American Journal of

# Biochemistry and Molecular Biology

ISSN 2150-4210

Academic
Journals Inc.

www.academicjournals.com

American Journal of Biochemistry and Molecular Biology 3 (1): 1-19, 2013
ISSN 2150-4210 / DOI: 10.3923/ajbmb.2013.1.19


# Insights into Steviol Glycoside Biosynthesis Pathway Enzymes Through Structural Homology Modeling

Praveen Guleria and Sudesh Kumar Yadav
Plant Metabolic Engineering and Nanobiology Laboratory, Biotechnology Division, CSIR-Institute of Himalayan Bioresource Technology, Palampur-176061 (HP), India

*Corresponding Author: Sudesh Kumar Yadav, Plant Metabolic Engineering and Nanobiology Laboratory, Biotechnology Division, CSIR-Institute of Himalayan Bioresource Technology, Palampur-176061 (HP), India*

## ABSTRACT

Steviol glycosides are the major secondary metabolites synthesized through steviol glycoside biosynthesis pathway operating in the leaves of *Stevia rebaudiana*. Present article documents the structural analysis of enzymes specific to steviol glycoside biosynthesis pathway, kaurenoic acid-13 hydoxylase (KAH) and three UDP-glycosyltransferases (UGT85C2, UGT74G1 and UGT76G1). The *in silico* protein structure prediction server SWISS MODEL was used to predict and evaluate the models. The secondary structure data of predicted model for KAH was in accordance to that of cytochrome P450s suggesting its nativeness to the respective superfamily. Similarly, the secondary structure data of target UGTs also showed conservation with the structural information of glycosyltransferases superfamily. PROCHECK and QMEAN Z-score evaluations suggested that the models predicted for the 4 query enzymatic proteins were of good quality. In addition, Ligand binding site analysis and molecular docking analysis was carried out for the predicted models. The following data suggested a possibility of the presence of an alternate pathway for the synthesis of steviol glycosides.

**Key words:** Binding site, comparative protein modeling, interaction, ligand, *Stevia rebaudiana*

## INTRODUCTION

Steviol glycosides are the diterpene secondary metabolites from *S. rebaudiana*. These are the glycosylated products of the precursor steviol (Richman *et al*., 2005). Steviol glycosides are used as dietary supplements as natural sweetner in various nations. It has been known that these glycosides are anti-diabetic, non-cariogenic and non-mutagenic (Yadav and Guleria, 2011). Steviol glycosides are synthesized in the leaves of *Stevia* via steviol glycoside biosynthesis pathway (Fig. 1). Steviol glycoside biosynthesis pathway comprises of 16 steps catalyzed by several enzymes. Among these, the last 5 steps are specific to steviol glycoside biosynthesis pathway (Yadav and Guleria, 2011; Guleria *et al*., 2011; Brandle and Telmer, 2007). These steps are catalyzed by enzymes Kaurenoic Acid-13 Hydroxylase (KAH) and four UDP-glycosyltransferases (UGTs) identified as UGT85C2, UGT74G1 and UGT76G1. One UGT is still to identify (Yadav and Guleria, 2011; Brandle and Telmer, 2007). Despite of the huge and wide prevailing importance, various aspects of this pathway are still hidden.

Formulating three dimensional structure of a protein is of great help in understanding its biochemical functions and molecular interaction properties (Bordoli *et al*., 2009). Protein structures are more conserved than protein or DNA sequences. So, *in silico* approaches are being used to

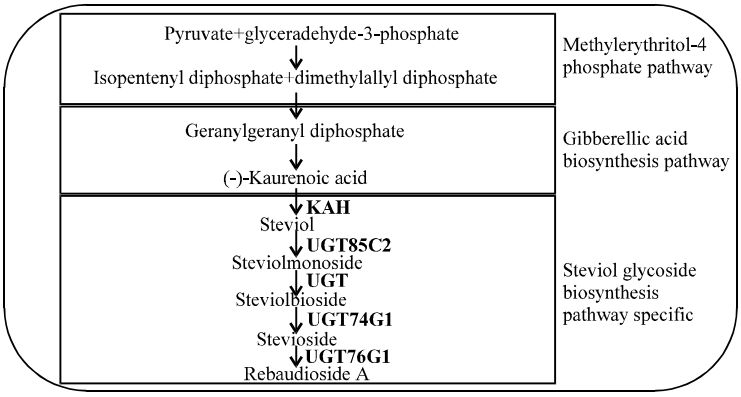



Fig. 1: Schematic representation of steviol glycoside biosynthesis pathway, Enzymes (bolded) are specific to steviol glycoside biosynthesis pathway

predict the structure of an unknown protein taking known three dimensional structures of related protein family as a template reference (Chithia and Lesk, 1986; Mugilan *et al*., 2010; Smith and Plazas, 2011; Joseph and Nair, 2012). Comparative homology modeling has been utilized to predict three dimensional structures for various plant proteins such as cytosolic glutamine synthetase from *Camellia sinensis* (Yadav, 2009), late embryogenesis abundant protein from *Arabidopsis thaliana* (Boobalan and Bharathi, 2010) lectins from mushroom (Khan and Khan, 2011) and thioredoxin from *Triticum aestivum* (Prabhavathi *et al*., 2011). Higher the sequence similarity, more significant is the structural identity and consequently, an increased reliability of the predicted model (Marti-Renom *et al*., 2000; Huq, 2008a, b).

In this study, protein structures of steviol glycoside biosynthesis pathway enzymes; Kaurenoic Acid-13 Hydoxylase (KAH) and three UDP-glycosyltransferases (UGT85C2, UGT74G1 and UGT76G1) were analyzed using *in silico* approaches. This is the first report documenting the computational elucidation of three dimensional models of these enzymatic proteins. These structures were further analyzed for the prediction of ligand binding sites and ligand interaction affinities.

## MATERIALS AND METHODS

**Retrieval of target sequences:** The amino acid sequences of the enzymatic proteins KAH, UGT85C2, UGT74G1 and UGT76G1 were obtained from NCBI (http://www.ncbi.nlm.nih.gov) Protein Database in FASTA format. It was made ascertain that three dimensional structures of these proteins were not available in Protein Data Bank (PDB). Accession number, protein lengths have been tabulated (supplementary Table 1 in appendix).

**Structure prediction and evaluation:** The three dimensional structures were predicted using SWISS-MODEL, an automated protein homology-modeling server (Schwede *et al*., 2003). The amino acid sequences of respective proteins were submitted to SWISS-MODEL server one-by-one and following steps were performed: domain annotation, template identification, automated modeling and structure assessment. Domain annotation determined the superfamily to which respective protein belonged as well as secondary structure elements of the target protein. Template identification predicted the possible templates for target sequence on the basis of target-template sequence similarity. Three dimensional structures were determined using automated modeling mode and the predicted models were evaluated using Structure Assessment tool of SWISS-MODEL server (Bordoli *et al*., 2009).

**Ligand binding site prediction:** Ligand binding sites of the evaluated models were predicted by submitting models to Q-SiteFinder server. This server depicts the energetically favorable ligand binding sites by using methyl probes at a grid resolution of 0.9 Å on a three dimensional grid encompassing the whole protein molecule (Laurie and Jackson, 2005).

**Molecular docking analysis:** Molecular docking between the predicted models corresponding to KAH, UGT85C2, UGT74G1, UGT76G1 and the pathway substrate molecules was performed using Molecular Docking server (http://www.dockingserver.com). It involves three steps: uploading of desired ligand molecule from NCBI PubChem, uploading of the query protein molecules through PDB and finally followed by docking. Each docking run was repeated twice to get best results (Bikadi and Hazai, 2009). Selected ligands were ent-kaurenoic acid, steviol, steviolmonoside, steviolbioside, stevioside and rebaudioside A. The uploaded proteins were three dimensional models predicted for enzymatic proteins KAH, UGT85C2, UGT74G1 and UGT76G1. Each protein was docked with every uploaded ligands.

## RESULTS AND DISCUSSION

**Homologue identification and secondary structure analysis:** Kaurenoic Acid-13 Hydroxylase (KAH) and three UDP glycosyltransferases (UGT85C2, UGT74G1 and UGT76G1) are the key enzymes of steviol glycoside biosynthesis pathway (Fig. 1). In this paper, SWISS-MODEL was used to predict structures of these enzymatic proteins. Domain annotation data indicated that KAH belongs to Cytochrome P450 superfamily. No plant enzymatic proteins belonging to P450 superfamily have been analyzed for their structures (Rupasinghe and Schuler, 2006). While domain annotation of UGT85C2, UGT74G1 and UGT76G1 documented that these enzymatic proteins belong to glycosyltransferases (GTs) superfamily and/or UDP-glucuronosyl/UDP-glucosyltransferase family. The structural analysis of GT proteins have been carried out for *Medicago truncatula* (Shao *et al*., 2005) and *Vitis vinifera* (Offen *et al*., 2006). Secondary structure predicts the α-helices and β-sheets present in the query protein. The analysis data of predicted secondary structures for all the four proteins are shown in Fig. 2 and their evaluations are presented in Supplementary Table 1 (Appendix). KAH possessed a higher number of α-helices (264) than β-sheets (37) or random coils (174) (Fig. 2a). Presence of increased number of helices than β-sheets is the conserved structural feature of Cytochrome P450s (Graham and Petersen, 1999; Stout, 2004). This suggested the nativeness of KAH to Cytochrome P450s and its helical nature. UGT85C2 possessed 220 random coils, 196 α-helices and 64 β-sheets (Fig. 2b). An increased number of random coils (213) than α-helices (189) or β-sheets (57) were also observed for UGT74G1 (Fig. 2c). Similarly, UGT76G1 showed 207 random coils, 189 α-helices and 61 β-sheets (Fig. 2d). The data documented the dominating character of random coils in predicted secondary structures of UGTs. This suggests that these three UGTs belong to the same family and possess coiled geometry.

**Model prediction and evaluation:** Template identification searched templates for the query sequences on the basis of significant sequence similarity. Retinoic acid bound cyanobacterial CYP120A1 protein (PDB ID 2ve3 chain A), a Cytochrome P450 was identified as template for KAH showing 33.9% of sequence similarity. 2vg8 chain A, N and O glucosyltransferase involved in xenobiotic metabolism of plants was identified as template for UGT74G1 with a highest sequence similarity of 26.6%. While 2pq6 chain A, crystal structure of *Medicago truncatula* UGT85H2, was

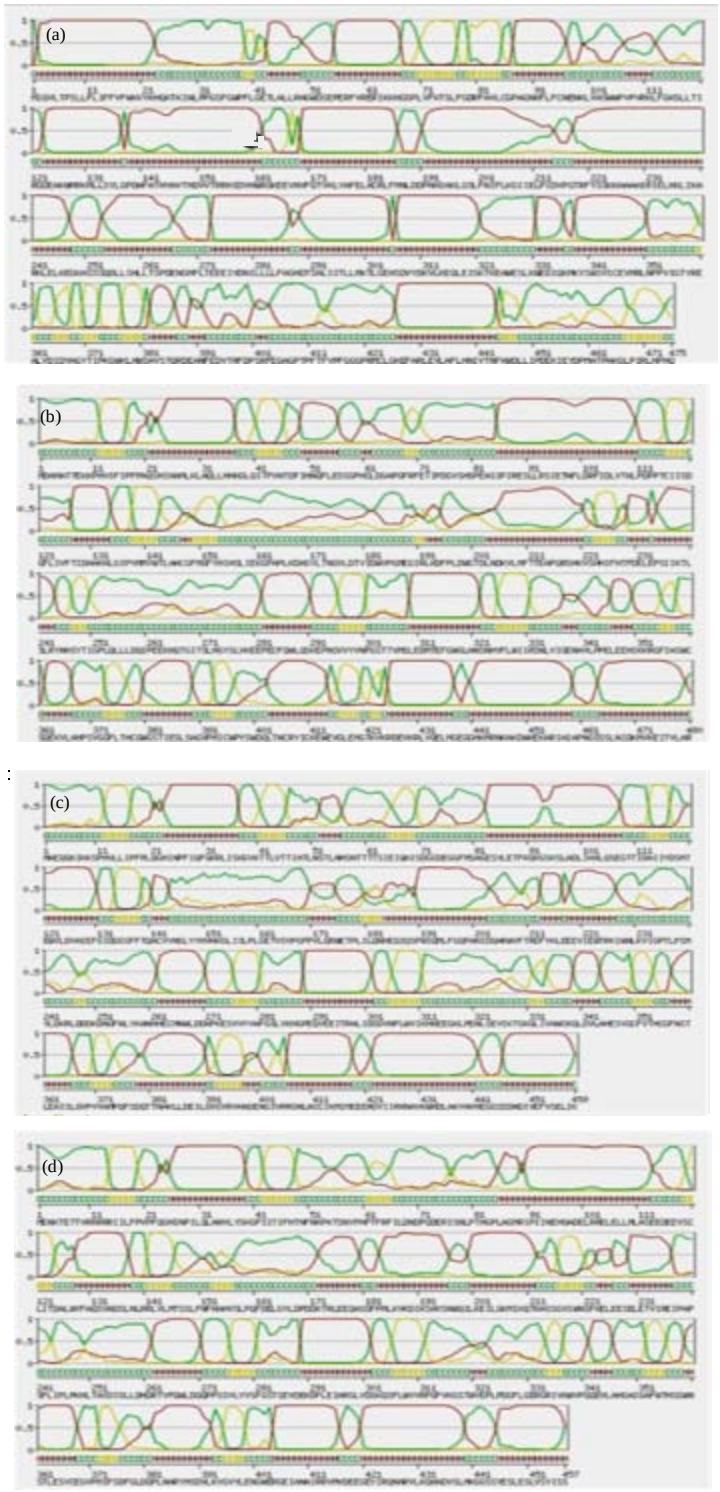



Fig. 2(a-d): Secondary structure for enzymes (a) KAH, (b) UGT85C2, (c) UGT74G1 and (d) UGT76G1. KAH possesses helical structure and UGTs possess coiled structure

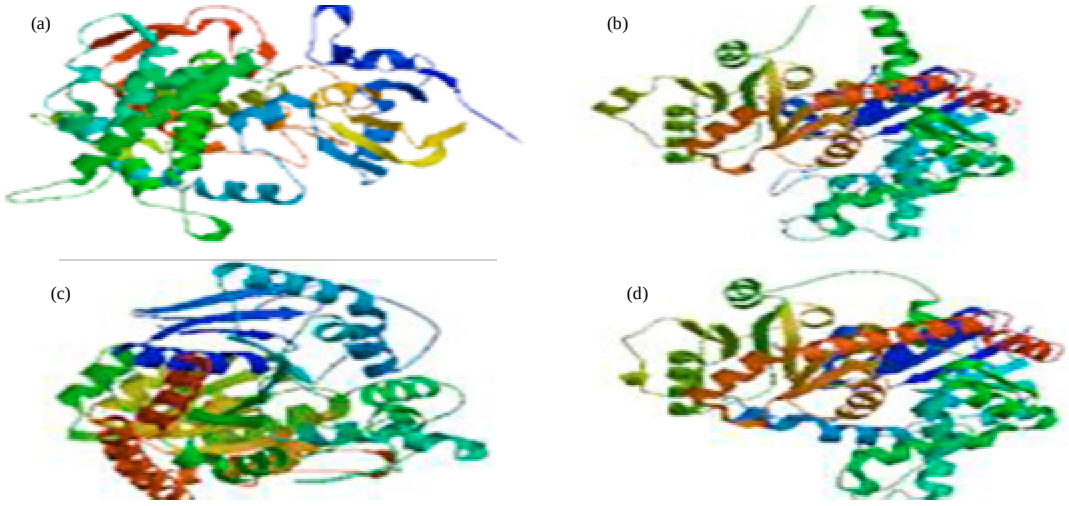



Fig. 3(a-d): Three dimensional model of (a) KAH, (b) UGT85C2, (c) UGT74G1 and (d) UGT76G1

identified as template for both UGT85C2 and UGT76G1 with a sequence similarity of 43.3 and 28%, respectively. For all the query proteins template-target sequence identity was more than 25%, hence suitable to conduct automated modeling (Schwede *et al*., 2003).

The three dimensional structure of enzymatic protein KAH was determined at 2.1 Å resolution by Automated modeling method of SWISS-MODEL server. The employed template was 2ve3 chain A, a cyanobacterial cytochrome P450. Out of total 476 residues, 447 residues were included by the software to constitute the modeled structure. The structure consisted of a single chain comprising of 17 α-helices and 12 β-strands arranged in four β-pleated sheets (Fig. 3a). This kind of arrangement is also a characteristic feature of Cytochrome P450 folds (Rupasinghe and Schuler, 2006; Kuhnel *et al*., 2008). Conserved secondary structure suggests that KAH belongs to cytochrome P450 superfamily. The enzymatic protein UGT85C2 was modeled at a resolution of 2.1Å by using template 2pq6 chain A, UGT85H2 from *M. truncatula*. A total of 468 residues out of 483 residues were included to constitute the three dimensional model. The predicted model possessed single chain consisting of 20 α-helices and 13 β-strands. The β-strands were arranged in two β-pleated sheets, one possessing seven stranded parallel β-strands and the other with six stranded parallel β-strands (Fig. 3b). This feature is common for the UGTs modeled till date using experimental approaches (Shao *et al*., 2005; Li *et al*., 2007). Results suggested the nativeness of computationally modeled UGT85C2 with the experimentally solved UGTs.

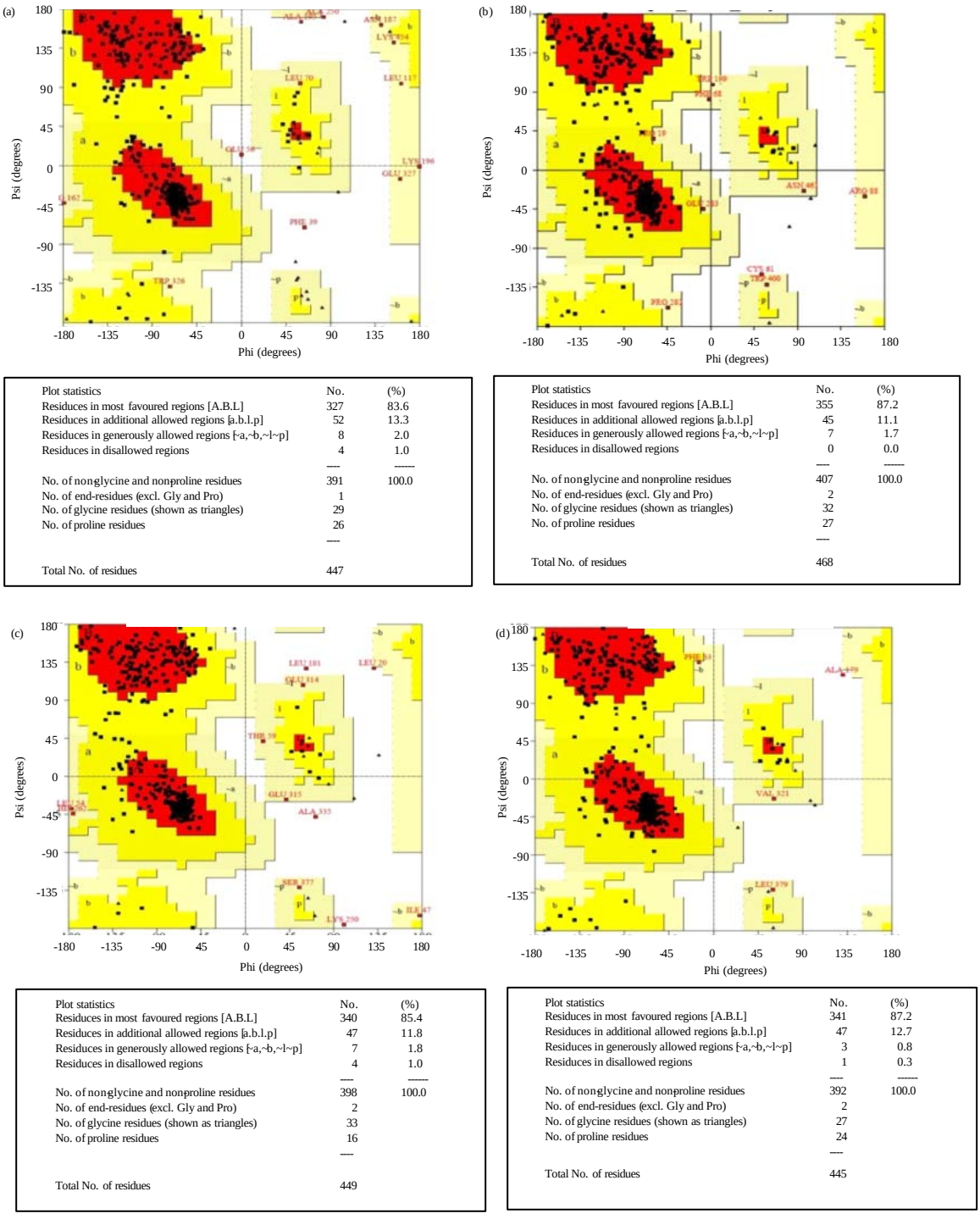
Similarly, automated modeling was carried out for the protein UGT74G1. This enzymatic protein was modeled on the basis of template 2vg8 chain A at 1.75 Å resolution. The predicted model consisted of 449 residues out of the total 460 residues. The geometry of model comprised of one chain, 16 α-helices, 13 β-strands arranged in 7 and 6 stranded parallel two β-pleated sheets (Fig. 3c). The geometry and topology of modeled UGT74G1 was similar to UGT85C2. Hence, similar to UGT85C2, the UGT74G1 was conserved for structural features with the experimentally solved UGTs (Shao *et al*., 2005; Brazier-Hicks *et al*., 2007). The other UDP glycosyltransferase protein, UGT76G1 was modeled by automated modeling on the basis of template utilized for UGT85C2, 2pq6 chain A. The predicted model was constituted of 445 residues out of 458 residues. The model

possessed single chain with 19 α-helices and 13 β-strands (Fig. 3d). Like UGT85C2 and UGT74G1, the β-strands of UGT76G1 were also arranged in the form of seven stranded and six stranded parallel two β-pleated sheets (Shao *et al*., 2005; Li *et al*., 2007). It was observed that the entire query UGTs possessed similar geometry of β-pleated sheets. Thus, it suggests that the studied UGTs; UGT85C2, UGT74G1 and UGT76G1 belong to same protein family.

The predicted models were assessed by evaluation through PROCHECK analysis (Laskowski *et al*., 1993). The number of residues in the most favored regions of Ramachandran plot deciphers the quality of predicted model. Another evaluating factor is the overall average value of G-factors. G-factors include the dihedral angles involved in phi-psi distributions, chi1-chi2 distribution, chi1, chi3-chi4, omega and main chain covalent forces constituting main chain bond angles and bond lengths. G-factors determine the degree of unusualness in the predicted model. The overall average values of G-factors below -0.5 and -1.0 corresponds to unusual and highly unusual properties of the model, respectively. Comparison of following parameters of the predicted models with respect to the template could help to assess the quality of the model.

The model for enzymatic protein KAH possessed 83.6% of the residues in the most favored regions of the Ramachandran plot (Fig. 4a), comparable to 89.6% for its template 2ve3A. The average value of G-factors was observed to be 0.03 (more than -0.5) for KAH and was 0.12 for its template 2ve3A. The data thus suggests that the model predicted for KAH is usual and worth use for representing the protein KAH. Ramachandran plot for UGT85C2 predicted model showed 87.2% of the residues in most favored regions (Fig. 4b), comparable to 89.6% for its template protein 2pq6A. The G-factor average value was 0.02 (more than -0.5) for UGT85C2 and was 0.35 for its template. Both the features of target protein were in close proximity to the template and within the required limits. Thus the model predicted for UGT85C2 was normal not unusual and can be used to represent the target protein. The model predicted for UGT74G1 showed 85.4% of residues in the most favored regions of Ramachandran plot (Fig. 4c), whereas 92.3% for its template 2vg8A. The average value of G-factor was -0.09 (more than -0.5) for UGT74G1. The G-factor was comparatively higher for this target protein to that of 0.11 for its template protein. In this case, both the features in target protein showed a higher variability from template protein. However, the G-factor was within limits which suggest that the predicted model is usual and can be used as a representative of the protein UGT74G1. The predicted three dimensional model for UGT76G1 possessed 87.0% of the residues in the most favored regions of Ramachandran plot (Fig. 4d), comparable to 89.6% for its template protein 2pq6A. The G-factor average was 0.02 (more than -0.5) for the modeled target protein and 0.35 for the template. Like UGT85C2, UGT76G1 was observed to be in close proximity of the template protein and the respective values were within limits. Hence, the predicted model was enough usual to represent structure of the target protein UGT76G1.

Recently, a new measure QMEAN Z-score has been introduced to determine the closeness of the computationally predicted models with the experimentally validated structures (Benkert *et al*., 2011). QMEAN score is a linear combination of six structural features: two distance dependent interaction potentials of mean force based on C-β atoms and on all atom types, torsion angle potential evaluating the local backbone geometry of the structure, solvation potential describing the burial status of residues and solvent accessibility in the form of SSE and ACC agreements. In order to calculate the QMEAN Z-score of a predicted model, the normalized raw scores of the model are compared to the scores obtained for a representative set of high resolution X-ray structures of similar size (number of residues of query protein ±10%). The output is obtained in the form of a




| Plot statistics | No. | (%) |
|---|---|---|
| Residues in most favoured regions [A.B.L] | 327 | 83.6 |
| Residues in additional allowed regions [a.b.l.p] | 52 | 13.3 |
| Residues in generously allowed regions [~a,~b,~l~p] | 8 | 2.0 |
| Residues in disallowed regions | 4 | 1.0 |
| No. of non-glycine and non-proline residues | 391 | 100.0 |
| No. of end-residues (excl. Gly and Pro) | 1 | |
| No. of glycine residues (shown as triangles) | 29 | |
| No. of proline residues | 26 | |
| Total No. of residues | 447 | |

| Plot statistics | No. | (%) |
|---|---|---|
| Residues in most favoured regions [A.B.L] | 355 | 87.2 |
| Residues in additional allowed regions [a.b.l.p] | 45 | 11.1 |
| Residues in generously allowed regions [~a,~b,~l~p] | 7 | 1.7 |
| Residues in disallowed regions | 0 | 0.0 |
| No. of non-glycine and non-proline residues | 407 | 100.0 |
| No. of end-residues (excl. Gly and Pro) | 2 | |
| No. of glycine residues (shown as triangles) | 32 | |
| No. of proline residues | 27 | |
| Total No. of residues | 468 | |

| Plot statistics | No. | (%) |
|---|---|---|
| Residues in most favoured regions [A.B.L] | 340 | 85.4 |
| Residues in additional allowed regions [a.b.l.p] | 47 | 11.8 |
| Residues in generously allowed regions [~a,~b,~l~p] | 7 | 1.8 |
| Residues in disallowed regions | 4 | 1.0 |
| No. of non-glycine and non-proline residues | 398 | 100.0 |
| No. of end-residues (excl. Gly and Pro) | 2 | |
| No. of glycine residues (shown as triangles) | 33 | |
| No. of proline residues | 16 | |
| Total No. of residues | 449 | |

| Plot statistics | No. | (%) |
|---|---|---|
| Residues in most favoured regions [A.B.L] | 341 | 87.2 |
| Residues in additional allowed regions [a.b.l.p] | 47 | 12.7 |
| Residues in generously allowed regions [~a,~b,~l~p] | 3 | 0.8 |
| Residues in disallowed regions | 1 | 0.3 |
| No. of non-glycine and non-proline residues | 392 | 100.0 |
| No. of end-residues (excl. Gly and Pro) | 2 | |
| No. of glycine residues (shown as triangles) | 27 | |
| No. of proline residues | 24 | |
| Total No. of residues | 445 | |

Fig. 4(a-b): Ramachandran plots for the predicted three dimensional models of (a) KAH, (b) UGT85C2, (c) UGT74G1 and (d) UGT76G1

model quality plot in which the query model is marked on normalized QMEAN score data obtained from high resolution structures of similar size. The predicted model of KAH has normalized QMEAN score less than 1 that lies within the prescribed limits (Fig. 5a). Thus the QMEAN score suggests

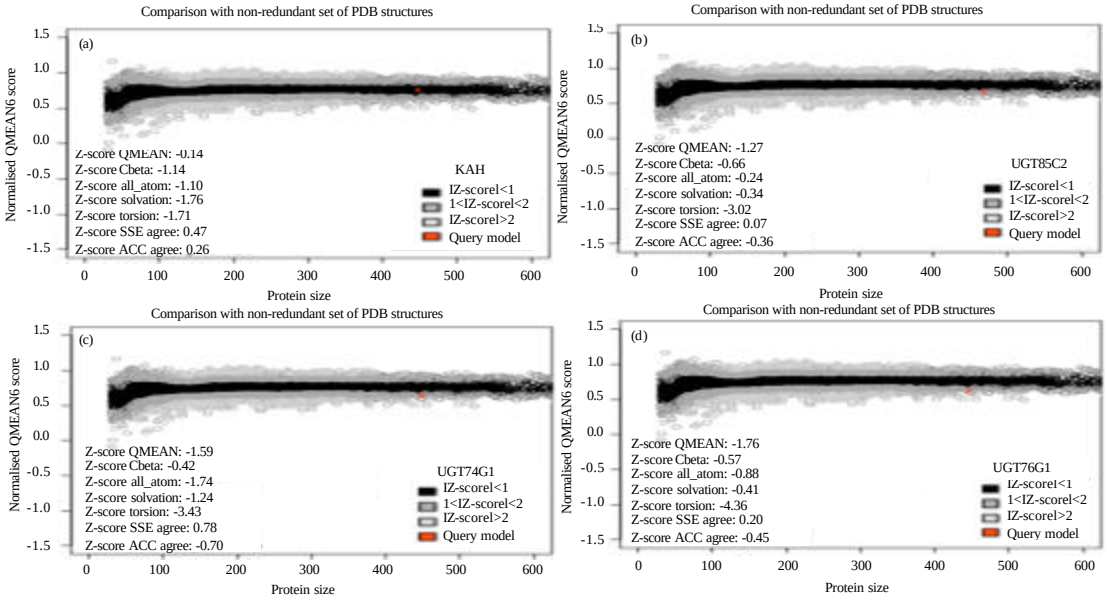



Fig. 5(a-d): Normalized QMEAN Z-scores for the proposed three dimensional models of (a) KAH, (b) UGT85C2, (c) UGT74G1 and (d) UGT76G1

that model is of good quality. The QMEAN score for UGT85C2 was predicted to be greater than 1 but less than 2 which was within the defined limits (Fig. 5b). Hence, the model was of good quality. The QMEAN score was very similar for both UGT74G1 (Fig. 5c) and UGT76G1 (Fig. 5d) to that of UGT85C2. Hence, models for both the enzymatic proteins UGT74G1 and UGT76G1 were also of good quality. Various quality assessment features suggested that the predicted models were of good quality and these models can be used to represent these enzymatic proteins KAH, UGT85C2, UGT74G1 and UGT76G1.

**Ligand binding site prediction:** In order to predict Ligand binding sites for the obtained evaluated models, the models were submitted to Q-SiteFinder server. Ten Ligand binding sites were predicted for each query model which were later arranged on the basis of total interaction energies. Out of predicted sites, the binding site with most favorable interaction energy, area and volume was identified as first predicted binding site. Ligand binding sites for the four query proteins are shown in Fig. 6. The red marked site was the first predicted site. The volume for predicted sites were 1221 cubic Angstroms for KAH (Fig. 6a), 300 cubic Angstroms for UGT85C2 (Fig. 6b), 1536 cubic Angstroms for UGT74G1 (Fig. 6c) and 513 cubic Angstroms for UGT76G1 (Fig. 6d). Various residues present in putative ligand binding sites of proteins KAH, UGT85C, UGT74G1 and UGT76G1 are shown in Supplementary Table 2 (Appendix).

**Molecular docking analysis:** Molecular docking predicts the stable protein-ligand interactions on the basis of protein-ligand complex geometries and binding energies (Shakyawar *et al.*, 2011). The assessment is based on binding affinity between the protein and ligand molecules. It allows the accurate prediction of binding geometry and binding energies. It has been known that good binding geometry prediction depends upon the accurate prediction of binding energy. Least is the binding energy, higher is the binding affinity. It has been found that the computationally predicted and

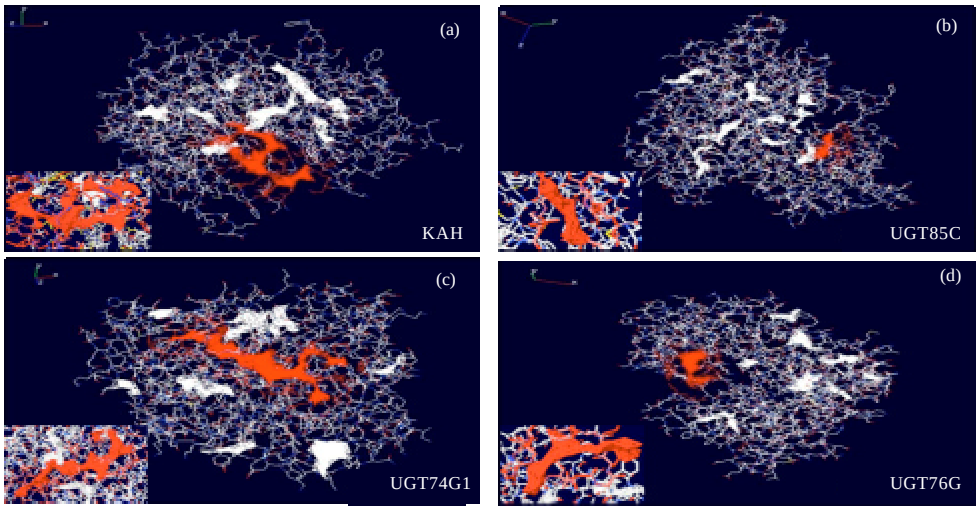



Fig. 6(a-d): Ligand binding sites predicted for three dimensional models of (a) KAH, (b) UGT85C2, (c) UGT74G1 and (d) UGT76G1. Ten sites were predicted for each protein. The site highlighted with red color is the first predicted site and a close view of the same has been shown as inset

experimentally solved binding energies are correlated, that further supports the computationally predicted data (Bikadi and Hazai, 2009; Lakshmi *et al.*, 2011). Molecular docking was carried out to analyze the affinity of the studied enzymatic proteins with the substrates of the steviol glycoside biosynthesis pathway (Supplementary Table 1).

First committed step of steviol glycoside biosynthesis pathway involves conversion of ent-kaurenoic acid to steviol by the activity of enzyme KAH. The interaction affinity of KAH, UGT85C2, UGT74G1 and UGT76G1 was evaluated for ent-kaurenoic acid and steviol. The model predicted for KAH showed highest affinity for the ligand steviol (-9.23 kcal $mol^{-1}$), followed by steviolmonoside (-9.07 kcal $mol^{-1}$) and ent-kaurenoic acid (-8.07 kcal $mol^{-1}$). UGT85C2 was observed to possess highest affinity for ent-kaurenoic acid (-7.30 kcal $mol^{-1}$), followed by steviolmonoside (-6.91 kcal $mol^{-1}$) and steviol (-6.32 kcal $mol^{-1}$). Similarly, UGT74G1 model showed highest affinity for ent-kaurenoic acid (-7.97 kcal $mol^{-1}$) followed by steviol (-7.19 kcal $mol^{-1}$) and steviolmonoside (-2.41 kcal $mol^{-1}$). The docking results for three dimensional model of UGT76G1 suggested its highest binding affinity for ent-kaurenoic acid (-7.11 kcal $mol^{-1}$), followed by steviol (-6.91 kcal $mol^{-1}$) and steviolmonoside (-6.48 kcal $mol^{-1}$). Results demonstrate that these enzymes possibly have the ability to interact with more than one ligands of steviol glycoside biosynthesis pathway. This is in well support by the presence of number of active binding sites on the protein surfaces. Data documents the probability of existence of alternative pathways or branching for the synthesis of various metabolites of steviol glycoside biosynthesis pathway.

This is the first report on prediction of three dimensional models for the enzymatic proteins KAH, UGT85C2, UGT74G1 and UGT76G1 from *Stevia rebaudiana*. From the evaluation data of PROCHECK and QMEAN Z-score, it was found that the predicted models were enough usual and good to represent the query proteins. Further, prediction of ligand binding sites for the predicted models of enzymatic proteins has opened way to carry out various manipulative studies that could facilitate metabolic engineering. Understanding three dimensional structures and ligand binding

sites provide the possibility of manipulating enzyme sequences for interaction with more ligands in addition to the already known ones. This study can help find way to increase the turn over production of steviol glycosides. Models for UGT85C2 and UGT76G1 were predicted on the basis of a common template, suggesting the possibility of single UGT that might be catalyzing two or more reactions by interacting with more ligands. Earlier study has also documented the multi-substrate activity of UGTs (Wang and Hou, 2009). This work suggested that the pathway may be modulated to specifically enhance the production of desired steviol glycosides to obtain final yield sweeter and less bitter.

## ACKNOWLEDGMENTS

We thank the Director, CSIR-IHBT, Palampur for his continuous support and encouragement. We would like to acknowledge the financial support from Council of Scientific and Industrial Research (CSIR), Govt of India. PG is also thankful to CSIR for proving fellowship as JRF.

## APPENDIX

Supplementary Table 1: Detailed description of models and their evaluation data for proteins, KAH, UGT85C2, UGT74G1 and UGT76G1

| Property/protein | KAH | UGT85C2 | UGT74G1 | UGT76G1 |
|---|---|---|---|---|
| NCBI Accession number | ACD93722.1 | AAR06922.1 | AAR06920.1 | AAR06912.1 |
| No. of amino acids | 476 | 483 | 460 | 458 |
| **Secondary structure prediction** | | | | |
| α-helices | 264 | 196 | 189 | 189 |
| Extended β-turns | 37 | 64 | 57 | 61 |
| Random coils | 174 | 220 | 213 | 207 |
| **Model prediction** | | | | |
| Template used | 2ve3A with atomic resolution of its X ray crystal structure being 2.1 A | 2pq6A with atomic resolution of its X ray crystal structure being 2.10 A | 2vg8A with atomic resolution of its X ray crystal structure being 1.75 A | 2pq6A with atomic resolution of its X ray crystal structure being 2.10 A |
| No. of residues in predicted model | 447 | 468 | 449 | 445 |
| No. of chains | 1 | 1 | 1 | 1 |
| No. of strands | 12 | 13 | 13 | 13 |
| No. of α-helices | 17 | 20 | 16 | 19 |
| No. of 3,10 helices | 3 | 6 | 3 | 4 |
| Topology | Mixed and antiparallel | Parallel | Parallel | Parallel |
| **Model evaluation: Procheck (Ramachandran plot) analysis** | | | | |
| Residues in most favored region (%) | 83.6 | 87.2 | 85.4 | 87.0 |
| Residues in additional allowed regions (%) | 13.3 | 11.1 | 11.8 | 12.0 |
| Residues in generously allowed regions (%) | 2.0 | 1.7 | 1.8 | 0.8 |
| Residues in disallowed regions (%) | 1.0 | 0.0 | 1.0 | 0.3 |
| **Ligand binding site prediction** | | | | |
| Protein volume (Cubic angstroms) | 42817 | 44251 | 41694 | 41998 |
| Prediction site volume (Cubic angstroms) | 1221 | 300 | 1536 | 513 |

Supplementary Table 2: Residues present in putative ligand binding sites of proteins KAH, UGT85C, UGT74G1, UGT76G1

| Potential ligand binding site residues | | | |
|---|---|---|---|
| KAH | 539 CE1 PHE 94 | 816 O MET 128 | 3174 C MET 422 |
| | 541 CZ PHE 94 | 817 CB MET 128 | 3175 O MET 422 |
| | 557 CA GLU 97 | 818 CG MET 128 | 3176 CB MET 422 |
| | 558 C GLU 97 | 821 N ARG 129 | 3180 N CYS 423 |
| | 559 O GLU 97 | 822 CA ARG 129 | 3181 CA CYS 423 |
| | 560 CB GLU 97 | 823 C ARG 129 | 3182 C CYS 423 |
| | 561 CG GLU 97 | 824 O ARG 129 | 3184 CB CYS 423 |
| | 562 CD GLU 97 | 825 CB ARG 129 | 3185 SG CYS 423 |
| | 563 OE1 GLU 97 | 826 CG ARG 129 | 3186 N LEU 424 |
| | 564 OE2 GLU 97 | 827 CD ARG 129 | 3187 CA LEU 424 |
| | 565 N ASN 98 | 828 NE ARG 129 | 3188 C LEU 424 |
| | 566 CA ASN 98 | 829 CZ ARG 129 | 3189 O LEU 424 |
| | 567 C ASN 98 | 830 NH1 ARG 129 | 3190 CB LEU 424 |
| | 568 O ASN 98 | 832 N LYS 130 | 3191 CG LEU 424 |
| | 569 CB ASN 98 | 837 CG LYS 130 | 3192 CD1 LEU 424 |
| | 570 CG ASN 98 | 838 CD LYS 130 | 3193 CD2 LEU 424 |
| | 572 ND2 ASN 98 | 839 CE LYS 130 | 3194 N GLY 425 |
| | 591 CA VAL 101 | 840 NZ LYS 130 | 3195 CA GLY 425 |
| | 592 C VAL 101 | 853 CB LEU 132 | 3196 C GLY 425 |
| | 593 O VAL 101 | 854 CG LEU 132 | 3197 O GLY 425 |
| | 594 CB VAL 101 | 855 CD1 LEU 132 | 3198 N LYS 426 |
| | 595 CG1 VAL 101 | 856 CD2 LEU 132 | 3199 CA LYS 426 |
| | 596 CG2 VAL 101 | 878 CD2 TYR 135 | 3217 CA PHE 428 |
| | 597 N ALA 102 | 880 CE2 TYR 135 | 3218 C PHE 428 |
| | 598 CA ALA 102 | 888 CG LEU 136 | 3219 O PHE 428 |
| | 599 C ALA 102 | 889 CD1 LEU 136 | 3220 CB PHE 428 |
| | 600 O ALA 102 | 890 CD2 LEU 136 | 2119 CB SER 291 |
| | 601 CB ALA 102 | 1292 CD2 PHE 185 | 2120 OG SER 291 |
| | 602 N SER 103 | 1293 CE1 PHE 185 | 2134 N SER 294 |
| | 603 CA SER 103 | 1294 CE2 PHE 185 | 2135 CA SER 294 |
| | 605 O SER 103 | 1295 CZ PHE 185 | 2138 CB SER 294 |
| | 606 CB SER 103 | 2027 CG ASN 279 | 2139 OG SER 294 |
| | 607 OG SER 103 | 2029 ND2 ASN 279 | 2588 CG MET 348 |
| | 628 CD1 TRP 105 | 2049 O LEU 282 | 2589 SD MET 348 |
| | 630 NE1 TRP 105 | 2051 CG LEU 282 | 2590 CE MET 348 |
| | 631 CE2 TRP 105 | 2052 CD1 LEU 282 | 2625 N PRO 353 |
| | 633 CZ2 TRP 105 | 2053 CD2 LEU 282 | 2626 CA PRO 353 |
| | 635 CH2 TRP 105 | 2054 N LEU 283 | 2629 CB PRO 353 |
| | 717 CA SER 116 | 2055 CA LEU 283 | 2630 CG PRO 353 |
| | 718 C SER 116 | 2056 C LEU 283 | 2631 CD PRO 353 |
| | 719 O SER 116 | 2057 O LEU 283 | 2632 N VAL 354 |
| | 720 CB SER 116 | 2058 CB LEU 283 | 2635 O VAL 354 |
| | 721 OG SER 116 | 2059 CG LEU 283 | 2636 CB VAL 354 |
| | 722 N LEU 117 | 2061 CD2 LEU 283 | 2637 CG1 VAL 354 |
| | 723 CA LEU 117 | 2081 N ALA 286 | 2638 CG2 VAL 354 |
| | 724 C LEU 117 | 2082 CA ALA 286 | 2655 CB THR 357 |
| | 725 O LEU 117 | 2083 C ALA 286 | 2656 OG1 THR 357 |
| | 726 CB LEU 117 | 2084 O ALA 286 | 2657 CG2 THR 357 |

Supplementary Table 2: Continue

| Potential ligand binding site residues | | | |
|---|---|---|---|
| | 727 CG LEU 117 | 2085 CB ALA 286 | 2675 CG ARG 359 |
| | 730 N LEU 118 | 2086 N GLY 287 | 2676 CD ARG 359 |
| | 731 CA LEU 118 | 2087 CA GLY 287 | 2677 NE ARG 359 |
| | 732 C LEU 118 | 2088 C GLY 287 | 2678 CZ ARG 359 |
| | 733 O LEU 118 | 2089 O GLY 287 | 2679 NH1 ARG 359 |
| | 734 CB LEU 118 | 2109 CA THR 290 | 2680 NH2 ARG 359 |
| | 735 CG LEU 118 | 2110 C THR 290 | 2701 CD1 LEU 362 |
| | 736 CD1 LEU 118 | 2111 O THR 290 | 2851 NE1 TRP 380 |
| | 737 CD2 LEU 118 | 2112 CB THR 290 | 2852 CE2 TRP 380 |
| | 747 C ILE 120 | 2113 OG1 THR 290 | 2854 CZ2 TRP 380 |
| | 748 O ILE 120 | 2114 CG2 THR 290 | 2855 CZ3 TRP 380 |
| | 750 CG1 ILE 120 | 2115 N SER 291 | 2856 CH2 TRP 380 |
| | 751 CG2 ILE 120 | 2116 CA SER 291 | 3120 O VAL 414 |
| | 752 CD1 ILE 120 | 2117 C SER 291 | 3125 CA PRO 415 |
| | 753 N ARG 121 | 2118 O SER 291 | 3126 C PRO 415 |
| | 754 CA ARG 121 | 3139 CE1 PHE 416 | 3127 O PRO 415 |
| | 755 C ARG 121 | 3140 CE2 PHE 416 | 3128 CB PRO 415 |
| | 756 O ARG 121 | 3141 CZ PHE 416 | 3131 N PHE 416 |
| | 757 CB ARG 121 | 3142 N GLY 417 | 3132 CA PHE 416 |
| | 758 CG ARG 121 | 3143 CA GLY 417 | 3133 C PHE 416 |
| | 759 CD ARG 121 | 3144 C GLY 417 | 3134 O PHE 416 |
| | 760 NE ARG 121 | 3146 N GLY 418 | 3135 CB PHE 416 |
| | 762 NH1 ARG121 | 3147 CA GLY 418 | 3136 CG PHE 416 |
| | 764 N GLY 122 | 3148 C GLY 418 | 3137 CD1 PHE 416 |
| | 765 CA GLY 122 | 3149 O GLY 418 | 3221 CG PHE 428 |
| | 766 C GLY 122 | 3155 CA PRO 420 | 3222 CD1 PHE 428 |
| | 767 O GLY 122 | 3156 C PRO 420 | 3223 CD2 PHE 428 |
| | 768 N ASP 123 | 3157 O PRO 420 | 3224 CE1 PHE 428 |
| | 786 CA ALA 125 | 3158 CB PRO 420 | 3225 CE2 PHE 428 |
| | 787 C ALA 125 | 3159 CG PRO 420 | 3226 CZ PHE 428 |
| | 788 O ALA 125 | 3160 CD PRO 420 | 3227 N ALA 429 |
| | 789 CB ALA 125 | 3161 N ARG 421 | 3228 CA ALA 429 |
| | 790 N LYS 126 | 3162 CA ARG 421 | 3229 C ALA 429 |
| | 791 CA LYS 126 | 3163 C ARG 421 | 3230 O ALA 429 |
| | 792 C LYS 126 | 3164 O ARG 421 | 3231 CB ALA 429 |
| | 793 O LYS 126 | 3165 CB ARG 421 | |
| | 794 CB LYS 126 | 3166 CG ARG 421 | |
| | 795 CG LYS 126 | 3167 CD ARG 421 | |
| | 796 CD LYS 126 | 3168 NE ARG 421 | |
| | 797 CE LYS 126 | 3169 CZ ARG 421 | |
| | 799 N TRP 127 | 3170 NH1 ARG 421 | |
| | 814 CA MET 128 | 3171 NH2 ARG 421 | |
| | 815 C MET 128 | 3172 N MET 422 | |
| | 173 CA MET 422 | | |
| UGT85C2 | 2326 CA ASN 300 | 2382 CE MET 307 | 2607 CG LEU 334 |
| | 2327 C ASN 300 | 2553 CD1 TRP 328 | 2608 CD1 LEU 334 |
| | 2329 CB ASN 300 | 2555 NE1 TRP 328 | 2610 N VAL 335 |
| | 2330 CG ASN 300 | 2556 CE2 TRP 328 | 2614 CB VAL 335 |

Supplementary Table 2: Continue

| Potential ligand binding site residues | | | |
|---|---|---|---|
| | 2331 OD1 ASN 300 | 2558 CZ2 TRP 328 | 2616 CG2 VAL 335 |
| | 2332 ND2 ASN 300 | 2562 CA ILE 329 | 2635 CD GLU 338 |
| | 2333 N PHE 301 | 2563 C ILE 329 | 2636 OE1 GLU 338 |
| | 2334 CA PHE 301 | 2564 O ILE 329 | 2637 OE2 GLU 338 |
| | 2335 C PHE 301 | 2565 CB ILE 329 | 2788 O ALA 357 |
| | 2336 O PHE 301 | 2566 CG1 ILE 329 | 2791 CA SER 358 |
| | 2337 CB PHE 301 | 2567 CG2 ILE 329 | 2792 C SER 358 |
| | 2338 CG PHE 301 | 2569 N ILE 330 | 2793 O SER 358 |
| | 2340 CD2 PHE 301 | 2570 CA ILE 330 | 2796 N TRP 359 |
| | 2344 N GLY 302 | 2571 C ILE 330 | 2797 CA TRP 359 |
| | 2345 CA GLY 302 | 2572 O ILE 330 | 2798 C TRP 359 |
| | 2346 C GLY 302 | 2573 CB ILE 330 | 2800 CB TRP 359 |
| | 2347 O GLY 302 | 2574 CG1 ILE 330 | 2801 CG TRP 359 |
| | 2348 N SER 303 | 2575 CG2 ILE 330 | 2802 CD1 TRP 359 |
| | 2349 CA SER 303 | 2576 CD1 ILE 330 | 2803 CD2 TRP 359 |
| | 2350 C SER 303 | 2577 N ARG 331 | 2804 NE1 TRP 359 |
| | 2351 O SER 303 | 2578 CA ARG 331 | 2805 CE2 TRP 359 |
| | 2352 CB SER 303 | 2579 C ARG 331 | 2810 N CYS 360 |
| | 2353 OG SER 303 | 2580 O ARG 331 | 2811 CA CYS 360 |
| | 2361 N THR 305 | 2581 CB ARG 331 | 2812 C CYS 360 |
| | 2363 C THR 305 | 2582 CG ARG 331 | 2813 O CYS 360 |
| | 2364 O THR 305 | 2583 CD ARG 331 | 2814 CB CYS 360 |
| | 2368 N VAL 306 | 2584 NE ARG 331 | 2815 SG CYS 360 |
| | 2369 CA VAL 306 | 2585 CZ ARG 331 | |
| | 2370 C VAL 306 | 2587 NH2 ARG 331 | |
| | 2372 CB VAL 306 | 2588 N SER 332 | |
| | 2373 CG1 VAL 306 | 2602 N LEU 334 | |
| | 2375 N MET 307 | 2603 CA LEU 334 | |
| | 2376 CA MET 307 | 2604 C LEU 334 | |
| | 379 CB MET 307 | 2606 CB LEU 334 | |
| | 2380 CG MET 307 | | |
| | 2381 SD MET 307 | | |
| UGT74G1 | 68 CD1 PHE 18 | 1326 O GLN 182 | 2615 CA LYS 338 |
| | 70 CE1 PHE 18 | 1327 CB GLN 182 | 2616 C LYS 338 |
| | 71 CE2 PHE 18 | 1328 CG GLN 182 | 2617 O LYS 338 |
| | 72 CZ PHE 18 | 1329 CD GLN 182 | 2618 CB LYS 338 |
| | 84 CB LEU 20 | 1330 OE1 GLN 182 | 2619 CG LYS 338 |
| | 85 CG LEU 20 | 1331 NE2 GLN 182 | 2620 CD LYS 338 |
| | 86 CD1 LEU 20 | 1332 N ASN 183 | 2621 CE LYS 338 |
| | 87 CD2 LEU 20 | 1333 CA ASN 183 | 2622 NZ LYS 338 |
| | 89 CA GLN 21 | 1334 C ASN 183 | 2623 N GLN 339 |
| | 90 C GLN 21 | 1335 O ASN 183 | 2624 CA GLN 339 |
| | 91 O GLN 21 | 1336 CB ASN 183 | 2625 C GLN 339 |
| | 92 CB GLN 21 | 1337 CG ASN 183 | 2627 CB GLN 339 |
| | 93 CG GLN 21 | 1338 OD1 ASN 183 | 2628 CG GLN 339 |
| | 94 CD GLN 21 | 1339 ND2 ASN 183 | 2629 CD GLN 339 |
| | 95 OE1 GLN 21 | 1340 N HIS 184 | 2630 OE1 GLN 339 |
| | 96 NE2 GLN 21 | 1341 CA HIS 184 | 2631 NE2 GLN 339 |

Supplementary Table 2: Continue

| Potential ligand binding site residues | | |
|---|---|---|
| 97 N GLY 22 | 1344 CB HIS 184 | 2632 N LEU 340 |
| 98 CA GLY 22 | 1345 CG HIS 184 | 2633 CA LEU 340 |
| 99 C GLY 22 | 1346 ND1 HIS 184 | 2634 C LEU 340 |
| 100 O GLY 22 | 1354 CB GLU 185 | 2636 CB LEU 340 |
| 101 N HIS 23 | 1355 CG GLU 185 | 2637 CG LEU 340 |
| 102 CA HIS 23 | 1356 CD GLU 185 | 2638 CD1 LEU 340 |
| 106 CG HIS 23 | 1357 OE1 GLU 185 | 2639 CD2 LEU 340 |
| 107 ND1 HIS 23 | 1358 OE2 GLU 185 | 2640 N ASP 341 |
| 108 CD2 HIS 23 | 1363 CB GLN 186 | 2641 CA ASP 341 |
| 109 CE1 HIS 23 | 1364 CG GLN 186 | 2644 CB ASP 341 |
| 110 NE2 HIS 23 | 1365 CD GLN 186 | 2645 CG ASP 341 |
| 113 C ILE 24 | 1366 OE1 GLN 186 | 2646 OD1 ASP 341 |
| 114 O ILE 24 | 1367 NE2 GLN 186 | 2647 OD2 ASP 341 |
| 115 CB ILE 24 | 1402 CB TRP 191 | 2740 CG HIS 354 |
| 116 CG1 ILE 24 | 1403 CG TRP 191 | 2741 ND1 HIS 354 |
| 118 CD1 ILE 24 | 1404 CD1 TRP 191 | 2742 CD2 HIS 354 |
| 119 N ASN 25 | 1405 CD2 TRP 191 | 2743 CE1 HIS 354 |
| 120 CA ASN 25 | 1406 NE1 TRP 191 | 2744 NE2 HIS 354 |
| 121 C ASN 25 | 1407 CE2 TRP 191 | 2747 C CYS 355 |
| 122 O ASN 25 | 1408 CE3 TRP 191 | 2748 O CYS 355 |
| 123 CB ASN 25 | 1409 CZ2 TRP 191 | 2751 N GLY 356 |
| 124 CG ASN 25 | 1410 CZ3 TRP 191 | 2752 CA GLY 356 |
| 125 OD1 ASN 25 | 1411 CH2 TRP 191 | 2753 C GLY 356 |
| 126 ND2 ASN 25 | 1440 CG LEU 195 | 2754 O GLY 356 |
| 133 CD PRO 26 | 1441 CD1 LEU 195 | 2755 N PHE 357 |
| 149 CB ILE 28 | 1442 CD2 LEU 195 | 2756 CA PHE 357 |
| 150 CG1 ILE 28 | 1475 CE1 PHE 199 | 2757 C PHE 357 |
| 152 CD1 ILE 28 | 1476 CE2 PHE 199 | 2759 CB PHE 357 |
| 185 NZ LYS 32 | 1477 CZ PHE 199 | 2760 CG PHE 357 |
| 341 O THR 53 | 1936 CA PHE 254 | 2762 CD2 PHE 357 |
| 345 N LEU 54 | 1937 C PHE 254 | 2766 N ASN 358 |
| 346 CA LEU 54 | 1938 O PHE 254 | 2767 CA ASN 358 |
| 347 C LEU 54 | 1940 CG PHE 254 | 2768 C ASN 358 |
| 348 O LEU 54 | 1941 CD1 PHE 254 | 2769 O ASN 358 |
| 351 CD1 LEU 54 | 1942 CD2 PHE 254 | 2770 CB ASN 358 |
| 361 N HIS 56 | 1943 CE1 PHE 254 | 2771 CG ASN 358 |
| 363 C HIS 56 | 1944 CE2 PHE 254 | 2772 OD1 ASN 358 |
| 364 O HIS 56 | 1945 CZ PHE 254 | 2773 ND2 ASN 358 |
| 366 CG HIS 56 | 1964 C TYR 257 | 2774 N SER 359 |
| 367 ND1 HIS 56 | 1965 O TYR 257 | 2775 CA SER 359 |
| 368 CD2 HIS 56 | 1967 CG TYR 257 | 2778 CB SER 359 |
| 369 CE1 HIS 56 | 1968 CD1 TYR 257 | 2779 OG SER 359 |
| 370 NE2 HIS 56 | 1969 CD2 TYR 257 | 2800 CG GLU 362 |
| 372 CA SER 57 | 1970 CE1 TYR 257 | 2801 CD GLU 362 |
| 373 C SER 57 | 1971 CE2 TYR 257 | 2802 OE1 GLU 362 |
| 374 O SER 57 | 1972 CZ TYR 257 | 2803 OE2 GLU 362 |
| 375 CB SER 57 | 1973 OH TYR 257 | 2896 CB PHE 376 |
| 376 OG SER 57 | 2148 CE2 TYR 278 | 2897 CG PHE 376 |

Supplementary Table 2: Continue

| Potential ligand binding site residues | | |
|---|---|---|
| 541 CA ALA 81 | 2149 CZ TYR 278 | 2898 CD1 PHE 376 |
| 542 C ALA 81 | 2150 OH TYR 278 | 2899 CD2 PHE 376 |
| 544 CB ALA 81 | 2159 CA ALA 280 | 2900 CE1 PHE 376 |
| 545 N GLY 82 | 2160 C ALA 280 | 2901 CE2 PHE 376 |
| 546 CA GLY 82 | 2162 CB ALA 280 | 2902 CZ PHE 376 |
| 547 C GLY 82 | 2163 N PHE 281 | 2903 N SER 377 |
| 548 O GLY 82 | 2165 C PHE 281 | 2904 CA SER 377 |
| 549 N GLU 83 | 2166 O PHE 281 | 2905 C SER 377 |
| 550 CA GLU 83 | 2174 N GLY 282 | 2906 O SER 377 |
| 554 CG GLU 83 | 2175 CA GLY 282 | 2907 CB SER 377 |
| 555 CD GLU 83 | 2176 C GLY 282 | 2908 OG SER 377 |
| 556 OE1 GLU 83 | 2177 O GLY 282 | 2909 N ASP 378 |
| 557 OE2 GLU 83 | 2178 N SER 283 | 2910 CA ASP 378 |
| 571 CD2 TYR 85 | 2179 CA SER 283 | 2911 C ASP 378 |
| 573 CE2 TYR 85 | 2180 C SER 283 | 2912 O ASP 378 |
| 574 CZ TYR 85 | 2181 O SER 283 | 2913 CB ASP 378 |
| 575 OH TYR 85 | 2182 CB SER 283 | 2914 CG ASP 378 |
| 581 CG LEU 86 | 2183 OG SER 283 | 2915 OD1 ASP 378 |
| 582 CD1 LEU 86 | 2184 N LEU 284 | 2916 OD2 ASP 378 |
| 827 CB SER 118 | 2185 CA LEU 284 | 2917 N GLN 379 |
| 828 OG SER 118 | 2186 C LEU 284 | 2918 CA GLN 379 |
| 834 CG MET 119 | 2187 O LEU 284 | 2921 CB GLN 379 |
| 835 SD MET 119 | 2188 CB LEU 284 | 2922 CG GLN 379 |
| 836 CE MET 119 | 2189 CG LEU 284 | 2923 CD GLN 379 |
| 983 CE1 PHE 138 | 2190 CD1 LEU 284 | 2925 NE2 GLN 379 |
| 986 N THR 139 | 2191 CD2 LEU 284 | 2931 OG1 THR 380 |
| 987 CA THR 139 | 2192 N VAL 285 | 2937 CB THR 381 |
| 988 C THR 139 | 2193 CA VAL 285 | 2938 OG1 THR 381 |
| 989 O THR 139 | 2196 CB VAL 285 | 2939 CG2 THR 381 |
| 990 CB THR 139 | 2198 CG2 VAL 285 | 2945 CG ASN 382 |
| 991 OG1 THR 139 | 2393 CB VAL 309 | 2946 OD1 ASN 382 |
| 992 CG2 THR 139 | 2394 CG1 VAL 309 | 2947 ND2 ASN 382 |
| 993 N GLN 140 | 2395 CG2 VAL 309 | |
| 994 CA GLN 140 | 2591 C ALA 335 | |
| 997 CB GLN 140 | 2592 O ALA 335 | |
| 998 CG GLN 140 | 2594 N TRP 336 | |
| 999 CD GLN 140 | 2595 CA TRP 336 | |
| 1000 OE1 GLN140 | 2596 C TRP 336 | |
| 1001 NE2 GLN 140 | 2598 CB TRP 336 | |
| 1024 CB VAL 144 | 2599 CG TRP 336 | |
| 1025 CG1 VAL144 | 2600 CD1 TRP 336 | |
| 1026 CG2 VAL144 | 2601 CD2 TRP 336 | |
| 1045 CB LEU 147 | 2602 NE1 TRP 336 | |
| 1046 CG LEU 147 | 2603 CE2 TRP 336 | |
| 1047 CD1 LEU147 | 2604 CE3 TRP 336 | |
| 1048 CD2 LEU 147 | 2605 CZ2 TRP 336 | |
| 1054 CG TYR 148 | 2606 CZ3 TRP 336 | |
| 1055 CD1 TYR 148 | 2607 CH2 TRP 336 | |

Supplementary Table 2: Continue

| Potential ligand binding site residues | | | |
|---|---|---|---|
| | 1057 CE1 TYR 148 | 2608 N CYS 337 | |
| | 1058 CE2 TYR 148 | 2609 CA CYS 337 | |
| | 1059 CZ TYR 148 | 2610 C CYS 337 | |
| | 1060 OH TYR 148 | 2611 O CYS 337 | |
| | 1239 CG LEU 172 | 2612 CB CYS 337 | |
| | 1240 CD1 LEU 172 | 2613 SG CYS 337 | |
| | 1241 CD2 LEU 172 | 2614 N LYS 338 | |
| | 1258 NE ARG 174 | | |
| | 1259 CZ ARG 174 | | |
| | 1260 NH1 ARG 174 | | |
| | 1261 NH2 ARG 174 | | |
| | 1308 CA ILE 180 | | |
| | 1309 C ILE 180 | | |
| | 1310 O ILE 180 | | |
| | 1311 CB ILE 180 | | |
| | 1312 CG1 ILE 180 | | |
| | 1313 CG2 ILE 180 | | |
| | 1314 CD1 ILE 180 | | |
| | 1316 CA LEU 181 | | |
| | 1317 C LEU 181 | | |
| | 1318 O LEU 181 | | |
| | 1319 CB LEU 181 | | |
| | 1323 N GLN 182 | | |
| | 1324 CA GLN 182 | | |
| | 1325 C GLN 182 | | |
| UGT76G1 | 559 CA ASN 78 | 2211 OG SER 285 | 2433 CB PRO 312 |
| | 560 C ASN 78 | 2212 N GLU 286 | 2434 CG PRO 312 |
| | 561 O ASN 78 | 2213 CA GLU 286 | 2435 CD PRO 312 |
| | 562 CB ASN 78 | 2214 C GLU 286 | 2440 N PHE 314 |
| | 563 CG ASN 78 | 2215 O GLU 286 | 2441 CA PHE 314 |
| | 564 OD1 ASN 78 | 2216 CB GLU 286 | 2442 C PHE 314 |
| | 565 ND2 ASN 78 | 2217 CG GLU 286 | 2444 CB PHE 314 |
| | 582 CA THR 81 | 2218 CD GLU 286 | 2445 CG PHE 314 |
| | 583 C THR 81 | 2219 OE1 GLU 286 | 2446 CD1 PHE 314 |
| | 585 CB THR 81 | 2220 OE2 GLU 286 | 2447 CD2 PHE 314 |
| | 586 OG1 THR 81 | 2221 N VAL 287 | 2448 CE1 PHE 314 |
| | 587 CG2 THR 81 | 2222 CA VAL 287 | 2449 CE2 PHE 314 |
| | 588 N HIS 82 | 2225 CB VAL 287 | 2450 CZ PHE 314 |
| | 593 CG HIS 82 | 2226 CG1 VAL 287 | 2451 N VAL 315 |
| | 594 ND1 HIS 82 | 2266 CB PHE 292 | 2455 CB VAL 315 |
| | 595 CD2 HIS 82 | 2267 CG PHE 292 | 2457 CG2 VAL 315 |
| | 596 CE1 HIS 82 | 2269 CD2 PHE 292 | 2617 O VAL 336 |
| | 597 NE2 HIS 82 | 2271 CE2 PHE 292 | 2622 CA LYS 337 |
| | 615 CD1 LEU 85 | 2272 CZ PHE 292 | 2623 C LYS 337 |
| | 2174 C SER 280 | 2297 CD1 ILE 295 | 2624 O LYS 337 |
| | 2176 CB SER 280 | 2398 NE1 TRP 308 | 2630 N TRP 338 |
| | 2177 OG SER 280 | 2399 CE2 TRP 308 | 2631 CA TRP 338 |
| | 2178 N PHE 281 | 2401 CZ2 TRP 308 | 2634 CB TRP 338 |

Supplementary Table 2: Continue

| Potential ligand binding site residues | | |
|---|---|---|
| 2179 CA PHE 281 | 2405 CA VAL 309 | 2635 CG TRP 338 |
| 2180 C PHE 281 | 2406 C VAL 309 | 2636 CD1 TRP 338 |
| 2181 O PHE 281 | 2407 O VAL 309 | 2638 NE1 TRP 338 |
| 2182 CB PHE 281 | 2408 CB VAL 309 | 2782 NE2 HIS 356 |
| 2183 CG PHE 281 | 2409 CG1 VAL 309 | |
| 2185 CD2 PHE 281 | 2410 CG2 VAL 309 | |
| 2187 CE2 PHE 281 | 2411 N VAL 310 | |
| 2189 N GLY 282 | 2412 CA VAL 310 | |
| 2190 CA GLY 282 | 2413 C VAL 310 | |
| 2191 C GLY 282 | 2414 O VAL 310 | |
| 2192 O GLY 282 | 2415 CB VAL 310 | |
| 2193 N SER 283 | 2416 CG1 VAL 310 | |
| 2194 CA SER 283 | 2417 CG2 VAL 310 | |
| 2195 C SER 283 | 2418 N ARG 311 | |
| 2196 O SER 283 | 2419 CA ARG 311 | |
| 2197 CB SER 283 | 2420 C ARG 311 | |
| 2198 OG SER 283 | 2421 O ARG 311 | |
| 2199 N THR 284 | 2422 CB ARG 311 | |
| 2200 CA THR 284 | 2423 CG ARG 311 | |
| 2201 C THR 284 | 2424 CD ARG 311 | |
| 2202 O THR 284 | 2425 NE ARG 311 | |
| 2206 N SER 285 | 2426 CZ ARG 311 | |
| 2207 CA SER 285 | 2428 NH2 ARG 311 | |
| 2208 C SER 285 | 2429 N PRO 312 | |
| 2209 O SER 285 | 2430 CA PRO 312 | |
| 2210 CB SER 285 | 2431 C PRO 312 | |
| 2432 O PRO 312 | | |

## REFERENCES

Benkert, P., M. Biasini and T. Schwede, 2011. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics, 27: 343-350.

Bikadi, Z. and E. Hazai, 2009. Application of the PM6 semi-empirical method to modeling proteins enhances docking accuracy of AutoDock. J. Cheminf., 1: 1-15.

Boobalan, C. and N. Bharathi, 2010. Comparative modeling and functional characterization of 3-D structure of a late embryogenesis abundant protein of *Arabidopsis thaliana*. J. Adv. Bioinf. Appl. Res., 1: 7-10.

Bordoli, L., F. Kiefeer, K. Arnold, P. Benkert, J. Battey and T. Schwede, 2009. Protein structure homology modeling using SWISS-MODEL workspace. Nature Protocols, 4: 1-13.

Brandle, J.E. and P.G. Telmer, 2007. Steviol glycoside biosynthesis. Phytochemistry, 68: 1855-1863.

Brazier-Hicks, M., W.A. Offen, M.C. Gershater, T.J. Revett and E.K. Lim *et al*., 2007. Characterization and engineering of the bifunctional N- and O-glucosyltransferase involved in xenobiotic metabolism in plants. PNAS, 104: 20238-20243.

Chithia, C. and A.M. Lesk, 1986. The relation between the divergence of sequence and structure in proteins. EMBO J., 5: 823-826.

Graham, S.E. and J.A. Petersen, 1999. How similar are P450s and what can their differences teach us? Arch. Biochem. Biophys., 369: 24-29.

Guleria, P., V. Kumar and S.K. Yadav, 2011. Effect of sucrose on steviol glycoside biosynthesis pathway in *Stevia rebaudiana*. Asian J. Plant Sci., 10: 401-407.

Huq, F., 2008a. Molecular modelling analysis of the metabolism of Maraviroc. Asian J. Biochem., 3: 79-89.

Huq, F., 2008b. Molecular modelling analysis of the metabolism of udenafil. Asian J. Biochem., 3: 109-117.

Joseph, B. and V.M. Nair, 2012. *In silico* analysis of Vi polysaccharide biosynthesis protein of *Salmonella typhi*. Am. J. Drug Discov. Develop., 2: 101-109.

Khan, F. and M.I. Khan, 2011. The mushroom lectins show three types of conserved domain in a bioinformatics analysis. Am. J. Biochem. Mol. Biol., 1: 375-388.

Kuhnel, K., N. Ke, M.J. Cryle, S.G. Sligar, M.A. Schuler and I. Schlichting, 2008. Crystal structure of substrate-free and retinoic acid-bound cyanobacterial Cytochrome P450 CYP120A1. Biochemistry, 47: 6552-6559.

Lakshmi, P.T.V., S. Radhika and A. Annamalai, 2011. Molecular docking analysis of phyto-ligands with multi drug resistant β-lactamases of *Staphylococcus aureus*. Trends Bioinform., 4: 23-34.

Laskowski, R.A., M.W. MacArthur, D.S. Moss and J.M. Thornton, 1993. PROCHECK: A program to check the stereochemical quality of protein structures. J. Applied Cryst., 26: 283-291.

Laurie, A.T.R. and R.M. Jackson, 2005. Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites. Bioinformatics, 21: 1908-1916.

Li, L., L.V. Modolo, L.E.L. Achnine, R.A. Dixon and X. Wang, 2007. Crystal atructure of *Medicago truncatula* UGT85H2-insights into the structural basis of a multifunctional (Iso) flavonoid glycosyltransferase. J. Mol. Biol., 370: 951-963.

Marti-Renom, M.A., A.C. Stuart, A. Fiser, R. Sanchez, F. Melo and A. Sali, 2000. Comparative protein structure modeling of genes and genomes. Ann. Rev. Biophys. Biomol. Struct., 29: 291-325.

Mugilan, A., Ajitha, M. Cathrin, M. Kumar, Devi and Thinagar, 2010. In silico secondary structure prediction method (Kalasalingam University structure prediction method) using comparative analysis. Trends Bioinform., 3: 11-19.

Offen, W., C. Martinez-Fleites, M. Yang, E. Kiat-Lim and B.G. Davis *et al*., 2006. Structure of a flavonoid glucosyltransferase reveals the basis for plant natural product modification. EMBO J., 25: 1396-1405.

Prabhavathi, M., K. Ashokkumar, N. Geetha and K.M. Saradha Devi, 2011. Homology modeling and structure prediction of thioredoxin (TRX) protein in wheat (*Triticum aestivum* L.). Int. J. Biosci., 1: 20-32.

Richman, A., A. Swanson, T. Humphrey, R. Chapman, B. McGarvey, R. Pocs and J. Brandle, 2005. Functional genomics uncovers three glycosyltransferases involved in the synthesis of the major sweet glucosides of *Stevia rebaudiana*. Plant J., 41: 56-67.

Rupasinghe, S. and M.A. Schuler, 2006. Homology modeling of plant cytochrome P450s. Phytochem. Rev., 5: 473-505.

Schwede, T., J. Kopp, N. Guex and M.C. Peitsch, 2003. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res., 31: 3381-3385.

Shakyawar, S.K., A. Goyal and V.K. Dubey, 2011. Database of *in silico* predicted potential drug target proteins in common bacterial human pathogens. Am. J. Drug Discovery Dev., 1: 70-74.

Shao, H., X. He, L. Achnine, J.W. Blount, R.A. Dixon and X. Wang, 2005. Crystal structures of a multifunctional triterpenes/flavonoid glycosyltransferases from *Medicago truccatula*. Plant Cell, 17: 3141-3154.

Smith, A.A. and M.C. Plazas, 2011. In silico characterization and homology modeling of cyanobacterial phosphoenolpyruvate carboxylase enzymes with computational tools and bioinformatics servers. Am. J. Biochem. Mol. Biol., 1: 319-336.

Stout, C.D., 2004. Cytochrome P450 conformational diversity. Structure, 12: 1921-1922.

Wang, J. and B. Hou, 2009. Glycosyltransferases: Key players involved in the modification of plant secondary metabolites. Front Biol. China, 4: 39-46.

Yadav, S.K., 2009. Computational structural analysis and kinetic studies of a cytosolic glutamine synthetase from *Camellia sinensis* (L.) O. Kuntze. Protein J., 28: 428-434.

Yadav, S.K. and P. Guleria, 2011. Steviol glycosides from *Stevia*: Biosynthesis pathway review and their application in foods and medicine. Crit. Rev. Food Sci. Nutr., (In Press).