

Full Length Research Paper

Prediction analysis of structure and function of granule-bound starch synthase gene and its encoding protein in cassava (*Manihot esculenta* Crantz)

Wei Lei^{#1,2}, Ruixia Yao^{#1} and Aimin Qiao^{1*}

¹College of Horticulture and Landscape Architecture, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China.

²The First affiliated Hospital, Guangdong Medical College, Zhanjiang 524000, China.

Accepted 12 September, 2011

Cassava (*Manihot esculenta* Crantz) is widely considered as one of the most valuable starch source plants with important food material and new energy production. Starch qua the important effectual caloric ingredient were synthesized mainly through a chain of enzymes, in which the committed-step enzymes granule-bound starch synthase (GBSS) were characterized and analyzed with multifarious bioinformatic tools and serves. The results were showed as following: the molecular structures and physicochemical properties of the Megess genes and encoding proteins were calculated; Megbss protein were localized to the cytosol lacking any transmembrane topological structure; the phylogram analysis suggested that the cassava and some crops producing starch were classified into a large groups according to significant functional association and genetic conservation. The secondary structure of the Megbss protein was mainly composed of α -helixes and random coils, and the tertiary structure were modeled successfully, with some key motifs included. Taken together, these results demonstrate that both of *Megbss* gene and its encoding protein from cassava have the typical molecular structure and function, and the study will lay theoretical foundation for molecular mechanism and genetic regulation researches of starch biosynthesis.

Keywords: Cassava, starch, granule-bound starch synthase, bioinformatics.

INTRODUCTION

Bioinformatics is the application of mathematics, statistics and computer technology to the field of molecular biology and evolutionary genetics. In the genome era, massive data from the DNA and RNA were obtained and then acquired disposal and management. In the post-genomic era, prediction and characterization of the protein structures, properties and functions become the hotspot of the current biology. Therefore, large numbers of computer softwares and online servers are provided for bioinformatics analyses (Sivakumar and Balaji, 2007). With the help of these tools, some important genes and

their encoding proteins, for example metabolic elements involved in the synthetic pathway of natural products, can be identified and analyzed on the level of computational simulation (Ling et al., 2007).

Starch, the major storage carbohydrate in plants, is used as an energy source of dormancy and growth. Due to its abundance in calories, it serves as one of the most important food resource for both human and animals and is applied in plenty of industrial processes, especially in current new energy strategy (Lin and Jeang, 2005; Hao et al., 2009). The proportion of the two components, amylose and amylopectin, determines the molecular structure, physicochemical properties and manufactural characteristics of starch (Jane and Shen, 1992; Yun, Matheson, 1992). The synthesis of both components depends on the transfer of glucose in a α -1,4 position from ADP-glucose to the non-reducing end of growing chains, and this key step is catalyzed by multiple forms of

*Corresponding author. E-mail: qiaoam@126.com or thdmast@gmail.com. Tel: +86-20-89003155. Fax: +86-20-89003155.

#These authors contributed equally to this work

starch synthase, in which the major granule-bound activity was called granule-bound starch synthase (GBSS) (Steven et al., 1998). GBSS (starch granule-bound ADP(UDP) glucose: α -1,4-D-glucan 4- α -glucosyl-transferase, EC 2.4.1.21), also denominated the waxy protein, is reported initially in dwarf beans, and subsequently described in potato, maize, wheat, cassava and other plants (Leloir et al., 1961; Echt and Schwartz, 1981; Yan et al., 2000). Although, GBSS catalyzes the elongation of both amylose and amylopectin *in vitro*, the mutations affecting GBSS genes in some plants resulted only in the decrease of amylose with the total amounts of starch remaining steady, so GBSS is considered as the key enzyme that determines the amylose content (Flipse et al., 1994). Amylose-free and low-amylose are significant for certain food and bioethanol industries.

Cassava (*Manihot esculenta* Crantz) is a root crop distributed in the tropical and subtropical areas as an important source of starch. With increasing demand of starch material in the modern industry, for example bio-fuel manufacture, the manipulation of properties and characteristics of cassava starch has become a focal issue in genetic breeding engineering (Xiao et al., 2009). The high-quality energy-type cassava depends on the reduced level of amylose, and herein GBSS is regarded as a useful regulation target. *gbss* gene from cassava (called *Megbss*) was cloned about ten years ago, however, little information is known about molecular structure and physico-chemical properties of *Megbss* gene and its encoding protein.

In present study, the bioinformatic analyses of *gbss* gene and its encoding protein from cassava were carried out. The results would lay theoretical foundation for molecular mechanism and genetic regulation researches of starch biosynthesis as well as the development and utilization of the correlative bioethanol resource plants in the future.

MATERIALS AND METHODS

The sequences with the complete coding regions (CDS) of *Megbss* was obtained from GenBank (Accession no.: X74160), and its corresponding amino acid sequence *Megbss* came from GenPept (Accession no.: CAA52273).

Comparative bioinformatic analyses of target sequences were performed online at the websites (<http://www.ncbi.nlm.nih.gov> and <http://www.expasy.org>). Molecular structures and physicochemical properties were obtained by ProtParam tools. Multiple alignment analysis, based on the full-length amino acid sequences, was performed with Vector NTI Suite 8 using default parameters (Lei et al., 2009). The subcellular location was predicted by TargetP 1.1 Server (Olof et al., 2000; Kristin and Siegfried, 2004). The cellular function, transmembrane helices and hydrophobicity in protein were analyzed by ProtFun 2.2 Server (Jensen et al., 2002; Jensen et al., 2003), TMHMM Server v.2.0 (Ikeda et al., 2002) and ProtScale (Kyte and Doolittle, 1982), respectively. The motifs and signal domains in protein were searched by PrositeScan (Combet et al., 2000) and Signal-HMM (Olof et al., 2007). Target proteins and their related sequences from other species were aligned with ClustalX (Thompson et al., 1997) and subsequently a phylogenetic tree was constructed by Neighbor-Joining method with 1000 replicates and

another tree was reconstructed by Maximum-Likelihood method with 1000 replicates, and reliability of each node was established by bootstrap methods using MEGA4.1 software, respectively (Saito and Nei, 1987; Kumar et al., 2001). The secondary structures of two UBGAT proteins were predicted by SIMPA96 online tool (Combet et al., 2000). And the homology-based three-dimensional (3D) structural modeling of UBGAT proteins was accomplished by Swiss-modeling (Guex and Peitsch, 1997; Schwede et al., 2003; Arnold et al., 2006; Benkert et al., 2011) and Accelrys ViewerLite 4.2 was used for 3D structure editing.

RESULTS AND DISCUSSION

Basal physicochemical properties of MeGBSS protein

Nucleotide acid sequence structure of *Megbss* gene was analyzed by Vector NTI Suite 8 software. The complete CDS possessed of typical coding structure including open reading frame (ORF), 5' untranslated region (UTR) and 3' UTR with poly (A) tail. Meanwhile, some physicochemical indices of *Megbss* protein were computed as follows: the formula was $C_{3013}H_{4745}N_{813}O_{861}S_{26}$, molecular weight 66968.3, isoelectric point (PI) 8.26, molar extinction coefficient 77185, estimated half-life 30 h *in vitro*, instability index 25.44, aliphatic index 90.15, grand average of hydropathicity (GRAVY) -0.097, and total number of negatively and positively charged residues was 65 and 68, respectively. Therefore, *Megbss* protein was classified as stable and polar one.

Analysis of subcellular localization and biochemical function

With the help of SignalP-HMM, TargetP 1.1 and TMHMM v2.0 online tools, MeGBSS protein was predicted to lack of transit peptide and signal domain (Figures 1 and 2). This suggested that *Megbss* was a non-secretory protein and lie in the cytoplasmic matrix without transmembrane topological structure, indicating the enzyme functioned and drove directly the flavonoid compounds biosynthesis in cytosol dispensing with transportation.

ProFun 2.2 Server analysis manifested the cellular function of *Megbss* protein may correlate to central intermediary metabolism, and this responds to the sub-cellular localization prediction, because central intermediary metabolite occurred usually within the cell.

As the specific functional element on the level of amino acid region, motif was treated as the focus target of protein structural biology. By the PrositeScan recognition, a series of patterns were found, including N-glycosylation site (27-30), cAMP- and cGMP-dependent protein kinase phosphorylation site (62-65), Protein kinase C phosphorylation site (568-570), Tyrosine kinase phosphorylation site (365-371), N-myristoylation site (106-111), Amidation site (300-303). After the multiple alignments by the software ClustalX, a phylogenetic tree was constructed in parallel with the Maximum Parsimony

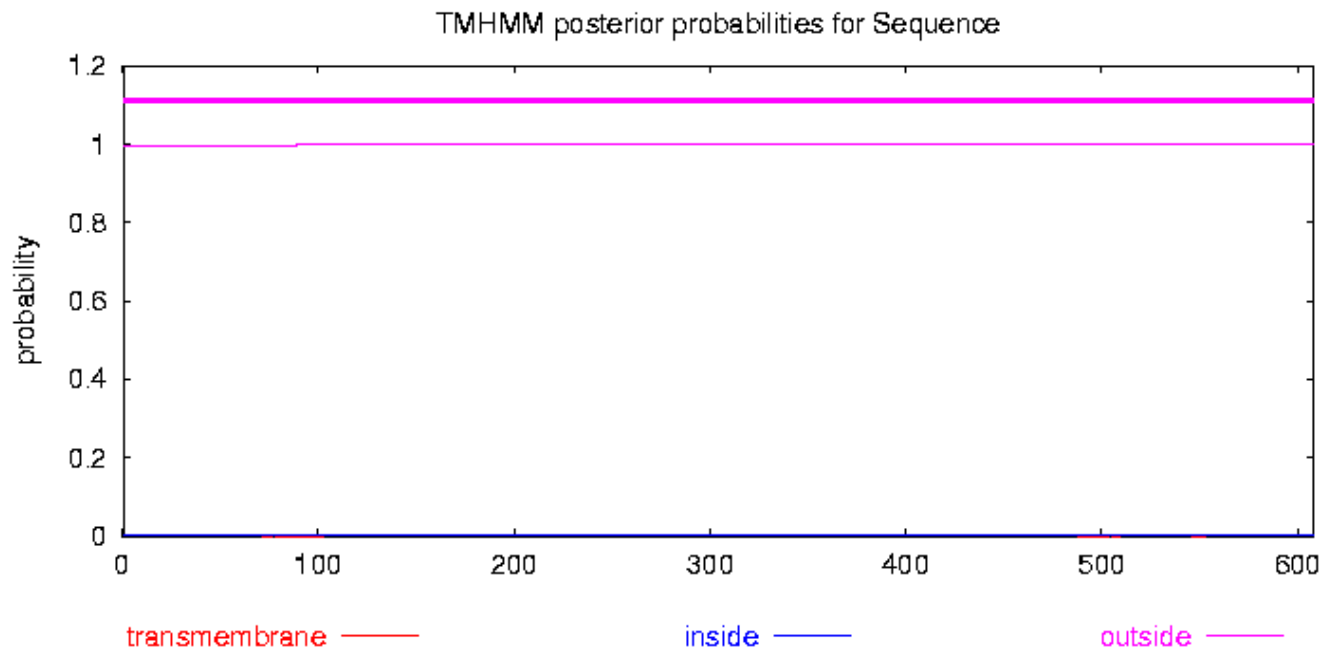


Figure 1. Prediction of transmembrane helices of MeGBSS protein.

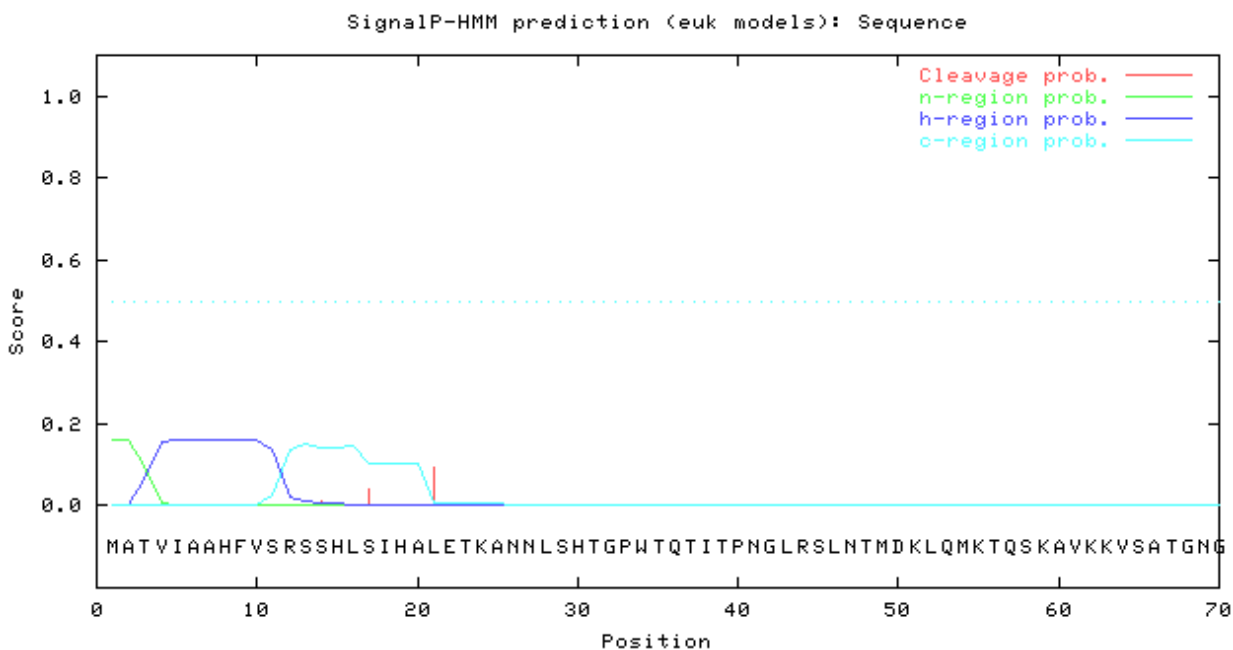


Figure 2. Prediction of signal domain of MeGBSS protein.

(MP) method (Figure 3). Cassava and some crops producing starch, such as *Ipomoea batatas*, *Oryza sativa*, *Sorghum bicolor* and *Triticum aestivum*, were clustered into one branch, implying GBSS protein have close relation with starch biosynthesis in plant kingdom and furthermore the gbss genes might have originated from

the same ancestor.

Establishment of secondary and tertiary structures

The secondary structure of MeGBSS protein was

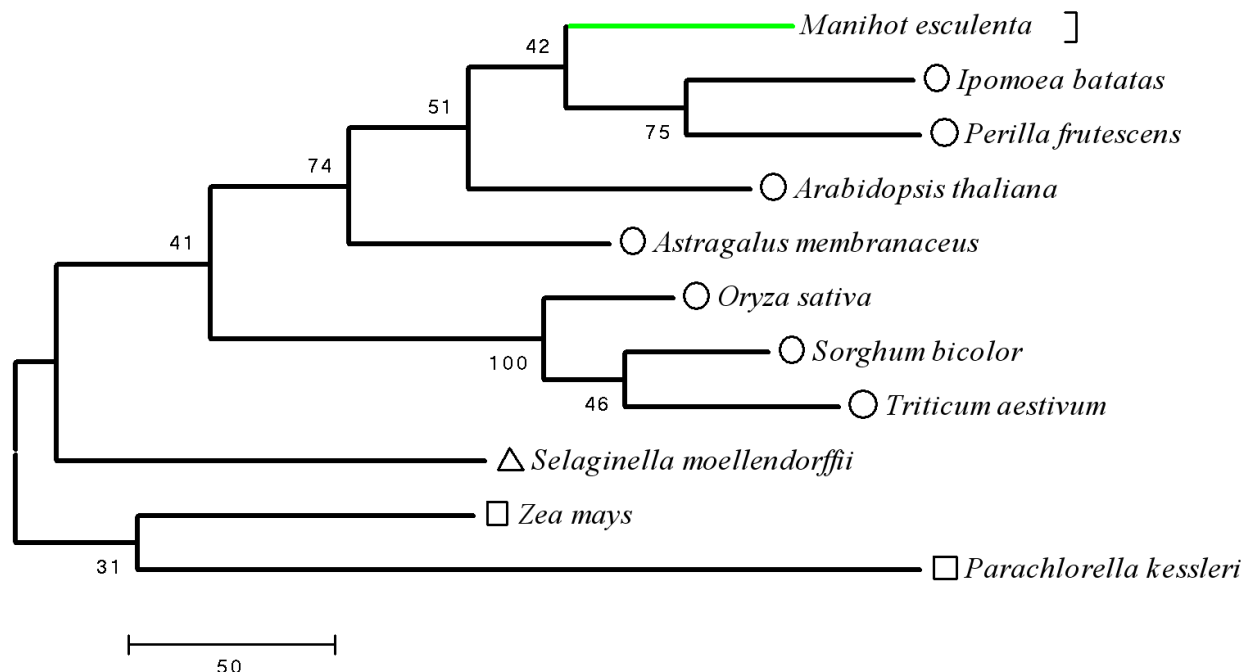


Figure 3. Phylogenetic tree analysis among MeGBSS and other GBSSs in plants. The Database accession number of the sequence used for the phylogenetic analysis: *Manihot esculenta* (GenPept accession NO.: CAA52273), *Zea mays* (GenPept accession NO.: NP_001106039), *Oryza sativa* (GenPept accession NO.: AAN77103), *Sorghum bicolor* (GenPept accession NO.: AAC49804), *Selaginella moellendorffii* (GenPept accession NO.: XP_002974509), *Ipomoea batatas* (GenPept accession NO.: AAA86423), *Arabidopsis thaliana* (GenPept accession NO.: AAM19783), *Triticum aestivum* (GenPept accession NO.: AAN03630), *Astragalus membranaceus* (GenPept accession NO.: AAC70779), *Perilla frutescens* (GenPept accession NO.: AAG43519), *Parachlorella kessleri* (GenPept accession NO.: AE79814).

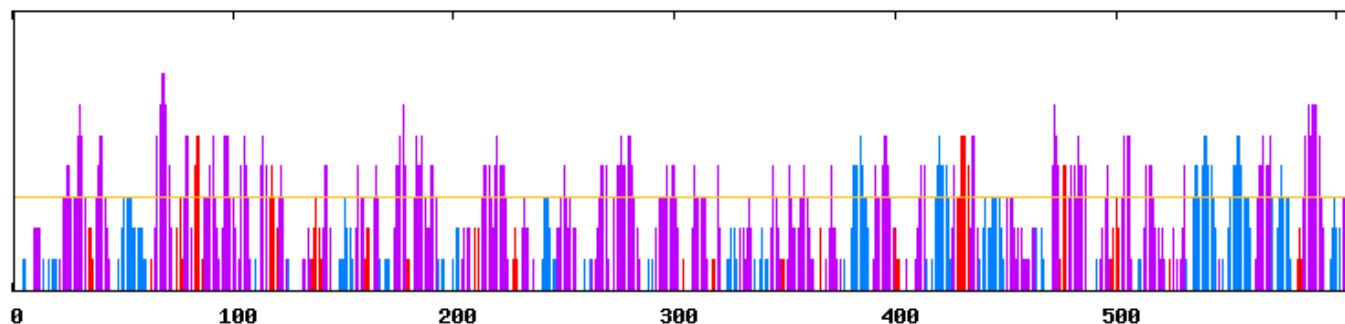





Figure 4. The secondary structure model of MeGBSS. The α -helix and extended strand were indicated as  and , respectively. Random coil was indicated as .

predicted by SIMPA96 online tool, thereinto, α -helix, extended strand and random coil shared 33.50, 13.63 and 52.71%, respectively (Figure 4).

Homology-based 3-D modeling of target proteins were implemented successfully using SWISS-MODEL (<http://swissmodel.expasy.org>) on the basis of the template from crystal structure of the glycogen synthase from *Agrobacterium tumefaciens* in complex with ADP, and the

substrate ADPG binding site was pointed (Figure 5) (Furukawa et al., 1990).

In order to select and identify the templates, two sensitive approaches were implemented: a profile blast and a hidden model-based template search. The profile for the query sequence and each model of the library were constructed from homologue series, which was chosen by iterative searches in the protein NMR

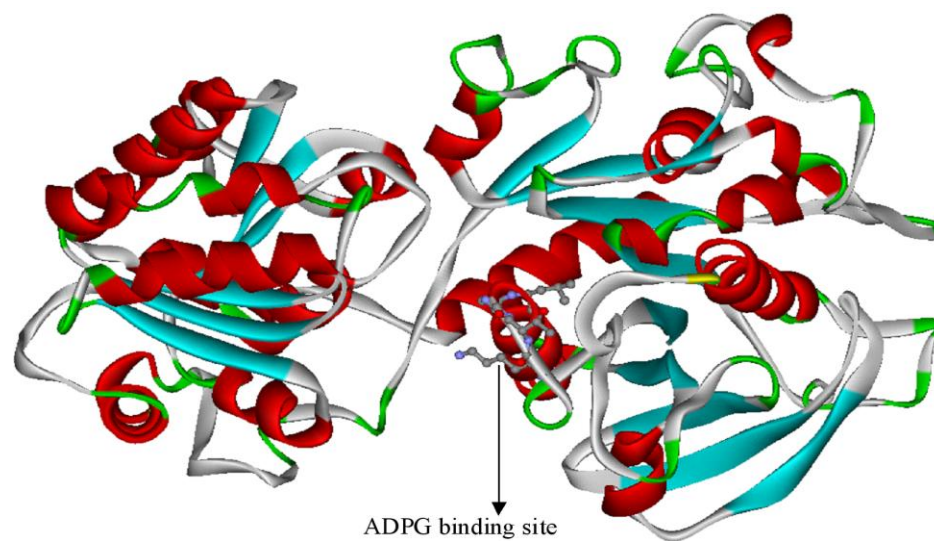
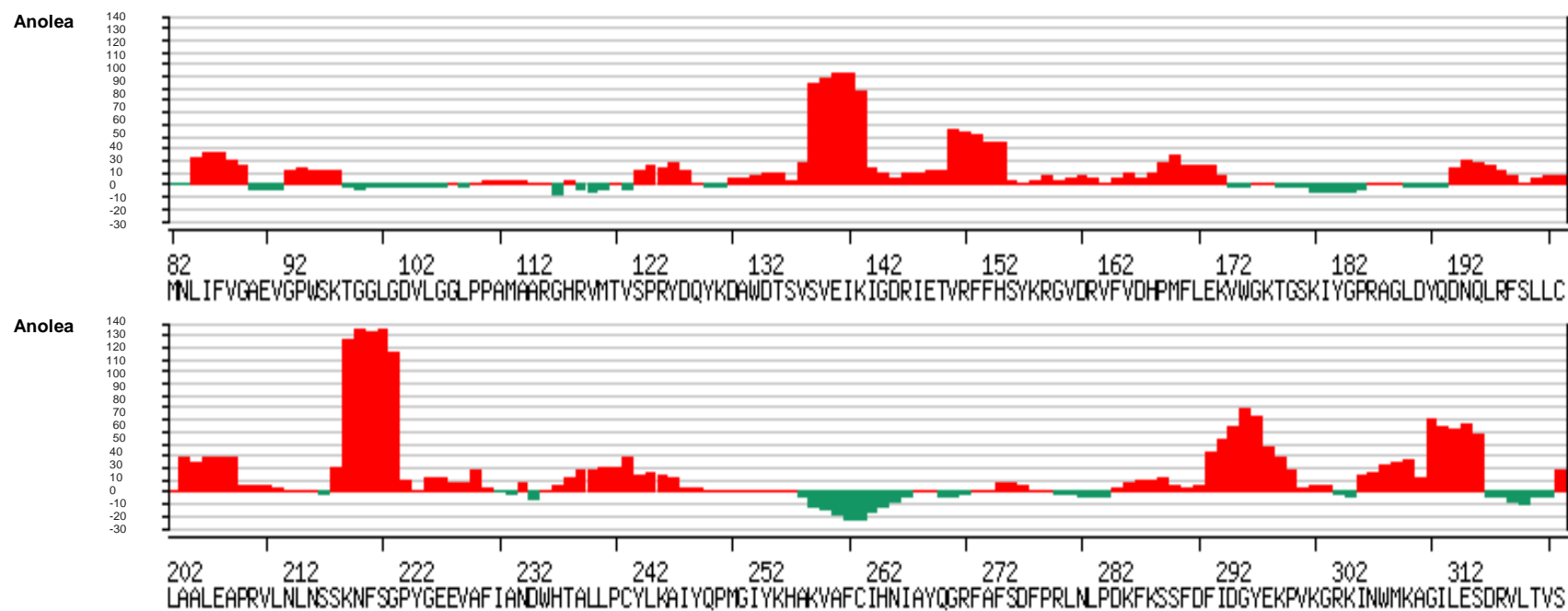


Figure 5. The 3D structural model of MeGBSS protein were established.



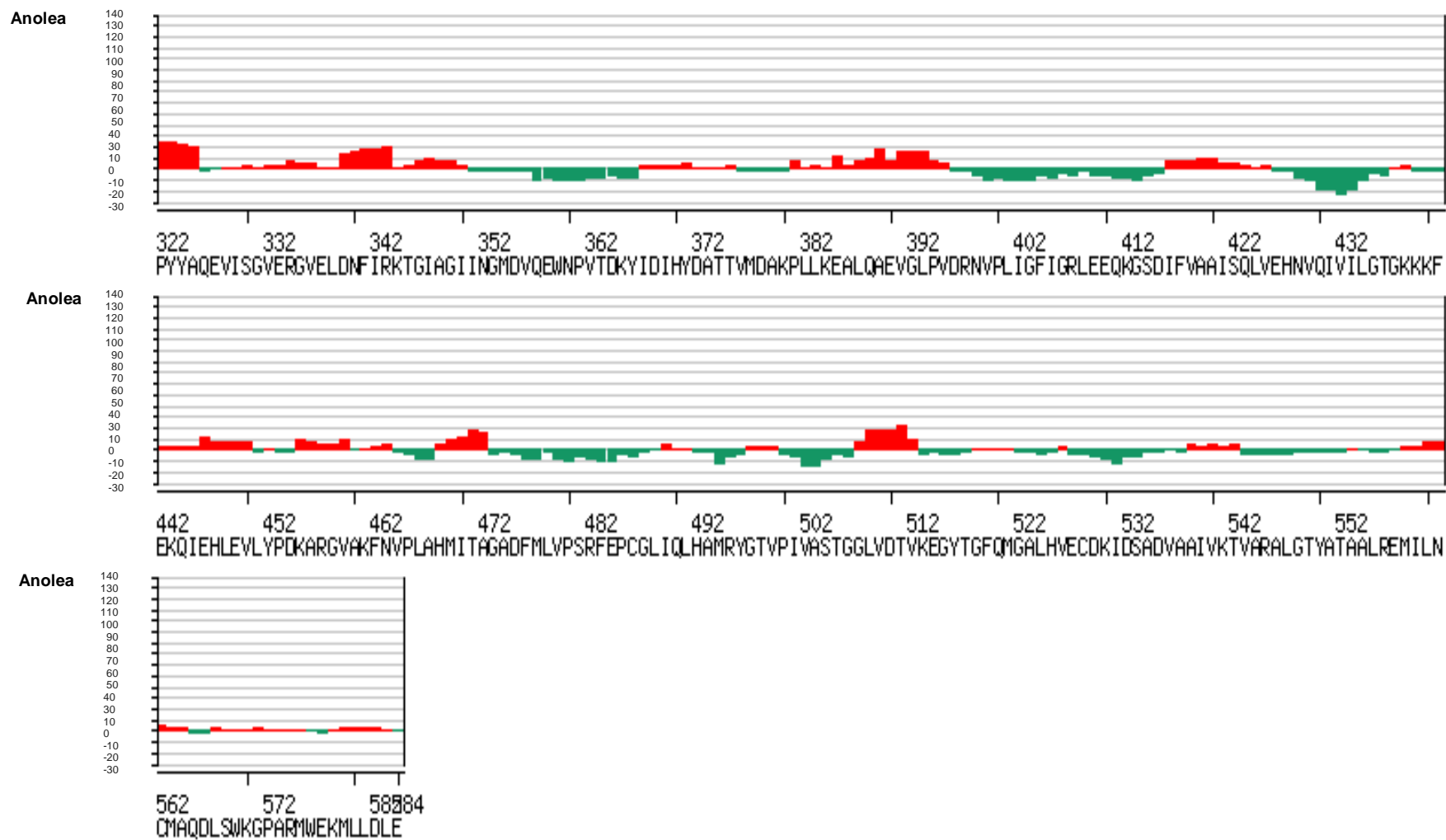


Figure 6. Local model quality estimation based on anolea method.

database, and then the target sequences were scored against the template HMM library for significant matches. Consequently, some segments with suitable template structures were

identified about MeGBSS amino acid sequences. Eventually, estimation of the models' quality was made by local one based on anole graph (Figure 6) and global one on QMEAN scores (Figure 7).

The starch source plant cassava is always used as the traditional food in some tropical areas and widely served the burgeoning biomass energy all over the world. Starch is caloric ingredient and

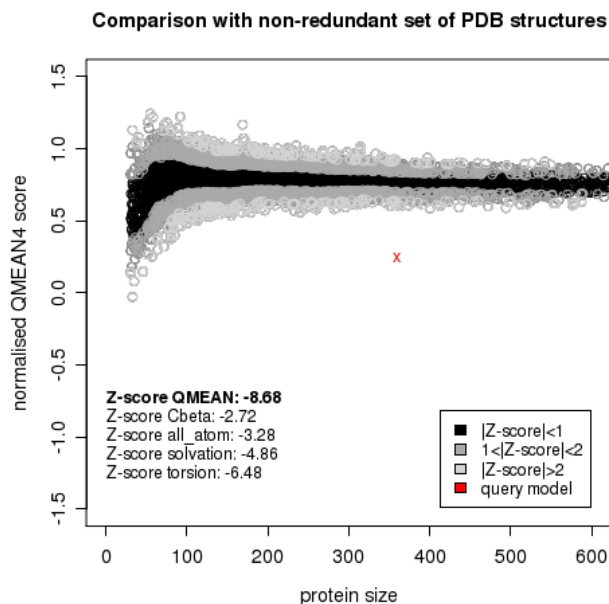


Figure 7. Global model quality estimation based on QMEAN4 method.

synthesized by a chain of enzymes, of which GBSS were considered as the crucial regulator. This study presented some important information about structural properties, biochemical function and expression profile of these genes and corresponding proteins by a series of computer softwares and databases. The results, such as 3-D modeling, functional motifs and systematic evolution and so on, revealed the initial molecular mechanism and reaction process which will be of significant use in providing important theoretical references for enzymology properties and genetic regulation researches of starch biosynthesis in cassava and development and utilization of its correlative biomass resource in the future.

REFERENCES

- Arnold K, Bordoli L, Kopp J, Schwede T (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modeling. *Bioinformatics* 22:195-201.
- Benkert P, Biasini M, Schwede T (2011). Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* 27:343-350.
- Combet C, Blanchet C, Geourjon C, Deléage G (2000). NPS@: Network protein sequence Analysis. *T.I.B.S.* 25:147-150.
- Echt GS, Schwartz D (1981). Evidence for the inclusion of controlling elements within the structural gene at the waxy locus in maize. *Genetics* 99:275-284.
- Flipse E, Huisman JG, De VBJ, Bergervoet JEM, Jacobsen E, Visser RGF (1994). Expression of a wild-type GBSS gene introduced into an amylose-free potato mutant by *Agrobacterium tumefaciens* and the inheritance of the inserts at the microsporoc level. *Theor. Appl. Genet.*, 88:369-375.
- Furukawa K, Tagaya M, Inouya M, Preiss J, Fukui T (1990). Identification of lysine 15 at the active site in *Escherichia coli* glycogen synthase. *J. Biol. Chem.* 265:2086-2090.
- Guex N, Peitsch MC (1997). SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* 18:2714-2723.
- Hao HY, Deng LK, Du JB (2009). Discussion on energy comprehensive utilization pattern in cassava fuel ethanol production process. *Cereal Food Ind.* 16:30-33.
- Ikeda M, Arai M, Lao DM (2002). Transmembrane topology prediction methods: A reassessment and improvement by a consensus method using a dataset of experimentally characterized transmembrane topologies. *In Silico Biol.* 2:19-33.
- Jane J, Shen JJ (1992). Internal structure of the potato starch granule revealed by chemical gelatinization. *Carbohydr. Res.* 247:279-290.
- Jensen JL, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Staerfeldt H, Rapacki K, Workman C (2002). Ab initio prediction of human orphan protein function from post-translational modifications and localization features. *J. Mol. Biol.* 319:1257-1265.
- Jensen JL, Staerfeldt HH, Brunak S (2003). Prediction of human protein function according to Gene Ontology categories. *Bioinformatics* 19:635-642.
- Kristin E, Siegfried H (2004). InvB is required for type III-dependent secretion of sopA in *Salmonella enterica* serovar typhimurium. *J. Bacteriol.* 186:1215-1219.
- Kumar SK, Tamamura K, Jakobsen IB, Nei M (2001). MEGA2: Molecular evolutionary genetics analysis software. *Bioinformatics* 17:1244-1245.
- Kyte J, Doolittle RF (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105-132.
- Lei W, Sun M, Luo KM, Shui XR, Sun YM, Tang SH (2009). Compute simulations to characterize chalcone synthase from *Scutellaria baicalensis* Georgi. *Mol. Boil.* 43:1012-1017.
- Leloir LF, De FMAR, Cardini CE (1961). Starch and oligosaccharide synthesis from uridine diphosphate glucose. *J. Biol. Chem.* 236: 636-641.
- Lin DG, Jeang CL (2005). Cloning, Expression, and Characterization of Soluble Starch Synthase I cDNA from Taro (*Colocasia esculenta* Var. *esculenta*). *J. Agric. Food Chem.* 53:7985-7990.
- Ling KH, Loo SS, Rosli R, Mariana NS, Mohamed R, Wan KL (2007). *In silico* identification and characterization of a putative phosphatidylinositol 4-phosphate 5-kinase (PIP5K) gene in *Eimeria tenella*. *In Silico Biol.* 7:115-121.
- Olof E, Henrik N, Søren B, Gunnar VH (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300:1005-1016.
- Olof E, Søren B, Gunnar VH, Henrik N (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* 2:953-971.
- Saito N, Nei M (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406-425.
- Schwede T, Kopp J, Guex N, Peitsch MC (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucl. Acids Res.* 31:3381-3385.
- Sivakumar K, Balaji SG (2007). *In silico* characterization of antifreeze proteins using computational tools and servers. *Proc. Indian Acad. Sci. Chem. Sci.* 119:571-579.
- Steven GB, Marion HBJ, Vande W, Richard GFV (1998). Progress in understanding the biosynthesis of amylose. *Trends Plant Sci.* 3:462-467.
- Thompson JD, Gibson TJ, Plewniaki F, Jeanmougin F, Higgins DG (1997). The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Res.* 25:4876-4882.
- Xiao Q, Guo YL, Kong H, He LK, Guo AP (2009). Cloning of Cassava *SBE I* Gene Fragment and Construction of its Storage Root Specific Antisense Expression Vectors. *Genomics Appl. Biol.* 28:255-261.
- Yan LL, Bhavne M, Fairclough R, Konik C, Rahman S, Appels R (2000). The genes encoding granule-bound starch synthases at the waxy loci of the A, B, and D progenitors of common wheat. *Genome* 43:264-272.
- Yun SH, Matheson NK (1992). Structural changes during development in the amylose and amylopectin fraction (separated by precipitation with concanavalin A) of starches from maize genotype. *Carbohydr. Res.* 270:85-101.