Probing Starch Biosynthesis Enzyme Isoforms by Visualization of Conserved Secondary Structure Patterns

Tayvich Vorapreeda[1], Weerayuth Kittichotirat[2], Asawin Meechai[3], Sakarindr Bhumiratana[1,3,4], Supapon Cheevadhanarak[4,5]

[1]*Biochemical Engineering and Pilot Plant Research and Development Unit, National Center for Genetic Engineering and Biotechnology & King Mongkut's University of Technology Thonburi, Bangkok, Thailand.*

[2]*Bioinformatics Program, King Mongkut's University of Technology Thonburi, Bangkok, Thailand.*

[3]*Department of Chemical Engineering, King Mongkut's University of Technology Thonburi, Bangkok, Thailand.*

[4]*Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi, Bangkok, Thailand.*

[5]*School of Bioresources and Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand.*

E-mail: supaponche@yahoo.com

ABSTRACT: Generally, enzymes in the starch biosynthesis pathway exist in many isoforms, contributing to the difficulties in the dissection of their specific roles in controlling starch properties. In this study, we present an algorithm as an alternative method to classify isoforms of starch biosynthesis enzymes based on their conserved secondary structures. Analysis of the predicted secondary structure of plant soluble starch synthase I (SSI) and soluble starch synthase II (SSII) demonstrates that these two classes of isoform can be reclassified into three subsets, SS-A, SS-B and SS-C, according to the differences in the secondary structure of the protein at C-terminus. SS-A reveals unique structural features that are conserved only in cereal plants, while those of SS-B are found in all plants and SS-C is restricted to barley. These findings enable us to increase the accuracy in the estimation of evolutionary distance between isoforms of starch synthases. Moreover, it facilitates the elucidation of correlations between the functions of each enzyme isoforms and the properties of starches. Our secondary structure analysis tool can be applicable to study the functions of other plant enzyme isoforms of economical importance.

1 INTRODUCION

Starch is the main carbohydrate reserve in crop plants that comprises of two major components, amylose and amylopectin. Although starch composes simply of glucose homopolymers but their structure is a complex, semicrystalline structure – the starch granule. Nowadays starch is an indispensable raw material in many industrials ranging from food industries to pharmaceuticals or agrochemicals industries. For instance, in agrochemical industries, starches are used for coating the plant seeds while in pharmaceutical industries, they are used in drug delivery. The utilization of starch mostly depends on its physical and chemical properties, which differ according to plant origins and their varieties, and the manufacturers often find them unsuitable for their industrial requirement. For example, rice starch is applicable for cosmetic uses owing to its very fine particles (average about 5 microns). However, it is reported to have problems with caking and stickiness. These problems are often handled by exposing the native form of starch to physical or chemical modification processes in ordered to tailor them appropriately according to the industrial needs. However, both physical and chemical modifications are cost intensive and generate high chemical waste. Moreover, large portion of starches are likely to be destroyed during the process. Thus, a major challenge would be to predict the effect of genetic changes on the functional properties of starch to produce designer starches that suit specific uses.

However, in order to perform this efficiently, a clear understanding about starch biosynthesis pathway is required. Despite the progress made by the scientists in analyzing and manipulating the mechanism of starch biosynthesis in plants, the roles of the key enzymes involved in this process are still not clearly understood. This is largely because the enzymes in starch biosynthesis pathway generally have a variety of isoforms that differ according to plant species, which increase the complexity of dissecting starch biosynthesis pathway. Because each form has distinct properties and plays a specific role in the synthesis of the starch polymers which are believed to affect starch properties, but their precise roles have not been identified [1-7]. For example, potato starch branching enzyme I (SBEI) was observed to be more active than starch branching enzyme II (SBEII) on an amylose substrate, whereas SBEII was observed to be more active than SBEI on an amylopectin substrate [8]. The existence of enzymes in multiple forms increases the complexity in understanding the starch biosynthesis process as well as the scope of a genetic engineer who try to manipulate them.

Currently, distinct classes of isoforms can be classified on the basis of similarity in amino acid sequence, molecular mass and antigenic properties [9-11]. With the current classification scheme, the relationship between different isoforms and the characteristics or the properties of starch produced as well as the starch biosynthesis process itself are still not yet clearly understood. One hypothesis is that the classification of isoforms based on

primary sequence may not be accurate [12-14], which Rost and Tian showed that enzymes function start to diverge quickly when the sequence identity below 70% [15, 16].

An alternative for classifying proteins (besides using primary sequence) is to classify them based on their structures [17, 18]. One very simple but yet strong reason is that structure is more highly conserved than primary sequence in evolution [19]. Classifying them by using correspondences in structures can therefore be an essential tool in the process of garnering information about both individuals and group of proteins in general. Besides that, structure classifying may also be able to detect remote homologies which might have been misclassified because of the divergence of amino acid sequence [20, 21]. However, from PDB databases [22], we found that there were no data about three-dimensional structure of key enzymes involve in starch biosynthesis pathway. Therefore, secondary structure level of enzyme is used to classify and characterize isoforms of enzyme in starch biosynthesis.

In recent studies, many groups of scientists have shown that secondary structures can be used to identify distantly related amino acid sequences. The explanation is that as the structures of related proteins are often more conserved than their sequences, this conservation also exists at the level of secondary structure [18, 23]. Przytycka and co-workers have constructed a protein taxonomy based on secondary structure, the results shown that the form of a tree in which proteins with similar secondary structure occupy neighboring leaves and it is largely in agreement with results from the structural classification of proteins (SCOP) [18]. In addition, Cid and co-workers have shown that comparison of predicted secondary structures can be used to reveal the presence of structural features that are conserved in one, the other or both families. These conservations can be the minimal sequence and structural features that may constitute the minimum catalytic unit of a group of related proteins [24]. Since the mechanisms by which distinct classes of isoform catalyze the same reaction and yet generate polymer variation are not fully understood, we present an alternative method for classifying and characterizing isoforms of starch biosynthesis enzymes based on predicted secondary structures may enable us to accurately estimate evolutionary distances between enzyme isoforms, specific functions or the effects of each enzyme isoforms on the starch properties.

2 METHODS

We have developed a program that can find out features from the secondary structure of starch synthase isoforms data input based on their similarity, we called this program "Secs_Miner". During each iteration, a Smith-Waterman dynamic programming alignment algorithm [25] together with our scoring model is used to calculate all pairwise similarity of secondary structure dataset. The highest score pair is then used to construct a representing pattern where it will be further considered in the successive rounds of the algorithm, instead of using its parental secondary structures (Fig. 1).

2.1 Scoring model

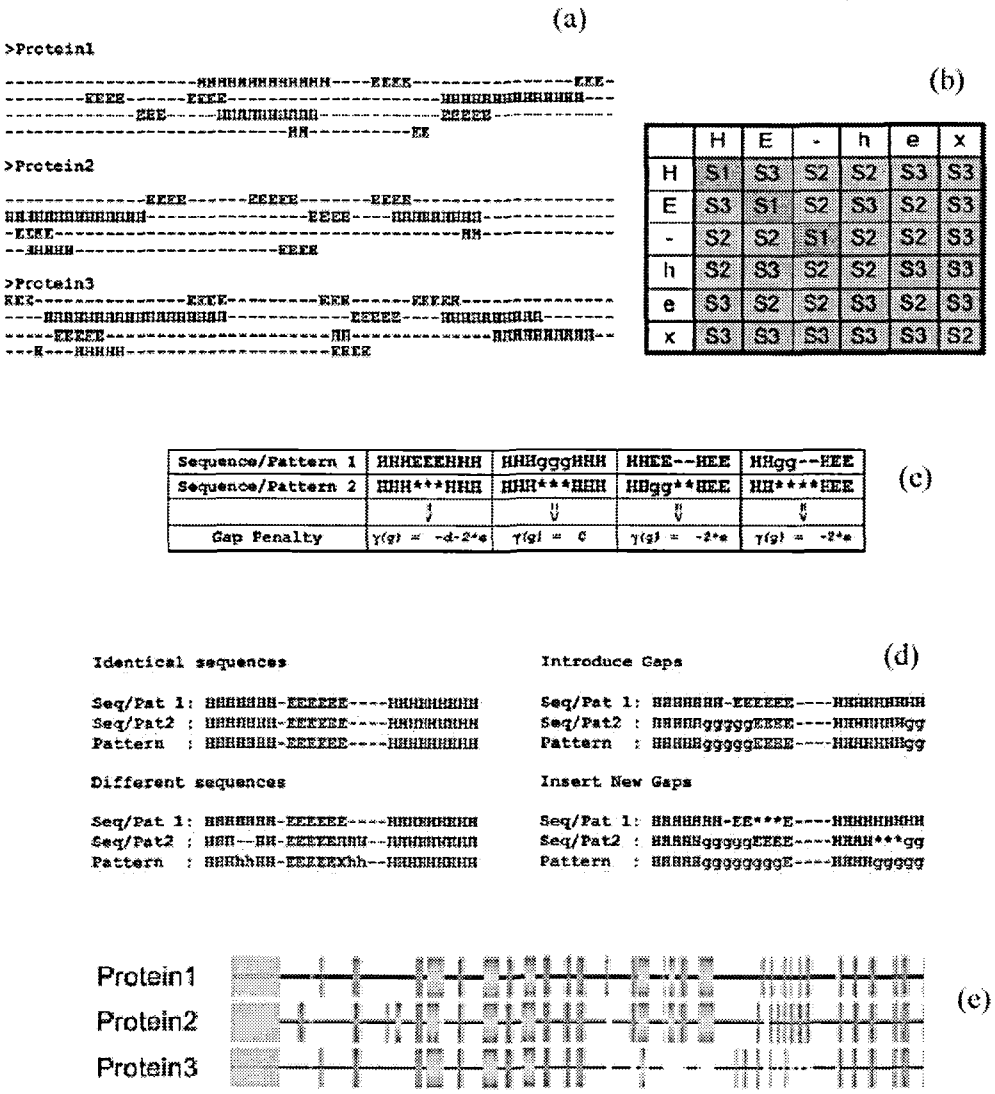
The protein secondary structure element to element substitution matrix was constructed for each type of secondary structure element composing of the 3 types of secondary structure residue (helices (H), extended strands (E), and interconnecting loops (-)) plus 3 secondary structure element classes (partial helices (h), partial extended strands (e), any residue (X)) where *S1*, *S2* and *S3* represent perfect match score, partial match score and mismatch score respectively (Fig. 1b). The values of all substitution matrix is assigned *S1* = 4, *S2* =1 and *S3* = 0. This scoring is built on the assumption that enzyme isoforms catalyze the same reaction or in another word, they have the same function. Therefore, most of the secondary structure elements are expected to be identical and we should more likely see large conservation substitutions in the alignment results and that this conservation is not likely to be observed by chance. As a result, the perfect match score (*S1*) is assigned with highest positive value (4 in this case). On the other hand, the less conservative substitutions (partial match score *S2*) are expected to be observed less frequently in the alignment results and therefore are assigned with smaller positive value (1 in this case). Lastly for non-conservative changes (mismatch score *S3*), we should expect to see this kind of substitution in real alignment very seldom and therefore is assigned with zero score term.

2.2 Gap penalties

Generally, the gap penalties used in this study is as shown in Eqs. 1.

$$\gamma(g) = -d-(g-1)e \quad (1)$$

Where $\gamma(g)$ represents gap penalty, which is the negative summation of the gap-open penalty (d) and the gap-extension penalty (e). In this study, the value of gap-open and gap-extension penalties are set to -2 and -1 respectively (Fig. 1c).



	H	E	-	h	e	x
H	S1	S3	S2	S2	S3	S3
E	S3	S1	S2	S3	S2	S3
-	S2	S2	S1	S2	S2	S3
h	S2	S3	S2	S2	S3	S3
e	S3	S2	S2	S3	S2	S3
x	S3	S3	S3	S3	S3	S2

Sequence/Pattern 1	HHHEEEEHHH	HHHgggHHH	HHEE--HEE	HHgg--EEE
Sequence/Pattern 2	HHH***HHH	HHH***HHH	HHgg**HEE	HH****EEE
	⇓	⇓	⇓	⇓
Gap Penalty	$\gamma(g) = -d-2*e$	$\gamma(g) = 0$	$\gamma(g) = -2*e$	$\gamma(g) = -2*e$

Figure 1: Step by step illustration of Secs_Miner program.
(a) Input data to be aligned showing secondary structures sequences where 'H' represents helices, 'E' represents extended strands or beta-sheets and '-' represents interconnecting loops. (b)Secondary structure element to element substitution matrix that was used to calculate the alignment where S1 represents perfect match score (4 points), S2 represents partial match score (1 point) and S3 represents mismatch score (0 point). (c) Gap penalties that will be issued to different types of alignment. (d) A representing pattern that is generated when any secondary structure element is aligned with identical secondary structure element and different secondary structure element that gaps and new gaps were previously introduced, respectively. (e) The resulting refined alignments for the three protein secondary structures.

2.3 Representing pattern construction

A representing pattern will be generated from the most similar pair of protein secondary structures based on the optimal pairwise alignment result and protein secondary structure class hierarchy. When any protein secondary structure element is aligned with identical element, that protein secondary structure element is placed in the representing pattern at the equivalent position. If secondary structure element is aligned with a different protein secondary structure element or class, the minimally inclusive class covering both aligned elements is placed at that position. If any character is aligned with a previously introduced gap (g) or a new gap (*), a gap character (g) is placed or inserted in the representing pattern at that position, respectively (Fig. 1d).

3 RESULTS & DISCUSSIONS

Starch synthase sequences were compiled from GenBank and PlantGDB [26] on the basis of BLAST analysis. The result represents 20 sequence-related proteins for each enzyme (we focused on starch synthase I (SSI) and starch synthase II (SSII)) from selected plants. Those of the dicotyledonous group were Arabidopsis (*A. thaliana)* and potato (*Solanum tuberosum*). The epitomes of the

monocotyledonous plant were sorghum (*Sorghum bicolor*), maize (*Zea mays*), wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), and rice (*Oryza sativa*). At the same time, these plants were also the models of starch biosynthesis in various organelles, such as Arabidopsis for the synthesis in leaves, potatoes and ceareal plants for those in tubers, and endosperms, respectively.

3.1 Predicted secondary structure features of starch synthases

The comparison of predicted secondary structure of plants SSI and SSII shows that these two classes of isoform can be reclassified into three groups based on secondary structures features (in this work, we name them "SS-A, SS-B and SS-C"), leading to the discovery of structural conservation and plant-specific features. SS-A displays three regions, I, II and III, that are conserved in cereal plants. While SS-B displays four regions, I, II, III and IV, that present in all plants, SS-C is found only in Barley that exhibits three conserved regions I, II, and IV (Fig. 2). This analysis revealed distinct secondary structure features that can be used as a structural marker to distinguish starch synthase of cereal plants from non-cereal ones.

The N-terminus is not conserved in predicted secondary structure for all three groups of protein, implicating that the N-terminus may not be essential for catalytic unit or catalytic activity of SSI and SSII. This finding is in agreement with the result from the study of maize starch synthase in that when the N-terminal of maize SSIIa and *SSIIb* were truncated, they do not affect the catalytic activity and kinetic properties.13]. In addition, in the case of potato starch synthase, it has been shown that the N-terminal arm of SSII did not determine the specific activity of this isoform, as removal of STAG6 did not significantly alter specific activity compared with the full-length SSII [2], again suggesting that the N-terminus is not catalytic unit for starch synthase.

Region I is displayed on three groups of protein and the predictions suggest the presence of an α-helix and three β-strands, and the intervening sequence contains a lys-ser(thr)-gly-gly consensus sequence for the ADP-glucose binding site for starch synthase [27]. This region might be an acquired feature of the starch synthase, in order to meet substrate requirements. Region II is directly linked to region I, which is the last region that is very similar for the three enzymes and the predictions suggest the presence of a β-strand followed by an α-helix, β-strand and α-helix, respectively. Region III and Region IV constitute the C-terminus of the two proteins, which region III and IV are absent in the SS-C and SS-A, respectively (Fig. 2). Each of the three classes of protein analyzed in this study has different features at the C-terminus.. Our results suggest that, C-terminus of starch synthase are more essential for specific catalytic activity and function than N-terminus. In the case of maize, it has been shown that, the C-terminal 450 residues of DU1 from maize when expressed in *Escherichia coli* showed to possess SS activity [28, 29]. Edwards suggested that the comparison sequence between granule bound starch synthase I (GBSSI) and soluble starch synthase II (SSII) from potato can be used to reveal the C-terminal region of GBSSI conferring most of the specific properties of this isoform, except its processive elongation of glucan chains. This region of GBSSI is distinct from the C-terminal region of other starch synthases [2]. Furthermore, a study of maize SSI glucan-binding affinities indicates that the entire C-terminal region of this enzyme is required for starch binding, while the N-terminus is not [5].. The predicted secondary structures of C-terminus of starch synthase shows difference of enzyme isoforms that should be mentioned that starch synthase, SS-A, is the unique isoform that exists in the starch biosynthesis process of the cereal plants and not present in non-cereal plant.

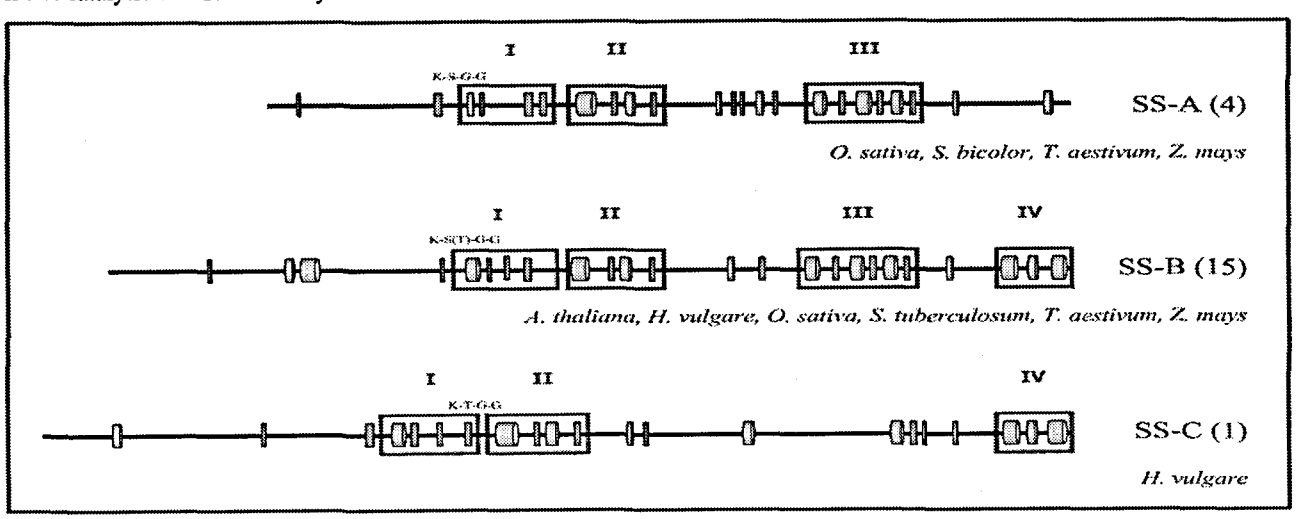


Figure 2: Alignment of predicted secondary structures of starch synthases from different plants, soluble starch synthase I and II can be divided into 3 subsets consist of SS-A, SS-B and SS-C, according to different in secondary structure. Thick barrels represent α-helices and the thin ones represent predicted β-sheets. The position of ADP-glucose binding sites are shown in region I. The numbers between brackets show the number of starch synthase tested.

3.2 Hypothesis

Our results indicate that the main difference between the three groups of starch synthase is the exhibit in the secondary structure of proteins between region III and IV at the C-terminus . The results from structure analysis propose that the SS-A presents in the starch biosynthesis pathway in the cereal plant and are absent in others. Previous studies suggested that the pathway of starch synthesis in the cereal endosperm is unique, and requires enzyme isoforms that are not present in other cereal tissues or non-cereal plants. Individual starch synthase isoforms are believed to have unique functions in starch synthesis but their precise roles have not been identified [7]. In addition, the number of chains per branched amylose molecule varies and is usually lower for cereal starches than those of non-cereal starches (e.g. branched corn amylose has an average of 5.3 chains whilst that from tapioca has 17.1) [30]. From the predicted secondary structure results, together with other information, we propose that soluble starch synthase, SS-A, may be associated with the construction of amylose which involved in the mechanism for the biosynthesis of branched amylose chains (as shown in Fig. 3). Divergence of the predicted secondary structure at C-terminus of SS-A, which is assumed to involve in the catalytic activity and binding affinities that plays a specific role in the transfer of glucose in an α-1,4 position from ADP-glucose to the non-reducing end of growing chains. Our study suggests that the genetic material for the synthesis of cereal starches differs from non-cereal starches.

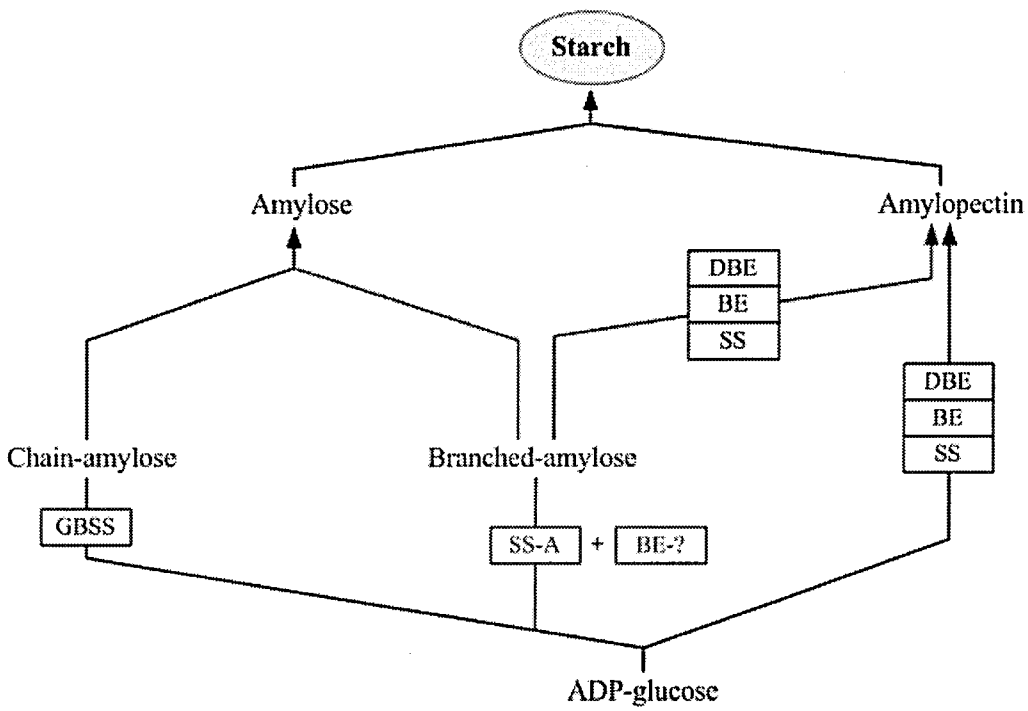


Figure 3: Proposed Pathway of starch biosynthesis in cereals predicted from secondary structure of starch synthases. The SS-A and branching enzyme working in collaborations are involved in the biosynthesis of branched amylose chains. Branched amylose is catalyzed the elongation of α-1,4-glucosidic and branched α-1,6-glucosidic bonds on amylopectin by other enzymes that involved in starch biosynthesis pathway such as soluble starch synthases, branching enzymes and debranching enzymes.

4 CONCLUSIONS

In this study, we describe a possibility of gaining more understanding about various enzyme isoforms in starch biosynthesis process by *ab initio* re-classification of isoforms that catalyze the same reaction into groups that have high residue correspondences to their predicted secondary structures. Since structure is more conserved than sequence in evolution and also closer to function, isoform members of this *ab initio* reclassification groups may process similar or identical chemical functions. As such we may be able to assign the functions of well-characterized isoform members of that group to new isoforms members whose specific functions are not known or not well understood. We also describe the possibility of analyzing common secondary structural pattern of related enzyme isoforms to visualize the important structural motifs of each group of structurally related enzyme isoforms. This may be useful in visualizing the important features of the enzymes such as the minimal catalytic units that have been conserved in evolution. On the secondary structure comparison analysis of SSI & SSII, we have found that the genetic materials for the synthesis of cereal starches differ from non-cereal starches. The information gained can be useful as a

guideline for plant molecular biologists who try to manipulate the starch biosynthesis process. Moreover, this *ab initio* classification can be applicable to study the enzyme isoforms of other economical plants of which their isoforms are hard to be differentiated.

REFERENCES

[1] A. Edwards, et al. Specificity of starch synthase isoforms from potato. European Journal of Biochemistry, 266: 724--736, 1999.

[3] M. Peng, M. Gao, M. Baga, P. Hucl, and R. N. Chibbar. Starch branching enzymes preferentially associated with A-type starch granules in wheat endosperm. Plant Physiology, 124(1): 265--272, 2000.

[4] P. L. Vrinten, and T. Nakamura. Wheat granule-bound starch synthase I and II are encoded by separate genes that are expressed in different tissues. Plant Physiology, 122(1): 255--264, 2000.

[5] P. D. Commuri, and P. L. Keeling. Chain-length specificities of maize starch synthase I enzyme: studies of glucan affinity and catalytic properties. The Plant Journal, 25(5): 475--486, 2001.

[6] A. M. Smith. The biosynthesis of starch granule. Biomarcomolecules, 2: 335--341, 2001.

[7] M. G. James, K. Denyer, and A. M. Myers. Starch synthesis in the cereal endosperm. Current Opinion in Plant Biology, 6: 215--222, 2003.

[8] U. Rydberg, L. Andersson, R. Andersson, P. Aman, and H. Larsson. Comparison of starch branching enzyme I and II from potato. European Journal of Biochemistry, 628: 6140--6145, 2001.

[9] J. Marshall, et al. Identification of the major starch synthase in the soluble fraction of potato tubers. The Plant Cell, 8: 1121--1135, 1996.

[10] M. E. Knight, et al. Molecular cloning of starch synthase I from maize (W64) endosperm and expression in *Escherichia coli*. The Plant Journal, 14(5): 613--622, 1998.

[11] T. Senoura, et al. Characterization of starch synthase I and II expressed in early developing seeds of kidney bean (*Phaseolus vulgaris* L.). Bioscience, Biotechnology and Biochemistry, 68(9): 1949--1960, 2004.

[12] A. Edwards, et al. Biochemical and Molecular Characterization of a novel starch synthase from potato tubers. The Plant Journal, 8(2): 283—294, 1995.

[13] J. M. Imparl-Radosevich, D. J. Nichols, P. Li, A. L. McLean, P. L. Keeling, and H. P. Guan. Analysis of purified starch synthases IIa and IIb: SS isoforms can be distinguished based on their kinetic properties. Archives of Biochemistry and Biophysics, 362: 131--138, 1999.

[14] A. Edwards, et al. Discrete form of amylose are synthesized by isoforms of GBSS in pea, The Plant Cell, 14: 1767--1785, 2002.

[15] B. Rost. Review: Protein secondary structure prediction continues to rise. Journal of Structural Biology, 134: 204--218, 1999.

[16] W. Tian, and J. Skolnick. How well is enzyme function conserved as a function of pairwise sequence identity?. Journal of Molecular Biology, 333: 863--882, 2003.

[17] J. F. Gibrat, T. Madej, and S. H. Bryant. Surprising similarities in structure comparison. Current Opinion in Structural Biology, 6: 377--385, 1996.

[18] T. Przytycka, R. Aurora, and G. D. Rose. A protein taxonomy based on secondary structure. Structural Biology, 6(7): 672--682, 1999.

[19] C. Chotia, and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. EMBO Journal, 5: 823--826, 1986.

[20] C. L. Tang, L. Xie, I. Y. Koh, S. Posy, E. Alexov, and B. Honig. On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. Journal of Molecular Biology, 334: 1043--1062, 2003.

[21] S. Goldsmith-Fischman, and B. Honig, Structural genomics: computational methods for structure analysis. Protein Science, 12: 1813--1821, 2003.

[22] D. Kihara and J. Skolnick. The PDB is a covering set of small protein structures. Journal of Molecular Biology, 334: 793--802, 2003.

[23] C. Geourjon, C. Combet, C. Blanchet, and G. Deleage. Identification of related proteins with weak sequence identity using secondary structure prediction. Protein Science, 10: 788--797, 2001.

[24] E. Cid, R. A. Geremia, A. A. Guinovart, and J. C. Ferrer. Glycogen synthase: towards a minimal catalytic unit?. FEBS Letters, 528: 5--11, 2002.

[25] T. F. Smith, and M. S. Waterman. Identification of common molecular subsequences. Journal of Molecular Biology, 147: 195--197, 1981.

[26] Q. Dong, S. D. Schlueter, and V. Brendel. PlantGDB, plant genome database and analysis tools. Nucleic Acid Research, 32: D354--D359, 2004.

[27] T. Baba, et al. Identification, cDNA cloning, and gene expression of soluble starch synthase in rice (*Oryza sativa* L.) immature seeds. Plant Physiology, 103: 565--573, 1993.

[28] H. Cao, et al. Identification of the soluble starch synthase activities of maize endosperm. Plant Physiology, 120: 205--215, 1999.

[29] M. Gao, et al.. Characterization of *dull1*, a maize gene coding for a novel starch synthase. The Plant Cell, 10: 399--412, 1998.

[30] Y. Takeda, S. Hizukuri, C. Takeda, and A. Suzuki. Structures of branched molecules of amyloses of various origins, and molar fractions of branched and unbranched molecules. Carbohydrate Research, 165(1): 139--145, 1987.