JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Article

pubs.acs.org/JAFC

Farinin: Characterization of a Novel Wheat Endosperm Protein Belonging to the Prolamin Superfamily

Donald D. Kasarda,* Elva Adalsteins, Ellen J.-L. Lew, Gerard R. Lazo, and Susan B. Altenbach

Western Regional Research Center, Agricultural Research Service, U.S. Department of Agriculture, 800 Buchanan Street, Albany, California 94710, United States

Supporting Information

ABSTRACT: Starch granule surface-associated proteins were separated by HPLC and identified by direct protein sequencing. Among the proteins identified was one that consisted of two polypeptide chains of 11 and 19 kDa linked by disulfide bonds. Sequencing of tryptic peptides from each of the polypeptides revealed similarities between some of the peptides and avenin-like b proteins encoded by partial cDNAs in NCBI. To identify a contiguous sequence that matched all of the peptides, contigs encoding three avenin-like b proteins were constructed from ESTs of the cultivar Butte 86. All peptide sequences were found in a protein encoded by one of these contigs that had not been identified previously. Protein and DNA sequences indicated that the two polypeptide chains were derived from a parent protein that had been cleaved at the C-terminal position of an asparagine residue. The name farinin is suggested for this protein and other avenin-like b proteins. Evolutionary relationships of the protein are discussed and a simple computer molecular model was constructed. On the basis of its sequence, the new protein was likely to be allergenic but unlikely to be active in celiac disease.

KEYWORDS: *Triticum aestivum, gluten proteins, protein sequences, DNA sequences, asparagine endoproteinase, farinins, purinins, wheat allergy, celiac disease*

INTRODUCTION

Wheat endosperm is made up of two main components, starch granules (about 70%) and storage proteins (about 12%). Lipids, other proteins, and nonstarch polysaccharides are also found in the endosperm, but are present in lesser proportions. Most of the storage proteins in the endosperm are classified as gluten, which is mainly responsible for the important viscoelastic properties of wheat flour doughs that are key to their unique ability to produce leavened breads. Gluten has two major subfractions: gliadins, monomeric proteins, and glutenins, polymeric proteins formed by intermolecular disulfide bonding of two main groups of subunits, the low-molecular-weight glutenin subunits (LMW-GS) and high-molecular-weight glutenin subunits (HMW-GS). Gliadins are divided into the α-gliadin, γ-gliadin, and ω-gliadin subfamilies.

ω-Gliadins have essentially a single domain made up of slightly diverged repeating sequences rich in glutamine and proline. ω-Gliadins usually have no cysteine residues and, hence, no disulfide bonds. α-Gliadins, γ-gliadins, and the LMW-GS have two major, approximately equal, domains—an N-terminal domain made up of repeating sequences somewhat similar to those of ω-gliadins and a C-terminal domain that is nonrepetitive, has a lower content of glutamine and proline, and includes three or four intramolecular disulfide bonds that link specific cysteine residues. This latter domain is therefore likely to be more structured and compact than the repeating sequence domain. The LMW-GS also have two cysteine residues that form intermolecular disulfide bonds—along with the six cysteines that form three intramolecular disulfide bonds.[1] The HMW-GS have a large central repeating sequence domain book-ended by small unique sequence domains at the N-terminal and C-terminal ends of the protein; these latter domains contain almost all the cysteine residues that participate in intra- and intermolecular bonding.

Additional grain protein components in the MW range below 30K that are gliadin-related have been described and sometimes called either low-molecular-weight gliadins,[2−4] or globulins.[4−6] The sequences of these smaller proteins indicate a genetic relationship to the unique sequence domains of gluten proteins and are notable for the absence of significant domains made up of repeats, whereas such repeating sequence domains are found in all traditional gluten proteins. Genetic relationships of these proteins to one another and to the gluten proteins have been explored to some extent,[3,7,8] but further work is needed to establish a clear context for them in relation to the traditional gluten proteins. Most of these nontraditional gluten proteins fall into the categories of avenin-like proteins or proteins with sequences similar to the 07h10 component of Anderson et al.,[3] which were classed as globulins by Gomez et al.[5,6] These latter proteins were also recognized by Skylas et al.[9] on the basis of N-terminal sequences similar to RTAWEPQH-, but not assigned to a class.

In a previous study, surface proteins from commercial wheat starches were extracted with the strongly denaturing solvent 2% sodium dodecylsulfate (SDS) and it was reported that starch surface proteins are a complex mixture of endosperm proteins that have become adsorbed to the starch surface, either during grain development or during the extraction and processing of the starch.[10] In a parallel study (reported here), a simple,

Received: December 13, 2012
Revised: February 1, 2013
Accepted: February 15, 2013
Published: February 15, 2013

ACS Publications This article not subject to U.S. Copyright. Published 2013 by the American Chemical Society

dx.doi.org/10.1021/jf3053466 | J. Agric. Food Chem. 2013, 61, 2407−2417

largely nondenaturing solvent, 0.1% trifluoroacetic acid (TFA), was investigated as an extractant for the wheat starch surface proteins. TFA is acidic and the denaturing effects (probably largely unfolding) of acidic solutions will be present, but TFA, unlike SDS, is not likely to bind strongly to proteins. This simple solvent is highly compatible with HPLC purification methods and essentially solvent-free proteins can readily be recovered from fractions by direct freeze-drying or vacuum drying in contrast to the more difficult task of removing detergent. This simple method was used to purify proteins from laboratory-prepared wheat starches derived from several different cultivars and the characterization of these proteins with particular focus on a novel avenin-like protein is described. We propose that this novel avenin-like protein be called a *farinin* and that the name might be applied to other avenin-like proteins, as well.

MATERIALS AND METHODS

Flour Samples and Laboratory Starch Preparation. Grain of the hard red winter wheat (*Triticum aestivum*) cv. Scout 66 was provided by Virgil A. Johnson, USDA and University of Nebraska. Grain of the hard red winter wheat cv. Cheyenne was provided by Calvin O. Qualset, University of California, Davis. Flour (pure endosperm) of the soft white winter club wheat (*T. aestivum*) cv. Paha was provided by Craig Morris and Arthur Bettge, USDA and Washington State University, Pullman, WA. Grain was milled with a Brabender Quadrumat Senior mill (C. W. Brabender Instruments, Inc., South Hackensack, NJ) after tempering to 15% moisture; yield was about 65%.

Starches from flours of single cultivars were prepared by the method of Wolf,[11] which is based on manual dough formation and washing of the starch granules from the dough with distilled water. The suspended starch granules were recovered by passing the suspension through Number 20XX nylon bolting cloth to remove gluten particles and other nonstarch material. After low-speed centrifugation (3000*g*) of the starch suspension, the upper, pigmented layer was scraped off the starch pellet and discarded to avoid contamination of the starch. The starch was resuspended in water and the process repeated 3 times. The final starch product was dried at room temperature with circulating air. Clumps were disrupted frequently to prevent formation of difficult-to-disperse aggregates.

Protein Extraction from Starch and Reverse-Phase HPLC. Proteins were extracted from the starch samples with 0.1% HPLC-purity TFA, usually at a 10:1 ratio of solvent to starch, but with ratios ranging from 4:1 to 10:1. The TFA protein extract was clarified by low-speed centrifugation, freeze-dried, and taken up in 3−5 mL of 6 M guanidinium chloride (ratio of solvent to TFA extract ranged from 40:1 to 65:1), filtered through a 5 μm filter to clarify the solution and to reduce viscosity (usually the filters had to be replaced multiple times as the solution tended to clog the filters, perhaps because of dissolved starch and other carbohydrates), and loaded onto the reverse-phase column. Reverse-phase HPLC was carried out with a Vydac (Hesperia, CA) 218TP54 C18 column (semipreparative) on Hewlett-Packard (San Jose, CA) or ThermoSeparations Instruments, Inc. (Riviera Beach, FL). Solvent A was 0.05% TFA; Solvent B was acetonitrile and 0.05% TFA. Gradients ranged from 10 to 50% acetonitrile to 10−90% acetonitrile as appropriate for particular separations. Monitoring of the separation was at a wavelength of 215 nm.

Electrophoresis and Protein Sequencing. The method used for 1D SDS−PAGE was basically that of Laemmli.[12] Sequencing of proteins and peptides was carried out by Edman degradation with an Applied Biosystems 477A sequencer (Applied Biosystems, Foster City, CA). The sequencer employed an online HPLC system for identification of the phenylthiohydantoin derivatives of the amino acids. The analysis of total extracts of Butte 86 flour by 2D electrophoresis and mass spectrometry was carried out as previously described.[13,14] The determination of the amino acid composition of one protein, eventually established to be tritin, was carried out at the University of California, Davis, Molecular Structure Facility using standard methods. A single MALDI-MS analysis of the peak 7 protein (tritin) was carried out by Dr. K. J. Wu (Charles Evans & Associates, Redwood City, CA) with a custom-built mass spectrometer and standard methods.

Derivation of Coding Sequences from ESTs Obtained from *T. aestivum* 'Butte 86'. Butte 86 ESTs with similarity to AF470351 and GU211171 were identified by searching with the BLASTN algorithm and downloaded from NCBI. ESTs were assembled with Lasergene Seqman Pro software (DNASTAR, Inc., Madison, WI) using the Classic Assembler with default settings except that the minimum match percentage was set to 98. Assemblies were inspected manually and mismatches that occurred in overlap regions of ESTs were resolved by examining quality scores for individual ESTs as detailed in Altenbach et al.[15] DNA consensus sequences, shown in Supporting Information Figure S1, were translated using functions within the Lasergene software. Cleavages of signal peptides were predicted using the SignalP 3.0 Server (http://www.cbs.dtu.dk/services/SignalP/). MW's and pI's of deduced proteins were calculated using the Protein Parameter tool (http://web.expasy.org/protparam/) found on the ExPASy Proteomics Server. Sequence alignments were performed using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) with default settings.

Molecular Modeling. Initial modeling of the farinin structure was carried out by the I-TASSER server (http://zhanglab.ccmb.med.umich.edu/I-TASSER/), which provides a composite approach to protein structure determination that is based on threading programs, ab initio modeling, and atomic level structure refinement.[16,17] The amino acid sequence used for the model was from this paper (farinin Bu-1). External restraints on the structure were included. These restraints corresponded to the hypothesized disulfide linkages based on the published arrangements for γ-gliadins;[18,19] a target distance of 2 Angstroms was set for the pairs of SG atoms proposed to participate in disulfide bonds. The constraints for gamma-gliadins were chosen because the disulfide arrangements are known for gamma-gliadins, but not for any other gluten protein that has 8 cysteines in the form of 4 disulfide bonds. Of the various possible models presented by I-TASSER, the model with the most disulfide linkages (9 linkages) was chosen for further testing and refinement. To test its stability, the model was heated to 295 K, subjected to an equilibration cycle of 5 ps, and energy minimized with the Quanta 4.0-CHARMm (version 23.1) software (Accelrys Software, San Diego, CA) running on a Silicon Graphics O_2 computer. The PDB (Protein Data Bank) coordinates of the resulting model were transferred to the program RasMol (http://rasmol.org/) for display and the corresponding RasMol cartoon representation was used for Figure 6.

Sequence Alignments, Comparisons, and Phylogenetic Analysis. Sequences were compared mainly with the NCBI BLAST program (blast.ncbi.nlm.nih.gov). In BLAST comparisons, default parameters were used except for two changes in the algorithm parameters: the parameter 'Compositional Adjustments' was set to 'No Adjustments' and no filtering of low-complexity regions was applied (personal communication, Peter Cooper, NCBI) in order to improve performance for gluten proteins. Phylogeny trees were constructed by combinations of programs at the Phylogeny project Web site (http://www.phylogeny.fr/; see: Dereeper et al.[20]).

RESULTS AND DISCUSSION

HPLC Fractionation of Proteins from Various Starches. An example of fractionation by HPLC of starch surface-associated proteins from laboratory-prepared starch of cv. Scout 66 is shown in Figure 1. Extraction of starch from cv. Cheyenne gave a similar pattern (data not shown). Major peaks (numbered in Figure 1) were collected and subjected to 1D SDS−PAGE, which showed a predominant band for each numbered peak fraction, along with several minor bands (data not shown). The proteins corresponding to the collected peaks were then subjected to N-terminal protein sequencing by

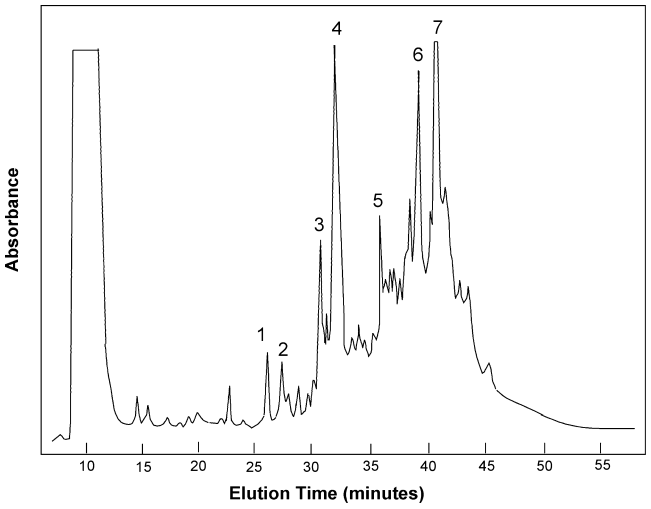


Figure 1. HPLC fractionation of proteins extracted with 0.1% TFA from starch of the cv. 'Scout 66.' The numbered peaks were collected and subjected to protein sequencing. Absorbance monitored at 215 nm.

Edman degradation for at least ten cycles. With the exception of peaks 6 and 7 (Figure 1), which were blocked to the Edman degradation, the protein fractions yielded significant sequence information (Table 1), usually with one clear major sequence evident. For Peaks 3 and 4, however, there were two major amino acids identified in almost equal amounts for each of the ten cycles analyzed (Table 1). Such a double sequence for a fractionated protein is characteristic of two polypeptide chains linked together. Identifications were made for Peaks 1, 2, and 5 by using the NCBI-BLASTP program (blast.ncbi.nlm.nih.gov) to search the NCBI nonredundant protein database. Peaks 1 and 2 were identified as α/β purothionin and Peak 5 was identified as a chitinase. Additional work was necessary to identify the proteins in Peaks 3, 4, 6, and 7. It was noted that HPLC patterns of extracts from commercial starches were much less well-defined (data not shown), which may be a consequence of protein damage associated with the harsher treatments accorded the commercial starches—especially high temperatures in the drying process.

The proteins corresponding to peaks analyzed from the experiment of Figure 1 were chosen on the basis of peak height, area, and degree of resolution. When more than one peak was identified for a given protein type, focus was usually on the largest of the peaks for characterization of the corresponding proteins—as was the case for Peak 4 versus Peak 3. In related experiments, evidence was noted that reanalysis of peaks by HPLC to enhance purity resolved minor peaks corresponding to other proteins (a rerun of Peak 4, for example, yielded, additionally, a well resolved peak that corresponded to a lipid transfer protein on the basis of its N-terminal sequence). It was clear that the method described here has the potential to resolve many more proteins than those selected for characterization in this study. Other proteins (mostly from cv. Paha) identified in similar experiments are shown in Table 2 and include peroxidase, lipid transfer protein, and puroindolines A and B. These were not characterized further.

Identification and Characterization of Peaks 6 and 7 as Tritin. The proteins of Peaks 6 and 7 yielded no sequence in the Edman degradation, indicating that they were blocked at their N-termini. The mass of the Peak 7 protein as measured by MALDI-MS was 30 560 Da. The proteins from Peaks 6 and 7 were digested with trypsin and the resulting peptides fractionated by HPLC. The following tryptic peptide sequences were obtained: WFHIVLK (Peaks 6 and 7), AQVNGWQDLS (Peak 7), EAVTTLLLMVHEATR (Peak 6) and QQMADAVTALYGR (Peak 7) indicating that Peaks 6 and 7 corresponded to the protein tritin[21] (BAA02948.1). Tritin, named by Coleman and Roberts[22] is a ribosome-inactivating protein from *T. aestivum*. Lowy et al.[23] purified a protein obtained from wheat starch by extraction with a sodium chloride solution. Their protein was blocked at the N-terminus, had a MW of about 30 000, was highly basic, was the principal protein in the NaCl extract, had no α-amylase activity, and did not have inhibitory effects on α-amylases from wheat or hog pancreas. The amino acid composition was close to that of our Peak 7 protein (data not shown). The unidentified protein of Lowy et al.[23] was almost certainly tritin.

The N-terminal residue of tritin is methionine; Habuka et al.[21] were able to sequence a cyanogen-bromide cleaved peptide beginning at residue 2. Consequently, it appears likely that a modified N-terminal methionine residue (such as *N*-acetyl methionine) is responsible for the failure of tritin to sequence in the Edman degradation.

Identification and Characterization of Peak 4 Proteins. Reduction of the Peak 4 protein followed by HPLC fractionation yielded two peaks. The SDS−PAGE patterns of unreduced Peak 4 protein and component peptides (designated A and B) resulting from reduction are shown in Figure 2. The apparent MW of the unreduced protein was approximately 24 000, while the A and B peptides had apparent MWs of 11 000 and 19 000, respectively. Sequencing of the proteins corresponding to the A and B peaks in the HPLC chromatogram gave predominantly a single amino acid at each cycle. Sufficient amounts of each peptide were prepared for tryptic digestion and HPLC purification of the tryptic peptides. The sequences of one tryptic peptide from the A component (T-A1) and three

Table 1. N-Terminal Sequence Analysis of Proteins Corresponding to Peaks in Figure 1[a]

peak	sequence	NCBI accession number	protein
1	KSXX(R/K)STLGR-	CAA65313, CAA65312	α,β-purothionin
2	KSXX(R/K)STLGR-	CAA65313, CAA65312	α,β-purothionin
3	(L/E)(E/P)(T/Q)(I/Q)(X/E)(S/A)(Q/H)-	AAP80612	farinin
4	(L/E)(E/P)(T/Q)(I/Q)(X/E)(S/A)(Q/H)-	ADA62375	farinin
5	SVSSVVSRRQFDRMLLHRND-	AAX83262	Chitinase, class II
6	*blocked to Edman degradation*	BAA02948	tritin
7	*blocked to Edman degradation*	BAA02948	tritin

[a]All proteins were from 'Scout 66'. Residues not identified in the Edman degradation, usually cysteine, were designated as X. Identifications of proteins in peaks 3, 4, 6, and 7 were made by digestion of proteins with trypsin, separation of peptides and sequencing of internal peptides (see text).

Table 2. Additional Proteins Identified in TFA Extracts of Wheat Starch[a]

cultivar	sequence	NCBI accession number	protein
Paha	AEPPVARGLS-	AAM88383	peroxidase I
Paha	KSXX(R/K)STLGR-	CAA65313, CAA65312	α,β-purothionin
Paha	IDCGHVDSLVRPCLSYVQGGPGPSGQCCD-[b]	CAH04989	type 1 nonspecific lipid transfer protein
Paha	DVAGGGGAQQ-	AAB28037	puroindoline A
Paha	EVGGGGGSQQ-	CAP20331	puroindoline B
Scout 66	EVGGGGGSQQCPQER-[b]	CAP20331	puroindoline B

[a]Residues not identified in the Edman degradation, usually cysteine, were designated as X. [b]Protein was alkylated before sequencing in order to identify cysteine residues.

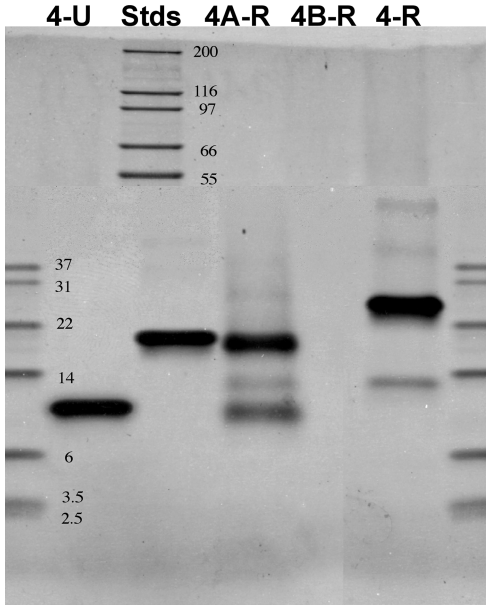


Figure 2. SDS–PAGE of Peak 4 and its component peptides: lane 4-U, purified Peak 4 (unreduced); lane 4A-R, A peptide purified from reduced Peak 4; lane 4B-R, B peptide purified from reduced Peak 4; lane 4-R, Peak 4 reduced.

tryptic peptides from the B component (T-B1, T-B2, and T-B3) were obtained (Figure 3). The best match in the NCBI nonredundant protein database to all but B-T3 was AAP80612, an unidentified protein from *T. aestivum* that was deduced from the cDNA sequence AF470351. B-T3 was an identical match to ADA62375, deduced from the partial cDNA sequence GU211171, corresponding to an avenin-like b protein. Because a contiguous sequence that matched all of the Peak 4 peptides was not obtained, the NCBI collection of ESTs was queried with AF470351 and GU211171 in hopes of identifying ESTs from a single cultivar that formed a contiguous sequence containing all of the peptides. Thirteen ESTs from the cv. Butte 86 were identified in the search (Supporting Information Table 1). These were assembled into three distinct contigs, referred to as Bu-1, Bu-2 and Bu-3 (Supporting Information Figure S1). The protein encoded by Bu-1 was a perfect match with all peptides identified from Peak 4 except T-B2 (Figure 3). T-B2 contained an arginine at position 29 that did not correspond to any sequence in NCBI.

On the basis of the protein encoded by Bu-1, the true mass of the mature Peak 4 protein (minus signal sequence) would be 29 978 Da, whereas that of the A peptide would be 11 164 and that of the B peptide, 18 832. The difference between the true MW and that determined by SDS–PAGE (24 000; Figure 2) most likely results from the unreduced protein having a more compact configuration as a consequence of it having multiple intramolecular disulfide bonds.

Comparison of our peptide sequences with the protein encoded by Bu-1 (Figure 3) led to the conclusion that the unreduced Peak 4 protein resulted from cleavage of a parent protein at an asparagine residue such that the two cleaved peptides remained attached to one another by one or more originally intramolecular (now interpeptide) disulfide bonds. It is interesting that both Bu-2 and Bu-3 contain a glutamine rather than an asparagine at this position, which indicates that the Bu-2 and Bu-3 forms are not cleaved. There was no indication in the preparations (from wheat starches) of the proteins corresponding to Bu-2 or Bu-3, which differ slightly in their N-terminal sequences from Bu-1, LETICSQGFG- versus LETTCSQGFG- (Figure 3); the difference of I (Ile) versus T (Thr) at position 4 would show clearly in N-terminal Edman sequencing. It may be that the cleaved Peak 4 protein is more strongly bound to starch granules than the intact proteins. Proteomic analyses[13,14] indicated that about 1% of the total protein in white flour from fertilized Butte 86 grain corresponded to contigs Bu-1, Bu-2, and Bu-3. The amounts of the noncleaved avenin-like proteins corresponding to Bu-2 and Bu-3 were about equal and slightly greater than that of Bu-1 (Peak 4), which was predominantly in the cleaved form. The genes for Bu1, Bu-2, and Bu3 may be located on the A, B, and D genomes of hexaploid wheat (one gene per genome), although we did not attempt to determine the chromosomal locations of these genes.

The N-terminal sequence of LETIC- indicates a possible secondary processing event in that the signal cleavage predicted by the Signal P server is between residues 19 and 20 (A and Q) so that the sequence we observed beginning with L might result from a secondary cleavage of the Q residue at position 20 from the sequence QLETIC, perhaps by the action of an aminopeptidase. Alternatively, the action of the signal cleavage peptidase may be degenerate, giving rise to two N-terminal sequences, LETIC- and QLETIC-. Direct amino acid sequencing would not proceed for the second sequence if the N-terminal glutamine was cyclized to the pyroglutamic form, which would not react with phenylisothiocyanate in the Edman degradation. In support of there being two N-terminal sequences, De Caro et al.[24] reported the N-terminal sequence for an avenin-like b protein from durum wheat as QLETTC-SQGFG- based on MS/MS sequencing, but the relative amounts of N-terminal sequences with and without Q (present as pyroE) as the first residue remains to be elucidated. The basis for this apparent degeneracy of the signal peptide cleavage is not understood, but it has also been observed by DuPont et al.[13] for LMW-GS.

Although Peak 4 protein from the cv. Scout 66 was used for most of the characterization work, proteins with identical

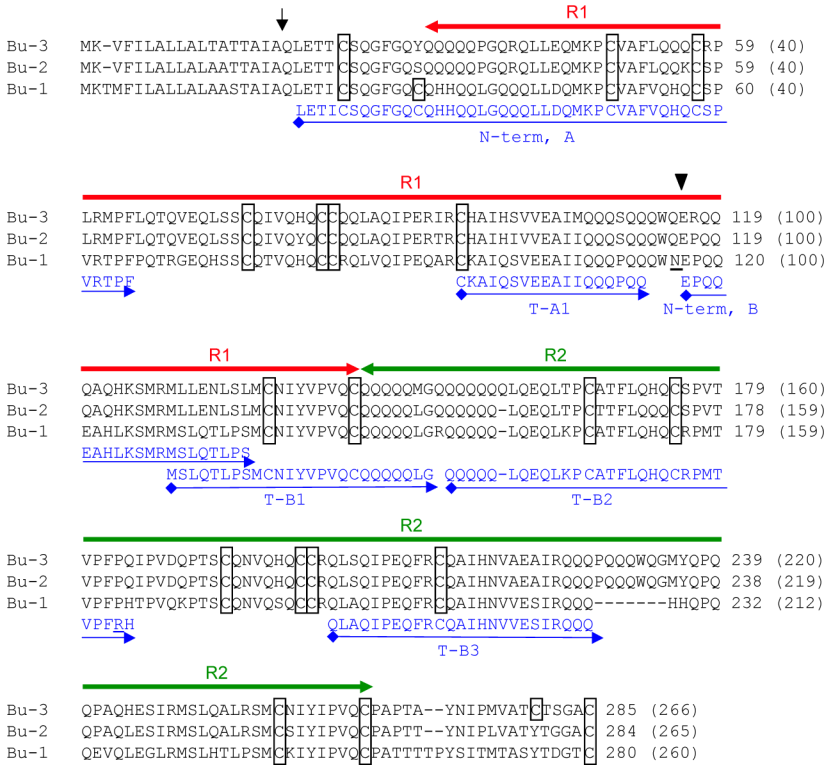


Figure 3. Comparison of the three protein sequences corresponding to the contig sequences, Bu-1, Bu-2, and Bu-3, derived from ESTs of cultivar Butte 86 with directly obtained sequences of peptides prepared from the Peak 4 protein of cultivar Scout 66. The red double-headed arrow indicates the sequence corresponding to domain R1. The green double-headed arrow indicates the sequence corresponding to the domain R2. The N-terminal and peptide sequences are in blue with T indicating peptides derived from tryptic digestion of the two intrinsic Peak 4 peptides, which are designated A and B. For example, TA-1 indicates a tryptic peptide derived from the intrinsic A peptide and T-B1 indicates a peptide derived from the intrinsic B peptide. The bold vertical arrowhead indicates the residue 96 asparagine (mature protein after signal cleavage) that is cleaved to yield the intrinsic peptides A and B. The vertical arrowhead indicates the predicted signal cleavage, which differs by one reside from the observed sequence in direct Edman sequencing. Sequence numbers are given at the end of each line for the complete sequences (including the signal sequence) followed by the sequence numbers (in parentheses) for the mature sequences.

double N-terminal sequences were found in extracts from starches prepared from the cvs. Cheyenne and Paha. Because of the need for fairly large amounts of protein, some preparations of purified Peak 4 from these other two cultivars were used in our amino acid sequencing analyses and thus the peptide sequences of Figure 3 might be considered as yielding a composite sequence. However, all the peptides we sequenced were identical to or showed only very minor differences from the Bu-1 sequence regardless of the cultivar used in their preparation.

Similarity to the Avenin-like b Proteins. The proteins encoded by Bu-1, Bu-2 and Bu-3 are similar to those of the avenin-like b proteins found in wheat and *Aegilops* species.[7,8] As is characteristic of the avenin-like b sequences, the proteins encoded by the Butte 86 contigs did not have a major repeating sequence domain such as is found in all the traditional gluten proteins. The reported avenin-like b sequences from wheat and *Aegilops* species did not have the asparagine residue found in Peak 4 and Bu-1 that is likely to be cleaved giving rise to the A and B chains of the protein. No entry in the NCBI database was identical to Bu-1, although Bu-2 and Bu-3 did have exact matches with avenin-like b sequences from wheat,[7] CAJ32659 and CAJ32655, respectively. The sequence for clone 12dc3 of Anderson et al.,[3] GU211171, apparently codes for a partial sequence of an avenin-like b protein; whereas the protein encoded by 09d3 of Anderson et al.,[3] the proteins of Clarke et al.,[4] and the LMW-gliadins of Salcedo et al.[2] apparently correspond to the closely related avenin-like a proteins described by Kan et al.[7]

The avenin-like b proteins have an unusual structure in which a particular domain that includes eight cysteine residues, similar to the unique sequence domain of gamma gliadins and other seed storage proteins,[25] is duplicated, resulting in two tandem domains[7] designated R1 and R2 (Figure 3). The R1 and R2 domains are also found in proteins encoded by Bu-1, Bu-2 and Bu-3; R1 and R2 have 60% identity in their aligned sequences (71% when positive positions were included). There was some variability in number (18 or 19) and/or position for the total number of cysteine residues of the avenin-like b sequences reported by Kan et al.[7] for the wheat cv. Cadenza and similar differences are found in the Butte 86 proteins. It is possible that the R1 and R2 domains resulted from an internal duplication;[7] both have a large segment deleted in comparison with avenin and gliadin sequences (data not shown) and it seems less likely that the two domains would both have this large deletion if the R1 and R2 domains had evolved separately. Of special note, the Bu-1 sequence (and the N-terminal tryptic peptide of the A peptide) has a cysteine residue at position 12 of the mature sequence that is missing from the avenin-like b proteins and contigs Bu-2 and Bu-3 (Figure 3), although an equivalent cysteine is found in the avenin-like a sequences of Kan et al.[7] The pair of cysteines at residues 5 and 12 found in Bu-1 is also characteristic of all avenin-like a sequences described by Kan et al.[7]

dx.doi.org/10.1021/jf3053466 | J. Agric. Food Chem. 2013, 61, 2407−2417

Avenin-like a sequences are similar to avenin-like b sequences except that one (R2) of the two seed storage protein domains (SS domain) is missing,[7] resulting in a lower MW of about 17 000 for the avenin-like a proteins and their SS domain has nine cysteines instead of the expected eight; the odd cysteine at the C-terminal position may not participate in the normal SS domain fold, but might link up with one of the two N-terminal cysteines of avenin-like a proteins that lie outside the SS domain.

Alternatively, the final cysteine might form an intermolecular disulfide bond. Kan et al.[7] speculated that avenin-like b proteins with an odd number of cysteines might be incorporated into polymers and this possibility was supported by De Caro et al.[24] who prepared their avenin-like b protein from a glutenin fraction. The Peak 4 protein, however, despite having 19 cysteine residues on the basis of the Bu-1 sequence (Figure 3) appeared to be present as a monomer since, unreduced, it gave a fairly distinct peak in the HPLC separation, with an SDS–PAGE-estimated MW of 24 000 for the purified protein (Figure 2). Possible disulfide arrangements will be discussed below in the modeling section and the likely free cysteine might be the final amino acid of the sequence. Because that region of the protein was not actually sequenced, the possibility of post-translational processing at the C-terminal end to remove the C-terminal cysteine cannot be ruled out.

Cleavage at Asparagine. DuPont et al.[13] in a proteomic analysis of a total (SDS + reducing agent) protein extract of flour from cv. Butte 86 found spots corresponding to: (a) the intact Bu-1 protein; (b) the 19 000 Da B peptide of Peak 4; and (c) the intact proteins encoded by Bu-2 and Bu-3. Thus, it appears that the cleavage of the protein corresponding to Bu-1 is partial. It is likely that an asparagine endopeptidase carries out the cleavage of the Bu-1 protein at residue 96 (an asparagine in the mature sequence; see Figure 3), giving rise to the A and B peptides. As expected for proteins corresponding to Bu-2 and Bu-3, which do not have asparagine at residue 96, no spots corresponding to the B peptide of Peak 4 (Figure 3) were observed in the proteomic study. The smaller A peptide of Bu-1/Peak 4 with mass of about 11 000 was not observed in the 2D patterns.

Cleavages at asparagine residues near the N-terminus of other wheat endosperm proteins have been reported. The mature (signal peptide-cleaved) D-type ω-gliadins (also known as ω-1,2 gliadins[26]) exhibit two different N-terminal sequences in approximately equal amounts: ARELNPQNKEL- and KELQSPQQSF-, with the latter sequence presumably resulting from cleavage of the former peptide at the asparagine in position 8;[27] the asparagine at 4 in the former peptide would not be expected to cleave because it is followed by proline residue. Cleavage at asparagine has also been proposed for the N-terminal processing of the immature polypeptide chain of certain LMW-GS.[28]

The cleavages at asparagines in the processing of endosperm proteins described above mostly have occurred close to the N-terminus. The cleavage of the Peak 4 protein, however, gave rise to two fairly large peptides, in accord with the suggestion here that the enzyme involved is an asparagine endopeptidase. A minor protein of wheat endosperm, triticin, that is related to the 12S storage globulins of legumes is also cleaved at an internal asparagine residue[29,30] and the protease involved in these cleavages may be related to the legumains,[31] which cleave storage globulins—primarily at asparagine residues.

Protein Nomenclature. The proteins characterized by Kan et al.[7] were designated "avenin-like," and this is correct, but it may be noted that their sequences are closer to some gliadins (see below: Sequence Comparisons, and Table 3) and could equally well be classed as gliadin-like. We suggest that a new name, *farinins* (from the Latin *far* = grain) be given to the avenin-like proteins of wheat including in the latter group, our Bu-1/peak 4 protein. Additionally, in order to add to the list of named endosperm proteins, we suggest that proteins of the 07h10 protein subfamily of Anderson et al.[3] be named *purinins* (from the Greek *puros* = grain[32]). The term *subfamily* was suggested by Xu and Messing[33] and we follow them in using this terminology. The purinins are even more closely related to the γ-gliadin C-terminal domain than to the avenin-like proteins (Table 3), while most other low-molecular-weight proteins of wheat, such as tritin, triticin, and the puroindolines are at best weakly related to the gliadins. The purinin proteins have been designated as LMW-gliadins,[3,4] or globulins.[5,6] The LMW-gliadins make up a somewhat heterogeneous group, but generally fall into the categories of farinins or purinins. A major exception is the group of α-amylase inhibitors, which will be discussed below. The purinins are coded by genes on group 1 chromosomes,[5] whereas the farinins are coded by genes on chromosomes of groups 4 and 7.[4,34]

Table 3. Similarity Comparisons of Seed Storage Protein Domain (SSP) and *Conserved I* Region Sequences (Con I) from the SSP Domain of Various Proteins by NCBI BLAST Multiple Alignment[a]

	score (bits)		identities (%)		positives (%)	
	SSP	Con 1	SSP	Con 1	SSP	Con 1
			Farinin R1 (query)			
avenin	93.6	66.2	37	49	50	65
γ-gliadin	85.5	51.6	37	38	48	50
α-gliadin	61.2	44.7	32	41	46	57
LMW-GS	75.9	66.6	34	48	49	68
purinin	75.9	56.6	36	45	53	65
farinin R2	140	94	62	61	74	76
			Avenin (query)			
farinin R1	93.6	66.2	37	49	50	65
γ-gliadin	157	90.5	56	68	66	77
α-gliadin	115	59.3	47	45	60	68
LMW-GS	133	77.8	53	61	63	75
purinin	130	81.3	49	57	62	78
farinin R2	99	78.6	38	57	48	69

[a] *Conserved I* sequences are shown in Supporting Information Figures S2, S3. Protein IDs: farinin (Bu-1); avenin (AAA32716); γ-gliadin (AAK84779); α-gliadin (ABQ52123); LMW-glutenin subunit (EU189095); purinin (ADA62372); farinin R2 (Bu-1).

Sequence Comparisons. It was noted by visual comparison that the sequences of the oat avenins were closer to γ-gliadins than to the farinins, and we sought to understand why NCBI BLAST, in particular, tends to give the highest similarity score to avenins when the query sequence is a farinin (Table 3)—this presumably being the basis for the name avenin-like having been assigned to the proteins. Possible problems may arise in BLAST and similar search programs, and in programs that delineate phylogenetic trees, when the proteins of interest have many insertions, deletions, and substitutions[35] as is the case for gluten proteins. Another possible complication arises from the predominance of glutamine and proline in gluten

proteins (personal communication, Nick Goldman, EMBL; personal communication, Peter Cooper, NIH) insofar as such skewed amino acid compositions may be in contradiction to certain assumptions on which the search and alignment algorithms are based. This might introduce a difficulty especially for the repeating sequence regions of the gluten proteins, where glutamine and proline residues strongly predominate. Repeating sequence domains of gluten proteins tend to have about 70 mol percent of glutamine plus proline. The nonrepeating sequence domains tend to have about 30 mol percent of glutamine plus proline, but even this lower proportion might, combined with the other difficulties mentioned above, lead to errors. These difficulties would not necessarily invalidate sequence alignments, for example, when the proteins being compared are limited to a single gluten subfamily, which have only small sequence variations (the γ-gliadins, for example[15]). However, when various subfamilies with greater sequence differences are compared and phylogenetic trees created, branch nodes and time frames between nodes of such trees may not be valid. Analysis of gluten proteins with various phylogeny programs sometimes gave different results depending on which programs were used and on the parameters assigned (data not shown). Any tree involving gluten proteins should be approached with caution as to its validity.

To investigate the question of why the NCBI BLAST program scores the farinins closest to oat avenins when the query sequence is an "avenin-like" protein, BLAST comparisons were made, when the query sequence was the farinin (Bu-1 R1 domain) and when it was an avenin, for the following sequences: (1) sequences corresponding to the C-terminal domains of a γ-gliadin, an α-gliadin, a LMW-GS, and the equivalent seed storage protein (SSP) domains of an avenin, a purinin, and the R1 or R2 domains of the farinin Bu-1; and (2) a short sequence tract (designated *Conserved I*) determined by visual inspection to correspond to the most highly conserved part of the sequences from 1). From the BLAST output, three numbers relating to sequence similarity are presented in Table 3: the BLAST score (bits), the number of sequence identities, and the number of positives (identities plus differences that result from single base changes in a codon). Repeating sequence domains, found in most gluten proteins (but not in the farinins, avenins, and purinins), were not included in our sequences because these repeating sequences are quite variable among subfamilies, probably recently evolved, and are not highly suitable for similarity comparisons.[33]

The C-terminal domains of seed storage proteins may be an ancestral type for sulfur-rich prolamins.[33,36,37] *Conserved I* (Con I) consists of an approximately 68-residue sequence that corresponded approximately to a combination of the A domain, the I_2 domain, and the B domain of γ-gliadins described by Shewry et al.[25] *Conserved I* included the first six cysteines of the usual eight found in SSP domains (for example, in γ-gliadins and avenins). However, because α-gliadins have only six cysteine residues, the equivalent tract includes only four cysteines, whereas in LMW-GS, which are missing the sixth cysteine typical of the γ-gliadins and avenins, the equivalent tract includes only 5 cysteines. An advantage of analyzing the *Conserved I* tract is that deletions are very much minimized; the comparisons and scoring of deletions are a troublesome part of sequence comparisons. The R1 and R2 domains of the Bu-1 farinin were analyzed separately. Single representatives (accessions) of each molecular subfamily were used because preliminary comparisons (data not shown) indicated that score variations among members of a subfamily were relatively small and not important to the points being made. The ω-gliadins were not included because they do not have unique sequence domains, being made up mainly of repeating sequences, which usually have no cysteine residues. The *Conserved I* tracts for γ-gliadin, farinin R1, and avenin are compared schematically in the context of the complete molecular structures in Supporting Information Figure S2 and the sequences of the *Conserved I* tracts are compared in Supporting Information Figure S3.

When farinin R1 was the search query sequence in a multiple alignment analysis of SSP domains, BLAST assigned the highest score to avenin, although γ-gliadins, and other gluten proteins gave only slightly lower scores (Table 3). Note that while the BLAST score was highest for avenin (93.6 vs 85.5 for γ-gliadin), identities were equal for γ-gliadins and avenin (37%), and positives were highest for purinin (53% vs 50% for avenin). For a corresponding comparison in which *Conserved I* sequences were compared with the farinin sequence again as query, avenin and LMW-GS gave high scores that were approximately the same (66.2 and 66.6, respectively). Identities were highest for avenin (49%) and LMW-GS (48%), whereas Positives were highest for LMW-GS (68%).

When avenin was the search query sequence in the multiple analysis, BLAST assigned the highest score (157) to γ-gliadin, whereas the score for farinin R1 was only 93.6. Identities were also highest for γ-gliadin (56% vs 37% for farinin R1), and this was also the case for positives (66% vs 50% for farinin R1). In the corresponding comparison with the *Conserved 1* sequences, the highest score was for γ-gliadin (90.5 vs 66.2 for farinin R1), and this was also the case for identities (68% vs 49% for farinin R1).

These results and visual examinations indicate that the farinin R1, R2 domains have approximately the same degree of similarity to avenin, γ-gliadin, α-gliadin, LMW-GS, and purinins (about 50% positives), but that avenin itself has a greater similarity to the other wheat proteins in Table 3 (about 60% positives). Thus, farinins are less avenin-like than the gliadins, and substitution of the name farinins for avenin-like as suggested here would be appropriate on that basis. In the case considered, visual analysis and BLAST were not in conflict despite the potential for problems in comparing gluten proteins by mathematical algorithms as discussed above.

Examination of Supporting Information Figure S3 suggests that the apparent conflict arises in part from there being a considerable number of identities among the proteins being compared so that minor differences assume an unexpected importance. Gaps that are present for the larger sequences (complete sequences or SSP domains) may also contribute to the differences that occur in query-dependent searching.

General Evolutionary Relationships. The farinins and purinins are smaller than traditional gluten proteins because they lack the significant repeating-sequence domain characteristic of the traditional gluten protein subfamilies, although the avenin-like b farinins migrate only slightly faster in SDS-PAGE than α- and γ-gliadins because of the duplication that gave rise to the R1 and R2 domains, which compensates to a large extent for the missing repeating sequence domain. The farinins and purinins, both lacking repeating sequence domains, while having considerable homology with gliadins and avenins, might be relics of earlier storage protein forms that predate development of the repeating sequence domains that are an important feature of the traditional gluten proteins. It cannot be

ruled out, however, that they may have evolved through loss of repeating sequence domains.

The addition of large repeating sequence domains, made up predominantly of glutamine and proline residues, to an ancestral SSP domain in the traditional gluten proteins, combined with extensive gene duplication, may represent the culmination of an evolutionary process that maximizes seed nitrogen in the form of amino acids that can be readily used directly in, or be readily transformed into, the amino acids and proteins needed by the developing embryo. The ω-gliadins, which consist essentially only of repeats (glutamine + proline make up about 70% of the amino acids in the B-genome associated ω5 gliadins), might epitomize this process in that domains less efficient at storing nitrogen have been lost. These latter domains presumably corresponded to one or more of the A, B, and C domains that were the basis for construction of the prolamin superfamily,[25,36] remnants of which are present in all the gluten proteins except the ω-gliadins. The predominance of glutamine and proline in the gluten proteins may be a consequence of lower energy requirements for their transformation into other amino acids needed for protein synthesis in the developing embryo.[38]

We suggest that farinins and purinins, as wheat endosperm proteins with a high degree of sequence identity to γ-gliadins, α-gliadins, LMW-GS, and avenins and having no known function, might also be classed as storage proteins, although the tendency of farinins and purinins to appear in the salt soluble fraction perhaps would pose some difficulties in calling them gluten proteins.

Another significant group of endosperm proteins that is difficult to classify is that of the α-amylase inhibitors. Although the α-amylase inhibitors are not usually thought of as "gluten" proteins because they differ in solubility properties (soluble in salt solutions), do not have a major repeating sequence domain, and apparently have a protective role as a consequence of their ability to inhibit certain insect α-amylases, they are present in fairly large amounts (approximately 4% of total endosperm protein[13,39]) and are likely to serve as storage proteins during the germination process.

A significant sequence similarity between the wheat α-globulin and the N-terminal domain of y-type HMW-GS has been reported by Gu et al.[40] This sequence similarity, combined with the wide distribution of the α-globulin gene across many different grass genomes, led Xu and Messing[33] to suggest that duplication of the α-globulin gene, followed by mutation of the duplicated gene, gave rise to the HMW-GS. Although the α-globulins may have some as yet unknown function, they might also serve a storage function in the endosperm. Xu and Messing[33] placed the α-globulins and HMW-GS in the same group (Category II) in their phylogenetic classification of grain proteins.

The similarity of the HMW-GS to the α-amylase inhibitors is of borderline significance in BLAST, but Cazalis et al.[41] using a fold-recognition approach, reported a significant similarity between the Dy10 N-terminal domain (Dy10NT) and α-amylase inhibitor 0.19 (for which a complete X-ray diffraction structure is available). They used the structural similarity to predict the folding of Dy10NT and its disulfide bond arrangement. The more likely of their two possibilities for arrangements of four of the five cysteines in the Dy10 N-terminal domain was in agreement with directly determined linkages (D. D. Kasarda et al., unpublished results). The predicted disulfide arrangement for Dy10NT differs from the arrangement predicted for γ-gliadins.[18,19] The relation of the α-amylase inhibitors to HMW-GS is also supported by a comparison of the sequences involving the CXC motif (CysXxxCys) characteristic of the wheat α-globulins and α-amylase inhibitors and the mutated equivalent in HMW-GS (Figure 4) with other gluten proteins. Accordingly, we suggest

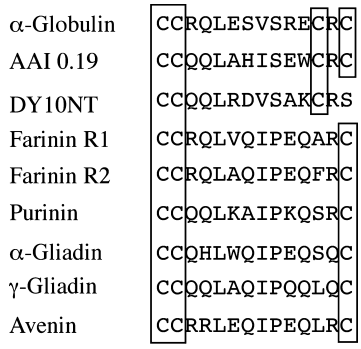


Figure 4. Comparison of sequences including the -CC- motif and the -CXC- motif (or its remnant). Protein IDs for α-globulin, Dy10, and AAI 0.19 were ABG68039, CAA31396, and P01085, respectively; other protein IDs as in Table 3.

that the α-amylase inhibitors should be included with α-globulins and HMW-GSs in Group III of the classification of Xu and Messing.[33] We suggest that the storage proteins of wheat have evolved from two different precursor pathways: one giving rise to the α-globulins, the HMW-GS, and the α-amylase inhibitors, the other giving rise to the LMW-GS, γ-gliadins, α-gliadins, ω-gliadins, farinins, and purinins. Divergence from an ancestral gene for the two lines may have occurred in different species with ultimate recombination during the evolution of the genus *Triticum*. This hypothesis would, however, be difficult to prove because there is no detailed understanding of the evolutionary pathways leading to the complex mixture of storage proteins in current day wheat species; ancestral species may have become extinct, and the fossil record for plants is minimal.

Wheat Quality Relationships. Glutenin polymers are generally considered the major contributor to variations in wheat quality and the HMW-GS show the strongest correlations in this regard. Although the molecular basis for these contributions is not well understood, it appears likely that it derives from the unusually large repeating sequence domains of HMW-GS combined with the ability of these subunits to form chain-extending and chain-branching intermolecular disulfide bonds with one another. The farinins make up about 1% of the total endosperm proteins.[13] It has been suggested that because some of the farinins have odd numbers of cysteine residues, they may be incorporated into the glutenin polymeric fraction and thereby contribute importantly to wheat quality variations.[7] Mamone et al.[42] have presented evidence that the avenin-like b proteins can be found in glutenin, and Chen et al.[43] have presented evidence for incorporation of avenin-like proteins into doughs during mixing, but at present it has not been clearly established that these low-abundance proteins are important contributors to wheat quality and its variation.

Despite the potential of avenin-like b proteins for forming intermolecular bonds, their low abundance combined with the absence of a repeating sequence domain is likely to minimize their contributions to dough viscoelasticity. Nevertheless,

further studies would be needed to evaluate the possibility that farinins have a significant effect on the mixing and/or baking quality of wheat.

Relation to Celiac Disease and Wheat Allergy. Celiac disease is an autoimmune-like condition triggered in susceptible individuals by the ingestion of gluten proteins. Peptides derived from gluten induce damage to the small intestinal mucosa, destroying villous structure and producing hypertrophy of crypt cells. Celiac disease is triggered in susceptible individuals by the ingestion of gluten proteins and almost all of the proteins in all the traditional gluten subfamilies carry at least one epitope (often several epitopes) active in celiac disease.[44] The farinins, because of their close sequence relationship to gliadins, might carry active epitopes. A screen of the farinin Bu-1 protein sequence did not, however, turn up any exact matches for epitopes known to stimulate T cells from celiac patients.[45] The absence of these epitopes results in part from the absence of a repeating sequence domain wherein most of the harmful epitopes reside.[46] All of the active epitopes in gluten proteins have not yet been characterized and a certain prediction for the absence of activity in celiac disease for farinin cannot be made. Perhaps farinins should be tested with patients, but at present it seems likely that these proteins are not active in celiac disease—or at most, minimally active.

It does seem likely, however, that farinins are allergenic. De Gregorio et al.[47] described a 19.4 kDa protein derived from an extract of whole meal wheat bread that reacted strongly with IgE of a serum pool from wheat food allergic patients. Based on short amino acid sequences, the peptide was identified as being derived from an avenin-like protein. The N-terminal sequence they obtained, EPQQEA, and its size correspond to the B peptide of Peak 4, described here, that results from intrinsic endopeptidase cleavage of the Bu-1 protein (Figure 3) and distinguishes the peptide from other avenin-like proteins. It is at least possible that only the B peptide of the Bu-1 protein is highly allergenic since the other avenin-like b proteins in wheat do not yield a peptide of corresponding size when reduced. Bu-1/Peak 4 differs significantly from Bu-2 and Bu-3 having 71% identities in a BLAST comparison with the latter two types, whereas Bu-2 and Bu-3 have 93% identities with one another. Bu-2 and Bu-3 are characteristic of previously characterized avenin-like proteins.[7,8,24]

Modeling of the Molecular Structure of the Bu-1/Peak 4 Protein. The Bu-1 protein corresponds to an avenin-like b protein and has 19 cysteine residues, but the corresponding protein extracted from starch behaved as a monomer upon SDS-PAGE carried out in a nonreducing buffer (Figure 2). This is a somewhat puzzling result since it is generally thought that an odd number of cysteines would result in incorporation of the molecule into the glutenin fraction.[7]

The Bu-1 protein has two complete storage protein domains[7] (R1, R2) each with eight cysteine residues apparently corresponding in homology or position (or both) to the cysteines of γ-gliadin or avenin. There are two additional cysteines, residues 5 and 12, near the N-terminus of the polypeptide chain that are outside the first domain, R1, of the mature protein; it also has a cysteine as the C-terminal amino acid of the protein shortly after the second domain (R2). Because of the close situation of cysteines 5 and 12, it seems reasonable to suggest that these might form an intramolecular disulfide bond and that the cysteines of the R1 and R2 domains might form sets of intramolecular disulfide bonds, similar to those found in γ-gliadin, for each of these two domains.[18] This leaves one cysteine (residue 260) free to react with other cysteines. This free cysteine would be expected to favor incorporation of the protein into the glutenin polymer. We cannot explain why the Bu-1 protein as extracted from the starch granule surface is a monomer. Perhaps there is post-translational processing of the protein at the C-terminal end that removes cysteine 260, which we cannot say for certain is present in our protein because the corresponding C-terminal peptide (Figure 3) was not obtained and sequenced directly. The proposed arrangement of disulfide linkages is shown in Figure 5. Chen et al.[43] proposed disulfide arrangements for an

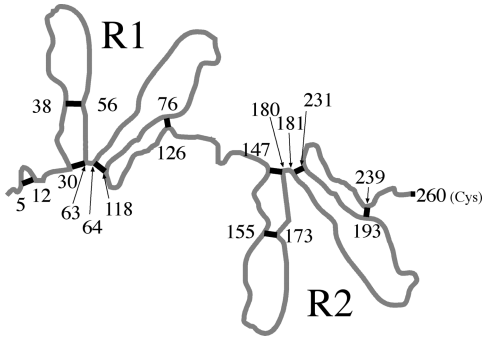


Figure 5. Proposed (hypothetical) disulfide arrangements for model of the farinin/Peak 4 Bu-1 protein. Numbering corresponds to the mature sequence beginning LETIC-. The final cysteine (residue 260) is not involved in disulfide linkage according to our model.

avenin-like b protein with 19 cysteine residues based on a predictive program. They obtained results considerably different from those put forward in our Figure 5. It is important to note, however, that only the Bu-1/Peak 4 protein has a complete complement of 8 cysteines in R1—other avenin-like b proteins seem to have only seven cysteines in R1; this would be expected to have an important effect on folding and the arrangement of disulfide linkages and our model applies only to proteins of the Bu-1/Peak 4 type.

A simple, computer-based molecular model of the parent Bu-1 protein, prior to cleavage at asparagine 96 is shown in Figure 6, which is based on the amino acid sequence and the proposed disulfide bond arrangements. The model shown in Figure 6 is based on the I-TASSER modeling system. It is notable for four α-helices and a number of turns. The threading component of I-TASSER found significant similarities to the structure of various α-amylase inhibitors. The R1 and R2 domains were distinct from one another with no main chain hydrogen bonding connecting them. Asparagine 96, the site of cleavage by the asparagine endoproteinase that gives rise to the disulfide linked, two peptide protein, is indicated in the model, where it appears as part of a turn structure that appears likely to be accessible to the endoprotease.

Although the model of Figure 6 is of the intact protein, it appears likely that the cleaved protein would have a similar structure. We speculate that the cleavage would occur after formation of all intramolecular disulfide bonds and these disulfide bonds would likely stabilize the structure.

No gluten protein has ever been crystallized and there are no three-dimensional structures available for any gluten protein that might be compared with our model; hence, modeling is the only window into possible structures available. Keck et al.[48] have suggested that the intramolecular disulfide bonds of gluten proteins are specific and this seems likely. Nevertheless, further

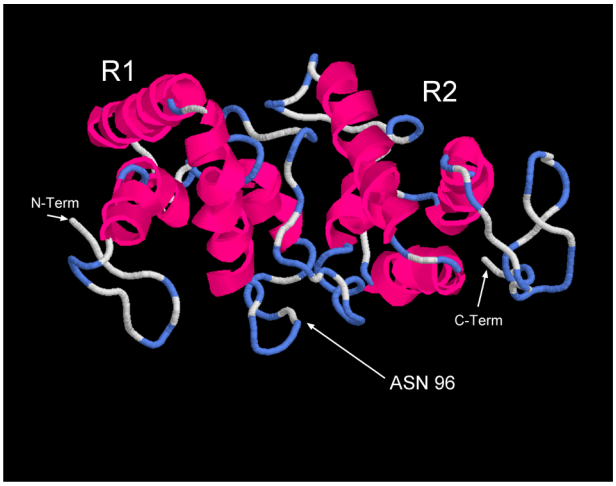


Figure 6. Computer molecular model of the farinin/Peak 4 Bu-1 protein based on the amino acid sequence of contig Bu-1 and the hypothetical disulfide arrangements illustrated in Figure 5. The R1 and R2 domains are indicated, along with the N-terminus of the protein chain (N-Term) and the C-terminus of the chain (C-Term). The position of the cleavage point at asparagine 96 is indicated. Alpha-helices are in red, turns in blue, and unassigned structure in white.

evidence in support of this contention would be desirable, as would attempts to determine the three-dimensional structures of gluten proteins having only intramolecular disulfide bonds. Farinin and purinin might be good candidates for crystallization insofar as they lack the repeating sequence domain that is likely to be flexible in structure and thereby interfere with crystal formation.

■ ASSOCIATED CONTENT

Supporting Information

Supplementary Table 1, identification of Butte 86 ESTs similar to AF470351 and GU211171 and assignment to contigs; Supplementary Figure S1, consensus sequences of Butte 86 contigs; Supplementary Figure S2, comparison of sequence domains for γ-gliadin, farinin R1, and avenin; Supplementary Figure S3, comparison of *Conserved 1* domain sequences for γ-gliadin, farinin R1, and avenin. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: donald.kasarda@ars.usda.gov. Tel: 510-525-9344. Fax: 510-559-5818.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Peter R. Shewry and Stefania Masci for reading the manuscript and for helpful suggestions.

■ REFERENCES

(1) D'Ovidio, R.; Masci, S. The low-molecular-weight glutenin subunits of wheat gluten. *J. Cereal Sci.* **2004**, *39*, 321−339.

(2) Salcedo, G.; Prada, J.; Aragoncillo, C. LMW gliadin-like proteins from wheat endosperm. *Phytochemistry* **1979**, *18*, 725−727.

(3) Anderson, O. D.; Hsia, C. C.; Adalsteins, A. E.; Lew, E. J.-L.; Kasarda, D. D. Identification of several new classes of low-molecular-weight wheat gliadin-related proteins and genes. *Theor. Appl. Genet.* **2001**, *103*, 307−315.

(4) Clarke, B. C.; Phongkham, T.; Gianibelli, M. C; Beasley, H.; Bekes, F. The characterization and mapping of a family of LMW-gliadin genes: effects on dough properties and bread volume. *Theor. Appl. Genet.* **2003**, *106*, 629−635.

(5) Gomez, L.; Sanchez-Monge, R.; Salcedo, G. A family of endosperm globulins encoded by genes located in group 1 chromosomes of wheat and related species. *Mol. Gen. Genet.* **1988**, *214*, 541−546.

(6) Gomez, L.; Sanchez-Monge, R.; Lopez-Otin, C.; Salcedo, G. Amino acid compositions and sequence analysis of the major low Mr globulins from *Triticum monococcum L.* endosperm. *J. Cereal Sci.* **1991**, *14*, 117−123.

(7) Kan, Y. C.; Wan, Y. F.; Beaudoin, F.; Leader, D. J.; Edwards, K.; Poole, R.; Wang, D. W.; Mitchell, R. A. C.; Shewry, P. R. Transcriptome analysis reveals differentially expressed storage protein transcripts in seeds of *Aegilops* and wheat. *J. Cereal Sci.* **2006**, *44*, 75−85.

(8) Chen, P.; Wang, C.; Li, K.; Chang, J.; Wang, Y.; Yang, G.; Shewry, P. R.; He, G. Cloning, expression and characterization of novel avenin-like genes in wheat and related species. *J. Cereal Sci.* **2008**, *48*, 734−740.

(9) Skylas, D. J.; Mackintosh, J. A.; Cordwell, S. J.; Balleal, D. J.; Walsh, B. J.; Harry, J.; Blumenthal, C.; Copeland, L.; Wrigley, C. W.; Rathmell, W. Proteome approach to the characterization of protein composition in the developing and mature wheat-grain endosperm. *J. Cereal Sci.* **2000**, *32*, 169−188.

(10) Kasarda, D. D.; DuPont, F. M.; Vensel, W. H.; Altenbach, S. B.; Lopez, R.; Tanaka, C. K.; Hurkman, W. J. Surface-associated proteins of wheat starch granules: suitability of wheat starch for celiac patients. *J. Agric. Food Chem.* **2008**, *56*, 10292−10302.

(11) Wolf, M. J. Wheat starch. In *Methods in Carbohydrate Chemistry*; Whistler, R. L., Ed.; Academic Press: New York, NY, 1964; Vol. *4*, pp 6−9.

(12) Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **1970**, *227*, 680−685.

(13) DuPont, F. M.; Vensel, W. H.; Tanaka, C. K.; Hurkman, W. J.; Altenbach, S. B. Deciphering the complexities of the wheat flour proteome using quantitative two-dimensional electrophoresis, three proteases and tandem mass spectrometry. *Proteome Sci.* **2011**, *9*, 10.

(14) Altenbach, S. B.; Tanaka, C. K.; Hurkman, W. J.; Whitehand, L. C.; Vensel, W. H.; DuPont, F. M. Differential effects of a post-anthesis fertilizer regimen on the wheat flour proteome determined by quantitative 2-DE. *Proteome Sci.* **2011b**, *9*, 46.

(15) Altenbach, S. B.; Vensel, W. H.; DuPont, F. M. Analysis of expressed sequence tags from a single wheat cultivar facilitates interpretation of tandem mass spectrometry data and discrimination of gamma gliadin proteins that may play different functional roles in flour. *BMC Plant Biol.* **2010**, *10*, 7.

(16) Roy, A.; Kucukural, A.; Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **2010**, *5*, 725−738.

(17) Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinf.* **2008**, *9*, 40.

(18) Köhler, P.; Belitz, H.-D.; Wieser, H. Disulphide bonds in wheat gluten: further cystine peptides from high-molecular-weight (HMW) and low-molecular-weight (LMW) subunits of glutenin and from γ-gliadins. *Z. Lebensm.−Unters. Forsch.* **1993**, *196*, 239−247.

(19) Müller, S.; Vensel, W. H.; Kasarda, D. D.; Köhler, P.; Wieser, H. Disulphide bonds of adjacent cysteine residues in low-molecular-weight subunits of wheat glutenin. *J. Cereal Sci.* **1998**, *27*, 109−116.

(20) Dereeper, A.; Guignon, V.; Blanc, G.; Audic, S.; Buffet, S.; Chevenet, F.; Dufayard, J.-F.; Guindon, S.; Lefort, V.; Lescot, M.; Claverie, J.-M.; Gascuel, O. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* **2008**, *36* (Suppl. 2), W465−W4699.

(21) Habuka, N.; Kataoka, J.; Miyano, M.; Tsuge, H.; Ago, H.; Noma, M. Nucleotide sequence of a genomic gene encoding tritin, a ribosome-inactivating protein from *Triticum aestivum*. *Plant Mol. Biol.* **1993**, *22*, 171−176.
(22) Coleman, W. H.; Roberts, W. K. Factor requirements for the tritin inactivation of animal cell ribosomes. *Biochm. Biophys. Acta* **1981**, *654*, 57−66.
(23) Lowy, G. D. A.; Sargeant, J. G.; Schofield, J. D. Wheat starch granule protein: The isolation and characterization of a salt-extractable protein from wheat granules. *J. Sci. Food Agric.* **1981**, *32*, 371−377.
(24) De Caro, S.; Ferranti, P.; Addeo, F.; Mamone, G. Isolation and characterization of avenin-like protein type B from durum wheat. *J. Cereal Sci.* **2010**, *52*, 426−431.
(25) Shewry, P. R.; Napier, J. A.; Tatham, A. S. Seed Storage Proteins: structures and biosynthesis. *Plant Cell* **1995**, *7*, 945−956.
(26) Kasarda, D. D.; Autran, J.-C.; Lew, E. J.-L.; Nimmo, C. C.; Shewry, P. R. N-terminal amino acid sequences of ω-gliadins and ω-secalins: Implications for the evolution of prolamin genes. *Biochim. Biophys. Acta* **1983**, *747*, 138−150.
(27) DuPont, F. M.; Vensel, W.; Encarnacao, T.; Chan, R.; Kasarda, D. D. Similarities of omega gliadins from *Triticum urartu* to those encoded on chromosome 1A of hexaploid wheat and evidence for their post-translational processing. *Theor. Appl. Genet.* **2004**, *108*, 1299−1308.
(28) Masci, S.; D'Ovidio, R.; Lafiandra, D.; Kasarda, D. D. Characterization of a low-molecular-weight glutenin subunit gene from bread wheat and the corresponding protein that represents a major subunit of the glutenin polymer. *Plant Phys.* **1998**, *118*, 1147−1158.
(29) Singh, N. K.; Shepherd, K. W.; Langridge, P.; Gruen, L. C.; Skerritt, J. H. Identification of legumin-like proteins in wheat. *Plant Mol. Biol.* **1988**, *11*, 633−639.
(30) Singh, N. K.; Donovan, G. R.; Carpenter, H. C.; Skerritt, J. H.; Langridge, P. Isolation and characterization of wheat triticin cDNA revealing a unique lysine rich repetitive domain. *Plant Mol. Biol.* **1993**, *22*, 227−237.
(31) Müntz, K.; Blattner, F. R.; Shutov, A. D. Legumains—a family of asparagine-specific cysteine endopeptidases involved in propolypeptide processing and protein breakdown in plants. *J. Plant Physiol.* **2002**, *259*, 1281−1293.
(32) Woodhouse, S. C. *English-Greek Dictionary: A Vocabulary of the Attic Language*; George Routledge & Sons, Ltd.: London, 1910; pp 1048; http://www.lib.uchicago.edu/efts/Woodhouse/.
(33) Xu, J. H.; Messing, J. Amplification of prolamin storage protein genes in different subfamilies of the Poaceae. *Theor. Appl. Genet.* **2009**, *119*, 1397−1412.
(34) Salcedo, G.; Prada, R.; Sanchez-Monge, R.; Aragoncillo, C. Aneuploid analysis of low molecular weight gliadins from wheat. *Theor. Appl. Genet.* **1980**, *56*, 65−69.
(35) Löytynoja, A.; Goldman, N. Uniting alignments and trees. *Science* **2009**, *324*, 1528−1529.
(36) Kreis, M.; Shewry, P. R.; Forde, B. G.; Forde, J.; Miflin, B. J. Structure and evolution of seed storage proteins and their genes with particular reference to those of wheat, barley and rye. In *Oxford Surveys of Plant Molecular and Cell Biology* ; Miflin, B. J., Ed.; Oxford University Press: Oxford, U.K., 1985; pp 253−317.
(37) Shewry, P. R.; Tatham, A. S. The prolamin storage proteins of seeds: structure and evolution. *Biochem. J.* **1990**, *267*, 1−12.
(38) Bhatia, C. R.; Rabson, R. Bioenergetic considerations in cereal breeding for protein improvemenrt. *Science* **1976**, *194*, 1418−1421.
(39) Altenbach, S. B.; Vensel, W. H.; DuPont, F. M.. The spectrum of low molecular weight alpha-amylase inhibitor genes expressed in the US bread wheat cultivar Butte 86. *BMC Res. Notes* **2011a**, *4*, 242.
(40) Gu, Y.; Wanjugi, H.; Coleman-Derr, D.; Kong, X.; Anderson, O. D. Conserved globulin gene across eight grass genomes identify fundamental units of the loci encoding seed storage proteins. *Funct. Integr. Genomics* **2010**, *10*, 111−122.
(41) Cazalis, R.; Aussenac, T.; Rhazi, L.; Marin, A.; Gibrat, J.-F. Homology modeling and molecular dynamics simulation of the N-terminal domain of wheat high molecular weight glutenin subunit 10. *Protein Sci.* **2003**, *12*, 34−43.
(42) Mamone, G.; De Caro, S.; Di Luccia, A.; Addeo, F.; Ferranti, C. Proteomic-based analytical approach for the characterization of glutenin subunits in durum wheat. *J. Mass Spectrom.* **2009**, *44*, 1709−1723.
(43) Chen, P.; Li, R.; Zhou, R.; He, G.; Shewry, P. R. Heterologous expression and dough mixing studies of a novel cysteine-rich avenin-like protein. *Cereal Res. Commun.* **2010**, *38*, 406−418.
(44) Jabri, B.; Kasarda, D. D.; Green, P. H. R. Innate and Adaptive Immunity: the Yin and Yang of celiac disease. *Immunol. Rev.* **2005**, *206*, 219−231.
(45) Sollid, L. M.; Qiao, S.-W.; Anderson, R. P.; Gianfrani, C.; Koning, F. Nomenclature and listing of celiac disease relevant gluten T-cell epitopes restricted by HLA-DQ molecules. *Immunogenetics* **2012**, *64* (6), 455−460.
(46) Arentz-Hansen, H.; McAdam, S. N.; Molberg, Ø.; Fleckenstein, B.; Lundin, K. E.; Jørgensen, T. J.; Jung, G.; Roepstorff, P.; Sollid, L. M. Celiac lesion T cells recognize epitopes that cluster in regions of gliadins rich in proline residues. *Gastroenterology* **2002**, *123*, 803−809.
(47) De Gregorio, M.; Armentia, A.; Diaz-Perale, A.; Palacin, A.; Duenas-Laita, A.; Martin, B.; Salcedo, G.; Sanchez-Monge, R. Salt-soluble proteins from wheat-derived foodstuffs show lower allergenic potency than those from raw flour. *J. Sci. Food Agric.* **2009**, *57*, 3325−3330.
(48) Keck, B.; Köhler, P.; Wieser, H. Disulphide bonds in wheat gluten: cystine peptides derived from gluten proteins following peptic and thermolytic digestion. *Z. Lebensm.−Unter. Forsch.* **1995**, *200*, 432−439.

Journal of Experimental Botany, Characterization of a novel wheat endosperm protein belonging to the prolamin superfamily, Donald D. Kasarda, Elva Adalsteins, Ellen J.-L. Lew, Gerard R. Lazo, and Susan B. Altenbach.

Supplementary Data

Supplementary Table 1. Identification of Butte 86 ESTs similar to AF470351 and GU211171 and assignment to contigs.

Accession Number[1]	Assignment
BQ804260	farinin Bu-3
BQ804477	farinin Bu-1
BQ804624	farinin Bu-2
BQ804715	farinin Bu-2
BQ804865	farinin Bu-3
BQ804917	farinin Bu-3
BQ805144	farinin Bu-1
BQ805627[2]	farinin Bu-3
BQ805826	farinin Bu-2
BQ806373	farinin Bu-3
BQ806517	farinin Bu-1
BQ838688	farinin Bu-1
BQ838933[3]	farinin Bu-1

[1] All ESTs except BQ805144 were identified by blastn searching of NCBI EST database with AF470351. BQ805144 was identified by blastn searching with GU211171.

[2] EST is similar to farinin Bu-3, but has a 12 bp deletion in the middle of the coding region.

[3] EST contains a premature stop codon.

Supplementary Figure S1. Consensus sequences of Butte 86 contigs.

Farinin Bu-1

GAGGTTGAGGCTAGCCACCATGAAGACAATGTTCATCCTCGCTCTCCTCGCCCTTGCAGCATCCACCGCCA
TTGCCCAATTGGAAACCATATGTAGCCAGGGCTTCGGACAATGCCAACACCACCAACAACTAGGGCAACAA
CAGTTGCTGGATCAGATGAAGCCATGTGTGGCTTTCGTACAACATCAGTGTAGCCCAGTGAGAACACCATT
CCCCCAAACACGGGGAGAGCAGCATAGCAGTTGCCAAACCGTGCAACACCAATGCTGTCGGCAGCTAGTGC
AGATCCCAGAACAAGCCCGGTGCAAGGCCATACAGAGCGTGGAAGAGGCTATCATTCAACAACAGCCCCAA
CAACAATGGAATGAGCCCCAACAGGAAGCACACCTTAAGAGCATGAGGATGTCGCTTCAGACCCTGCCGTC
TATGTGCAACATCTACGTCCCGGTACAATGCCAGCAACAGCAACAACTGGGGCGACAACAACAACAACAGT
TGCAGGAGCAGTTAAAACCGTGTGCGACATTCCTACAACATCAATGTAGGCCAATGACAGTGCCATTCCCG
CATACACCAGTGCAGAAGCCCACCAGCTGCCAGAACGTGCAGTCCCAATGCTGCCGGCAGCTAGCACAGAT
CCCAGAGCAATTCCGCTGCCAAGCCATTCATAATGTGGTAGAGTCTATCAGGCAACAACAACATCACCAAC
CACAACAGGAAGTACAACTTGAGGGCCTGAGGATGTCACTTCACACCCTACCGTCGATGTGCAAAATCTAC
ATCCCTGTACAATGCCCAGCCACCACCACCACCCCCTATAGCATTACCATGACAGCTAGCTATACCGATGG
TACCTGCTAGTTGATCAATATAGCAGTGGCAAAATAAAATGTCATGCATCATCATGTGTGCCTAGTACTAA
TTGCAACATGGAAATAATGAACAGATGTTCCTGAATGAGGCACTTGTCTAAATAAAGCTACATGTTTGCAT
GATAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

Farinin Bu-2

TTGAGGCTAGCCACCATGAAGGTCTTCATCCTGGCTCTCCTTGCCCTCGCAGCAACCACCGCCATTGCCCA
GTTGGAAACCACATGTAGCCAGGGCTTTGGACAATCCCAACAACAACAACAACCTGGTCAACGACAGTTGC
TGGAGCAGATGAAGCCGTGTGTGGCATTCCTGCAACAAAAGTGTAGCCCACTGAGAATGCCATTCCTCCAG
ACACAAGTCGAGCAACTTAGCAGCTGCCAGATCGTGCAATACCAATGCTGCCAGCAGTTGGCGCAGATCCC

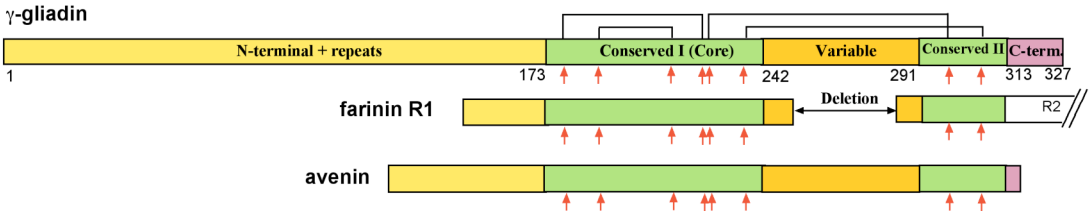
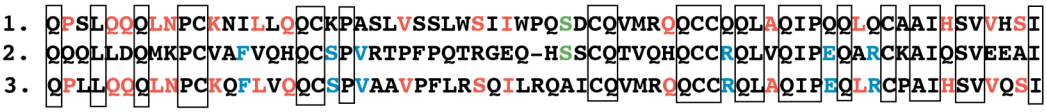
[illegible]

CTATCATTGATGTGC GGCAGGAGCCCCAACAGCAAGCACAACATAAGAGCATGAGGATGTTGCTTGAGAACC
ACAACAGTTGCAGGA AACATCTACGTCCCGGTACAATGCCAGCAGCAACAACAACTGGGGCAACAACAACAA
CATTCCCCCAGATAC GCAGTTGACACCGTGCACGACATTCCTGCAACAGCAGTGTAGCCCAGTGACAGTGCC
TCACAGATCCCAGAG CGGTGGATCAGCCTACCAGCTGCCAAAATGTGCAGCACCAATGTTGCCGACAGCTAT
CCAACAACAATGGCA CAATTCCGTTGCCAAGCCATCCACAACGTGGCGGAGGCTATCAGGCAACAACAACCC
AGGCCCTGCGGTCGA GGGTATGTACCAGCCCCAGCAGCCAGCACAACTTGAGAGCATTAGGATGTCGCTTCA
TTGGTGGCTACCTAC TGTGTAGCATCTACATCCCAGTACAATGCCCCGCCCCCACCACCTACAACATTCCAT
ATCATCATGTGTGGC ACTGGTGGTGCGTGCTAGATGATCTATGTAGCGGTGGCAAAATAAAAATGTCATGCA
CTGAATAAAAAAAAA AAGTACTAGTTGCAATATGGAAATGATGAATAAATGTTCCTGAATAAAGCATTTGTC
AAAAAAAAA

Farinin Bu-3

TCGCCCTCACAGCAA CAACAGCAAAGCAAACTTGAGGCTAGCCACCATGAAGGTCTTCATCCTGGCTCTCCT
CAACAGCAACAACCT CCACCGCCATTGCCCAGTTGGAAACCACATGTAGCCAGGGCTTCGGACAATACCAAC
GTGTAGGCCACTGAG GGGCAACGACAGTTGCTGGAGCAGATGAAGCCGTGTGTGGCATTCCTGCAACAACAC
ACCAATGCTGCCAGC AATGCCATTCCTCCAGACACAGGTGGAGCAGCTTAGCAGCTGCCAAATCGTGCAACA
GCTATCATGCAACAA AGCTAGCGCAGATCCCAGAGCGAATCCGGTGCCATGCCATCCACAGCGTGGTAGAGC
GATGTTGCTTGAGAA CAATCCCAACAACAATGGCAGGAGCGCCAACAGCAAGCACAACATAAGAGCATGAGC
TGGGGCAACAACAAC CCTATCATTGATGTGCAACATCTATGTCCCGGTACAATGCCAGCAGCAACAACAAAT
TGTAGCCCAGTGACA AACAACAACAGTTGCAGGAGCAGTTGACACCGTGTGCGACATTCCTCCAACATCAGT
CAATGTTGCCGACA GTGCCATTCCCCCAGATACCGGTGGATCAGCCTACCAGCTGCCAGAATGTGCAGCAC
TATCAGGCAACAAC GCTATCACAGATCCCGGAGCAATTTCGTTGCCAAGCCATCCATAACGTGGCTGAGGC
AGCATTCGGATGTCG AACCCCAACAACAATGGCAGGGTATGTACCAGCCGCAGCAGCCAGCACAACATGAGA
ACCGCCTACAACAT CTTCAGGCCCTGCGGTCGATGTGCAACATCTACATACCGGTACAATGCCCCGCCCCC
AAAATAAAAATGCC TCCCATGGTGGCTACCTGCACTAGTGGTGCATGCTAGATGATCAATGTAGCGGCCGA
CCTGAATAAAGCAT ATGCATCATCATGTGTGGCAAATACTAGTTGCAATATGGAAATGATGAATAAACGTT
TTGCCTGAATAAAAAAAAAAAAAAAA

Supplementary Figure S2. Comparison of sequence domains (mature proteins) for γ-**gliadin** (AAK84779), N-terminal+R1 of **farinin** (Bu-1), and **avenin** (AAA32716). The sequence domains are represented by boxes and the length of the combined domains in a protein is approximately scaled to the length of the protein polypeptide chain, although only the R1 domain is shown for the farinin protein. Positions of cysteine residues proposed to participate in disulfide linkages (Keck et al., 1995) are indicated by red arrows and the proposed disulfide linkages are indicated for the γ-gliadin. The total number of residues for the γ-gliadin is indicated along withnumbers designating residues at the beginning or end of domains. The **Conserved I** sequences are compared in Figure S3.



Supplementary Figure S3. Comparison of **Conserved I** sequences of: **1.** γ**-gliadin** (AAK84779), **2. farinin** (Bu-1), and **3. avenin** (AAA32716). Triple identities are boxed. Identities between farinin and avenin only are in blue. Identities between farinin and γ-gliadin only are in green. Identities between γ-gliadin and avenin only are in red.