JPP JOURNAL OF PROTEINS AND PROTEOMICS
3(1), January-June 2012, pp. 31-38

COMPARATIVE MODELLING AND LIGAND BINDING SITE PREDICTION OF A FAMILY 43 GLYCOSIDE HYDROLASE FROM *Clostridium thermocellum*

Shadab Ahmed, Rahul Charan, Arabinda Ghosh and Arun Goyal*

Department of Biotechnology, Indian Institute of Technology Guwahati, Guwahati-781039, Assam, India

Abstract: The phylogenetic analysis of *Clostridium thermocellum* family 43 glycoside hydrolase (*Ct*GH43) showed close evolutionary relation with carbohydrate binding family 6 proteins from *C. cellulolyticum, C. papyrosolvens, C. cellulyticum,* and *A. cellulyticum*. Comparative modeling of *Ct*GH43 was performed based on crystal structures with PDB IDs 3C7F, 1YIF, 1YRZ, 2EXH and 1WL7. The structure having lowest MODELLER objective function was selected. The three-dimensional structure revealed typical 5-fold β–propeller architecture. Energy minimization and validation of predicted model with VERIFY 3D indicated acceptability of the proposed atomic structure. The Ramachandran plot analysis by RAMPAGE confirmed that family 43 glycoside hydrolase (*Ct*GH43) contains little or negligible segments of helices. It also showed that out of 301 residues, 267 (89.3%) were in most favoured region, 23 (7.7%) were in allowed region and 9 (3.0%) were in outlier region. IUPred analysis of *Ct*GH43 showed no disordered region. Active site analysis showed presence of two Asp and one Glu, assumed to form a catalytic triad. This study gives us information about three-dimensional structure and reaffirms the fact that it has the similar core 5-fold β–propeller architecture and so probably has the same inverting mechanism of action with the formation of above mentioned catalytic triad for catalysis of polysaccharides.

Keywords**:** *Ct*GH43; glycoside hydrolase; *Clostridium thermocellum*; comparative modelling; 5-fold β-propeller.

Introduction

Traditionally plant cell walls are found to contain natural renewable energy sources, like cellulose, hemicellulose and lignin. Hemicelluloses are heteropolysaccharides like glucans, mannans, arabinans, and xylans. Xylan is the most common among them, constituting up to 30–35% of the total cell wall dry mass (Corrall, 2006). *Clostridium thremocellum* is an anaerobic, thermophilic and cellulolytic, gram-positive bacterium showing high level of cellulose utilization (Johnson *et al.*, 1982) and degrades the cellulosic materials by a large multi-enzymes system called the cellulosome comprising of nearly 20 different catalytic subunits ranging in size from about 40 to 180 kDa (Lamed *et al.* 1983; Bayer *et al.*, 1983; Henrissat, 1991; Henrissat and Bairoch, 1996; Lynd *et al.*, 2002; Demain, Newcomb and Wu, 2005; Fontes and Gilbert, 2010). Currently, more than 17,000 glycosidase sequences comprising of more than 113 families are known, and the sequence-based classification of their catalytic domains into glycoside hydrolase (GH) families and clans is available on the continuously updated Carbohydrate-Active enzymes (CAZy) server (*http://www.cazy.org/fam/acc_GH.html*). Glycoside hydrolases belonging to family 43 (GH43) were found to have β-xylosidase, β-1,3-xylosidase, α-L-arabinofuranosidase, arabinase, xylanase, galactan 1,3-β-galactosidase etc activities (*http://www.cazy.org/GH43.html*). Hydrolysis of the glycosidic bond can be carried out by any one of two mechanisms, leading to either retention or inversion of the anomeric configuration of the substrate (Henrissat, 1991; Henrissat and Bairoch, 1993; Davies and Henrissat, 1995; Lairson *et al.*, 2008).

Corresponding Author: **Arun Goyal**
E-mail: arungoyl@iitg.ernet.in

The α-L-arabinofuranosidases (GH43, GH51, GH54 and GH62) in combination with other lignocellulose degrading enzymes, are gaining importance in various agro-industrial processes (Numan and Bhosle, 2006). A few important applications include production of important medicinal compounds, improvement of the wine flavours bread quality, pulp treatment, juice clarification, and production of bioethanol and the synthesis of oligosaccharides (McCleary *et al.*, 1988; Gubitz, 1997; Gobbetti *et al.*, 2000; Zaldivar *et al.*, 2001; Remond *et al.*, 2004). Recently *Ct*GH43 (α-L-arabinofuranosidases) in combination with *Candida shehatae* was shown to utilize the pentose sugars for bio-ethanol production from waste material like mango and poplar leaves (Ahmed *et al.*, 2012). It is thus important to investigate these enzymes from a structural perspective for a better understanding of structure-function relationship that can help improve their myriad applications.

The secondary structures of proteins are generally stabilized by hydrogen bonds. The common examples are the alpha-helix and beta-sheet (Martin, 2004; Rost and Sanders, 1993). MODELLER is a very popular program for comparative model building (Sanchez and Sali, 2000; Marti-Renom *et al.*, 2000; Sali and Blundell, 1993). It first finds one or more template (s) and generates an alignment between template and the target sequence. Molecular dynamics (MD) force field are used integrating Newton's laws of motion but some restraints were imposed additionally named as probability density function (PDF) from Cα distances calculated from templates (Sanchez and Sali, 2000; Marti-Renom *et al.*, 2000; Sali and Blundell, 1993; Fiser *et al.*, 2000). If more than one template is used then highly conserved regions have greater restraints compared to those that vary more. The predicted model is then refined by energy minimization (Sali and Blundell, 1993; Fiser *et al.*, 2000). Recent versions of MODELLER automatically derive the restraints from the known related structures and their alignment with the target sequence. In the Ramachandran plot (RC plot), systematically the torsion angles around Cα-N phi (φ) and the Cα-C psi (ψ) of the constituent amino acid residues are varied to determine the most stable conformations. RC plot is basically a conformation chart, which tells us the values of φ and ψ that are sterically possible, and so displays permissible areas and forbidden areas. In the plot the core or allowed regions are the permissible areas for φ and ψ angle pairs for residues in a protein (Ramachandran *et al.*, 1963; Ramachandran and Sasishekaran, 1968; Lovel *et al.*, 2003). Most of the pairs must be in favoured regions of the plot and only a few shall be in "disallowed" regions (Lovel *et al.*, 2003; Ramachandran and Sasishekaran, 1968). The globular proteins are composed of amino acids which have the potential to form favorable interactions; but intrinsically unstructured proteins (IUPs) adopt no stable structure as their amino acid composition does not allow enough favorable interactions to take place (Dosztanyi *et al.*, 2005).

Q-SiteFinder (*http://www.bioinformatics.leeds.ac.uk/qsitefinder*) has been used for ligand binding site prediction by binding the hydrophobic probes ($-CH_3$) to the protein (s), and finding the clusters with the most favourable binding energy among all the residues that are then placed in rank order of the likelihood of being a binding site. THEMATICS (*http://pfweb.chem.neu.edu/thematics/submit.html*) is a method which is used for predicting the catalytic site & binding sites from 3-D structure where the charge on an ionizable residue is assumed to be on a particular atom of the ionizable group, designated as the 'charge center'; for example in the case of Asp or Glu residues, the charge is assumed to be associated to the central C atom of the carboxyl group (Ondrechen *et al.*, 2001). Using the coordinates of these charge centres, clusters of ionizable residues are defined such that the distance of the 'charge centre' of a particular residue must be within 9Å of at least one other charge centre from another residue in the cluster. These clusters are the *THEMATICS positive clusters.* Clusters with two or more members are considered predictive. The globular proteins are composed of amino acids that have the potential to form a large number of favorable interactions, whereas intrinsically unstructured proteins (IUPs) adopt no stable structure because their amino acid composition does not allow sufficient favorable interactions to form (Dosztanyi *et al.*, 2005). With this assumption IUPred was employed to check for any such regions in protein.

This work describes the 3-dimensional model building of *Ct*GH43 and identification of the active site and further active site residues that may play a key role in the catalysis or breakdown of the polysaccharides.

Materials and Methods

Phylogenetic analysis - Phylogenetic analysis for finding evolutionary relation was based on pBLAST from MPI toolkit server (*http://toolkit.tuebingen.mpg.de/*), Germany. MSA of above hits was done using ClustalW (*http://www.ch.embnet.org/ software/ClustalW.html*) and viewed with the help of Java alignment editor called Jalview (*http://www.jalview.org/*).

Secondary structure prediction - PSI-PRED VIEW was used for the secondary structure prediction of various turns, helices and coils that may be present in *Ct*GH43 (*http://bioinf2.cs.ucl.ac.uk/psiout/a43f2d904ff1ff82.psi.pdf*).

Homology modelling and validation - For model building, the *Ct*GH43 sequence was fed to HHpred (Soding, 2005) for finding the homologs or suitable templates. HHpred is the first server that is based on the pair wise comparison of hidden markov models (HMM) for homology (Krogh *et al.*, 1994). Multiple sequence alignment (MSA) revealed suitability of templates with significant E-value (≤ 0.001), used for model building. Tertiary structure is generally stabilized by non-local interactions, most commonly by formation of hydrophobic core and also through salt bridges, hydrogen bonds, disulfide bonds and even post-translational modifications. The 3-dimensional structure was modelled by using a versatile program called MODELLER (9v2) from Max-Planck Institute, Department of Developmental Biology; MPI toolkit link (*http://toolkit.tuebingen.mpg.de/*) and further, the model was validated by VERIFY3D (Luthy *et al.*, 1992). The best fitted model based on energy minimization using SWISS- MODEL Swiss-Pdb Viewer (*http://www.expasy.org/spdbv/*) having lowest value of MODELLER objective function was selected and visualized using a software called PyMOL (*http://www.pymol.org*) and RasMol (*http://www.biomed.curtin.edu.au/biochem/help/download. html*). The Ramachandran plot (RC) plot with the help of RAMPAGE server (*http://dicsoft1.physics.iisc.ernet.in/rp*) and RC plot server (*www.cryst.bioc.cam.ac.uk/rampage*) shows various residues categorized under favoured, allowed, and in disallowed regions based on residues occupying permissible and forbidden areas of plot as per their φ and ψ torsion angle values (Ramachandran *et al.*, 1963; Lovel *et al.*, 2003; Ramachandran and Sasishekaran, 1968).

Catalytic and Ligand binding site prediction - Q-SiteFinder and THMEATICS were used for catalytic site and ligand binding site prediction.

Protein disorder analysis - IUPred was employed to check for intrinsically unstructured regions in protein (Dosztanyi, Csizmok, Tompa and Simon, 2005).

Results and Discussion

A phylogenetic tree was constructed to investigate the evolutionary inter-relationships among various species or other entities that are believed to have a common ancestor and it depicts close resemblance of *Ct*GH43 with Carbohydrate binding family 6 from different *Clostridium* species *and Acetovibrio cellulolyticans* (Fig. 1). MSA based on BLASTp (Fig. 2) by ClustalW and edited using Jalview (*http://www.jalview.org/*) showed that it has high conservation of sequences with Carbohydrate binding family 6 from *Clostridium cellulolyticum, C. papyrosolvens, C. cellulyticum,* and *Acetovibrio cellulyticum.* In the figure, high consensus portions shown in dark grey are identical amino acids in all 5 sequences. In light grey are the conserved amino acids in at least 4 out of 5 sequences and the neutral portions are shown in white and gaps are included for improving the alignment (Fig. 2). PSI-PRED analysis of *Ct*GH43 sequence (Fig. 3) shows many strands and coils along with their confidence level for such occurrence but shows little or no helices as usual with Clan F (5-fold β-propeller) structures (Henrissat, 1991; Davies and Henrissat, 1995; Henrissat and Baroch, 1996;). The HHpred analysis (using HMM-HMM comparison) for homology search gave 9 hits having significant probability score and better sequence similarity as compared to others. The 3-dimensional structure prediction of *Ct*GH43 using the templates from PDB was done (Fig. 4A) by utilizing a web service provided by Max-Planck

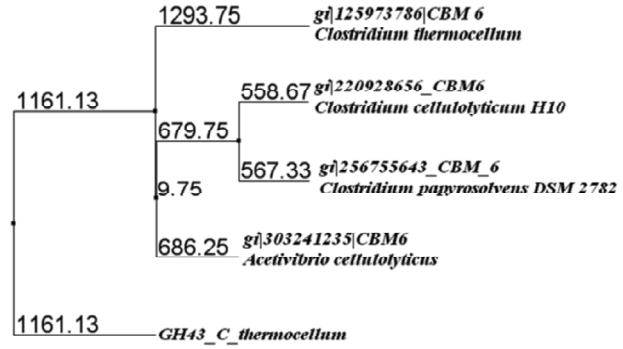

Figure 1: **Phylogenetic tree generated (showing distance) using pBLAST showing *Ct*GH43 and its closest neighbour, carbohydrate binding module family 6 (CBM6) from *Clostridium spp.* and *Acetovibrio spp.***

```
                                                10        20
GH43_C_thermo/1-301                             ----------
gi|125973786|CBM/1-679                        1 MPKKLKKIIKLCSMVFVIGILT 22
gi|220928656_CBM6/1-535                       1 ---MFKKIKKFSVFMVALLFLS 19
gi|256755643_CBM_6/1-536                      1 ---MSKRVKKLSILLVVCILVS 19
gi|303241235|CBM6/1-540                       1 ---MFKKFSRVWVIVFTISIMI 19

                                        30        40
GH43_C_thermo/1-301                     1 -------GLPIIRFSHSVLLQT 15
gi|125973786|CBM/1-679                 23 LLLPEKGAADYPIFSQRFTADP 44
gi|220928656_CBM6/1-535                20 VYSFNLVFADYPIFYQRYTADP 41
gi|256755643_CBM_6/1-536               20 ISGINTVFADYPIFYQRYTADP 41
gi|303241235|CBM6/1-540                20 FLIPERGLADYPIFSQRYTADP 41

                                  50        60
GH43_C_thermo/1-301            16 QPQSCNGRLYIYCSHDSDATPG 37
gi|125973786|CBM/1-679         45 AAVVYNGRLYIYCSHDSDATPG 66
gi|220928656_CBM6/1-535        42 SGLEANGRLYLYSSHDVYDP-N 62
gi|256755643_CBM_6/1-536       42 SGIEANGRLYLYCSHDVYDP-S 62
gi|303241235|CBM6/1-540        42 AAVVYNGRLYVYCSHDADATDG 63

                            70        80
GH43_C_thermo/1-301         38 QSTYNIPDITCISTDDLKNWTD 59
gi|125973786|CBM/1-679      67 QSTYNIPDITCISTDDLKNWTD 88
gi|220928656_CBM6/1-535     63 KPGYIMNDITCISTDDLKNWTD 84
gi|256755643_CBM_6/1-536    63 NPGYKMNDITCISTDDLKNWTD 84
gi|303241235|CBM6/1-540     64 QKTYNIPDITCISTDDLKNWTD 85

                               90        100
GH43_C_thermo/1-301            60 HGEVFNAKRDSRWASVSWAPSI 81
gi|125973786|CBM/1-679         89 HGEVFNAKRDSRWASVSWAPSI 110
gi|220928656_CBM6/1-535        85 HGEVFKASG---WASLSWAPVV 103
gi|256755643_CBM_6/1-536       85 HGEVFKASG---WASLSWAPTV 103
gi|303241235|CBM6/1-540        86 HGEVFNAKKDSSWAGVSWAPSI 107

                                     120       130
GH43_C_thermo/1-301            82 VYRNNKFYLYYGNGGNGIGVAV 103
gi|125973786|CBM/1-679        111 VYRNNKFYLYYGNGGNGIGVAV 132
gi|220928656_CBM6/1-535       104 VAKNNKYYMYFGNGAGGIGVSV 125
gi|256755643_CBM_6/1-536      104 VAKNNKYYMYFGNGGGSIGVAV 125
gi|303241235|CBM6/1-540       108 VYRNNKFYLYYGNGGNGIGVAV 129

                                140       150
GH43_C_thermo/1-301            104 SDSPTGPFKDPLPGPLVSWNTP 125
gi|125973786|CBM/1-679         133 SDSPTGPFKDPLPGPLVSWNTP 154
gi|220928656_CBM6/1-535        126 SDSPTGPFKDALGKALINGSTP 147
gi|256755643_CBM_6/1-536       126 SDSPTGPFKDALGKALITGSTP 147
gi|303241235|CBM6/1-540        130 SDSPIGPFKDPLPSALIIGNIP 151
```

Figure 2: **Multiple sequence alignment analysis of *Ct*GH43 using ClustalW. The analysis showed similarities with matching sequences *viz.* similar proteins from CBM6 family from *Clostridium cellulolyticum, Clostridium papyrosolvens, Clostridium cellulyticum, Acetovibrio cellulyticum.* Identical amino acids in all 5 sequences are shown in dark grey. In light grey are the conserved amino acids in at least 4 out of 5 sequences and the neutral portions shown are in white and gaps are given for alignment improvement.**

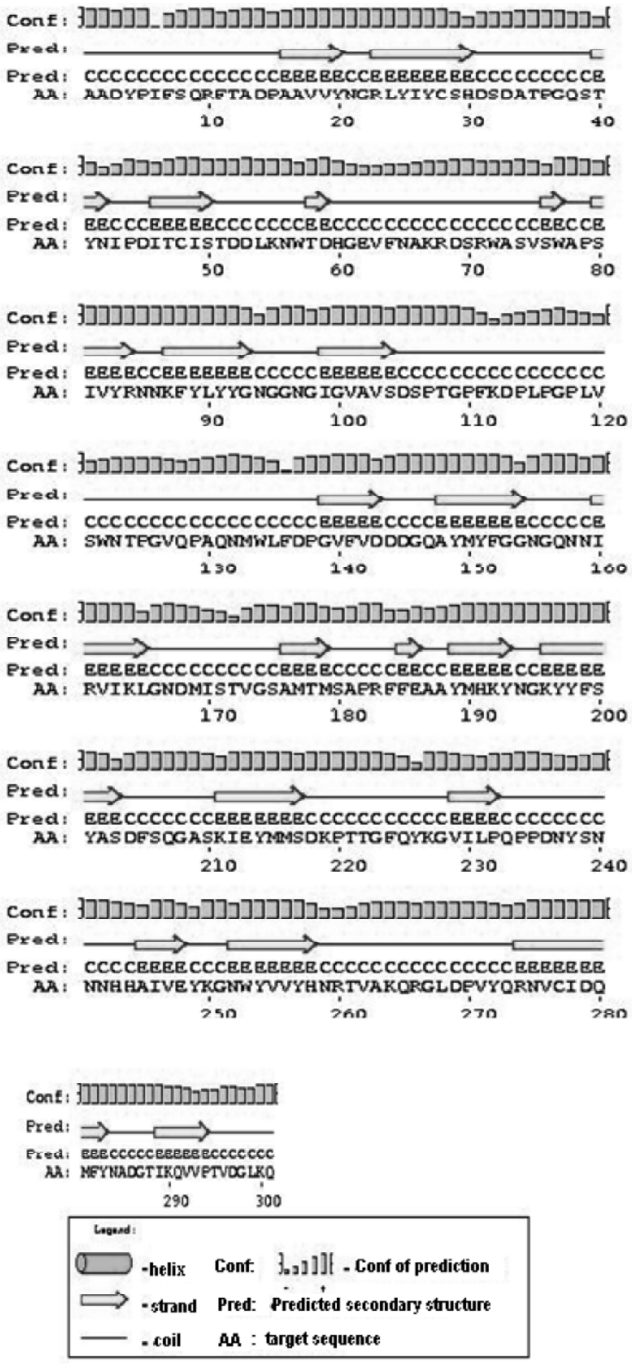

Figure **3: Secondary structure of *Ct*GH43 using PSI-PRED VIEW software showing beta strands (shown by arrow) and coils (as continuous line) with confidence level of prediction. Cylindrical signs denotes helices, arrow strands and black line denote coils.**

Institute, Developmental Biology, Germany (*http://toolkit.tuebingen.mpg.de/*). MODELLER 9v2 was used for homology modeling. Comparative modeling was done utilizing the following templates: arabinoxylan arabinofuranohydrolases (3C7F), β-1,4-xylosidase (1YIF), xylan β-1,4-

xylosidase (1YRZ), β-D-xylosidase (2EXH), endo-1,5-α-L-arabinanase (1WL7) from the PDB (Bernstein, Koetzle and Williams, 1977; Berman, Westbrook and Feng, 2000; Berman *et al.*, 2007; Marchler-Bauer, Lu and Anderson, 2011). These templates or 3-D structures helped in three dimensional model building. The characteristic 5-fold β–propeller core can be easily identified from 3-D model of *Ct*GH43. Fig. 4B shows the residues involved at active site (Asp, Glu, Asp) of the enzyme, which may play important role in the activity. VERIFY3D validation of predicted structure of *Ct*GH43 confirmed that the model is acceptable (Fig. 5). The energy minimization of pdb file of *Ct*GH43 was done using Swiss-Pdb Viewer version 4.0 (Guex and Peitsch, 1997).

The 3-D structure of *Ct*GH43 showed little or no segments of helices (Fig. 6). Similarly,

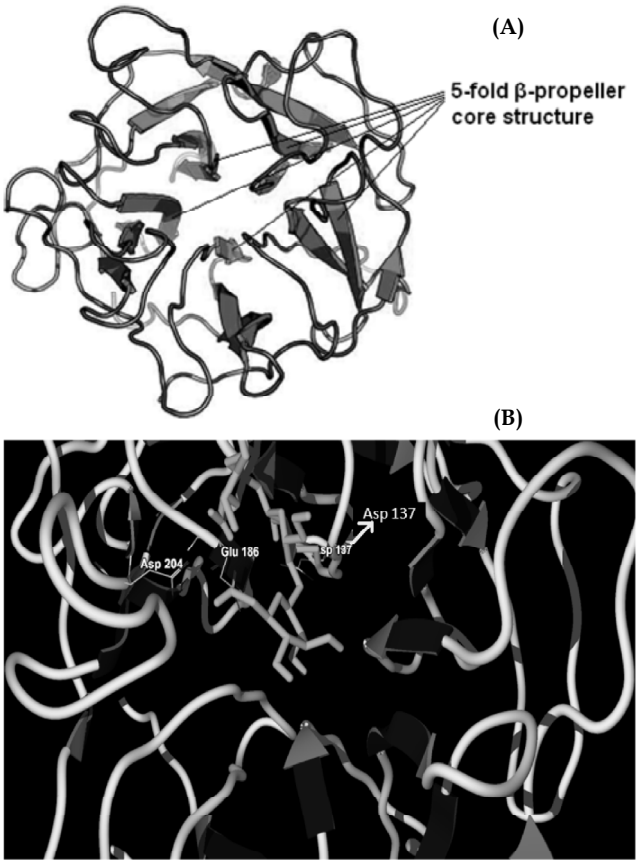

Figure 4: **(A) Three dimensional structure of *Ct*GH43 developed utilizing MODELLER 9v2 and visualized using PYMOL software showing the characteristic fold (5 fold β-sheets in grey, coils in greyish black).**

(B) Three dimensional structure of *Ct*GH43 revealing all the residues that are present at active site. It shows typical architecture where an Asp is the catalytic general base, a Glu as the catalytic general acid an Asp as *p*Ka modulator.

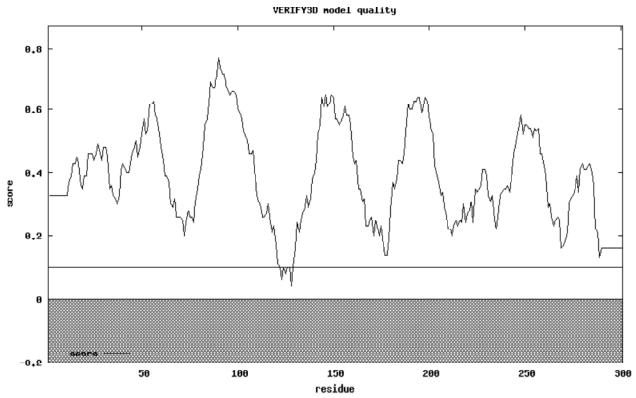

Figure 5: **Quality assessment of predicted 3-D model by VERIFY3D software (very few residues cross the light grayish line having negative value implying predicted model is acceptable).**

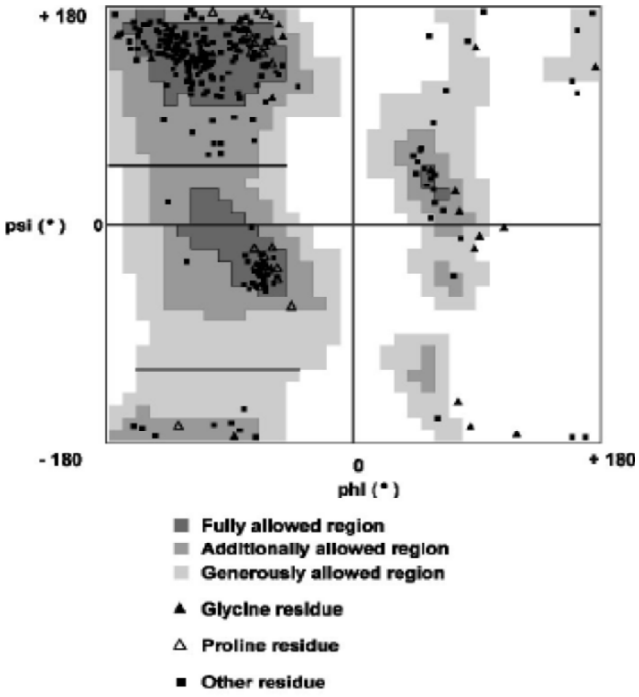

Figure 6: **Ramachandran plot for *Ct*GH43 from the online server available from Indian Institute of Science (IISc), Bangalore. It shows the left-handed as well as right-handed helices, beta sheets, disallowed and allowed areas. Black areas are allowed only for Gly residues.**

repetitive values in the region (Fig. 6) of phi = -110 to -140 and psi = +110 to +130 gave extended chains with conformations that allow interactions between closely folded parallel segments (β-sheet structures). β-sheets in folded proteins are twisted rather than planar, with a right-handed twist of 0°-30° between strands. β-sheets can consists entirely of parallel or mixture of the two. The structure of *Ct*GH43 is composed mostly of β-sheets with 178 residues and the RC plot showed

a broad range of values in the (-110°, +130°) regions (Fig. 6). The same figure shows only 4 residues in 3_{10} helical regions, which are present in first quadrant of conformational map (Fig 6). The RAMPAGE plot indicates that out of the 301 residues, 267 (89%) were in favoured region, 23 (8%) were in allowed region and 9 (3%) were in disallowed region implying that the predicted model is quite acceptable (Fig. 7). Ramachandran plot for general, glycine, pre-proline and proline amino acid residues was also done and it showed the glycine, pre-proline and proline residues of *Ct*GH43 occupying allowed regions and also those glycine residues in disallowed region (Fig. 8). Ramachandran Z-score was found to be -1.706; indicating how well the backbone conformations of all residues corresponded to the known allowed areas in the Ramachandran plot as per expected ranges for well-refined structures (Ramachandran, Sasisekharan and Ramakrishnan, 1966). The IUPred result indicates that there are only few residues that show tendency for disorderedness with index more than 0.5, concluding that protein does not have any significant disordered region (Fig. 9).

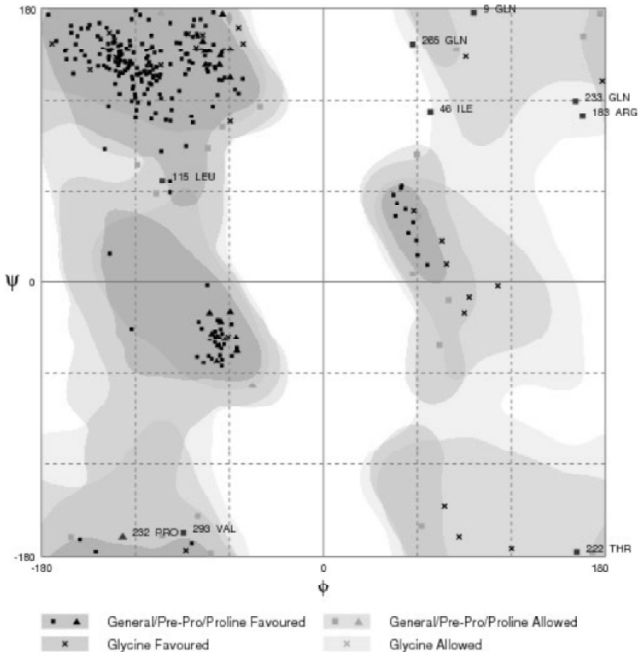

Figure 7: **Ramachandran plot analysis of *Ct*GH43 using RAMPAGE software. It shows the various residues lying in favoured, allowed and disallowed region and the glycine residues (267 residues are in favoured region, 23 in allowed region and 9 (≤ 2.9%) in disallowed region so that ≥ 97% residues have allowed conformations).**

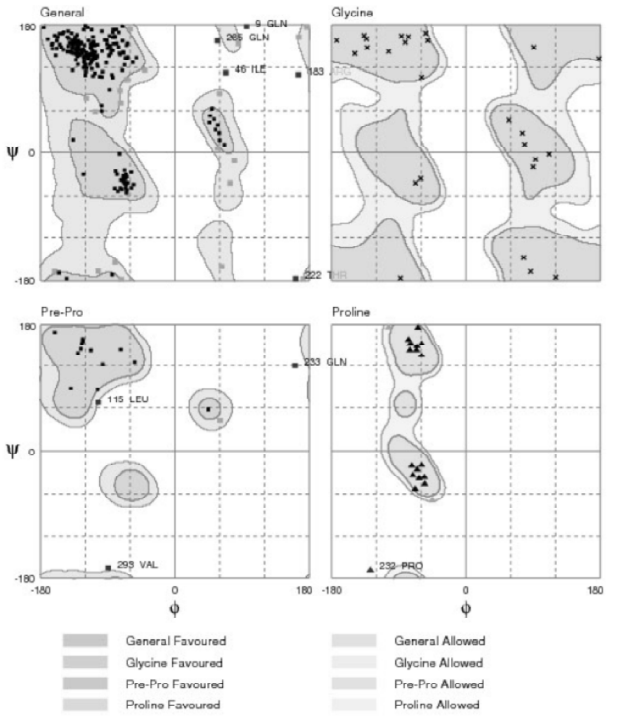

Figure 8: **Ramachandran plot analysis of *Ct*GH43 for general, Gly, pre-Pro, and Pro residues using RAMPAGE. The conformations and location of each of the above is shown in individual plots having heading as general, Glycine, pre-Pro and Pro.**

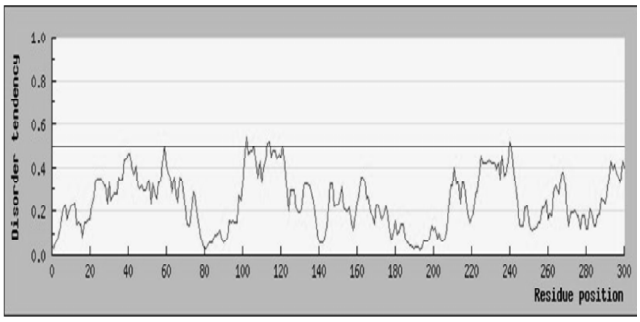

Figure 9: **Protein Disorder Analysis plot using IUPred.**

Conclusion

From the above results we can conclude that the *Ct*GH43 secondary structure has predominantly β-sheets with little or no helices. The 3-dimensional structure predicted using MODELLER confirms the above and shows the expected 5 fold β-propeller in the model. The RC plot analysis showed most residues in the permissible areas and indicated that the model is of good quality. Further validation by VERIFY3D reinstated the fact that predicted model has a stable conformation and is acceptable. Little or no disordered regions were found when analyzed by

IUPred. These facts can be further used for functional characterization by performing active site prediction, ligand binding or docking and other such studies on this protein. The active site among other residues show that aspartate is the catalytic general base, a glutamate as the catalytic general acid and another aspartate (Fig. 4B) that may be responsible for *p*Ka modulation and orienting the catalytic general acid. Close analysis of the PDB structures 2EXH, 1WL7, 3C7E, 1YIF, 1UV4 and 3KIU revealed that all these proteins have 5 fold β-propeller core believed to be the catalytic domain. All the above structures have two Asp and a single Glu at the active site forming a catalytic triad of corboxylates which binds to the substrate or ligand. *Ct*GH43 also shows similar architecture with the presence of Asp 137, Asp 204, Glu 186 at the active site where Glu acts as the catalytic acid; one Asp as base and another Asp is responsible for pKa modulation and orienting the catalytic acid.

Acknowledgments

The research work has been supported by a grant from Department of Biotechnology, Ministry of Science and Technology, New Delhi to AG.

Abbreviations

*Ct*GH43, family 43 glycoside hydrolase from *C. thermocellum*; GH51, GH54 and GH62, family 51, 54 and 62 glycoside hydrolases from *C. thermocellum*; MSA, multiple sequence alignment; 3-D, three dimensional; RC plot, Ramachandran plot; Z-score, standard deviations away from the mean; HMM, Hidden Markov model; THEMATICS, Theoretical microscopic titration curves.

References

Bayer, E.A. Kenig, R. and Lamed, R. (1983). Adherence of *Clostridium thermocellum* to cellulose, *J. Bacteriol. 156*, 828-36.

Berman, H. M. Henrick, K. Nakamura, H. Markley, J. Bourne, P.E. Westbrook, J. (2007). Realism about PDB. *Nat. Biotechnol. 25*, 845-846.

Berman, H.M. Westbrook, J. and Feng, Z. (2000). The Protein Data Bank. *Nucleic Acids Res. 28*, 235-242.

Bernstein, F. C. Koetzle, T. F. and Williams G. J. B. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol. 112*, 535-542.

Corrall, O.L. and Ortega, F.V. (2006). Xylanases in Advances in Agricultural and Food Biotechnology. (eds. Guevara-González, R.G. and Torres-Pacheco, I.), Research Signpost, Trivandrum, Kerala, 305-322.

Das, S. P., Ravindran, R., Ahmed, S., Das, D., Goyal, G., Fontes, C.M.G.A., and Goyal, A. (2012) Bioethanol Production Involving Recombinant *C. thermocellum* Hydrolytic Hemicellulase and Fermentative Microbes. *Appl. Biochem. Biotechnol.* DOI 10.1007/s12010-012-9618-7.

Davies, G. & Henrissat, B. (1995). Structures and mechanisms of glycosyl hydrolases. *Structure 3*, 853-857.

Demain, A.L. Newcomb, M. and Wu, J.H. (2005). Cellulase, clostridia, and ethanol. *Microbiol. Mol. Biol. R. 69*, 124-154.

Dosztányi, Z. Csizmók, V. Tompa, P. and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics 21*, 3433-3434.

Fiser, A., Do, R. K. and Sali, A. (2000) Modeling of loops in protein structures. *Protein Sci. 9*, 1753-1773.

Fontes, C.M.G.A. & Gilbert, H.J. (2010). Cellulosome: Highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates, *Annu. Rev. Biochem. 79*, 655-681.

Gobbetti, M., Lavermicocca, P., Minervini, F., De Angelis, M. and Corsetti, A. (2000) Arabinose fermentation by *Lactobacillus plantarum* in sourdough with added pentosans and a-L-arabinofuranosidase: a tool to increase the production of acetic acid. *J. Appl. Microbiol. 88*, 317-324.

Gu¨bitz, G.M., Haltrich, D., Latal, B., Steiner, W. (1997) Mode of depolymerisation of hemicellulose by various mannanases and xylanases in relation to their ability to bleach softwood pulp. *Appl. Microbiol. Biotechnol. 47*, 658–621.

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

Henrissat, B. (1991). A classification of glycosyl hydrolases based on amino-acid sequence similarities. *Biochem. J. 280*, 309-316.

Henrissat, B. and Bairoch, A. (1993). New families in the classification of glycosyl hydrolases based on amino-acid sequence similarities. *Biochem. J. 293*, 781-788.

Henrissat, B. and Bairoch, A. (1996). Updating the sequence-based classification of glycosyl hydrolases. *Biochem. J. 316*, 695-696.

Johnson, E. A. Sakajoh, M. Halliwell, G. Madia, A. & Demain, A. L. (1982). Saccharification of complex cellulosic substrates by cellulase system from Clostridium thermocellum, *Appl. Environ. Microb. 43*, 1125-1132.

Krogh, A. Brown, M. Mian, I.S. Sjolander, K. and Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol. 235*, 1501-1531.

Lairson, L.L. Henrissat, B. Davies, G.J. & Withers, S.G. (2008). Glycosyltransferases: Structures, Functions and Mechanisms. *Annu. Rev. Biochem. 77*, 521-555.

Lamed, R., Setter, E., and Bayer, E. A. (1983). Characterization of a cellulose-binding, cellulase-containing complex in *Clostridium thermocellum*. *J. Bacteriol. 156*, 828-836.

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

Lüthy, R., Bowie, J. U. and Eisenberg, D. (1992). Assessment of protein models with three dimensional profiles. *Nature 356*, 83-85.

Lynd, L.R., Weimer, P.J., van Zyl, W.H. and Pretorious, I.S. (2002). Microbial cellulose utilization: fundamentals and biotechnology, *Microbiol. Mol. Biol. R. 66*, 506-577.

Marchler-Bauer, A., Lu, S. and Anderson, J.B. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res. 39*, 225-229.

Martin, A.C.R. (2004). Comparative modelling. Bioinformatics: Genes, proteins and computers (eds. Orengo, C.A., Jones, D.T. and Thornton, J.M.), New York, BIOS Scientific Publishers, 135-150.

Marti-Renom, M.A., Stuart, A., Fiser, A., Sánchez, R., Melo, F., and Sali, A. (2000), Comparative protein structure modeling of genes and genomes. *Ann. Rev. Bioph. Biom. 29*, 291-325.

McCleary, B.V., Harrington, J. and Allen, H. (1988) Enzymic solutions to polysaccharide related industrial problems. (eds. Phillips, G.O., Williams, P.A. and Wedlock, D.J.) Gums and stabilizers for the industry. IRL Press, Oxford, 51–62.

Ondrechen, M.J., Clifton, J.G. and Ringe, D. (2001). THEMATICS: A Simple Computational Predictor of Enzyme Function from Structure. *P. Natl. Acad. Sci.* USA. *98*, 12473-12478.

Ramachandran, G. N., Ramakrishnan, C. and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol. 7*, 95-99.

Ramachandran, G.N. and Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Adv. Prot. Chem. 23*, 283-437.

Ramachandran, G.N. Sasisekharan, V. and Ramakrishnan, C. (1966). Molecular structure of polyglycine II. Biochemi. *Biophysic. Act. 112*, 168-170.

Rémond, C., Plantier-Royon, R., Aubry, N., Maes, E., Bliardc, C. and O'Donohue, M. J. (2004) Synthesis of pentose-containing disaccharides using a thermostable a-L-arabinofuranosidase. *Carbohydr. Res. 339*, 2019–2025.

Rost, B. and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol. 202*, 865-884.

Sali, A. and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol. 234*, 779-815.

Sánchez, R. and Sali, A. (2000). Comparative protein structure modeling: Introduction and practical examples with MODELLER in Protein Structure Prediction: Methods and Protocols. (ed. D. M. Webster), New Jersey, Totowa, Humana Press, 97-129.

Söding, J. (2005). HHpred: Protein homology detection by HMM-HMM comparison. *Bioinformatics 21*, 951-960.

Zaldivar, J., Nielsen, J. and Olsson, L. (2001) Fuel ethanol production from lignocellulose: a challenge for metabolic engineering and process integration. *Appl. Microbiol. Biotechnol. 56*, 17-34.