Structural and Functional Analysis of ADP-Glucose Pyrophosphorylase (AGPase)

Dissertation submitted in fulfilment of the requirement of

BACHELORS OF TECHNOLOGY

IN

BIOINFORMATICS

BY

SARTHAK SOOD (191902)

PUSHPIT THAKUR (191904)

Under the supervision of

DR. RAJ KUMAR

2023

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY WAKNAGHAT, SOLAN

CERTIFICATE

This is to certify that the work being presented in the major project entitled "**Structural and functional analysis of wheat ADP-glucose pyrophosphorylase (AGPase)**" in partial fulfilment for the award of degree of B. Tech. in Bioinformatics and submitted to the Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by **Sarthak Sood** and **Pushpit Thakur** during the time period from August 2022 to May 2023 under the supervision of **Dr. Raj Kumar**, Assistant Professor, Jaypee University of Information Technology, Waknaghat.. This work has not been submitted partially or wholly to any other University or Institute for the award for this or any other degree or diploma.

29/05/2023

(Supervisor's Signature)

Dr. Raj Kumar
Assistant Professor,
Department of Biotechnology & Bioinformatics
Jaypee University of Information and Technology (JUIT)
Waknaghat, Solan, H.P - 173234

Date: 29th May, 2023

DECLARATION BY THE STUDENT

We hereby declare that the work presented in this report entitled "Structural and functional analysis of wheat ADP-glucose pyrophosphorylase (AGPase)" in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Bioinformatics** submitted in the Department of Biotechnology & Bioinformatics, Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from August 2022 to May 2023 under the supervision of **Dr. Raj Kumar** (Assistant Professor, Department of Biotechnology and Bioinformatics). We also authenticate that we have carried out the above mentioned project work under the proficiency stream.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

(Student Signature)
Sarthak Sood (191902)

(Student Signature)
Pushpit Thakur (191904)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

29/05/2023

(Supervisor's Signature)
Supervisor's name: **Dr. Raj Kumar**
Designation: Assistant Professor
Department name: Department of Biotechnology & Bioinformatics
Dated: 29th May, 2023

Acknowledgement

We wish to express our sincere gratitude and appreciation for everyone who helped us complete this project. We would especially like to express our gratitude to **Dr. Raj Kumar** sir, our project supervisor, and **Dr. Sudhir Kumar** sir, head of the Department of Biotechnology and Bioinformatics, for allowing us to work on the project.

Moreover, we would like to convey our gratitude to the Bioinformatics Laboratory staff for allowing us to use their tools and resources.

It was really helpful to us, and we learned many interesting facts regarding the structural and functional analysis of ADP-glucose pyrophosphorylase. We were also enlightened to various methods of working with proteins on various software's like the Discovery Studio and Pymol before using them in actual molecular dynamics (MD) simulations.

Sarthak Sood (191902)

Pushpit Thakur (191904)

Plagiarism Certificate

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: 29/05/2023

Type of Document (Tick): PhD Thesis | M.Tech Dissertation/ Report | B.Tech Project Report | Paper

Name: SARTHAK, PUSHPIT Department: BIOINFORMATICS Enrolment No 191902, 191904

Contact No. 70186-43803, 82199-39427 E-mail. 191902@juitsolan.in, 191904@juitsolan.in

Name of the Supervisor: DR. RAJ KUMAR

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): STRUCTURAL AND FUNCTIONAL ANALYSIS OF ADP-GLUCOSE PYROPHOSPHORYLASE (AGPase)

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

- Total No. of Pages = 41
- Total No. of Preliminary pages = 12
- Total No. of pages accommodate bibliography/references = 3

Sarthak Goel

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found Similarity Index at 16 (%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

(Signature of Guide/Supervisor) 29/05/2023

29/05/2023

Signature of HOD

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Generated Plagiarism Report Details (Title, Abstract & Chapters)	
	• All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String		Word Counts	
Report Generated on			Character Counts	
		Submission ID	Total Pages Scanned	
			File Size	

Checked by
Name & Signature

Librarian

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com

Table of contents

List of abbreviations

AGPase	ADP glucose pyrophosphorylase
Sh2	Shrunken-2
Bt2	Brittle-2
LS	Large subunit
SS	Small subunit
AA	Amino acid
MDS	Molecular dynamic simulations
DS	Discovery studio
PE	Potential energy
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
RG	Radius of gyration
PCA	Principal component analysis

List of figures

List of tables

Abstract

ADP-glucose pyrophosphorylase (AGPase) is a key enzyme and is made up of two large subunits (LS) and two small subunits (SS). This enzyme plays a crucial role in starch biosynthesis. AGPase is found in both the cytosol and the plastid. The functional unit of AGPase is as a homotetramer having 4 identical monomers. The length of a single monomer is 451bp. AGPase is encoded by two genes, Shrunken-2 (Sh2) and Brittle-2 (Bt2). The Shrunken-2 (Sh2) gene encodes regulatory larger subunits and the Brittle-2 (Bt2) gene encodes smaller subunits having catalytic properties. There is information present related to AGPase and its genes, but adequate information about the variation and diversity in the structure, evolution and function is not present.

Three-dimensional (3D) structure of wheat AGPase was not available in the protein data bank (PDB) and hence needed to be modelled from an orthologous sequence that was found by running Blast on the LS and SS of wheat. The results showed that the potato AGPase sequence was the most similar to wheat AGPase. Therefore, potato AGPase was used as a template for building the 3D structure of wheat AGPase. The template structure had multiple gaps which were filled using Discovery Studio (DS) software. The template structure was optimised by energy minimization using steepest descent algorithm and running 100 ns molecular dynamics (MD) simulations. Potential energy, RMSD, RMSF, radius of gyration, and average structure were calculated from simulation trajectories. The MD simulations results showed that the protein system stayed stable throughout and did not show any major conformational changes. The optimized potato AGPase was used as a template for modelling of wheat AGPase using homology modelling technique. The final model was validated using Ramachandran plot. The validation results showed that majority of the residues although a small number of residues were identified as outliners falling in the disallowed regions. The model provided the initial structure for further analysis of structure and functional understanding of wheat AGPase.

CHAPTER 1

INTRODUCTION

Introduction

Starch or amylum is a polymeric carbohydrate made up of many glucose units connected by glycosidic linkages. Most green plants synthesise this polysaccharide as a means of storing energy. It is the most abundant carbohydrate in human diets worldwide and is present in particularly abundant levels in common foods like wheat, potatoes, maize (corn) and rice. Pure starch- a white, flavourless, and odourless powder is insoluble in both cold water and alcohol. It is made up of two different kinds of molecules: branched amylopectin and linear and helical amylose. Depending on the plant, starch often has a weight ratio of 75 to 80% amylopectin and 20 to 25% amylose [1]. Glycogen is a more intricately branched form of amylopectin, which serves as an animal's energy reserve.

1.1 Etymology

The Germanic source of the term "starch" has the meanings "strong, rigid, strengthen, stiffen." Strength in modern German has a long history of use in the textile industry, specifically in the size of yarn for weaving and the starching of linen. It is also related to the Greek word for starch, "amylon," which means "not milled." It offers the root amyl, which functions as a prefix for a number of 5-carbon compounds associated with or generated from starch (e.g. amyl alcohol).

1.2 Biosynthesis

Using the enzyme glucose-1-phosphate adenylyltransferase, plants first convert glucose 1-phosphate to ADP-glucose to make starch. Energy in the form of ATP is needed for this step. ADP is subsequently released and amylose is produced when the enzyme starch synthase joins the ADP-glucose to a developing chain of glucose residues via a 1,4-alpha glycosidic bond. Similar to how UDP-glucose is put to the non-reducing end of glycogen during glycogen synthesis, the ADP-glucose is almost certainly added to the non-reducing end of the amylose polymer [2]. The branched amylopectin is produced by the starch branching enzyme, which adds 1,6-alpha glycosidic linkages between the amylose chains. Some of these branches are cut off by the enzyme isoamylase, which breaks down starch. These enzymes come in a variety of isoforms, which makes the manufacturing process extremely complex. Glycogen and

amylopectin both have similar structures, however the former contains around one branch point for every ten 1,4-alpha linkages while the latter has one for every thirty. The diagram for starch biosynthesis is shown in the Figure 1.2.1. While mammals and fungi produce glycogen from UDP-glucose, bacteria typically produce glycogen from ADP-glucose. Amylopectin is produced from ADP-glucose.

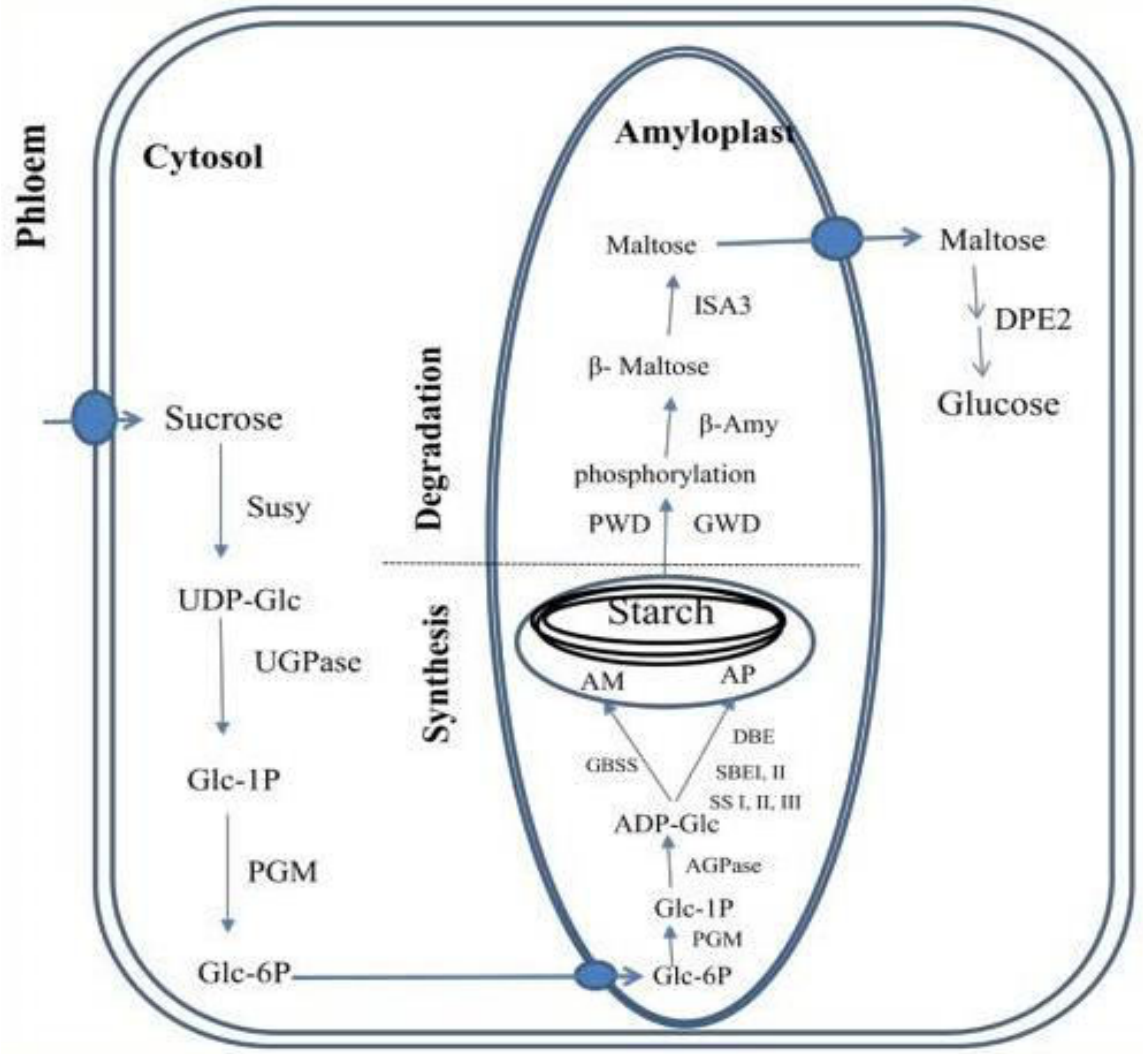


Fig 1.2. 1 Starch biosynthetic pathway

The endosperm cell imports sucrose, which is then converted to hexose phosphates, usually by sucrose synthase. These hexose phosphates act as basic components for the synthesis of carbohydrates, proteins, and oils. Hexose phosphate is primarily required for starch production in developing endosperms. These energy-dense reactions also require ATP, which is produced by

glycolysis and oxidative phosphorylation. The amount and molecular makeup of the starch produced can be influenced by enzyme or transporter activities that influence ATP and hexose phosphate availability prior to the production of ADPglucose.

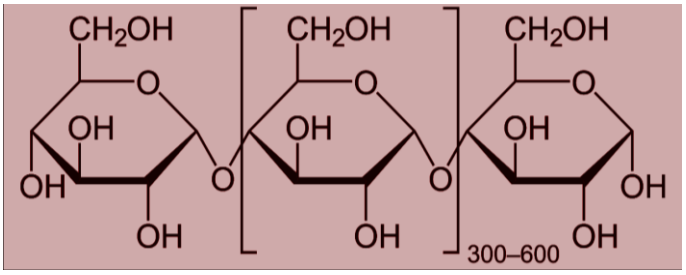


Fig 1.2. 2 Structure of amylose molecule

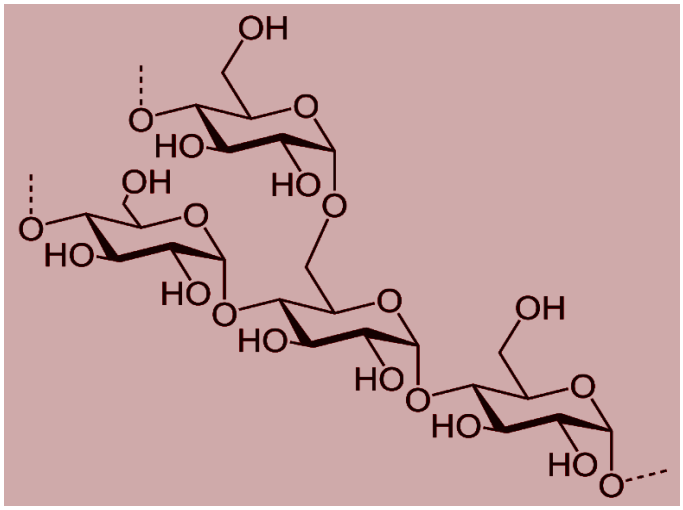


Fig 1.2. 3 Structure of amylopectin molecule

1.3 Structure and role of AGPase in starch biosynthesis

The first important regulatory and rate-limiting enzyme in plants' starch production is called ADP-glucose pyrophosphorylase (AGPase).[3]. The only precursor molecules used by AGPase to create the fundamental components of starch, ADP-glucose (ADP-G) moieties are glucose-1-phosphate (G-1-P) and ATP.[4] This enzyme regulates the amount of starch in other plant organs including the endosperm of cereal grains. The activity of AGPase is inhibited by a number of abiotic stressors, including heat and drought, which significantly reduce crop yield.[5] As a

result, attempts have been made to genetically modify the AGPase in maize to produce its thermostable forms. However, more research is needed to fully understand the precise nature of the metabolic and physiological changes brought on by the genetic engineering of AGPase.[6] Both the cytosol and plastids contain AGPase. ADP-glucose (ADP-G) is mostly produced in cereals in the cytosol of a developing endosperm cells, however some ADP-G is also produced in the plastids. However, AGPase-mediated ADP-G synthesis largely takes place in the plastids of dicotyledonous plants.

The freshly synthesised AGPase subunits are transported into the plastids by transit peptides, which are also encoded by the plastidial AGPase. With two comparable small subunits, SS (α2), and two similar big subunits, LS (β2), both plastidial and cytosolic AGPases work as heterotetramers (α2β2). The nuclear genes that encode the SS and LS are different, and their molecular weights are just slightly different.

The SS performs both the catalytic and the regulatory functions, while the LS mostly performs regulatory activities. According to the respective functions of the two subunits, the LS in cereal seed endosperm is crucial in determining grain yield [7], whilst the SS in leaves adds to plant biomass. Most cereals have been shown to contain one of two forms of AGPase SS genes, Type 1 or Type 2. Two transcripts are produced by type 1 genes: cytosolic endosperm and leaf plastidial. Similar to Type 1 genes, Type 2 genes produce an endosperm plastidial protein.

AGPase LS and SS monomer units are made up of a C-terminal β-helix domain and an N-terminal catalytic domain. The basic structural component of the catalytic domain is a Rossmann-like seven-stranded sheet that is mostly parallel but mixed, and is covered in helices. Strong hydrophobic contacts are made between the N-terminal catalytic domain and the C-terminal β-helix domain via α helix [8].

Although there is a wealth of knowledge on AGPase and its genes, there is a lack of in-depth knowledge regarding the diversity and variance in the structure, function, and evolution of the fundamental genes among various plant species.

CHAPTER 2

REVIEW OF LITERATURE

The very first committed and rate-limiting step in bacterial glycogen and plant starch production is catalysed by the enzyme ADP-glucose pyrophosphorylase. It is the enzyme site that controls the accumulation of storage polysaccharides in bacteria and plants [3]. Metabolites of energy flux allosterically activate or inhibit it. The C-terminal catalytic domain adopts a left-handed parallel b helix and is engaged in cooperative allosteric control and a special oligomerization, while the N-terminal catalytic domain mimics a dinucleotide-binding Rossman fold. The N-terminus, the glucose-1-phosphate binding site, and the ATP-binding site are just a few of the distinct regions of the enzyme which are involved in the communication between the regulator-binding sites and the active site[7]. These structures give insight on the enzyme's allosteric regulation and catalytic mechanism.

An essential enzyme for the manufacture of starch in plants is called ADP-glucose pyrophosphorylase (AGPase). There are two genes that encode homotetrameric plant AGPase: Shrunken-2 (Sh2) and Brittle-2 (Bt2). The Bt2 gene encodes a smaller subunit with catalytic properties, while the Sh2 gene encodes the larger regulatory subunit. Without altering the reading frame, a specific mutation in the Sh2 gene led to the addition of 2 extra amino acid residues—serine and tyrosine—at a specific position at the carboxyl end which caused an increase in seed weight of up to 11–17% [9]. Despite the rice and maize subunits' being 93% similar structurally, no change in seed weight was seen after the same insertion into the bigger AGPase enzyme subunit in rice. In the case of mutated maize but not in the case of mutated rice, a segment of secondary structure six amino acid long residues got decreased exactly before the point of insertion of additional amino acids (serine and tyrosine) [12]. The subtle difference in the secondary structure that results from the six-residue sequence may therefore be the main causes of the functional variance between the engineered bigger subunits in the two plants [18].

Since a protein's 3D structure determines how well it functions, the main goal of this research is to computationally model the 3D structure of the wheat LS and SS of AGPase in order to come up with a potential response to the aforementioned objection [15].

Most organisms store their carbon and energy stores as α -1,4-polyglucans, which include starch and glycogen. Utilising an activated form of glucose is necessary for the biosynthesis of

polyglucans. Eukaryotic heterotrophic microbes, mammals, and fungi use UDP-glucose (UDPGlc), while bacteria and photosynthetic eukaryotes use ADP-glucose (ADP-Glc). ADP-Glc synthesis, which is catalysed by ADP-glucose pyrophosphorylase, is the primary regulating step of the route in both plants and bacteria.[9]

ADP-Glc α and β subunits, also known as small and large subunits, respectively, combine to form an α2β2 heterotetramer in all eukaryotes that have been defined as having PPases [15]. The β component of higher plant AGPase is less conserved (50–60% identity), but the α subunit is highly conserved (85–95% identity).[10] The α and β subunits of potato tuber AGPase are 53% identical.

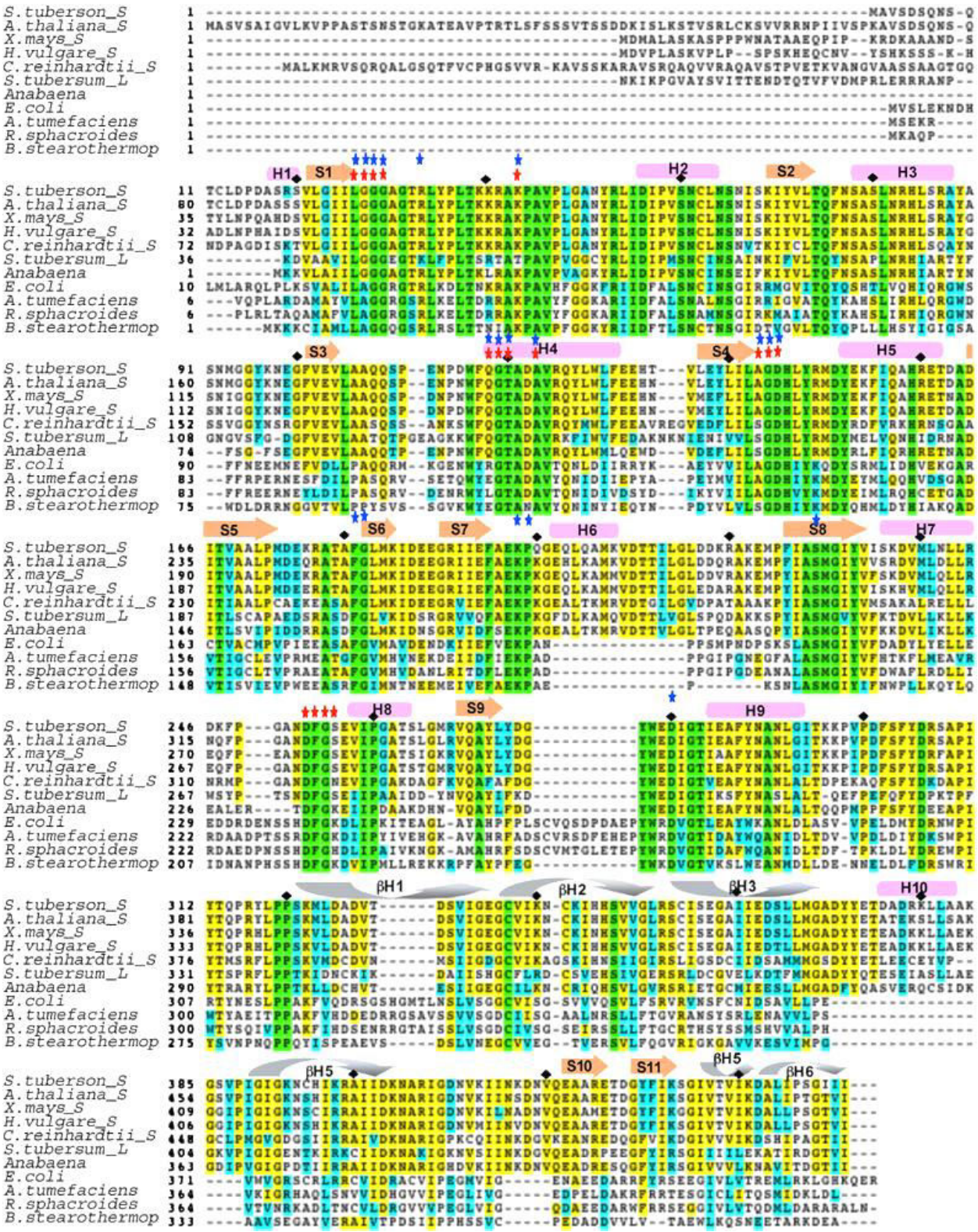


Fig 2.0.1 AGPase sequence alignment across several various species [15]. Cylinders represent helices; straight block arrows represent beta strands; curved block arrows

represent the turns in the beta helix domain; blue stars represent residues interacting with ADP-Glc; red stars represent residues interacting with ATP. For the sequence abbreviations refer to abbreviations at the beginning. All green residues are identical across the different species, yellow residues are identical across most of the sequences, and cyan residues are homologous across most of the sequences.

2.1 Identifying the "True" orthologs of the genes for ls and ss

To find "real" orthologs for the genes that encode LS and SS in seven monocots and three dicot plant species, the full-length cDNA sequences of the maize Sh2 (encoding LS) and the Bt2 (encoding SS) were utilised as references in tBLASTx.[11] (fig 2.1.1 & 2.1.2)

The following criteria were used to determine which sequences were "true" orthologs:

(i)There should be a high level of query coverage and sequence identity across the protein length.

(ii) Presence of all domains and motifs present in the original query sequence.

(iii) Conservation of the relative size and sequence interval among motifs and domains of the query sequence with those of different species.

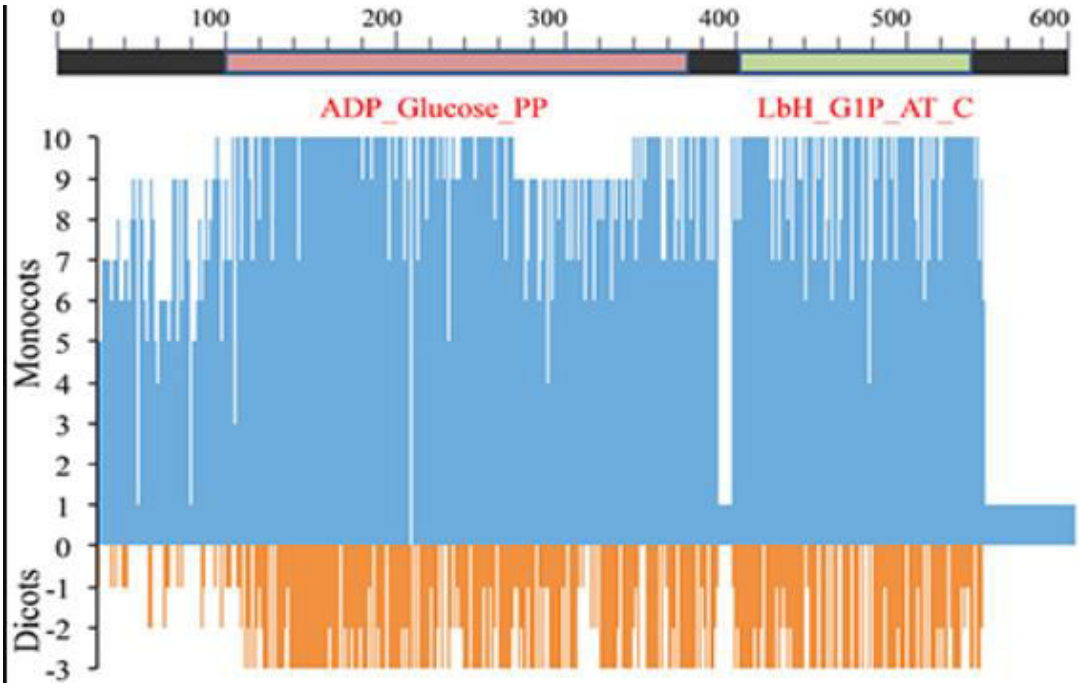


Fig 2.1. 1 Similarity of the AGPase LS amino acid sequence to the consensus sequence in 8 monocots (including wheat homoeologous on group 1 chromosomes) and 3 dicots [7]

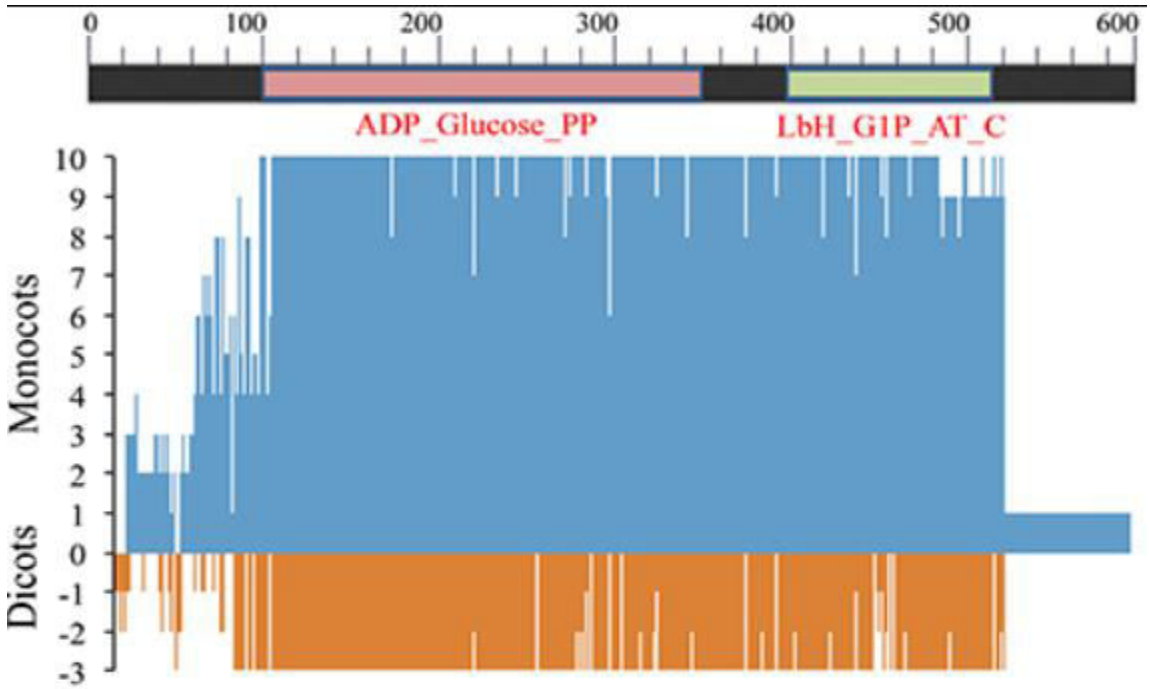


Fig 2.1. 2 Similarity of the AGPase SS amino acid sequence to the consensus sequence in 8 monocots (including wheat homoeologous on group 1 chromosomes) and 3 dicots [7]

2.2 Protein sequence analysis

Using Geneious software, multiple sequence alignment of the amino acid (aa) sequences from all the orthologs was used to create the consensus amino acid (aa) sequence for LS and SS. The consensus sequence was generated using an aa at a specific place that is present in most of the orthologs [12]. The consensus sequence was determined using aa from the maize reference if a different aa was present at a specific location. A scale of 0–10 was used to compare monocots and dicots, with each scale having a different value for the homoeologous relationship to wheat [9]. Zero meant there was no similarity at all to the consensus, whereas monocots and dicots had values of 10 and 3, respectively, which suggested that all of the species' aa had been conserved.

2.3 3D Structure Analysis of the LS and the SS

Swiss-Model was employed in an automated manner to create the 3D structures of the LS and SS from aa sequences [14]. The following servers were used to verify the 3D structures for all the species thus generated using geometric and energetic methods: (fig 2.3.1)

(a) Structure Analysis and Verification Server (SAVES) in

(i) uses PROCHECK to determine the fraction of aa that fall in the preferred region compared to other regions. [13]

(ii) VERIFY3D to assess whether an atomic model (3D) and its own aa sequence are compatible.[14]

(iii) ERRAT to examine the statistics of unbonded interactions among various atom types [15]

(b) Swiss-Model server with tool for structural evaluation.

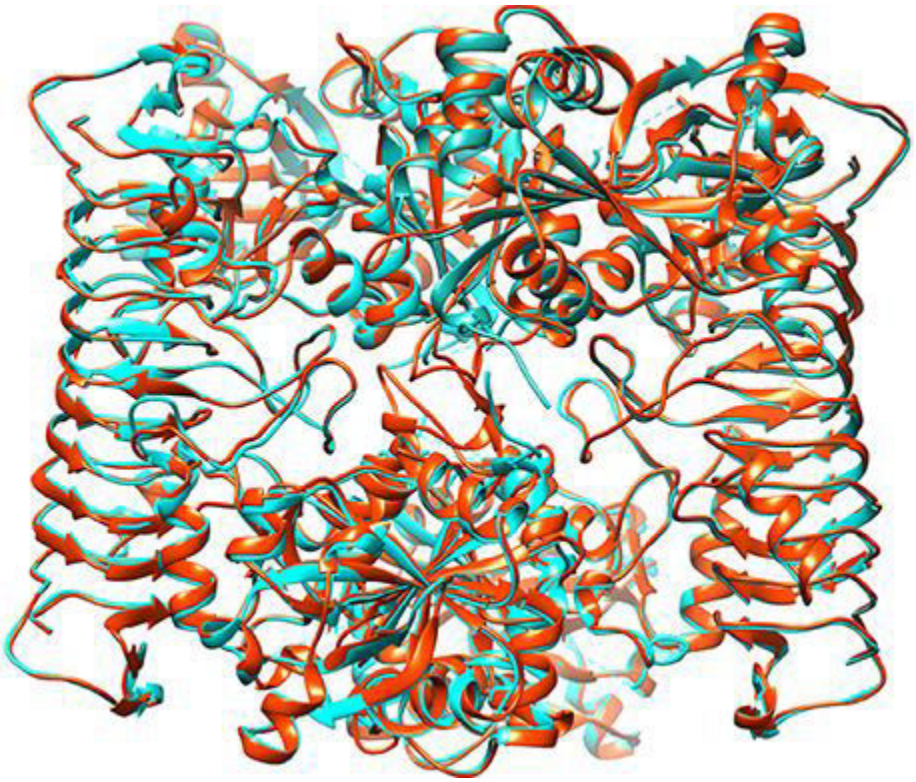


Fig 2.3. 1 Predicted superimposed Structure of Wheat AGPase (Orange Coloured) over potato AGPase (Cyan Coloured) [8]

CHAPTER 3

MATERIALS AND METHODS

3.1 Protein sequence analysis

Since no PDB model is present on wheat AGPase our first priority was to search for the Protein Sequence for the same on NCBI Protein Servers.The primary focus was to find the Protein sequences for the LS and SS of wheat AGPase along with their GenBank IDs for further analysis of these subunits. These IDs would then be used to find sequences which are similar to our target sequences for model building. BLAST results showed that potato was the sequence most similar to wheat. Our results were also backed by the previous studies on AGPase across different species which shows that the SS is generally more conserved compared to the LS.

3.2 Protein structure modelling

Since the structure of wheat AGPase was not readily available we needed to model the structure from the most closely related sequence i.e. potato. But since the structure of potato(1YP2) on the RCSB website (https://www.rcsb.org/) had gaps in chains, that structure needed to be modelled so as to get proper results of the analysis. The structure was modelled using loop modelling method on the discovery studio client. To check for the correctness in the modelled structure a 30ns MDs was run on the monomer. The studies and MDS were first done on the monomer because the target protein is a homotetramer made up of 4 identical protein monomers, meaning the results of the monomer and the tetramer should be comparable for a correctly modelled protein. Since the results of that simulation were accurate a 100ns simulation was run on the complete tetramer.

3.3 Molecular dynamics simulation

Molecular dynamics simulations were done on the monomer as well as the tetramer which gives us some detailed information about the structure and how the protein binding occurs. Simulations were carried out in a cubic box as the unit cell, with aqueous solution present in it, at a min distance of 1.0nm to satisfy the minimum image convention. The energy minimization was done to ensure that there were no steric clashes. Equilibration was carried out under 2 steps: constant number of particles, volume, and temperature (NVT) and constant number of particles, pressure, and temperature (NPT) using molecular dynamics parameters (MDP). The NVT was done at 500 ps (picosecond) for monomer and 1 ns (nanosecond) for tetramer. The NPT was done at 1 ns for

monomer and 2 ns for tetramer. Simulations were done under periodic boundary conditions (PBC) which help to avoid edge effects in the system present. At last production MD was done at 30ns for the monomer and 100 ns for the complete tetramer. Using visual molecular dynamics (VMD) software which is a visualisation tool results were analysed. For the analysis, protein was centred and PBC was removed from the system. Water was removed from the final file to visualise only the protein after the complete simulation was run on it. Xmgrace plotting tool was used for root mean square deviation (RMSD) calculation using the gmx rms command. For calculating root mean square fluctuation (RMSF) gmx rmsf/ command was used. For calculating radius of gyration (RG) gmx gyrate command is used.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Finding the protein sequence of Wheat (*Triticum aestivum*)

Due to unavailability of a modelled wheat AGPase structure we ran a search on the NCBI protein database to find the corresponding sequence for the enzyme. The results are shown in Table 4.1.1

GenBank ID	Sequence
GenBank: ABD66657.1 (Large subunit) >ABD66657.1 plastid ADP-glucose pyrophosphorylase large subunit [*Triticum aestivum*]	MDLRVAAPASVAAAARRGVLGCARVRP LQGRRQCRPSVRVSVATTESAAAAAAV AASADEDEETTNPRTVVAVILGGGAGTR LFPLTKRRAKPAVPIGGAYRLIDVPMSN CINSGINKVYVLTQFNSASLNRHLSRAY NFSNGVGFGDGFVEVLAATQRPGSEGKT WFQGTADAVRQFAWLFDDAKSKDIEDV LILSGDHLYRMDYMDFVQSHRQRDAGIS ICCLPIDGSRASDFGLMKIDDTGRVISFSE KPRGADLKAMQVDTTLLLPKEEAEKKP YIASMGVYIFKKEILLNLLRWRFPTANDF GSEIIPAAAREINVKAYLFNDYWEDIGTI KSFFEANLALAEQPSKFSFYDASKPMYT SRRNLPPSMISGSKITDSIISHGCFLDKCR VEHSVVGIRSRIGSNVHLKDTVMLGADF YETDMERGDQLAEGKVPIGIGENTSIQN CIIDKNARIGKNVTIANAEGVQESDRASE GFHIRSGITVVLKNSVIADGLVI
GenBank: AAU50665.1 (Small submit) >AAU50665.1 ADP-glucose pyrophosphorylase small subunit (plastid) [*Triticum aestivum*]	MAMATAMAATYGAPITAPAPAAFSPRR AAGGRRVRAVTARPRPLFSPRAVSDSRN SQTCLDPDASTSVLGIILGGGAGTRLYPL TKKRAKPAVPLGANYRLIDIPVSNCLNS NVSKIYVLTQFNSASLNRHLSRAYGNNI GGYKNDGFVEVLAAQQSPESPDWFQGT ADAVRQYLWLFEEHNVMEFLILAGDHL YRMDYQKFIQAHRETDADITVAALPMD EERATASGLMKIDDEGRIVEFSEKPKGEK LKAMMVDTTILGLDSERAKELPYIASMG IYVFSKDAMLRLLRDNFPSANDFGSEVIP GATEIGMRVQAYLYDGYWEDIGTIEAFY NANLGITKKPVPDFSFYDRSAPIYTQSRY LPPSKVLNADVTDSVIGEGCVINHCTINH SVVGLRSCISEGAVIEDSLLMGADYYETE NDKKVLSESGGIPIGIGKNAHIRKAIIDKN ARIGENVKIINVDDIQEASRESDGYFIKSG IVTVIKDALIPSGTVI

Table 4.1.1 Table showing the GenBank IDs of LS and SS of wheat

4.2 Searching for structures similar to wheat AGPase using BLAST

After getting the protein sequence of the LS and SS the next step is to find a structure which is similar to the target sequence. This structure obtained will be utilised as a template for model building. BLAST results showed that potato was the sequence most similar to wheat. The query coverage for the LS of potato was 86% with E-value being 2e-179 and a percentage identity of 58.85 percent. The query coverage for the SS of potato was 90% with an E-value of 0.0 and percentage identity of 86.67 percent. These results were also backed by the previous studies on AGPase across different species which shows that the SS is generally more conserved compared to the LS.

4.2.1 Running BLAST-p on the LS and SS of Wheat

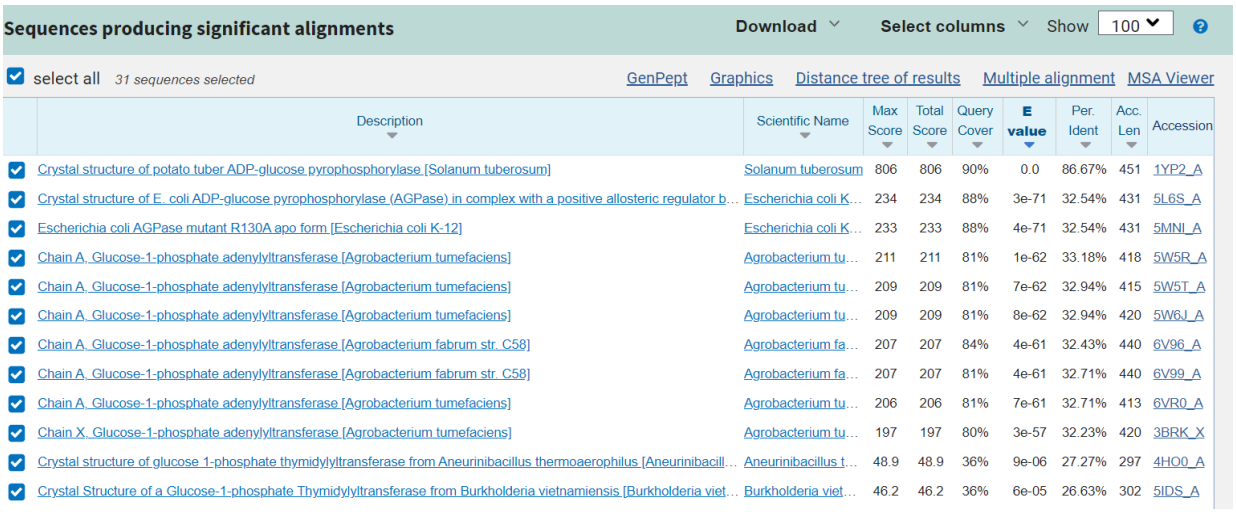
The results obtained after running BLAST on LS and SS are shown in Fig 4.2.1.1 and 4.2.1.2 respectively.

Sequences producing significant alignments Download Select columns Show 100

select all *33 sequences selected* GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Crystal structure of potato tuber ADP-glucose pyrophosphorylase [Solanum tuberosum]	Solanum tuberosum	512	512	86%	2e-179	56.85%	451	1YP2_A
Escherichia coli AGPase mutant R130A apo form [Escherichia coli K-12]	Escherichia coli K...	220	220	80%	6e-66	32.94%	431	5MNI_A
Crystal structure of E. coli ADP-glucose pyrophosphorylase (AGPase) in complex with a positive allosteric regulator b...	Escherichia coli K...	219	219	80%	1e-65	32.94%	431	5L6S_A
Chain A, Glucose-1-phosphate adenylyltransferase [Agrobacterium tumefaciens]	Agrobacterium tu...	195	195	79%	1e-56	32.38%	418	5W5R_A
Chain A, Glucose-1-phosphate adenylyltransferase [Agrobacterium tumefaciens]	Agrobacterium tu...	193	193	79%	6e-56	32.14%	415	5W5T_A
Chain A, Glucose-1-phosphate adenylyltransferase [Agrobacterium tumefaciens]	Agrobacterium tu...	193	193	79%	7e-56	32.14%	420	5W6J_A
Chain A, Glucose-1-phosphate adenylyltransferase [Agrobacterium fabrum str. C58]	Agrobacterium fa...	192	192	79%	3e-55	31.75%	440	6V96_A
Chain A, Glucose-1-phosphate adenylyltransferase [Agrobacterium fabrum str. C58]	Agrobacterium fa...	192	192	79%	4e-55	31.75%	440	6V99_A
Chain A, Glucose-1-phosphate adenylyltransferase [Agrobacterium tumefaciens]	Agrobacterium tu...	191	191	79%	5e-55	31.90%	413	6VR0_A
Chain X, Glucose-1-phosphate adenylyltransferase [Agrobacterium tumefaciens]	Agrobacterium tu...	181	181	79%	2e-51	31.19%	420	3BRK_X
Chain E, Mannose-1-phosphate guanyltransferase beta [Homo sapiens]	Homo sapiens	57.4	57.4	70%	2e-08	22.19%	360	7D72_E
Crystal structure of glucose 1-phosphate thymidylyltransferase from Aneurinibacillus thermoaerophilus [Aneurinibacill...	Aneurinibacillus t...	53.9	53.9	54%	2e-07	22.68%	297	4HO0_A
The Structural Basis Of The Catalytic Mechanism And Regulation Of Glucose-1-Phosphate Thymidylyltransferase (R...	Pseudomonas ae...	50.4	50.4	55%	3e-06	22.18%	293	1FXO_A
Pseudomonas aeruginosa RmlA in complex with allosteric inhibitor [Pseudomonas aeruginosa PAO1]	Pseudomonas ae...	50.4	50.4	55%	3e-06	22.18%	302	4ARW_A

Fig 4.2.1. A BLAST analysis of the LS of Wheat



Sequences producing significant alignments Download Select columns Show 100

select all *31 sequences selected* GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Crystal structure of potato tuber ADP-glucose pyrophosphorylase [Solanum tuberosum]	Solanum tuberosum	806	806	90%	0.0	86.67%	451	1YP2_A
Crystal structure of E. coli ADP-glucose pyrophosphorylase (AGPase) in complex with a positive allosteric regulator b…	Escherichia coli K…	234	234	88%	3e-71	32.54%	431	5L6S_A
Escherichia coli AGPase mutant R130A apo form [Escherichia coli K-12]	Escherichia coli K…	233	233	88%	4e-71	32.54%	431	5MNI_A
Chain A, Glucose-1-phosphate adenylyltransferase [Agrobacterium tumefaciens]	Agrobacterium tu…	211	211	81%	1e-62	33.18%	418	5W5R_A
Chain A, Glucose-1-phosphate adenylyltransferase [Agrobacterium tumefaciens]	Agrobacterium tu…	209	209	81%	7e-62	32.94%	415	5W5T_A
Chain A, Glucose-1-phosphate adenylyltransferase [Agrobacterium tumefaciens]	Agrobacterium tu…	209	209	81%	8e-62	32.94%	420	5W6J_A
Chain A, Glucose-1-phosphate adenylyltransferase [Agrobacterium fabrum str. C58]	Agrobacterium fa…	207	207	84%	4e-61	32.43%	440	6V96_A
Chain A, Glucose-1-phosphate adenylyltransferase [Agrobacterium fabrum str. C58]	Agrobacterium fa…	207	207	81%	4e-61	32.71%	440	6V99_A
Chain A, Glucose-1-phosphate adenylyltransferase [Agrobacterium tumefaciens]	Agrobacterium tu…	206	206	81%	7e-61	32.71%	413	6VR0_A
Chain X, Glucose-1-phosphate adenylyltransferase [Agrobacterium tumefaciens]	Agrobacterium tu…	197	197	80%	3e-57	32.23%	420	3BRK_X
Crystal structure of glucose 1-phosphate thymidylyltransferase from Aneurinibacillus thermoaerophilus [Aneurinibacill…	Aneurinibacillus t…	48.9	48.9	36%	9e-06	27.27%	297	4HO0_A
Crystal Structure of a Glucose-1-phosphate Thymidylyltransferase from Burkholderia vietnamiensis [Burkholderia viet…	Burkholderia viet…	46.2	46.2	36%	6e-05	26.63%	302	5IDS_A

Fig 4.2.1.2 B BLAST analysis of the SS of Wheat

4.2.2 Inference from BLAST analysis

Running BLAST-p on ADP-glucose pyrophosphorylase large subunit of wheat showed that the sequence is 56.85% identical to the AGPase of Solanum tuberosum. The sequence was also found to be 32.94% identical to E. coli AGPase mutant apofrom.

The abovementioned findings are consistent with the bulk of research on the subject, which indicates that the large sequence is typically less conserved than the small subunit.

Running BLAST-p on ADP-glucose pyrophosphorylase small subunit of wheat showed that the sequence is 86.67% identical to the AGPase of *Solanum tuberosum*. The sequence was also found to be 32.54% identical to E.coli crystal structure AGPase.

The abovementioned findings are consistent with the bulk of research on the subject, which indicates that the small sequence is typically more conserved than the large subunit.

4.3 Potato AGPase (PDB ID - 1YP2) the protein most similar to Wheat AGPase

Since the BLAST query returned Potato Tuber AGPase as the most structurally similar to our target sequence the next step was to analyse the Structure of potato to gain an insight into the proper functioning of the protein.

4.3.1 Inference based on literature

The potato AGPase is a homotetramer having 4 identical monomers. The different chains in the potato tuber AGPase are labelled A B C and D Each chain is made up of Approximately 451 Amino Acids. Each chain is Divided into a LS and a SS. In a single monomer of the homotetramer the LS is approximately 320 Residue long while the SS is 130 Residue long. The LS spans from AA 1 to 320 while the SS spans from AA 321 to 451.

4.4 Visualising the Potato AGpase structure in Discovery Studio

4.4.1 Structure of the LS

The Structure of the large subunit of monomeric form of Potato AGPase is shown below.

The residues in pink represent the large subunit.

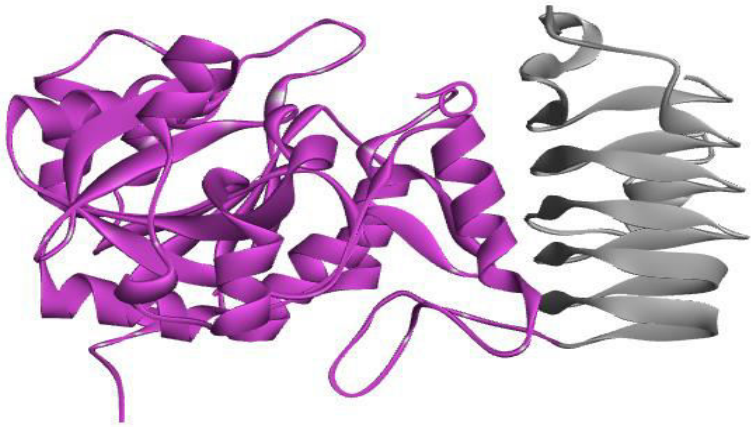


Fig 4.4.1 LS of potato monomer

4.4.2 Structure of the SS

The Structure of the Small subunit of monomeric form of Potato AGPase is shown below.

The residues in blue represent the small subunit.

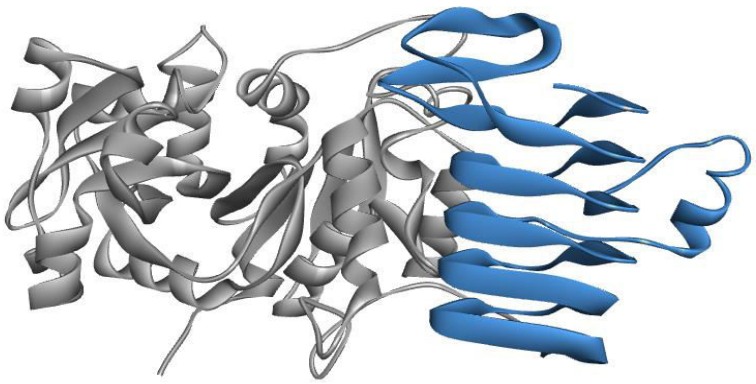


Fig 4.4. 2 SS of potato monomer

4.4.3 Complete Structure of potato AGPase

The complete homotetramer of potato AGPase showing all the 4 chains in different colours is shown below. The Structure was downloaded from rcsb.pdb and visualised on Discovery Studio Client.

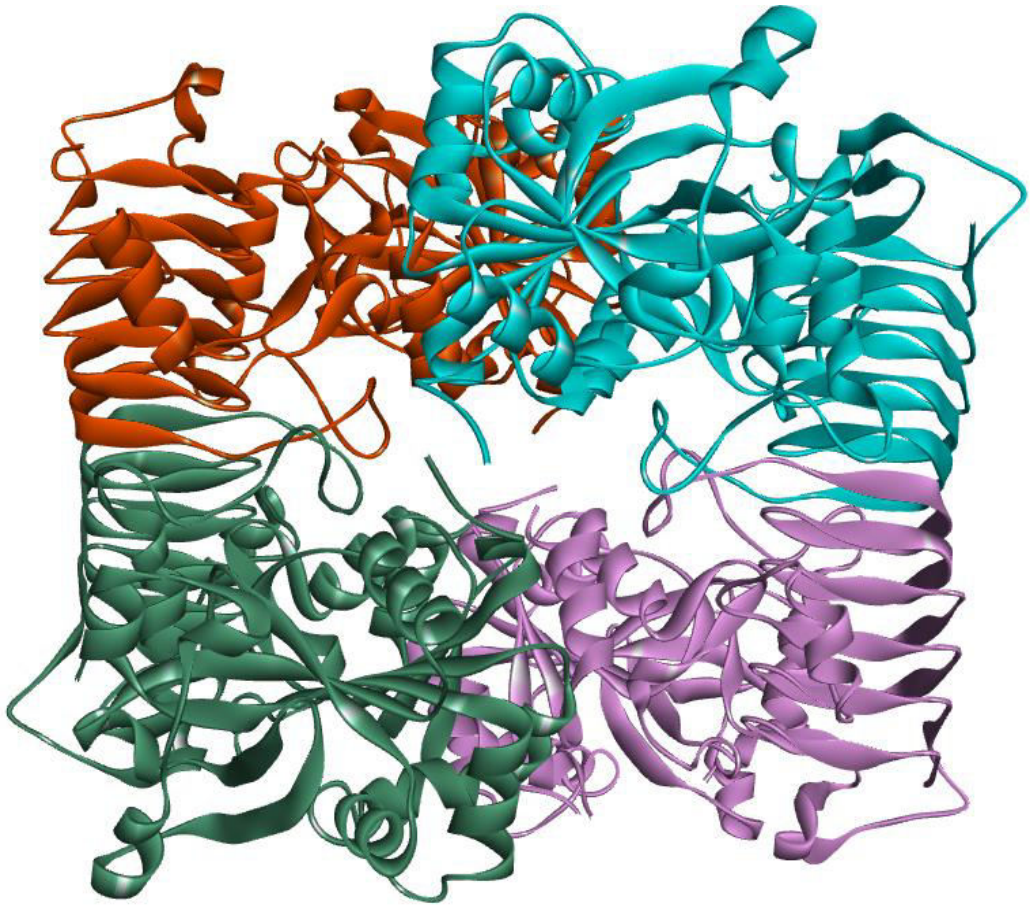


Fig 4.4.3.1 Crystal structure of potato AGPase (unmodelled), Chain A : Blue Coloured, Chain B : Pink Coloured, Chain C : Green Coloured, Chain D : Red Coloured

4.5 Running MD simulations on a single monomer of 1YP2

Since the potato AGPase is a homotetramer made up of 4 identical chains / monomers the first ideal step would be to run the MD simulations on a single monomer. This step is necessary for 2 important reasons;

- First, the protein structure of potato AGPase (1YP2) available on RCSB.pdb is not a complete model and the structure has gaps which need to be filled for proper MDS and analysis.

- Second, since the protein is a homotetramer the results obtained from the MD simulations of the monomer and the tetramer should be comparable, from which we can infer that the protein that was modelled is accurate and can be used for further analysis.

The MDS performed on the protein is done on the LINUX command line using GROMACS 2021.

4.5.1 Analysis of MDS performed on the monomer of potato AGPase

The molecular dynamic simulations provide a dynamic behaviour of the system as a function of time (potato AGPase in this case) which helps us to know the behaviour of the protein and how that protein changes over time. The monomer of 1YP2 was subject to a 1ns NVT and NPT equilibration with the production MD of 30 ns.

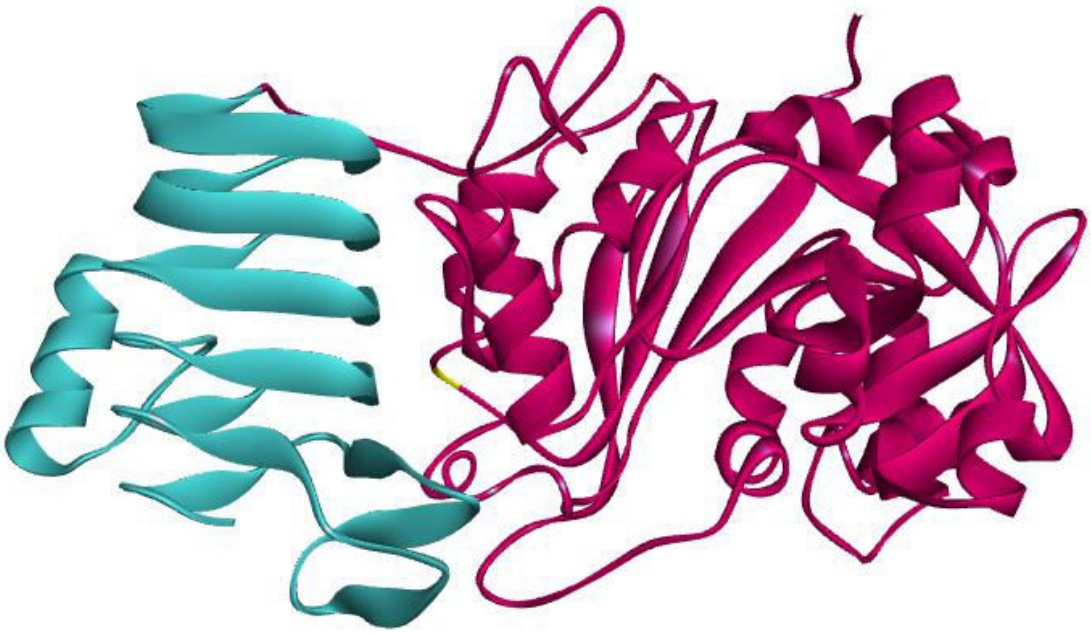


Fig 4.5.1.1 The modelled monomer of 1YP2 used in MDS (LS in pink and SS in cyan)

4.5.2 Observations and inference of the results

The. xvg,. xtc and the trajectory files generated from the MDS are used to analyse the protein The RMSD (Root Mean Square Deviation) RMSF (Root Mean Square

Fluctuation) RG (Radius of Gyration) and the Potential energy graphs of the system were generated after running a production MD of 30 ns on the system.

4.5.2.1 Potential Energy

The potential energy of the system is a combination of all the inter as well as intra atomic forces such as bond stretching, angle bending, van der Waals and electrostatic forces. A stable system should ideally have a low and constant PE for the entire duration of the simulations. Large fluctuations and drifts in the PE usually means that the system is unstable whereas an increasing PE means that the system is not properly equilibrated and there are errors in the input parameters. The PE of our system for the entire duration of the simulations was relatively constant and remained stable with little to no fluctuations. The initial decrease in PE is due to the system coming to the most stable configuration (Lower the PE more stable the system)

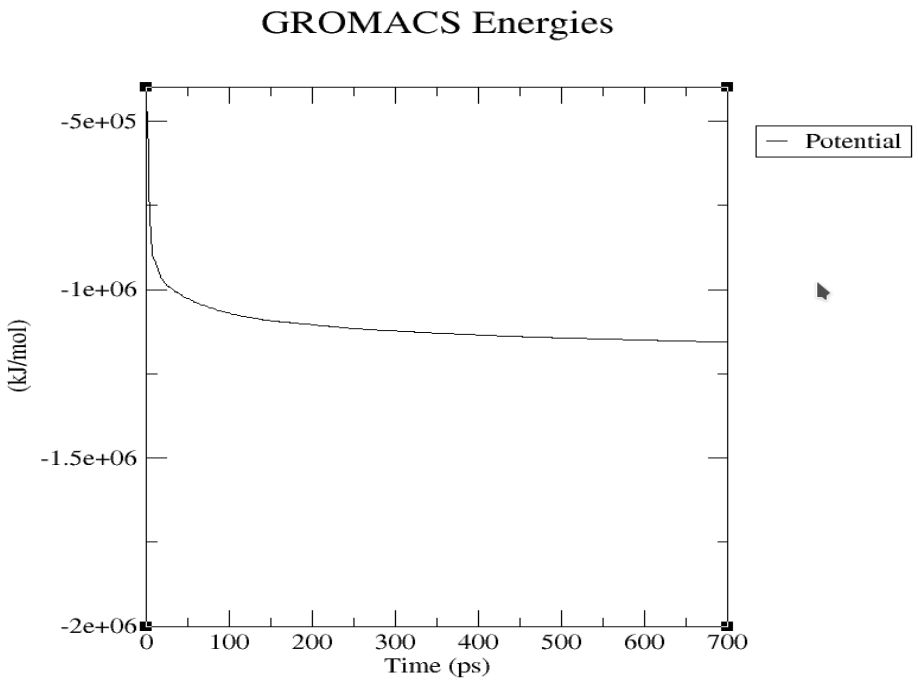


Fig 4.5.2.1 Potential energy graph of the system

4.5.2.2 RMSD

The RMSD is used to analyse the stability of simulated protein structure . It identifies the flexible regions and structural changes occurring during simulation. RMSD is generally used to compare the structures of simulations and to validate methods. Acceptable range of RMSD value is between 0.1-0.5nm. The RMSD value of the protein for the entire duration of the simulation was in the range of 0.25nm and 0.36nm. The initial increase in RMSD values for a few nanoseconds indicates the initial adjustment of the system towards a least potential energy state.

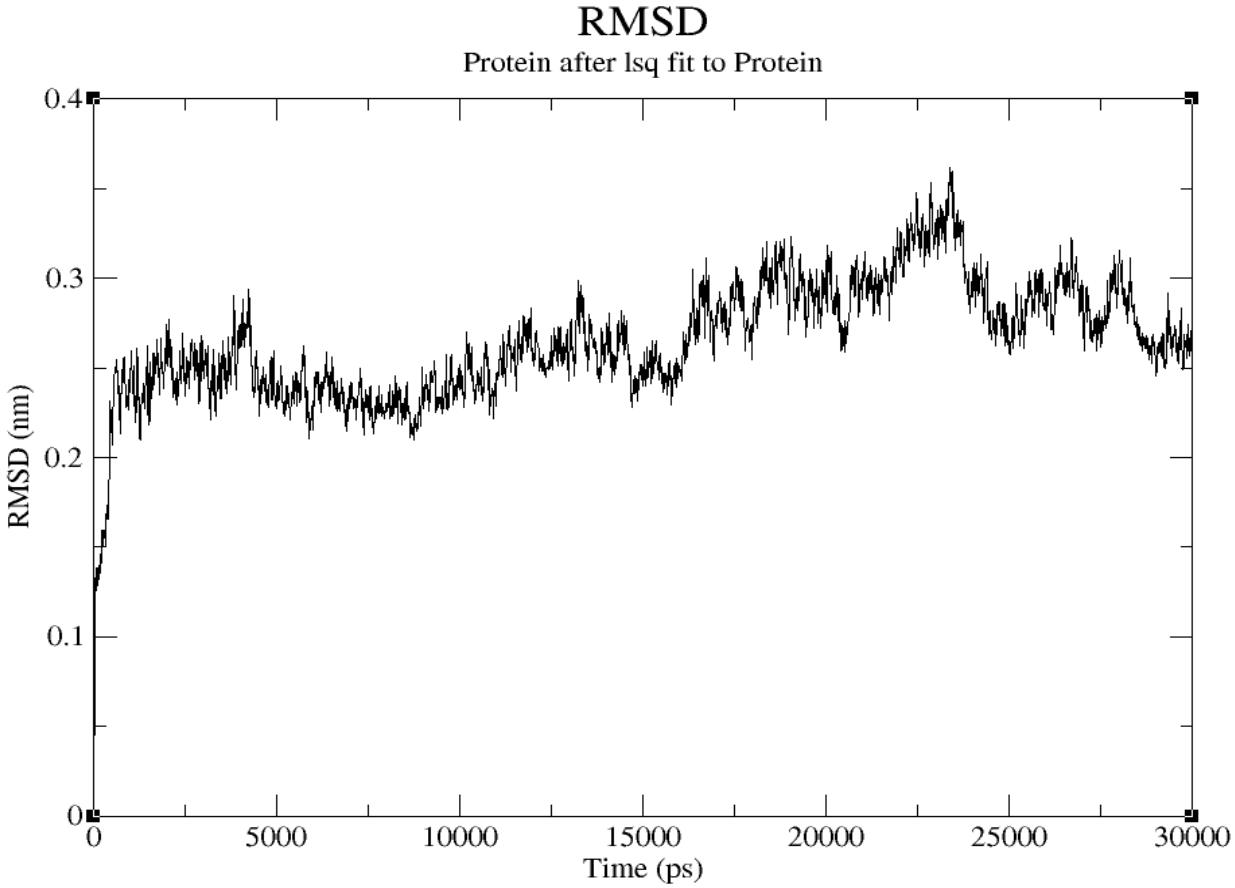


Fig 4.5.2.2 Graph of RMSD(nm) vs Time(ps)

4.5.2.3 RMSF

Root mean square fluctuation gives us average deviation of atoms in a protein from their average position in a simulation and is used to analyse the flexibility and dynamics of protein or of the system. This RMSF can help in understanding the structural basis and also can be used to see the effect of the mutations. If the RMSF is less than 0.5nm the structure is considered to be rigid whereas RMSF value greater than 1nm is considered a standard for a flexible structure. The RMSF value of the protein for the entire duration of the simulation is less than 0.5nm meaning that the protein is rigid and not undergoing major conformational changes , such as a protein domain becoming more flexible due to a binding event. From the below given RMSF graph we can also infer that the protein simulations are stable and there are no errors in the input parameters.

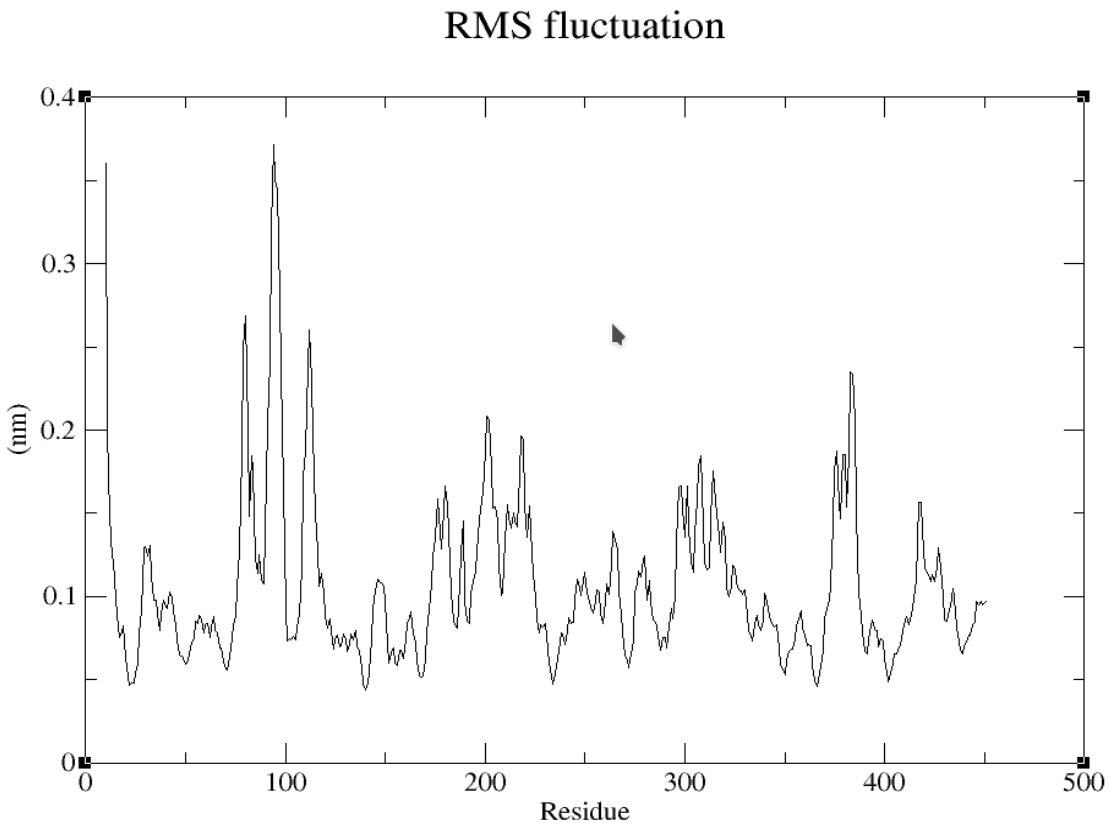


Fig 4.5.2.3 Graph of RMSF Fluctuation

4.5.2.4 Radius of Gyration

The radius of gyration (RG) is the spatial distribution of the systems' mass / charge around the centre of mass. The RG graphs can be used to assess the compactness of the molecules. A graph with low RG generally means that the system is compact and higher RG indicates extended or unfolded system. Throughout the duration of the simulation if the RG is increasing, we can infer that the system is becoming more extended/unfolded whereas a decrease in RG means that the system is becoming compact. Throughout the duration of our simulation the RG is maintaining a steady state between 2.37 and 2.45 with no significant fluctuations. We can infer from those results that the system is not undergoing any compactness or extended change after the simulation is getting completed.

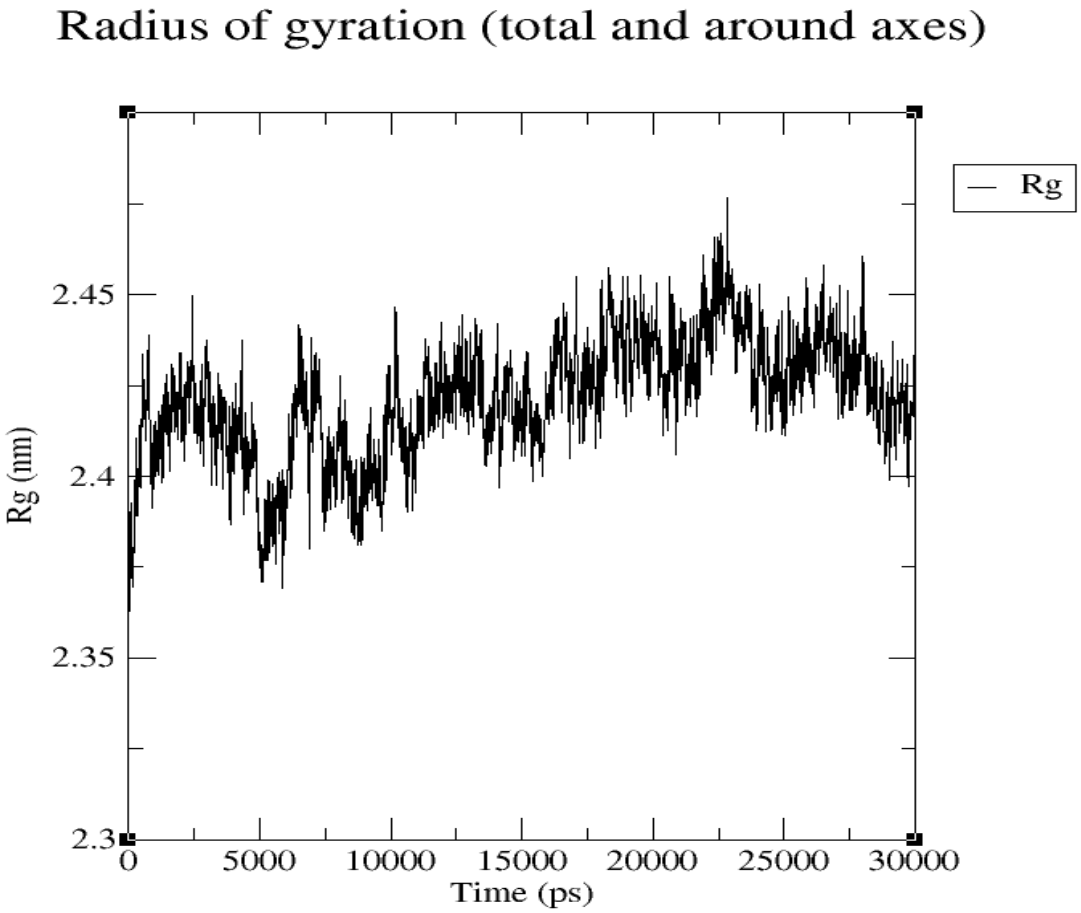


Fig 4.5.2.4 Graph of radius of gyration of the system

4.5.3 Structural comparison between 1YP2 and modelled tetramer of potato AGPase

From the results of the monomer MD simulation analysis we can infer that the protein is not undergoing any kind of significant changes and the input parameters are also accurate. Hence the tetramer was modelled using 1YP2 as the crystal structure and the same was visualised using Discovery Studio Client.

The structure was visualised two ways:

- First, by superimposing the two tetramers on top of one another.
- Second, by a side comparison which shows the modelled parts along with the gaps in the protein.

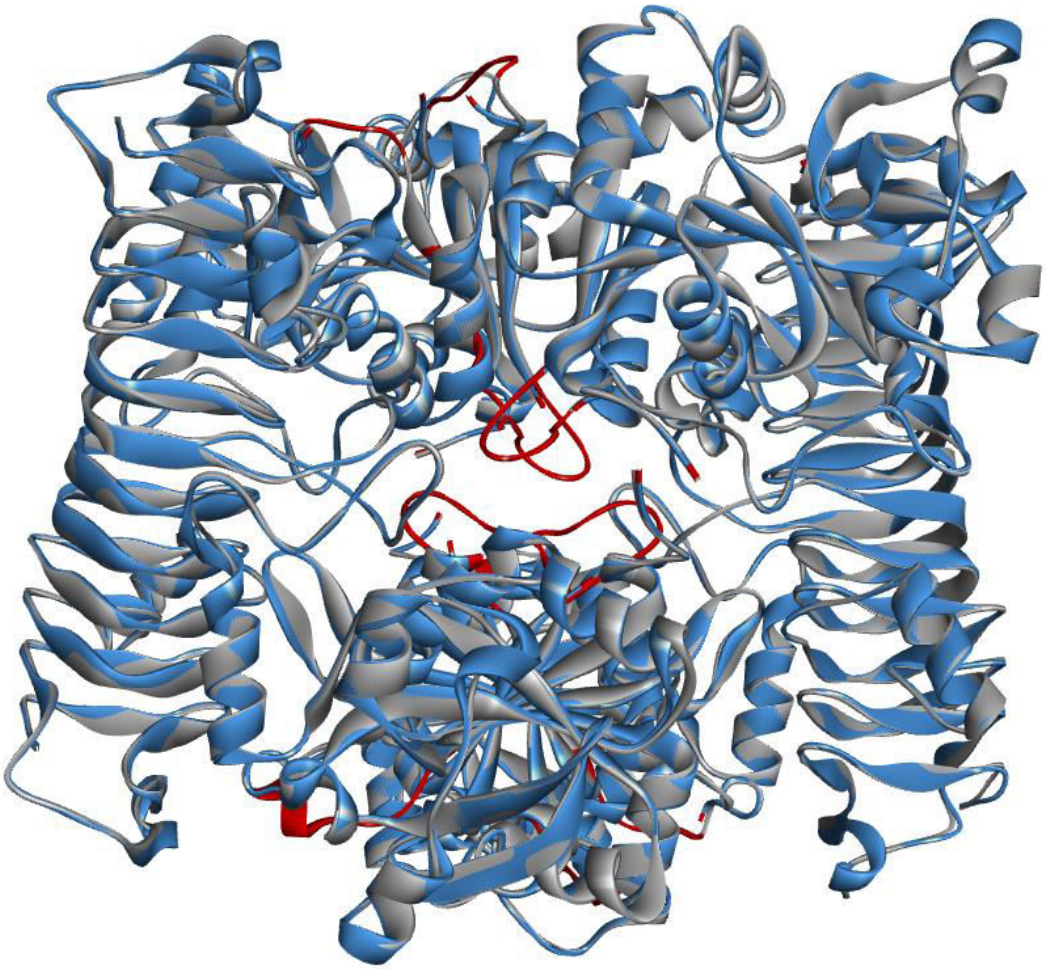


Fig 4.5.3.1 A superimposed 1YP2 and modelled protein tetramer

The potato AGpase model downloaded from rcsb.pdb is shown in blue colour. The modelled protein is coloured grey. The part of the protein coloured red shows the gaps that were present in the protein and were modelled for proper MD Simulation analysis.

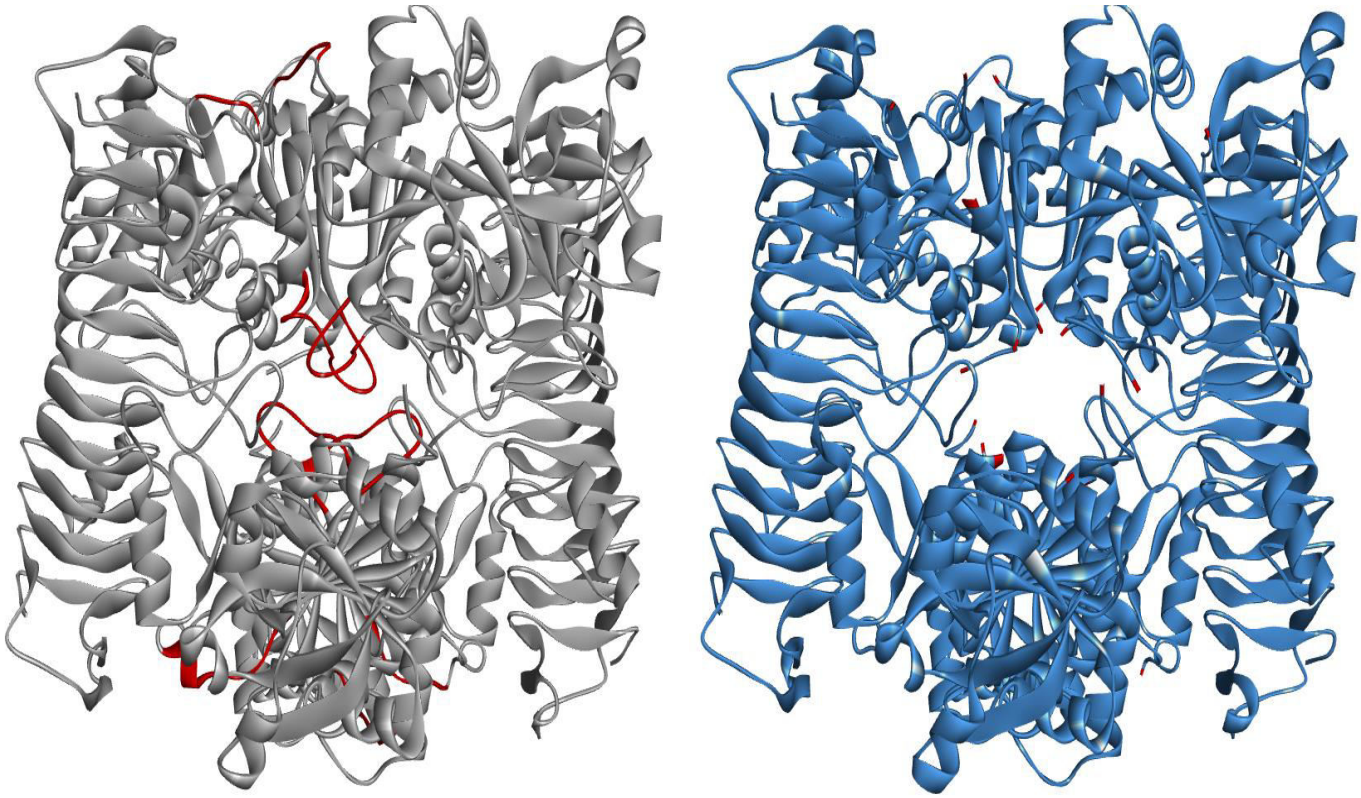


Fig 4.5.3.2 A side by side representation of the modelled protein (grey) with the potato crystal structure 1YP2 (blue)

Through this side-by-side representation we can see the gap in the protein that was modelled for getting accurate results from MD Simulations.

4.6 Running MD Simulations on the entire tetramer of modelled protein

Since the results of our monomer simulations showed no anomaly, we could conclude from that, that the modelled structure was accurate. Hence the MDS on the entire tetrameric protein could be performed and the results of the same were used for production MD of 100ns. The MDS performed on the protein is done on the LINUX command line using GROMACS 2021.

4.6.1 RMSD

The RMSD is used to analyse the stability of simulated protein structure . It identifies the flexible regions and structural changes occurring during simulation. RMSD is generally used to compare the structures of simulations and to validate methods. Acceptable range of RMSD value is between 0.1-0.5nm. The RMSD value of the protein for the entire duration of the simulation was in the range of 0.19nm and 0.38nm. The initial increase in RMSD values for a few nanoseconds indicates the initial adjustment of the system towards a least potential energy state. The RMSD of the entire system is shown in fig 4.6.1.1

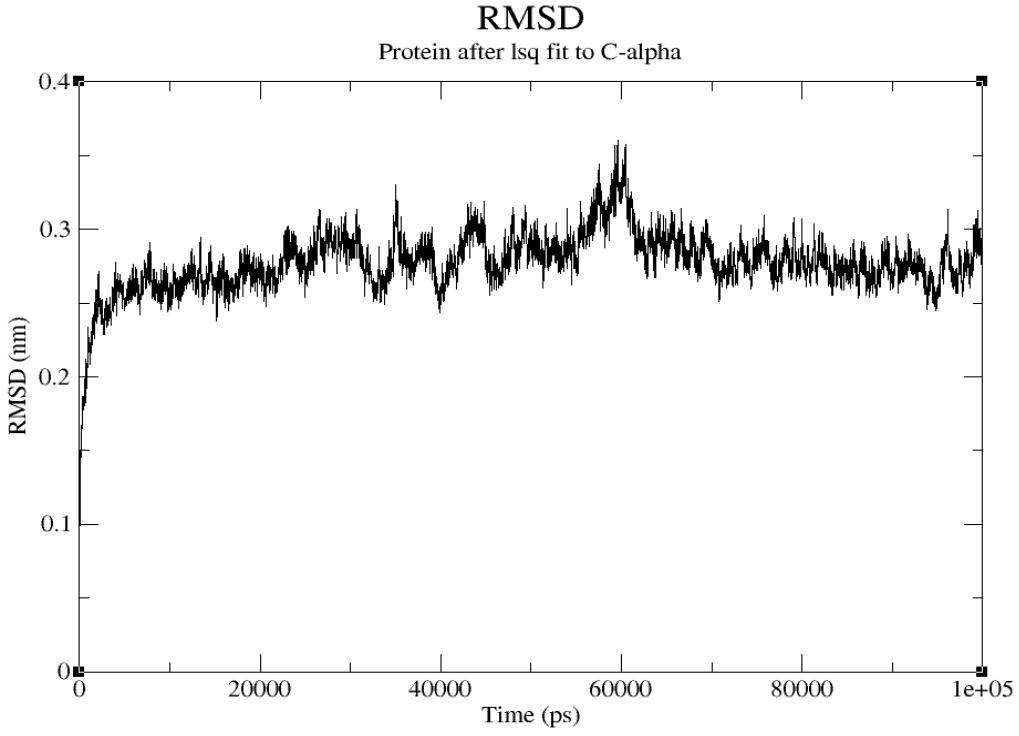


Fig 4.6.1.1 Graph of RMSD of the complete tetramer

4.6.2 2-D projection of trajectory

The graph of a 2D projection of a trajectory can be used to learn about the motion and behaviour of the system under study during a molecular dynamics simulation. It allows one to observe and analyse the behaviour of the atoms or molecules inside the simulation. The trajectory graph can provide details about the system's overall motion. It can show translational motion, for instance, if the graph's points move in a particular direction, rotational motion, for example, if the points move in a circular pattern, or complex motion patterns, for example, if the points move in an intricate way.

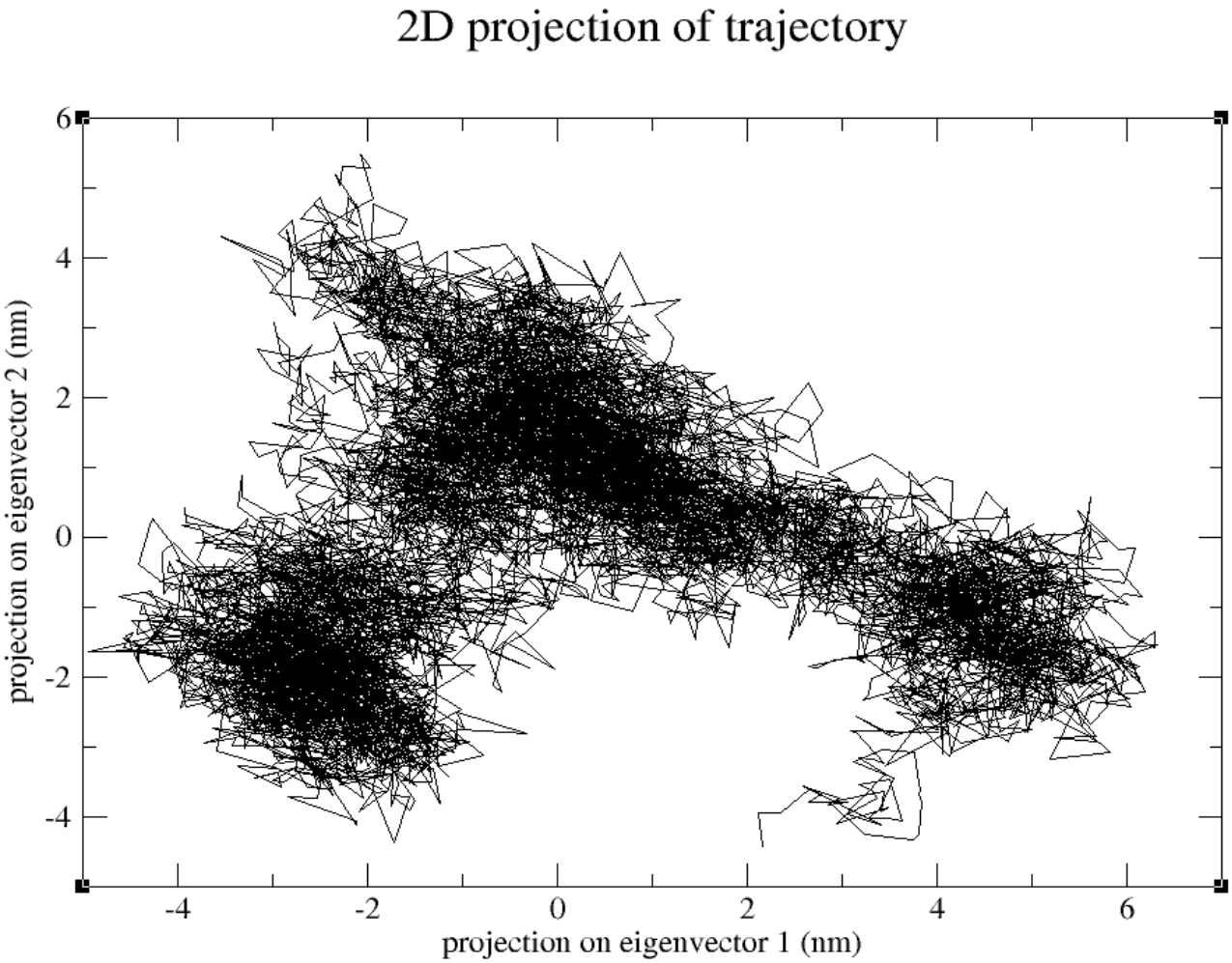


Fig 6.5.2.1 Graph of 2-D projection of trajectory

4.6.3 Average Structure Calculation

A representative conformation or arrangement of atoms that characterises the overall shape and arrangement of the protein in its original state is referred to as the average structure of a protein. It is an approximation of the protein's structure, considering the inherent flexibility and dynamics of biomolecules. To calculate average structure using GROMACS involves numerically solving the equations of motions of atoms of a protein.

From the graph of RMSD we calculated the frame which had the least RMSD and that frame was used for average structure calculation. The average structure obtained is shown in the below figure 4.6.3.1.

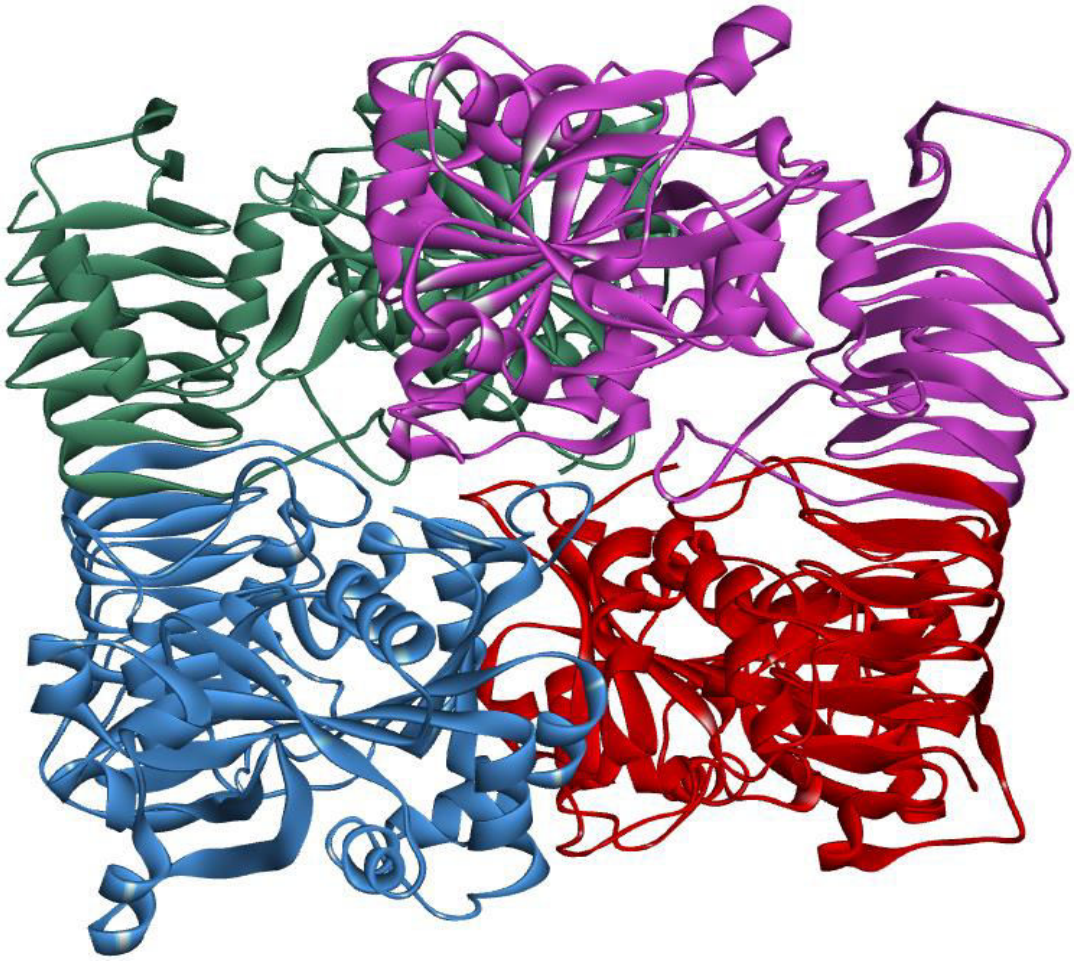


Fig 4.6.3.1 Average structure of the modelled protein obtained after PCA analysis

4.6.4 Homology Modelling

Homology modelling basically refers to building a 3-D model of the protein from the amino acid sequence. It is one of the most accurate computational methods used for structure prediction of the proteins. For the modelled protein used in our research we did a homology modelling using DS and created 10 homology models. The basic steps followed in homology modelling were aligning the sequences followed by building the homology models. For the parameters no. of models was set to 10 and cut overhang to TRUE. The model resolution was set to low and then the models were obtained. The best model among that was used for superimposition. Superimposition was done with the modelled potato crystal structure on the Discovery studio. The results of the same are shown in the Table 4.6.4.1 And Fig 4.6.4below. After getting the homology modelling results a Ramachandran plot was obtained on the DS (fig 4.6.4.3). From this plot we can conclude that the protein structure that was modelled for the crystal structure of potato using ab-initio methods is a good structure as most of the set of torsion angles lie in the allowed areas and only some torsion angles were in the forbidden region.

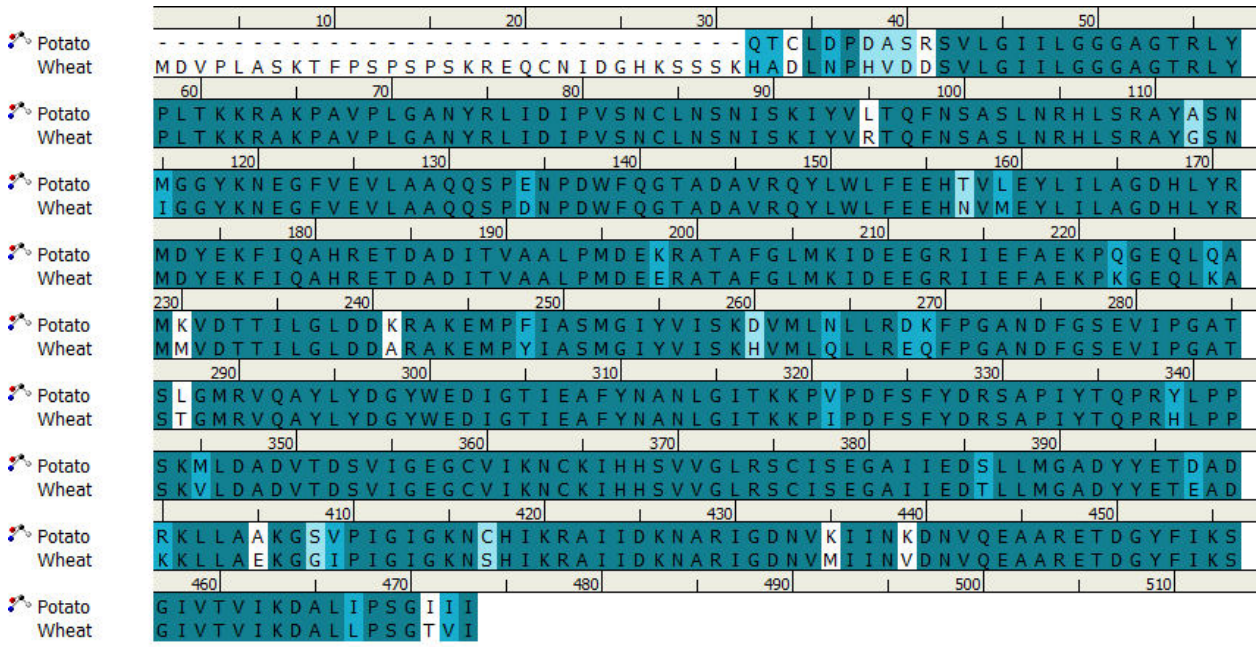


Fig 4.6.4.1 Alignment between the template (potato) and the target (wheat)

Model Scores			
Name	**PDF Total Energy**	**PDF Physical Energy**	**DOPE Score**
Wheat.M0008	2628.2756	1395.415319	-52705.917969
Wheat.M0009	2721.4358	1389.973568	-52717.925781
Wheat.M0010	2745.8857	1381.26676761	-52488.890625
Wheat.M0001	2778.0459	1390.6611018	-52770.359375
Wheat.M0003	2782.2280	1416.83010331	-52577.820313
Wheat.M0005	2783.2446	1393.35503906	-52652.054688
Wheat.M0007	2786.4458	1402.2183594	-52501.761719
Wheat.M0002	2790.3064	1391.44755455297	-52576.253906
Wheat.M0004	2806.1287	1402.9541788	-52700.531250
Wheat.M0006	2855.1318	1419.20553	-52643.183594

Table 4.6.4.1 Table showing model scores of homology modelling

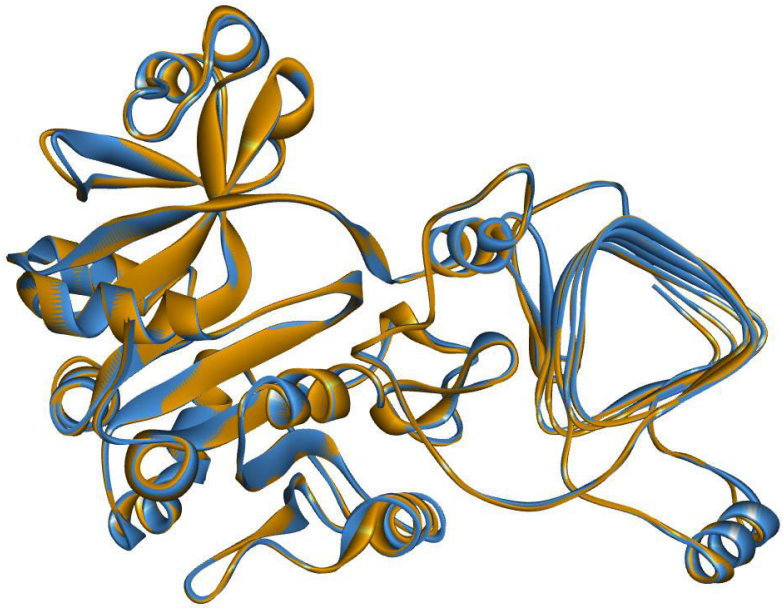


Fig 4.6.4.2 Superimposed structures of the best fit homology model of the target i.e., wheat (in blue) over the template i.e., potato (in brown)

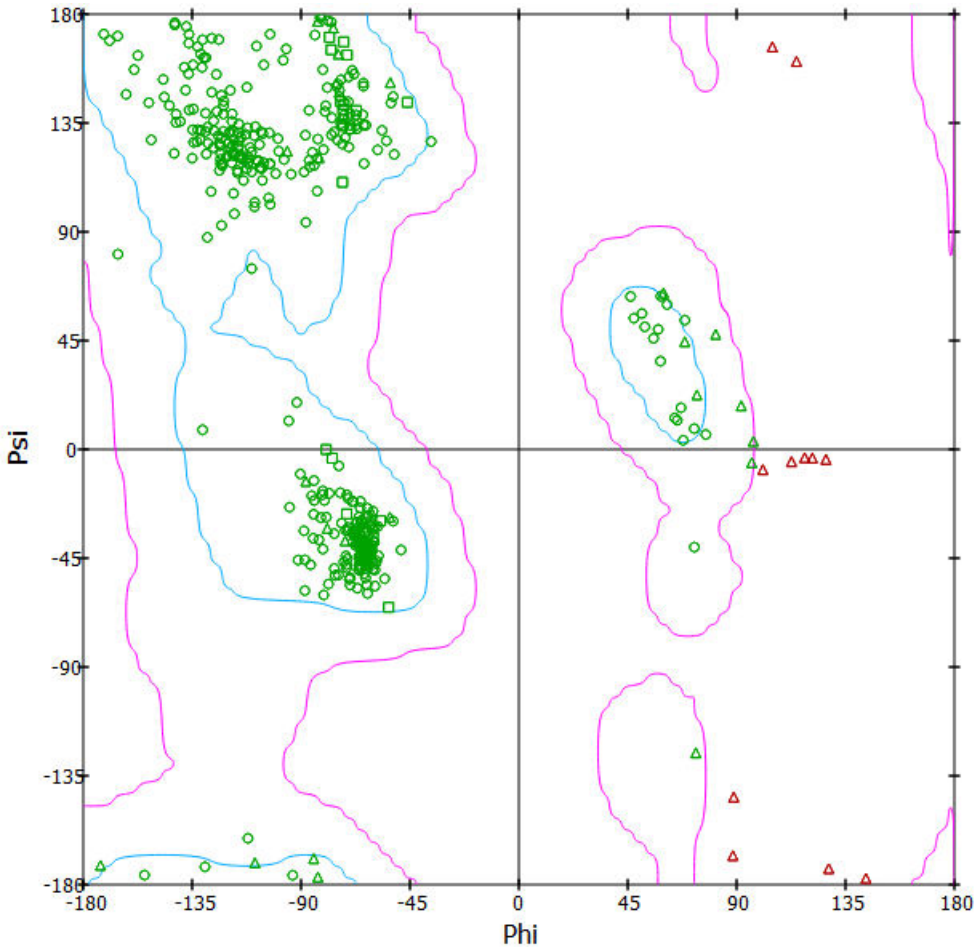


Fig 4.6.4.3 The Ramachandran plot obtained after homology modelling. Most of the torsion angles are in the allowed region hence the structure modelled is of good quality.

CHAPTER 5

CONCLUSION

5.1 Conclusion

Starch is a complex carbohydrate and plays an important role in a healthy and balanced diet. Starch is considered a staple ingredient in more than fifty percent of the peoples' diet across the globe. AGPase being the first rate limiting enzyme in starch biosynthesis plays an important role in the amount of starch produced by the plants.

In this project we did analysis on the AGPase so as to model the crystal structure of AGPase specific to wheat. Since the wheat AGPase was most similar to the potato AGPase, our first goal was to run a MDS on the potato AGPase. But since the potato crystal structure had gaps those gaps had to be modelled before running any kind of MDS to obtain accurate results. The MDS was run both on the potato monomer as well as the tetramer. The results of both the simulations showed no abnormal behaviour in the in-vitro analysis of the modelled structure. Hence taking that modelled crystal structure as the template homology modelling was done to obtain the crystal structure of our target protein i.e., wheat AGPase. To validate the homology modelling results superimposition between the potato and wheat AGpase was done on DS. A Ramachandran plot was also generated to further validate the modelled structure. The results of all the project work showed that the modelled crystal structure is accurate and the homology model obtained from those results can be used in the MDS specific to wheat.

5.2 Future Prospects

- Comparative protein modelling is thought to be one of the most accurate ways to develop models and comprehend function in the absence of experimental 3D structures.
- Based on the literature, we intend to create a 3D model of wheat AGPase LS and SS that will be utilised to gain insight into the homotetrameric structure of wheat AGPase.
- Following the creation of the 3D model, MD simulations utilising Gromacs will be used for refining.
- Choosing mutants after MDS.
- Mutagenesis done in-silico.
- Analysis of Mutants' structure and function.

References

1. Brown, W. H.; Poon, T. (2005). Introduction to organic chemistry (3rd ed.). Wiley. p. 604. ISBN 978-0-471-44451-0.

2. Nelson, D. (2013) Lehninger Principles of Biochemistry, 6th ed., W.H. Freeman and Company.

3. Smidansky, E. D., Martin, J. M., Hannah, L. C., Fischer, A. M., and Giroux, M. J. (2003). Seed yield and plant biomass increases in rice are conferred by deregulation of endosperm ADP-glucose pyrophosphorylase. Planta 216, 656–664. doi: 10.1007/s00425-002-0897-z

4. Soliman, A., Ayele, B. T., and Daayf, F. (2014). Biochemical and molecular characterization of barley plastidial ADP-glucose transporter (HvBT1). PLoS ONE 9:e98524. doi: 10.1371/journal.pone.0098524

5. Boehlein, S. K., Shaw, J. R., Stewart, J. D., Sullivan, B., and Hannah, L. S. (2015). Enhancing the heat stability and kinetic parameters of the maize endosperm ADP-glucose pyrophosphorylase using iterative saturation mutagenesis. Arch. Biochem. Biophys. 568, 28–37. doi: 10.1016/j.abb.2015.01.008

6. Hannah, L. C., Shaw, J. R., Giroux, M. J., Reyss, A., Prioul, J. L., Bae, J. M., et al. (2001). Maize genes encoding the small subunit of ADP-glucose pyrophosphorylase. Plant Physiol. 127, 173–183. doi: 10.1104/pp.127.1.173

7. Lee, S. K., Hwang, S. K., Han, M., Eom, J. S., Kang, H. G., Han, Y., et al. (2007). Identification of the ADP-glucose pyrophosphorylase isoforms essential for starch synthesis in the leaf and seed endosperm of rice (Oryza sativa L.). Plant Mol. Biol. 65, 531–546. doi: 10.1007/s11103-007-9153-z

8. Rani, S., Sharma, P., Sharma, A., and Chatrath, R. (2013). Comparative computational analysis of ADP Glucose pyrophosphorylase in plants. Bioinformation. 9, 572–576. doi: 10.6026/97320630009572.

9. Preiss J, Shen L, Greenberg E, Gentner N (1966) Biosynthesis of bacterial glycogen. IV. Activation and inhibition of the adenosine diphosphate glucose pyrophosphorylase of Escherichia coli B00. Biochemistry 5: 1833–1845

10. Nakata PA, Greene TW, Anderson JM, Smith-White BJ, Okita TW, Preiss J (1991) Comparison of the primary sequences of two potato tuber ADP-glucose pyrophosphorylase subunits. Plant Mol Biol 17: 1089–1093

11. Dhaliwal, A. K., Mohan, A., and Gill, K. S. (2014). Comparative analysis of ABCB1 reveals novel structural and functional conservation between monocots and dicots. Front. Plant Sci. 5:657. doi: 10.3389/fpls.2014.00657

12. Navarro, C., Moore, J., Ott, A., Baumert, E., Mohan, A., Gill, K. S., et al. (2015). Evolutionary, comparative and functional analyses of the brassinosteroid receptor gene, BRI1, in wheat and its relation to other plant genomes. PLoS ONE 10:e0127544. doi: 10.1371/journal.pone.0127544

13. Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. 26, 283–291. doi: 10.1107/S0021889892009944

14. Eisenberg, D., Lüthy, R., and Bowie, J. U. (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol. 277, 396–404. doi: 10.1016/S0076-6879(97)77022-8

15. Colovos, C., and Yeates, T. O. (1993). Verification of protein structures: patterns of non-bonded atomic interactions. Protein Sci. 2, 1511–1519. doi: 10.1002/pro.5560020916

16. V. Verma, S. Agrawal, A. Kumar, and J. P. Jaiswal, “Starch content and activities of starch biosynthetic enzymes in wheat, rice and millets,” Journal of Pharmacognosy and Phytochemistry, vol. 9, no. 4, pp. 1211–1218, Jan. 2020, doi: 10.22271/phyto.2020.v9.i4q.11883.

17. G. Yunsong et al., “Expression and regulation of genes involved in the reserve starch biosynthesis pathway in hexaploid wheat (*Triticum aestivum* L.),” Crop Journal, vol. 9, no. 2, pp. 440–455, Apr. 2021, doi: 10.1016/j.cj.2020.08.002.

18. W. Adly et al., “Somaclonal Variation for Genetic Improvement of Starch Accumulation in Potato (*Solanum tuberosum*) Tubers,” Plants, vol. 12, no. 2, p. 232, Jan. 2023, doi: 10.3390/plants12020232.

19. X. Ma, X. Ouyang, D. Liu, and A. Zhang, “The 218th amino acid change of Ser to Ala in TaAGPS-7A increases enzyme activity and grain weight in bread wheat,” Crop Journal, vol. 11, no. 1, pp. 140–147, Aug. 2022, doi: 10.1016/j.cj.2022.07.006