GLOBAL JOURNAL OF BIO-SCIENCE AND BIOTECHNOLOGY

www.scienceandnature.org

COMPARATIVE MODELING OF CELLULASE IN *ASPERGILLUS TERREUS*

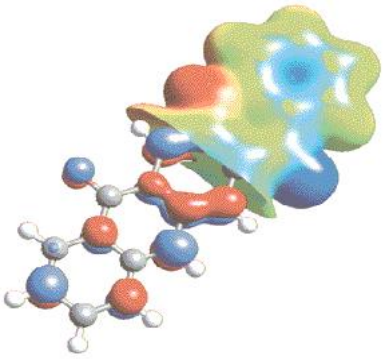
S. Maragathavalli, S.V. Megha, S. Brindha, V. Karthikeyan & B. Annadurai
Research and development centre, Bharathiyar University, Coimbatore-641 041

ABSTRACT
The widely distributed hydrolytic enzyme of wide application is cellulose. It is involved in conversion of biomass into simpler sugars. Homology modeling is used to predict the 3D structure of a unknown protein based on the known structure of a similar protein. During evolution, sequence changes much faster than structure. It is possible to identify the 3D structure by looking at a molecule with some sequence identity. One can predict the 3D structure with how much sequence identity is needed with a certain number of aligned residues, to reach the safe homology modeling zone for a sequence of under residues, for example, a sequence identity of 40% is sufficient for structure prediction. When the sequence identity false in the safe homology modeling zone. We can assume that the 3D structure of both sequences is the same. The known structure is called target homology modeling of the target structure can be done in seven steps namely (template recognition) FASTA, from and BLAST from EMBL – EBI and NCBI and initiation alignment by using PDB with the help of BLAST alignment correction using clustal W (SCRs), Back Bone generation, loop modeling, side chain modeling, model optimization and model validation. By this, the cellulose protein is modeled by SPDB Viewer with the help of template, lib4 derived from PDB and visualized in rosmol finally the modeled cellulose protein is checked by the WHAT IF SERVER and the results had showed.

KEY WORDS: Cellulase, SWISS-MODEL, PDB, BLAST, FASTA, SMTL.

INTRODUCTION

Homology modeling is the attempt to create a 3 dimensional protein structures given its amino acid sequence and a structural template. The other name for homology modeling is 'comparative modeling'. If the sequence similarity between the unknown and the template are sufficiently high (>50%), the procedure can automate with reasonable results. Threading techniques have also be used with good results for molecules that are structurally similar.



The Homology Module allows you to build a 3D model of a protein based on the 3D structure or structures of one or more homologous proteins. The protein with the undetermined structure that you want to model is called the "model", "unknown", or "sequence" protein. The protein(s) with known 3D structures is/are are referred to as the "reference" or "real" protein(s).

Why homology modeling

- Rate of structure solving through NMR or X-ray is slow compared to the deposition of DNA and Protein sequences
- Crystallization is the bottleneck (time in months)
- No generic receipe for crystallization
- Membrane proteins are difficult to crystallize
- 30% of proteome of living things

Knowledge of 3D structure is essential for the understanding of the protein function

- Structural information enhances our understanding of protein-protein or protein-DNA interactions
- Primary and Structural information of proteins are stored in different places (SWISS-PROT, PDB)
- Rate of publication of primary sequence has increased dramatically
- Swiss-Prot Release 41.24 of 19-Sep-2003 134,343 entries; PDB as of 16-Sep-03 has 22,516 structures
- Traditional structure solving methods are slow-NMR
- NMR limited by size of the molecule; Use of H signatures creates noise

Knowledge of three-dimensional structure is a prerequisite for the rational design of site-directed mutations in a protein and can be of great importance for the design of drugs. Structural information often greatly enhances our understanding of how proteins function and how they interact with each other or it can, for example, explain antigenic behaviour, DNA binding specificity, etc. X-ray crystallography and NMR spectroscopy are the only ways to obtain detailed structural information. Unfortunately, these techniques involve elaborate technical procedures

and many proteins fail to crystallize at all and/or cannot be obtained or dissolved in large enough quantities for NMR measurements. The size of the protein is also a limiting factor for NMR. In the absence of experimental data, model-building on the basis of the known three dimensional structure of a homologous protein is at present the only reliable method to obtain structural information. Comparisons of the tertiary structures of homologous proteins have shown that three-dimensional structures have been better conserved during evolution than protein primary structures, and massive analysis of databases holding results of these three dimensional comparison methods, as well as a large number of well-studied examples [e.g. 50-58] indicate the feasibility of model building by homology.

Steps Involved

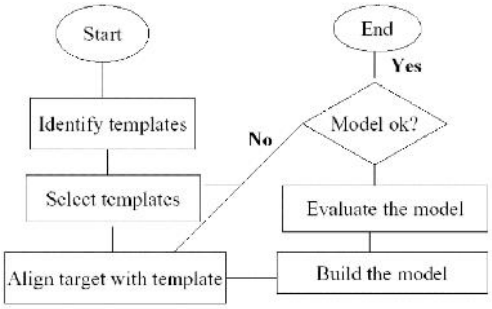
- Similiarity Search (sequence or structural)
- Sequence Alignment
- Structural Alignment
- Selecting the Templates
- Model Building
- Evaluating the Model(s)

Although it is simpler to use a single reference protein, It is much more reliable to use several reference proteins. This is because comparison of the several known structures allows you to identify regions of structural conservation in addition to regions of sequence conservation. There are many programs available for homology modelling. The easiest method is to use a webserver such as

-SWISS-MODELLER

3D-Jigsaw homology or What-IF

Procedures for Homology Modeling



A homology modeling routine needs three items of input:

1. The sequence of the protein with unknown 3D structure, the "target sequence".
2. A 3D template is chosen by virtue of having the highest sequence identity with the target sequence. The 3D structure of the template must be determined by reliable empirical methods such as crystallography or NMR, and is typically a published atomic coordinate "PDB" file from the Protein Data Bank.

Uses of homology models

Successful predictions based on homology models have been reviewed by Baker and Sali. The positions of conserved regions of the protein surface can help to identify putative active sites and binding pockets. If the ligand is known to be charged, the binding site may be predicted by searching the surface for a cluster of complementary charges. The size of a ligand may be predicted from the volume of the putative binding pocket. In one case, relative affinities of a series of ligands have been predicted. Such predicions are useful to guide mutagenesis experiments.

Cellulase enzyme is commercially has wide application. Hence, an urgent need for this enzyme to be characterized in all aspect to understand the structural and functional relations. The presence study is to identify the more efficient celllulolytic enzyme producing microorganism for Biopolyshing using the computational analysis Protein sequence of cellulase is retrieved from NCBI and where subjected to Protparam to analyze physicochemical parameters, Secondary structure prediction using GOR IV and SOPMA, Homology modeling (Swiss model), Phylogenetic analysis and active site prediction site by SCFBIO.

Sequence retrieval and alignment

Cellulase protein sequence of *Clostridium thermocellum* [AAA23226.1] was retrieved from the National Center for Biotechnology (NCBI) and made as the query sequence for the structure, properties prediction and modeling. Blastp was performed and obtained nine similar sequences of different strains. Clustal W multiple sequence alignment was done for those sequences using BioEdit5.0.

Goto:http://swissmodel.expasy.org/repository/http://swissmodel.expasy.org/workspace/http://www.proteinmodelportal.or

Three dimensional protein structures are crucial for understanding protein function at a molecular level. In recent years, tremendous progress in experimental techniques for large-scale protein structure determination by X-ray crystallography and NMR has been achieved. Structural genomic efforts have contributed significantly to the elucidation of novel protein structures (Levitt, 2007), and to the development of technologies, which have increased the speed and success rate at which structures can be determined and lowered the cost of the experiments (Slabinski *et al.*, 2007, Manjasetty *et al.*, 2008). However, the number of known protein sequences grows at an ever higher rate as large-scale sequencing projects, such as the Global Ocean Sampling expedition, are producing sequence data at an unprecedented rate (Yooseph, 2007). Consequently, the last release of the UniProt (Bairoch *et al.*, 2005) protein knowledgebase (version 14.0) contained more than 6.5 million sequences, which is about 100 times the number protein structures currently deposited in the Protein Data Bank (Berman *et*

al., 2007) (~53 000, September 2008). For the foreseeable future, stable and reliable computational approaches for protein structure modelling will therefore be required to derive structural information for the majority of proteins, and a broad variety of *insilico* methods for protein structure prediction has been developed in recent years.

Homology (or comparative) modelling techniques have been shown to provide the most accurate models in cases, where experimental structures related to the protein of interest were available. Although the number of protein sequence families increases at a rate that is linear or almost linear with the addition of new sequences (Yooseph *et al*., 2007), the number of distinct protein folds in nature is limited (Levitt, 2007; Chothiac, 1992) and the growth in the complexity of protein families appears as a result of the combination of domains (Levitt, 2007). Achieving complete structural coverage of whole proteomes (on the level of individual soluble domain structures) by combining experimental and comparative modelling techniques therefore appears to be a realistic goal, and is already been pursued, *e.g.* by the Joint Center for Structural Genomics for the small model organism *Thermotogamaritima* (JCSG) (McCleverty *et al*., 2008, Xu *et al*., 2008). Assessment of the accuracy of methods for protein structure prediction, *e.g.* during the bi-annual CASP (Critical Assessment of Techniques for Protein Structure Prediction) experiments (Kopp *et al*., 2007; Kryshtafovych *et al*., 2007) or the automated EVA project (Koh *et al*., 2003), has demonstrated that comparative protein structure modelling is currently the most accurate technique for prediction of the 3D structure of proteins. During the CASP7 experiment, it became apparent that the best fully automated modelling methods have improved to a level where they challenge most human predictors in producing the most accurate models (Battey *et al*., 2007; Soding, 2005; Zhang, 2007). Nowadays, comparative protein structure models are often sufficiently accurate to be employed for a wide spectrum of biomedical applications, such as structure based drug design (Hillisch *et al.*, 2004; Tan *et al.*, 2008; Thorsteinsdottir et al.,2006; Vangrevelinghe, 2003), functional characterization of diverse members of a protein family (Murray *et al.*, 2005), or rational protein engineering, e.g. the humanization of therapeutic antibodies, or to study functional properties of proteins (Lippow *et al*., 2007; Junne *et al*., 2006; Peitch 2002; Trammontano 2008; Li *et al*., 2007).

Here, we describe the SWISS MODEL Repository, a database of annotated protein structure models generated by the SWISS MODEL Pipeline, and a set of associated web-based services that facilitate protein structure modelling and assessment. We emphasize the improvements of the SWISS MODEL Repository which have been implemented since our last report (Koop and Schwede, 2006). These include a new pipeline for template selection, the integration with interactive tools in the SWISS MODEL Workspace, the programmatic access via DAS (distributed annotation system) (Jenkinson *et al*., 2008), the implementation of a reference frame for protein sequences based on md5 cryptographic hashes, and the integration with the Protein Model Portal (http://www. proteinmodelportal.org) of the PSI Structural Genomics Knowledge Base (Berman *et al*., 2008; Berman, 2008).

Homology modelling

The SWISS MODEL Repository contains models that are calculated using a fully automated homology modelling pipeline. Homology modelling typically consists of the following steps: selection of a suitable template, alignment of target sequence and template structure, model building, energy minimization and/or refinement and model quality assessment. This requires a set of specialized software tools as well as up-to-date sequence and structure databases. The SWISS MODEL pipeline (version 8.9) integrates these steps into a fully automated workflow by combining the required programs in a PERL based framework. Since template search and selection is a crucial step for successful model building, we have implemented a hierarchical template search and selection protocol, which is sufficiently fast to be used for automated large-scale modelling, sensitive in detecting low homology targets, and accurate in correctly identifying close target structures. In the first step, segments of the target sequence sharing close similarity to known protein structures are identified using a conservative BLAST (Altchul *et al*., 1997) search with restrictive parameters [*E*-value cut-off: 10^{-5}, 60% minimum sequence identity to sequences of the SWISS MODEL Template Library SMTL (Arnold *et al*., 2006). This ensures that information about close sequence relationships is not dispersed by the subsequent profile based search strategies (Sadowski and Jones, 2007). If regions of the target sequence remain uncovered, in the second step a search for suitable templates is performed against a library of Hidden Markov Models for SMTL using HHSearch (Soding, 2005).Templates resulting from both steps are ranked according to their *E*-value, sequence identity to the target, resolution and structure quality. From this ranked list, the best templates are progressively selected to maximize the length of the modelled region of the protein. New templates are added if they significantly increase the coverage of the target sequence (spanning at least 25 consecutive residues), or new information is gained (*e.g.* templates spanning several domains help to infer relative domain orientation). For each selected target–template alignment, 3D models are calculated using ProModII (Guex and peitch, 1997) and energy minimized using the Gromos force field (Van Gunsteren *et al*., 1996). The quality of the resulting model is assessed using the ANOLEA mean force potential (Melo and Feytmans, 1998).

Depending on the size of the protein and the evolutionary distance to the template, model building can be relatively time consuming. Therefore, comprehensive databases of precomputed models (Koop and Schwede 2006; Koop and Schwede 2004; Pieperu *et al*., 2006) have been developed in order to be able to cross-link real-time model information with other biological data resources, such as sequence databases or genome browsers.

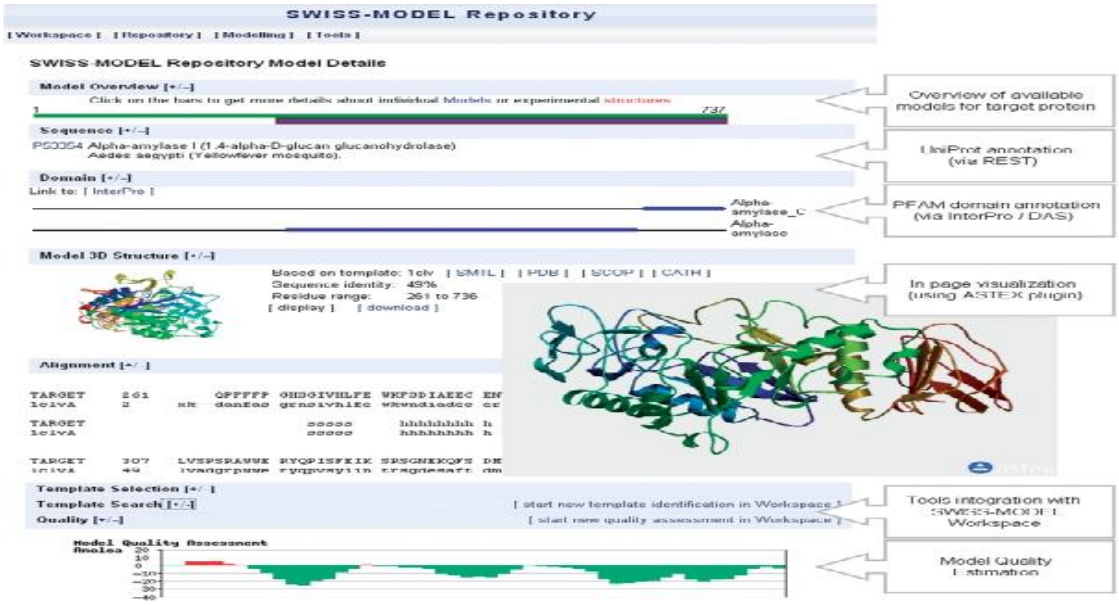
Model database

The SWISS MODEL Repository is a relational database of models generated by the automated SWISS MODEL pipeline based on protein sequences from the UniProt database (Bairoch *et al*., 2007). Within the database, model target sequences are uniquely identified by their md5 cryptographic hash of the full length raw amino acid

sequence. This mechanism allows the redundancy in protein sequence databases entries to be reduced, and facilitates cross-referencing with databases using different accession code systems. Mapping between UniProt and various database accession code systems to our md5 based reference system is derived from the iProClass database (Huang *et al*., 2007). Regular updates are performed for all protein sequences in the SwissProt database (Boulet *et al*., 2007), as well as complete proteomes of several model organisms (*Homo sapiens, Mus musculus, Rattus norvegicus, Drosophila melanogaster, Arabidopsis thaliana, Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Caenorhabditis elegans* and *Hepaci virus*). Incremental updates are performed on a regular basis in order to both include new target sequences from the UniProt database and to take advantage of newly available template structures, whereas full updates are required when major improvements to the underlying modelling algorithms have been made. The current SWISS MODEL Repository release contains 3.45 million models for 2.72 million unique sequences, built on 26 185 different template structures (34 540 chains), covering 48.8% of the entries from UniProt (14.0), and more specifically 65.4% of the unique sequences of Swiss-Prot (56.0), the manually annotated section of the UniProt knowledgebase. The size of the models ranges from 25 up to 2059 residues (*e.g.* fatty acid synthase β-subunit from *Thermomyces lanuginosus*) with an average model length of 221 residues.

Graphical user web interface

The web interface at http://swissmodel. expasy. org/ repository/ provides the main entry point to the SWISSMODEL Repository. Models for specific proteins can be queried using different database accession codes (*e.g.* UniProt AC and ID, GenBank, IPI, Refseq) or directly with the protein amino acid sequence (or fragments thereof, *e.g.* for a specific domain). For a given target protein, a graphical overview illustrating the segments for which models (or experimental structures) are available is shown (Figure 1). Functional and domain annotation for the target protein is retrieved dynamically in real time using web service protocols to ensure that the annotation information is up-to-date.



UniProt annotation of the target protein is retrieved via REST queries (http://www.uniprot.org). Structural domains in the target protein are annotated by PFAM domain assignment (Fim *et al*., 2008), which is retrieved dynamically by querying the InterPro (Mulder and Apweiler, 2008) database using the DAS protocol (Jenkinson et al 2008). The md5-based reference frame for target proteins allows to update the database accession mappings in between modelling release cycles. This ensures that cross references with functional annotation resources such as InterPro correspond to proteins of identical primary sequence, thereby avoiding commonly observed problems with incorrect cross-references as a result of instable accession codes or asynchronous updates of different data resources. Finally, for each model, a summary page provides information on the modelling process (template selection and alignment), model quality assessment by ANOLEA (Melo and Feytmans, 1998) and Gromos (Van Gunsteren *et al*., 1996), and in page visualization of the structure using the Astex Viewer (Hartshom, 2002) plugin. Typical view of a SWISSMODEL Repository entry. For the UniProt entry P53354, the α-amylase I (EC 3.2.1.1; 1,4-α-D-glucanglucanohydrolase) from

Integration with SWISS-MODEL Workspace

The SWISSMODEL Repository is a large-scale database of precomputed 3D models. Often however, one may be interested in performing additional analyses either on the models themselves, or on the underlying protein target sequence. We have therefore implemented a tight link between the entries of the SWISSMODEL Repository and the corresponding modules in the SWISSMODEL Workspace, which provides an interactive web based, personalized working environment (Arnold *et al*., 2006;

Guex and Peitch 1997; Schwede *et al*., 2003). Besides the functionality for building protein models it provides various modules to assess protein structures and models. The estimation of the quality of a protein model is an important step to assess its usefulness for specific applications. In particular, models based on template structures sharing low sequence identity require careful evaluation. Therefore, entries from the Repository can be directly submitted to the Workspace for quality assessment using different global and local quality scores such as DFire (Zhou and Zhou, 2002), ProQRes (Wallner and Elofsson, 2006) or QMEAN (Benkert *et al*., 2008).

The default output format for models in the Repository is the project file for the program Deep View (Guex and Peitsch 1997); this program allows the underlying alignments to be adjusted manually and for the request to be resubmitted to Workspace for modelling. While new protein structures are deposited in the PDB on a daily basis, the respective modelling update cycles are more infrequent, resulting in a delay in the incorporation of new templates. The Repository therefore links directly to the corresponding template search module in Workspace, which allows searches for newly released templates to be performed. The direct cross-linking between Repository and Workspace allows combining the advantages of the database of pre-computed models with the flexibility of an interactive modelling system.

The DAS-Server of the SWISSMODEL Repository is based on the DAS/1 standard and can be queried by primary UniProt accession codes or md5hashs of the corresponding sequences. Individual models for a query sequence ('SEGMENT') are annotated as 'FEATURE', with information about the start and stop position in the target sequence, template-sequence identity and the URL to the corresponding SWISSMODEL Repository entry. The DAS service allows the SWISSMODEL Repository to be cross linked with other resources using the same standards, *e.g.* genome browsers. The SWISSMODEL Repository DAS service is accessible at http://swissmodel. expasy.org/service/das/swissmodel/.

The protein model portal

One of the major bottlenecks in the use of protein models is that, unlike for experimental structures, modelling resources are heterogeneous and distributed over numerous servers. However, it is often beneficial for the user to directly compare the results of different modelling methods for the same protein. We have therefore developed the protein model portal (PMP) as a component of the PSI structural genomics knowledge base (Berman *et al*., 2008, Berman 2008). This resource provides access to all structures in the PDB, functional annotations, homology models, structural genomics protein target tracking information, available protocols and the potential to obtain DNA materials for many of the targets. The PMP currently provides access to several million pre-built models from four PSI centers, ModBase (Pieper *et al*., 2006) and SWISS-MODEL Repository (Koop and Schwede, 2004, 2006).

Future Direction

SWISSMODEL Repository will be updated regularly to reflect the growth of the sequence and structure databases. Future releases of SWISSMODEL Repository will include models of oligomeric assemblies, as well as models including essential co-factors, metal ions and structural ligands. Structural clustering of the Swiss Model Template Library will also allow us to routinely include ensembles of models for such proteins, which undergo extensive domain movements.

Secondary structure and physicochemical characterization cellulose

The sequences obtained were analyzed using various softwares available in the ExPASyserver (Web: Proteomic tools Expasy). The GOR IV analysis was performed to understand the presence of helices, beta turns and coils in the protein structure (Bioinformatic Tools for Protein structure analysis and visualsation). Self-optimized prediction method with alignment (SOPMA) analysis was done for analyzing the structural components (Geourjon and Deleage, 1994). Comparison was made between the GOR IV and SOPMA analysis results. ProtParam software analysis was done to understand about the amino acid composition, molecular weight, instability index, aliphatic index and grand average of hydropathicity (GRAVY) (Web: Proteomic tools. Expasy). Hydropathy plot analysis for all cellulase sequences was performed and the nature of amino acid residues was studied using ProtScale (Web: Proteomic tools. Expasy).based on Kyte and Doolittle scale.

Homology modeling of cellulase

Homology models were predicted using SWISS-MODEL (Arnold *et al*., 2006; Kiefer *et al*., 2009 and Peitsch, 1995) and the quality was analysed using VMD 1. 9.1 (Humphery *et al*., 1996). RMSD values were calculated using the RMSD calculator and the best homology model was selected. Ramachandran plot for the best predicted model was depicted by RAMPAGE software (Lovell *et al*., 2003).

Phylogenetic analysis

Phylogenetic relation among the aligned cellulase sequences obtained from Blastp were analyzed based on neighbor joining method (Saitou and Nei, 1987) using MEGA 4.0 (Tamura et al, 2007). The cellulase sequence of *C. thermocellum* [AAA23226.1] was considered as the root taxon for the analysis. Confidence level was analyzed using bootstrap of 1000 replications.

Activity validation by active site comparison:

Active sites of the predicted models and the template were analyzed using Automated Active Site prediction AADS server of SCFBio (Tanya *et al*., 2011). Amino acid compositions of all the cavities were analyzed and the frequency of amino acid occurrence in the cavities of each models were analyzed.

Blast analysis and sequence retrieval:

The cellulase protein sequence of *Clostridium thermocellum* [AAA23226.1] was used as query sequence and nine sequences were obtained by performing Blastp. Multiple sequence alignment was done in BioEdit software and further used for phylogenetic analysis in MEGA.

Secondary structure and physicochemical analysis

SOPMA and GOR IV were used to predict the secondary structure, percentage of alpha, extended and random coils of cellulase producing microorganism are presented Table

1 (see supplementary material). SOPMA analysis for the structure prediction was also done and obtained the percentage of alpha, extended, beta and random coils (Table 1). The secondary structure indicates whether a given amino acid lies in a helix, strand or coil (Jyotsna *et al*., 2010; Ojeiru et al, 2010). SOPMA was used for structure prediction of cellulase protein (Pradeep *et al*., 2012). Random coil dominates the other forms in the cellulase analyzed by SOPMA and GOR IV. It was identified that random coils of *M.abomyces* (58.72%) and *T. longibrachiatum* (57.88%) were dominant compared to other forms. However, followed by random coils, extended forms ranging from (10%-27%) was dominant over α and β helix. All the cellulases analyzed, α-helix was ranging from (13%-37%) dominates β-helix, which had less percentage of conformation (4%-10%).

ProtParam analysis was performed and the number of amino acid residues, molecular weight, pI value, aliphatic index and GRAVY index was obtained for each sequence Table 2 (see supplementary material).

Comparison of the amino acid residue occurrence in cellulase sequences were done and the most dominant residues were highlighted Table 3 (see supplementary material). It was found that molecular weight ranging from 25-127 kDa and it was higher in *C. thermocellum* (83 kDa) and lower in *M. albomyces* (25kDa). Comparing to the eukaryotic cellulase available, the higher aliphatic index of up to 97.51 was noted in *T. subterraneus* strains which indicate their stability over a wide range of temperatures. GRAVY value was negative in all species studied. It was notable that the bacterial strains had lower GRAVY values indicating the better possibilities of aqueous interaction. pI value showed that cellulase is acidic in nature. *T. subterraneus* had a slightly neutral pI value and the highest GRAVY value. Generally it was observed that towards acidic pI values the GRAVY tends to be low. In eukaryotic cellulases, the occurrence of α helices was found to be too low. In case of *A. bisporus*, α helices was similar to that of lower taxonomic groups. Moreover these cellulases possess higher percentages of random coils. A general pattern of inverse relationship between the percentage of occurrence of α helices and random coils were observed in both higher and lower taxonomic levels.

Cellulase of *M. albomyces*, *T. longibrachiatum* and *R. flavefaciens FD-1* was classified as unstable (II > 40) with an instability index (II) of 53.54, 55.23 and 54.34 respectively. It is notable that the *M. albomyces* and *T. longibrachiatum* are eukaryotic isolates and possess the least percentage of alpha helices in their structure. *P. haloplanktis* and *R. flavefaciens FD-1* with dominant amino acid residues Asn (10.1%) and Ser (11.6%) respectively which are hydrophilic residues, all the other sequences had ALA and GLY as dominant residues which are hydrophobic in nature. ALA was dominant in cellulases of *A. bisporus, C. thermocellum, P. carotovarum, Saccharophagus sp*. and *T. subterraneus* whereas, Gly was dominant for *C. thermocellum, M. albomyces* and *T. longibrachiatum*.

Homology model validation

SWISS MODEL was used to predict the homology model of the cellulase sequences and the protein structure quality was analyzed. RMSD values for the models were calculated and the model with least value i.e. the best predicted model is shown in (Figure 1). Ramachandran plot for the model was constructed using RAMPAGE software. Residue B 169 -LEU belonged to outlier region and the number of residues in the allowed and favoured region was very close to the expected values. It was observed that 94.8% of residues were in favored region and 5.5% in allowed region. It was found that 0.2% was found in outlier region.

FIGURE 1

Homology Model of *Pseudoalteromonas haloplanktis* cellulase based on template 1tvn predicted using SWISS MODEL. The model showed least RMSD value compared to other models

Phylogenetic analysis

Phylogenetic tree was constructed using the ten sequences based on neighbour joining method with reference sequence *C. thermocellum* [AAA23226.1] as a root (Figure 2). It was observed that the cellulase of *T. subterraneus* [ZP_07835928.1] was found to be more related to the eukaryotic cellulases. *T. longibrachiatum* [CAA43059.1], *T. subterraneus* [ZP_ 07835928.1], *M. albomyces* [CAD56665.1], *A. bisporus* [CAA83971.1], *C. thermocellum* [CAA43035.1] were belonging to same group. It can be implied that cellulase sequence of *T. subterraneus* and *C. thermocellum* were much similar to eukaryotic cellulase and it is not much evolved from the *C. thermocellum* [AAA23226.1] cellulase sequence. But the higher boot strap values for the other sequences supports its divergence from the root sequence. However, all the taxa of the group belonged to prokaryotic origin. There was no much influence for evolutionary divergence of the sequence with respect to variations in secondary structure. Neighbor joining tree showing evolutionary relationship among cellulase sequences of different origin were depicted using MEGA 4.0.Boot strap values are depicted at the nodes and branch lengths are also shown. *Clostridium thermocellum*.

Compared to bacterial cellulases, fungal cellulases are widely used. Moreover, the cellulolytic activities are high for fungal cellulases. Highest cellulase activity for *C. thermocellum* was 12.05IU/ml. *P. haloplanktis* being a psychrophilic bacterium the cellulase obtained is cold adaptable. Cellulase from the former has conserved five amino acid residues in their active sites (Garsoux *et al*., 2004). *C. thermocellum* is a thermophilic bacteria and its cellulase has a better heat stability. It is known to be ethanogenic strain and cellulase from this source has high commercial applications (Xu *et al*., 2001). Cysteine residues contribute to protein thermal stability (Xu *et al*.,

2001. Amongst fungi, species of *Trichoderma* and *Aspergillus* are well known for cellulolytic potential (Lynd *et al*., 2002). Apart from the above, other fungi used for cellulase production are *Humicola* and *Aspergillus sp*. (Ghori *et al*., 2011). Hydropathy plot for the cellulase sequence was constructed using ProtScale based on Kyte and Doolittle and the hydrophilicity and hydrophobicity nature was observed from the plot. It was observed that the majority of the residues were belonging to the hydrophilic regions confirming the interaction of the enzymes in aqueous medium. Aliphatic residues namely ALA, LEU, ILE and VAL were among the hydrophobic residues in the profile. Similarly, Phe which is an aromatic residue and sulfur containing residues MET and CYS were the other residues belonging to hydrophobic regions of ProtScale profile.

Active site prediction based on active site

Active sites for each model and template were predicted using Active Site prediction server and tabulated. It was found that *T. longibrachiatum* had most number of cavities (192). *C. thermocellum* [AAA23226.1] had 84 cavities which were very close to template with 85 cavities. Comparison of amino acid residues present in the cavities of each models were made. It was inferred that THR rich active sites may be favouring the enzyme activity in extreme environments and ASN rich cavities may be contributing towards better enzyme activity. Among the analysed models, 4 models and the template was found to possess ASN as the dominant residue in its active sites. Both *C. thermocellum* and *R. flavefaciens* FD-1 cellulases had LYS rich active sites. ARG was dominant in active sites of *M. albomyces* [CAD56665.1] and *T. subterraneus DSM* 13965[ZP_07835928.1] cellulases. However *P. haloplanktis*, an extremophile had THR dominant active sites. In *T. longibrachiatum* ASN and THR was found to be dominant in active sites with a frequency of 10.58. It is clearly notable that the hydrophilic amino acid residues are high in the active sites of these enzyme structures ensuring their interaction with substrate in aqueous phase. However the least found residue was CYS which assures stable interaction and bonding. Though the frequency of CYS was too low, it was found in both *C. thermocellum* and 3 eukaryotic cellulases. So this result validates the higher cellulolytic activity and *T. longibrachiatum* could be the source of most active cellulase from the present study.

These studies provide an insight for better prospecting of cellulolytic isolates from the environment for various industrial applications. Among the microbial cellulase used in the present work, *T. longibrachiatum* cellulase was found to be best with high number of active sites.

RESULTS & DISCUSSION

Modeling of the cellulasess

The aim of comparative homology modeling is to build 3D model for a protein. In the case of whole protein of unknown structure (target) based on one or more related protein of known structures (templates). The necessary conditions for getting a useful model are that the similarity between the target sequence and the template structure is detectable and that correct alignment between them constructed. This approach to structure prediction is possible because a small change in the protein sequence results in a small change in 3D structure.

Description of the complete procedure

1. Retrieve the template from 3D-PSSM-FOLD RECOGNITION SERVER

The 3D-PSSM server is designed to take a protein sequence of interest to you, and attempt to predict its 3-dimensional structure and its probable function. We have library of known protein structures onto each of which your sequence is “threaded” and scored for compatibility. We use a variety of scoring components;

1D-PSSM (sequence profile built from relatively close homologues), 3D-PSSMs (more general profiles containing more remote homologues), matching of secondary structure elements, and propensities of the residues in your query sequence to occupy varying levels of solvent accessibility.

Submitted your query

If you select “recognize a fold “from the home page menu you are presented with a submission form. Enter your sequence and your e-mail address.

The result of scanning your sequence against our fold library will be returned to you by e-mail, usually within 10-20 minutes.

Downloading results

The top frame is where most of the important information residues. At the top of the page is a link to allow you to download these results for viewing offline.

Please Note: Your results will only reside on the server for 5 days.

2.Aligning the target sequence with the template structure:

Once the best template has been selected the sequence of the template and the query was aligned using clustalW in order to insert gaps while modeling wherever necessary. The sequence alignment option was selected and the gaps were inserted manually with the help of clustalW output. The output for alignment of template and query is also selected from 3D-PSSM output link.

Output

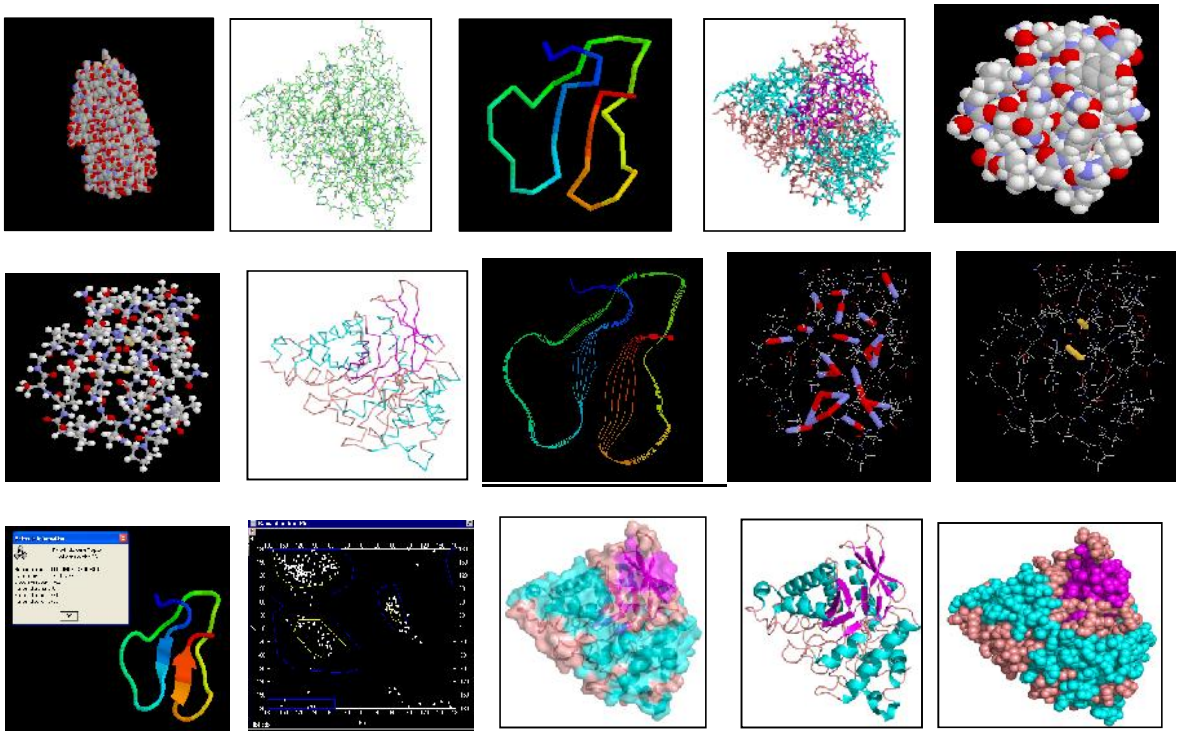
Aligning the target sequence with the template structure:

Once the best template has been selected the sequence of the template and the query was aligned using clustalW in order to insert gaps while modeling wherever necessary. The sequence alignment option was selected and the gaps were inserted manually with the help of clustalW output. The output for alignment of template and query is also selected from 3D-PSSM output link.

Loop modeling

The next steps, the amino acid, which are output of the allowed region of ramachandran plot, are checked after the model is obtained from the Swiss-model server. The ramachandran plot was constructed G.N.Ramachandran who used computer model of small polypeptides to systematically vary phi and psi angle with the objective of finding stable conformations. The following amino acids were found to be output of the region in the ramachandran plot in the model obtained ILU-297, GLU-298, Glysine and proline have not been considered in this case even if it

occurs in the disallowed region of ramachandran plot. For each conformation, the structure was examined for close contact between atoms. Atoms were treated as hard spheres with dimension corresponding. To their van der Waal's radii. Therefore, phi and psi angles, which cause spheres to collide, correspond to sterically disallowed conformation of the polypeptide backbone.



Visualization of template protein structure selected by 3dpssm server (1ib4) 64% and extracted from pdb (template)

WHAT IF Template Structure Check

Introduction

A set of WHAT IF checks will be run on the template structure.

Methods

Modelling proteins by homology is becoming a routine technique and many people rely on black box like WWW based modelling servers as their only source of structural information. The fully automatic Swiss-Model server is a good example. The server listed above is less automatic, and therefore more aimed at the experienced modeler. However the model is made, one needs to get an impression about the quality of the template and the quality of the model. This server checks the template structure for you, the next server is meant for the model. The difference is that in the model server several of the typical Xray checks (symmetry contacts; B-factors) are switched of whereas they are switched on in this server.

If your template is a standard PDB file, you can find the check report in the PDBREPORT database of structure validation reports.

What if Report

Report of protein analysis by the

What if program **********

Date : 2015-03-17

This report was created by WHAT IF version 20050215-1726

This document contains a report of findings by the WHAT IF program during the analysis of one or more proteins. It contains a separate section for each of the proteins that have been analyzed. Each reported fact has an assigned severity, one of:

* Error: severe errors encountered during the analyses. Items marked as errors are considered severe problems requiring immediate attention.
* Warning: Either less severe problems or uncommon structural features. These still need special attention.
* Note: Statistical values, plots, or other verbose results of tests and analyses that have been performed.
  If alternate conformations are present, only the first is evaluated.

Hydrogen atoms are only included if explicitly requested, and even then they are not used by all checks.

Legend

Some notations need a little explanation:

RESIDUE: Residues in tables are normally given in 3-5 parts:

- A number. This is the internal sequence number of the residue used by WHAT IF.
- The residue name. Normally this is a three letter amino acid name.
- The sequence number, between brackets. This is the residue number as it was given in the input file. It can be followed by the insertion code.

- The chain identifier. A single character, if no chain identifier was given in the input file, this will be invisible.
- A model number (only for NMR structures).

Z-VALUE: To indicate the normality of a score, the score may be expressed as a Z-value or Z-score. This is just the number of standard deviations that the score deviates from the expected value. A property of Z-values is that the root-mean-square of a group of Z-values (the RMS Z-value) is expected to be 1.0. Z-values above 4.0 and below -4.0 are very uncommon. If a Z-score is used in
WHAT IF, the accompanying text will explain how the expected value and standard deviation were obtained?

 # 1 # Note: No strange inter-chain connections detected
No covalent bonds have been detected between molecules with non-identical chain identifiers

 # 2 # Note: No duplicate atom names
All atom names seem adequately unique

 # 3 # Error: Missing unit cell information
No SCALE matrix is given in the PDB file.

 # 4 # Note: Proposal for corrected SCALE matrix
A corrected SCALE matrix has been derived.
Proposed scale matrix
0.017825 0.000000 0.005552
0.000000 0.010373 0.000000
0.000000 0.000000 0.018130

 # 38 # Note: Ramachandran plot
In this Ramachandran plot X-signs represent glycines, squares represent prolines and small plus-signs represent the other residues. If too many plus-signs fall outside the contoured areas then the molecule is poorly refined (or worse).In a color picture, the residues that are part of a helix are shown in blue, strand residues in red. "Allowed" regions for helical residues are drawn in blue, for strand residues in red, and for all other residues in green. In the TeX file, a plot has been inserted here Chain without chain identifier

 # 39 # Note: Inside/Outside residue distribution normal
The distribution of residue types over the inside and the outside of the protein is normal inside/outside RMS Z-score: 1.007

 # 61 # Note: Summary report for users of a structure. This is an overall summary of the quality of the structure as compared with current reliable structures. This summary is most useful for biologists seeking a good structure to use for modeling calculations. The second part of the table mostly gives an impression of how well the model conforms to common refinement constraint values. The first part of the table shows a number of constraint-independent quality indicators.

Structure Z-scores, positive is better than average:

1st generation packing quality	:	-0.758
2nd generation packing quality	:	-2.802
Ramachandran plot appearance	:	-1.843
chi-1/chi-2rotamer normality	:	-0.853
Backbone conformation (poor)	:	-3.939

RMS Z-scores, should be close to 1.0:

Bond lengths	:	0.717
Bond angles	:	1.001
Omega angle restraints	:	0.947
Side chain planarity	:	0.120 (tight)
Improper dihedral distribution	:	0.944
Inside/Outside distribution	:	1.007

The cellulase protein is modeled by spdb viewer with the help of template 1ib4 derived from PDB and visualized in rasmol. Finally the modeled Cellulase protein is checked by the WHATIF SERVER and result has showed.
The protein modeling was done according to the following tools, 3DPSS on line server, spdb and rasmol. It is found that the light of this investigation throught out and further analysis of gene sequencing of Cellulase.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Arnold, K., Bordoli, L., Kopp, J., Schwede, T. (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics. 22:195–201.

Bairoch, A., Apweiler, R., Wu CH, Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. (2007) The Universal Protein Resource (UniProt) Nucleic Acids Res.; 33:D154–D159.

Battey, J.N., Kopp, J., Bordoli, L., Read, R.J., Clarke, N.D., Schwede, T. (2007) Automated server predictions in CASP7. Proteins; 69 (Suppl 8):68–82.

Benkert, P., Tosatto, S.C., Schomburg, D. (2008) QMEAN: a comprehensive scoring function for model quality assessment. Proteins.71:261–277.

Berman, H.M., Westbrook, J.D., Gabanyi, M.J., Tao, Y., Shah, R., Kouranov, A., Schwede, T., Arnold, K., Kiefer, F., Bordoli, L. (2008) PSI structural genomics knowledge base. Nucleic Acids Res. in press.

Berman, H.M. (2008) Harnessing knowledge from structural genomics. Structure. 16:16–18.

Berman, H., Henrick, K., Nakamura, H., Markley, J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. 35:D301–D303.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig H. (2000) The Protein Data Bank. Nucleic Acids Res, 28: 235–242.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bairoch, A. (2007) UniProtKB/Swiss-Prot: The manually annotated section of the UniProtKnowledgeBase. Methods Mol. Biol.406:89–112.

Chothia, C. (1992) Proteins. One thousand families for the molecular biologist. Nature; 357:543–544.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R, Sonnhammer, E.L. (2008) The Pfam protein families. database. Nucleic Acids Res. 36:D281–D288.

Geneviève Garsoux, Josette Lamotte, Charles Gerday, and Georges Feller (2004) Kinetic and structural optimization to catalysis at low temperatures in a psychrophilic cellulase

from the Antarctic bacterium *Pseudoalteromonas haloplanktis* Biochem J. 2004 Dec 1; 384(2): 247–253.

Ghori Muhammad ishfaq, Sibtain Ahmed Muhammad Aslam Malana and Amer Jamil (2011) Corn stover-enhanced cellulase production by Aspergillus niger NRRL 567. 10(31), pp. 5878-5886.

Guex, N. & Peitsch, M. C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis, 18:2714–2723.

Geourjon, C., Deléage, G. (1994) SOPM: a self-optimized method for protein secondary structure prediction. Protein Eng. :7(2):157-64.

Hartshorn, M. J. Astex Viewer: 2002 A visualisation aid for structure-based drug design. J. Comput. Aided Mol. Des.16:871–881

Hillisch, A., Pineda, L.F., Hilgenfeld, R. (2004) Utility of homology models in the drug discovery process. Drug Discov. Today; 9:659–669.

Huang, H., Hu, Z.Z., Arighi, C.N., Wu, C.H. (2007) Integration of bioinformatics resources for functional analysis of gene expression and proteomic data. Front Biosci.12:5071–5088.

Humphrey, W., Dalke, A. & Schulten, K. (1996) VMD: visual molecular dynamics. *Journal of molecular graphics*, *14*(1), 33-38.

Junne, T., Schwede, T., Goder, V., Spiess, M. (2006) The plug domain of yeast Sec61p is important for efficient protein translocation, but is not essential for cell viability. Mol. Biol. Cell. 17:4063–4068.

Jenkinson, A. M., Albrecht, M., Birney, E., Blankenburg H., Down, T., Finn, R.D., Hermjakob, H., Hubbard, T.J., Jimenez, R.C., Jones, P. (2008) Integrating biological data–the distributed annotation system. BMC Bio informatics, 9(Suppl 8):S3.

Jyotsna, C., Ashish, P., Shailendra, G., Verma, M. K. (2010) Homology Modeling and Binding Site Identification of 1 deoxy d- xylulose 5 phosphate Reductoisomerase of Plasmodium Falciparum: New drugtarget for Plasmodium falciparum. *International Journal of Engineering Science and Technology*, 2 (8), 3468-3472.

Kiefer, F., Arnold, K., Künzli, M., Bordoli, L., Schwede, T. (2009) The SWISS-MODEL Repository and associated resources. Nucleic Acids Res. 2009 Jan;37(Database issue):D387-92. doi: 10.1093/nar/gkn750. Epub 2008 Oct 18.

Kopp, J., Bordoli, L., Battey, J.N., Kiefer, F., Schwede, T. (2007) Assessment of CASP7 predictions for template-based modeling targets. Proteins; 69(Suppl 8):38–56.

Kryshtafovych, A., Fidelis, K., Moult, J. (2007) Progress from CASP6 to CASP7. Proteins; 69(Suppl 8):194–207.

Koh, I.Y., Eyrich, V.A., Marti-Renom, M.A., Przybylski D., Madhusudhan, M.S., Eswar, N., Grana, O., Pazos, F., Valencia, A., Sali, A. (2003) EVA: evaluation of protein structure prediction servers. Nucleic Acids Res.; 31:3311–3315.

Kopp, J., Schwede, T. (2006) The SWISS-MODEL Repository: new features and functionalities. Nucleic Acids Res. 34:D315–D318.

Kopp, J., Schwede, T. (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. Nucleic Acids Res.32:D230–D234.

Lee, R. Lynd, Paul J. Weimer, Willem H. van Zyl, and Isak S. Pretorius (2002) Microbial Cellulose Utilization: Fundamentals and Biotechnology,66(3) 506-577.

Lovell, S.C., Davis, I.W. & Arendal, W.B. (2003) Structure validation by Ca geometry: j/y and Cb deviation. Proteins 50, 437-450.

Levitt, M. (2007) Growth of novel protein structural data. Proc. Natl Acad. Sci. USA; 104:3183–3188.

Lippow SM, Wittrup KD, Tidor B. 2007 Computational design of antibodyaffinity improvement beyond in vivo maturation. Nat. Biotechnol.;25:1171–1176.

Li Y, Drummond DA, Sawayama AM, Snow CD, Bloom JD, Arnold FH. 2007 A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. Nat. Biotechnol;25:1051–1056.

Lynd, L. R., Weimer, P. J., Van Zyl, W. H., & Pretorius, I. S. 2002. Microbial cellulose utilization: fundamentals and biotechnology. Microbiology and molecular biology reviews, 66(3), 506-577.

Manjasetty, B.A., Turnbull, A.P., Panjikar, S., Bussow, K., Chance, M.R. (2008) Automated technologies and novel techniques to accelerate protein crystallography for structural genomics. Proteomics.8:612–625

McCleverty, C.J., Columbus, L., Kreusch, A., Lesley, S.A. (2008) Structure and ligand binding of the soluble domain of a Thermotogamaritima membrane protein of unknown function TM1634. Protein Sci., 17:869–877.

Melo, F., Feytmans, E. (1998) Assessing protein structures with non-local atomic interaction energy. J. Mol. Biol.277:1141–1152.

Mulder, N.J., Apweiler, R. (2008) The InterPro database and tools for protein domain analysis. Curr. Protoc. Bioinformatics.

Murray, P.S., Li, Z., Wang, J., Tang, C.L., Honig, B., Murray, D. (2005) Retroviral matrix domains share electrostatic homology: models for membrane binding function throughout the viral life cycle. Structure, 13:1521–1531.

Ojeiru, F.E., Kazuya, T., Yuki, H., Mohammed, S.M. and Shunsuke, M. (2010) Circular Dichroism Studies on C-terminal Zinc Finger Domain of Transcription Factor GATA-2. Yonago Actamedica, 53, 25–28.

Pieper, U., Eswar, N., Davis, F.P., Braberg, H., Madhusudhan, M.S., Rossi, A., Marti-Renom, M., Karchin, R., Webb, B.M., Eramian, D. (2006) MODBASE: a database

of annotated comparative protein structure models and associated resources. Nucleic Acids Res.34:D291–D295.

Peitsch, M.C. (1995) Protein Modeling by email.*Nature Biotechnology* **13**, 658 – 660.

Peitsch, M.C. (2002) About the use of protein models. Bioinformatics.18:934–938.

Poly, M. (1997) Synthesis, structure and function of poly-alpha-amino acids--the simplest of protein models. Cell Mol Life Sci. 1997 Oct;53(10):780-9.

Pradeep, N.V., Anupama, Vidyashree, K.G., Lakshmi, P. (2012) In silico Characterization of Industrial Important Cellulases using Computational tool.*Adv Life Sci Tech.* 4: 8-15.

Soding, J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics. 21: 951– 960.

Schwede, T., Kopp, J., Guex, N., Peitsch (2003) MC. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res. 31:3381–3385.

Saitou, N. & Nei, M. (1987) Theneighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol & Evol.* 1987 4**:** 406 [PMID: 3447015]

Slabinski, L., Jaroszewski, L., Rodrigues, A.P., Rychlewski, L., Wilson, I.A., Lesley, S.A., Godzik, A. (2007) The challenge of protein structure determination—lessons from structural genomics. Protein Sci., 16:2472–2482.

Sadowski, M.I., Jones, D.T. (2007) Benchmarking template selection and model quality assessment for high-resolution comparative modeling. Proteins.69:476–485.

Tamura, K., Dudley, J., Nei, M., Kumar, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. Aug; 24 (8): 1596-9. Epub 2007 May 7

Tanya Singh, Biswas, D. and Jayaram B. (2011) AADS - An Automated Active Site Identification, Docking, and Scoring Protocol for Protein Targets Based on Physicochemical Descriptors, J. Chem. Inf. Model. 51, 2515–2527.

Tan, E.S., Groban, E .S., Jacobson, M.P., Scanlan, T.S. (2008) Toward deciphering the code to aminergic G protein-coupled receptor drug design. Chem. Biol. 15: 343–353.

Thorsteinsdottir, H.B,, Schwede, T., Zoete, V., Meuwly, M. (2006) How inaccuracies in protein structure models affect estimates of protein- ligand interactions: computational analysis of HIV-I protease inhibitor binding. Proteins; 65:407–423.

Tramontano, A. (2008) The biological applications of protein models. In: Schwede T, Peitsch MC, editors. Computational Structural Biology. World Scientific Publishing, Singapore.

Thornton, Angela M. & Ethan M. Shevach (2000) "Suppressor effector function of CD4+CD25+ immune regulatory T cells is antigen nonspecific." The Journal of Immunology 164.1: 183-190.

Van Gunsteren, W.F., Billeter, S.R., Eising, A., Hünenberger, P.H., Krüger, P., Mark, A.E., Scott, W.R.P., TironiI, G. (1996) Biomolecular Simulations: The GROMOS96 Manual and User Guide. Zürich: VdFHoch schulverlag ETHZ;

Vangrevelinghe, E., Zimmermann, K., Schoepfer, J., Portmann, R., Fabbro, D., Furet, P. (2003) Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. J. Med. Chem; 46:2656–2662.

Wallner B, and Elofsson A. 2006 Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Sci.15:900–913.

Xu Bingze, Jan-Christer Jansonand Daniel Sellos (2001) Cloning and sequencing of a molluscan endo-beta1,4glucanase gene from the blue mussel, *Mytilus edulis*. Eur. J. Biochem. 268, 3718-3727

Xu, Q., Kozbial, P., McMullan, D., Krishna, S.S., Brittain, S.M., Ficarro, S.B., Di Donato, M., Miller, M.D., Abdubek, P., Axelrod, H.L. (2008) Crystal structure of an ADP-ribosylated protein with a cytidine deaminase-like fold, but unknown function (TM1506), from Thermoto gamaritima at 2.70 A resolution. Proteins; 71:1546–1552.

Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J.A., Heidelberg, K.B., Manning, G., Li, W. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol., 5:e16.

Zhang, Y. (2007) Template-based modeling and free modeling by I-TASSER in CASP7. Proteins. ;69 (Suppl 8) :108–117.

Zhang, C. & DeLisi, C. (2001)"Protein folds: molecular systematics in three dimensions." Cellular and Molecular Life Sciences CMLS 58.1: 72-79.

Zhou, H. & Zhou, Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. 11:2714–2726.