UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
INSTITUTO DE CIÊNCIAS E TECNOLOGIA DAS ÁGUAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIODIVERSIDADE

ELVIS SANTOS LEONARDO

EVOLUÇÃO MOLECULAR E ANÁLISE ESTRUTURAL DA ENZIMA *GRANULE-BOUND STARCH SYNTHASE* (GBSS) EM ANGIOSPERMAS

SANTARÉM – PA

2023

UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
INSTITUTO DE CIÊNCIAS E TECNOLOGIA DAS ÁGUAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIODIVERSIDADE

ELVIS SANTOS LEONARDO

EVOLUÇÃO MOLECULAR E ANÁLISE ESTRUTURAL DA ENZIMA *GRANULE-BOUND STARCH SYNTHASE* (GBSS) EM ANGIOSPERMAS

Dissertação apresentada ao Programa de Pós-graduação em Biodiversidade da Universidade Federal do Oeste do Pará, com requisito para obtenção de grau de Mestre em Biodiversidade.

Orientador: Prof. Dr. Thiago Jose de Carvalho André
Universidade Federal de Brasília
Coorientador: Prof. Dr. Kauê Santana da Costa
Universidade Federal do Oeste do Pará

SANTARÉM – PA

2023

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema Integrado Bibliotecas – SIBI/UFOPA

L581e Leonardo, Elvis Santos

Evolução molecular e análise estrutural da enzima Granule-Bound Starch Synthase (GBSS) em angiospermas / Elvis Santos Leonardo – Santarém, 2023.

40 f.: il.

Orientadora: Thiago Jose de Carvalho André

Coorientador: Kauê Santana da Costa

Dissertação (Mestrado) – Universidade Federal do Oeste do Pará, Pró–reitoria de Pesquisa, Pós Graduação e Inovação Tecnológica, Instituto de Ciências e Tecnologia das Águas, Programa de Pós-Graduação em Biodiversidade.

1. Abordagens computacionais. 2. Angiospermas. 3. História evolutiva. 4. Síntese de amido. I. André, Thiago Jose de Carvalho, *orient*. II. Costa, KauÍ Santana da. III.Título.

CDD: 23 ed. 572.838

Bibliotecário-Documentalista: Ronne Clayton de Castro Gonçalves – CRB-2/1410

ATA Nº 26

Em acordo com o Regimento do Programa de Pós Graduação em Biodiversidade da Universidade Federal do Oeste do Pará, a dissertação de mestrado é julgada por uma Banca Avaliadora não presencial, constituída por cinco avaliadores, sendo um deles obrigatoriamente externo ao curso, com título de doutor (Artigo 56 do referido regimento). O acadêmico é considerado aprovado quando ao menos três membros avaliadores emitirem pareceres aprovado. Alternativamente, o discente será dispensado da banca avaliação da dissertação, quando comprovar o aceite ou publicação de pelo menos um artigo resultante da sua dissertação, como primeiro autor, em co-autoria com orientador, ou orientador e coorientador quando o orientador for um docente colaborador, em periódico indexado com percentil mínimo de 75 (setenta e cinco) ou superior referente às métricas mais recentes do maior percentil utilizado pelo Journal Citation Reports (Clarivate) ou pelo Scientific Journal Rankings (Scimago), cabendo ao discente apenas a apresentação pública do trabalho (Artigo 58). O discente que teve sua dissertação aprovada deverá apresentá-la em sessão pública com duração de até 50 (cinquenta) minutos obrigatoriamente até no máximo 15 (quinze) dias após a aprovação, e no prazo máximo de vínculo com o curso, ou seja, 24 (vinte e quatro) meses após o início do primeiro semestre letivo do discente no curso (artigo 64). Assim, aos oito dias do mês de dezembro do ano de dois mil e vinte e dois, às quatorze horas, de forma remota através da plataforma GoogleMeet, instalou-se a apresentação de seminário público da dissertação de mestrado do aluno ELVIS SANTOS LEONARDO. Deu-se início a abertura dos trabalhos, onde o Professor Dr. THIAGO JOSE DE CARVALHO ANDRE, após esclarecer as normativas de tramitação da defesa e seminário público, de imediato solicitou a candidata que iniciasse a apresentação da dissertação, intitulada "EVOLUÇÃO MOLECULAR E ANÁLISE ESTRUTURAL DA ENZIMA GRANULE-BOUND STARCH SYNTHASE (GBSS) EM ANGIOSPERMAS". Concluída a exposição, o professor comunicou o discente que a versão final da dissertação deverá ser entregue ao programa, no prazo de 60 dias; contendo as modificações sugeridas pela banca examinadora e constante nos formulários de avaliação da banca. A banca examinadora foi composta pelos examinadores professores doutores listados abaixo. Os pareceres assinados seguem em sequência.

Thiago Jose de Carvalho André
Orientador

Elvis Santos Leonardo
Discente

Dr. JOÃO PAULO MATOS SANTOS LIMA, UFRN

Examinador Externo à Instituição

Dr. GABRIEL IKETANI COELHO, UFOPA

Examinador Externo ao Programa

Dra. THAIS ELIAS ALMEIDA, UFOPA

Examinadora Interna

ELVIS SANTOS LEONARDO

Mestrando

AGRADECIMENTOS

Em primeiro lugar agradeço a Deus por permitir trilhar esse novo caminho da minha vida acadêmica, tudo isso não seria possível sem ele. Este Trabalho é dedicado a você, familiar ou amigo que contribuiu muito na minha caminhada. Sem vocês eu não conseguiria.

Agradeço ao meu orientador, Thiago Jose de Carvalho André por ter me dado a oportunidade de trabalhar nesse projeto, e ao meu coorientador Kauê Santana da Costa por auxiliar no projeto, sua ajuda foi fundamental para a finalização do trabalho, agradeço todos os ensinamentos repassados de ambas as áreas de atuação, a junção de conhecimento que vocês me proporcionaram foi extraordinário.

Agradeço ao programa de Pós-graduação em Biodiversidade e a Universidade Federal do Oeste do Pará pela oportunidade e suporte para realização desta qualificação profissional. Agradeço à Coordenação de Aperfeiçoamento Pessoal de Nível Superior (CAPES) pelo fornecimento da bolsa de mestrado e ao Programa de Apoio ao Desenvolvimento Acadêmico pelos recursos que auxiliaram a execução das atividades. Ao núcleo de professores que contribuíram durante todo o decorrer do curso, o conhecimento agregado foi essencial para minha formação.

“Uma vez que as proteínas participam de um jeito ou de outro em todos os processos químicos no organismo vivo, deve-se esperar informações altamente significativas para a química biológica a partir da elucidação de sua estrutura e de suas alterações.”

Emil Fischer

RESUMO

O principal carboidrato de armazenamento em plantas é o amido, um recurso natural de importância global para alimentação que faz parte de alimentos para humanos, ração para animais de criação, além de matéria prima para diversas indústrias. No entanto, não há muitos estudos sobre a enzima que sintetiza este carboidrato, tanto em relação sobre sua estrutura e função como sua história evolutiva no contexto geral das angiospermas. Este estudo utiliza uma análise filogenética abrangente do gene Waxy que codificam a enzima GBSS (*Granule-Bound Starch Synthse*) para compreender melhor sua história evolutiva. Além disso, a estrutura proteica contém duas isoformas. Para determinar as estruturas foi utilizado modelagem por homologia, além de alinhamento estrutural para verificar regiões conservadas. A análise filogenética mostrou que o gene que codifica a enzima GBSS é bastante conservado em angiospermas, separando bem grandes grupos filogenéticos como monocotiledôneas e dicotiledôneas. A enzima contém duas isoformas, GBSSI e GBSSII e essa diversificação pode estar relacionada a processos de duplicação nos genomas de plantas. Apesar da estrutura geral das isoformas ser similar, há variação estrutural na sequência de aminoácidos. Este estudo fornece uma compreensão abrangente da história evolutiva e estrutural da principal enzima envolvida na síntese do amido e esses resultados devem apoiar estudos futuros que visam aumentar a compreensão da biossíntese de amido e a divergência evolutiva e funcional da GBSS em plantas.

Palavras-Chave: Abordagens computacionais. Angiospermas. História evolutiva. Síntese de amido.

ABSTRACT

The main storage carbohydrate in plants is starch, a natural resource of global importance for consumption that is part of human food, stock animal feed, as well as raw material for several industries. However, there are not many studies on the enzyme that synthesizes this carbohydrate, both in terms of its structure and function and its evolutionary history in the general context of angiosperms. This study uses a comprehensive phylogenetic analysis of the Waxy gene encoding the GBSS (Granule-Bound Starch Synthase) enzyme to better understand its evolutionary history. Furthermore, the protein structure contains two isoforms. To determine the structures, homology modeling was used, in addition to structural alignment to verify conserved regions. Phylogenetic analysis showed that the gene encoding the GBSS enzyme is highly conserved in angiosperms, separating large phylogenetic groups such as monocotyledons and dicotyledons. The enzyme contains two isoforms, GBSSI and GBSSII, and this diversification may be related to duplication processes in plant genomes. Although the general structure of the isoforms is similar, there is structural variation in the amino acid sequence. This study provides a comprehensive understanding of the evolutionary and structural history of the main enzyme involved in starch synthesis and these results should support future studies that aim to increase the understanding of starch biosynthesis and the evolutionary and functional divergence of GBSS in plants.

Key words: angiosperms. computational approaches. evolutionary history. starch synthesis.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO GERAL

Molecular and structural evolution of the Granule-Bound Starch Synthase (GBSS) in flowering plants

O que é a pesquisa?

O estudo investigou a evolução molecular do gene que codifica a enzima (GBSS) evolvida na síntese de amido em plantas, além de realizar análises estruturais sobre a enzima utilizando métodos *in sílico*, ou seja, programas e algoritmos de computador buscando analisar a estrutura proteica e propor uma hipótese filogenética para compreender a história evolutiva desse gene em plantas.

O amido é o principal carboidrato de armazenamento em plantas. Este açúcar natural é muito importante para as plantas pois serve de reserva de energia fazendo parte importante dos processos metabólicos e manutenção desses organismos, além de fazer parte da nutrição humana fazendo parte dos constituintes dos cereais, tubérculos e frutas.

O processo de síntese do amido é principalmente realizado através da ação da enzima GBSS que é codificada pelo gene Waxy que catalisa o processo de síntese na transferência de resíduos de glicose, atuando na formação de amilose e amilopectina, cadeias lineares e ramificadas respectivamente.

Atualmente o estudo sobre o gene e a enzima GBSS ainda são poucos e estão focados em pequenos grupos de plantas, sendo que ainda não há um estudo mais amplo envolvendo a diversidade de linhagens de angiospermas. Assim, este estudo pode auxiliar numa compressão mais ampla das classificações em diferentes níveis, além de informações estruturais importantes relacionados a ação enzimática.

Como foi feita?

Nesta pesquisa, foram utilizados modelos matemáticos e programas de computador para analisar de forma comparativa as sequências de gene Waxy que foram obtidas através da internet em bancos de dados digitais. Também foram utilizadas sequencias de aminoácidos das enzimas para se chegar as estruturas proteicas através de modelagem comparativa, também conhecida como modelagem por homologia, uma abordagem que também utiliza programas de computador e servidor via web (Figura 1).

Figura 1. Passos da análise filogenética.

A parte de modelagem por homologia segue a mesma lógica da análise filogenética, mudando apenas o passo 3 onde já se tem uma abordagem matemática e computacional estabelecida e no passo 4 há a criação de um modelo que será interpretado e analisado no passo 5.

A análise filogenética parte da ideia de homologia (biologicamente a semelhança de características entres os organismos, nesse caso o gene Waxy, possuem uma origem comum, advindo de um ancestral comum) então este gene pode revelar a história evolutiva e mudanças na estrutura da enzima GBSS.

Os resultados mostraram que o gene que codifica a enzima GBSS é muito similar em diferentes tipos de plantas. Assim, dizemos que o gene é conservado. Dois grandes grupos de plantas conhecidos como monocotiledôneas e dicotiledôneas possuem suas respectivas linhagens divergentes de GBSS. O gene GBSS possui uma linhagem ancestral GBSSI da qual a linhagem GBSSII evoluiu, o que parece ter tido origem em dicotiledôneas, tendo origem entre o ancestral mais recente entre rosídeas e superasterídeas.

Os resultados da modelagem estrutural *in sílico* resultaram em dois modelos teóricos GBSSI e GBSSII pertencem ao mesmo sistema de classificação de família proteica e domínio estrutural, mas apresentam algumas características estruturais diferentes, como número de estruturas secundárias e local de atividade enzimática. Além disso, a forma GBSSI da enzima interage com outros cofatores de ligação no seu sítio

enzimática além do ADP (Adenosina difosfato, composto orgânico importante no metabolismo celular).

Qual a importância?

Esses achados fornecem uma visão mais ampla sobre os eventos históricos evolutivos envolvendo o gene Waxy que codifica a enzima GBSS. O processo de diversificação desse gene no decorrer do tempo e o surgimento de um gene alternativo GBSSII provavelmente foi vital para o processo de síntese de amido em alguns grupos plantas.

Nas ciências biológicas, a área de estudo denominada filogenética é um ramo que utiliza de conhecimentos da área de exatas, estatística, onde é possível verificar relações evolutivas entre os organismos. Para este estudo, são utilizadas informações genéticas (genes ou genoma) que, por meio de aplicativos de computador, podem ser processadas e analisadas gerando modelos que podem explicar a história evolutiva de um organismo de interesse.

Autores

Elvis Santos Leonardo

Thiago Jose de Carvalho Andre

Kauê Santana da Costa

Título original

Evolução molecular e análise estrutural da enzima *Granule-Gound Gtarch Synthase* (GBSS) em angiospermas

Instituição

Universidade Federal do Oeste do Pará

Financiador

Coordenação de Aperfeiçoamento Pessoal de Nível Superior - CAPES

Sugestões de leitura

Gu, C., Wang, L., Zhang, L., Liu, Y., Yang, M., Yuan, Z., Li, S., & Han, Y. 2013. Characterization of Genes Encoding Granule-Bound Starch Synthase in Sacred Lotus Reveals Phylogenetic Affinity of Nelumbo to Proteales. Plant Molecular Biology Reporter, 31(5), 1157–1165.

Matioli, S. Russo., & Fernandes, F. M. de Campos. (Org) 2012. Biologia molecular e evolução. 2. ed. São Paulo: Holos, Editor, 2012. 256p.

VERLI, H. (2014). Bioinformática da Biologia à Flexibilidade Molecular. In: Universidade Federal do Rio Grande do Sul (org). Sociedade Brasileira de Bioquímica e Biologia Molecular. 1. ed. São Paulo. Pp 80-115.

VERLI, H. (2014). Bioinformática da Biologia à Flexibilidade Molecular. In: Universidade Federal do Rio Grande do Sul (org). Sociedade Brasileira de Bioquímica e Biologia Molecular. 1 ed. São Paulo. Pp 147-172.

CAPÍTULO ÚNICO

Manuscrito Submetido ao periódico *Jornal of Molecular Evolution*. As normas indicadas para redação de artigos pela revista estão disponíveis no link: https://www.springer.com/journal/239/submission-guidelines

Molecular and structural evolution of the Granule-Bound Starch Synthase (GBSS) in flowering plants

Elvis Santos Leonardo[1]

Kauê Santana da Costa[2]

Thiago Jose de Carvalho Andre[1,3]

1 – Programa de Pós-Graduação em Biodiversidade, Universidade Federal do Oeste do Pará, Rua Vera Paz s/n, Salé, Santarém (PA), Brazil, 68040-255

2 – Universidade Federal do Oeste do Pará, Instituto de Biodiversidade e Florestas, Rua Vera Paz s/n, Salé, Santarém (PA), Brazil, 68040-255

3 – Universidade de Brasília, Instituto de Ciências Biológicas, Departamento de Botânica, Campus Universitário Darcy Ribeiro, Asa Norte, Brasília (DF), Brazil, 70910-900

Molecular and structural evolution of the Granule-Bound Starch Synthase (GBSS) in flowering plants

Elvis Santos Leonardo[1*], Kauê Santana[2], Thiago André[1,3]

1 – Programa de Pós-Graduação em Biodiversidade, Universidade Federal do Oeste do Pará, Rua Vera Paz s/n, Salé, Santarém (PA), Brazil, 68040-255

2 – Universidade Federal do Oeste do Pará, Instituto de Biodiversidade e Florestas, Rua Vera Paz s/n, Salé, Santarém (PA), Brazil, 68040-255

3 – Universidade de Brasília, Instituto de Ciências Biológicas, Departamento de Botânica, Campus Universitário Darcy Ribeiro, Asa Norte, Brasília (DF), Brazil, 70910-900

* Correspondence author: elvis.etrr@gmail.com

***This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001**

Abstract

Starch is the main storage carbohydrate in plants, being a natural resource of global importance for human consumption and raw material for some industries. Granule-Bound Starch Synthase (GBSS) is the main enzyme responsible for the synthesis of starch. Nevertheless, this crucial enzyme is still poorly understood, both in terms of structure and function and its evolutionary history in the general context of flowering plants. This study applies a comprehensive phylogenetic analysis of genes encoding the GBSS enzyme to better understand its evolutionary history and structural constraints. To determine structural evolution, computational approaches were used to verify conserved regions, through structural alignment. Phylogenetic analysis showed that the gene encoding the GBSS enzyme is highly conserved in angiosperms and that the enzyme contains two isoforms, GBSSI and GBSSII. This molecular diversification may be related to duplication processes in plant genomes. Their structures, despite being similar, show structural variation, deletions, and amino acid mutations. This study provides a more comprehensive understanding

of the evolutionary and structural history of the main enzyme involved in starch synthesis, and these data should support future studies that aim to increase our understanding of starch biosynthesis and the evolutionary and functional divergence of GBSS in plants. We also highlight that caution regarding isoforms is needed when using GBSS for reconstructing phylogenetic trees.

Key words: starch synthesis, evolutionary history, computational approaches, angiosperms.

1 Introduction

In the process of starch synthesis, the main gene involved is known as Waxy, which encodes the enzyme Granule-bound starch synthase (GBSS) (Miao et al. 2014; Li et al. 2019). Molecular evolution and computational analysis of this synthesis in plants have been studied to understand the metabolism and evolutionary relationships of the main enzymes involved. Due to the low number of copies of the gene that encodes this enzyme, it has been widely used to study phylogenetic and evolutionary relationships in plants (Cheng et al. 2012). However, most of these studies are focused on small phylogenetic groups (e.g., Cheng et al. 2012; Li et al. 2012). There are still few computational studies on the structure of this enzyme and functional aspects that can bring a better understanding of its function (e.g., Momma and Fujimoto 2012; Cuesta-Seijo et al. 2013). Thus, evolutionary and theoretical computational studies can bring new insights into the evolutionary history of this enzyme and important structural aspects, such as classification and protein domains that are related to enzyme function, as well as to what possible events promoted the diversification of this enzyme.

Starch is the main storage carbohydrate produced by plants (Seung et al. 2020). It is widely used in human nutrition and is one of the main constituents of cereals, tubers, legumes, and fruits (Miao et al. 2014; do Carmo et al. 2020; Seung et al. 2020). It is found in the form of semi-crystalline insoluble granules and contains in its constitution two types of polymers: amylopectin and amylose (Cheng et al. 2012; Miao et al. 2014; do Carmo et al. 2020; Seung et al. 2020), which vary in their relative proportions. Amylose consists of a linear polymer, formed by glucose residues that are linked by α-1,4 bonds. Amylopectin also contains linear chains with α-1,4 bonds plus branched chains with α-1,6 bonds (Kato et al. 2019; Seung et al. 2020). The production of this carbohydrate in plants is notorious due to the amount produced and stored. The use of amylose for plant growth and survival remains unclear (Seung et al. 2020).

The synthesis of amylose is relatively simple and is catalyzed by the enzyme GBSS, which transfers glucose residues from ADP-Glucose producing long chains of amylose, and which also acts on the side chains (Cheng et al. 2012). In most cereals studied so far, GBSS consists of two isoforms, GBSSI and GBSSII. The GBSSI gene seems to be expressed exclusively in storage tissues, such as endosperm and seed embryos, while the GBSSII gene is expressed in tissues such as the leaf, stem, root, and pericarp (Vrinten and Nakamura 2000; Dian et al. 2003; Cheng et al. 2012). In addition, there may be a difference in the expression profiles of further isoforms, GBSSIa and GBSSIb genes in cereals and other monocots (Vrinten and Nakamura 2000; Dian et al. 2003; Cheng et al. 2012). For instance, in peas, these two GBSSI isoforms, GBSSIa and GBSSIb, are differentially expressed, where a high expression of GBSSIa is found in embryos, while GBSSIb is highly expressed in leaves (Edwards et al. 2002)

The isoenzymes involved in starch synthesis have been widely reported in monocots and dicots. A phylogenetic analysis of the GBSS gene by Lu et al. (2012) showed that the monocots form a group containing a GBSSI isoform and the dicots, a GBSSII isoform. This suggests that there was a divergence in the enzyme responsible for starch synthesis in plants. Another study to evaluate the variation of the GBSSI enzyme in a group of Poaceae showed that there is conservation between exons/introns (Shapter et al. 2009). The characterization of a new gene in *Amaranthus cruentus* L. (CrGBSSIb) and its phylogenetic

105 analysis found results that corroborated studies already carried out by Cheng et al (2012), where GBSSII
106 form a group in dicots and GBSSI in monocots (Park et al. 2017).

107 A study on the evolution of the GBSS genes of angiosperms showed a duplication 251 million
108 years ago (Lu et al. 2012; Cheng et al. 2012). Other studies involving GBSS and other genes that participate
109 in starch synthesis indicate that there were one or two gene duplication processes before grasses diverged
110 (Cheng et al. 2012). In monocots, these genes are divided into two gene families: GBSSI and GBSSII (Miao
111 et al. 2014). Furthermore, there are some structural similarities between GBSS and soluble starch synthases
112 (SSs), which may be related to common ancestry and duplication events (Qu et al. 2018).

113 Although there are studies on phylogenetic relationships using genes that synthesize starch in
114 plants, including the GBSS enzyme, comprehensive studies involving all sequences known for flowering
115 plants are still scarce. In addition, structural analysis of these enzymes, together with gene trees, can
116 enhance our understanding of the evolutionary aspects of this gene in flowering plants. Here, we present
117 the phylogenetic analysis of the GBSS enzyme and its isoforms I and II to understand the evolutionary
118 history of this enzyme in the context of flowering plants' evolution, in addition to a structural *in silico*
119 elucidation of the isoforms and structural comparison. We further classify protein families and domains.

120

121 **2 Material and methods**

122 2.1 Data acquisition

123 Data on the gene encoding the enzyme Granule-bound starch synthase (GBSS) were obtained from
124 NCBI (https://www.ncbi.nlm.nih.gov/) (Benson et al. 1990). An extensive search of the maximum number
125 of taxa available in the database was carried out with the aid of the algorithm BLAST (Altschul et al. 1997)
126 through the reference sequence NM103023.4 from *Arabidopsis thaliana* L. Heynh. General BLAST
127 parameters: maximum target sequences: 100, short queries: automatically adjusted for short input
128 sequences, expectation threshold: 0.05, word size: 28, maximum matches in a query range: 0.
129 Match/mismatch scores: 1, -2, gap cost: linear. Filter: low complexity regions, mask: mask for lookup table
130 only.

131 After obtaining information about GBSS, coding DNA sequences (CDSs) of the enzyme were
132 obtained. Only complete CDSs were selected for analysis. As a complement to the information already
133 obtained in the database, we recovered sequences of taxa with unannotated genomes, from an optimized
134 BLASTn using the same reference sequence, NM103023.4. In this case, only genomes of taxa without any
135 representative recovered by the first search were added. General BLASTn parameters: maximum target
136 sequences: 100, short queries: automatically adjusted for short input sequences, expectation threshold: 0.05,

word size: 11, maximum matches in a query range: 0. Match/mismatch scores: 2, -3, gap cost: existence: 5, extension: 2. Filter: low complexity regions, mask: mask for lookup table only.

To annotate the coding and non-coding regions, exons, and introns of the recovered genome fragments, we used the program AUGUSTUS (Stanke and Waack 2003), which is used for gene prediction through genomic fragments through a web server (https://bioinf.uni-greifswald.de/augustus/) (Stanke et al. 2004). This genetic prediction tool can use predictions with annotated reference information and *ab initio* (Stanke et al. 2006, 2008). All predicted genes were validated through the BLAST algorithm against the NCBI database to verify similarity with other GBSS genes already predicted.

In the end, 732 accessions were obtained, including CDSs of GBSSI and GBSSII, in addition to 101 genes isolated from unannotated genomes. We eliminated repeated data, very short sequences (< 500 bp) which resulted in 353 sequences of GBSS genes that were then used in subsequent analyses.

2.2 Phylogenetic analysis

For the phylogenetic analysis of the GBSS gene in flowering plants, an alignment was first performed using the MUSCLE algorithm (Edgar 2004), implemented in Geneious Prime® version 20221.1 (https://www.geneious.com/prime/) using the default settings parameters. Gaps were kept as predicted by the alignment algorithm. Phylogenetic analyzes were performed on the W-IQ-TREE web server (Trifinopoulos et al. 2016). The construction of the phylogenetic tree was based on maximum likelihood (ML = maximum likelihood) (Felsenstein 1981), where the substitution model was automated with an ultrafast bootstrap with a value of 3,000 with a maximum of 2,000 iterations. The minimum correlation coefficient was kept at 0.99. To support the branches, the SH-aLRT test was used with 2000 replicates (Guindon et al. 2010) together with the approximate Bayes test (Anisimova et al. 2011). The best nucleotide substitution model for the dataset was the GTR+F+I+G4.

2.3 In silico structural modeling

The structure of the GBSS enzyme was obtained by comparative modeling on the SWISS-MODEL server (Arnold et al. 2006; Biasini et al. 2014) (https://swissmodel.expasy.org/) The structure of the GBSS enzyme was obtained by comparative modeling on the SWISS-MODEL server (Arnold et al. 2006; Biasini et al. 2014) (https://swissmodel.expasy.org/). The modeling process steps are: (1) searching for models, (2)

aligning the target sequence with the model, (3) building the model, and (4) evaluating the quality (Arnold et al. 2006; Waterhouse et al. 2018). Eleven taxa were selected for modeling (Supplementary Material S2). For structural comparison, taxa that share the two isoforms GBSSI and GBSSII were chosen (Table 1).

The search for models was carried out on the SWISS-MODEL server, and two structures were found, for the GBSSI isoform the *Oryza sativa subsp, japonica* S. kato structure was used (PDB ID: 3VUE, string A, resolution 2.7 Å). For the GBSSII isoform, we selected the structure of *Hordeum vulgare* L. (PDB ID: 4HLN, strand A, resolution 2.7 Å). All the structures modeled had the regions of loops refined using a script from the MODELLER program (Webb and Sali 2017).

The theoretical models were validated by stereochemical analysis using the Ramachandran plot (Lovell et al. 2003), on the SWISS-MODEL server itself using MolProbity version 4.4 (Davis et al. 2007; Chen et al. 2010). The quality of the stereochemistry of the models was performed as a function of the phi and psi angles. Quality was also inferred through the QMEAN (Benkert et al. 2008, 2011), which is a composite score that employs average strength statistical potentials (Sippl 1993). To verify possible errors in the structures, an analysis was performed on the ProSa-Web server (Wiederstein and Sippl 2007) (https://prosa.services.came.sbg.ac.at/prosa.php), which is a widely used tool to check errors in three-dimensional protein structures. Error recognition can be done both in theoretical and experimentally elucidated structures (Wiederstein and Sippl 2007). The quality of the models was inferred by the z score and an energy graph.

The theoretical structures of the GBSS enzyme were aligned using the USCF Chimera software (Pettersen et al. 2004). Through the structural alignment, a theoretical structure of *Prunus avium* (L.) L. GBSSI and *Musa acuminata* Colla. GBSSII was selected. This choice was made based on the lowest value of the mean squared deviation (RMSD). The structural comparison made it possible to identify and compare conserved and non-conserved structural motifs, in addition to structural differences. Through these data, it was possible to infer relevant structural aspects of the molecular function, as well as differences in the catalytic sites of the enzyme.

To identify the protein family and domain, InterPro was used, which is a resource available from the European Institute of Bioinformatics, the EMBL-EBI (Goujon et al. 2010; McWilliam et al. 2013). The GenomeNet database was also used, which provides a MOTIF tool that also performs a functional search through the amino acid sequence (Kanehisa 2002). The sequences used in the search were the same used in

the comparative modeling together (Table 1). All sequences were aligned and compared to verify the conserved regions.

3 Results

3.1 Phylogenetic analysis

Alignment using the MUSCLE algorithm resulted in an aligned matrix of 8,945 bp of coding regions, with 13 exons recovered. Phylogenetic analysis, considering the well-known history of flowering plants relationships, shows that the GBSS gene virtually reflects this history; particularly the clades including monocots (branch support with 71.3% bootstrap) and eudicots (branch support 100% bootstrap) (Figure 1). The region in violet in Figure 1 shows that diversification of GBSS II occurred in the most recent common ancestor between rosids and superasterids. Overall, these results show that the GBSS gene had an ancestral GBSSI lineage from which the GBSSII lineage evolved. The complete phylogeny with the classification of groups and taxa is in Supplementary Material S1.

3.2 In silico structural modeling

The two isoforms of the GBSS enzyme are similar by having a low RMSD. The theoretical GBSSI structural isoform presented mainly 16 beta-sheet and 18 alpha-helix, and only the structure of *Solanum tuberosum* presented 15 beta-sheet and 19 alpha-helix. The GBSSII isoform presented 18 beta sheets and 21 alpha-helices. The structures in *Prunus avium* and *Solanum tuberosum* are also of the GBSSII isoform, but they had three more Alpha-helix structures. Therefore, we conclude that the GBSSI and GBSSII isoforms are different.

In addition to having different secondary structures, the GBSSII isoform contains a greater number of loops, and it's supposed binding site is located on the outside, while the GBSSI isoform contains a cavity between the two N-terminal and C-terminal regions forming a pit (Figure 2 and 3). In total, 11 theoretical structural models were obtained, all had stereochemistry values above 92%, good local energy parameters, approximate z-score values of experimental models, and energy graph showed favorable values (Supplementary Material S2).

The structural alignment resulted in low RMSD values in relation to the theoretical GBSSI structures compared to the 3VUE. In a model of *Oryza sativa japonica*, the structures are well correlated, there was also low RMSD between the theoretical GBSSII structures in relation to the 4HLN. A model of *Hordeum vulgare* shows an excellent structural relationship. The RMSD between the GBSSI and II structures showed higher values (Table 3) due to structural differences such as the number of Beta-Sheet and Alpha-helix, Supplementary Material S3.

The theoretical GBSSI structures presented, for the most part, 16 Beta-sheet and 18 Alpha-helix. Only the structure of *Solanum tuberosum* presented 15 Beta-sheet and 19 Alpha-helix. The GBSSII structures presented 18 Beta-Sheets and 21 Alpha-helices, and the *Prunus avium* and *Solanum tuberosum* structures presented 3 more Alpha-helices structures.

Here, we present the theoretical structure of the *Prunus avium* Granule-Bound Starch Synthase I (PaGBSSI) enzyme. The enzyme has 504 amino acids and two domains. The enzyme belongs to the Bacteria/Plant glycogen synthesizer family (InterPro: IPR011835). The starch synthesis domain is found in the N-terminal region from residue 94-354, and the Glycosyl transferase domain, family 1 is found in the C-terminal region from residue 408-524, compared to the structure of *Oryza sativa japonica* GBSSI (OsGBSSI), the domains have good structural overlap (Figure 2). Both domains have overlapping β-α-β structures.

The structural alignment showed that ADP-binding residues were conserved in PaGBSSI. The ADP-complexed structure shows an electron density map along with residue interactions (Fig. 3B). The structure of PaGBSSI with the KTGGL motif can be seen on the loop part, in violet (Fig. 3A). All residues that bind to ADP in (OsGBSSI) are conserved in the same position, except for residue His502, where in the structure of *Oryza sativa* there is a Gln493.

The theoretical structure of the Granule-Bound Starch Synthase II enzyme from *Musa acuminata* (MaGBSSII) contains 473 amino acids and belongs to the Bacteria/Plant glycogen synthesizer family (InterPro: IPR011835). The enzyme also contains the two domains present in the PaGBSSI protein, the starch synthesis domain is found in the N-terminal region going from residue 285-530, and the glycosyl transferase domain, family 1 is in the C-terminal region and goes from residue 567-697. The structural alignment of the MaGBSSII with the SSI of *Hordeum vulgare* (HvSSI) shows a good overlap of the two

250 structural domains (Figure 4). The structural comparison showed that the HvSSI residues are all conserved
251 in the MaGBSSII structure. The structure bounds to maltopentaose (Figure 5).

252 In the comparison using the amino acid sequences was possible to observe the isoforms sharing
253 the same family and protein domains presented in the structure of PaGBSSI and MaGBSSII. Although they
254 underwent a process of diversification between GBSSI and GBSSII during evolution, they still share the
255 family and protein domains.

256

257 **4 Discussion**

258 In this study, we sought to analyze the molecular evolution of the enzyme Granule-Bound Starch
259 Synthase (GBSS) in flowering plants from a phylogenetic and structural approach. Phylogenetic analysis
260 showed that the GBSS gene is conserved in flowering plants, reflecting the classification of large
261 phylogenetic groups. The GBSSII isoform is a lineage that evolved from an ancestral GBSSI lineage. The
262 GBSSI and GBSSII enzyme isoforms retain the same protein family and the same domains, but the binding
263 site is different. In addition, the GBSSI isoform has an affinity with other binding cofactors besides ADP.
264 We conclude that the GBSS enzyme in plants is conserved and presents different isoforms with different
265 structural characteristics, shapes, and binding sites.

266 Phylogenetic analysis of coding DNA sequences (CDSs) using the Maximum Likelihood method
267 (Felsenstein 1981), showed conservation in relation to large phylogenetic groups, such as Commelinids,
268 Superasterids, and Superrosids, in addition to the clustering of two major phylogenetic classification groups
269 of plants, monocots, and eudicots. These data agree with the classification made by other phylogenetic
270 studies using other molecular markers already used in phylogenetic studies, such as the APG III and APGIV
271 system (Bremer et al. 2009; Chase et al. 2016), highlighting the usefulness of GBSS to recover phylogenetic
272 histories in flowering plant groups.

273 The GBSS enzyme diversification process that occurred between the two large groups monocots
274 and eudicots may be associated with a major duplication event of the entire genomes (Cheng et al. 2012).
275 About 150 to 270 million years ago, there was a whole genome duplication event that is strongly associated
276 with this GBSS enzyme diversification process in plants (Qu et al. 2018). Thus, the duplication of the
277 ancestral GBSS gene in angiosperms was probably the result of a duplication of the entire genome (Cheng
278 et al. 2012; Qu et al. 2018).

A typical GT-B structure and a peculiar Rossmann fold present in OsGBSSI are identified in the N and C-terminal domain (MOMMA and FUJIMOTO 2012). Another important structural feature observed in PaGBSSI is the presence of a KTGGL motif in the N-terminal region. Studies with *Escherichia coli* glycogen synthase complexed with ADP and maltooligosaccharides showed a closed dynamic movement with two Rossmann fold domains associated with the KTGGL motif (Momma and Fujimoto 2012). The mean square deviation value of the atomic positions was 0.340 Å, suggesting that the structure PaGBSSI is very similar to the structure OsGBSSI and probably has a closed active state. The way ADP binds to the structure and involves the N-terminal and C-terminal domains may be a factor that contributes to the closed state of the enzyme (Momma and Fujimoto 2012). All residues that bind to ADP are conserved in the structure PaGBSSI, except for the change from Gly493 to His502 (Figure 3). This change can alter the binding stability of ADP at the binding site.

The theoretical structure of GBSSII compared to the structure of HvSSI also had high structural overlap, as the HvSSI adopts a characteristic GT-B fold (a double Rossmann fold) and it is possible to verify that the MaGBSSII structure also has the same characteristic (Cuesta-Seijo et al. 2013). The MaGBSSII structure, unlike the GBSSI structure of PaGBSSI, does not have a KTGGL domain, despite having high similarity and an RMSD below 1Å. In contrast, the large N-terminal region present in MaGbSSII and HvSSI is absent in PaGBSSI. This region is characterized by an extensive loop and may be related to the dynamics of the structure after binding with a cofactor (Cuesta-Seijo et al. 2013). The structure of MaGBSSII has an active site characteristic of GT5 as well as HvSSI, the amino acid residues of both domains have a closed conformation (Cuesta-Seijo et al. 2013).

Here we report the evolutionary history of the gene encoding the enzyme Granule-Bound Starch Synthase (GBSS) in the general context of flowering plants. The Waxy gene encoding the enzyme GBSS is well conserved in flowering plants and may be potentially useful to investigate phylogenetic relationships in plants, but it is necessary to be cautious because of the presence of two isoforms. In particular, some taxa of monocots are within the eudicots clades, nested within the GBSSII divergence, likely due to convergence processes acting on this gene. The structures of the PaGBSSI and MaGBSSII isoforms elucidated *in silico* provide clues as to how possibly the same binding mechanisms can form with the ADP and pentasaccharide molecule can bind to GBSSI and GBSSII respectively, the two isoforms share the same motifs and protein domains and presenting an N-terminal and C-terminal domain with Rossmann folds characteristic of these classes of proteins. The PaGBSSI structure presents a KTGGL domain in the N-terminal region, while this

309 motif is absent in the MaGBSSII isoform. In both structures, the residues that bind to cofactors are the
310 same, except in the PaGBSSI structure where there is a change from Gly to His. Together, these findings
311 provide new insight that increases our understanding of the evolutionary history, structure, and function of
312 the main enzyme involved in starch synthesis in plants.

313

314 **5 References**

315 Altschul SF, Madden TL, Schäffer AA, et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein
316 database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/NAR/25.17.3389

317 Anisimova M, Gil M, Dufayard JF, et al (2011) Survey of Branch Support Methods Demonstrates Accuracy,
318 Power, and Robustness of Fast Likelihood-based Approximation Schemes. Syst Biol 60:685–699.
319 https://doi.org/10.1093/SYSBIO/SYR041

320 Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for
321 protein structure homology modeling. Bioinformatics 22:195–201.
322 https://doi.org/10.1093/bioinformatics/bti770

323 Benkert P, Biasini M, Schwede T (2011) Toward the estimation of the absolute quality of individual protein
324 structure models. Bioinformatics 27:343–350. https://doi.org/10.1093/BIOINFORMATICS/BTQ662

325 Benkert P, Tosatto SCE, Schomburg D (2008) QMEAN: A comprehensive scoring function for model quality
326 assessment. Proteins 71:261–277. https://doi.org/10.1002/PROT.21715

327 Benson D, Boguski M, Lipman DJ, Ostell J (1990) The National Center for Biotechnology Information.
328 Genomics 6:389–391. https://doi.org/10.1016/0888-7543(90)90583-G

329 Biasini M, Bienert S, Waterhouse A, et al (2014) SWISS-MODEL: modeling protein tertiary and quaternary
330 structure using evolutionary information. - PubMed - NCBI. Nucleic Acids Res 42:W252–W258

331 Bremer B, Bremer K, Chase MW, et al (2009) An update of the Angiosperm Phylogeny Group classification for
332 the orders and families of flowering plants: APG III. Botanical Journal of the Linnean Society 161:105–
333 121. https://doi.org/10.1111/j.1095-8339.2009.00996.x

334 Chase MW, Christenhusz MJM, Fay MF, et al (2016) An update of the Angiosperm Phylogeny Group
335 classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean
336 Society 181:1–20. https://doi.org/10.1111/boj.12385

337 Chen VB, Arendall WB, Headd JJ, et al (2010) MolProbity: All-atom structure validation for macromolecular
338 crystallography. Acta Crystallogr D Biol Crystallogr 66:12–21.
339 https://doi.org/10.1107/S0907444909042073

340 Cheng J, Khan MA, Qiu W-M, et al (2012) Diversification of Genes Encoding Granule-Bound Starch Synthase
341 in Monocots and Dicots Is Marked by Multiple Genome-Wide Duplication Events. PLoS One 7:e30088.
342 https://doi.org/10.1371/journal.pone.0030088

343 Cuesta-Seijo JA, Nielsen MM, Marri L, et al (2013) Structure of starch synthase I from barley: insight into
344 regulatory mechanisms of starch synthase activity. Acta Crystallogr D Biol Crystallogr 69:1013–1025.
345 https://doi.org/10.1107/S090744491300440X

346 Davis IW, Leaver-Fay A, Chen VB, et al (2007) MolProbity: All-atom contacts and structure validation for
347 proteins and nucleic acids. Nucleic Acids Res 35:. https://doi.org/10.1093/nar/gkm216

348 Dian W, Jiang H, Chen Q, et al (2003) Cloning and characterization of the granule-bound starch synthase II gene
349 in rice: Gene expression is regulated by the nitrogen level, sugar and circadian rhythm. Planta 218:261–
350 268. https://doi.org/10.1007/s00425-003-1101-9

351 do Carmo CD, Sousa MBE, dos Santos Silva PP, et al (2020) Identification and validation of mutation points
352 associated with waxy phenotype in cassava. BMC Plant Biol 20:1–12. https://doi.org/10.1186/s12870-020-
353 02379-3

354 Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids
355 Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340

356 Edwards A, Vincken JP, Suurs LCJM, et al (2002) Discrete forms of amylose are synthesized by isoforms of
357 GBSSI in pea. Plant Cell 14:1767–1785. https://doi.org/10.1105/tpc.002907

358 Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol
359 17:368–376. https://doi.org/10.1007/BF01734359

360 Goujon M, McWilliam H, Li W, et al (2010) A new bioinformatics analysis tools framework at EMBL-EBI.
361 Nucleic Acids Res 38:. https://doi.org/10.1093/nar/gkq313

362 Guindon S, Dufayard JF, Lefort V, et al (2010) New Algorithms and Methods to Estimate Maximum-Likelihood
363 Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol 59:307–321.
364 https://doi.org/10.1093/SYSBIO/SYQ010

365 Kanehisa M (2002) The KEGG databases at GenomeNet. Nucleic Acids Res 30:42–46.
366 https://doi.org/10.1093/nar/30.1.42

367 Kato K, Suzuki Y, Hosaka Y, et al (2019) Effect of high temperature on starch biosynthetic enzymes and starch
368 structure in japonica rice cultivar ‘Akitakomachi’ (Oryza sativa L.) endosperm and palatability of cooked
369 rice. J Cereal Sci 87:209–214. https://doi.org/10.1016/j.jcs.2019.04.001

370 Li C, Li Q-G, Dunwell JM, Zhang Y-M (2012) Divergent Evolutionary Pattern of Starch Biosynthetic Pathway
371 Genes in Grasses and Dicots. Mol Biol Evol 29:3227–3236. https://doi.org/10.1093/molbev/mss131

372 Li Q, Pan Z, Liu J, et al (2019) A mutation in Waxy gene affects amylose content, starch granules and kernel
373 characteristics of barley (Hordeum vulgare). Plant Breeding 138:513–523.
374 https://doi.org/10.1111/pbr.12695

375 Lovell SC, Davis IW, Arendall WB, et al (2003) Structure validation by Calpha geometry: phi,psi and Cbeta
376 deviation. Proteins 50:437–50. https://doi.org/10.1002/prot.10286

377 Lu Y, Li L, Zhou Y, et al (2012) Cloning and Characterization of the Wx Gene Encoding a Granule-Bound Starch
378 Synthase in Lotus (Nelumbo nucifera Gaertn). Plant Mol Biol Report 30:1210–1217.
379 https://doi.org/10.1007/s11105-012-0430-x

380 McWilliam H, Li W, Uludag M, et al (2013) Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids
381 Res 41:. https://doi.org/10.1093/nar/gkt376

382 Miao H, Sun P, Liu W, et al (2014) Identification of genes encoding granule-bound starch synthase involved in
383 amylose metabolism in banana fruit. PLoS One 9:e88077. https://doi.org/10.1371/journal.pone.0088077

384 MOMMA M, FUJIMOTO Z (2012) Interdomain Disulfide Bridge in the Rice Granule Bound Starch Synthase I
385 Catalytic Domain as Elucidated by X-Ray Structure Analysis. Biosci Biotechnol Biochem 76:1591–1595.
386 https://doi.org/10.1271/bbb.120305

387 Park YJ, Nishikawa T, Matsushima K, Nemoto K (2017) Characterization of a new granule-bound starch synthase
388 gene found in amaranth grains (Amaranthus cruentus L.). Molecular Breeding 37:.
389 https://doi.org/10.1007/s11032-017-0712-y

390 Pettersen EF, Goddard TD, Huang CC, et al (2004) UCSF Chimera - A visualization system for exploratory
391 research and analysis. J Comput Chem 25:1605–1612. https://doi.org/10.1002/jcc.20084

392 Qu J, Xu S, Zhang Z, et al (2018) Evolutionary, structural and expression analysis of core genes involved in
393 starch synthesis. Sci Rep 8:12736. https://doi.org/10.1038/s41598-018-30411-y

394 Seung D, Echevarría-Poza A, Steuernagel B, Smith AM (2020) Natural polymorphisms in Arabidopsis result in
395 wide variation or loss of the amylose component of StARCH. Plant Physiol 182:870–881.
396 https://doi.org/10.1104/pp.19.01062

397 Shapter FM, Eggler P, Lee LS, Henry RJ (2009) Variation in Granule Bound Starch Synthase I (GBSSI) loci
398 amongst Australian wild cereal relatives (Poaceae). J Cereal Sci 49:4–11.
399 https://doi.org/10.1016/j.jcs.2008.06.013

400 Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins: Structure, Function,
401 and Bioinformatics 17:355–362. https://doi.org/10.1002/PROT.340170404

402 Stanke M, Diekhans M, Baertsch R, Haussler D (2008) Using native and syntenically mapped cDNA alignments
403 to improve de novo gene finding. Bioinformatics 24:637–644.
404 https://doi.org/10.1093/BIOINFORMATICS/BTN013

405 Stanke M, Steinkamp R, Waack S, Morgenstern B (2004) AUGUSTUS: a web server for gene finding in
406 eukaryotes. Nucleic Acids Res 32:W309–W312. https://doi.org/10.1093/nar/gkh379

407 Stanke M, Tzvetkova A, Morgenstern B (2006) AUGUSTUS at EGASP: using EST, protein and genomic
408 alignments for improved gene prediction in the human genome. Genome Biology 2006 7:1 7:1–8.
409 https://doi.org/10.1186/GB-2006-7-S1-S11

410 Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel.
411 Bioinformatics 19:ii215–ii225. https://doi.org/10.1093/BIOINFORMATICS/BTG1080

412 Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ (2016) W-IQ-TREE: a fast online phylogenetic tool for
413 maximum likelihood analysis. Nucleic Acids Res 44:W232–W235. https://doi.org/10.1093/nar/gkw256

414 Vrinten PL, Nakamura T (2000) Wheat Granule-Bound Starch Synthase I and II Are Encoded by Separate Genes
415 That Are Expressed in Different Tissues. Plant Physiol 122:255–264. https://doi.org/10.1104/pp.122.1.255

416 Waterhouse A, Bertoni M, Bienert S, et al (2018) SWISS-MODEL: homology modelling of protein structures
417 and complexes. Nucleic Acids Res 46:W296–W303. https://doi.org/10.1093/nar/gky427

418 Webb B, Sali A (2017) Protein Structure Modeling with MODELLER. In: Methods in Molecular Biology. pp
419 39–54

420 Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-
421 dimensional structures of proteins. Nucleic Acids Res 35:W407–W410.
422 https://doi.org/10.1093/NAR/GKM290

423

424 **Tables**

425 Table 1. Taxa used in *in silico* structural modeling.

Species	Order	Higher classification	
Musa acuminata GBSSII *Musa acuminata GBSSI*	Zingiberales	Comelinids	Monocots
Prunus avium GBSSI *Prunus avium GBSSII*	Rosales	Superrosids	Eudicots
Glycine soja GBSSI *Glycine soja GBSSII*	Fabales		
Solanum tuberosum GBSSI *Solanum tuberosum GBSSII*	Solanales	Superasterids	
Arabidopsis thaliana GBSSI	Brassicales		
Coffea euginoides GBSSI *Coffea euginoides GBSSI*	Gentianales		

426

427 Table 2. SWISS-MODEL web search results: identity values, coverage, and resolution of the model
428 structures. All accession models were obtained by x-ray.

Accession	Species	Enzyme	Template	Identity	Coverage	Resolution
NP_174566.1	*Arabidopsis thaliana*	GBSSI	3VUE.A	69.45	0.86	2.7 Å
XP_027182769.1	*Coffea eugenioides*	GBSSI	3VUE.A	70.78	0.86	2.7 Å
XP_028182166.1	*Glycine soja*	GBSSI	3VUE.A	69.83	0.87	2.7 Å
AHA51121.1	*Musa acuminata*	GBSSI	3VUE.A	73.43	0.76	2.7 Å
XP_021822644.1	*Prunus avium*	GBSSI	3VUE.A	70.97	0.85	2.7 Å
XP_015162571.1	*Solanum tuberosum*	GBSSI	3VUE.A	69.45	0.83	2.7 Å
XP_027153219.1	*Coffea eugenioides*	GBSSII	4HLN.A	50.82	0.62	2.7 Å

XP_028195575.1	*Glycine soja*	GBSSII	4HLN.A	52.47	0.63	2.7 Å
AHA51125.1	*Musa acuminata*	GBSSII	4HLN.A	50.53	0.62	2.7 Å
XP_021812457.1	*Prunus avium*	GBSSII	4HLN.A	50.41	0.63	2.7 Å
NP_001274977.1	*Solanum tuberosum*	GBSSII	4HLN.A	50.92	0.63	2.7 Å

Table 3. RMSD values result from structural alignment between targets and models.

Targets + GBSSI (3VUF)	**Enzyme**	**RMSD Å**	**Targets + SS1 (4HLN)**	**RMSD Å**	**Enzyme**
Arabidopsis thaliana	GBSSI	0.340	*Arabidopsis thaliana*	0.884	GBSSI
Coffea eugenioides	GBSSI	0.344	*Coffea eugenioides*	0.883	GBSSI
Glycine soja	GBSSI	0.360	*Glycine soja*	0.854	GBSSI
Musa acuminata	GBSSI	0.345	*Musa acuminata*	0.862	GBSSI
Prunus avium	GBSSI	0.340	*Prunus avium*	0.838	GBSSI
Solanum tuberosum	GBSSI	0.365	*Solanum tuberosum*	0.867	GBSSI
Hordeum vulgare 4HLN	SSI	0.903	*Oryza sativa Japonica*	0.903	GBSSI
Coffea eugenioides	GBSSII	0.961	*Coffea eugenioides*	0.259	GBSSII
Glycine soja	GBSSII	0.910	*Glycine soja*	0.388	GBSSII
Musa acuminata	GBSSII	0.894	*Musa acuminata*	0.204	GBSSII
Prunus avium	GBSSII	0.910	*Prunus avium*	0.220	GBSSII
Solanum tuberosum	GBSSII	0.948	*Solanum tuberosum*	0.466	GBSSII

Figures

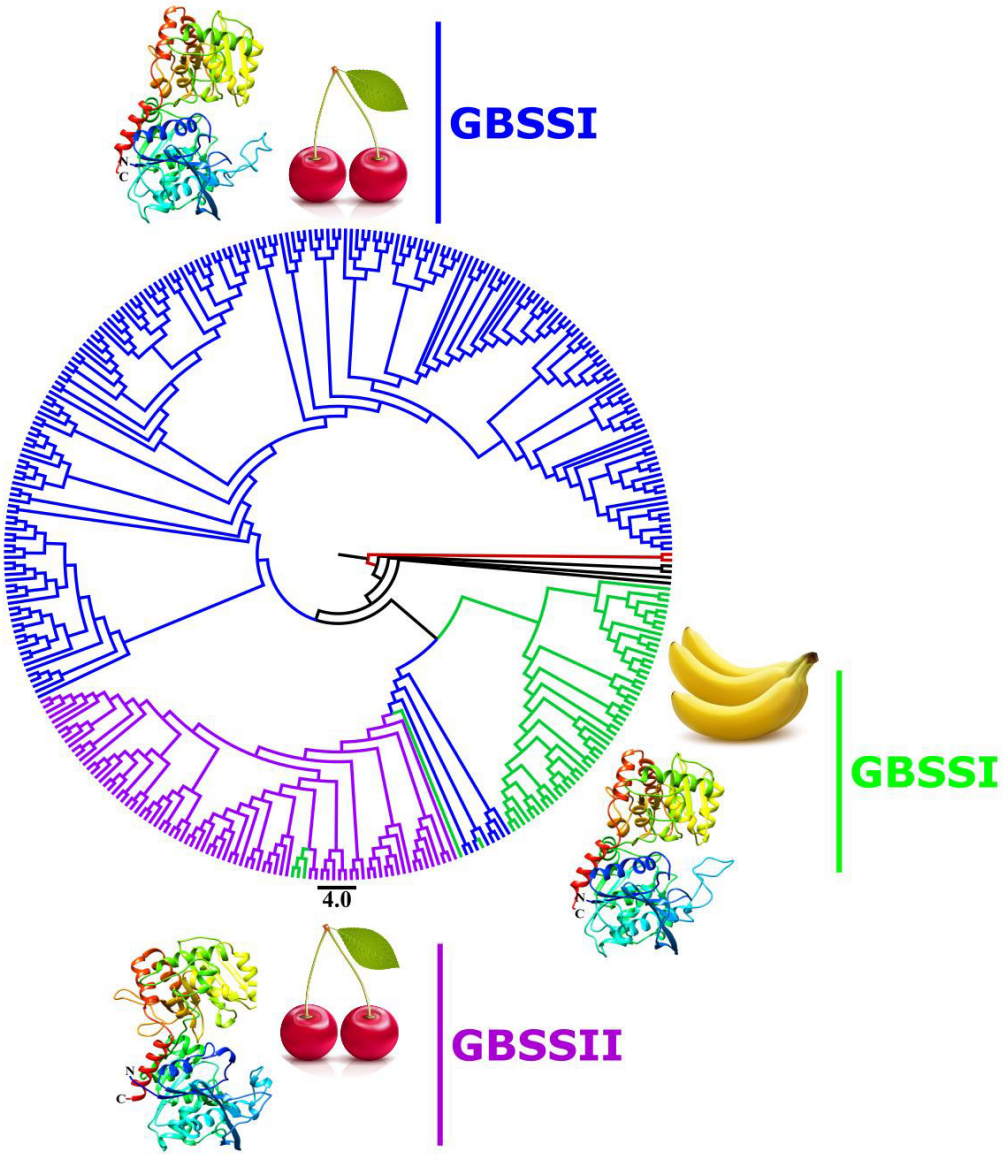


Figure 1 Maximum likelihood phylogenetic analysis of the GBSS enzyme in flowering plants. The tree is derived from the alignment of nucleotide sequences of CDS coding regions. In blue are the taxa belonging to the classification of eudicots, in green are the monocots, and in violet are the taxa that share the GBSSII isoform. There is a representation of two organisms, *Musa acuminata* in monocots and *Prunus avium* in eudicots. The structures of GBSSI are more similar to each other than compared to GBSSII from the same taxa. Bootstrap values for all branches can be seen in Supplementary Material S1.

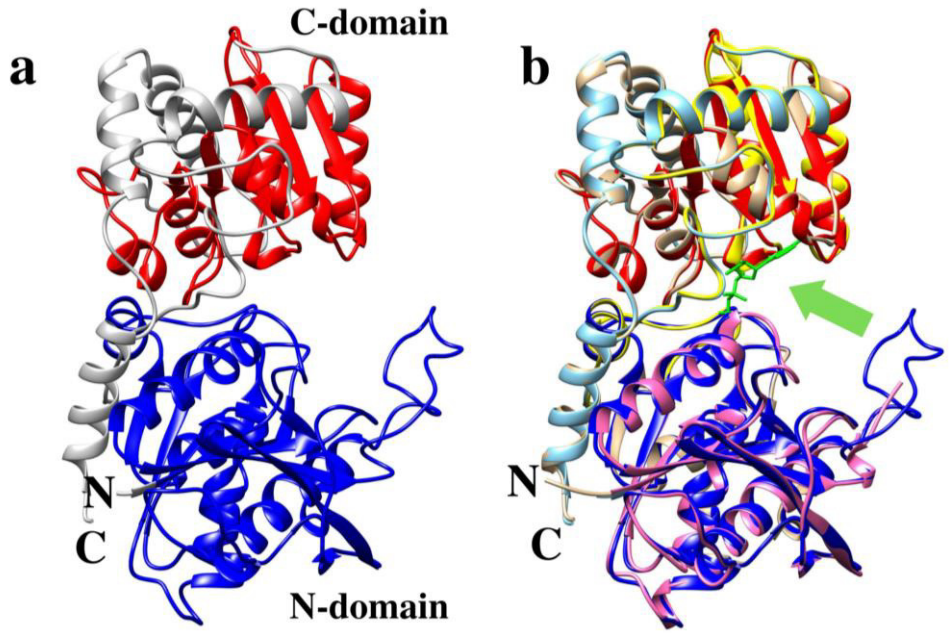


Figure 2 Representation of the theoretical model and structural alignment of GBSS. (**a**) PaGBSSI structure: the region in blue represents the region of the N-terminal domain and in red the C-terminal domain. (**b**) view of the structural alignment between PaGBSSI and OsGBSSI. The N-terminal and C-terminal domains are represented by pink and yellow respectively. The green arrow shows the Adenosine Diphosphate molecule linked to the OsGBSSI structure.

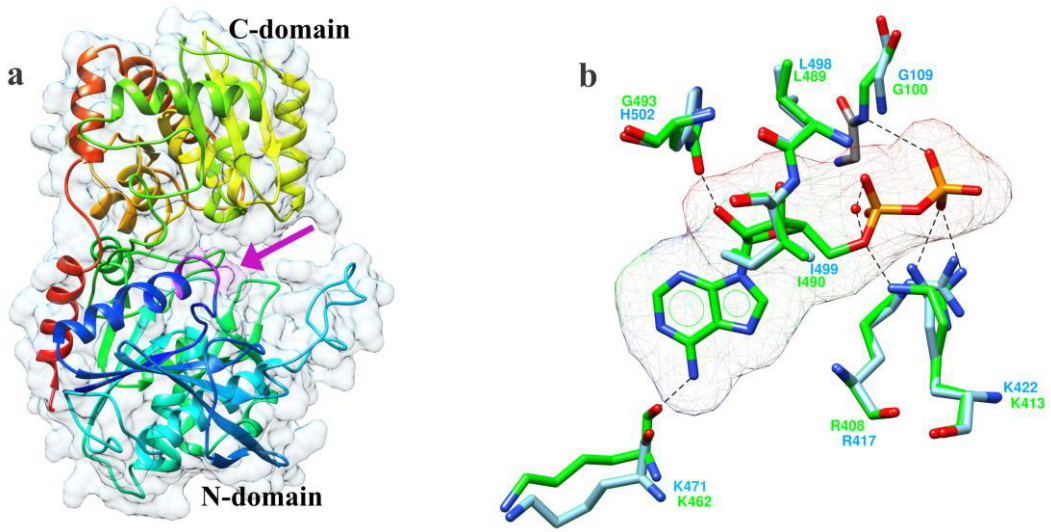


Figure 3 PaGBSSI structure with KTGGL motif and comparison with ADP-complexed structure of OsGBSSI. (**a**) density map of the PaGBSSI structure with protein motif in the N-terminal region. The motif

is highlighted with a violet-colored region and arrow. (**b**) superposition of OsGBSSI structure (green color residues, PDB 3VUF) complexed with ADP with electron density map. Prunus avium GBSSI residues are shown in blue.

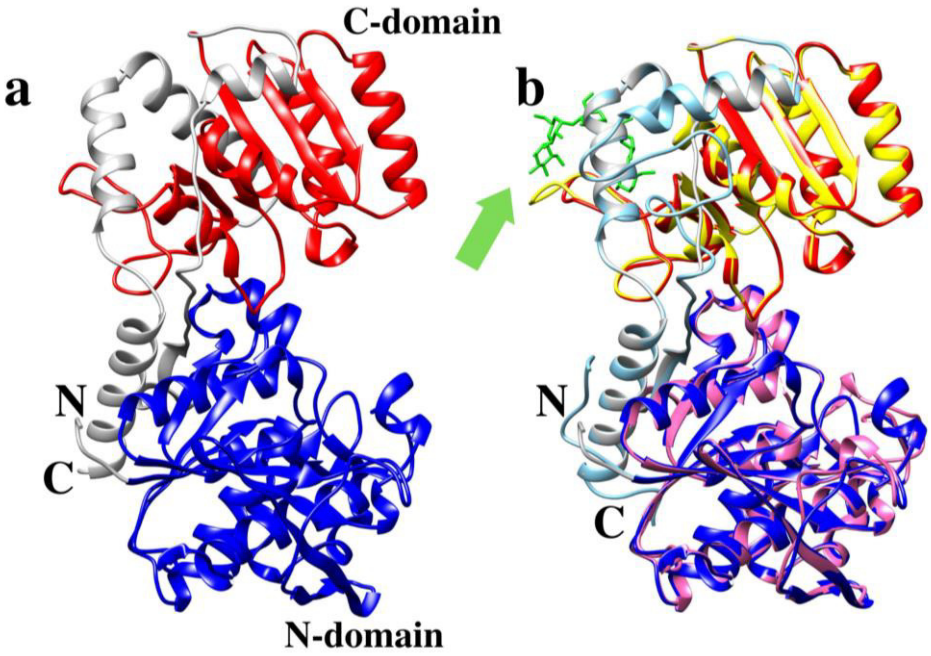


Figure 4 Representation of the theoretical model and structural alignment. (**a**) MaGBSSII structure. The region in blue represents the N-terminal domain and in red the C-terminal domain. (**b**) view of the structural alignment between PaGBSSI and HvSSI. The N-terminal and C-terminal domains are represented by pink and yellow respectively. The green arrow shows the Adenosine Diphosphate molecule linked to the HvSSI structure.

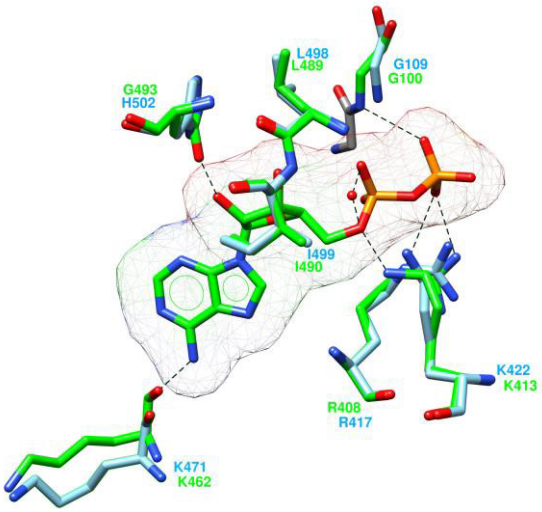


Figure 5 Pentasaccharide with an electron density map complexed to the SSI protein structure. Residues in green belong to the SSI structure of HvSSI (PDB 4 HLN), and residues in blue are from the GBSSII structure of MaGBSSII.

499 **Supplementary Material S1**

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

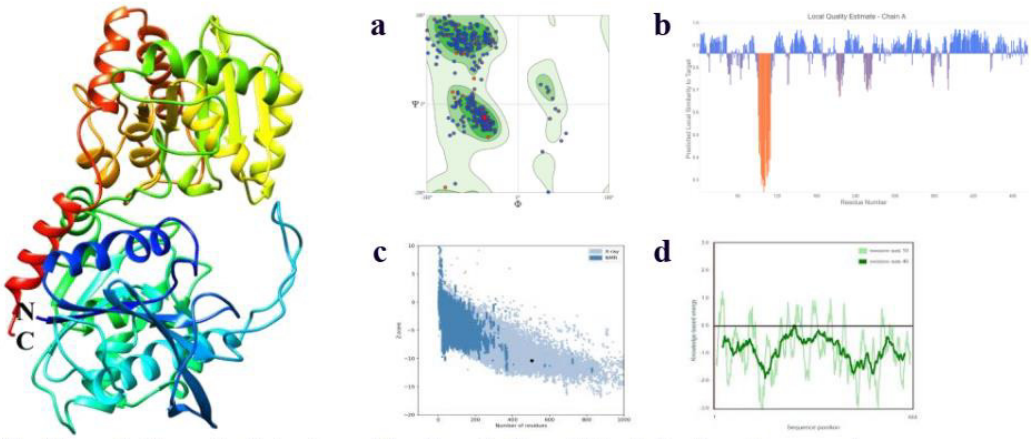
526

527

528

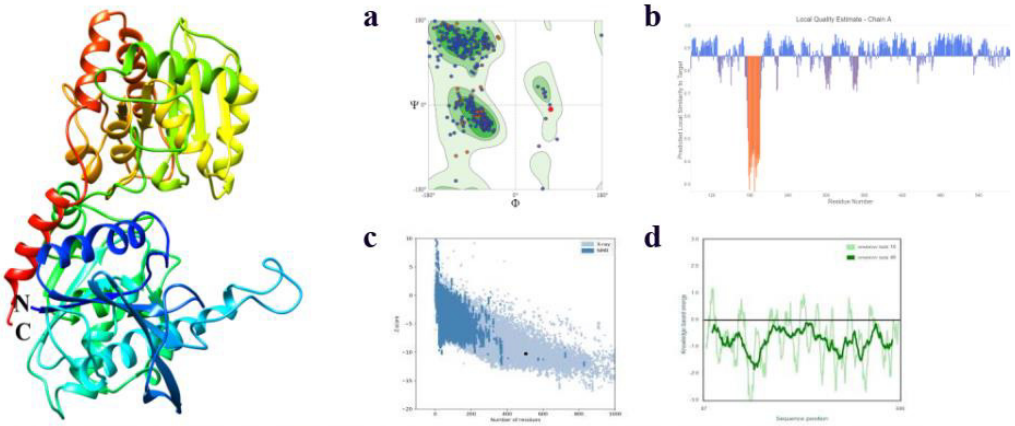
Angiosperms
Monocots
Diversification of GBSSII
Diversification of monocots and eudicotsEudicots
Marchantiales
Selaginellales
Nymphaeales
Amborellales
Magnoliales
Arecales
Poales
Commelinids
Dioscoreales
Asparagales
Pandanales
Liliales
Commelinales
Zingiberales
Laurales
Cupressales
Piperales
Zingiberales
Curcurbitales
Myrtales
Brassicales
Poales
Monocots
Granule-Bound Stach Synthase GBSSI
Rosales
Fabids
Curcurbitales
Fagales
Vitales
Caryophyllales
Asparagales
Arecales
Zingiberales
Ericales
Apiales
Asterales
Lamiales
Gentianales
Solanales
Superasterids
Sapindales
Malvales
Myrtales
Celastrales
Malpighiales
Fabales
Superrosids
Granule-Bound Stach Synthase GBSSII
Caryophyllales
Vitales
Saxifragales
Ericales
Asterales
Superasterids
Apiales
Sapindales
Fabales
Rosids
Fagales
Rosales
Granule-Bound Stach Synthase GBSSI
Proteales
Trochodendrales
Ranunculares
Proteales
Malpighiales
Myrtales
Cornales
Ericales
Caryophyllales
Cornales
Asterales
Dipsacales
Garryales
Aquifoliales
Boraginales
Lamiales
Gentianales
Superasterids
Eudicots
Solanales
Santalales
Saxifragales
Vitales
Myrtales
Fagales
Rosales
Cucurbitales
Fabales
Superrosids
Oxalidales
Sapindales
Malvales
Sapindales
Malpighiales
Celastrales
Brassicales
3.0

529 **Supplementary Material S2**



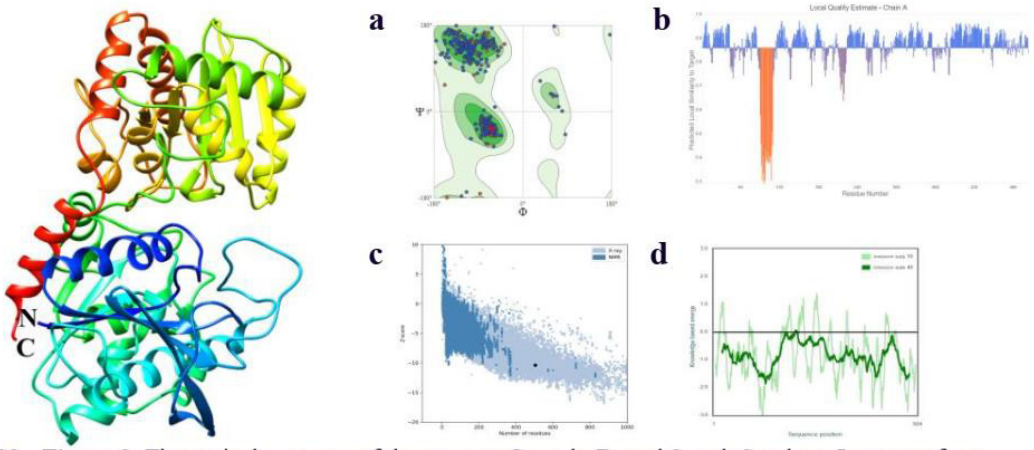
S2 - Figure 1. Theoretical structure of the Granule-Bound Starch Synthase I enzyme from *Arabidopsis thaliana*. (**a**) Stereochemical analysis, (**b**) QNEMAN composite score analysis, (**c**) Z-score, (**d**) Energy graph.

530



FS2 - Figure 2. Theoretical structure of the enzyme Granule-Bound Starch Synthase I enzyme from *Coffea eugenioides*. (**a**) Stereochemical analysis, (**b**) QNEMAN composite score analysis, (**c**) Z-score, (**d**) Energy graph.

531

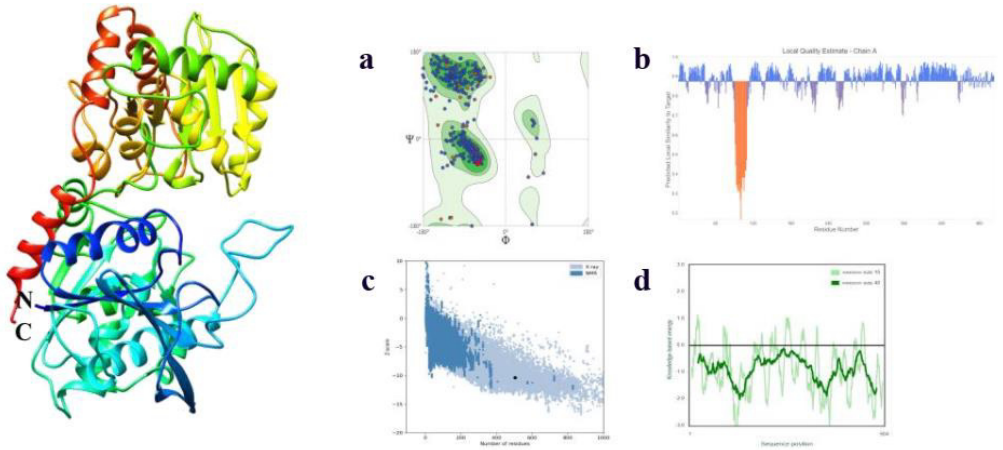


S2 - Figure 3. Theoretical structure of the enzyme Granule-Bound Starch Synthase I enzyme from *Glycine soja*. (**a**) Stereochemical analysis, (**b**) QNEMAN composite score analysis, (**c**) Z-score, (**d**) Energy graph.

532

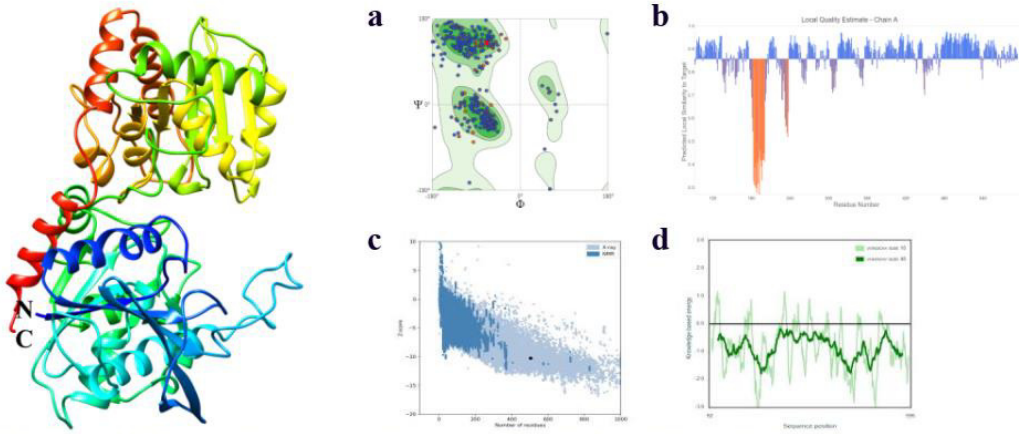
533

534



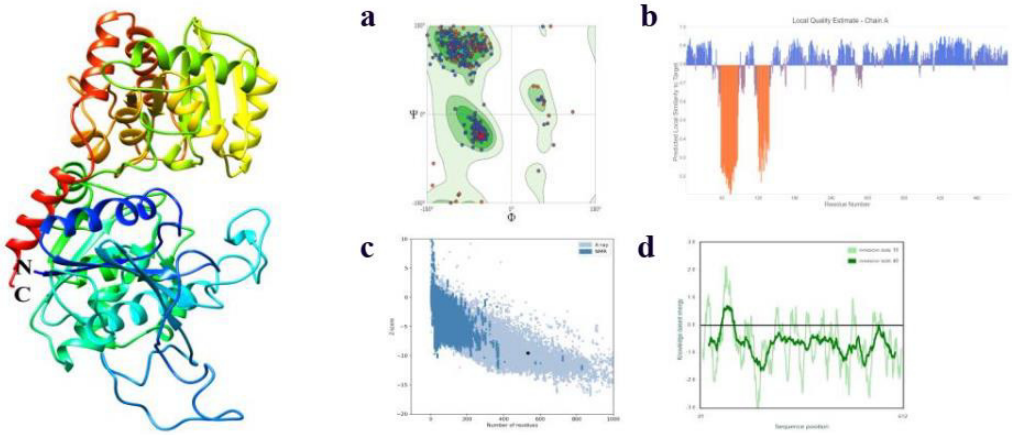
S2 - Figure 4. Theoretical structure of the enzyme Granule-Bound Starch Synthase I enzyme from *Musa acuminata*. (**a**) Stereochemical analysis, (**b**) QNEMAN composite score analysis, (**c**) Z-score, (**d**) Energy graph.

535



S2 - Figure 5. Theoretical structure of the enzyme Granule-Bound Starch Synthase I enzyme from *Prunus avium*. (**a**) Stereochemical analysis, (**b**) QNEMAN composite score analysis, (**c**) Z-score, (**d**) Energy graph.

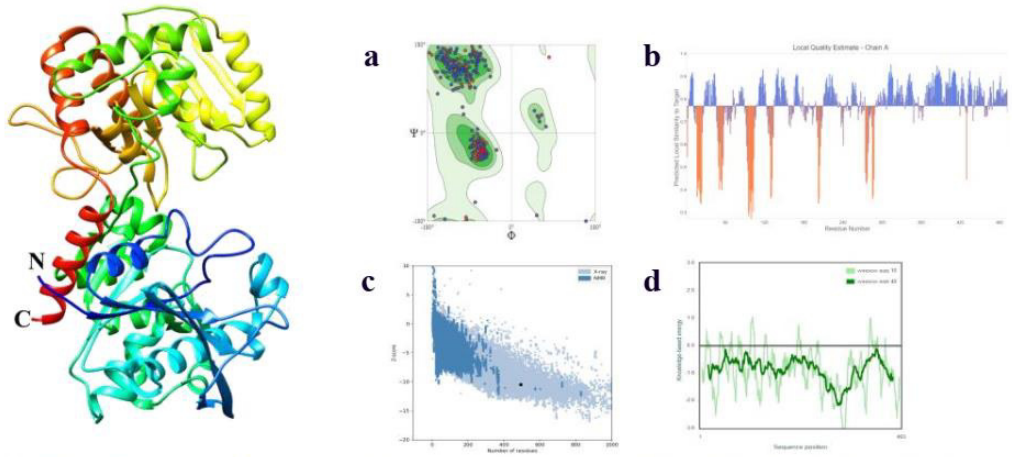
536



S2 - Figure 6. Theoretical structure of the enzyme Granule-Bound Starch Synthase I enzyme from *Solanum tuberosum*. (**a**) Stereochemical analysis, (**b**) QNEMAN composite score analysis, (**c**) Z-score, (**d**) Energy graph.

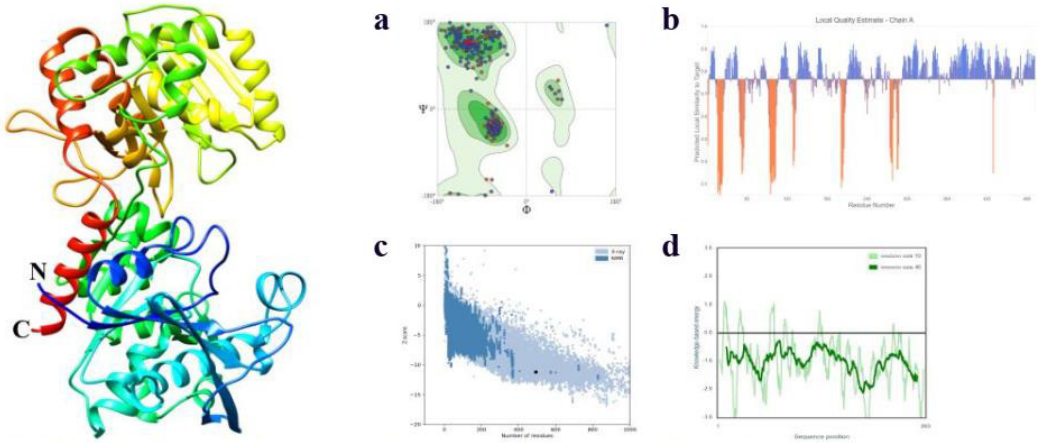
537

538



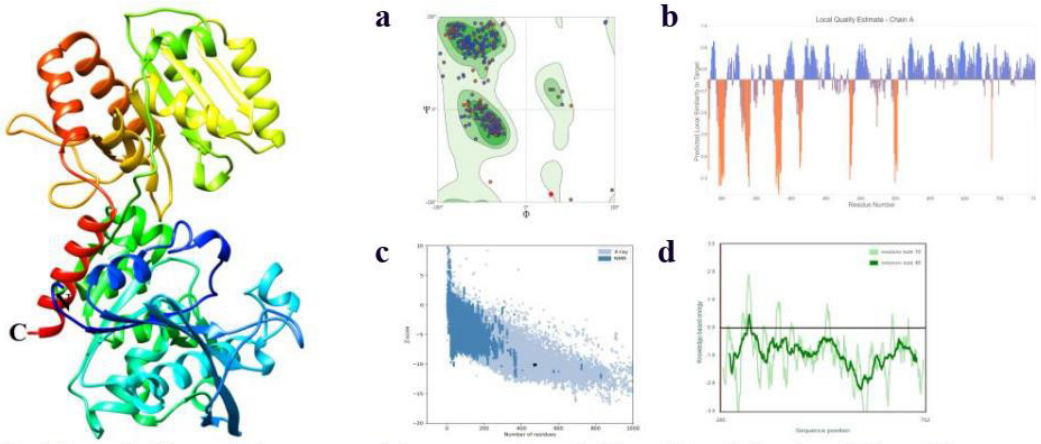
S2 - Figure 7. Theoretical structure of the enzyme Granule-Bound Starch Synthase II from *Coffea eugenioides*. (**a**) Stereochemical analysis, (**b**) QNEMAN composite score analysis, (**c**) Z-score, (**d**) Energy graph.

539



S2 - Figure 8. Theoretical structure of the enzyme Granule-Bound Starch Synthase II from *Glycine soja*. (**a**) Stereochemical analysis, (**b**) QNEMAN composite score analysis, (**c**) Z-score, (**d**) Energy graph.

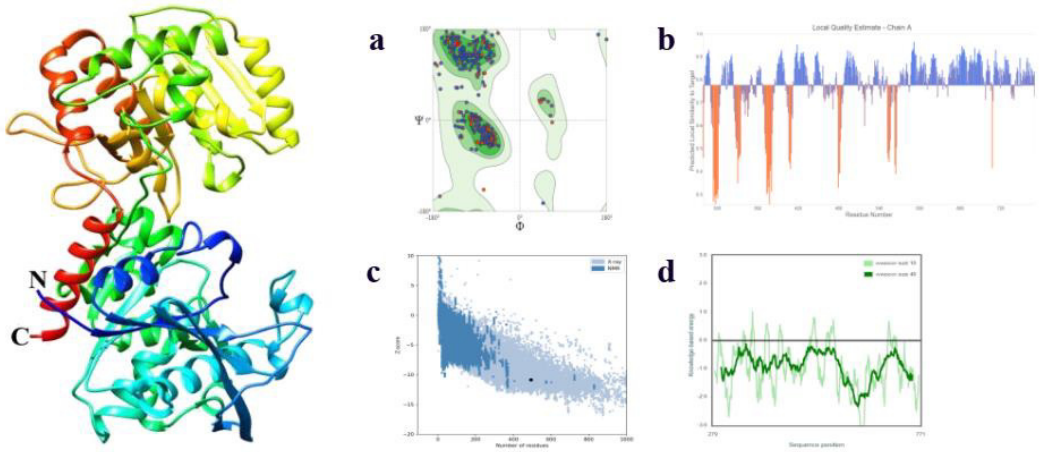
540



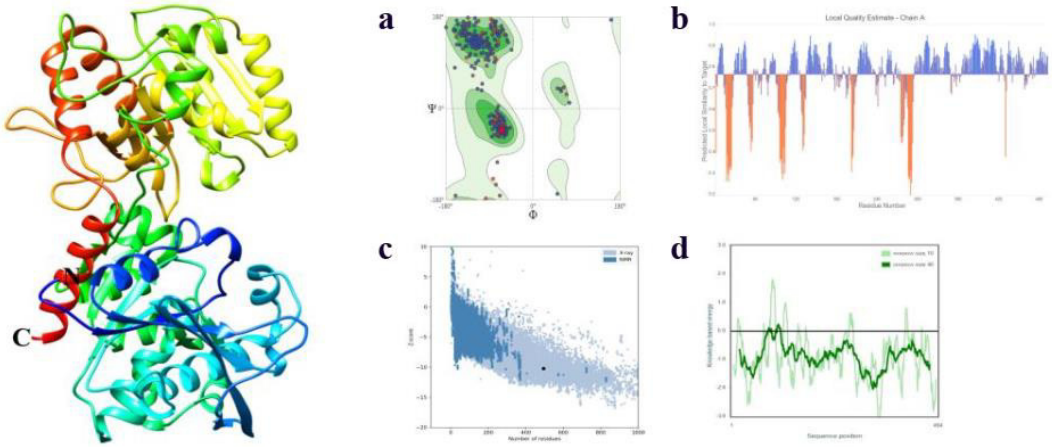
S2 - Figure 9. Theoretical structure of the enzyme Granule-Bound Starch Synthase II from *Musa acuminata*. (**a**) Stereochemical analysis, (**b**) QNEMAN composite score analysis, (**c**) Z-score, (**d**) Energy graph.

541

542

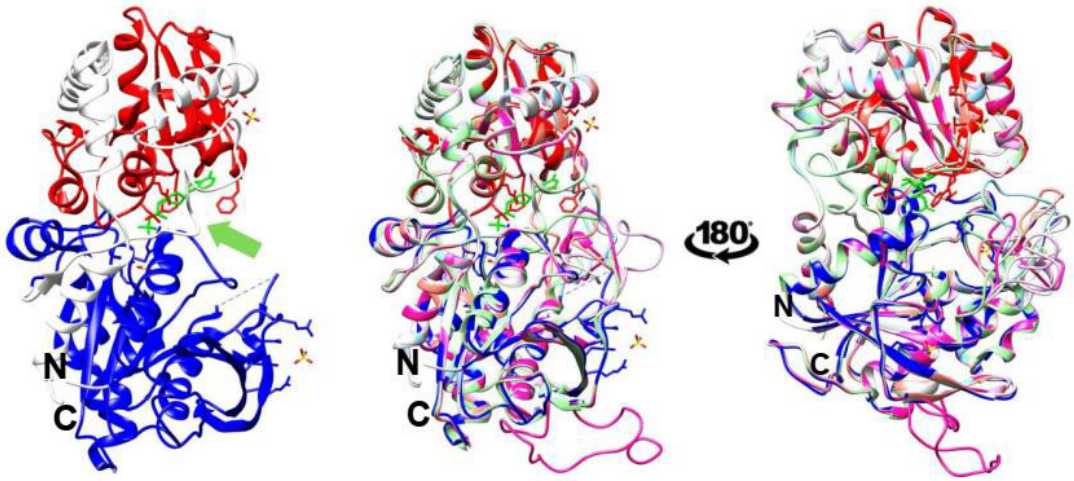


S2 - Figure 10. Theoretical structure of the enzyme Granule-Bound Starch Synthase II from *Prunus avium*. (**a**) Stereochemical analysis, (**b**) QNEMAN composite score analysis, (**c**) Z-score, (**d**) Energy graph.



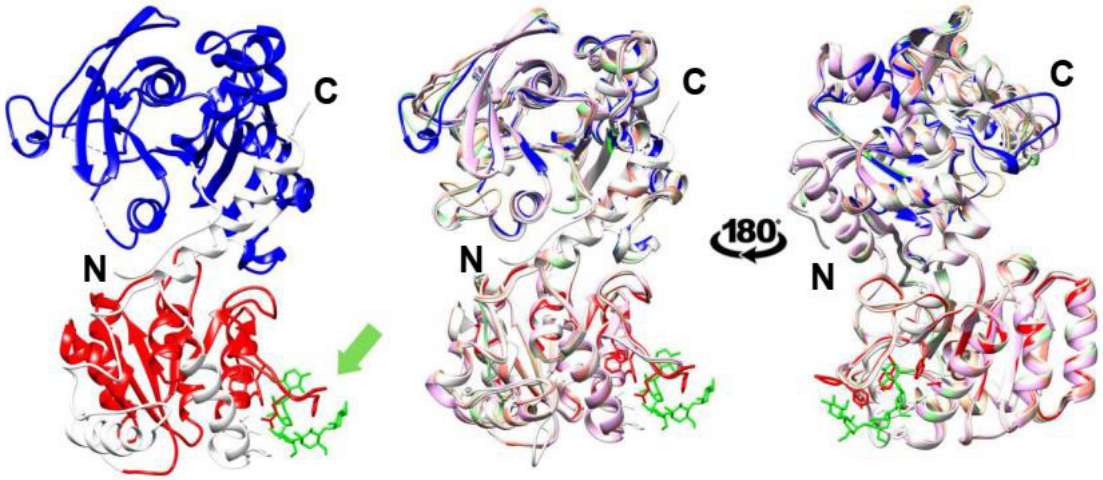
S2 - Figure 11. Theoretical structure of the enzyme Granule-Bound Starch Synthase II from *Solanum tuberosum*. (**a**) Stereochemical analysis, (**b**) QNEMAN composite score analysis, (**c**) Z-score, (**d**) Energy graph.

556 **Supplementary Material S3**



S3 - Figure 1. Structural alignment of the model structure of *Oriza sativa* 3VUE.A GBSSI with predicted theoretical structures of GBSSI. The green arrow shows the ADP cofactor complexed with the structure. The catalytic domain of starch synthesis is identified in blue (INTERPRO: IPR013534). In red, the Glycosyl transferase family 1 domain (INTERPRO: IPR01296).

557



S3 - Figure 2. Structural alignment of the model structure of *Hordeum vulgare* 4HLN.A SSI with predicted theoretical structures of GBSSII. The green arrow shows the maltooligosaccharide bound to the surface of the structure. The catalytic domain of starch synthesis is identified in blue (INTERPRO: IPR013534). In red, the Glycosyl transferase family 1 domain (INTERPRO: IPR01296).

558

Acadêmicas.

Comentários à coordenação do PPGBEES:

O tema do trabalho é interessante e a abordagem utilizada pelo discente está adequada ao cumprimento dos objetivos apresentados. No entanto, percebem-se vários equívocos na realização da metodologia e na interpretação dos resultados que, na visão do presente parecerista, não suportam as conclusões apresentadas no final do trabalho. No manuscrito enviado também é possível observar equívocos em conceitos importantes (parálogos, ortólogos, isoformas, etc.). Faltam detalhes importantes na descrição da metodologia empregada, como: características do alinhamento utilizados, limite de corte entre parálogos e ortólogos, tratamento dos gaps, teste de hipóteses alternativas, método de enraizamento, uso das sequências de aminoácidos, verificação de códons, tratamento de regiões da proteína sem cobertura de moldes, etc. Embora as figuras e tabelas estejam bem apresentadas, as legendas precisam de uma maior quantidade de informações. O resumo do trabalho precisa de uma profunda revisão pois não está bem descrito. A escrita em língua inglesa está compreensível, no entanto, uma revisão é recomendada. Tais correções são essenciais à qualidade do trabalho e por tal motivo, o meu parecer é **APROVADO COM CORREÇÕES**. Ressalto que as correções são necessárias e me coloco a disposição (email ou vídeoconferência) para caso o discente tenha dúvidas ou questionamentos em relação ao meu parecer ou as correções encaminhadas em anexo (na forma de anotações no arquivo .pdf enviado).

Avaliação final do projeto de dissertação de mestrado

I - Aprovada (X) – Com correções.

Aprovada: indica que o revisor aprova a dissertação sem ou com correções. Na existência de correções, estas devem ser indicadas nos comentários à coordenação e/ou no próprio documento da dissertação.

IV - Reprovada ( )

Reprovada: indica que a dissertação não é adequada.

Nome do membro da banca: João Paulo Matos Santos Lima (UFRN)

Data: 30/11/2022

Assinatura:

Art. 65. Após a apresentação da dissertação em sessão pública, o discente terá até 60 dias corridos para entregar a versão final da dissertação, contendo a ficha catalográfica, conforme artigo 60 deste regimento, sob pena de não diplomação até que a versão final seja devidamente submetida no Sistema de Gestão de Atividades Acadêmicas.

Comentários à coordenação do PPGBEES:

Junto aos documentos para avaliação foi enviado o arquivo pdf “Regras basicas para as Introducoes Gerais”. Este tópico é obrigatório? Pois não o encontrei na dissertação enviada.

Avaliação final do projeto de dissertação de mestrado

I - Aprovada (X)

Aprovada: indica que o revisor aprova a dissertação sem ou com correções. Na existência de correções, estas devem ser indicadas nos comentários à coordenação e/ou no próprio documento da dissertação.

IV - Reprovada ( )

Reprovada: indica que a dissertação não é adequada.

Nome do membro da banca: Gabriel Iketani Coelho

Data: 28/11/2022

Assinatura:

gov.br Documento assinado digitalmente
GABRIEL IKETANI COELHO
Data: 28/11/2022 18:22:56-0300
Verifique em https://verificador.iti.br

Comentários à coordenação do PPGBEES:

Embora a dissertação apresente um artigo adequado a um produto esperado do mestrado no programa, com aderência e potencial de publicação, a dissertação apresentada está incompleta e não deve ser aprovada pelo colegiado. O discente não incluiu a introdução geral, obrigatório pelo formato de dissertação do PPGBEES. Esse formato foi aprovado pelo colegiado do curso, que reconhece a importância da apresentação da introdução geral tanto quanto exercício de escrita de divulgação científica por parte do mestre a ser formado, quanto do texto em si como parte de compreensão acessível que constará na versão final do texto. A aprovação do texto do discente faltando essa parte crucial da dissertação abre um precedente que pode invalidar o formato atual da dissertação do Programa.
Como esse pré-requisito pode não estar claro para membros avaliadores da banca que sejam externos ao curso, eu sugiro ao colegiado, que reprove a dissertação, dando ao discente o prazo regimental para apresentação de uma versão que se adeque às exigências do curso, e vinculando a aprovação (caso o número total de pareceres já emitidos para a dissertação seja positivo) à apresentação da dissertação com a introdução geral.

À exceção da ausência da introdução geral, nos demais quesitos a dissertação apresenta um artigo na área de evolução, com potencial de alto impacto, e necessita apenas de pequenas adequações. Caso a dissertação estivesse completa, ou seja, com a introdução geral, a minha recomendação seria a aprovação. Alguns pontos a serem destacados são:

Com relação à escrita, há algumas sugestões de correções e melhorias no texto. Algumas frases estão truncadas ou incompletas. Nos resultados, há uma parte que deveria estar em Materiais e Métodos. A citação e respectivas referências precisam ser padronizadas. Todos os comentários estão detalhados no arquivo da dissertação que acompanha essa avaliação.

Avaliação final do projeto de dissertação de mestrado

I - Aprovada ( )

Aprovada: indica que o revisor aprova a dissertação sem ou com correções. Na existência de correções, estas devem ser indicadas nos comentários à coordenação e/ou no próprio documento da dissertação.

IV - Reprovada (X)

Reprovada: indica que a dissertação não é adequada.

Nome do membro da banca: Thaís Elias Almeida

Data: 28 de novembro de 2022

Assinatura: