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

GEETIKA JETHRA1, A K MISHRA2, P S PANDEY3 and H CHANDRASEKHARAN4

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
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 (Cuzzetto 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:

\[
\text{PPA-I} \rightarrow \text{SP3} \rightarrow \text{PPA-II} \rightarrow \text{SPARKS} \rightarrow \text{PROSPECT2} \rightarrow \text{SAM-T02} \rightarrow \text{PAINT} \rightarrow \text{HHSEARCH} \rightarrow \text{FUGUE}
\]

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).

\[
\text{Score} (l, j) = \frac{Z(l, j)}{Z(0)} \times \frac{\text{conf}(l)}{\text{conf}(j)} \times \frac{\text{seq_id}(l, j)}{\text{seq_id}(j)}
\]

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
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
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 \( \phi \) and \( \psi \) 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 5(a) Function prediction: EC Number based

Fig 5(b) Function prediction: GO terms
CAA35597.1 is divided into three parts: consensus prediction of GO terms, the function of CAA35597.1. As depicted through EC number, GO terms is considered as the best predicted model out of five for respectively (Yellow mealworm) and "strand" of wheat is similar to 1tmqB and 1bluA of query protein are shown as "strand" in yellow. Ligand atoms are binding site (Fig 5c). The backbone of the model in the subsection of the function prediction, predicts the ligand 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.

**REFERENCES**


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