Journal of Biomolecular Structure and Dynamics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tbsd20

Comprehensive amino acid composition analysis of seed storage proteins of cereals and legumes: identification and understanding of intrinsically disordered and allergenic peptides

Pratibha Sharma, Lakshay Malhotra & Rajinder K. Dhamija

To cite this article: Pratibha Sharma, Lakshay Malhotra & Rajinder K. Dhamija (04 Jan 2024): Comprehensive amino acid composition analysis of seed storage proteins of cereals and legumes: identification and understanding of intrinsically disordered and allergenic peptides, Journal of Biomolecular Structure and Dynamics, DOI: 10.1080/07391102.2023.2300126

To link to this article: https://doi.org/10.1080/07391102.2023.2300126

JOURNAL OF BIOMOLECULAR STRUCTURE AND DYNAMICS
https://doi.org/10.1080/07391102.2023.2300126

Taylor & Francis
Taylor & Francis Group

Comprehensive amino acid composition analysis of seed storage proteins of cereals and legumes: identification and understanding of intrinsically disordered and allergenic peptides

Pratibha Sharma[a,b], Lakshay Malhotra[b,c] and Rajinder K. Dhamija[d]

[a]Human Behaviour Department, Institute of Human Behaviour and Allied Sciences, New Delhi, India; [b]Department of Biophysics, All India Institute of Medical Sciences, New Delhi, India; [c]Department of Biochemistry, Sri Venkateswara College, University of Delhi, New Delhi, India; [d]Institute of Human Behaviour and Allied Sciences, Delhi, India

Communicated by Ramaswamy H. Sarma

ABSTRACT

The seed storage proteins of cereal and legumes are the primary source of amino acids which are required for sustaining the nitrogen and carbon demands during germination and growth. Humans derive most of their dietary proteins from storage proteins in form of a wide variety of foods, for consumption. The amino acid content of most of these proteins is biased and the need for this biasness is not understood. The high abundance of proline, glutamine, and cysteine in cereals makes the gluten fraction viscoelastic. The cereal proteins have less charge and legume proteins have more charge on them. Their non-polar amino acid distribution has large variations. These characteristics are strongly responsible for the partial and complete unfolding of several domains of the storage proteins. Many of the storage proteins share a highly conserved structural feature within the cupin superfamily spread across all kingdoms of life. The intrinsically disordered viscoelastic proteins help in making dough which is vital for the quality of bread. Unfolded regions harbor more immunogenic sequences and cause food-related allergies and intolerance. We have discussed these properties in terms of comparison of cereal and legume storage protein sequences and allergy. Our study supports the findings that large disordered regions contain allergen-representative peptides. Interestingly, a high number of allergen-representative peptides were cleavable by digestive enzymes. Furthermore, unfolded storage proteins mimic microbial immunogens to induce a memory immune response. Results findings can be used to guide the understanding of immunological characteristics of storage proteins and may assist in treatment decisions for food allergy.

ARTICLE HISTORY
Received 16 September 2023
Accepted 18 December 2023

KEYWORDS
Seed storage protein; intrinsically disordered; amino acids; allergenic peptides; cereal; legume

Introduction

The seeds of cereals and legumes fulfill most of the human protein requirements. Neither of them provides a balanced assortment of amino acids for human needs. Lysine is present in low and high quantities in cereals and legumes. Methionine is present in high and low quantities in cereals and legumes (Navarro, 2012; Shewry et al., 1995; Shewry & Halford, 2002). The gluten proteins of cereals confer viscoelastic properties to the dough. This property enables different food products made from them. Knowledge of these seed storage proteins is of practical and academic interest. The viscoelastic property of the dough depends on the structure of the gluten. Efforts are on to change the composition to get soft or tough bread. The content and composition of seed storage proteins are controlled by a genetic and physiological makeup (Triboi et al., 2000; Zhu & Khan, 2001). Seed proteins have evolved to support the germination process. It remains to be seen the extent to which the composition of the seed storage proteins will be changed to get essential amino acids or soft bread without compromising the plant growth.

Among the cereals, wheat, rye, and barley contain gluten. Oats may have gluten in small amounts, while maize does not contain gluten (Michaelsen et al., 2009). Immunological reactions to gluten can damage the lining of the small intestine, a condition known as celiac disease (Kagnoff, 2005). Small peptides like the pentapeptides QQQPP, QQPFP, and PQQPF in wheat glutenin and gliadin bind to IgE antibodies (Tanabe, 2007; Xue et al., 2011). Legume proteins do not have a high content of proline or glutamine as cereal proteins. Yet, they have allergenic peptides. Food allergens have two main characteristics such as the resistance to proteolysis due to intra-molecular disulphide bonds and stability in acidic conditions. This increases the probability of these epitopes reaching the intestinal mucosa where they can be absorbed and elicit immune reactions (Astwood et al., 1996; Bannon et al., 2003; Kim et al., 2007; Maleki et al., 2000). A very high content of proline makes gluten digestibility difficult for human digestive proteolytic enzymes such as

CONTACT Pratibha Sharma pratibha.aiims@gmail.com Human Behaviour Department, Institute of Human Behaviour and Allied Sciences, , Dilshad Garden, New Delhi-110095, India.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/07391102.2023.2300126.

carboxypeptidase, chymotrypsin, elastase, pepsin, and trypsin (Hausch et al., 2002; Helmerhorst et al., 2010). Numerous disulfide linkages confer compactness to the protein molecules to the extent that they are protected from extensive proteolysis. Disulfide intact Ara h2 (peanut allergen) fragments contain the immune-dominant IgE-binding epitopes intact despite the enzymatic digestion by trypsin, chymotrypsin, and pepsin (Sen et al., 2002). A complete understanding of these food allergens concerning protein sequence disorder is not available. In our study, we have compared the amino acid frequency of seed storage proteins of cereals and legumes with the overall plant database proteins. We have reported the presence of essential, and non-essential amino acids and biasedness in amino acid compositions. We have computationally predicted the secondary structure, tertiary structures, and disorder in all the protein sequences, calculated charge, and hydrophobicity compared with complete structured proteins by X-ray crystallography from the Protein data bank (PDB). In our study, newer approaches like protein-disordered regions containing the allergenic sites are considered. We have further checked digestive enzyme cleavable sites in allergenic representative peptides. Furthermore, immune responses concerning microbial infections are determined. The result findings are important for understanding disorderliness which is associated with allergic-immune responses in humans.

Methodology

Search strategies and collection of protein sequences

The data for seed storage proteins were obtained from Universal Protein Resource (http://www.uniprot.org/), Protein Data Bank (https://www.rcsb.org/), National Centre for Biotechnology Information (https://www.ncbi.nlm.nih.gov/), and through a literature search in PubMed. All possible combinations of search keywords with "seed", "storage", "nutrient reservoir, "legume", and "cereal" were used. We also performed searches in Uniprot using the keywords Poaceae (cereal) [4479] and Fabaceae (legume) [3803]. Only the proteins having their function as seed storage were selected. This was confirmed in the literature searches. To ensure not having redundancy in data, proteins were checked in protein BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) and Multiple Sequence Alignment (https://www.genome.jp/tools-bin/clustalw). This non-redundant set of proteins resulted in a database of seed storage proteins from cereals and legumes (Table S1). The plant proteins sequence data (kingdom Viridiplantae) was downloaded consisting of nearly 14.2 million entries. A third set of 150 randomly selected proteins from PDB data were taken to form the folded protein set. These proteins are involved in different pathways of synthesis; metabolism, ultra structure, communication, and catalysis. A fourth set of sequences of NCBI database proteins randomly selected from all the organisms were downloaded from the NCBI FTP site. The NCBI data contains FASTA sequence data corresponding to approximately 30 million protein sequences.

Amino acids composition of proteins

The amino acid composition of each protein was calculated using an Expasy tool Protparam (http://web.expasy.org/protparam/) (Gasteiger et al., 2005). Physio-chemical properties such as molecular weight, theoretical pI, aliphatic index, and total number of positively (Arg + Lys), and negatively charged (Asp + Glu) amino acids were calculated for each protein. The pI is the pH at which the amino acid does not migrate in an electric field. Expasy calculates the pI and molecular weight based on Compute pI/Mw parameters. Bioperl package (Bio::Tools::Protparam available at: https://metacpan.org/pod/BioPerl) was used to collate the Protparam parameters of individual proteins in the datasets of plant proteins, PDB set, and the database of seed storage proteins (Stajich et al., 2002). The mean hydrophobicity of proteins was calculated using Kyte and Doolittle methods (https://web.expasy.org/protscale/) (Kyte & Doolittle, 1982). Each amino acid has a hydropathy score. Kyte-Doolittle hydropathy scale is widely used to calculate hydrophobicity representing an individual value for each amino acid of a protein. The scale combines data on water-vapor transfer free energies of different residues and their exposure distribution in a set of structurally resolved proteins. A positive score means hydrophobic and a negative score means hydrophilic. The mean hydrophobicity is defined as the sum of the normalized hydrophobicity of all residues divided by the number of residues in the polypeptide. The mean net charge is defined as the total charges at pH 7.0 divided by the total number of residues (Uversky et al., 2000).

Intrinsic disorder prediction

Disorder in each protein was calculated using tools Foldindex (https://fold.proteopedia.org/cgi-bin/findex), TopIDP (https://www.disprot.org/), and PONDR (http://www.pondr.com/cgi-bin/pondr.cgi) (Linding et al., 2003; Prilusky et al., 2005; Sickmeier et al., 2007). These three online software tools are widely used to predict disorders in protein. For seed storage proteins the percent disorder was calculated as the percentage of the total number of disordered amino acids (score above 0.5) divided by the total number of amino acids in the protein). JPred4 software (http://www.compbio.dundee.ac.uk/jpred/index.html/) was used to predict the secondary structure of seed storage proteins (Drozdetskiy et al., 2015). The percent secondary structure was calculated as the percentage of the sum of alpha helix and beta sheet occurrence (reliability of prediction score $\geq$5 on a 0–9 scale) in the protein sequence divided by the total number of amino acids in the protein. The I-TASSER (Iterative Threading ASSEmbly Refinement) tool was used to predict protein three-dimensional structure based on the hierarchical approach of threading (Yang et al., 2015).

Multiple sequence alignment and phylogenetic tree

The identified seed storage proteins from cereal and legumes were aligned using multiple sequence alignment tools

available online (https://www.genome.jp/tools-bin/clustalw) and CLC sequence viewer 8.0 software (CLC bio LLC, Cambridge, MA, USA) (Knudsen et al., 2011). The phylogenetic tree was also constructed using the above tools with bootstrapping of 1000 replicates.

Allergenic peptides and enzyme indigestibility

The Allermatch tool (https://www.allermatch.org/) was used to find the IgE binding sites in the proteins (exact word match search). The similarities were searched for small identical stretches of at least six contiguous amino acids (Fiers et al., 2004). Another allergy prediction tool, Algpred (http://crdd.osdd.net/raghava/algpred/), was also used for finding the peptides showing homology to known epitopes and allergen-representative peptides (ARPs). Algpred allows the prediction of allergens based on the similarity of known epitope with any region of the protein. It facilitates BLAST search against 2890 allergen-representative peptides and assigns a protein allergen if it has a BLAST hit (Björklund et al., 2005; Saha & Raghava, 2006). All the resulting small allergenic peptide sequences from seed storage proteins are listed in Table S2 and Battais et al. (2005); Maruyama et al. (1998); Matsuo et al. (2005); Tanabe (2001); Tatham et al. (1990). The allergens of cereal and legume seed storage proteins were in silico cleaved by pepsin, trypsin, and chymotrypsin by using the Expasy tool Peptide cutter (http://web.expasy.org/peptide_cutter/) (Gasteiger et al., 2005). An overview of the above methodology and tools used are shown in Figure 1.

Result and discussion

A total of 78 seed storage proteins were collated, out of which 38 proteins belong to cereals and 40 proteins belong to legumes. It was found that these proteins not only have huge diversification in length ranging between 150 and 850 amino acid residues but also have large variations in isoelectric point (pI) which was found in the range of pH 4.5–9.5 (Table S1). The percent composition of amino acid along with the molecular parameters was calculated for each seed storage protein. Thereafter, the composition of cereal and legume seed storage proteins was compared with each other (Figure 2 and Table 1).

From the amino acid composition analysis, it was found that the cereal proteins are highly abundant in Gln, Pro, and Cys. Barley, rye, rice, wheat, and maize have a very high proline content than rice, oat, and sorghum. ω-Gliadins, ω-secalins, and c-hordeins of wheat, rye, and barley were found to have very high glutamine content (35–40%). Interestingly, the Gln + Pro content of these proteins ranges from 63-67%. Eleven proteins in cereal contain Gln + Pro above the 45% level as Gamma-gliadin, B hordein, Glutenin low and high molecular weight, Glutenin, Alpha/beta-gliadin, Alpha-gliadin, Secalin, gamma secalin, C hordein, Omega secalin and Omega gliadin. However, only two cereal proteins Globulin-1S allele and Serpin-Z4 have $\leq$5% Gln (Figure 2 and Table S1). On the other hand, the legume proteins have higher amounts of Glu, Arg, Asp, Asn, Lys, Leu, and Ilu than cereals. Conglutin proteins, legumin A, glycinin G2, phaseolin, and beta-conglycinin contain Glu or Arg at more than 10% in content. Only 3 legume proteins contain >10% Gln including Glycinin G2, Conglutin, and Conglutin delta (Figure 2 and Table S1).

Further, the composition of cereal and legume proteins was compared with the proteins of the plant kingdom (Figure 3) and universal proteins (Table 1) along with the propensities of amino acids to form helices and strands (Prevelige & Fasman, 1989). From this analysis, it was observed that the non-polar amino acids are present at lower concentrations in cereals and legumes than in plant and universal proteins. In contrast, the polar components are also present in lesser quantities in cereals compensated by Gln and Pro contents. Gln, Arg, Glu, Ser, and Asn were found more in legume proteins than plant and universal proteins. Additionally, legume proteins have less

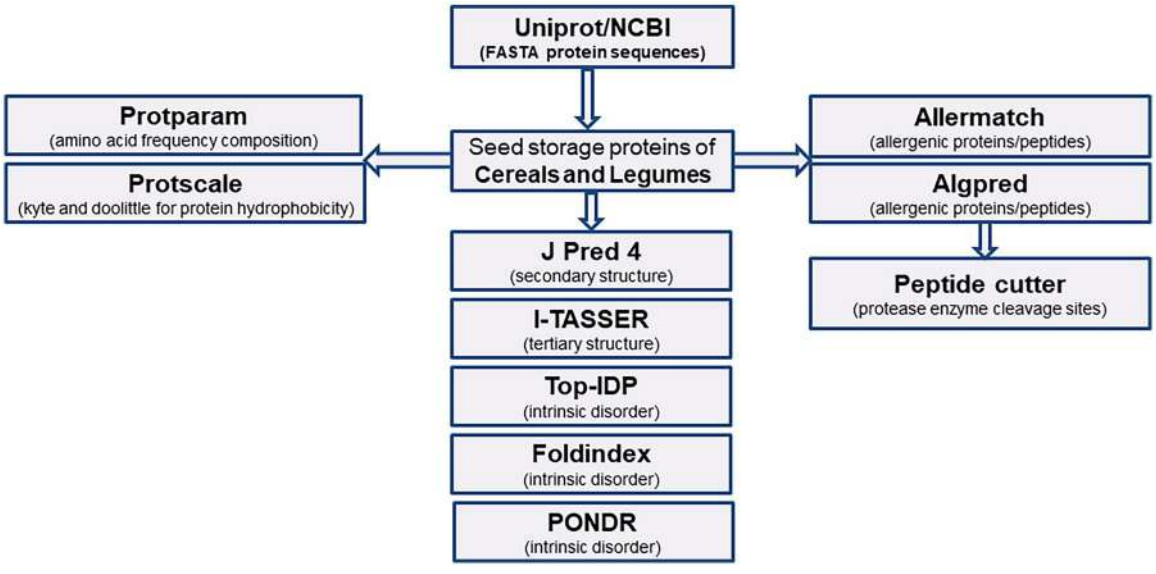

Figure 1. Schematic representation of methodology used for the amino acid composition analysis, identification and understanding of intrinsically disordered and allergenic peptides of seed storage proteins found in cereals and legumes.

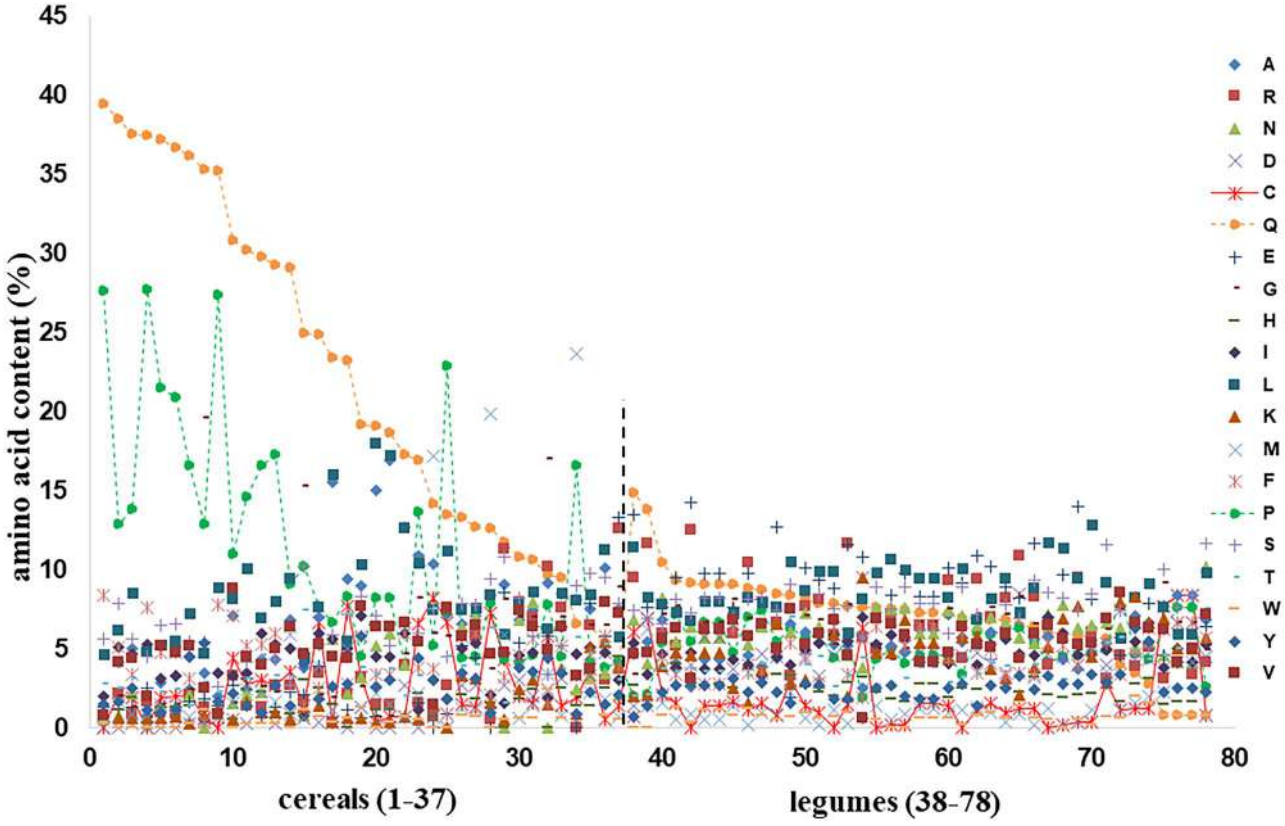

Figure 2. Amino acid composition of individual seed storage proteins of cereal and legumes (see Table S1).

Table 1. Amino acid frequency of occurrence in proteins of different groups.

AA	NCBI random proteins ($n = 30$ million)	Plants proteins ($n = 14.2$ million)	Cereals SSPs ($n = 37$)	Legumes SSPs ($n = 41$)	Helix*	Strand*
G	6.96	7.14	6.00 (↓)	6.30	B	i
D	4.93	5.15	1.29 (↓)	4.58	i	i
C	1.83	1.76	**2.21** (↑)	1.33 (↓)	I	h
Y	2.96	2.92	2.82	2.50 (↓)	b	h
N	4.16	3.90	2.77 (↓)	**6.52** (↑)	B	b
K	5.36	5.42	1.63 (↓)	4.73 (↓)	I	b
S	6.82	8.03	6.54 (↓)	7.77 (↑)	I	b
W	1.45	1.37	0.55 (↓)	0.64 (↓)	h	h
P	4.78	5.12	**11.4** (↑)	4.89	B	b
H	2.46	2.48	1.72 (↓)	2.44	h	b
L	10.06	9.68	7.86 (↓)	9.12 (↓)	H	h
I	6.15	5.14	3.89 (↓)	5.04 (↓)	I	H
A	8.20	8.07	6.46 (↓)	5.34 (↓)	H	I
V	6.61	6.57	5.07 (↓)	5.76 (↓)	h	H
F	4.16	4.23	4.37	4.56	h	h
E	5.66	6.07	3.19 (↓)	**9.23** (↑)	H	B
T	5.61	4.96	3.69 (↓)	3.61 (↓)	i	h
R	5.35	5.93	3.57 (↓)	7.15 (↑)	i	i
Q	3.64	3.60	**22.53** (↑)	**7.14** (↑)	h	h
M	2.74	2.36	2.34 (↓)	1.21 (↓)	h	H

For ≥10% change in amino acid frequency of seed storage protein compared to the NCBI database (increase and decrease is indicated). The amino acid frequencies for randomly selected NCBI (redundant) and UniProt databases differ at the decimal places. The Chou-Fasman helix/sheet propensities of amino acids to form helices and strands are indicated (H = Strong Former, h = Former, I = Weak Former, i = Indifferent, B = Strong Breaker, b = Breaker. Adopted from Prevelige and Fasman (1989). The percentage composition of amino acids is compared for seed storage proteins of cereal and legumes with a complete plant proteins database covering >60,000 reported plant species (Uniprot) and universal proteins. Amino acids rich in nitrogen are N, Q and R. Q is abundant in cereals followed by legumes. N and R are abundant in legumes. The changes in frequency high or low compared to the universal database are shown in bold text.

amount of Gln compared to cereal but higher than plants and universal proteins. Interestingly, the essential amino acids Lys, Met, Thr, His, Val, Leu, and Trp were not abundantly found in cereals and legumes when compared to the average content of all plants (Figure 3).

It is well known that the amino acid Gln can function as a protein-protein cross-linker catalyzed by glutamine transaminase (Orsi et al., 2001). Gln can form interchain hydrogen bonds, therefore contributing to protein's elastic properties. The high content of Gln decreases the solubility of the proteins. On the other hand, the amino acid Cys cross-links low molecular weight glutenin subunits through disulfide bridges (Lombardi et al., 2009; Orsi et al., 2001). In cereals, the amino acid Cys was mostly found segregated

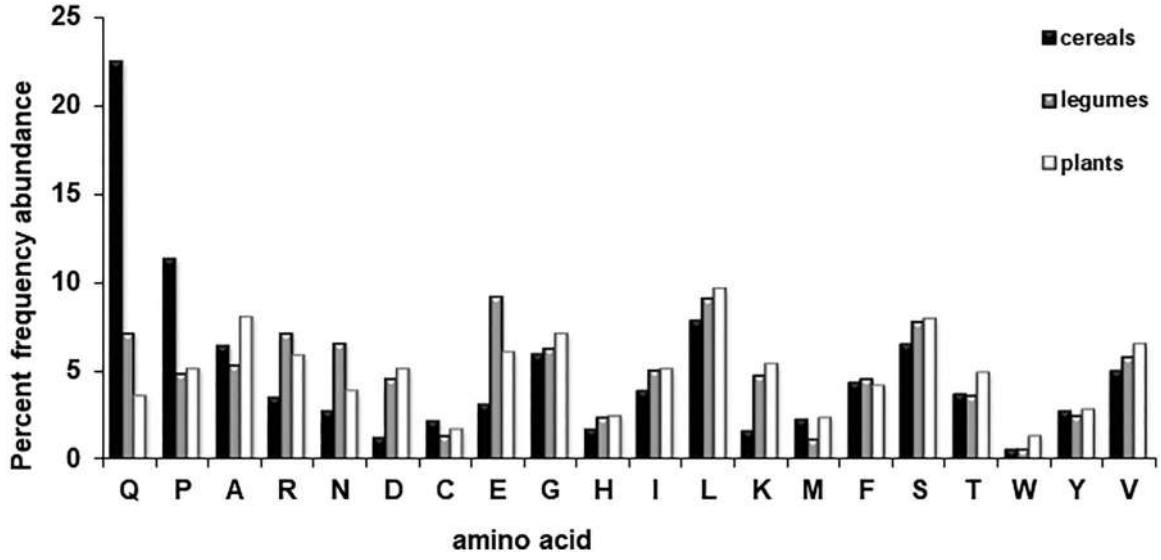

Figure 3. Percent frequency abundance of amino acids in complete plant proteins, cereal, and legume seed storage proteins. H, I, L, K, M, F, T, W, and V are the amino acids essential for the human diet.

towards N and C terminals. The major seed vacuolar storage proteins, (i.e., 7S and 11S globulins), as well as many ancillary storage proteins, form dimers, trimers, and tetramers in the ER lumen shortly after synthesis (Chrispeels et al., 1982). Mutant proteins that are unable to form the correct quaternary structures are retained and degraded in the ER (Vitale & Denecke, 1999). The 26-kDa protein, α-TIP, is an aquaporin, or water channel (Maurel et al., 1995), providing a functional explanation for the desiccation and hydration phases of the seed's life cycle. Accumulation of α-TIP occurs during late seed maturation, primarily after the subdivision of the vegetative vacuole is completed (Johnson et al., 1989).

Pattern and repeats of amino acids among proteins with IDPs

The sequence analysis of cereal proteins showed that the amino acid Gln and Pro can occur in different ways such as highly clustered, randomly repeated, or scattered throughout the sequence. In maize proteins, nine PPP repeats were found in zein gamma protein whereas the prolines were scattered throughout the sequence in zein delta. It was found that cereal proteins like Omega gliadin, C hordein, B hordein, alpha/beta gliadin, glutenin HMW, and glutenin LMW have abundant proline residues scattered throughout their sequence. They were rich in glutamine also with repeats of 15 residues. Wheat LMW glutenin protein contained PP repeats 19 times, QQQ repeats 27 times, and one stretch of 14 Qs. On the other hand, Glutenin HMW contained more scattered Qs and QQ repeats 91 times. Longer glutamine amino acid repeats containing 3–14Qs were most common among proteins alpha/beta gliadin, glutenin LMW, Glutenin HMW, alpha-gliadin, kafirin proteins, secalin, avenin, glutelin type b 5 of rice. Tri-peptide repeats of $(PXX)_n$, $(PQQ)_n$, $(QQP)_n$, $(PPQ)_n$, or $(PXQ)_n$ type were observed in most seed storage proteins. The tandem repeats of the above motifs resulted in polyproline II helices and reduced conformational variability (Jin et al., 2009). Wheat, barley, maize, and rye SSPs being rich in Q and P amino acids had these repeats. These repeats were observed in cereals as scattered as well as tandemly. The Q amino acid residue helps in the compaction of polyproline II helices as occurs in many biological systems, like biominerals.

Protein folding pattern

The folding of protein plays an important role in the regulation of post-translational protein modifications like glycosylation. The presence of proline in place of asparagine was observed which may inhibit the N-glycosylation process of protein folding. 11S globulins are non-glycosylated proteins that form hexameric structures with conserved β-barrel domains (Shewry et al., 1995). Poly-L-glutamine alone can form β-stands held together by a hydrogen bond between the amide groups. Q repeats may form polar zippers which help in protein-protein interactions. The expanded Q repeats increases the protein's affinity for interacting with other protein with similar repeats and helps in the formation of dimers/trimers of β-pleated sheets (Stott et al., 1995). The single glutamine repeats are sufficient to form β-hairpin structures.

Since most of the cereal and legume seed storage proteins lack three-dimensional (3D) structures. Therefore, we have predicted the 3D structures as shown in Table S3. We have mentioned the folds observed and structural differences in these proteins. The seed storage proteins are classified here as albumin and globulins including classes—2S albumins, 7S and 11S globulins, protease inhibitors, gliadin/glutenin family, prolamin family, serpin family, and zein family. Cereal prolamins are rich in proline, glutamine, and cysteine. Prolamins and albumins were found rich in alpha-helices. Whereas globulins were found to have one or more number of beta-barrels. Protease inhibitors have mixed alpha and beta-sheet folds. Other structural features of 3_{10} and Pi-helices are also observed.

Disorder and secondary structure

Intrinsically disordered regions (IDR) and intrinsically disordered proteins (IDP) were observed more in the seed storage

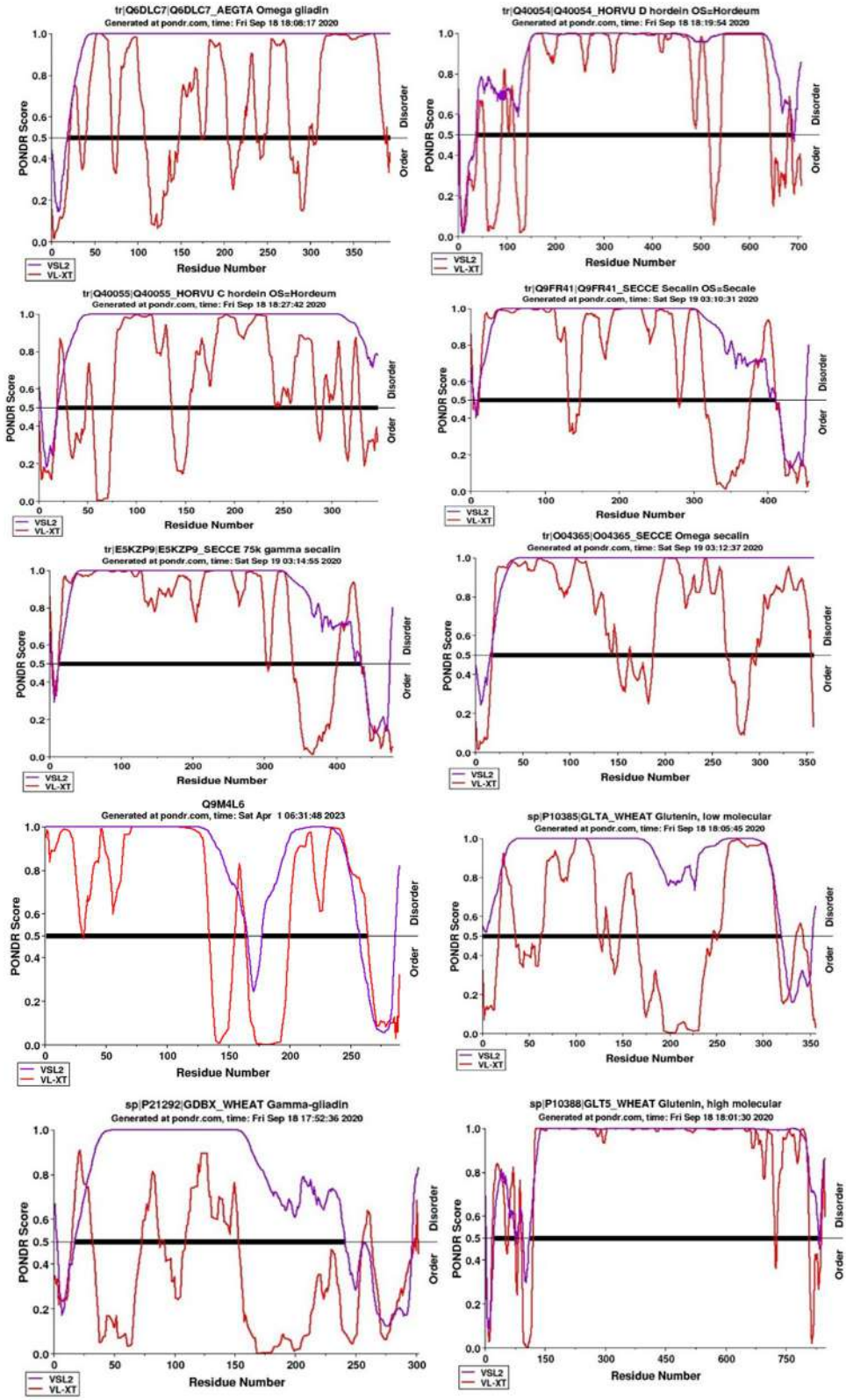

Figure 4. Disorder patterns of selected seed storage proteins. These proteins are mostly or fully disordered as predicted by PONDR VL-XT (red-line)/VSL2 (violet-line) predictors for Omega gliadin Q6DLC7 (*Aegilops tauschii*); D hordein Q40054, C hordein Q40055 (*Hordeum vulgare*); secalin Q9FR41, Gamma secalin E5KZP9, Omega secalin O04365 (*Secale cereal*); alpha-gliadin Q9M4L6, glutenin low molecular weight P10385, Gamma-gliadin P21292, and glutenin high molecular weight P10388 (*Triticum aestivum*). Residues >0.5 thresholds are considered as disordered. Long and short stretches of disorders are shown.

proteins of cereals as compared to the legumes. Gln and Pro are the major disorder-promoting residues in cereals. In contrast, Gln, Arg, Glu, Asn, and Ser were in legume proteins. Thirteen proteins contained <20% disorder. Ten proteins contained nearly > 80% disorder (Figure 4). Other proteins contained either disordered terminal regions or repeats of unstructured regions (Table S1). Globulins, serpin, conglutins, and albumins were found with > 30% predicted secondary

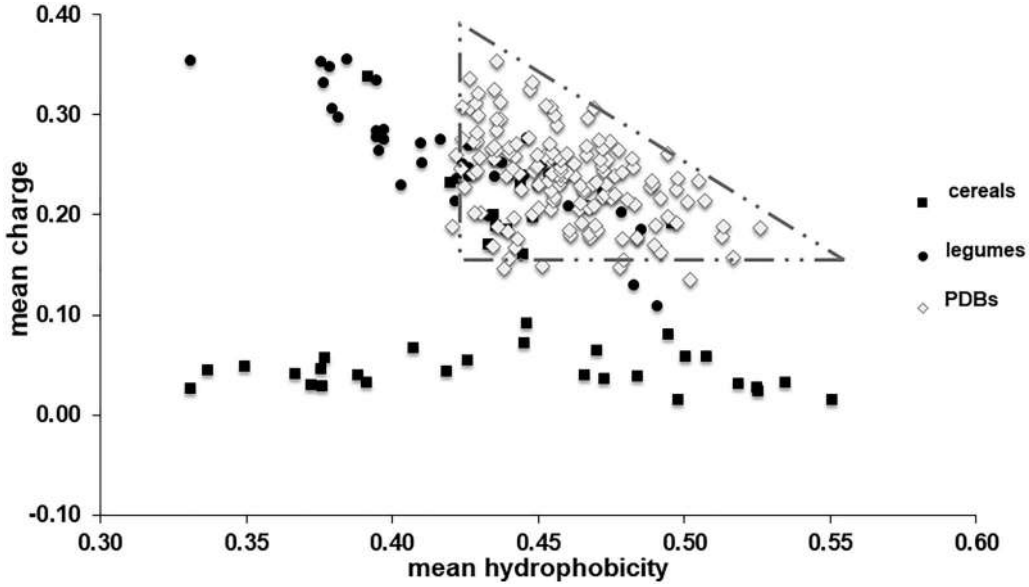

Figure 5. Comparison of the mean net charge and mean hydrophobicity for legume seed storage proteins, cereal seed storage proteins and the set of 150 proteins collected from PDB as natively structured (seed storage proteins are shown in dark square as cereal and legume as dark circle and PDB data proteins are shown in grey rhombus).

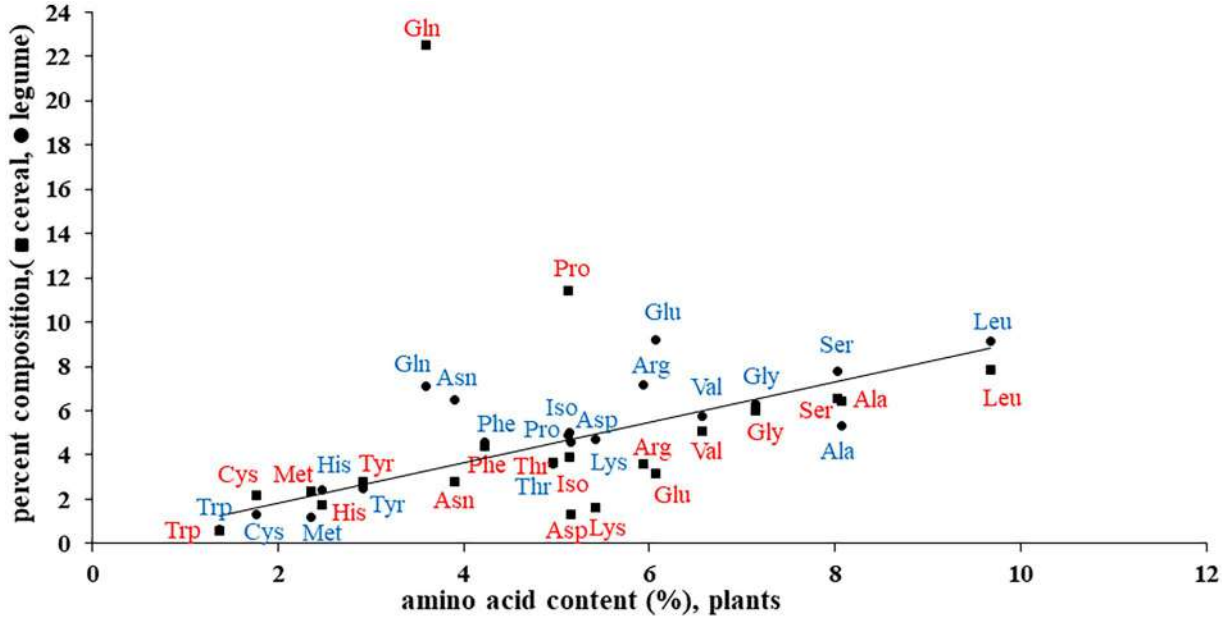

Figure 6. Percent universal plants frequency compared with that of cereal and legume seed storage proteins.

structure among seed storage proteins. Nine proteins had <10% secondary structure in them (Table S1). In the discussion here, we have considered the proteins as disordered if they have >80% disorder and <30% secondary structure. Ordered proteins have <20% disorder and >30% secondary structure. Legume proteins have a higher content of secondary structure. In the charge-hydrophobicity (C-H) plot (Uversky et al., 2000), the structural proteins were observed in a separate cluster with a mean net charge between 0.13 and 0.35 and mean hydrophobicity between 0.43 and 0.53 (Figure 5). Both hydrophobicity and charge are necessary for a protein to fold. Apparently, proteins cannot fold completely outside this space. Most of the legume and cereal proteins were observed outside the space. Three-fourths of the cereal proteins had net charge <0.1 and two-thirds of legume proteins had hydrophobicity <0.43 are unfolded. Thus, either a large hydrophobic or high charge characteristic alone is not sufficient to ensure complete protein folding. The high net charge leads to charge-charge repulsions and low hydrophobicity will not form the compact structures in unfolded proteins (Xue et al., 2011). Around one-third of the seed storage proteins occupied the folded space of the C-H plot. In case of cereal proteins, the mean charge did not vary much (between 0.02 to 0.06), the disorder decreases as the mean hydrophobicity increases from 0.33 to 0.55. The secondary structure increases from 10 to 20% as the hydrophobicity increases from 0.4 to 0.55. If all the cereal storage proteins are considered, this correlation was insignificant (r^2

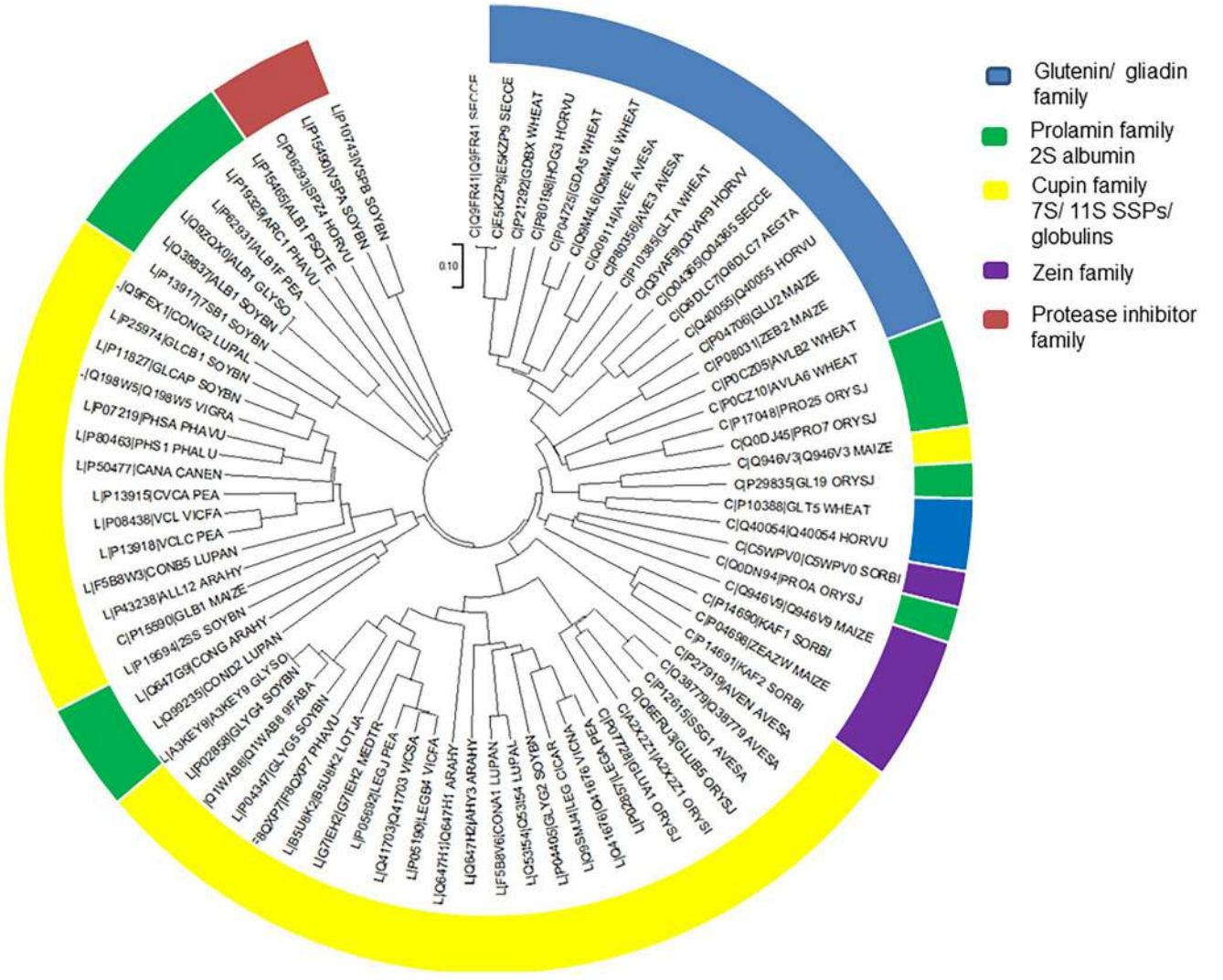

Figure 7. Phylogenetic analyses of seed storage proteins. Phylogenetic tree constructed for seed storage proteins of cereals and legumes using ClustalX2 after multiple sequence alignment with clustal W and CLC sequence viewer using neighbor-joining method with 1000 bootstrap replicates. The cereal and legume proteins are denoted by 'C' and 'L,' protein accession number and organism name respectively. All the families are highlighted including Cupin family (11S & 7S globulins), cereal prolamins, 2S albumins, gutenin/gliadin family, zein family and protease inhibitor family.

= 0.0065). This correlation was not significant for folded proteins ($r^2 = 0.1822$). The legume proteins carry charges within the range of 0.13 and 0.35 sufficient for folding. The legume proteins having hydrophobicity >0.43 are in the folded region. The correlation of the C-H relationship for legume proteins is significant ($r^2 = 0.7503$). The proteins of cereals and legumes in the folded region had a low content of Gln and Pro. It is expected that the C-H characteristics should innately reflect in the secondary structure and disorder. These three parameters, namely, secondary structure, disorder, and C-H characteristics did not correlate among themselves (Figure 6). It could be due to that the effects of individual amino acids and their proximity effects on the secondary structure have not been reflected completely in the predictions. However, some of the effects of the abundance of Gln and Pro appear to be evident. The presence of Pro and Gln in large amounts influences the structures of the storage proteins. The repeat occurrence of Pro enables the polyproline II helix (PP II) formation (Stapley & Creamer, 1999). Gln has the second maximum propensity for PP II. Short polyglutamine chains can also form polyproline II helical structures (Chellgren et al., 2006; Kelly et al., 2001). The sequence repeats of Pro and Gln throughout the sequences allow the cereal proteins to take up extended form. In the extended form, Gln residues form interchain H-bonds. Pieces of evidence indicate that the proline residues enable the glutamines to form H-bonds by keeping them in an open and extended conformation (Darnell et al., 2007). The peptide-solvent interactions are important factors in stabilizing the PP II structure (Kelly et al., 2001; Sreerama & Woody, 1999). The protein regions rich in both glutamine and glycine are more mobile (loop forming) (Alberti et al., 2001). Hydration produced a marked increase in chain mobility, especially above a threshold water content of about 37% w/w. The changes in secondary structure are very noticeable in water contents corresponding to the NMR mobility threshold (Belton et al., 1995).

During germination, the storage proteins provide nitrogen and carbon. They can hold water. The requirement for the viscoelastic proteins in the seed is not fully understood.

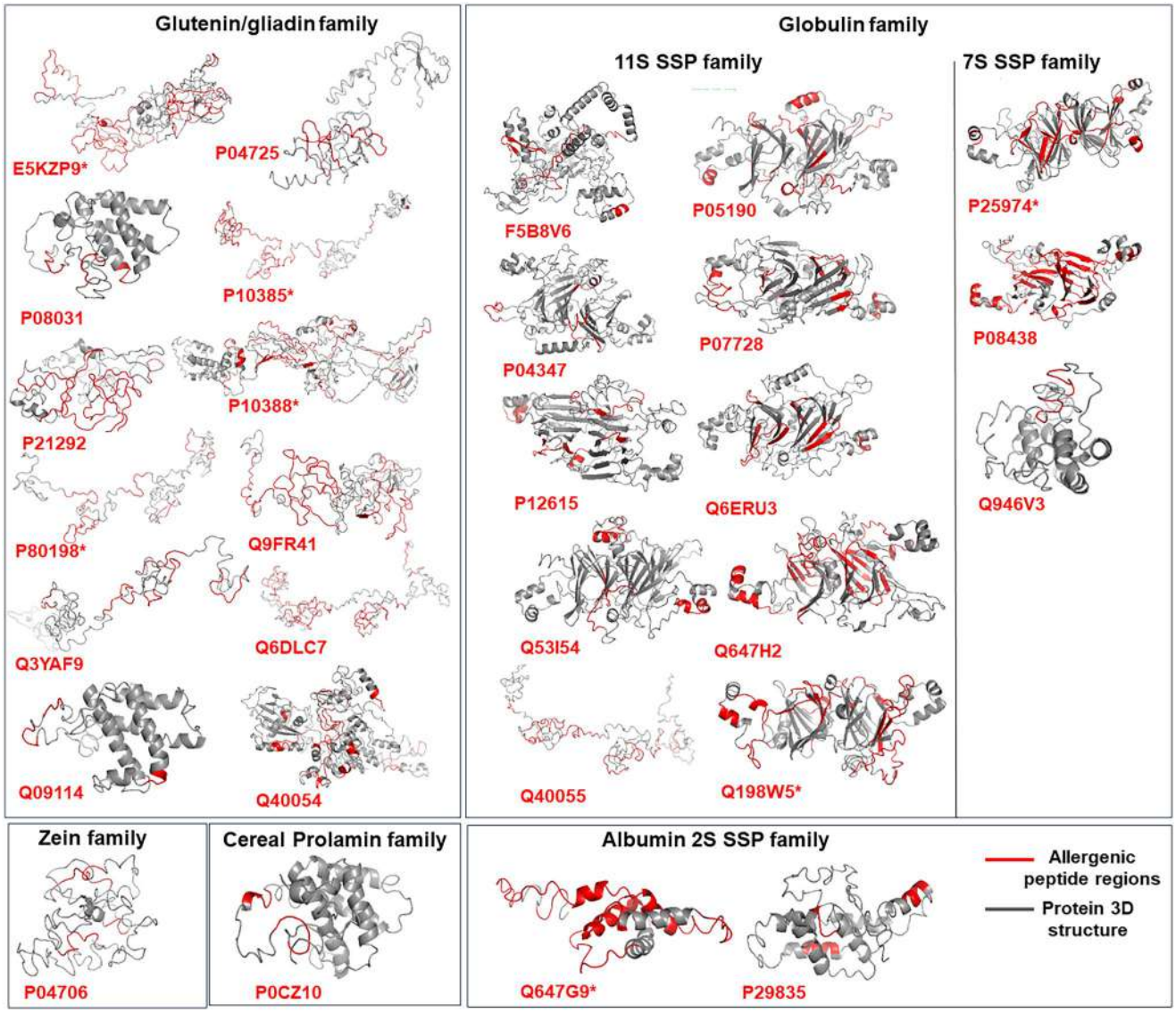

Figure 8. Structure based classification of seed storage proteins. The three-dimensional (3D) structures were modelled using i-TASSER platform. Complete structures of each protein are represented along with allergenic peptides. The allergenic peptide regions correspond to the sequences represented in Table S2 and 3D structure correspond to Table S3 (*indicates the protein as known allergen).

The genetic and environmental factors cross-talk to determine the protein accumulation, amount, and size distribution of polymeric proteins with little plasticity (Malik et al., 2011). Enriching seed storage proteins in essential amino acids would be at the cost of the contents of Gln, Pro, and Cys. This will have implications for the 'bread' forming ability. Seed proteins do not evolve to contain essential amino acids. Genetic modifications aimed to enrich essential amino acids may give instability to cereal seed germination.

The complete phylogenetic analysis and multiple sequence alignments of cereal and legume SSPs from different species are represented in Figure 7 (Tables S4 and S5). These proteins belong to the Cupin family (7S/11S globulins), albumin family (2S), cereal prolamins, glutenin/gliadin family, protease inhibitors, and zein family. These families trigger allergic reactions (Jain, 2023). Based on the hierarchical approach to protein structure prediction and structure-based function annotation models we have classified the identified SSPs. We observed the unfolded regions of these proteins carrying a high number of allergenic peptides. In Figure 8, we have highlighted the allergenic peptides in the modeled three-dimensional structure. We have clearly shown that most of the allergenic peptides lie on the unstructured regions of these proteins. Although a few legume proteins have PDB structures (Table S1), the allergenic peptides belong mainly to random coil or loop regions.

Allergy and enzyme digestibility

Some of the seed storage proteins are known as the major food allergens (King et al., 2009; Navarro, 2012). These proteins contain a very high number of IgE binding sites. Allergenic peptides rich in proline and glutamine amino acids were abundantly found in cereals. Whereas in legumes most allergenic peptides were found rich in R, N, E, and L. There were 72 seed storage proteins containing allergenic peptides. Six proteins were observed to have no allergenic peptides. A total of 697 non-redundant allergenic peptides

Table 2. Enzyme digestibility/non-digestibility of allergenic peptides of cereal and legume seed storage protein.

	Cereal SSPs ($n = 37$)		Legume SSPs ($n = 41$)	
Allergenic proteins	35		37	
Allergenic peptides	272		796	
Non-redundant allergenic peptides	165		532	
Digestible allergenic peptides (pepsin, trypsin, chymotrypsin)	105		464	
Non-digestible allergenic peptides (%)	60 (36.4)		68 (12.8)	
	Percent frequency distribution of amino acids			
Amino acid	Enzyme cleavable[#]	Enzyme non-cleavable[#]	Enzyme cleavable[#]	Enzyme non-cleavable[#]
A	7.9	8.9	6.2	7.4
R	5.6	0.3	7.4	0.2
N	3.8	1.0	8.7	3.2
D	1.8	0.8	4.2	9.3
C	3.2	0.0	1.3	1.3
Q	22.3	43.4	6.7	5.3
E	4.0	0.0	8.9	38.6
G	5.4	6.9	6.5	6.3
H	1.7	0.0	1.3	1.7
I	3.1	2.6	5.1	7.2
L	8.7	0.5	9.8	0.0
K	0.8	0.0	4.7	0.0
M	0.3	0.0	0.5	0.2
F	5.2	3.8	6.2	0.0
P	11.0	21.7	5.5	5.5
S	3.2	4.3	6.3	4.9
T	1.4	2.3	2.5	3.2
W	1.4	0.0	0.6	0.2
Y	1.9	1.3	2.1	0.0
V	7.2	2.3	5.4	5.7

Enzymes considered for cleavage of allergenic peptides are pepsin (pH-1.3), pepsin (pH > 2), trypsin, chymotrypsin high-specificity and chymotrypsin low-specificity available at expasy tool peptide cutter. Allergenic peptides are found in both cereals and legumes SSPs. Allergenic peptides rich in P, Q amino acids and R, N, E amino acids in cereal and legumes respectively are disorder-promoting amino acids. These amino acids forms patterns in the allergens like PXX, QQP, PQQ, PQX etc. in cereals and patterns with R, N, E, D amino acids are observed in legumes. Limited proteolysis of allergenic fragments is represented here.

were predicted (Table 2 and Table S2). Cleavability by digestive (human) endo-proteases was predicted. 36.4% (60 out of 165 peptides) and 12.8% (68 out of 532 peptides) of cereal and legume allergenic peptides were not cleavable by pepsin, trypsin, and chymotrypsin (Table 2). The legumes have a higher proportion of protease specificity amino acids. They have less proportion of non-cleavable peptides than cereals. The composition of the non-cleavable peptides is very different from the cleavable allergenic peptides. The Gln and Pro residues make up to 65% content of the cereal non-cleavable peptides while the legume non-cleavable proteins contain only 10% Gln and Pro content. Cereal non-cleavable peptides have no acidic residues while the legume proteins were found with 48% of them. The residues Leu, Trp, and Met were completely absent in the non-cleavable peptides. Gln, Pro, and acidic residues are not specific to any endoproteases. Proline occurring at P2 and P2′ positions of the scissile bond are inhibitory to pepsin. Proline occurring at P1′ position is inhibitory to trypsin (Gasteiger et al., 2005). These properties may enable the prediction of non-cleavable allergenic peptides. It would be of great interest to know the scissile amino acid specificities of the aspartic and cysteine proteinases of the germinating wheat (Capocchi et al., 2000). Most of the allergenic peptides are cleavable (Table 2). Resistance to enzyme digestion could be one of the causes of food allergy (Bannon et al., 2003). It is evident, however, that the digestion stability alone cannot be used to define an unknown protein as an allergen. We observed more phosphorylation sites of amino acids S, T, and Y in legume SSPs compared to cereal SSPs. The role of posttranslational modifications in allergy and other immune responses is not reflected in this analysis. Seed storage proteins undergo deamidation reactions. This is linked to celiac disease (Sjöström et al., 1998). They are also cross-linked either in the seeds or after ingested. The cross-linked peptides can change their antigenic character (Dieterich et al., 1997). However, two other intrinsic properties of the storage proteins are not considered with their allergenicity. Firstly, these proteins contain large disordered regions in their sequences. More than 70% of the allergen-representative peptides and binding sites are present in the disordered regions of allergenic proteins (Xue et al., 2011). This may help in predicting the allergenic sites. Secondly, many of the storage proteins and proteins from fungal and bacterial pathogens share a common “cupin fold” in their structure (Dunwell et al., 2000). This raises the possibility of seed storage proteins mimicking either the sequences or conformations of microbial immunogens to induce a memory immune response (Table S6). This cross-reactive injury can result in the absence of the immunogen in a hit-and-run manner.

Our study incorporates the use of various bioinformatics tools and software in predicting the secondary and tertiary structures, disorderliness, and allergenic peptides of seed storage proteins from cereal and legumes. Therefore, the further use of *in vitro* and *in vivo* based studies will help to develop a better understanding of these proteins in relation to their structure, digestibility, and allergenicity.

The biasedness in amino acid frequencies among cereal and legumes as discussed in our manuscript is related to their unstructure, hydrophobicity, allergenicity, and predicted immune response. Presently, there is no such evidence that describes the details of seed storage protein in terms of amino acid frequency, composition, secondary/tertiary structure predictions, disorder, enzyme digestibility, allergenicity,

and possible folds-based classification. However, the evidence at the present stage is only predicted. The pieces of evidence presented in our study are crucial. The result findings may help researchers understand the structure, amino acid propensities, protein interactions, behavior of allergenic molecules, and immunological characteristics of storage proteins.

Conclusion

The seed storage proteins of cereals and legumes are the main source of protein for human consumption. The cereal proteins have high proline content to assume PP II conformation. This extended form enables hydration of the main chain and crosslinking by glutamine and cysteine. Hydrated and cross-linked proteins are viscoelastic. The legume proteins lack this property because they have less proline, cysteine, and glutamine. The modified amino acids and protein crosslinking changes the immunogenic property of liberated peptides. Predicting the immunological characteristics of these proteins will help a large human population with food allergies.

Acknowledgements

We would like to acknowledge Dr. A. Srinivasan, Ex-Professor, Department of Biophysics, AIIMS, Delhi for initial intellectual assistance for this study.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

We would like to acknowledge the fellowship grant from Department of Health Research, Ministry of Health & Family Welfare (R.12013/07/2019-HR) to PS at the time of writing this manuscript. However, no any funding has been received for conducting, writing, and publication of this work.

Data availability statement

The data supporting the findings of this manuscript are available within the paper and its supplementary files. Additional data will be made available from the corresponding author upon reasonable request.

References

Alberti, E., Humpfer, E., Spraul, M., Gilbert, S. M., Tatham, A. S., Shewry, P. R., & Gil, A. M. (2001). A high resolution (1)H magic angle spinning NMR study of a high-M(r) subunit of wheat glutenin. *Biopolymers*, *58*(1), 33–45. https://doi.org/10.1002/1097-0282(200101)58:1<33::AID-BIP40>3.0.CO;2-X

Astwood, J. D., Leach, J. N., & Fuchs, R. L. (1996). Stability of food allergens to digestion in vitro. *Nature Biotechnology*, *14*(10), 1269–1273. https://doi.org/10.1038/nbt1096-1269

Bannon, G., Fu, T.-J., Kimber, I., & Hinton, D. M. (2003). Protein digestibility and relevance to allergenicity. *Environmental Health Perspectives*, *111*(8), 1122–1124. https://doi.org/10.1289/ehp.5812

Battais, F., Mothes, T., Moneret-Vautrin, D. A., Pineau, F., Kanny, G., Popineau, Y., Bodinier, M., & Denery-Papini, S. (2005). Identification of IgE-binding epitopes on gliadins for patients with food allergy to wheat. *Allergy*, *60*(6), 815–821. https://doi.org/10.1111/j.1398-9995.2005.00795.x

Belton, P. S., Colquhoun, I. J., Grant, A., Wellner, N., Field, J. M., Shewry, P. R., & Tatham, A. S. (1995). FTIR and NMR studies on the hydration of a high-M(r) subunit of glutenin. *International Journal of Biological Macromolecules*, *17*(2), 74–80. https://doi.org/10.1016/0141-8130(95)93520-8

Björklund, A. K., Soeria-Atmadja, D., Zorzet, A., Hammerling, U., & Gustafsson, M. G. (2005). Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins. *Bioinformatics (Oxford, England)*, *21*(1), 39–50. https://doi.org/10.1093/bioinformatics/bth477

Capocchi, A., Cinollo, M., Galleschi, L., Saviozzi, F., Calucci, L., Pinzino, C., & Zandomeneghi, M. (2000). Degradation of gluten by proteases from dry and germinating wheat (*Triticum durum*) seeds: An in vitro approach to storage protein mobilization. *Journal of Agricultural and Food Chemistry*, *48*(12), 6271–6279. https://doi.org/10.1021/jf0006170

Chellgren, B. W., Miller, A. F., & Creamer, T. P. (2006). Evidence for polyproline II helical structure in short polyglutamine tracts. *Journal of Molecular Biology*, *361*(2), 362–371. https://doi.org/10.1016/j.jmb.2006.06.044

Chrispeels, M. J., Higgins, T. J., Craig, S., & Spencer, D. (1982). Role of the endoplasmic reticulum in the synthesis of reserve proteins and the kinetics of their transport to protein bodies in developing pea cotyledons. *The Journal of Cell Biology*, *93*(1), 5–14. https://doi.org/10.1083/jcb.93.1.5

Darnell, G., Orgel, J. P. R. O., Pahl, R., & Meredith, S. C. (2007). Flanking polyproline sequences inhibit beta-sheet structure in polyglutamine segments by inducing PPII-like helix structure. *Journal of Molecular Biology*, *374*(3), 688–704. https://doi.org/10.1016/j.jmb.2007.09.023

Dieterich, W., Ehnis, T., Bauer, M., Donner, P., Volta, U., Riecken, E. O., & Schuppan, D. (1997). Identification of tissue transglutaminase as the autoantigen of celiac disease. *Nature Medicine*, *3*(7), 797–801. https://doi.org/10.1038/nm0797-797

Drozdetskiy, A., Cole, C., Procter, J., & Barton, G. J. (2015). JPred4: A protein secondary structure prediction server. *Nucleic Acids Research*, *43*(W1), W389–94. https://doi.org/10.1093/nar/gkv332

Dunwell, J. M., Khuri, S., & Gane, P. J. (2000). Microbial relatives of the seed storage proteins of higher plants: Conservation of structure and diversification of function during evolution of the cupin superfamily. *Microbiology and Molecular Biology Reviews: MMBR*, *64*(1), 153–179. https://doi.org/10.1128/MMBR.64.1.153-179.2000

Fiers, M. W. E. J., Kleter, G. A., Nijland, H., Peijnenburg, A. A. C. M., Nap, J. P., & van Ham, R. C. H. J. (2004). Allermatch, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. *BMC Bioinformatics*, *5*(1), 133. https://doi.org/10.1186/1471-2105-5-133

Gasteiger, E., Hoogland, C., Gattiker, A., et al. (2005). *Protein identification and analysis tools on the ExPASy server*. Humana Press.

Hausch, F., Shan, L., Santiago, N. A., Gray, G. M., & Khosla, C. (2002). Intestinal digestive resistance of immunodominant gliadin peptides. *American Journal of Physiology. Gastrointestinal and Liver Physiology*, *283*(4), G996–G1003. https://doi.org/10.1152/ajpgi.00136.2002

Helmerhorst, E. J., Zamakhchari, M., Schuppan, D., & Oppenheim, F. G. (2010). Discovery of a novel and rich source of gluten-degrading microbial enzymes in the oral cavity. *PloS ONE*, *5*(10), e13264. https://doi.org/10.1371/journal.pone.0013264

Jain, A. (2023). Seed storage protein, functional diversity and association with allergy. *Allergies*, *3*(1), 25–38. https://doi.org/10.3390/allergies3010003

Jin, T., Ito, Y., Luan, X., Dangaria, S., Walker, C., Allen, M., Kulkarni, A., Gibson, C., Braatz, R., Liao, X., & Diekwisch, T. G. H. (2009). Elongated polyproline motifs facilitate enamel evolution through matrix subunit compaction. *PLoS Biology*, *7*(12), e1000262. https://doi.org/10.1371/journal.pbio.1000262

Johnson, K. D., Herman, E. M., & Chrispeels, M. J. (1989). An abundant, highly conserved tonoplast protein in seeds. *Plant Physiology*, *91*(3), 1006–1013. https://doi.org/10.1104/pp.91.3.1006

Kagnoff, M. F. (2005). Overview and pathogenesis of celiac disease. *Gastroenterology*, *128*(4 Suppl 1), S10–S18. https://doi.org/10.1053/j.gastro.2005.02.008

Kelly, M. A., Chellgren, B. W., Rucker, A. L., Troutman, J. M., Fried, M. G., Miller, A. F., & Creamer, T. P. (2001). Host-guest study of left-handed polyproline II helix formation. *Biochemistry*, *40*(48), 14376–14383. https://doi.org/10.1021/bi011043a

Kim, S. B., Ki, K. S., Khan, M. A., Lee, W. S., Lee, H. J., Ahn, B. S., & Kim, H. S. (2007). Peptide and tryptic hydrolysis of native and whey protein to reduce its allergenicity. *Journal of Dairy Science*, *90*(9), 4043–4050. https://doi.org/10.3168/jds.2007-0169

King, T. P., Hoffman, D., Lowenstein, H., Marsh, D. G., Platts-Mills, T. A., & Thomas, W. (2009). Allergen nomenclature. *International Archives of Allergy and Immunology*, *105*(3), 224–233. https://doi.org/10.1159/000236761

Knudsen, B., Knudsen, T., Flensborg, M., Sandmann, H., Heltzen, M., Andersen, A., Dickenson, M., Bardram, J., Steffensen, P. J., & Mønsted, S. (2011). Clc sequence viewer. A/S Cb, Version 6(2). https://resources.qiagenbioinformatics.com/manuals/clcsequenceviewer/current/index.php?manual=Introduction_CLC_Sequence_Viewer.html

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, *157*(1), 105–132. https://doi.org/10.1016/0022-2836(82)90515-0

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., & Russell, R. B. (2003). Protein disorder prediction: Implications for structural proteomics. *Structure (London, England: 1993)*, *11*(11), 1453–1459. https://doi.org/10.1016/j.str.2003.10.002

Lombardi, A., Barbante, A., Cristina, P. D., Rosiello, D., Castellazzi, C. L., Sbano, L., Masci, S., & Ceriotti, A. (2009). A relaxed specificity in interchain disulfide bond formation characterizes the assembly of a low-molecular-weight glutenin subunit in the endoplasmic reticulum. *Plant Physiology*, *149*(1), 412–423. https://doi.org/10.1104/pp.108.127761

Maleki, S. J., Kopper, R. A., Shin, D. S., Park, C. W., Compadre, C. M., Sampson, H., Burks, A. W., & Bannon, G. A. (2000). Structure of the major peanut allergen Ara h 1 may protect IgE-binding epitopes from degradation. *Journal of Immunology (Baltimore, Md.: 1950)*, *164*(11), 5844–5849. https://doi.org/10.4049/jimmunol.164.11.5844

Malik, A. H., Prieto-Linde, M. L., & Kuktaite, R. (2011). Individual and interactive effects of cultivar maturation time, nitrogen regime and temperature level on accumulation of wheat grain proteins. *Journal of the Science of Food and Agriculture*. *91*(12), 2192–2200.

Maruyama, N., Ichise, K., Katsube, T., Kishimoto, T., Kawase, S., Matsumura, Y., Takeuchi, Y., Sawada, T., & Utsumi, S. (1998). Identification of major wheat allergens by means of the *Escherichia coli* expression system. *European Journal of Biochemistry*, *255*(3), 739–745. https://doi.org/10.1046/j.1432-1327.1998.2550739.x

Matsuo, H., Kohno, K., Niihara, H., & Morita, E. (2005). Specific IgE determination to epitope peptides of ω-5 gliadin and high molecular weight glutenin subunit is a useful tool for diagnosis of wheat-dependent exercise-induced anaphylaxis. *Journal of Immunology (Baltimore, Md.: 1950)*, *175*(12), 8116–8122. https://doi.org/10.4049/jimmunol.175.12.8116

Maurel, C., Kado, R. T., Guern, J., & Chrispeels, M. J. (1995). Phosphorylation regulates the water channel activity of the seed-specific aquaporin alpha-TIP. *The EMBO Journal*, *14*(13), 3028–3035. https://doi.org/10.1002/j.1460-2075.1995.tb07305.x

Michaelsen, K. F., Hoppe, C., Roos, N., Kaestel, P., Stougaard, M., Lauritzen, L., Mølgaard, C., Girma, T., & Friis, H. (2009). Choice of foods and ingredients for moderately malnourished children 6 months to 5 years of age. *Food and Nutrition Bulletin*, *30*(3 Suppl), S343–S404. https://doi.org/10.1177/15648265090303S303

Navarro, V. (2012). The world health situation. *FAO Cereal Supply Demand Brief. World Food Situation*, *34*(1), 1–10.

Orsi, A., Sparvoli, F., & Ceriotti, A. (2001). Role of individual disulfide bonds in the structural maturation of a low molecular weight glutenin subunit. *The Journal of Biological Chemistry*, *276*(34), 32322–32329. https://doi.org/10.1074/jbc.M103833200

Prevelige, P., Jr., & Fasman, G. D. (1989). *Chou-Fasman prediction of the secondary structure of proteins: The Chou-Fasman-Prevelige algorithm. InPrediction of protein structure and the principles of protein conformation* (pp. 391–416). Springer US.

Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E. H., Man, O., Beckmann, J. S., Silman, I., & Sussman, J. L. (2005). Foldindex: A simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics (Oxford, England)*, *21*(16), 3435–3438. https://doi.org/10.1093/bioinformatics/bti537

Saha, S., & Raghava, G. P. (2006). Algpred: Prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Research*, *34*(Web Server issue), W202–W209. https://doi.org/10.1093/nar/gkl343

Sen, M., Kopper, R., Pons, L., Abraham, E. C., Burks, A. W., & Bannon, G. A. (2002). Protein structure plays a critical role in peanut allergen stability and may determine immunodominant IgE binding epitopes. *Journal of Immunology (Baltimore, Md.: 1950)*, *169*(2), 882–887. https://doi.org/10.4049/jimmunol.169.2.882

Shewry, P. R., & Halford, N. G. (2002). Cereal seed storage proteins: Structures, properties and role in grain utilization. *Journal of Experimental Botany*, *53*(370), 947–958. https://doi.org/10.1093/jexbot/53.370.947

Shewry, P. R., Napier, J. A., & Tatham, A. S. (1995). Seed storage proteins: Structures and biosynthesis. *The Plant Cell*, *7*(7), 945–956. https://doi.org/10.1105/tpc.7.7.945

Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V. N., Obradovic, Z., & Dunker, A. K. (2007). Disprot: The database of disordered proteins. *Nucleic Acids Research*, *35*(Database issue), D786–93. https://doi.org/10.1093/nar/gkl893

Sjöström, H., Lundin, K. E., Molberg, O., Körner, R., McAdam, S. N., Anthonsen, D., Quarsten, H., Norén, O., Roepstorff, P., Thorsby, E., & Sollid, L. M. (1998). Identification of a gliadin T-cell epitope in coeliac disease: General importance of gliadin deamidation for intestinal T-cell recognition. *Scandinavian Journal of Immunology*, *48*(2), 111–115. https://doi.org/10.1046/j.1365-3083.1998.00397.x

Sreerama, N., & Woody, R. W. (1999). Molecular dynamics simulations of polypeptide conformations in water: A comparison of alpha, beta, and poly(pro)II conformations. *Proteins: Structure, Function, and Genetics*, *36*(4), 400–406. https://doi.org/10.1002/(SICI)1097-0134(19990901)36:4<400::AID-PROT3>3.0.CO;2-B

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., & Birney, E. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, *12*(10), 1611–1618. https://doi.org/10.1101/gr.361602

Stapley, B. J., & Creamer, T. P. (1999). A survey of left-handed polyproline II helices. *Protein Science: A Publication of the Protein Society*, *8*(3), 587–595. https://doi.org/10.1110/ps.8.3.587

Stott, K., Blackburn, J. M., Butler, P. J., & Perutz, M. (1995). Incorporation of glutamine repeats makes protein oligomerize: Implications for neurodegenerative diseases. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(14), 6509–6513. https://doi.org/10.1073/pnas.92.14.6509

Tanabe, S. (2001). Identification of wheat allergens. *InInternet Symposium on Food Allergens*, *3*(4), 163–170.

Tanabe, S. (2007). Epitope peptides and immunotherapy. *Current Protein & Peptide Science*, *8*(1), 109–118. https://doi.org/10.2174/138920307779941569

Tatham, A. S., Marsh, M. N., Wieser, H., & Shewry, P. R. (1990). Conformational studies of peptides corresponding to the coeliac-activating regions of wheat α-gliadin. *The Biochemical Journal*, *270*(2), 313–318. https://doi.org/10.1042/bj2700313

Triboi, E., Abad, A., Michelena, A., Lloveras, J., Ollier, J. L., & Daniel, C. (2000). Environmental effects on the quality of two wheat genotypes:

1. Quantitative and qualitative variation of storage proteins. *European Journal of Agronomy*, *13*(1), 47–64. https://doi.org/10.1016/S1161-0301(00)00059-9

Uversky, V. N., Gillespie, J. R., & Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Genetics*, *41*(3), 415–427. https://doi.org/10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7

Vitale, A., & Denecke, J. (1999). The endoplasmic reticulum—Gateway of the secretory pathway. *Plant Cell*. *11*(4), 615–628. https://doi.org/10.2307/3870888

Xue, B., Soeria-Atmadja, D., Gustafsson, M. G., Hammerling, U., Dunker, A. K., & Uversky, V. N. (2011). Abundance and functional roles of intrinsic disorder in allergenic proteins and allergen representative peptides. *Proteins*, *79*(9), 2595–2606. https://doi.org/10.1002/prot.23077

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: Protein structure and function prediction. *Nature Methods*, *12*(1), 7–8. https://doi.org/10.1038/nmeth.3213

Zhu, J., & Khan, K. (2001). Effects of genotype and environment on glutenin polymers and breadmaking quality. *Cereal Chemistry*, *78*(2), 125–130. https://doi.org/10.1094/CCHEM.2001.78.2.125