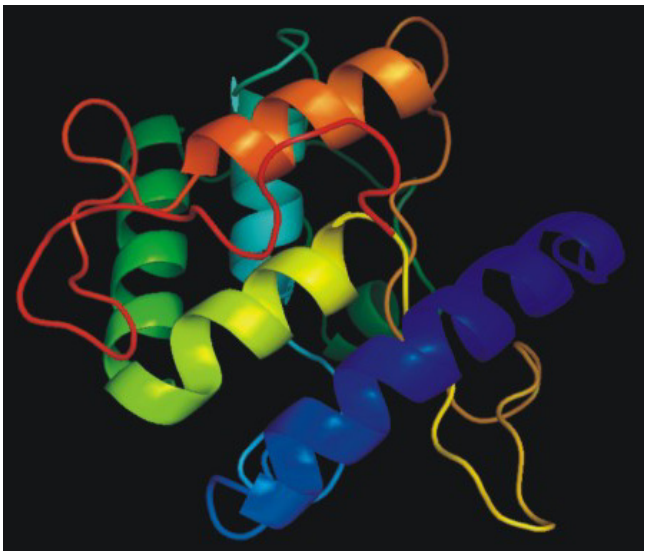


Fig 3 Best predicted model selected based on C-score

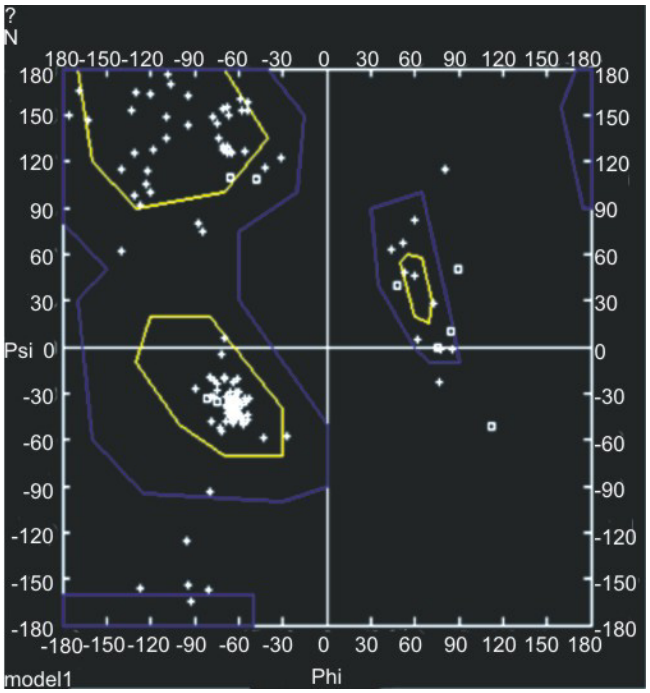
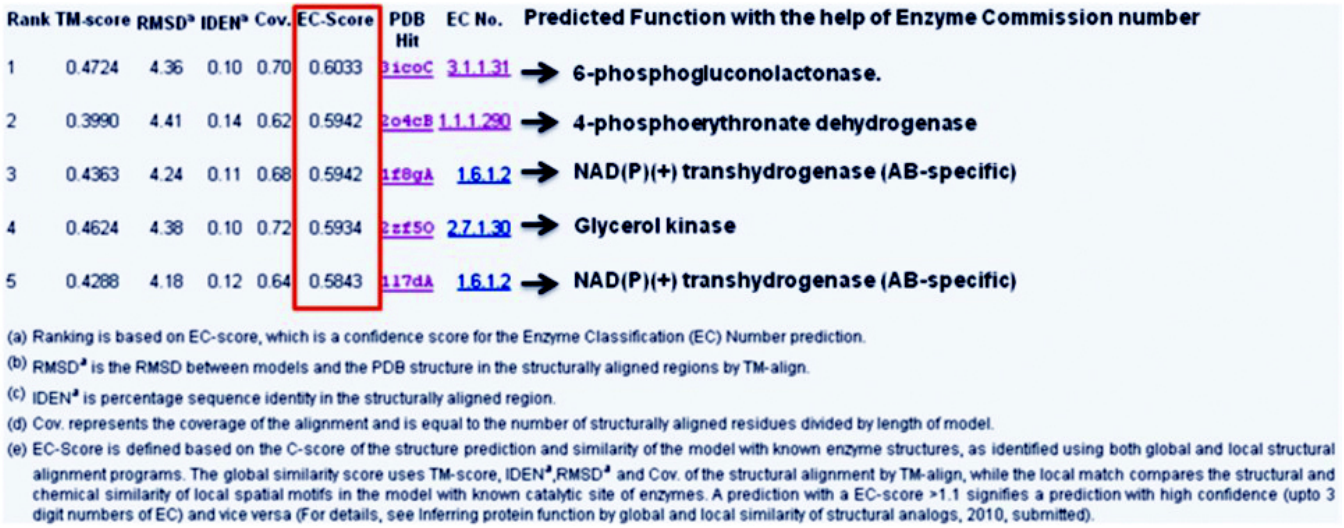


Fig 4 Ramachandran plot showing distribution of amino acids for the predicted Model of CAA35597.1



Rank	TM-score	RMSD[a]	IDEN[a]	Cov.	EC-Score	PDB Hit	EC No.	Predicted Function with the help of Enzyme Commission number
1	0.4724	4.36	0.10	0.70	0.6033	3icoC	3.1.1.31	6-phosphogluconolactonase.
2	0.3990	4.41	0.14	0.62	0.5942	2o4cB	1.1.1.290	4-phosphoerythronate dehydrogenase
3	0.4363	4.24	0.11	0.68	0.5942	1f8gA	1.6.1.2	NAD(P)(+) transhydrogenase (AB-specific)
4	0.4624	4.38	0.10	0.72	0.5934	2zf5O	2.7.1.30	Glycerol kinase
5	0.4288	4.18	0.12	0.64	0.5843	1l7dA	1.6.1.2	NAD(P)(+) transhydrogenase (AB-specific)

(a) Ranking is based on EC-score, which is a confidence score for the Enzyme Classification (EC) Number prediction.

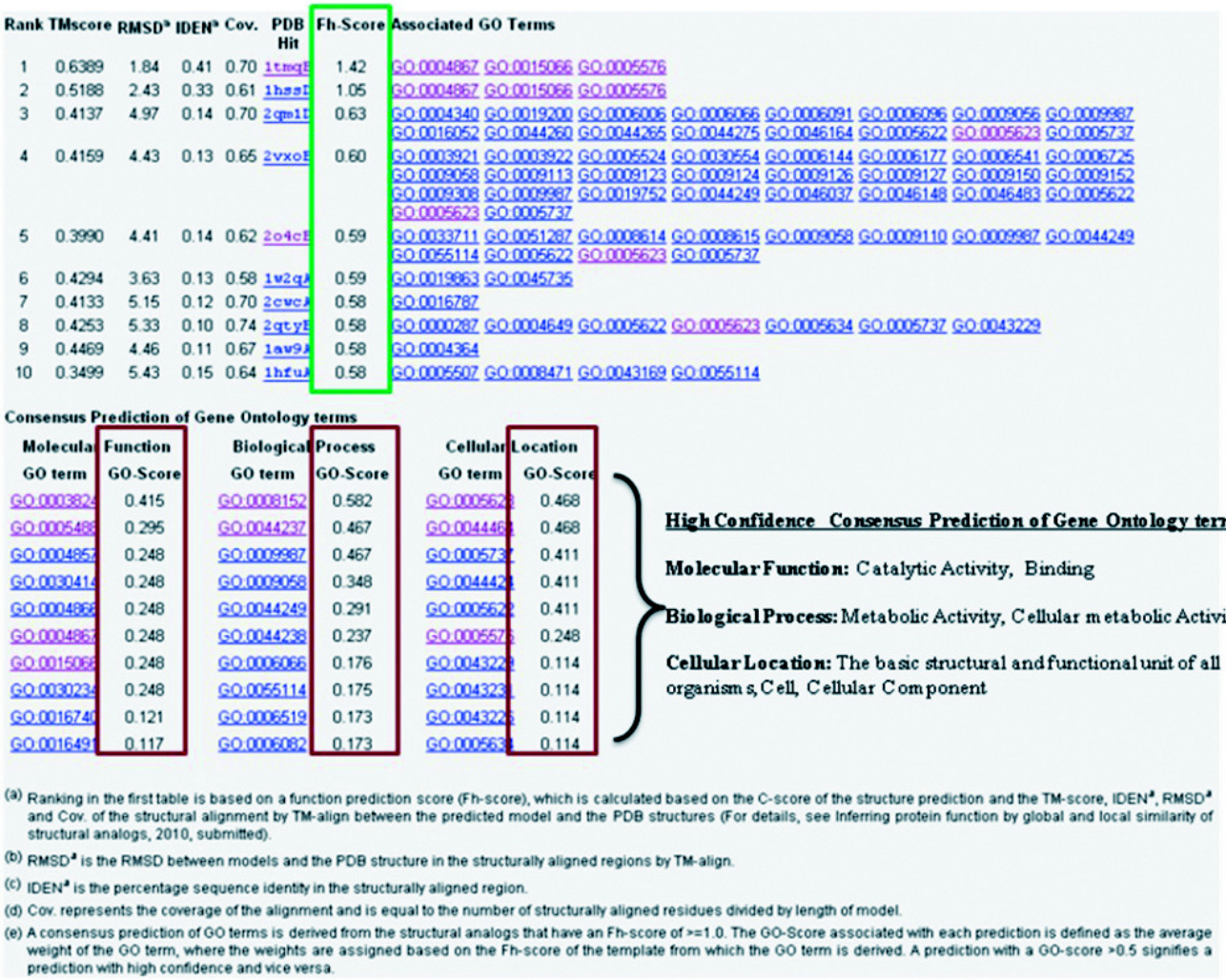
(b) RMSD[a] is the RMSD between models and the PDB structure in the structurally aligned regions by TM-align.

(c) IDEN[a] is percentage sequence identity in the structurally aligned region.

(d) Cov. represents the coverage of the alignment and is equal to the number of structurally aligned residues divided by length of model.

(e) EC-Score is defined based on the C-score of the structure prediction and similarity of the model with known enzyme structures, as identified using both global and local structural alignment programs. The global similarity score uses TM-score, IDEN[a],RMSD[a] and Cov. of the structural alignment by TM-align, while the local match compares the structural and chemical similarity of local spatial motifs in the model with known catalytic site of enzymes. A prediction with a EC-score >1.1 signifies a prediction with high confidence (upto 3 digit numbers of EC) and vice versa (For details, see Inferring protein function by global and local similarity of structural analogs, 2010, submitted).

Fig 5(a) Function prediction: EC Number based



Rank	TMscore	RMSD[a]	IDEN[a]	Cov.	PDB Hit	Fh-Score	Associated GO Terms
1	0.6389	1.84	0.41	0.70	1tmqB	1.42	GO:0004867 GO:0015066 GO:0005576
2	0.5188	2.43	0.33	0.61	1hssB	1.05	GO:0004867 GO:0015066 GO:0005576
3	0.4137	4.97	0.14	0.70	2qm1B	0.63	GO:0004340 GO:0019200 GO:0006006 GO:0006066 GO:0006091 GO:0006096 GO:0009056 GO:0009987 GO:0016052 GO:0044260 GO:0044265 GO:0044275 GO:0046164 GO:0005622 GO:0005623 GO:0005737
4	0.4159	4.43	0.13	0.65	2vxoB	0.60	GO:0003921 GO:0003922 GO:0005524 GO:0030554 GO:0006144 GO:0006177 GO:0006541 GO:0006725 GO:0009058 GO:0009113 GO:0009123 GO:0009124 GO:0009126 GO:0009127 GO:0009150 GO:0009152 GO:0009308 GO:0009987 GO:0019752 GO:0044249 GO:0046037 GO:0046148 GO:0046483 GO:0005622 GO:0005623 GO:0005737
5	0.3990	4.41	0.14	0.62	2o4cB	0.59	GO:0033711 GO:0051287 GO:0008614 GO:0008615 GO:0009058 GO:0009110 GO:0009987 GO:0044249 GO:0055114 GO:0005622 GO:0005623 GO:0005737
6	0.4294	3.63	0.13	0.58	1v2qA	0.59	GO:0019863 GO:0045735
7	0.4133	5.15	0.12	0.70	2cwcA	0.58	GO:0016787
8	0.4253	5.33	0.10	0.74	2qtyB	0.58	GO:0000287 GO:0004649 GO:0005622 GO:0005623 GO:0005634 GO:0005737 GO:0043229
9	0.4469	4.46	0.11	0.67	1aw9A	0.58	GO:0004364
10	0.3499	5.43	0.15	0.64	1hfuA	0.58	GO:0005507 GO:0008471 GO:0043169 GO:0055114

Consensus Prediction of Gene Ontology terms

Molecular Function		Biological Process		Cellular Location	
GO term	GO-Score	GO term	GO-Score	GO term	GO-Score
GO:0003824	0.415	GO:0008152	0.582	GO:0005622	0.468
GO:0005488	0.295	GO:0044237	0.467	GO:0044464	0.468
GO:0004857	0.248	GO:0009987	0.467	GO:0005737	0.411
GO:0030414	0.248	GO:0009058	0.348	GO:0044424	0.411
GO:0004866	0.248	GO:0044249	0.291	GO:0005623	0.411
GO:0004867	0.248	GO:0044238	0.237	GO:0005575	0.248
GO:0015066	0.248	GO:0006066	0.176	GO:0043229	0.114
GO:0030234	0.248	GO:0055114	0.175	GO:0043231	0.114
GO:0016740	0.121	GO:0006519	0.173	GO:0043227	0.114
GO:0016491	0.117	GO:0006082	0.173	GO:0005634	0.114

High Confidence Consensus Prediction of Gene Ontology terms:

Molecular Function: Catalytic Activity, Binding

Biological Process: Metabolic Activity, Cellular metabolic Activity

Cellular Location: The basic structural and functional unit of all organisms, Cell, Cellular Component

(a) Ranking in the first table is based on a function prediction score (Fh-score), which is calculated based on the C-score of the structure prediction and the TM-score, IDEN[a], RMSD[a] and Cov. of the structural alignment by TM-align between the predicted model and the PDB structures (For details, see Inferring protein function by global and local similarity of structural analogs, 2010, submitted).

(b) RMSD[a] is the RMSD between models and the PDB structure in the structurally aligned regions by TM-align.

(c) IDEN[a] is the percentage sequence identity in the structurally aligned region.

(d) Cov. represents the coverage of the alignment and is equal to the number of structurally aligned residues divided by length of model.

(e) A consensus prediction of GO terms is derived from the structural analogs that have an Fh-score of >=1.0. The GO-Score associated with each prediction is defined as the average weight of the GO term, where the weights are assigned based on the Fh-score of the template from which the GO term is derived. A prediction with a GO-score >0.5 signifies a prediction with high confidence and vice versa.

Fig 5(b) Function prediction: GO terms

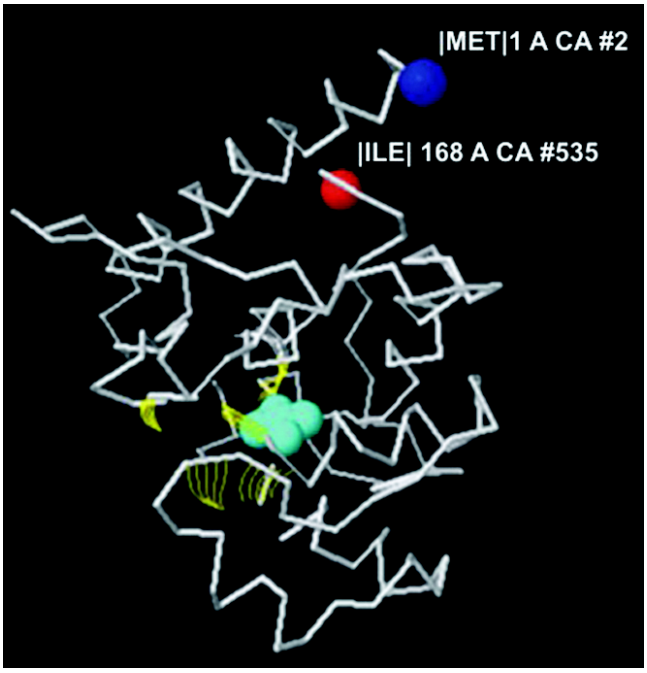


Fig 5 (c): Predicted binding site

Abbreviations: ARG: arginine, LEU: leucine, GLU: glutamic acid, CYS: cysteine, ALA: alanine

the function are just kept for reference (Fig 5a) but does not give accurate result and therefore, other parameters are checked. In GO terms Fh-score is a strong indicator of functional similarity between the predicted structure and detected analogs. Taking under consideration Fh-score cutoff of 0.8, an overall accuracy of 56% is achieved. However, because the function of proteins is multi-faceted and the unanimity of functionalities of the identified analogs usually yields a more reliable prediction, a consensus between a GO term and its ancestor terms in the ontology has been proven to be a more reliable indicator of the GO terms. Predicted consensus are predicted from the functional analogs having Fh-score >1.0. For predicted function (Fig 5b). Third subsection of the function prediction, predicts the ligand binding site (Fig 5c). The backbone of the model in the image is shown in white solid lines. Ligand atoms are highlighted in cyan, whereas the binding sites residues of the query protein are shown as "strand" in yellow.

The above analyses show that the unknown query protein of wheat is similar to 1tmqB and 1bluA of *Tenebrio molitor* (Yellow mealworm) and *Allochromatium vinosum* respectively**.** The 3D structure having maximum C-value -0.521 is considered as the best predicted model out of five for CAA35597.1. As depicted through EC number, GO terms and consensus prediction of GO terms, the function of CAA35597.1 is divided into three parts:

- *Molecular function:* Catalytic activity, binding, inhibitor activity
- *Biological process:* Metabolic activity, cellular metabolic activity
- *Cellular location:* The basic structural and functional unit of all organisms, cell, cellular component.

According to protein ontology analysis the predicted protein is involved in catalytic, binding and inhibitory activities therefore further binding site analysis was performed. Based on TM-Score (0.4389), BS-Score (1.1056), Identity (0.0460) and RMSD (5.12), I-TASSER predicted ALA as a ligand (shown in Cyan), which interacts with 36: ARG 39: LEU 86: GLU 89: GLU 95: ARG and 96: CYS (shown in "strand" in yellow see Fig 5c).

TM-score is used to rank the structural analogs. The structural analogs with a TM-score > 0.5 can be used for determining the structure class/protein family of the predicted query protein. BS-score is a measure of local sequence and structural similarity between template's binding site and predicted binding site in the query structure based on large scale benchmarking analysis, binding site prediction with BS-score >1.1 signify prediction with high confidence. Identity is the percentage sequence identity in the structurally aligned region. The various scores calculated for the query protein structure were within the favourable range.

REFERCENCES

Altschul S F *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3 389–402.

Arakaki A K *et al.* 2004. Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics* **20**: 1 087–96.

Battey J N *et al.* 2007. Automated server predictions in CASP7. *Proteins* **69:** 68–82.

Becker O M *et al.* 2006. An integrated *in silico* 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. *Journal of Medicinal Chemistry* **49**: 3 116–35.

Brenner S E *et al*. 2000. Expectations from structural genomics. *Protein Science* **9**: 197–200.

Brylinski M *et al*. 2008. Q-Dock: low-resolution flexible ligand docking with pocket-specific threading restraints. *Journal of Computational Chemistry* **29**(10): 1 574–88.

Cozzetto D *et al.* 2009. Evaluation of template-based models in CASP8 with standard measures. *Proteins* **77**(9): 18–28.

Ekins S *et al*. 2007. *In silico* pharmacology for drug discovery: applications to targets and beyond. *British Journal of Pharmacology* **152**: 21–37.

Jauch R *et al*. 2007. Assessment of CASP7 structure predictions for template free targets. *Proteins* **69**: 57–67.

Kopp J *et al*. 2007. Assessment of CASP7 predictions for template-based modeling targets. *Proteins* **69**: 38–56.

Kumar P A *et al.* 2011 Biotechnology and crop improvement. *Indian Journal of Agricultural Sciences* **81**(9): 787–800.

Liwo A *et al*. 1999. Protein structure prediction by global optimization of a potential energy function. *Proceeding of the National Academy of Sciences USA* **96:** 5 482–5.

Malmstrom L *et al.* 2007. Superfamily assignments for the yeast

proteome through integration of structure prediction with the gene ontology. *PLoS Biology* **5**: 725–38.

Roy A *et al*. 2011. A protocol for computer-based protein structure and function prediction. *Journal of Visualized Experiments* **57**, e3259.

Roy A *et al*. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols* **5**: 725–38.

Roy A *et al*. 2009. Molecular and structural basis of drift in the functions of closely-related homologous enzyme domains: implications for function annotation based on homology searches and structural genomics. *In Silico Biology* **9**: 41–55.

Simons K T *et al*. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology* **268**: 209–25.

Skolnick J *et al*. 2000. Structural genomics and its importance for gene function analysis. *Nature Biotechnology* **18**: 283–7.

Stevens R C *et al*. 2001. Global efforts in structural genomics. *Science* **294**: 89–92.

Terwilliger T C *et al*. 1998. Class-directed structure determination: foundation for a protein structure initiative. *Protein Science* **7**: 1 851–6.

Wu S *et al*. 2007. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biology* **5**(17): 1–10.

Yuan X *et al*. 2003. Ab initio protein structure prediction using pathway models. *Comparative and Functional Genomics* **4**: 397–401.

Zhang Y *et al*. 2006. Structure modeling of all identified G protein-coupled receptors in the human genome. *PLoS Computational Biology* **2**(2): 88–99.

Zhang Y *et al*. 2003. TOUCHSTONE II: a new approach to Ab-initio protein structure prediction. *Biophysical Journal* **85**: 1 145–64.

Zhang Y. 2007. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **69**: 108–17.

Zhang Y. 2009. Protein structureprediction: when is it useful? *Current Opinion in Structural Biology* **19**: 145–55.