Deciphering putatively transmembrane and secreted human gut microbial Prolyl endopeptidases: Therapeutic implications for Celiac Disease

A Project Report

submitted by

KAVYA BAI O P

MT23247

in partial fulfilment of the requirements

for the award of the degree of

MASTER OF TECHNOLOGY

COMPUTATIONAL BIOLOGY

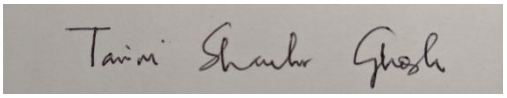
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

24-07-2025

THESIS CERTIFICATE

This is to certify that the thesis titled **Deciphering putatively transmembrane and secreted human gut microbial Prolyl endopeptidases: Therapeutic implications for Celiac Disease** submitted by **KAVYA BAI OP**, to the Indraprastha Institute of Information Technology Delhi (IIIT-Delhi) for the award of the degree of **MASTER OF TECHNOLOGY**, is a bona fide record of the research work done by her under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

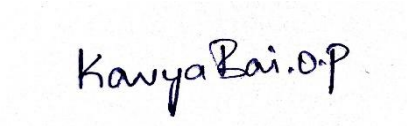


Dr. Tarini Shankar Ghosh
Thesis Supervisor
Assistant Professor
Dept. of Computational Biology
IIIT Delhi, 110020

Place: New Delhi

ACKNOWLEDGEMENTS

I extend my sincere gratitude to Dr. Tarini Shankar Ghosh for his invaluable guidance, and to our dedicated team for their technical assistance and collaboration. Special thanks to the institution for providing the resources and environment necessary for this project's success. With their support, I've embarked on a journey of discovery, aiming to make a meaningful impact in our field.



Abstract

Celiac disease (CeD) is a chronic immune-mediated enteropathy triggered by the ingestion of gluten, a storage protein complex found in wheat, rye, and barley. Gluten comprises glutenin and gliadins, which are rich in proline and glutamine residues. The high proline content makes these proteins resistant to degradation by human gastrointestinal and brush-border proteases. While a strict gluten-free diet is currently the primary treatment for CeD, maintaining such a diet is often challenging and can significantly impact patients' quality of life. Therefore, alternative or supplementary therapeutic approaches are urgently needed.

Certain gut-associated microbes have been reported to produce Prolyl Endopeptidases (PEPs), which can hydrolyze gluten peptides. In this study, we screened 10,903 bacterial PEPs from diverse taxa including Firmicutes, Bacteroidetes, Proteobacteria, Cyanobacteria, Fungi, and Animals. Using HMMER, these sequences were compared against curated reference databases comprising 115,692 transmembrane proteins and 40,344 secreted proteins from 169 gut-associated species. This analysis identified 372 transmembrane and 1,250 secreted PEP homologs. Among them, 31 transmembrane and 36 secreted PEPs from animal-associated species were selected for further analysis due to their closer similarity to human enzymes.

The PEP from Sphingomonas capsulata—a component of the oral enzyme therapy latiglutenase—was used as a reference. Through comparative sequence and domain analysis, structural modeling, and molecular docking, ten PEPs (4 transmembrane and 6 secreted) were identified with high structural and functional similarity. Docking studies showed strong binding affinity to gluten peptides, with scores ranging from −7.2 to −10.2 kcal/mol.

Furthermore, a prevalence analysis using five publicly available gut microbiome datasets comprising 414 samples (CeD and controls) revealed that the significant species harboring these enzymes were notably less abundant in CeD patients. This suggests a potential association between disease progression and the depletion of gluten-degrading microbial species.

Together, our findings highlight a set of microbial enzymes with substantial gluten-degrading potential, which are underrepresented in individuals with Celiac disease. These enzymes present promising candidates for the development of next-generation enzyme-based therapies for CeD management.

Keywords: Celiac Disease, Prolyl Endopeptidases (PEPs), Gluten intolerance, Transmembrane Proteins, Signal peptides, Enzyme therapeutics.

Content

1	**Introduction**	8-9
2	**Methods**	10
2.1	Literature review	10
2.2	Methods	10
2.2.1	Identifying of PEPs which has potential to degrade gluten	12
2.2.2	MSA of significant species with the Reference and Structure Prediction	13
2.2.3	Docking Analysis of Inhibitor and Substrate	13
2.2.4	Prevalence of significant species in Celiac-Control Study	14
3	**Results**	14
3.1	Homologs Results: Identification of protein homologs from the HMMER	14
3.2	Phylogenetic trees of Transmembrane and secreted microbial species with the reference species	19
3.3	The stable Confirmation of the modelled Prolyl Endopeptidase structure	20
3.4	Structural Validation of the modelled structure of Prolyl Endopeptidase	21
3.5	Docking of the predicted PEP Enzyme and the peptide causing Celiac	21
3.6	Prevalence Analysis of the significant species across the Globe and in Celiac Control Studies.	24
4	**Conclusion and Future Scope**	25
4.1	Conclusion	25
4.2	Future Perspective	26

List of Figures

Fig No	Name of the figure	Page no
1.1	Modulation of gut microbiome on celiac disease	9
2.1	Workflow for the identification of PEPs from Various Taxonomical groups	10
3.1	Taxonomy Identification of Transmembrane microbial species of Animal Taxonomy group	16
3.2	Taxonomy Identification of Secreted microbial species of Animal Taxonomy group	16
3.3	Phylogenetic tree of Transmembrane proteins from Animal taxonomy group microbial species	17
3.4	Phylogenetic tree of secreted proteins from Animal taxonomy group microbial species	18
3.5	MSA of closely related species to the crystal structure of reference protein representing the conservation of active site and other stabilizing residues.	19
3.6	Functional Domain involved in Gliadin Degradation	20
3.7	Superimposed structure of Modelled structure with the reference structure	20
3.8	Structural validation of modelled PEP	21
3.9	The representative structure of Docking of modelled PEP with the epitope.	21
3.10	Global Prevalence of Microbial Species Encoding Transmembrane Proteins	22
3.11	Global Prevalence of Microbial Species Encoding Secreted Proteins	23
3.12	Prevalence of Species in Celiac-Control study (Transmembrane proteins)	24
3.13	Prevalence of Species in Celiac-Control study (Secreted proteins)	24

List of Tables

Table No	Name of table	Page no
2.1	sequence retrieval and Data preprocess	11
2.2	Criteria for the selection of the sequence for the structure prediction	13
3.1	Number of Homologs retrieved after running HMMER	15

Chapter 1

Introduction

Serine proteases, also known as serine endopeptidases are enzymes that hydrolyse peptide bonds in proteins, where a serine residue acts as the nucleophilic amino acid at the active site[1]. Prolyl Endopeptidases are also type of serine proteases which specifically cleave proline-X peptide bonds which are widely distributed in all the kingdoms of life. PEPs differ in specificities for small peptide substrates and from the trypsin-subtilisin-like serine proteases in their catalytic triad arrangement. This enzyme Hydrolyse peptides consisting of less than 30 residues of proline from the carboxy terminal end[2].

Recent Studies showed that PEPs have an essential role in the treatment of diseases such as Alzheimer's, amnesia, depression, cancer and celiac disease[3]. Bacterially PEPs are pharmaceutically necessary enzymes, their functional and evolutionary details are not fully understood. This study attempts to address their activity in the human gut microbiota. These enzymes are of 680-700 residues in length in most of the species[2].Studies on crystal structure of Bacterial PEPs from *Spingomonas capsulata*, *Myxococcus xanthus* and *Aeromonas punctata* have revealed two-domain architecture consist of catalytic α/β hydrolase domain and a β-propeller domain[4]. According to the MEROPS[5] database the PEPs are classified into the S9 family. The S9 family is divided into four subcategories where different members have different substrate specificity. The active site residues are in the order Ser, Asp and His in sequence.

Celiac Disease is caused due to the Gluten peptides which are composed of gliadins and glutenin and these peptides are rich in proline residues which are resistant to the hydrolysis caused by the intestinal enzymes. In addition to the ingestion of gluten, the development of celiac disease requires genetic susceptibility and the disorder almost exclusively occurs in individuals with the human leukocyte antigen (HLA)-DQ2 and/or HLA- DQ8 haplotypes[6]. However, as only a fraction of HLA- DQ2-positive and/or HLA- DQ8-positive individuals consuming gluten develop the disorder, it is likely that other genetic and/or environmental factors play a role in the disease onset[6]. As these peptides are resistant to degradation caused by enzymes like trypsin, Chymotrypsin present in intestine as a result, various long gliadin peptides are generated in the gastrointestinal tract that are capable of activating the detrimental immune responses seen in patients with celiac disease. The Gliadin exposure increases the release of Zonulin (a protein that modulates intestinal permeability by regulating tight junctions) Binding of Zonulin to chemokine receptor (CXCR3) triggers secondary messenger system that leads to disengagement of Zona Occludens from tight junction[7]. As a result Gluten crosses the intestinal epithelial barrier and enters the lamina propria. Tissue Transglutaminase 2(TG2) is a crucial autoantigen in celiac disease, a chronic inflammatory enteropathy triggered by gluten ingestion in genetically susceptible individuals. TG2 is mainly expressed in the lamina propria, the connective tissue layer beneath the intestinal epithelium[7]. Its expression is upregulated in the lamina propria during active celiac disease, TG2 converts distinct glutamine residues in gluten peptides to glutamic acid in a deamidation reaction[8]. Deamidation enhances the binding of gluten peptides by increasing their affinity to human leukocyte antigen (HLA)-DQ2 on antigen-presenting cells. These are presented to CD4 + T cells, it releases pro-inflammatory cytokines

driving inflammation, crypt hyperplasia and the destruction of intestinal villi. Cytotoxic CD8+ T cells are recruited to attack intestinal epithelial cells exacerbating Villous Atrophy and disrupting intestinal lining. In normal Individuals sufficient Microbial Prolyl Endopeptidase activity ensures that most gliadin is broken into smaller fragments where that can be absorbed without triggering an immune response and its excreted through feces.

The modulation created by Gut microbiome has a crucial role in Celiac Disease with both positive and negative effects. The microbes in the gut can secrete Gluten degradation enzymes like Prolyl Endopeptidase, Acyl-aminoacyl peptidase, Oligopeptidase B. There are bacterial species with negative effects like Pseudomonas aeruginosa that have been linked to the production of immunogenic peptides. In celiac conditions Dysbiosis can disrupt tight junctions and increase the intestinal permeability. Comparative Studies have been carried out on 3 Microbial Prolyl Endopeptidase from Flavobacterium meningosepticum (FM) and Sphingomonas capsulate (SC) and a novel enzyme from Myxococcus xanthus (MX)[2]. These enzymes were interrogated with reference chromogenic substrates, as well as two related gluten peptides (PQPQLPYPQPQLP and LQLQPFPQPQLPYPQPQLPYPQPQLPYPQPQPF), believed to play a key role in coeliac sprue pathogenesis. SC hydrolysed PQPQLPYPQPQLP well, but had negligible activity against LQLQPFPQPQLPYPQPQLPYPQPQLPYPQPQPF[9]. In contrast, the FM and MX peptidases cleaved both substrates, although the FM enzyme acted more rapidly on LQLQPFPQPQLPYPQPQLPYPQPQLPYPQPQPF than MX[2]. Whereas the FM enzyme showed a preference for Pro–Gln bonds, SC cleaved both Pro–Gln and Pro–Tyr bonds with comparable efficiency, and MX had a modest preference for Pro–(Tyr/Phe) sites over Pro–Gln sites.

Our study aimed to identify the microbial species in gut which are having PEP Activity against the Gluten protein so that those can be given as probiotics to the individuals suffering with Celiac as a complementary treatment, because Strict Gluten free Diet can lead to nutrient deficiency and also it can't be ensured that all the products would be tagged as gluten free.

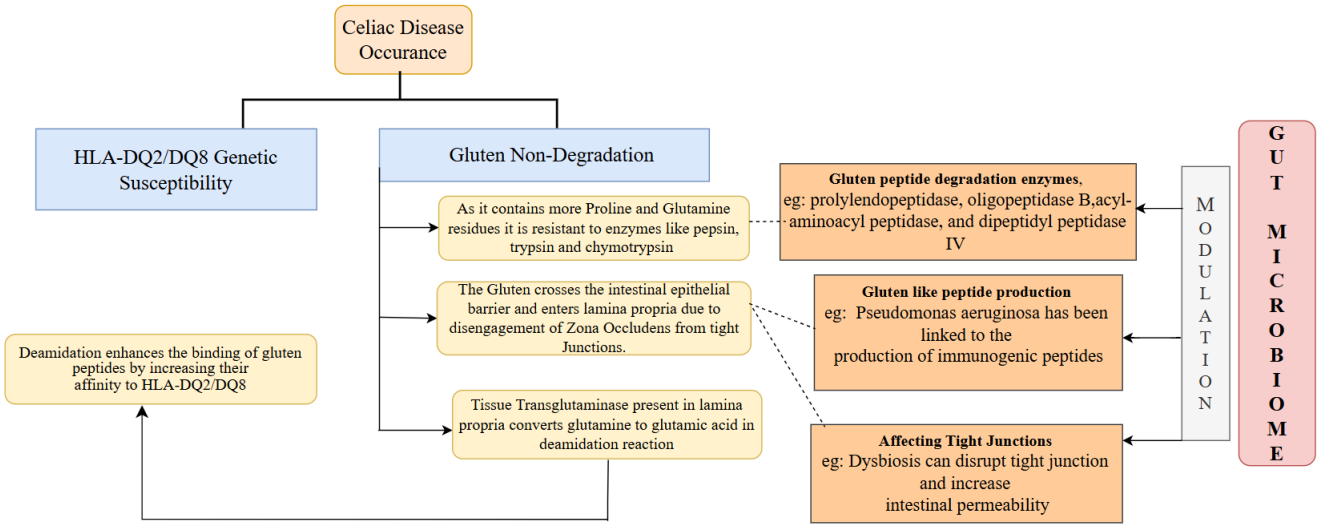


Figure 1.1 Modulation of gut microbiome on celiac disease

Chapter 2

METHODS

2.1 Literature review

In our investigation of microbial prolyl endopeptidases, we conducted an extensive literature search to understand the role of gut microbial prolyl endopeptidases in degrading gluten. Our focus also involved in understanding the role of gluten in pathogenesis of celiac disease and the resistance caused by them to the normal intestinal enzymes like trypsin and chymotrypsin. We utilized search terms to locate microbial PEPs, employing keywords like "Prolyl Endopeptidases", "Gliadin", "gut microbiome", "Celiac", "mechanism". By using these specific terms, we aimed to efficiently identify relevant literature containing information on PEPs involved in genotoxic processes.

2.2 Methods

Our research had two primary objectives. Firstly, we sought to identify the Gut microbial PEPs present across a variety of microbial species, employing sophisticated bioinformatics methodologies. Secondly we aimed to understand the structural and functional properties of the identified PEPs and comparing it with the already found PEP from *Sphingomonas capsulata* to know how similar it's in structure and with respect to functional properties. This section will provide a detailed methodology of the two main objectives.

2.2.1 Identifying of PEPs which has potential to degrade gluten

Identifying gut microbial PEPs that has potential to degrade the gluten from 6 different Taxonomic groups like Proteobacteria, Firmicutes, Cyanobacteria, CFB from the prokaryotic group and Fungi and Animal from the eukaryotic group as the Human gut mainly composed of the microbial species belonging to Firmicutes and Bacteroides.

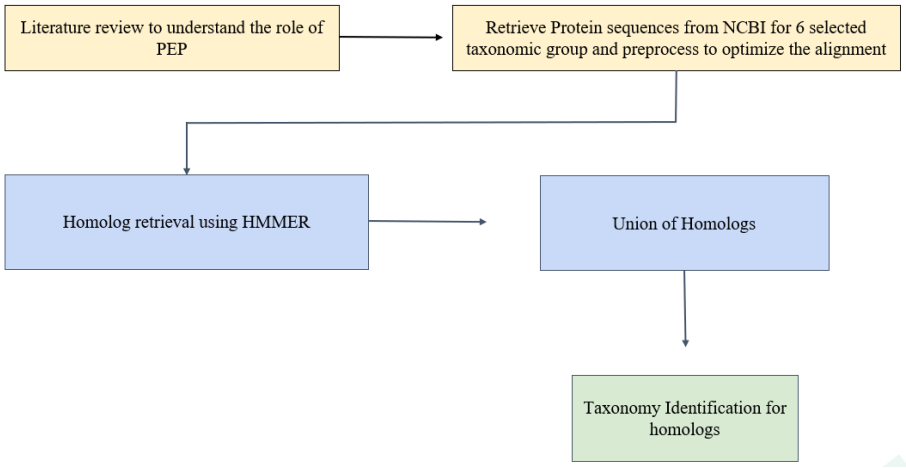


Figure 2.2 Workflow for the identification of PEPs from Various Taxonomical groups

Prolyl endopeptidase protein sequences from all the six taxonomic groups were retrieved from NCBI database. The sequences varied in length with lowest of 100 amino acids residues to 8000 amino acids as this would make the alignment more gapped, so the proteins which were of 600 and above amino acids length were considered to make the alignment optimum.

Sr No	Taxonomic groups	Number of fasta	After preprocessing
1	Proteobacteria	11208	4789
2	CFB(Cytophaga-Flavobacterium-Bacteroides)	1094	943
3	Cyanobacteria	128	121
4	Firmicutes	236	36
5	Animals	4926	4560
6	Fungi	564	454

Table 2.1 sequence retrieval and Data preprocess

After data preprocess the selected sequences were aligned using MAFFT.

#mafft input_sequences.fasta > aligned_sequences.aln

After alignment the .aln files were used as input files to build the HMM profiles using HMM build command.

hmmbuild protein_model.hmm aligned_sequences.aln

The built profiles were used to search for the homologs against the reference database using HMM search command.

 # hmmsearch --tblout search_results.tbl my_protein_model.hmm target_sequences.fasta

The reference database condidered here to retrieve the homologs were Transmembrane proteins and signal peptides from 169 gut associated core key stone species. Totally it had 115692 Transmembrane proteins and 40344 secreted proteins. This reference databases allows selective identification of PEPs that are either secreted or surface-exposed, making them functionally accessible to gluten peptides in the gut lumen. This ensures that the retrieved PEPs are more likely to degrade gluten extracellularly, where the immune-reactive peptides exist in Celiac patients.

After the homologs were retrived from the reference database the all the homologs that were common in all the 6 taxonomic groups were considered. For all the common Homologs belonging to all the 6 taxonomic groups the Taxonomic Lineage was retrieved from NCBI taxonomy browser it provides the information about the Domain,phylum,class,order,family,genus and species. Sankey plots was made to show the flow of the lineages.

2.2.2 MSA of significant species with the Reference Protein and Structure Prediction

Reference Protein (Prolyl Endopeptidase) was considered from ***Sphingomonas capsulata*** as the PEP from this organism is already has been used in the enzymatic treatments like Latiglutenase where it has been given to patients with Celiac Disease under the name ALV003. This latiglutenase contain two components one is PEP from this species and glutamine-specific endoprotease (EP-B2 from barley). This was considered as reference enzyme because it is stable under acidic conditions also. This molecule has been under clincal trials.

Considering this as reference molecule the Homologs that were belonging to Animal taxonomic group from both transmembrane and secreted proteins were considered for the analysis. There were 31 homologs retrieved in transmembrane and 36 retrieved in Secreted proteins. These were aligned with the reference protein using CLUSTAL OMEGA. The species which were closely related to the reference were considered and we looked for the Active site and domain conservation. Prosite was used for determining the Active site and Interproscan was used for the analysis of Beta propellar Domain and Gauge domain.

The top 3 closely related species to the reference from Transmembrane and secreted proteins were selected for the structure prediction and Docking Studies.Homology Modelling (Swiss Model) was done for the structure prediction of proteins. The template that was choosen for building the structure was the PEP from ***Aeromonas caviae*** which is known in its both open and closed form. This domain (closed form: PDB ID 3IVM) showed around 50% identity to the selected sequences and was hence considered as a template for structure modelling of the sequences.

Sr No	Bacterial Species	Coverage	Sequence Identity	GMQE	Qmean Disco
1	*Bacteroides_intestinalis*	85%	53.52%	0.82	0.81
2	*Alistipes_shahii*	86%	54.05%	0.81	0.80
3	*Prevotella_ovulorum*	85%	49.34%	0.81	0.79
4	*Alistipes_finegoldii*	87%	51.91%	0.82	0.80
5	*Alistipes_onderdonkii*	85%	51.76%	0.82	0.72
6	*Alistipes_putredenis*	86%	50.00%	0.82	0.79

Table 2.2 Criteria for the selection of the sequence for the structure prediction.

The best structures with good coverage, Sequence Identity and quality index was selected. These structures were superimposed with the reference structure to know how deviating is the predicted structure from the reference structure. The root mean square deviation was calculated by aligning the Beta propellar and gauge domain of the predicted structure with the reference structure.

To Validate the predicted structures by checking the steric clashes and to look for number of amino acids in allowed and disallowed region was done by plotting Ramachandran plots using saves server, PROCHECK tool was used for this purpose. Per residue energy of the Modelled structure was given by PROSA web tool which provides the Z score which suggests whether the obtained model is reliable and close to experimental determined structures and also the PROSA tool provides the energy plot of the model quality across the sequence.

2.2.3 Docking Analysis of Inhibitor and Substrate

After Validating the structures, to know how good the gliadin peptide can bind to the predicted structure Docking analysis was performed between the epitope that is responsible for causing celiac disease and the predicted protein structures. Autodock vina was used to perform docking for all the selected modelled structures and the epitope of gliadin. The visualization was done using Pymol software.

2.2.4 Prevalence of significant species in Celiac-Control Study

Datasets of case controlled studies were collected from different literatures. It resulted in 5 Datasets where one dataset belonged to whole genome sequencing data and 4 were belonging to 16s sequencing data. Metaphlan 4 was run on the whole genome sequencing data to get the species classification and the abundance, for 16s data SPINGO was run. Once after getting the

species abundance table Prevalence of the significant species was calculated by linking it with the metadata to check how is the prevalence of these species varying with the normal and celiac conditions.

Global prevalence of these significant species which had the PEP enzyme were also calculated which included 140 studies related to different countries. To look for the global prevalence of these significant species.

This prevalence study allows us to understand the how are these gut keystone species are distributed world-wide and how prevalent they are across the world. By knowing this prevalence we can relate it to the Celiac condition as the gut microbiome is linked to celiac occurrence, and occurs where the dysbiosis takes place.

Chapter 3

Results

3.1 Homologs Results: Identification of protein homologs from the HMMER

A thorough investigation of literature yielded valuable insights into the molecular mechanisms like how the Prolyl endopeptidase is involved in degrading the gluten. We also got to know about how is gliadin becoming resistant to the normal enzymatic activity and getting accumulated in the lamina propria of the intestine.

HMMER yielded in 372 homologs for transmembrane proteins and 1250 for secreted proteins.

Sr No	Organism	Number of fasta Sequences	Number of homologs in Transmembrane Species	Number of homogs in SignalP species
1	Proteobacteria	4789	71	265
2	CFB(Cytophaga-Flavobacterium-Bacteroides)	943	63	234
3	Cyanobacteria	121	23	96
4	Firmicutes	36	71	229
5	Animals	4560	71	236
6	Fungi	454	73	190

Table 3.1 Number of Homologs retrieved after running HMMER

After retrieving the homologs from all the 6 taxonomic groups the homologs retrieved for Animal taxonomic group were only considered as it will be more similar and identical to the human enzymes.

But before considering the homologs from Animals group we have combined all the homologs that were common in all the 6 taxonomic groups. After combining the common homologs the Taxonomical lineage was retrieved from NCBI Taxonomy browser which had phylum, class, order, family and genus, species.

Sankey plot was plotted to visually represent the taxonomical hierarchy of the homologs. In both transmembrane and secreted proteins most of the species were belonging to Bacteroides and were belonging to bifidobacterium genus.

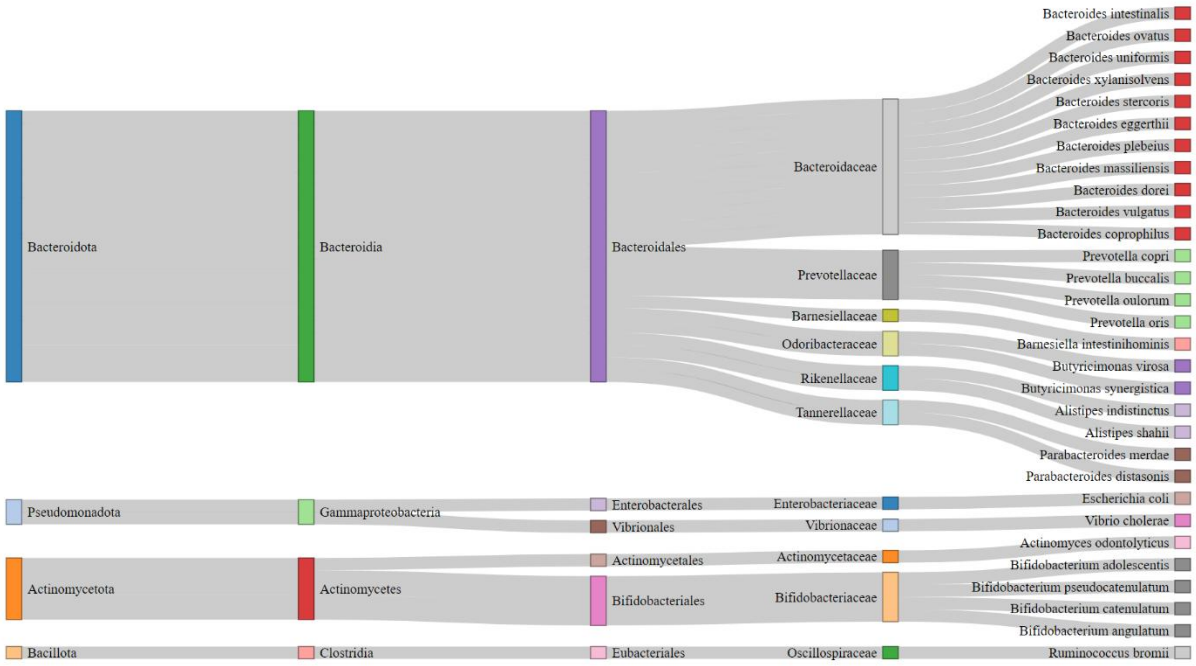


Figure 3.1 Taxonomy Identification of Transmembrane microbial species of Animal Taxonomy group

The above figure shows the taxonomical flow of the homologs that were generated for the transmembrane proteins which clearly shows most of the species were belonging to Bacteroidota phylum which is the major phylum of the human gut microbiome. The hits that were generated were belonging to species which are known to be potential probiotics.

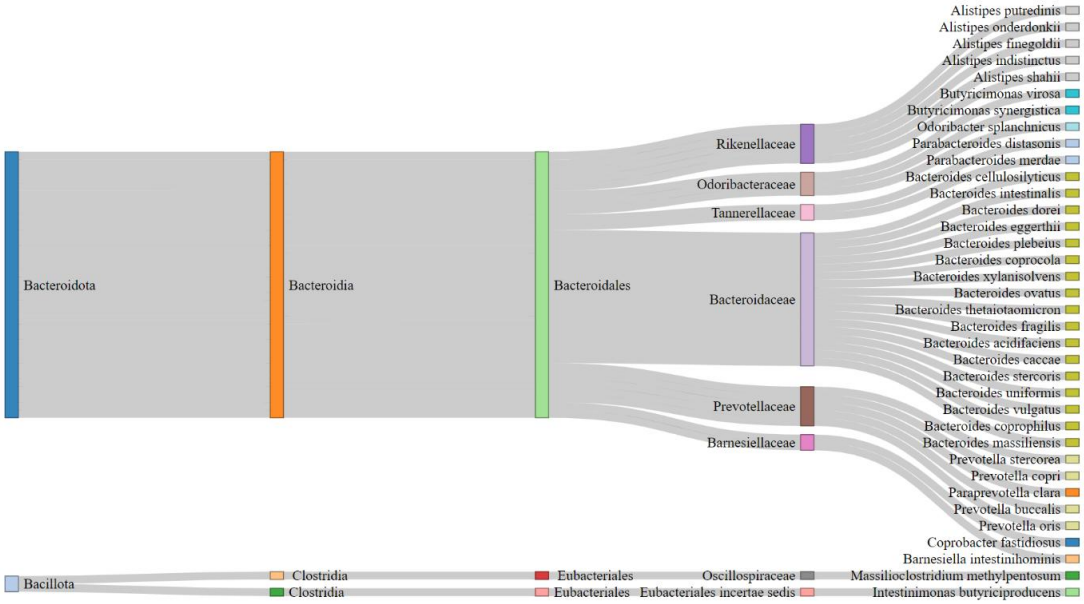


Figure 3.2 Taxonomy Identification of Secreted microbial species of Animal Taxonomy group

So this shows that Prolyl endopeptidases from these species will not be harmful as all these species are reported to be commensal and good bacterium, so there will not be any complications.

The homologs that were retrieved for the secreted proteins were also mostly from Bacteroidota phylum, and genus belonging to Bacteroides and Prevotella, Alistipes which are all also commensal and beneficial bacteria.

3.2 Phylogenetic trees of Transmembrane and secreted microbial species with the reference species

The homologs that were retrieved for Animal taxonomic group were considered from both transmembrane and secreted proteins and it was aligned with the Prolyl endopeptidase of reference protein from the species *Sphingomonas capsulate* to see which of the transmembrane and secreted proteins from the microbial species is closely related to the reference protein. The phylogenetic tree was constructed to analyse the evolutionary relationships between transmembrane and secreted proteases identified from gut microbial species and a reference protein.

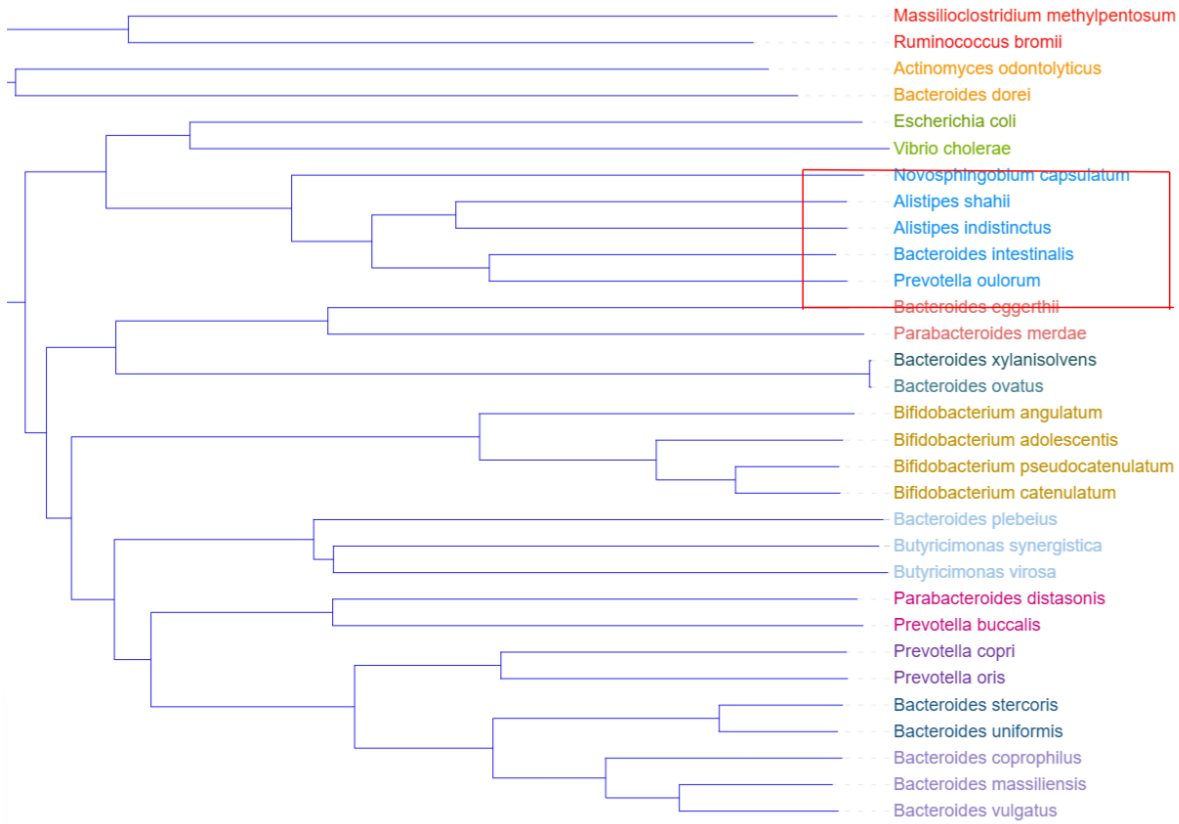


Figure 3.3 Phylogenetic tree of Transmembrane proteins from Animal taxonomy group microbial species

From the tree we can infer that the transmembrane proteins from ***Alistipes shahii, Alistipes indistinctus, Bacteroides intestinalis and Prevotella ovulorum*** are clustering closely with the reference protein from *Sphingomonas capsulate* this suggests they have high sequence similarity and conserved domains to reference protein so they can mimic the reference enzyme.

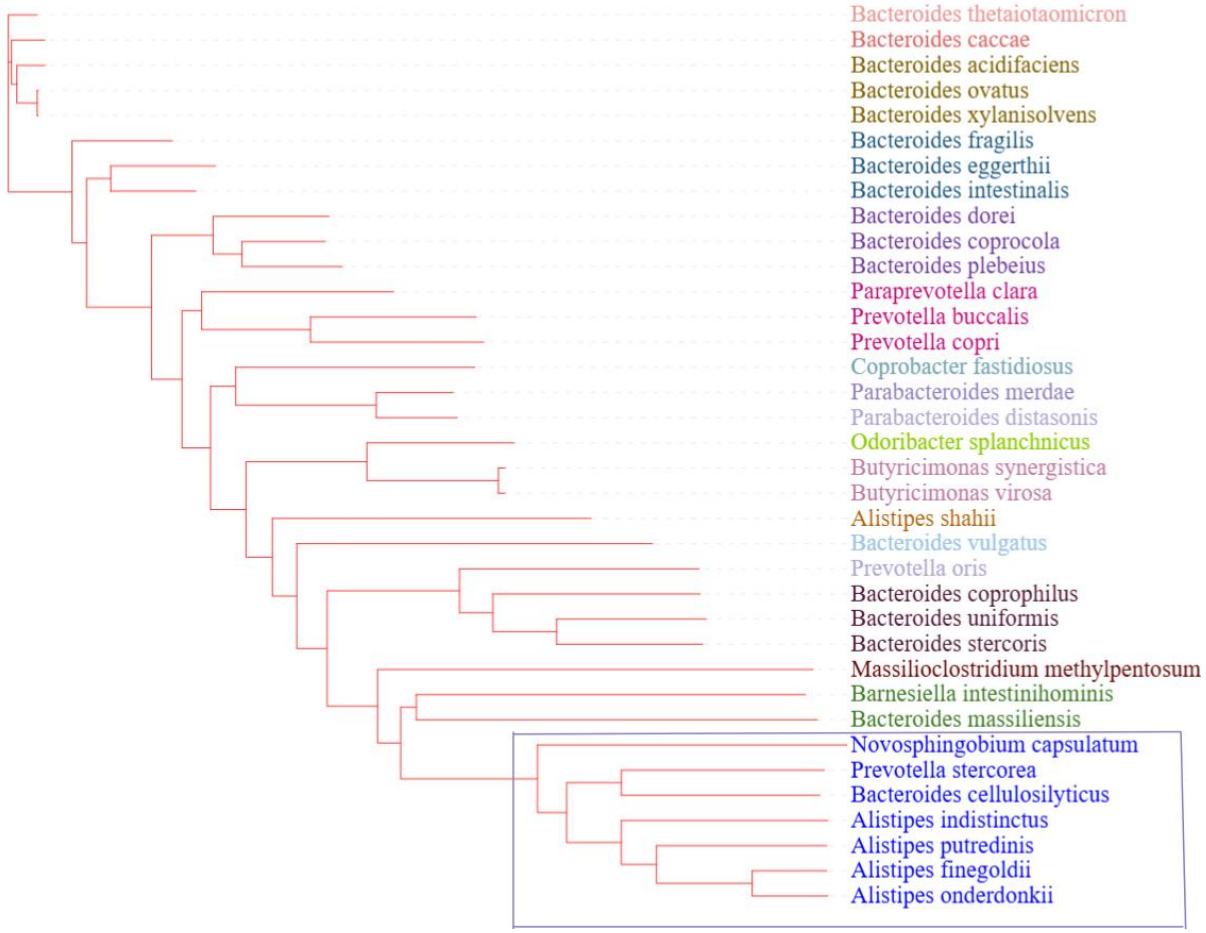


Figure 3.4 Phylogenetic tree of Secreted proteins from Animal taxonomy group microbial species

The above phylogenetic tree represents the evolutionary relationships of secreted proteins from the gut keystone microbial species with the reference protein. For the secreted proteins microbial species like ***Prevotella stercorea, Bacteroides cellulosilyticus, Alistipes indistinctus, Alistipes putredinis, Alistipes finegoldii and Alistipes onderdonkii*** was forming close clusters with the reference protein which shows that these species have conserved domains related to activity and the function so that it can mimic the reference protein.

The closely clustered microbial species from transmembrane and secreted proteins was aligned with the reference protein to look for the conserved active site residues and the conservation of functional domains that are required to bind to the peptide and cleave it.

The species that were getting closely clustered with the reference protein with the Prolyl endopeptidase domain was potential probiotics which it can be given as oral supplements containing bacteria as these enzymes have ability to degrade the gliadin peptide which is involved in the pathogenesis of celiac disease. The multiple sequence alignment of these

microbial species with the reference protein shows that active site domain and the catalytic domain was conserved along with the reference protein.

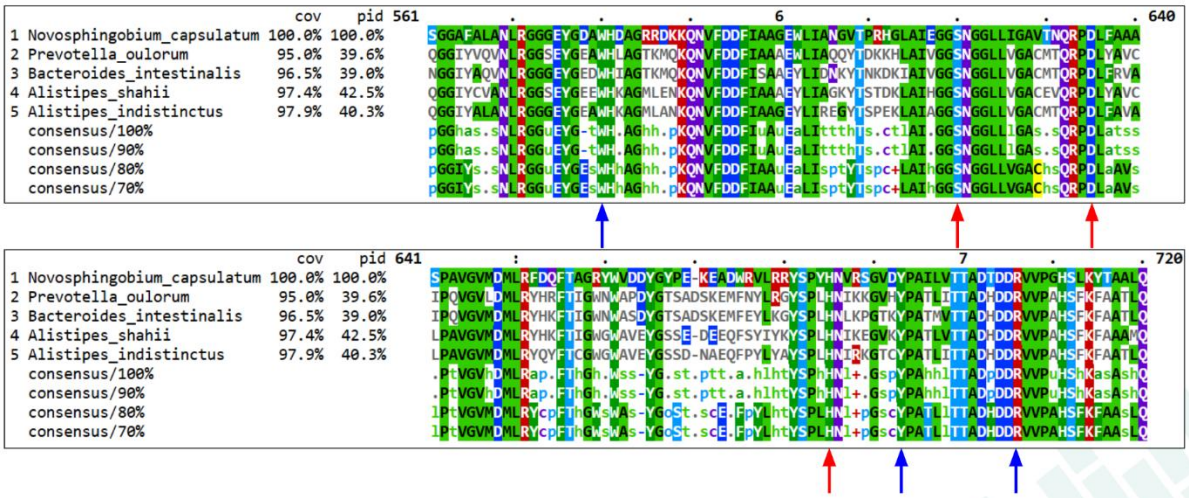


Figure 3.5 MSA of closely related species to the crystal structure of reference protein representing the conservation of active site and other stabilizing residues.

The above figure shows the Multiple sequence alignment of prolyl endopeptidases of selected sequences of the relevant section which is representing the conservation of active site and other stabilizing residues involved in the catalytic activity. The active site residues are marked in red (serine, Histidine and aspartic acid) and the other stabilizing residues are marked in blue (Tryptophan, Tyrosine and Arginine). All the closely related sequences were conserved with coverage of 95% and identity percentage of 35%.

Functional Domain Analysis was performed using Interproscan which resulted in the presence and conservation of functional domains like Peptidase guage domain where initially the peptide binds at this site and the catalytic site where the peptide gets cleaved.

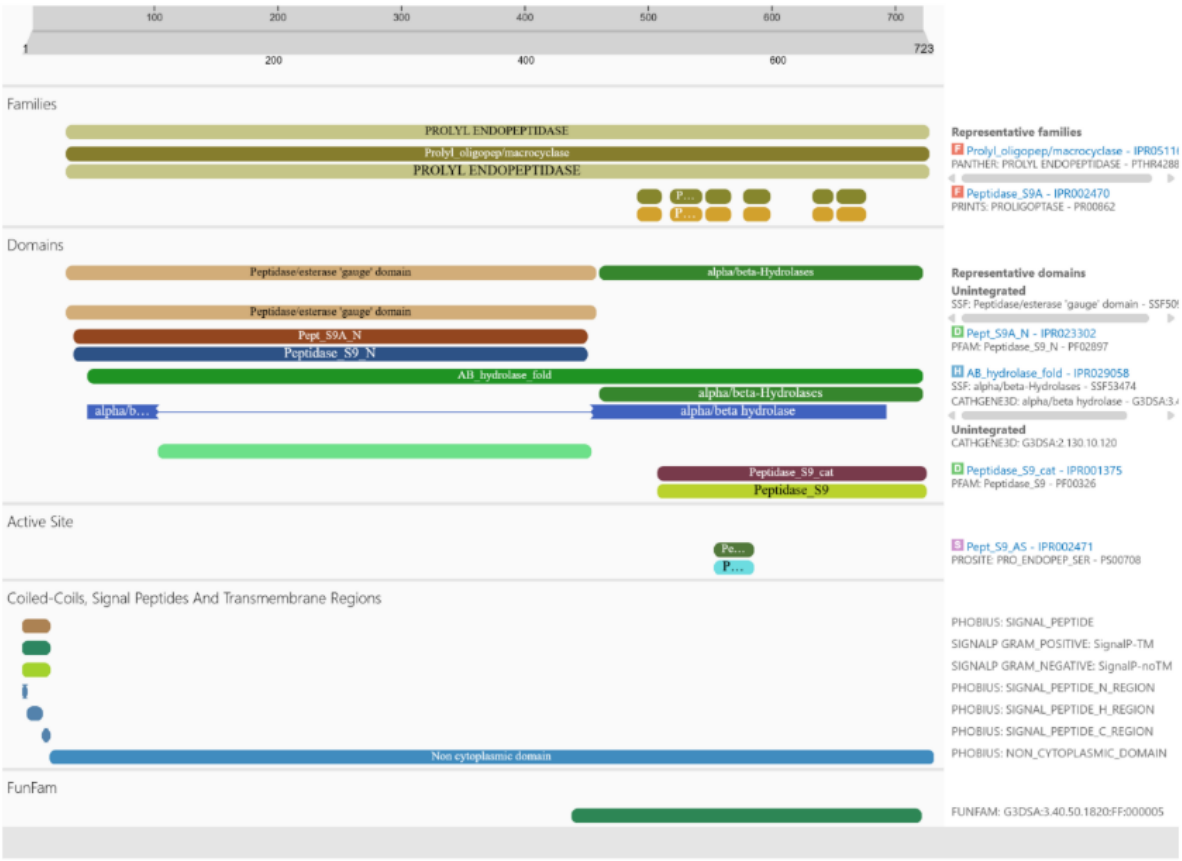


Figure 3.6 Functional Domain involved in Gliadin Degradation

3.3 The stable Confirmation of the modelled Prolyl Endopeptidase structure

The top 3 proteins from the transmembrane and secreted microbial species were considered for modelling the structure. The structures that were modelled was superimposed with the reference protein from *Sphingomonas capsulata.* The catalytic domain and the gauge domain of the reference protein was superimposed with the catalytic and gauge domain of the modelled structure to see how much deviating is the modelled structure from the reference. The root mean square deviation for the superimposed structure with respect to active site was found to be 0.357 Angstorm which indicates that the predicted structure is quite similar to the reference structure.

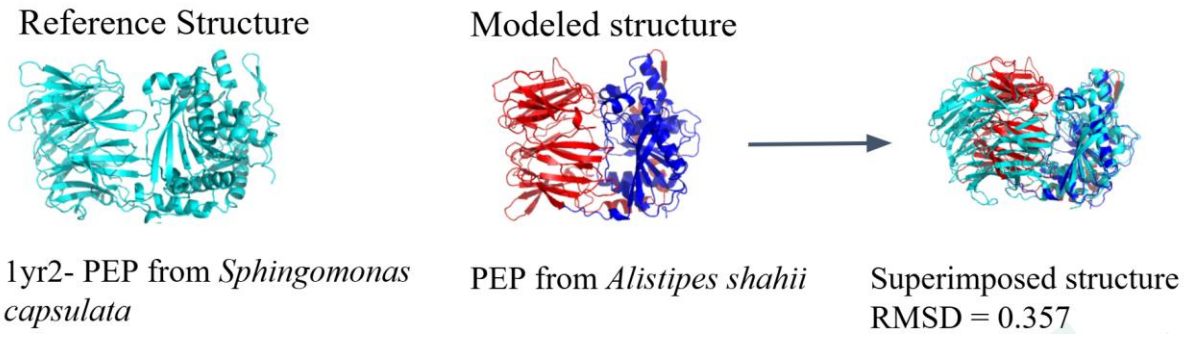
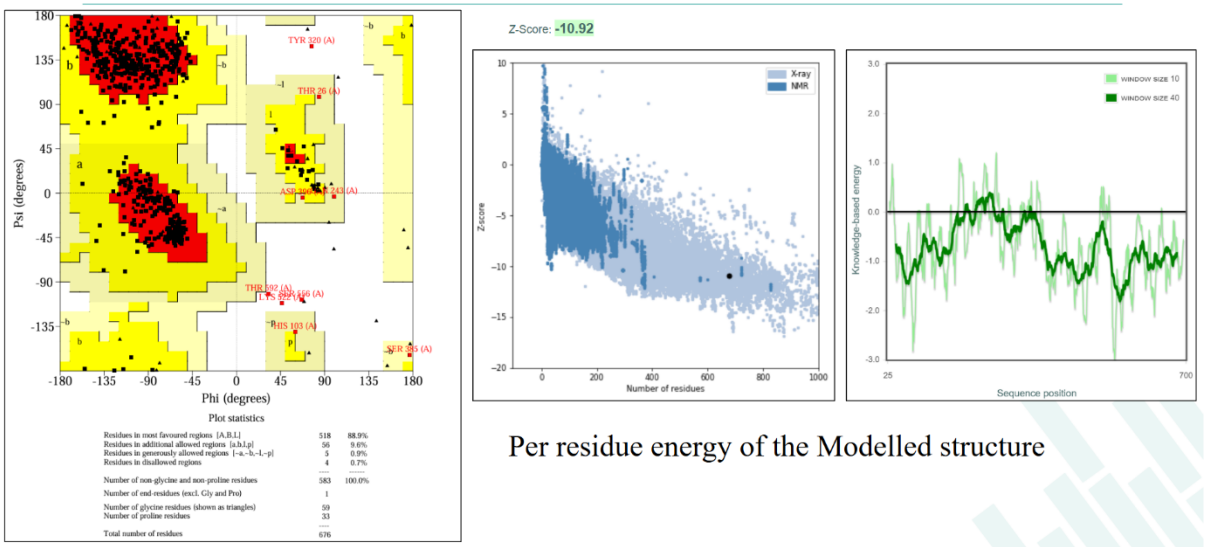


Figure 3.7: Superimposed structure of Modelled structure with the reference structure

3.4 Structural Validation of the modelled structure of Prolyl Endopeptidase

The structures that were modelled was validated by checking the number of residues in the disallowed region and allowed region by plotting the Ramachandran plot using the saves server. This analysis was performed for all the predicted structures and it was found that there were less than 1% of the residues present in disallowed region and rest 99% residues were falling in allowed region. This plots shows the allowed confirmation space of an amino acid in the protein. Prosa servers was used to get the Z score plot which tells whether the obtained model is located within the proteins determined by experimental structures of X ray or NMR. It also provides the energy plot of model quality across the sequence where x axis represents the amino acid sequence position and the y axis represents the energy value for each residue. The positive energy values indicates unfavourable interaction, where as negative indicates favourable interactions which can be seen in well folded protein region. In our predicted structures most of the residues were falling in favourable region and all the structures has the Z score around -10.



Per residue energy of the Modelled structure

Figure 3.8: Structural validation of modelled PEP

3.5 Docking Analysis of the predicted PEP Enzyme and the peptide causing Celiac

The celiac causing epitope which was rich in proline containing residues was choosen to dock with the modelled structures using Autodock Vina. This analysis was performed for top 3 closely related microbial species of transmembrane and secreted proteins with the epitope of gliadin peptide. The binding affinity for all these structures were around -7.2 to -10.2 kcal/mol which suggests that there is a strong interaction between substrate and the inhibitor. The binding residues were choosen for the distance of 4 angstorms and it noted to be serine, tyrosine, Lysine, Threonine. The below figure represents the 2D representation of the docking of predicted structure of microbial species *Alistipes onderdonkii* with the gliadin epitope which had the binding affinity -10.2 kcal/mol.

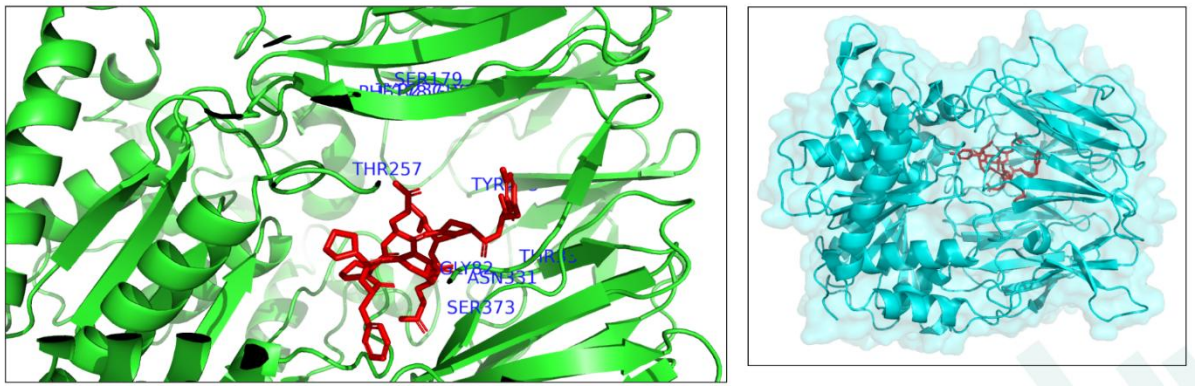


Figure 3.9 The representative structure of Docking of modelled PEP with the epitope.

3.6 Prevalence Analysis of the significant species across the Globe and in Celiac – Control Studies.

The microbial species which were retrieved from HMMER for Animal taxonomic groups from Transmembrane and secreted proteins were considered for analysing the prevalence. 140 global studies were considered from different countries to know how prevalent they are across the globe.

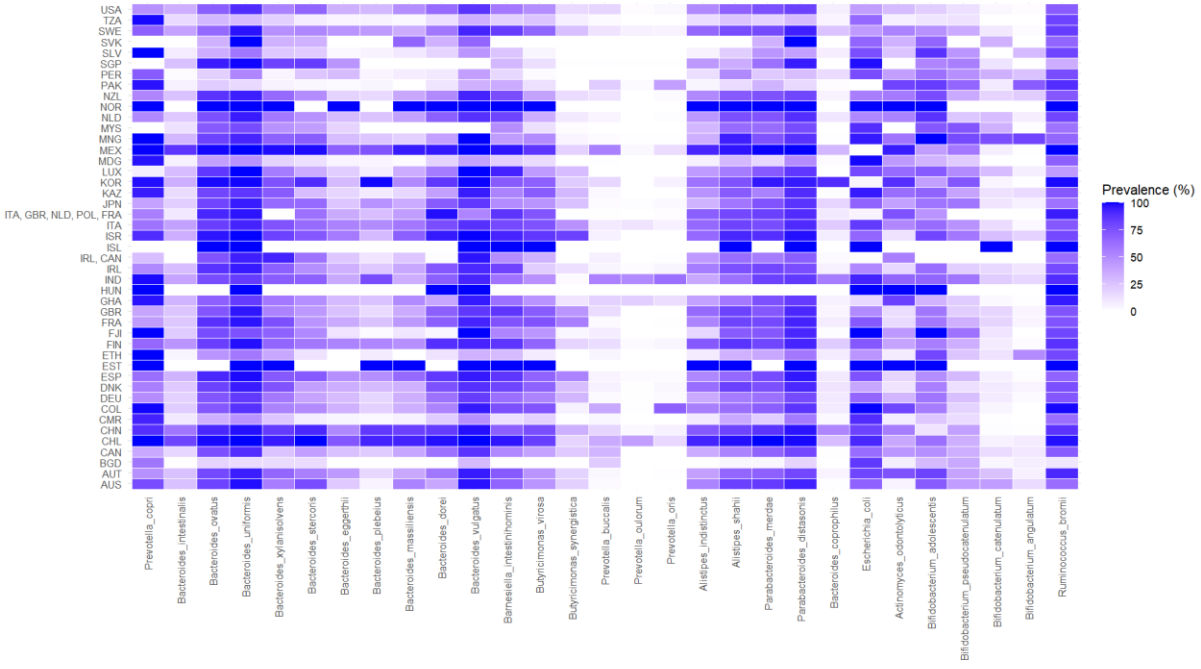


Figure 3.10 Global Prevalence of Microbial Species Encoding Transmembrane Proteins

In the above figure we see that species like *Prevotella oulorum, Prevotella oris , Bifidobacterium angulatum, Bifidobacterium catenulatum* and *Bacteroides coprophilus* which are crucial for gut health are depleting in most of the countries which clearly indicates the gut dysbiosis and countries like USA, Mexico, India, Finland and Australia have Celiac disease with the prevalence of 1 to 4 percent according to world celiac disease rates.

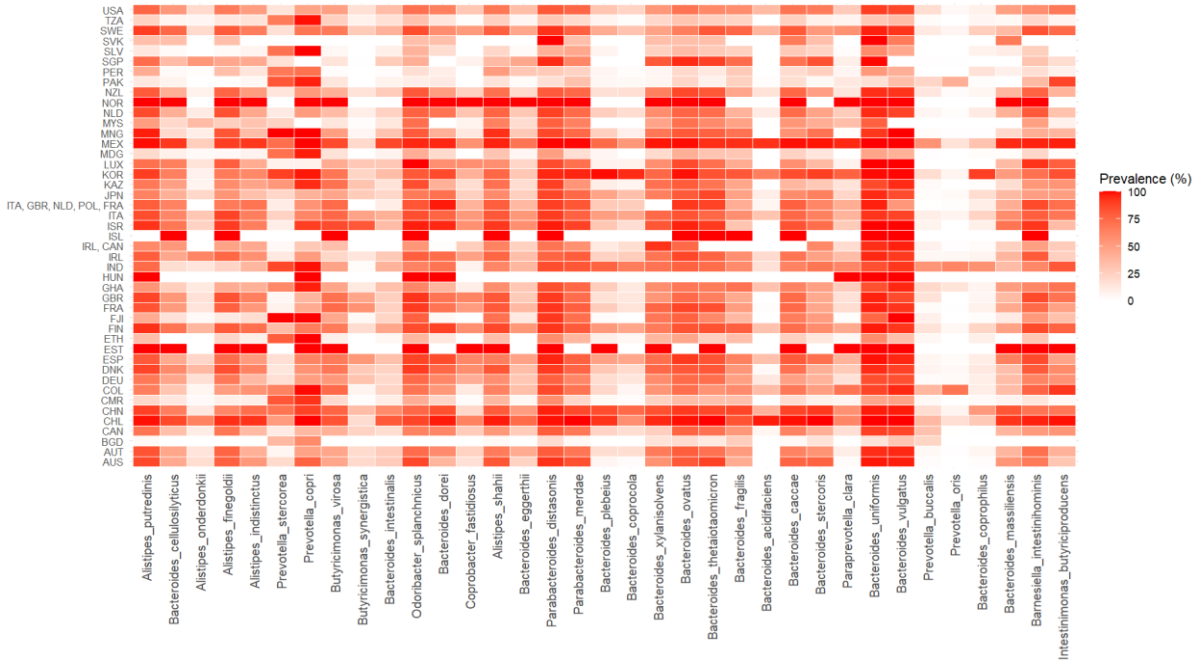


Figure 3.11 Global Prevalence of Microbial Species Encoding Secreted Proteins

The global prevalence of few Microbial species belonging to Bacteroides genus which were encoding Secreted proteins were seen to be getting reduced at most of the countries which also indicates the gut dysbiosis and increase in the prevalence of Celiac Disease.

To build the connection between the gut dysbiosis and celiac occurrence the prevalence was calculated for significant microbial species encoding transmembrane proteins and secreted proteins. 5 datasets were collected which had 419 samples totally, where one dataset belonged to whole genome sequencing data and 4 datasets belonged to 16s data. After running metaphlan on WGS data and SPINGO on 16s Dataset which provided us the species abundance table. Metadata from all the 5 datasets was collected for prevalence analysis.

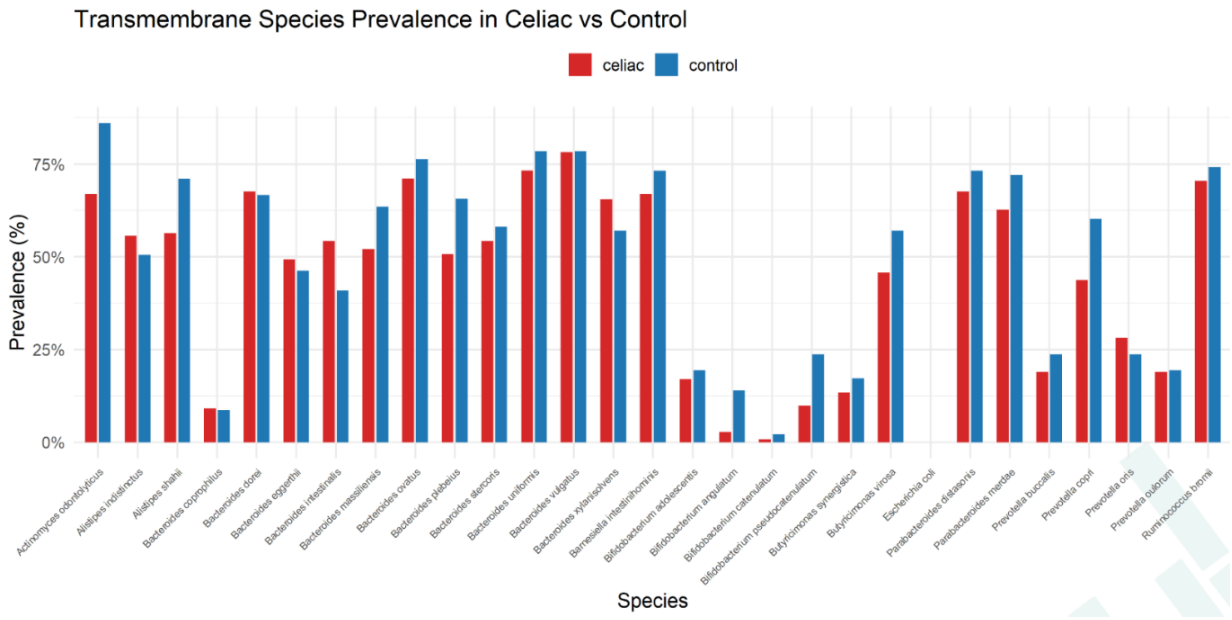


Figure 3.12 Prevalence of Species in Celiac-Control study (Transmembrane proteins)

The significant microbial species are less prevalent in celiac condition compared to control which clearly shows gut microbiome has connection with the celiac occurrence. The gut dysbiosis leads to depletion of beneficial microbes which leads to occurrence of celiac disease.

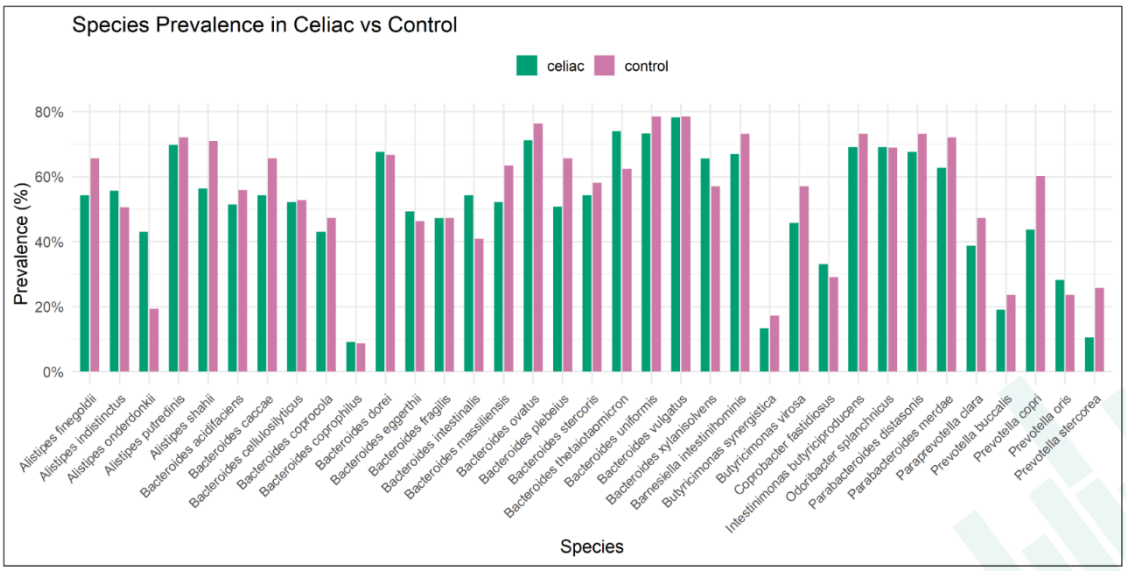


Figure 3.13 Prevalence of Species in Celiac-Control study (Secreted proteins)

Chapter 4

Conclusion and Future Scope

4.1 Conclusion

In this study we have performed analysis on Prolyl Endopeptidases from various Gut Associated Core Keystone Microbial Species which encodes Transmembrane and Secreted proteins. Our study revealed that the homologs that were retrieved from HMMER was mostly belonging to Bacteroides and firmicutes phylum as our gut microbiome is mainly composed of species belonging to this phylum. The species that were retrieved after running HMMER was mostly commensal which are usually present in the human gut and all these significant species has the conserved active sites and functional domains like gauge domain to fit the peptide and catalytic site was also conserved which is required to cleave the proline residues.

The reference PEP was considered from *Sphingomonas capsulata* as the enzyme from this species was used in oral therapeutics. Then we have considered the microbial species from both the transmembrane and secreted proteins which were forming close clusters with the reference which resulted in commensal and potential probiotic species from Genus *Bacteroides* and *Alistipes*. The predicted structures were also closely related to the crystal structure of the reference and Docking studies was performed to know how well the epitope of gliadin peptide can bind to the modelled structures it showed that there is good binding affinity that ranged between -7.2 to -10.2 kcal/mol.

The prevalence analysis of the significant species across the globe and case control studies suggested that the gut dysbiosis can be linked to the celiac occurrence and also highlights a potential link between microbial enzymatic capacity and disease progression.

4.2 Future Perspectives

This study presents the Prolyl Endopeptidases from gut microbial species which are commensal and have ability to degrade gliadin peptide. All these analyses that has been carried out to conclude this was all insilico methods using Bioinformatics approaches. To make this study more relevant Simulations studies need to be carried out on these modelled structures with the epitope to look for the stability of binding under different environments and Experimental validation including in vitro gluten degradation assays, enzyme kinetics, and in vivo testing will be crucial to confirm their therapeutic potential.

Enzyme engineering like mutating the enzyme can be done on the identified PEPs to enhance the binding efficiency, stability and to deliver under certain physiological conditions.

References

1. L. Hedstrom, “Serine protease mechanism and specificity,” Chem Rev, vol. 102, no. 12, pp. 4501–4524, Dec. 2002, doi: 10.1021/cr000033x.
2. L. Shan, T. Marti, L. M. Sollid, G. M. Gray, and C. Khosla, “Comparative biochemical analysis of three bacterial prolyl endopeptidases: implications for coeliac sprue,” Biochem J, vol. 383, no. Pt 2, pp. 311–318, Oct. 2004, doi: 10.1042/BJ20040907.
3. A. Penttinen, J. Tenorio-Laranga, A. Siikanen, M. Morawski, S. Rossner, and J. A. García-Horsman, “Prolyl oligopeptidase: a rising star on the stage of neuroinflammation research,” CNS Neurol Disord Drug Targets, vol. 10, no. 3, pp. 340–348, May 2011, doi: 10.2174/187152711794653742.
4. “Fermentation, purification, formulation, and pharmacological evaluation of a prolyl endopeptidase from Myxococcus xanthus: Implications for Celiac Sprue therapy - Gass - 2005 - Biotechnology and Bioengineering - Wiley Online Library.” Accessed: Jun. 19, 2025. [Online]. Available: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/bit.20643
5. N. D. Rawlings, A. J. Barrett, P. D. Thomas, X. Huang, A. Bateman, and R. D. Finn, “The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database,” Nucleic Acids Research, vol. 46, no. D1, pp. D624–D632, Jan. 2018, doi: 10.1093/nar/gkx1134.
6. D. Dewar, S. P. Pereira, and P. J. Ciclitira, “The pathogenesis of coeliac disease,” Int J Biochem Cell Biol, vol. 36, no. 1, pp. 17–24, Jan. 2004, doi: 10.1016/s1357-2725(03)00239-5.
7. S. Drago et al., “Gliadin, zonulin and gut permeability: Effects on celiac and non-celiac intestinal mucosa and intestinal cell lines,” Scandinavian Journal of Gastroenterology, vol. 41, no. 4, pp. 408–419, Jan. 2006, doi: 10.1080/00365520500235334.
8. A. Fasano et al., “Zonulin, a newly discovered modulator of intestinal permeability, and its expression in coeliac disease,” The Lancet, vol. 355, no. 9214, pp. 1518–1519, Apr. 2000, doi: 10.1016/S0140-6736(00)02169-3.
9. “Structural Basis for Gluten Intolerance in Celiac Sprue | Science.” Accessed: Jun. 21, 2025. [Online]. Available: https://www.science.org/doi/10.1126/science.1074129