MS/MS identification strategies for the arginine and lysine deficient proline-rich glutamine-rich wheat storage proteins

William Vensel, Frances DuPont, Stacia Sloane and Susan Altenbach, Western Regional Research Center Albany, CA

INTRODUCTION

1. Wheat gluten proteins represent ~80% of the protein in wheat flour and confer the unique viscoelastic properties that enable the production of bread, noodles and many other food products [1].
2. The gluten proteins are made up of a large number of proteins with very similar sequences that fall into five major groups, the α-, γ- and ω-gliadins and the high- and low-molecular weight glutenin subunits (HMW-GS and LMW-GS).
3. Protein within the α-, γ- and ω-gliadin groups are monomeric proteins that are associated with extensibility properties while proteins from HMW-GS and LMW-GS groups are linked by cysteine residues into large polymers that confer elasticity properties to wheat flour doughs.
4. All gluten proteins have high percentages of proline and glutamine and low percentages of arginine and lysine making it difficult to produce suitable and sufficient tryptic fragments for mass spectrometry (Table 1).
5. There is considerable heterogeneity of gluten protein sequences among different wheat cultivars.
6. Individual proteins within the major gluten protein groups respond differentially to elevated temperature and fertilizer during grain development, altering the protein composition and functionality of wheat flour [2,3]. To better understand how crop management and environment impact wheat flour quality, we developed MS\MS methods to improve identification of these highly similar proteins in the US wheat cultivar Butte 86.

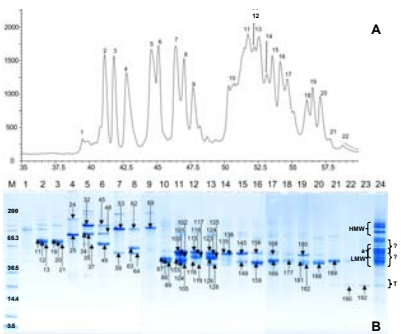
Table 1.

Protein Types	MW Range	Representative protein	Selected Amino Acids				
			%C	%P	%Q	%K	%R
α-gliadin	38000-42000	Butte #4 [5]	2.0	15.7	32.5	1.7	0.3
γ-gliadin	36000-44000	Butte #2 [5]	2.4	16.2	33.9	1.2	0.9
ω-gliadin	44000-78000	BAE2038 (omega-5)	0.0	18.9	50.8	0.9	0.9
		AAG1770	0.0	26.7	37.1	1.0	0.3
LMW-GS	29000-55000	ABC84363 (Butte #1)	2.6	12.7	31.6	0.6	2.6
HMW-GS	67000-88000	P10388 (Dx5)	0.6	13.1	34.9	0.9	1.0
		P10387 (Dy10)	1.0	10.6	31.7	1.2	2.1

METHODS

1. Total protein from each flour sample was extracted with SDS under reducing conditions, precipitated with acetone and analyzed by RP-HPLC (Fig 1 A) and then by 1DE (Fig 1B).
2. Gel bands were excised from multiple gels and separately digested with Thermolysin, Chymotrypsin and Trypsin.
3. Peptides were analyzed by MS/MS, using a QSTAR Pulsar *i* quadrupole time of flight mass spectrometer with a nano-electrospray source [4].

Figure 1.



DATA ANALYSIS

1. Initial searches of spectra were carried out against the "SuperWheat" (S_W) database using MASCOT and X!Tandem.
2. The "Superwheat" database contained 2,049,765 entries and included proteins encoded by all NCBI nonredundant green plant sequences, proteins encoded by contigs from three public EST assemblies, and proteins encoded by contigs assembled from ESTs from the wheat cultivar Butte 86 that was used for these experiments [5].
3. Search engine results from Mascot, X!Tandem or both Mascot and X!Tandem were analyzed separately for each enzyme or combination of enzymes using Scaffold [6] (Table 2 & Fig 2).

Table 2. Effect of search engine and enzyme on the total number of peptides detected for each gluten protein group.

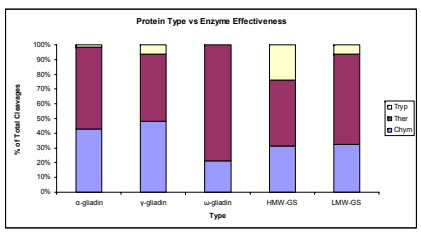
Enzyme	Chymotrypsin			Thermolysin			Trypsin			Chymo+ Thermo	All
Protein Type	Mascot	X!Tandem	Both	Mascot	X!Tandem	Both	Mascot	X!Tandem	Both	Both	Both
α-gliadin	53	141	191	24	216	249	3	2	7	440	447
γ-gliadin	32	91	126	18	123	119	18	29	17	245	262
HMW-GS	18	289	253	185	343	362	132	115	193	615	808
LMW-GS	40	105	141	64	207	264	29	5	28	405	433
ω-gliadin	2	38	19	8	65	71	0	0	0	90	90
Total	145	664	730	299	954	1065	182	151	245	1795	2040

Trypsin was the least effective enzyme.

Combining the results from all 3 enzymes and both search engines resulted in the most identifications.

X!Tandem identified more peptides than Mascot especially when digestions were performed with chymotrypsin and thermolysin.

Figure 2. % CLEAVAGE BY ENZYME AND PROTEIN TYPE



Enzyme effectiveness varied with the protein type. Trypsin was most effective with the HMW-GS and least effective with the gliadins and LMW-GS.

EFFECTS OF DATABASE SIZE AND COMPOSITION ON PROTEIN DISCOVERY

After analysis of S_W in Scaffold, a "subset" database was exported from Scaffold that included the sequences of all proteins that matched spectra in the dataset (Scaffold settings: 20% Protein Probability, 1 peptide and 0% peptide probability). The subset database contained 1,500 entries.

1. Eight additional databases were then constructed. Decoy sequences were Target sequences that were either reversed, randomized, or sequences from the unrelated organisms *Escherichia coli* or *Methanosarcina mazei*.
   a. Four databases contained ~30,000 sequences. One included ~15,000 *Triticum aestivum* sequences from NCBI nr plus an equal number of reversed sequences as a decoy and the Butte 86 sequences. The other three databases included the *Triticum aestivum* sequences from NCBI nr, the subset database sequences, any Butte 86 sequences that were not present in the subset database and an equal number of decoy sequences. Decoy databases are shown in Table 3.
   b. Four other databases contained ~3,200 sequences. These included the subset database, any Butte 86 sequences not present in the subset database and an equal number of decoy sequences. Decoy databases are as indicated in Table 3.
   c. The false discovery rate (FDR) was calculated as: FDR= (2 * (# decoy proteins) / (# of all proteins) [7].
3. Searches of spectra were carried out against the eight databases using MASCOT and X!Tandem.

Table 3. Number of proteins identified within each gluten protein group when searches of spectra were carried out against different databases.

	Sequence #	2,049,765	25,926	28,908	28,906	29,294	3,191	3,152	3,192	3,192
	Database	S_W	Ta_B86*	Ta_B86	Ta_B86	Ta_B86	Subset	Subset	Subset	Subset
	Decoy	none	Reverse	Reverse	Random	E. coli	M. mazei	E. coli	Random	Reverse
	α-gliadin	12	18	17	21	16	25	21	22	22
A	γ-gliadin	10	15	15	15	18	13	14	14	15
L	ω-gliadin	6	4	5	6	6	9	9	10	6
L	HMW-GS	7	6	10	8	7	7	7	8	7
	LMW-GS	13	16	15	17	18	16	14	14	14
	Total	48	59	62	67	65	70	65	68	64
	FDR	ND	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B	α-gliadin	2	7	3	3	3	8	7	6	6
8	γ-gliadin	3	5	4	5	5	7	6	5	7
6	ω-gliadin	2	2	1	1	1	4	3	3	2
	LMW-GS	5	2	2	2	2	3	3	4	2
	Total	12	16	10	11	11	22	19	18	17

Filter settings for Scaffold were: 95% Protein probability, 2 peptides and 90% Peptide probability.

RESULTS

1. The database size influenced the number of hits.
   48 gluten proteins passed the filter setting for the initial S_W database search. 12 of these were sequences from Butte 86.
2. A greater number of gluten proteins were identified when the database size was reduced to less than 30,000 sequences. 59 gluten proteins were identified using the Ta_B86* database (that did not contain the Subset database) and 16 of these were assigned to Butte 86 sequences.
3. Adding the "subset" database to the search (Ta_B86 Reverse Decoy) increased the total protein number passing the filter to 62, but decreased the number of Butte 86 proteins identified.
4. The number of proteins passing the filter ranged from 64 to 70 when subset database searches were performed and more of those proteins were assigned to sequences from Butte 86. The greatest number of Butte 86 sequences was obtained when *M. mazei* was used as the decoy.

CONCLUSIONS

To maximize the number of cultivar-specific wheat gluten proteins identified by MS/MS:

Digest proteins separately with chymotrypsin, thermolysin and trypsin.

Use both MASCOT and X!Tandem search engines for analysis of spectra.

Perform first pass searches on a large SuperWheat database followed by second pass searches on a subset database that contains *M. mazei* or random decoy sequences.

REFERENCES

[1] Shewry, P. R., Underwood, C., Wan, Y., Lovegrove, A., et al., Storage product synthesis and accumulation in developing grains of wheat. J. of Cereal Science 2009, 50, 106-112.
[2] DuPont, F. M., Hurkman, W. J., Vensel, W. H., Chan, R., et al., Differential accumulation of sulfur-rich and sulfur-poor wheat flour proteins is affected by temperature and mineral nutrition during grain development. Journal of Cereal Science 2006, 44, 101-112.
[3] Altenbach, S. B., DuPont, F. M., Kothari, K. M., Chan, R., et al., Temperature, water and fertilizer influence the timing of key events during grain development in a US spring wheat. Journal of Cereal Science 2003, 37, 9-20.
[4] Vensel, W. H., Tanaka, C. K., Cai, N., Wong, J. H., et al., Developmental changes in the metabolic protein profiles of wheat endosperm. Proteomics 2005, 5, 1594-1611.
[5] Altenbach, S. B., Vensel, Wiliam H. and DuPont, Frances M. . (2009) Analysis of expressed sequence tags from a single wheat cultivar facilitates interpretation of tandem mass spectrometry data and discrimination of gamma gliadin proteins that play different functional roles in flour, BMC Plant Biology.
[6] Searle, B. C., Turner, M., Nesvizhskii, A. I., Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *Journal of Proteome Research* 2008, *7*, 245-253.
[7] Kall, L., Storey, J. D., MacCoss, M. J., Noble, W. S., Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of Proteome Research* 2008, *7*, 29-34.