BIOCHEMICAL STUDIES OF CEREAL PROLAMINS FROM SORGHUM AND WHEAT

by

CHRISTOPHER L. MILLER

B.S., Kansas State University, 2003
M.S., Kansas State University, 2008

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Biochemistry and Molecular Biophysics Graduate Group
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2013

Abstract

Prolamins are the alcohol soluble storage proteins found in the endosperm of seeds from cereals and related grasses. The physical and biochemical properties of prolamins vary between species; and due to their relative abundance can greatly affect the properties and healthfulness of foods from those sources. In this work I investigate peptides from the high molecular weight glutenin of wheat, which is linked to dough elasticity and finished product quality.

Using 2D NMR I determined the three-dimensional structure for the repeat peptide Ac-GQQPGQG-Am, which makes up ~50% of the 700 residue central domain. The structure was found to be a flexible β-hairpin with a type II β-turn across residues QPGQ. The NMR structure was later compared to 33 proteins with known three-dimensional structure carrying the exact sequence (backbone RMSD=0.802Å). This finding provides useful insight into the structure of high molecular weight glutenin and the molecular nature gluten elasticity.

Alternatively, I studied the kafirin storage prolamins from sorghum, which do not have important physical properties, but are poorly digestible by humans and livestock. Improving digestibility of sorghum could significantly impact human health and nutrition in countries where sorghum is a dietary staple. In this work I devised a unique protocol to isolate kafirins under both non-reducing and reducing conditions.

I studied kafirin extracts using SDS-PAGE, HPLC and MALDI-TOF MS, then purified β-kafirin, for the first ever characterization of this single protein. Past studies implicate β-kafirin as a source of poor digestibility due to extensive intermolecular disulfide cross-linking. Contrary to this claim I found more than 50% of β-kafirin was extractable without reducing agents. I used chymotrypsin to digest pure β-kafirin and map 10 cysteine residues to 5 intra-molecular disulfide bonds. Precise pairings have yet to be determined although the protein is largely intact after 12 hours of digestion. This work challenges us to think about sorghum protein body formation and the mechanism that leads to disulfide cross-linking during seed desiccation at maturity.

BIOCHEMICAL STUDIES OF PROLAMINS FROM SORGHUM AND WHEAT

by

CHRISTOPHER L. MILLER

B.S., Kansas State University, 2003
M.S., Kansas State University, 2008

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Biochemistry and Molecular Biophysics Graduate Group
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2013

Approved by:

Major Professor
Dr. Gerald R. Reeck

Abstract

Prolamins are the alcohol soluble storage proteins found in the endosperm of seeds from cereals and related grasses. The physical and biochemical properties of prolamins vary between species; and due to their relative abundance can greatly affect the properties and healthfulness of foods from those sources. In this work I investigate peptides from the high molecular weight glutenin of wheat, which is linked to dough elasticity and finished product quality.

Using 2D NMR I determined the three-dimensional structure for the repeat peptide Ac-GQQPGQG-Am, which makes up ~50% of the 700 residue central domain. The structure was found to be a flexible β-hairpin with a type II β-turn across residues QPGQ. The NMR structure was later compared to 33 proteins with known three-dimensional structure carrying the exact sequence (backbone RMSD=0.802Å). This finding provides useful insight into the structure of high molecular weight glutenin and the molecular nature gluten elasticity.

Alternatively, I studied the kafirin storage prolamins from sorghum, which do not have important physical properties, but are poorly digestible by humans and livestock. Improving digestibility of sorghum could significantly impact human health and nutrition in countries where sorghum is a dietary staple. In this work I devised a unique protocol to isolate kafirins under both non-reducing and reducing conditions.

I studied kafirin extracts using SDS-PAGE, HPLC and MALDI-TOF MS, then purified β-kafirin, for the first ever characterization of this single protein. Past studies implicate β-kafirin as a source of poor digestibility due to extensive intermolecular disulfide cross-linking. Contrary to this claim I found more than 50% of β-kafirin was extractable without reducing agents. I used chymotrypsin to digest pure β-kafirin and map 10 cysteine residues to 5 intra-molecular disulfide bonds. Precise pairings have yet to be determined although the protein is largely intact after 12 hours of digestion. This work challenges us to think about sorghum protein body formation and the mechanism that leads to disulfide cross-linking during seed desiccation at maturity.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Thank you to my wife Brandi, for your patience and support throughout this process. Without you none of this would be possible. To Jerry, thank you for taking a risk on something different and setting out on this journey together. To Dirk, thank you for valuing my effort to pursue a Ph.D. while a faculty member and for you generous financial contributions to my program. To my many friends and family who have supported me through this process I am eternally grateful.

Dedication

To Braden and Sophia. May the curiosity for science always fill your minds.

Prelude

Cereals and their storage proteins are among the most important to all humankind not only for the energy they provide but also for their useful mechanical properties in countless industrial and renewable materials. Cereals are diverse and abundant and in the case of wheat and maize have become ubiquitous, finding their way into almost every facet of human existence from fuels to plastics and beyond.

Despite the benefits derived from cereals they are not perfect and several negative outcomes have been recognized. For wheat and sorghum this relates to digestibility of their storage proteins. Sorghum is inherently difficult for humans and animals to digest, which can lead to undernourishment in cultures where sorghum is a dietary staple. While wheat gluten is a known allergen that must be strictly avoided by those with celiac disease. These are complex problems that will only be solved after we understand the storage proteins at a fundamental level.

The very nature of cereal storage proteins make them difficult to study. They are found within a complex mixture of proteins and biological materials often forming disulfide-linked networks in the seed endosperm. This complexity has limited our ability to study intact pure proteins and therefore the biological role of many storage proteins is unknown.

To study cereals and their proteins scientists take different approaches. One approach is to study components within a flour or dough system to understand how each affects the properties of the entire system. This is common in the science of dough rheology where each component affects the bulk material properties of the system. A second approach, and my own, is a reductionist approach, which attempts to isolate individual components from the complex system to study their properties in pure form. Both approaches are needed to understand biochemical origin of the most important cereal properties such as gluten elasticity and protein digestibility.

Chapter 1 - Isolation and Characterization of β-Kafirin from *Sorghum bicolor* cv. BTx623

Introduction

Sorghum Kafirins

The alcohol-soluble proteins (prolamins) from sorghum were first isolated by Johns & Brewster (1916) from Kafir sorghum. In this early work Johns and Brewster reported the compositional similarities between kafirin and maize zein proteins, a comparison that has been made repeatedly ever since.

Protein Body Formation

Sorghum kafirins and maize zeins are similarly synthesized in the endoplasmic reticulum from which membrane bound protein bodies form (Krishnan et al. 1989; Larkins & Hurkman 1978; Taylor et al. 1984). Lending & Larkins (1989) studied zein localization within maize protein bodies, and found α-zein localized to the inner protein body while β- and γ-zein were found at the surface. Shull et al. (1992) found a similar pattern for kafirins in sorghum protein bodies also using immuno-localization techniques.

Shull et al. (1991) proposed nomenclature for identifying sorghum kafirins based on cross-reactivity between kafirin SDS-PAGE bands and antibodies raised against maize zeins. The nomenclature α-, β-, γ-, and δ-kafirin for sorghum is consistent with the naming for maize zeins (Shull & Watterson 1991). The β-zein antibody in this study reacted with the 20kDa kafirin band and had no reactivity to lower molecular weight bands. Regardless, the researchers proposed the 20-, 18-, and 16kDa alcohol soluble proteins to be named β-kafirin.

Sorghum Protein Digestibility

Despite similarities in protein bodies and storage proteins, one contrast between sorghum and maize is protein digestibility. Kurien et al. (1960) made early observations on poor digestibility of sorghum proteins while studying the diets of undernourished adolescent human males in India. In this work they replaced rice with sorghum in the diets, which resulted in lower nitrogen uptake in the body. MacLean et al. (1981) repeated this finding in preschool age children and found 46% nitrogen absorption from sorghum diets compared to 81% absorption from wheat and 73% from maize. Axtell et al. (1981) studied sorghum protein digestibility using rats but were unable to reproduce the results of Maclean and coworkers. Axtell et al. (1981) then developed *in vitro* pepsin digestion methods, which Mertz et al. (1984) correlated to previously observed human digestibility of sorghum and other cereals.

Hamaker et al. (1987) used *in vitro* pepsin digestion to show uncooked sorghum was 80% digestible, similar to digestibility of uncooked maize (83.4%), barley (93.2%), rice (91.1%) and wheat (91.3%). They also demonstrated cooking dramatically lowered sorghum digestibility to 56.3% a phenomenon not observed in the other grains. Hamaker et al. (1987) hypothesized cooking decreased digestibility of sorghum by increasing the amount of disulfide cross-linked kafirins and demonstrated a 25% increase in digestibility of sorghum protein by cooking the flour along with mercaptoethanol (Hamaker et al. 1987).

Using antibodies reactive against α-, β-, and γ-kafirin Oria, Hamaker, & Shull (1995a) demonstrated for uncooked sorghum flour, 30% of α-kafirins, 15.3% of β-kafirin and 13.5% of γ-kafirin remained after 2hrs digestion with pepsin. After cooking, the amount of undigested kafirin increased to 47.9% for α-kafirins, 41.7% of β-kafirin and 28.1% of γ-kafirin. Based on findings from Oria et al. (1995a), Weaver et al. (1998) hypothesized that poor digestibility of sorghum protein resulted from disulfide cross-linking between β- and γ-kafirin at the protein body surface, an effect compounded by cooking.

Sequence of the Sorghum Genome

After publication of the sorghum genome sequence a more complete understanding of sorghum kafirins was possible (Paterson et al. 2009). The genome reveals a single gene for β-kafirin and γ-kafirin with a predicted mass (minus signal peptide) of 18.837 kDa and 20.401 kDa respectively (Belton et al. 2006). This is in contrast to Oria, Hamaker, & Shull (1995a) who reported antibody cross-reactivity to three SDS-PAGE bands for β-kafirin at 16 kDa, 18 kDa and 20 kDa and one band for γ-kafirin at 28 kDa.

Maize zeins were named based on their differential solubility in alcohol and those names were transferred to sorghum kafirins based on cross-reactivity to zein antibodies. In either case the given names α-, β-, and γ-kafirin were originally a classification of extractability in alcohol and not a reflection of genetic similarity. Coincidently, the sorghum genome revealed, as Song

(2004) had reported, multiple tandem and highly similar genes exist for α-kafirin in sorghum, which are unrelated by sequence to β-kafirin and γ-kafirin. β-Kafirin and γ-kafirin have weak sequence similarity (19.5% pairwise identity) although neither structure nor function is known. Figure 1 shows a dendogram grouping α-kafirins separately from the unrelated β- and γ-kafirin, which have sequence similarity to each other.

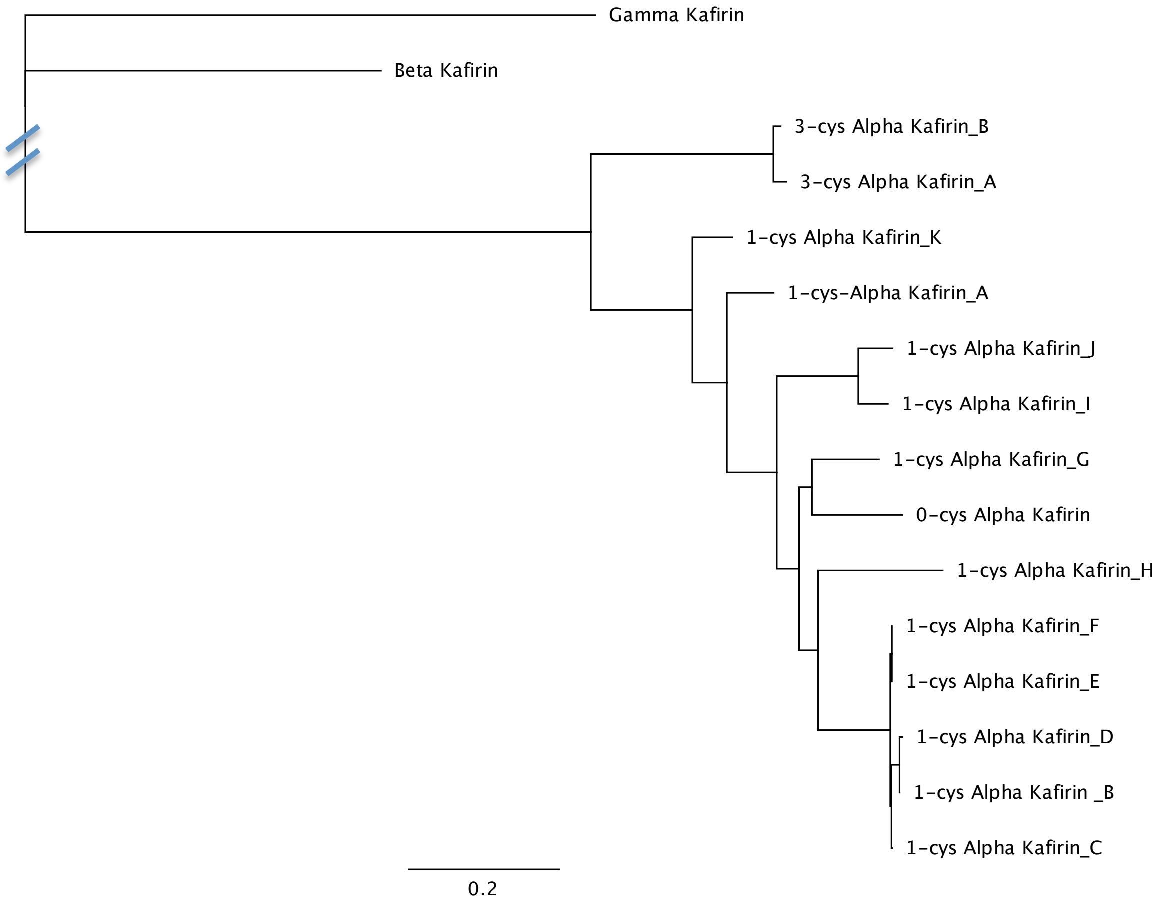


Figure 1 Dendogram for α-, β-, and γ-kafirin based on amino acid sequence. All α-kafirins are related by sequence within the α-kafirins, but are unrelated to β-, and γ-kafirin by sequence. β-Kafirin and γ-kafirin have only weak sequence similarity also shown in **Figure 3**.

Oria et al. (1995a) working in the Hamaker lab hypothesized poor digestibility of sorghum protein was caused by disulfide cross-linking between β- and γ-kafirin. Oria et al. (1995a) based their hypothesis from observations of kafirin localization by Shull et al. (1992), their own *in vitro* digestion studies, and the high cysteine content of both β- and γ-kafirin. Figure 2 shows β- and γ-kafirin sequences and Figure 3 shows their sequence alignment.

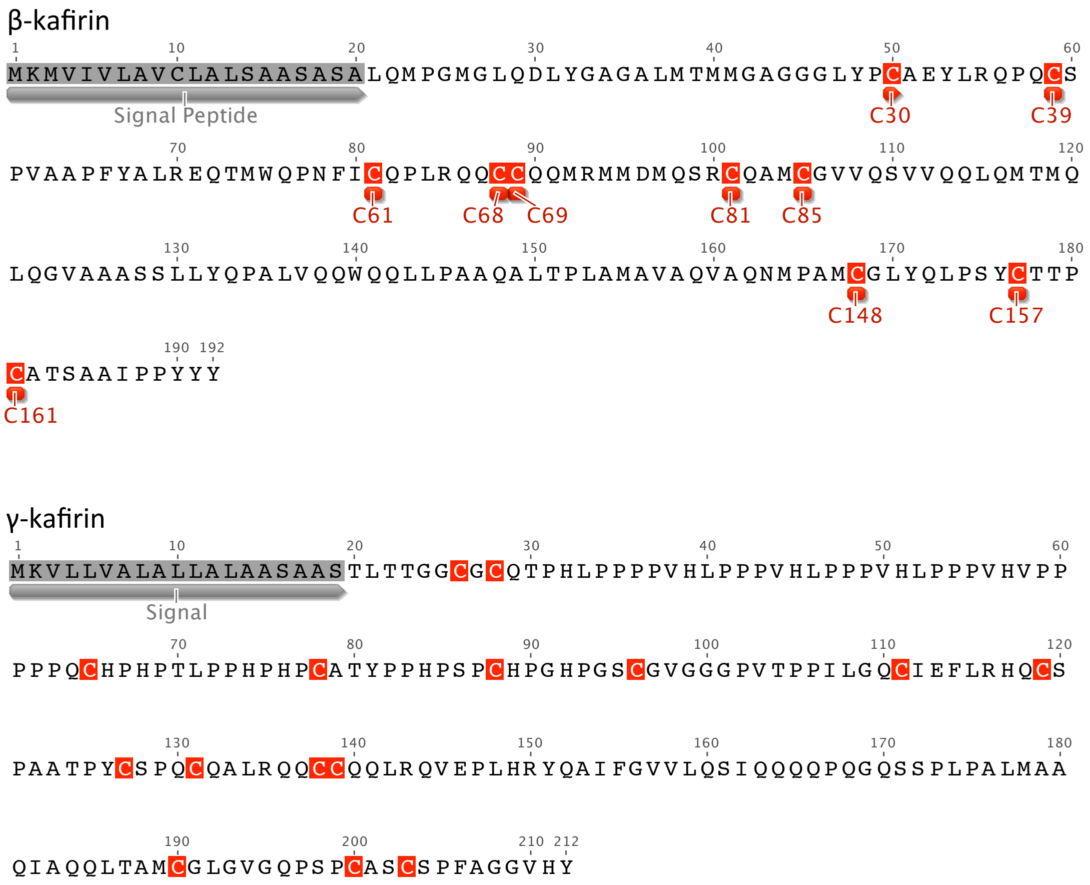
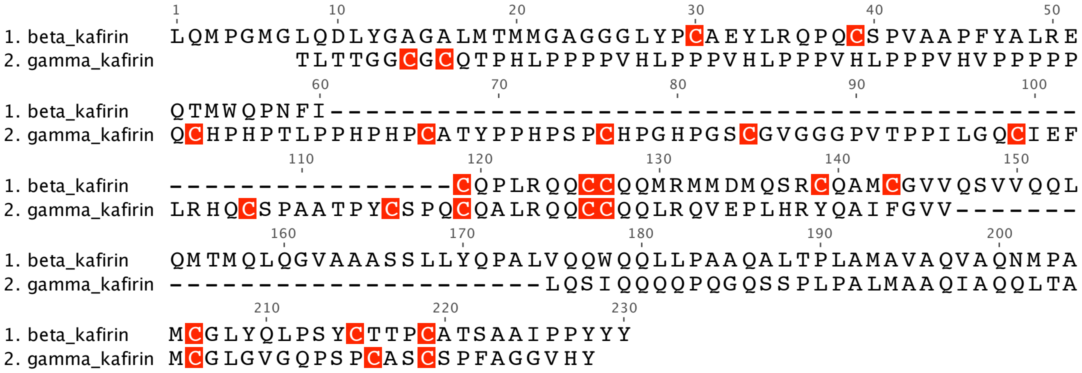

β-kafirin
MKMVIVLAVCLALSAASASALQMPGMGLQDLYGAGALMTMMGAGGGLYPCAEYLRQPQCS
Signal Peptide
C30
C39
PVAAPFYALREQTMWQPNFICQPLRQQCCQQMRMMDMQSRCQAMCGVVQSVVQQLQMTMQ
C61
C68
C69
C81
C85
LQGVAAASSLLYQPALVQQWQQLLPAAQALTPLAMAVAQVAQNMPAMCGLYQLPSYCTTP
C148
C157
CATSAAIPPYYY
C161
γ-kafirin
MKVLLVALALLALAASAASTLTTGGCGCQTPHLPPPPVHLPPPVHLPPPVHLPPPVHVPP
Signal
PPPQCHPHPTLPPHPHPCATYPPHPSPCHPGHPGSCGVGGGPVTPPILGQCIEFLRHQCS
PAATPYCSPQCQALRQQCCQQLRQVEPLHRYQAIFGVVLQSIQQQQPQGQSSPLPALMAA
QIAQQLTAMCGLGVGQPSPCASCSPFAGGVHY

Figure 2 Amino acid sequences for β- and γ-kafirin. β-Kafirin (top) and γ-kafirin (bottom) are shown with signal peptide in grey and cysteine residues in red, annotations below β-kafirin indicate cysteine numbering after removing signal peptide sequence.



```
                 1        10        20        30        40        50
1. beta_kafirin  LQMPGMGLQDLYGAGALMTMMGAGGGLYPCAEYLRQPQCSPVAAPFYALRE
2. gamma_kafirin        TLTTGGCGCQTPHLPPPPVHLPPPVHLPPPVHLPPPVHVPPPPP
                         60        70        80        90        100
1. beta_kafirin  QTMWQPNFI------------------------------------------
2. gamma_kafirin QCHPHPTLPPHPHPCATYPPHPSPCHPGHPGSCGVGGGPVTPPILGQCIEF
                        110       120       130       140       150
1. beta_kafirin  ----------------CQPLRQQCCQQMRMMDMQSRCQAMCGVVQSVVQQL
2. gamma_kafirin LRHQCSPAATPYCSPQCQALRQQCCQQLRQVEPLHRYQAIFGVV-------
                       160       170       180       190       200
1. beta_kafirin  QMTMQLQGVAAASSLLYQPALVQQWQQLLPAAQALTPLAMAVAQVAQNMPA
2. gamma_kafirin ---------------------LQSIQQQQPQGQSSPLPALMAAQIAQQLTA
                      210       220       230
1. beta_kafirin  MCGLYQLPSYCTTPCATSAAIPPYYY
2. gamma_kafirin MCGLGVGQPSPCASCSPFAGGVHY
```

Figure 3 Sequence alignment for β- and γ-kafirin. Only 19.5% pairwise identity is observed between β-, and γ-kafirin. Sequence similarity is highest in the C-terminal end of the amino acid sequences.

Disulfide-linked Kafirin Complexes

In biochemical terms disulfide cross-linking describes proteins linked by intermolecular disulfide bonds into units ranging in size from dimers to large multi-protein complexes. When these complexes or networks become very large the term cross-link takes on a definition from material science or polymer science. In this case cross-linking is defined by ramified or branched chains of disulfide-linked proteins with at least two intermolecular disulfide bonds. Specifically, if one disulfide bond were broken the protein would not be released from the network. By this definition a linear chain of disulfide-linked proteins would not constitute a cross-linked network of proteins. In the following studies disulfide cross-linking is taken biochemically to mean disulfide-linked complexes, since no information is known about the nature of the disulfide cross-links.

Mazhar & Chandrashekar (1993) investigated disulfide cross-linking throughout seed development and found β- and γ-kafirin were extractable without reducing agents from 10-40 days after half bloom. After this point through maturation and desiccation, decreasing amounts of β- and γ-kafirin were extractable without reducing agents (Mazhar & Chandrashekar 1993). Complete extraction was achieved with the use of reducing agents.

This finding led Oria et al. (1995b) to study digestibility of sorghum flour also taken at intervals of seed development. They found no significant reduction in digestibility between cooked and uncooked sorghum flour up to 35 days after half bloom. However, after this stage through desiccation, decreased digestibility was observed for cooked sorghum samples, with lowest digestibility occurring in flour from mature desiccated seeds (Oria et al. 1995b).

To explore the nature of kafirin cross-linking El Nour et al. (1998) extracted kafirins under non-reducing conditions followed by extraction with reducing agents. Using non-reducing conditions they were able to remove single α-, β-, and γ-kafirins plus disulfide-linked dimers, trimers and multimers. The non-reduced extracts were analyzed using two-dimensional gel electrophoresis under non-reducing conditions in the first dimension and reducing conditions in the second. Interestingly, El Nour and coworkers detected no β-kafirin within the disulfide-linked complexes.

El Nour et al. (1998) demonstrated that additional β-kafirin was extractable with the use of reducing agents. But they were not able to determine to what, if anything, the β-kafirin may have been disulfide-linked (El Nour et al. 1998).

Weaver et al. (1998) discovered a highly digestible mutant sorghum cultivar P721Q and using this cultivar, Oria et al. (2000) applied electron microscopy to reveal protein bodies with an unusual folded surface. Oria et al. (2000) then used antibody staining to locate γ-kafirin at the base of each fold, unexpectedly it showed no β-kafirin located in this area (Oria et al. 2000).

This research using high-digestible mutant cultivars indicates genetic control over protein body superstructure, which affects digestibility, and presumably γ-kafirin localization. Winn et al. (2009) studied a related high-digestible mutant P850029 and mapped the high-digestible trait to two loci on chromosome 1. Although the kafirin protein sequences were not provided for the high-digestible mutants the kafirin genes are located on chromosomes 5 (α-kafirin), 9 (β-kafirin), and 2 (γ-kafirin) indicating the high-digestible mutant is possibly not a kafirin-mutant cultivar.

Despite these findings the prevailing theory of protein digestibility in sorghum is that disulfide cross-linking between β- and γ-kafirin at the protein body surface prevents access to more easily digested α-kafirin, a phenomenon believed to be compounded by boiling.

My following work is the first to demonstrate purification of an individual kafirin protein. I found greater than 50% of β-kafirin is extractable in 47.5% ethanol without reducing agents and no evidence of β-kafirin disulfide cross-linked complexes. My work is the first to identify β-kafirin in sorghum by sequence. And chymotrypsin digestion with mass spectral analysis indicates 10 cysteine residues are present in 5 intra-molecular disulfide bonds.

Materials and Methods

Sorghum Flour

I worked with the sorghum cultivar BTx623, because its genome has been sequenced (Paterson et al. 2009). The seeds were provided by Tesfaye Tesso from the Department of Agronomy at Kansas State University.

To begin the study I decorticated 20 grams of seed using a Seedburo Barley Pearler (Seedburo, Des Plaines, IL) fitted with a 30 grit carborundum stone and No. 7 mesh screen. I then milled the decorticated seed using a Quaker City Grinding Mill (QCGS, LLC. Phoenixville, PA) model 4-E attrition style mill. The milled flour had a particle size distribution with 100% less than 420 microns, 76.7% greater than 140 microns and 23.3% les than 140 microns. The flour moisture content was measured to be 11.11% using a Metler Toledo MJ33 moisture analyzer (Metler-Toledo, Columbus, OH).

The protein content of the sorghum flour was 13.97% on a dry matter basis and measured using the LECO system (LECO Corporation, St. Joseph, MI), which determines protein content based on combustible nitrogen. To account for non-protein nitrogen a correction factor of 6.25 was used.

Protein Extraction

I investigated multiple extraction techniques during this research based on methods published by Park & Bean (2003), Hamaker et al. (1995) and others, which are well documented by de Mesa-Stonestreet et al. (2010). In these previously published protocols researchers optimized total protein or total kafirin extraction for quantification, rather than isolating pure proteins for characterization, which was the goal of my work. Nonetheless, I followed general extraction schemes of removing albumins and globulins with 0.5M NaCl followed by removal of water-soluble components with deionized water (pH 6.5).

After removal of water-, and salt-soluble proteins I extracted the ethanol-soluble kafirins in ethanol ranging in concentration from 40%-70% *v/v* deionized water pH 6.5. The kafirin protein remaining after exhaustive extraction with ethanol could be removed with the use of detergents and reducing agents. In this case, I used 1% *w/v* SDS[i] plus 20mM TCEP-HCl[ii]. I studied other reducing agents including DTT[iii] and β-mercaptoethanol, however the advantage of TCEP is its effectiveness over a broad range of pH from pH 1.5-8.5.

In this work I used a flour-to-solvent ratio of 1:20 *w/v* with the typical flour sample size being 50.0 mg. After combining the flour and solvent in a 1.5 mL micro-centrifuge tube the samples were vortexed continuously for 3.0 minutes using a Vortex Genie (Scientific Industries, Bohemia, NY) equipped with a micro-tube foam insert. After each extraction I centrifuged the sample for 3.0 min at 8,000 x g then removed the supernatant (750 µL) for analysis. Each extraction was repeated up to five times by re-suspending the pellet in fresh solvent followed by 3.0 min vortexing and 3.0 min centrifugation. The fresh extracts from each consecutive extraction were either analyzed immediately or lyophilized and stored at -20°C for later study.

RP-HPLC

RP-HPLC[iv] was performed with an HP 1100 HPLC equipped with an auto-sampler and UV/VIS diode-array multiple wavelength detector (Agilent, Santa Clara, CA). RP-HPLC has been used successfully in the study of kafirins and a variety of column choices have been employed (Bean et al. 2011; Blackwell & Bean 2012). In this research I used a Zorbax SB300 C8 column with dimensions of 4.6 x 250mm and 5µm particle size (Agilent, Santa Clara, CA).

In all of my work the mobile phase solvents were A; deionized water plus 0.1% trifluoroacetic acid *v/*v and B; 95% acetonitrile:water plus 0.07% trifluoroacetic acid *v/v.* The acetonitrile and trifluoroacetic acid used in these experiments were both Optima HPLC/MS grade chemicals (Thermo Fisher Scientific, Waltham, MA).

The sample is injected at time 0.0 with the column pre-equilibrated to 30% B and flow rate of 0.1mL/min. Solvent B is held constant at 30% while the flow rate increases from 0.1-1.0 mL/min over 3.0 minutes. Slow isocratic loading allows larger sample sizes to be injected (up to 100µL) while maintaining high chromatographic resolution. At 3.0 minutes solvent B gradually increases at a rate of 0.667%/min for 42.0 minutes reaching a final concentration of 58% B at 45.0 minutes. The column temperature was maintained at 50°C.

The eluent stream was monitored at multiple wavelengths 220nm, 254nm, 260nm, 280nm, and 321nm. Fractions were collected and dried using a Savant SpeedVac Plus concentrator (Thermo Fisher Scientific, Waltham, MA). Dried samples were stored at -20°C for later analysis by mass spectrometry and gel electrophoresis.

SDS-PAGE

Dried samples were rehydrated in 30.0 μL Laemmli Sample Buffer (BioRad, Hercules, CA) then the entire amount was loaded into the loading well. For all SDS-PAGE experiments I used NuPAGE 4-12% Bis-Tris precast gels in the 1.5mm 10 well format and BenchMark™ Protein Ladder standards with a molecular weight range from 10-220 kDa. Duplicate samples were run with the addition of DTT to reduce disulfide cross-links between kafirins. Gels were run at 200V constant voltage following the manufacturers' instructions.

Mass Spectrometry

I used a Bruker Daltronics Ultraflex III MALDI-TOF/TOF[v] mass spectrometer (Bruker, Bremen, Germany) located in the Biotechnology Core Lab at Kansas State University. This mass spectrometer uses a nitrogen laser operating at 337nm.

I rehydrated the lyophilized samples with 25.0 µL of 70% *v/v* acetonitrile/deionized water plus 0.1% trifluoroacetic acid. In each MALDI-TOF MS measurement samples were affixed to the target using the dried droplet method (Chapman 2000). In this method 1.0 µL of matrix solution is applied onto the MALDI-TOF MS target then immediately 1.0 µL of analyte is added and together allowed to dry fully.

Various matrices were needed to effectively ionize the mixture of high and low molecular weight proteins and small molecules in the extracts. I used sinapinic acid[vi] to study larger intact proteins, CHCA[vii] for small molecules and DHB[viii] for small proteins and peptides. I used 1,5 DAN[ix] for sequencing and positive identification of β-kafirin using in-source decay methods (Fukuyama et al. 2006).

All matrix solutions were prepared using 70% acetonitrile *v/v* ultrapure water plus 0.1% trifluoroacetic acid with a matrix concentration of 20 mg/mL.

In all experiments including the in-source decay method the MALI-MS detector was operated in linear positive mode. I collected data at a resolution of 2.0 GS/s with a detector gain of 23x. Using a large laser focus and frequency range of 22-50Hz, data was accumulated from 1000 shot bursts averaging 10,000 shots per sample. Matrix peaks were suppressed by gating ions with *m/z* of 1000 Da or less.

The MALDI-TOF MS was calibrated using Bruker Protein Calibration Standard I.

Table 1 MALDI-TOF MS Calibration Matrix Table Protein Standard I

Proteins	Charge State	Average *m/z*
Insulin	$[M+H]^+$	5734.51
Ubiquitin	$[M+H]^+$	8565..76
Cytochrom C	$[M+H]^+$	12360.97
Myoglobin	$[M+H]^+$	16952.30
Cytochrom C	$[M+2H]^{2+}$	6180.99
Myoglobin	$[M+2H]^{2+}$	8476.65

Circular Dichroism Spectroscopy

I used a JASCO J-815 series circular dichroism spectrometer to study the secondary structure of β-kafirin isolated and purified in earlier experiments. β-Kafirin is not soluble in aqueous buffers therefore I chose to rehydrate the lyophilized protein in 47.5% ethanol *v/v* ultrapure water with pH 6.5. I collected CD^x spectra for both the sample and solvent blank using a quartz cuvette with 3.0mm path-length and total volume of 300.0 μL. For each sample I collected five consecutive CD spectra at a rate of 50nm/min with absorbance measured from 260-190 nm in 1 nm increments. Deconvolution and secondary structure prediction were done with CDSSTR, SELCON3 and CONTIN (Sreerama & Woody 2000)

Proteolysis of β-Kafirin

To study intramolecular disulfide bond patterns of β-kafirin I digested β-kafirin using mass spectrometry grade chymotrypsin (Fisher-Thermo, Waltham, MA). Chymotrypsin cleaves immediately C-terminal to bulky hydrophobic residues Phe, Tyr and Trp with high specificity and at Leu and Met with less specificity. Since chymotrypsin has activity in high concentration organic solvents (Klibanov 2001) I prepared 50mM Tris-HCl (pH 8.0), 2mM calcium chloride digestion buffer in 80% acetonitrile *v/v* ultrapure water.

Digesting in 80% acetonitrile was an advantage since β-Kafirin has poor solubility in aqueous buffers and high solubility in acetonitrile >50%. I rehydrated lyophilized β-kafirin in

50μL digestion buffer then added 2μL of chymotrypsin with a concentration of 1μg/μL. I confirmed chymotrypsin activity using bovine serum albumin and monitored chymotrypsin auto-digestion using chymotrypsin blank digests. The digestion temperature was controlled at 35°C using an Eppendorf Thermomixer R (Eppendorf, Hamburg, Germany). I monitored the progress of digestion at various time intervals up to 24 hours using MALDI-TOF MS with SA[xi] and DHB as matrices.

Non-reduced fragments from cleavage of β-kafirin by chymotrypsin were measured using MALDI-TOF MS to determine their mass. I then reduced the sample using 10mM DTT (Fisher-Thermo, Waltham, MA) and measured the reduced fragment mass using MALDI-TOF MS. Disulfide mapping was done by comparing both reduced and non-reduced mass spectra to the predicted sites for cleavage of β-kafirin by chymotrypsin.

Sequence Alignment

I used a package of programs designed for comparing proteins and families of proteins with weak sequence similarity. The software called SEQALIGN was developed in the Reeck lab by Kirk Clark (SEQALIGN 1991) and is not publicly available. SEQALIGN is an implementation of programs written by David Teller, based on an algorithm of Sankoff & Cedergren (1983). When comparing two sequences SEQALIGN randomizes each sequence and performs a sequence alignment based on the algorithm from Needleman & Wunsch (1970). A scoring matrix is applied to compute the alignment score for the randomized and non-randomized pairing.

The randomization and alignment is done repeatedly to generate a test statistic for the comparison at a given number of gaps. The *z*-value indicates the standard deviations a paired alignment score falls from the score of the randomized alignment. A paired alignment with higher z-score gives evidence the alignment is unlikely due to chance.

The SEQALIGN program recursively generates the test statistic for alignments with zero gaps, through any number of specified gaps. For all alignments adding gaps will always increase the alignment score, but Clark (1991) found a sharp increase in the alignment score between related proteins when the optimum gap number is found.

For my analysis of the kafirin proteins I used the McLachlan scoring matrix (McLachlan 1971). Each test statistic was generated from 500 randomized pairings with 0-30 gaps.

Homology Modeling

To build homology models for β-kafirin I used the software I-TASSER developed by Zhang (2008). I-TASSER uses multiple threading alignment to align a target sequence to the sequence of proteins with known three-dimensional structure. In automated mode I-TASSER searches a non-redundant set of proteins from the Brookhaven Protein Data Bank (PDB). In directed mode the user provides a template structure and sequence alignment for the target and template proteins.

In directed mode I provided the PDB structure file for CHFI[xii] (PDB:1BEA), a maize protein with known three-dimensional structure solved by Behnke et al. (1998) in the Reeck lab. The sequence alignment for β-kafirin and CHFI was generated using SEQALIGN.

I-TASSER computes a score to judge the accuracy of the modeled structure called c-score. The c-score is generated by comparing the convergence parameters of the modeled structure against the convergence parameters for proteins of known structure also solved using I-TASSER. It provides an estimate of RMSD that is well correlated with the RMSD for modeled proteins of known structure (Zhang 2008).

The c-score ranges from -5.0 – 2.0 and more positive scores have higher predicted accuracy.

Results

Kafirin Protein Extraction

The alcohol soluble kafirins from sorghum α-, β-, and γ-kafirin equal 78%, 8%, and 13% of total protein respectively as reported by Watterson et al. (1993). ESTs exist for δ-kafirin in sorghum although δ-kafirin has not been detected at the protein level (Izquierdo & Godwin 2005). My aim in this work was to isolate and purify kafirins, specifically β-kafirin, for further characterization. I extracted kafirins from sorghum flour using 70% ethanol without reducing agents shown in Figure 4. This method produced an extract containing α-kafirin, β-kafirin and other unidentified proteins. Dimers of α-kafirin were reduced with 20mM DTT after extraction and during SDS-PAGE.

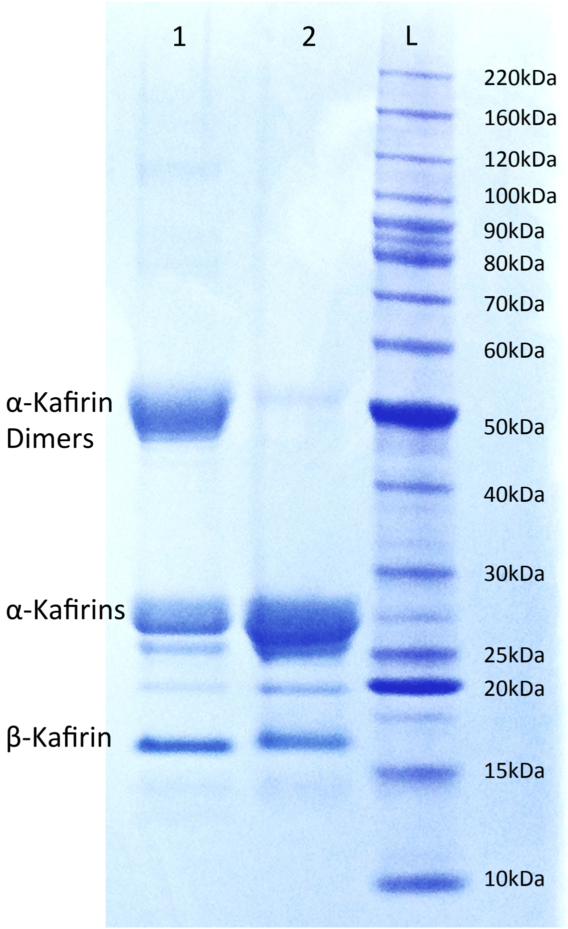


Figure 4 Sorghum kafirins extracted in 70% Ethanol without reducing agents. (Lane 1) Non-reduced sorghum kafirins extracted in 70% ethanol and run in non-reducing conditions. (Lane 2) Non-reduced sorghum kafirins extracted in 70% ethanol and reduced with 20mM DTT after extraction. (Lane L) Protein molecular weight standards. Identification was made by in-gel-digest and MS/MS peptide mass fingerprinting.

When the 70% ethanol- 2% SDS extract is analyzed using HPLC a larger number of components are resolved Figure 5. β-Kafirin elutes at 38 min in the non-reduced extract and single α-kafirins and dimers with α-kafirin elute in a series of broad overlapping peaks between 40-47 min. When the extract is reduced with 20mM DTT the overlapping peaks resolve into sharp peaks containing 1 or 2 components determined by MALDI-TOF MS. After reduction with DTT β-kafirin elution is delayed to 45 minutes.

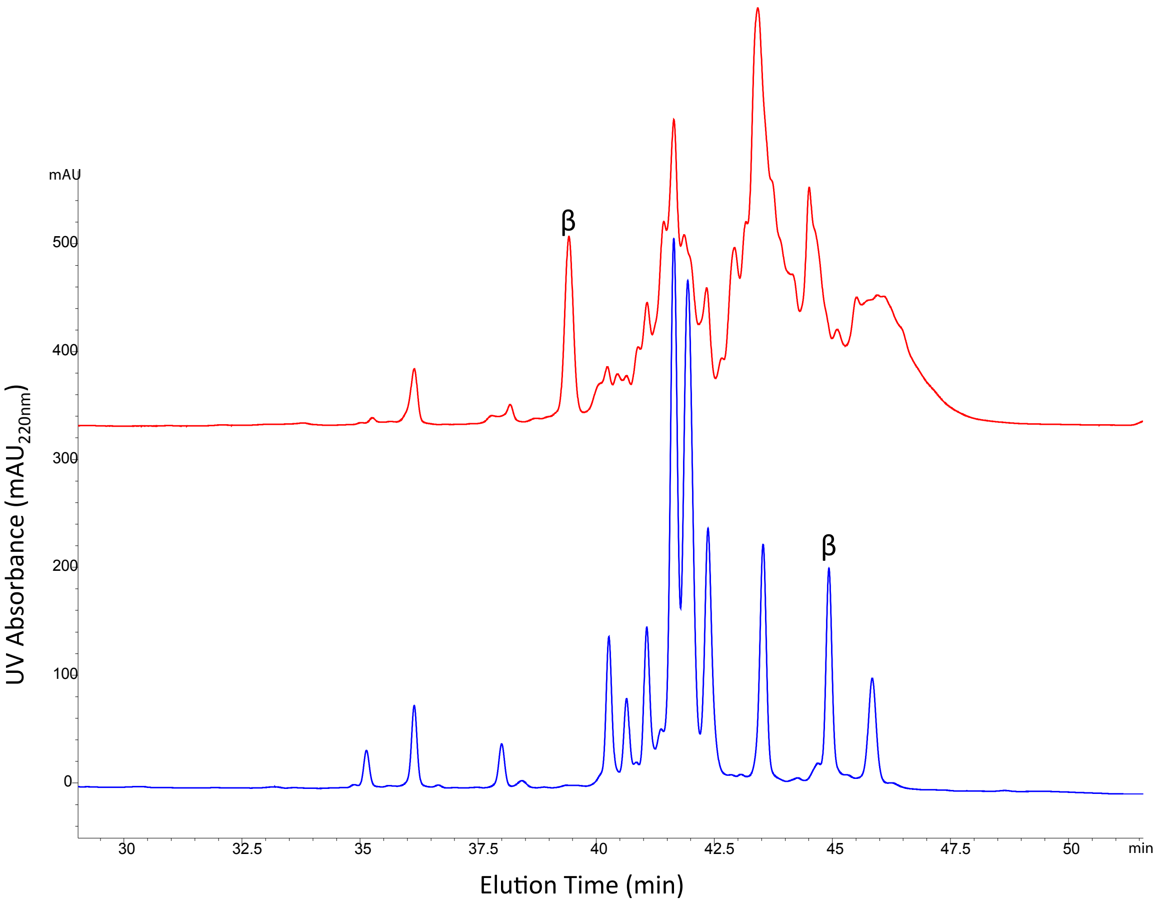
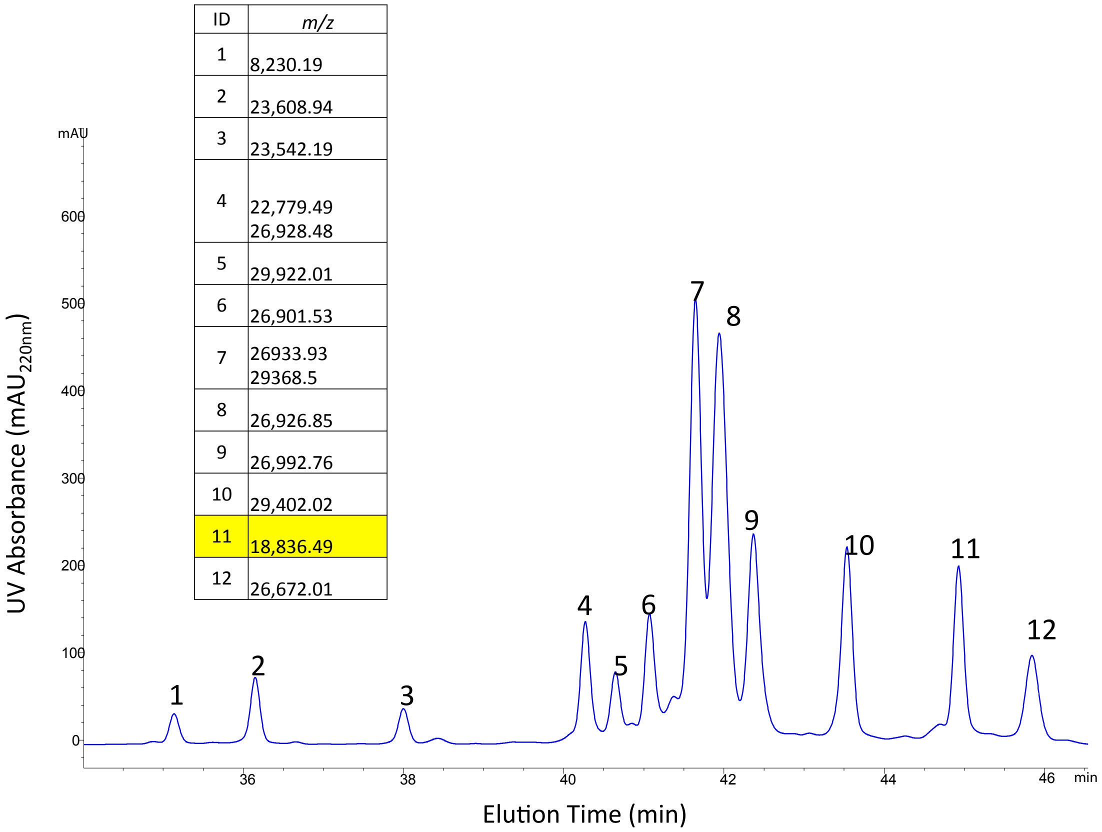


Figure 5 HPLC Chromatogram of sorghum flour extracted in 70% Ethanol plus 2% SDS. 70% ethanol- 2% SDS extract without reducing agents (top red). 70% ethanol 2% SDS extract reduced with 20mM DTT after extraction (bottom blue). β-Kafirin (β) has a delayed elution time after reduction with DTT, which indicates a structural change after reducing disulfide bonds. The β-kafirin peak height and area is unchanged after reduction indicating β-kafirin is not contained in the broad peaks (40-47 min top), which are resolved after reduction with DTT.

I collected each peak eluted during HPLC and used MALDI-TOF MS to determine the mass of each component contained in those peaks. Figure 6 shows a chromatogram for the 70% ethanol 2% SDS extract annotated with numbers based on elution time. Peak number 1 was not present in the non-reduced extract indicating it may have been released from a disulfide-linked complex after reduction with DTT. The mass for peak 1 (8.2kDa) is not consistent with the mass for known kafirins and has not yet been identified. Peaks 2-10 and 12 are consistent with the predicted mass for α-kafirins (23-29kDa) and peak 11 contains β-kafirin, which has a predicted mass of 18.837kDa. Interestingly peaks 4 and 7 each contain two proteins with apparently similar chromatographic properties albeit different mass.



ID	*m/z*
1	8,230.19
2	23,608.94
3	23,542.19
4	22,779.49 26,928.48
5	29,922.01
6	26,901.53
7	26933.93 29368.5
8	26,926.85
9	26,992.76
10	29,402.02
11	18,836.49
12	26,672.01

Figure 6 HPLC chromatogram for kafirins extracted in 70% ethanol plus 2%SDS then reduced with 20mM DTT. Peaks are numbered by elution order (1-12). (Inset) Table of protein mass for proteins contained in each peak measured by MALDI-TOF MS. (Peak 1) Unidentified protein mass 8,230 Da, (Peaks 2-10) Consistent with predicted mass for α-kafirins 23,000-29,000 Da (Peak 11) β-kafirin 18,837 Da.

Purification of β-kafirin

I optimized the extraction of β-kafirin from sorghum flour by using 47.5% ethanol without reducing agents. The HPLC chromatogram for the 47.5% EtOH extract in Figure 7 shows the sample greatly enriched in β-kafirin compared to α-kafirin and other alcohol soluble proteins.

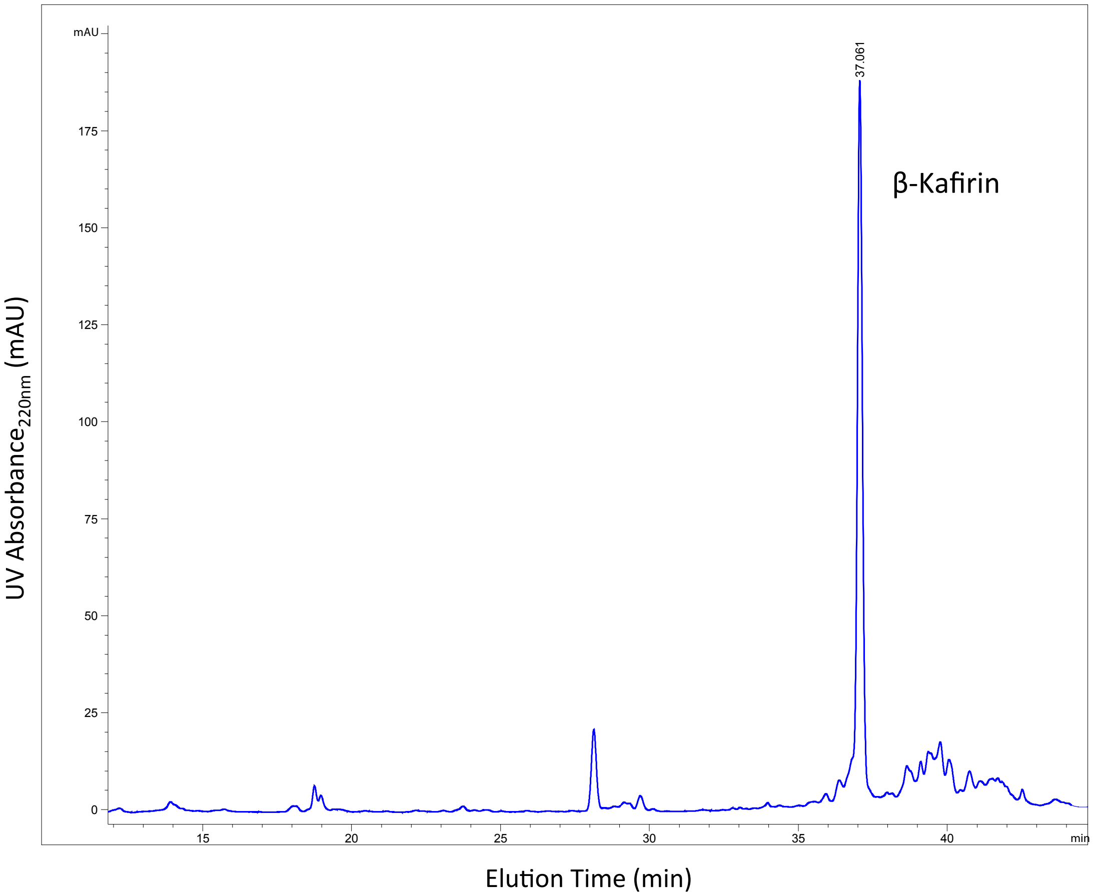


Figure 7 HPLC chromatogram of sorghum flour extracted with 47.5% ethanol. β-Kafirin is enriched in this sample compared to other proteins allowing for purification by HPLC.

I used this extract to study the chromatographic properties for β-kafirin after reduction with DTT. The results shown in Figure 8 confirm earlier observations of delayed elution after reduction with DTT. This result provides evidence for intra-molecular disulfide bonds, which presumably make β-kafirin compact. After reduction with DTT β-kafirin exhibits more hydrophobic behavior during chromatography. There is no evidence for β-β dimers or larger β-kafirin complexes in the 47.5% extract, which would appear as increased peak height and area after reduction with DTT.

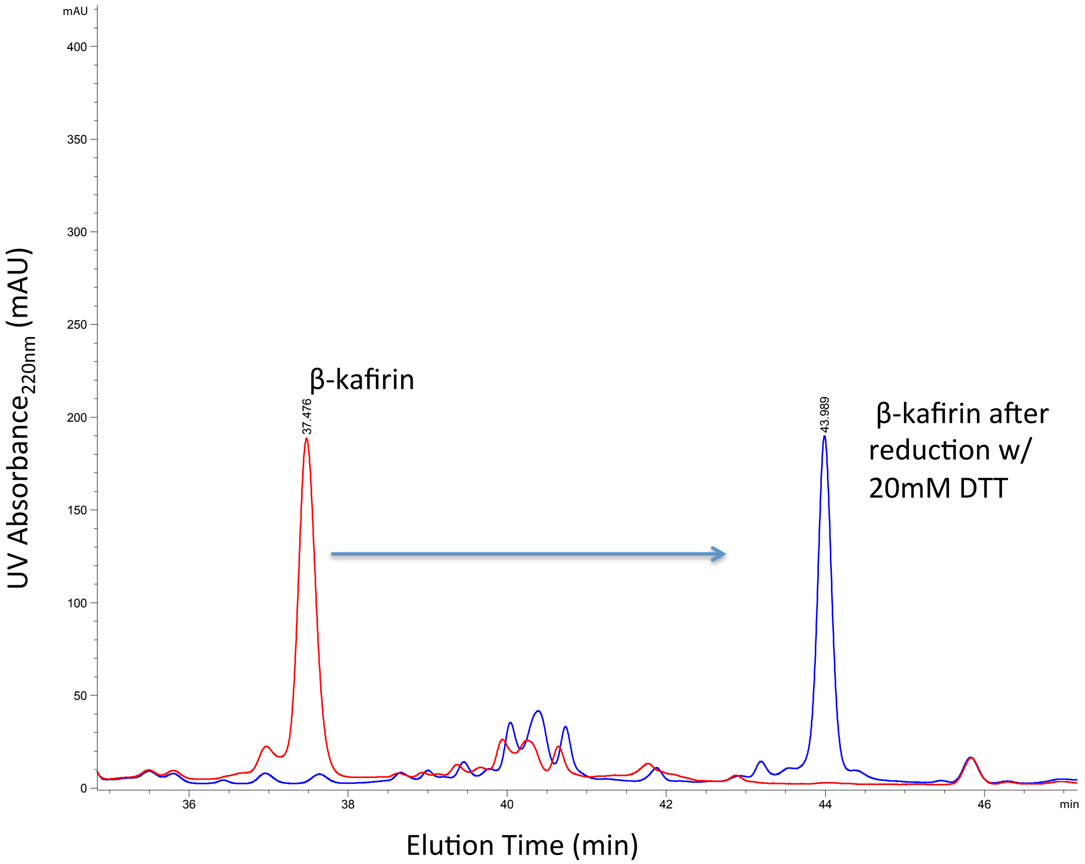


Figure 8 HPLC chromatogram β-kafirin extracted in 47.5% ethanol. β-Kafirin extracted in 47.5% ethanol without reducing agents (red) elutes at 37.4 min. After reducing the 47.5% ethanol extract with 20mM DTT β-kafirin elutes at 43.9 min under identical chromatographic conditions.

The 47.5% ethanol extract was used to purify a larger quantity of β-kafirin by pooling extracts and lyophilizing for later characterization. The pooled β-kafirin was further purified by HPLC under different chromatographic conditions than used for isolation. Figure 9 shows the single purified β-kafirin peak from HPLC, which also migrates as a single band between 15-20kDa when analyzed using SDS-PAGE. The β-kafirin peak collected in this final stage of purification had a mass of 18,837.49 Da measured by MALDI-TOF MS Figure 10, which is precisely the predicted mass for β-kafirin 18,837.

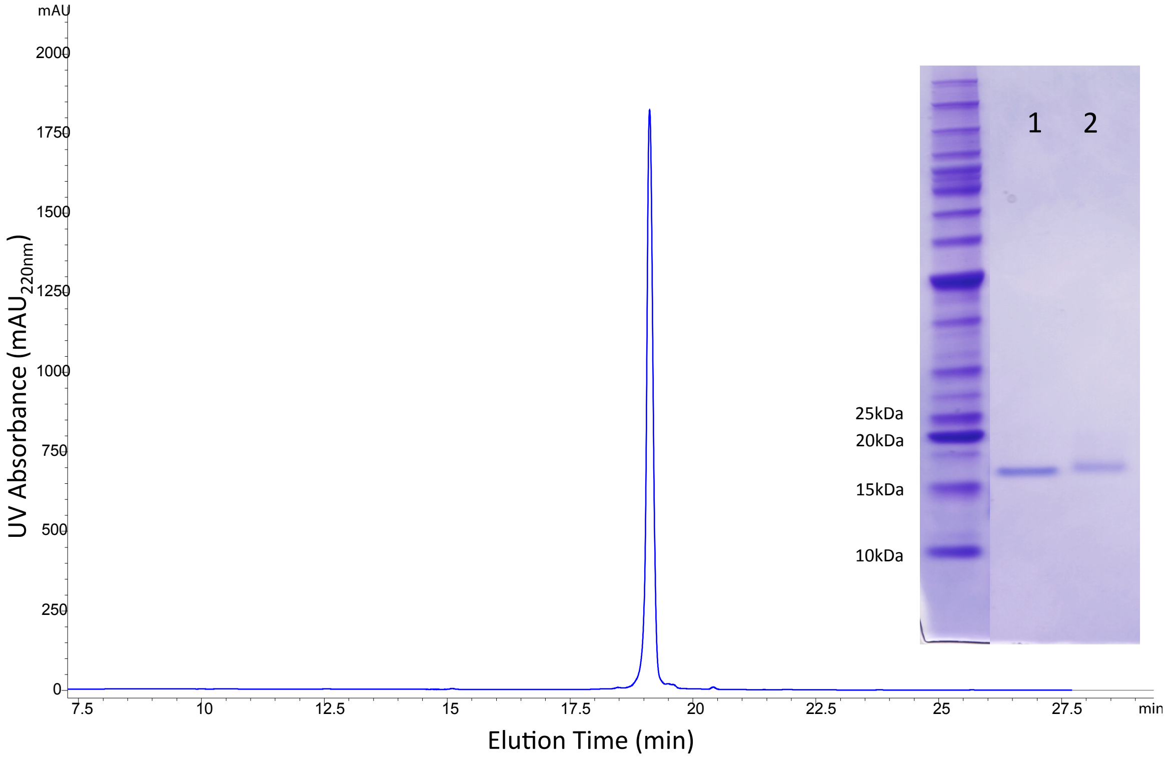


Figure 9 β-kafirin after purification from 47.5% ethanol extract using HPLC. HPLC chromatogram of purified β-kafirin, showing a single peak eluting at 19min, run under different chromatographic conditions than used for isolation. SDS-PAGE of β-kafirin purified by HPLC (Inset Lane 1) β-kafirin migrates in a single band run under non-reducing conditions (Inset Lane 2) β-kafirin migrates to a single band with higher apparent mass after reduction in 20mM DTT.

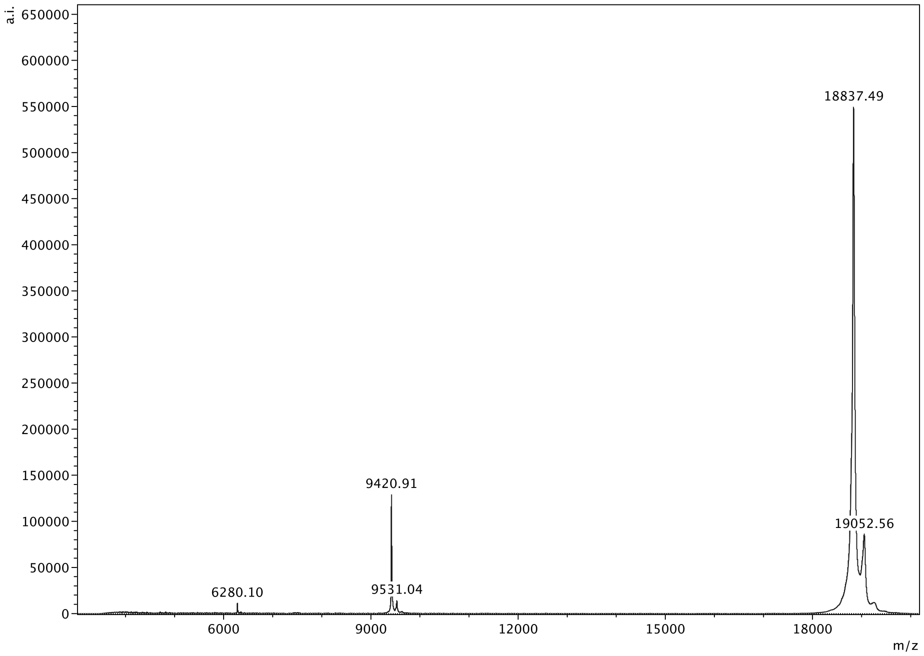
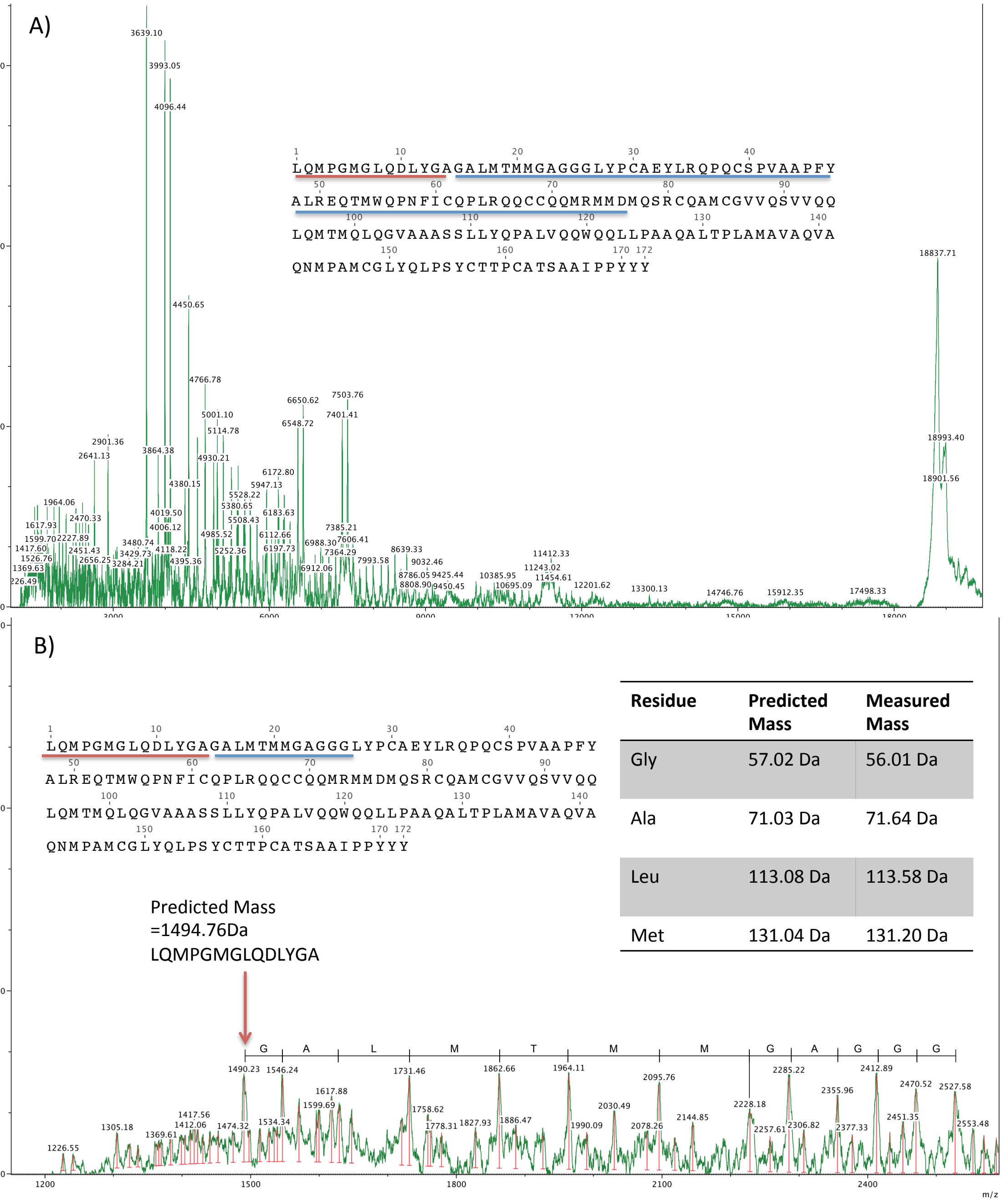


Figure 10 MALDI-TOF MS mass spectrum of β-kafirin extracted in 47.5% ethanol and purified with HPLC. Three charge states are observed for β-kafirin +1*m/z* (18,837 Da), +2*m/z* (9,420 Da) and +3*m/z* (6,280 Da).

β-Kafirin Identification

Purified β-kafirin was positively identified by in-source-decay with using the matrix 1,5-DAN. The fragment ions shown in Figure 11 provide 44% sequence coverage.



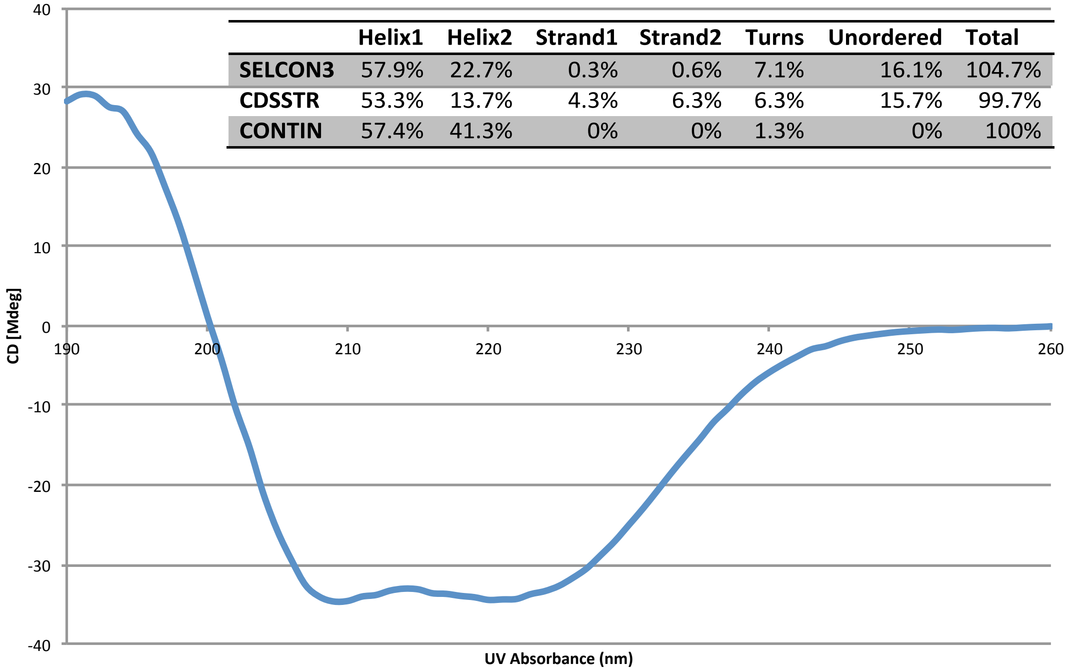
Residue	Predicted Mass	Measured Mass
Gly	57.02 Da	56.01 Da
Ala	71.03 Da	71.64 Da
Leu	113.08 Da	113.58 Da
Met	131.04 Da	131.20 Da

Figure 11 MALDI-TOF MS In-Source-Decay identification of β-kafirin using 1,5-DAN. (A) Mass spectrum showing 44% sequence coverage from the N-terminal to Glu-76. (B) Expanded view showing deduced N-terminal amino acid sequence by mass subtraction. Peaks represent fragment ions resulting from random cleavage of β-kafirin by 1,5-DAN.

The expanded mass spectra showing the observed the entire sequence coverage by MALDI-TOF MS in-source-decay can be seen in Appendix Figure A.1

β-Kafirin Characterization

Purified β-kafirin was used to characterize the intact protein and also served as a standard to identify β-kafirin in SDS-PAGE, HPLC and MALDI-TOF MS. I began by using CD to measure the secondary structure composition of β-kafirin in 47.5% EtOH. The CD spectrum shown in Figure 12 was used along with secondary structure prediction software SELCON3, CDSSTR and CONTIN to determine the secondary structure content of β-kafirin. Each program produced comparable results indicating strong α-helix (Helix1) content between 53-57%. The prediction also suggests 13-22% Helix2 (distorted helix) and up to 16% unordered structure. This finding confirms our earlier hypothesis that β-kafirin has a regularly folded structure.



	Helix1	Helix2	Strand1	Strand2	Turns	Unordered	Total
SELCON3	57.9%	22.7%	0.3%	0.6%	7.1%	16.1%	104.7%
CDSSTR	53.3%	13.7%	4.3%	6.3%	6.3%	15.7%	99.7%
CONTIN	57.4%	41.3%	0%	0%	1.3%	0%	100%

Figure 12 CD Spectrum for β-kafirin in 47.5% ethanol pH 6.5. Secondary structure prediction programs SELCON3, CDSSTR and CONTIN were used to predict secondary structure content from the CD spectrum for β-kafirin. The predictions (table inset) are in close agreement and indicate 53-57% alpha-helix character with up to 16% unordered structure.

To determine the amount of β-kafirin extractable without reducing agents I did sequential extractions with 47.5% EtOH and analyzed the extracts using SDS-PAGE Figure 13 and HPLC (not shown). I determined 53.4 % of β-kafirin is extractable without reducing agents calculated by integrating peak area for β-kafirin in each extraction shown in Table 2. Repeated experiments indicate this amount could be higher than 60% (data not shown).

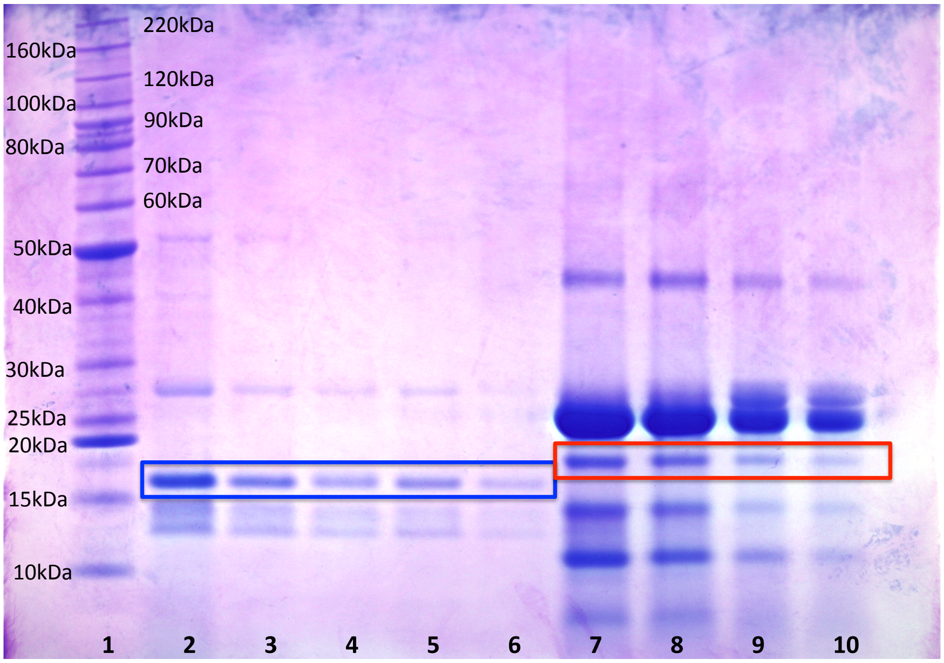


Figure 13 SDS-PAGE sequential extraction with 47.5% ethanol under non-reducing and reducing conditions. (lane 1) molecular weight standards. (lanes 2-6) Sorghum flour extracted sequentially with 47.5% ethanol for 5 rounds. (lanes 7-10) Extraction rounds 6-9 with 47.5% ethanol plus 20mM DTT. β-Kafirin migrates between 15-20kDa in the non-reduced extract run in non-reducing conditions (blue box). β-Kafirin migrates to a higher apparent mass when reduced during extraction and run under reducing conditions (red box).

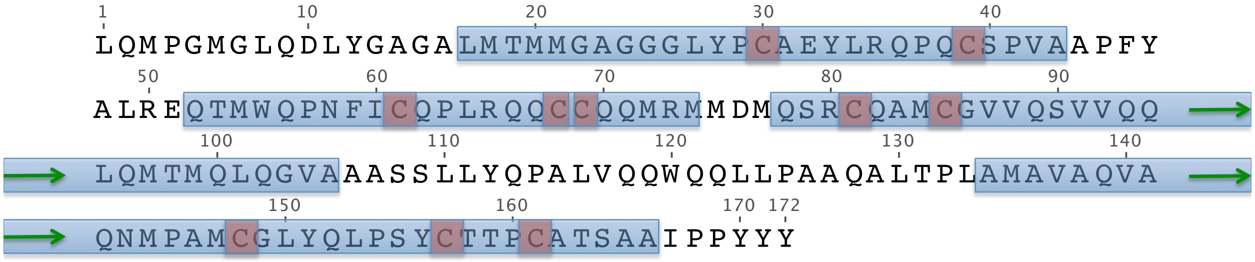
Table 2 Integrated area for β-kafirin peak during sequential extraction with 47.5% EtOH.

Extraction	*Peak Area*	*% of Total β-kafirin*
1	1333	28.38%
2	510.6	10.87%
3	279.6	5.95%
4	165.9	3.53%
5	93.5	1.99%
6	125.4	2.67%
Total	2508	53.40%
7	874.3	18.61%
8	533.3	11.35%
9	407.9	8.68%
10	373.3	7.95%
Total	2188.8	46.60%
Grand Total	4696.8	100.00%

Intramolecular disulfide bonds

To study the state of cysteine residues in β-kafirin I did the Ellman's reagent test for free sulfhydryls in both intact and urea-denatured β-kafirin. The results of these tests were negative in both samples indicating all cysteine residues were in the oxidized state within intramolecular disulfide bonds.

To test this hypothesis and in attempt to map the disulfide linkages I digested β-kafirin with chymotrypsin. The limit digest was monitored over 24 hrs using MALDI-TOF MS to determine the mass of fragments released. After 16hrs of digestion with chymotrypsin, the limit digest remained largely intact with a measured mass between 12-14kDa (not shown). After reducing the partially digested β-kafirin with DTT the disulfide cross-linked fragments were released into single peptide chains. A schematic diagram showing the fragments detected by MALDI-TOF MS is shown in Figure 14. I was unable to determine the precise disulfide pairings using the limit digest method although they are cross-linked in such a way that cleaved fragments are not released by digestion.

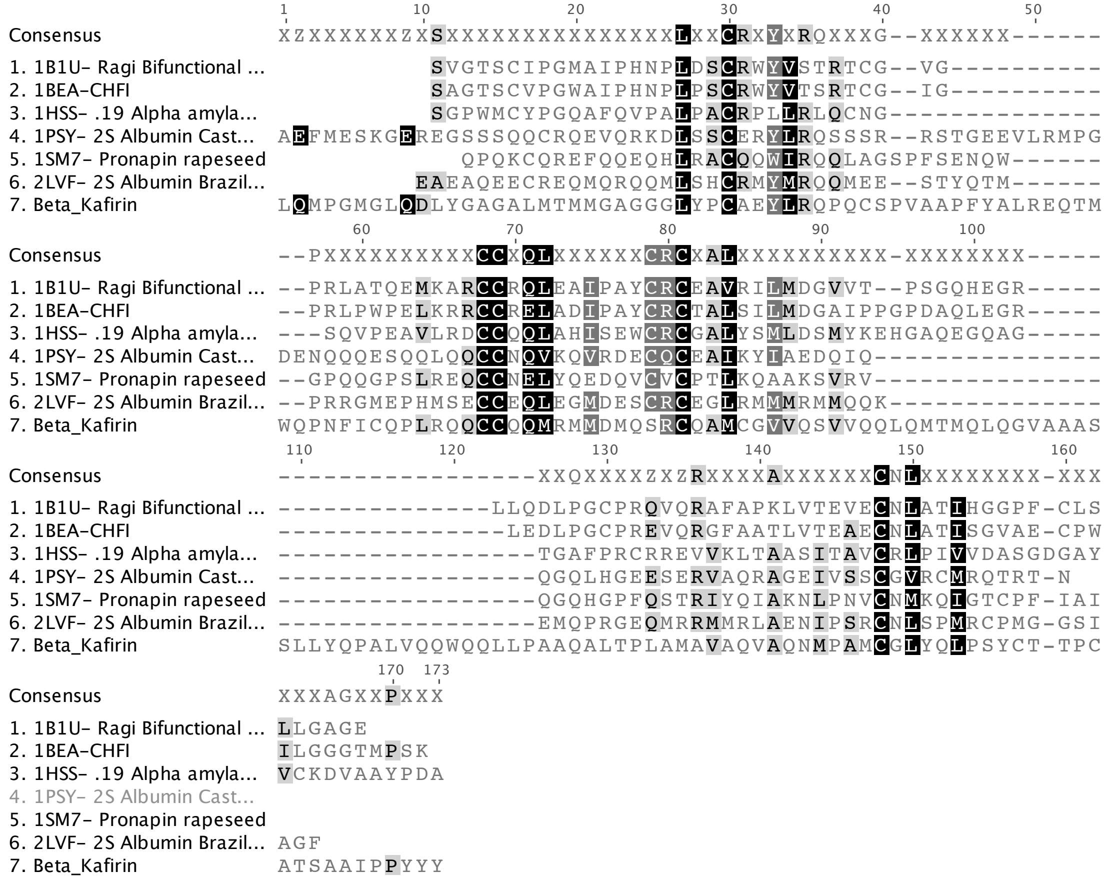


1
10
20
30
40
LQMPGMGLQDLYGAGALMTMMGAGGGLYPCAEYLRQPQCSPVAAPFY
50
60
70
80
90
ALREQTMWQPNFICQPLRQQCCQQMRMMDMQSRCQAMCGVVQSVVQQ
100
110
120
130
140
LQMTMQLQGVAAASSLLYQPALVQQWQQLLPAAQALTPLAMAVAQVA
150
160
170
172
QNMPAMCGLYQLPSYCTTPCATSAAIPPYYY

Figure 14 Schematic diagram showing β-kafirin fragments remaining after digestion with chymotrypsin. The chymotrypsin limit digest was measured by MALDI-TOF MS after 2, 16 and 24 hrs. The undigested mass remaining after 16hrs ranged from 12-14kDa, which included multiple missed cleavages. After reduction with DTT the disulfide linked fragments shown in blue were released.

Homology Modeling β-kafirin

I performed a BLAST search for proteins with sequence similarity to β-kafirin and known three-dimensional structure. The multiple sequence alignments shown in Figure 15 highlight the 5 conserved cysteine residues between all members.



```
                                1        10        20        30        40        50
Consensus                       XZXXXXXXZXSXXXXXXXXXXXXXXXLXXCRXYXRQXXXG--XXXXXX------

1. 1B1U- Ragi Bifunctional ...            SVGTSCIPGMAIPHNPLDSCRWYVSTRTCG--VG----------
2. 1BEA-CHFI                              SAGTSCVPGWAIPHNPLPSCRWYVTSRTCG--IG----------
3. 1HSS- .19 Alpha amyla...               SGPWMCYPGQAFQVPALPACRPLLRLQCNG--------------
4. 1PSY- 2S Albumin Cast...     AEFMESKGEREGSSSQQCRQEVQRKDLSSCERYLRQSSSR--RSTGEEVLRMPG
5. 1SM7- Pronapin rapeseed                  QPQKCQREFQQEQHLRACQQWIRQQLAGSPFSENQW------
6. 2LVF- 2S Albumin Brazil...            EAEAQEECREQMQRQQMLSHCRMYMRQQMEE--STYQTM------
7. Beta_Kafirin                 LQMPGMGLQDLYGAGALMTMMGAGGGLYPCAEYLRQPQCSPVAAPFYALREQTM
                                     60        70        80        90        100
Consensus                       --PXXXXXXXXXXCCXQLXXXXXXCRCXALXXXXXXXXXX-XXXXXXXX-----

1. 1B1U- Ragi Bifunctional ...  --PRLATQEMKARCCRQLEAIPAYCRCEAVRILMDGVVT--PSGQHEGR-----
2. 1BEA-CHFI                    --PRLPWPELKRRCCRELADIPAYCRCTALSILMDGAIPPGPDAQLEGR-----
3. 1HSS- .19 Alpha amyla...     ---SQVPEAVLRDCCQQLAHISEWCRCGALYSMLDSMYKEHGAQEGQAG-----
4. 1PSY- 2S Albumin Cast...     DENQQQESQQLQQCCNQVKQVRDECQCEAIKYIAEDQIQ---------------
5. 1SM7- Pronapin rapeseed      --GPQQGPSLREQCCNELYQEDQVCVCPTLKQAAKSVRV---------------
6. 2LVF- 2S Albumin Brazil...    --PRRGMEPHMSECCEQLEGMDESCRCEGLRMMMRMMQQK--------------
7. Beta_Kafirin                 WQPNFICQPLRQQCCQQMRMMDMQSRCQAMCGVVQSVVQQLQMTMQLQGVAAAS
                                 110       120       130       140       150       160
Consensus                       -----------------XXQXXXXZXZRXXXXAXXXXXXCNLXXXXXXXX-XXX

1. 1B1U- Ragi Bifunctional ...  --------------LLQDLPGCPRQVQRAFAPKLVTEVECNLATIHGGPF-CLS
2. 1BEA-CHFI                    ---------------LEDLPGCPREVQRGFAATLVTEAECNLATISGVAE-CPW
3. 1HSS- .19 Alpha amyla...     -----------------TGAFPRCRREVVKLTAASITAVCRLPIVVDASGDGAY
4. 1PSY- 2S Albumin Cast...     -----------------QGQLHGEESERVAQRAGEIVSSCGVRCMRQTRT-N
5. 1SM7- Pronapin rapeseed      -----------------QGQHGPFQSTRIYQIAKNLPNVCNMKQIGTCPF-IAI
6. 2LVF- 2S Albumin Brazil...    -----------------EMQPRGEQMRRMMRLAENIPSRCNLSPMRCPMG-GSI
7. Beta_Kafirin                 SLLYQPALVQQWQQLLPAAQALTPLAMAVAQVAQNMPAMCGLYQLPSYCT-TPC
                                       170 173
Consensus                       XXXAGXXPXXX

1. 1B1U- Ragi Bifunctional ...  LLGAGE
2. 1BEA-CHFI                    ILGGGTMPSK
3. 1HSS- .19 Alpha amyla...     VCKDVAAYPDA
4. 1PSY- 2S Albumin Cast...
5. 1SM7- Pronapin rapeseed
6. 2LVF- 2S Albumin Brazil...   AGF
7. Beta_Kafirin                 ATSAAIPPYYY
```

Figure 15 Multiple sequence alignment for β-kafirin and proteins with sequence similarity and known three-dimensional structure. Although these proteins only have 17.7% pairwise identity 5 of 10 cysteine residues are strictly conserved in all members.

Using SEQALIGN I did multiple pairwise comparisons between β-kafirin and each protein found in the BLAST search. Each alignment had statistically significant although weak similarity and therefore I chose the top two hits CHFI (PBD:1BEA) and Pronapin (PDB:1SM7) as template structures for model building. I compared the β-kafirin homology models using the I-

TASSER c-score and found using CHFI as the template gave c-score = -0.82, ExpRMSD = 6.8+/-4.0 versus Pronapin template c-score = -0.98, ExpRMSD = 7.1+/-4.2. Using c-score as a criteria we chose to continue modeling β-kafirin from CHFI a protein whose structure had been solved in the Reeck lab.

I built several models using various gapped alignments and best model c-score was achieved using the sequence alignment with three gaps shown in Figure 16.

```
        1         .         +         .         +         .         +         .         +         .         +   50
A Beta Kafirin:   L Q M P G M G L Q D L Y G A G A L M T M M G A G G G L Y P C A E Y L R Q P Q C S P V A A P F Y A L R
B CHFI        :   - - - - - - - - - - - - - - - - - - - - - S A G T S C V P G W A I P H N P L P S - - - C R W Y V T S
                                                            + * * - + - + * - - + + - + + * + - *       - + + * + + +

       51         .         +         .         +         .         +         .         +         .         +   100
A Beta Kafirin:   E Q T M W Q P N F I C Q P L R Q Q C C Q Q M R M M D M Q S R C Q A M C G V V Q S V V Q Q L Q M T M Q
B CHFI        :   R T C G I G P R L P W P E M K R R C C R E L A D I P A Y C R C T A L - - - - - - - - - - - - - - - -
                  + + - - + - * + + - - + + + + + + * * + + + - - + + + - - * * + * +

      101         .         +         .         +         .         +         .         +         .         +   150
A Beta Kafirin:   L Q G V A A A S S L L Y Q P A L V Q Q W Q Q L L P A A Q A L T P - - - - - - - - - - L A M A V A Q V
B CHFI        :   - - - - - - - - S I L M D G A I P P G P D A Q L E G R L E D L P G C P R E V Q R G F A A T L V T E A
                                  * + * - + + * + - + - - + + + * + + - + + - + *                     - * + - * + + +

       51         .         +         .         +         .         +         .         +         .         +   182
A Beta Kafirin:   A Q N M P A M C G L Y Q L P S Y C T T P C A T S A A I P P Y Y Y
B        CHFI:    E C N L A T I S G V A E C P W I L G G G T M P S - - - - - - - -
                  + - * + + + + - * + - + - * + + - - - + - + + *
```

Figure 16 SEQALIGN pairwise alignment between β-kafirin and CHFI. This 3-gap alignment was used with the CHFI three-dimensional structure PDB:1BEA, to produce homology models for β-kafirin using I-TASSER.

The homology model for β-kafirin is shown superimposed with CHFI in Figure 17 and predicts a 4-helix bundle held by intramolecular disulfide bonds. Outside the 4-helix bundle β-kafirin is predicted to contain unordered structure. The region within the 4-helix bundle shows a close similarity to the same structural region in CHFI.

The cysteine residues C30, C68, C69, C81 and C148 are strictly conserved in the sequence alignment and together with C157 form 3 disulfide bonds in the model. C85 and C161 are conserved in three-dimensional space, which are predicted to form a 4th disulfide bond. The final disulfide pair C39 and C61 are not shown with perfect disulfide alignment in the model, though they are in close proximity and may form the 5th disulfide bond in β-kafirin. The model view showing inferred disulfide bonds can be seen in Figure 18.

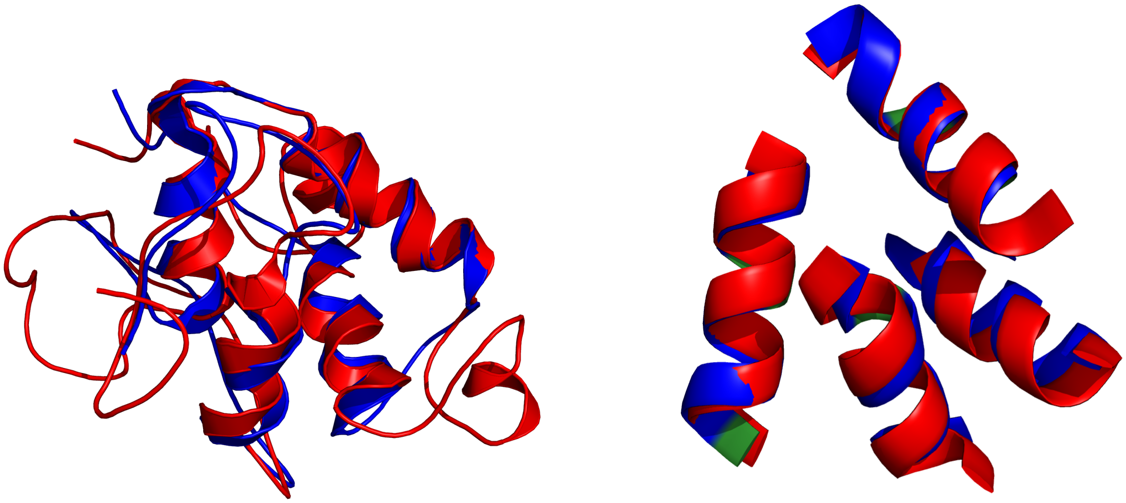


Figure 17 Homology model for β-kafirin superimposed over CHFI. β-Kafirin homology model (red) was created using I-TASSER using sequence alignments and the solved three-dimensional structure for CHFI (blue).

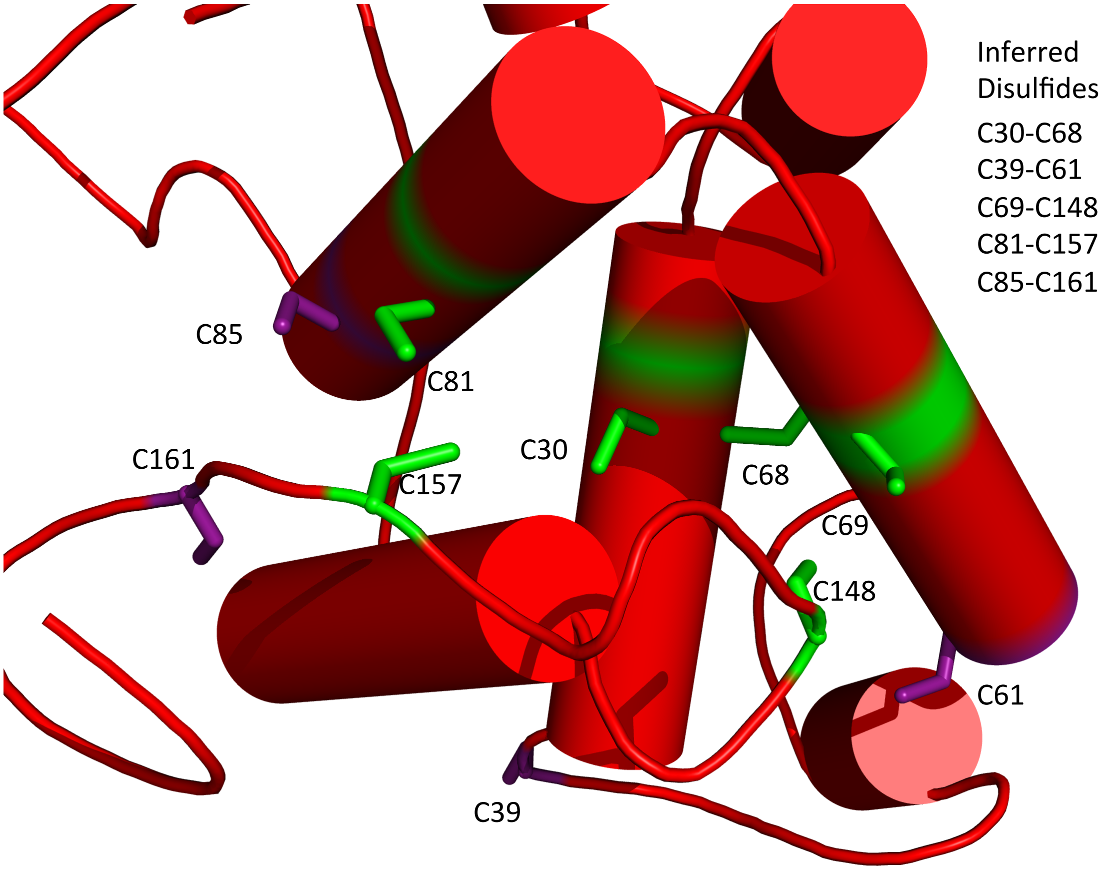


Figure 18 β-kafirin homology model based CHFI from maize. Cysteine residues C30, C68, C69, C81, C148, are strictly conserved in the sequence alignment and with C157 for 3 disulfide bond pairs (green). Cysteine residues C85 and C161 are not conserved in the sequence alignment but share close proximity in space and are predicted to form a disulfide pair. The 5th disulfide bond pair C39 and C61 is shown apart in the model, but may be allowed to form a disulfide bond with further model refinement.

Discussion

My research shows β-kafirin is a regularly folded protein in 47.5% ethanol with predominantly α-helix secondary structure. More than 50% of β-kafirin in mature sorghum is extractable in 47.5% ethanol without reducing agents. But total β-kafirin extraction from sorghum flour is only achieved with the use of reducing agents.

I found no evidence for disulfide-linked complexes containing β-kafirin in extracts without reducing agents regardless of solvent conditions. El Nour et al. (1998) reported a similar result, which supports my findings. This is a paradoxical situation, since 100% extraction of β-kafirin was unachievable without reducing agents regardless of solvent conditions. The inability to extract β-kafirin without reducing agents does not confirm β-kafirin is part of a large disulfide cross-linked network, and further investigation is needed to determine, to what, if anything β-kafirin may be linked.

Mazhar & Chandrashekar (1993) studied the extractability of sorghum kafirins throughout seed development. They found kafirins were extractable without reducing agents up to 40 days after half bloom, a point when kafirin content equals the amount at maturity. They found after this point and throughout desiccation to maturity, decreasing amounts of kafirin were extractable without reducing agents.

Oriaet et al. (1995b) linked this phenomenon to decreased digestibility of cooked sorghum across the same development span. It seems that a mechanism of seed dormancy may be triggered at 35-40 days after half-bloom that continues to maturity. Oria et al. (1995b) linked their findings of poor digestibility to moisture content rather than developmental days. Mazhar & Chandrashekar (1993) could have tested this by collecting seed at early stages and allowing it to air dry slowly at room temperature rather than freeze-drying, which are not equivalent drying processes. It is unclear whether the apparent cross-linking occurs simply from desiccation or rather by a directed mechanism of seed dormancy.

Weaver et al. (1998) discovered a mutant sorghum cultivar with highly digestible kafirin proteins and Oria et al. (2000) suggested the higher digestibility was due to deformed protein body structure. Winn et al. (2009) mapped the high digestible trait to chromosome 1, which points to unknown components possibly related protein body superstructure formation. In my studies of non-reduced extracts I found proteins released after reduction with DTT that were not consistent with known kafirins. It was outside the scope of this research, however studies to identify previously unknown components of the protein body may provide new insight into the phenomenon of poor digestibility.

After purifying β-kafirin I conducted experiments showing no free sulfhydryls in the extractable β-kafirin. And digestion with chymotrypsin indicated 5 intramolecular disulfide bonds. I observed evidence of a folded protein with CD, and both HPLC and SDS-PAGE provided further support for internal disulfide bonds. Disulfide bond mapping is a complex mathematical exercise when 10 cysteine residues are present.

The nature of disulfide bonds in β-kafirin is such that after cleavage by chymotrypsin the fragments are not released. The cross-linked bundle remaining after digestion contained more than 50% of the starting mass for β-kafirin. However since β-kafirin is only 6-8% of the total protein this could not account for the 50% reduction in digestibility observed by Kurien et al. (1960).

Weaver et al. (1998) hypothesized the poor digestibility of sorghum was due to disulfide cross-linking between β- and γ-kafirin at the surface of protein bodies. Yet when Hamaker et al. (1987) cooked sorghum flour with reducing agents digestibility only increased by 25%. It seems as though some other mechanism may be involved which may include more common plant defense mechanisms such as protease inhibitors or simply aggregation.

Boiling is thought to be a harsh condition, which normally denatures most proteins. However, certain proteins such as CHFI can withstand such harsh conditions and return to their natively folded conformation after boiling. Part of the strength in CHFI comes from the disulfide linked 4-helix bundle structure. This structure has been repeated in nature and my model of β-kafirin predicts a three-dimensional structure similar to CHFI even though their sequences are weakly similar.

β-kafirin has a stretch of hydrophobic residues absent in CHFI and the model places these residues at the surface of β-kafirin. This may be why β-kafirin is only extracted with organic solvents when CHFI is water-soluble. More research to study the insoluble aggregates after cooking, may provide information about the complex system of cooked flour, which contains thousands of unique components.

In this work my aim was to purify and characterize β-kafirin to determine its role in poor sorghum protein digestibility. The outcome of my work questions whether β-kafirin plays any role beyond the poorly digestible 4-helix bundle held by intramolecular disulfide bonds. Kumar et al. (2012), created an experimental sorghum cultivar with down-regulated γ-kafirin and observed no changes in protein body superstructure or improved digestibility. Their finding along with my own results suggest we rethink entirely the nature of poor sorghum protein digestibility and the mechanisms that lead to this phenomenon.

Future Work

My future work on sorghum kafirins will be to continue studying the disulfide-bonding pattern of β-kafirin. And to determine whether β-kafirin remaining after extraction without reducing agents is part of a larger disulfide linked network. If β-kafirin is discovered as part of a larger disulfide linked network then understanding the mechanism that leads to this disulfide exchange may be important for improving sorghum digestibility.

Using my extraction protocol for enriched β-kafirin I will scale the extraction to generate sufficient quantities for crystallizing the purified β-kafirin. Once crystallization is achieved I will solve the three-dimensional structure for β-kafirin using X-ray crystallography.

In my work I have observed several abundant proteins extracted along with kafirins, although having masses inconsistent with known kafirins . I will identify those proteins using MALDI-TOF MS in-source-decay, and then determine their role in the protein body superstructure.

To study kafirin localization I will produce antibodies for each kafirin and several other proteins. Different from previous researchers our kafirin extracts for antibody production will include individual β- and γ-kafirin and if possible α-kafirins. Each kafirin will be identified by sequencing along with SDS-PAGE band mobility.

Chapter 2 - NMR Structure of Glutenin Repeat Unit Peptide GQQPGQG from High Molecular Weight Glutenin 1Dx5 from Common Bread Wheat *Triticum aestivum*

Introduction

Flour from the endosperm of common bread wheat *Triticum aestivum* is unique among the cereal grains because when mixed with water it can form dough suitable for baking yeast-leavened products. The properties of dough from wheat flour have been described as viscoelastic, that is, elastic resistance to extension under force, and viscous flow at rest.

Viscoelasticity has been well studied and attributed to the prolamin storage proteins found in wheat endosperm. The prolamins of wheat and other cereal grasses are named as such due to their high proline and glutamine content (Shewry 1990). In wheat they are deposited in protein bodies (Sabelli & Larkins 2009) and are a major component of wheat gluten. The term gluten is a moniker for the insoluble mass of prolamins and non-prolamins, which form after mixing wheat flour into dough and may contain more than 100 unique polypeptides (Shewry 2009).

The term prolamin was coined by Osborne who studied the storage proteins of wheat and devised some of the earliest classifications of wheat proteins based on their solubility in water (albumins), saline (globulins) and alcohol solutions (prolamins) (Osborne 1909). The century of work following Osborne's classification has revealed the complexity of wheat endosperm proteins, and solubility alone is not sufficient for categorizing the various proteins (Gianibelli et al. 2001).

The gluten proteins are broadly classified as either glutenin or gliadin and these groupings have multiple sub-groups. The glutenin proteins exist in two basic forms (unrelated by sequence) known as high molecular weight glutenins and low molecular weight glutenins. The remaining prolamins are called gliadin and have at least 3 subclasses α/β, γ and ω-gliadins. The glutenin proteins form a disulfide-linked network in the endosperm, reaching molecular weights into the millions of daltons, while the gliadins are typically monomeric.

Genetic studies of developing and mature wheat seeds revealed as many as 6000 active genes based on mRNA transcripts, although two dimensional gel electrophoresis revealed only ~1700 proteins (Skylas et al. 2005). The authors attributed this discrepancy in observed proteins to limitations in the method of identifying proteins by two-dimensional electrophoresis. After reaching maturity the number of total seed proteins resolved using two-dimensional electrophoresis had dropped to ~1200 (Skylas et al. 2005). When Dupont et al. studied the proteome of wheat flour versus the entire seed they resolved only 476 proteins by two-dimensional gel electrophoresis of which 233 were identified with mass spectrometry (Dupont et al. 2011).

One challenge in studying wheat genomics and proteomics is the enormous size of the wheat genome (16,000 Mb) compared to other cereals such as rice (420Mb), maize (2500Mb), barley (4800Mb) (Skylas et al. 2005) and sorghum (730Mb) (Paterson et al. 2009). Common bread wheat is a hexaploid species with three largely redundant genomes AABBDD. The hybridization between tetraploid grass wild emmer *Triticum dicoccoides,* and the diploid species *Triticum tauschii* may have occurred as few as 10,000 years ago resulting in the species *Triticum aestivum,* which we know as common bread wheat (Shewry et al. 2003).

The endosperm proteins in wheat are of great interest because of their impact on end-use product quality (Payne 1987), and therefore decades of work have been devoted to understanding their role and interactions (Khan & Shewry 2009). Wheat endosperm contains ~80% of the total seed protein (Shewry 2009), and researchers are working to quantify the various components.

Glutenins can be divided into two groups based on molecular weight and are known as high molecular weight glutenins and low molecular weight glutenins. Together the high and low molecular weight glutenins make up ~50% of endosperm protein mass (Wieser 2007). The high molecular weight glutenins are particularly important as they have been closely linked to dough quality (Pirozi et al. 2008), despite making up only 10-12% of the endosperm protein mass (Wieser 2007).

Genes encoding the high molecular weight glutenin proteins are found on the long arm of chromosomes 1A, 1B and 1D (McIntosh et al. 2003). Two tightly linked genes are present at two loci coding for separate high molecular weight glutenin proteins termed x-type and y-type

(Payne 1987). In total, six unique high molecular weight glutenin genes are possible. However due to silencing of 1Ay in all cultivars and the silencing of 1Ax and 1Bx in some others, there may be only 3-5 high molecular weight glutenin genes expressed (McIntosh et al. 2003).

The low molecular weight glutenins and gliadins make up nearly 90% of the endosperm protein mass (Wieser 2007). The low molecular weight glutenins are found in disulfide-linked networks similar to those of the high molecular weight glutenins. As stated earlier, the low molecular weight glutenins, whose individual molecular weights range from ~30-40kDa have no sequence similarity to high molecular weight glutenins.

There may be as many as 20 separate (but highly similar) low molecular weight glutenin genes expressed in the wheat endosperm (Dupont et al. 2011). The genes have been mapped to the short arm of group 1A, 1B, and 1D chromosomes (Wrigley et al. 2006). The loci termed Glu-A3, Glu-B3 and Glu-D3 are tightly linked to the γ-gliadin and ω-gliadin loci also on the short arm of group 1 chromosomes 1A, 1B and 1D (Dupont et al. 2011). As many as 29 γ-gliadin genes (Qi et al. 2009) and 5 ω-gliadin genes (Anderson et al. 2009), have been identified. These encode proteins that make up ~30% and ~10% of the gluten protein mass respectively (Wieser 2007).

The α/β-gliadin genes are located on the short arm of group 6 chromosomes 6A, 6B and 6D, in complex loci containing up to 150 α/β-gliadin genes (Anderson et al. 1997). More recent studies (Gu et al. 2004) suggest at least 50% of the α/β-gliadin genes are silent but still only 22 mature proteins were uniquely identifiable using two dimensional gel-electrophoresis and MS/MS (Dupont et al. 2011). The α/β-gliadin proteins make up ~30% of gluten proteins in wheat (Wieser 2007).

Overall the gluten forming proteins in wheat are a complex group of proteins that have not been fully described. Undoubtedly there are components, which have yet to be discovered, and the role of each protein in relation to wheat quality has not been determined. A summary of the gluten protein components and estimates of abundance are given in Table 3.

Table 3 Summary of gluten forming proteins from wheat.

Gluten Fraction	*% of Gluten by weight*	*Gene #*	*Loci*
High Molecular Weight Glutenin	~10-12%	3-5	
x-type	4-9%	2-3	Glu-A1, Glu-B1, Glu-D1
γ-tpye	3-4%	1-2	Glu-A2, Glu-B2, Glu-D2
Low Molecular Weight Glutenin	~25%	>20	Glu-A3, Glu-B3, Glu-D3
Gliadin	~60-70%		
α/β-Gliadin	28-35%	150+	Gli-A2, Gli-B2, Gli-D2
γ-Gliadin	23-31%	29	Gli-A1, Gli-B1, Gli-D1
ω-Gliadin	3-6%	5	Gli-A3, Gli-B3, Gli-D3

Summarized from previous work Dupont et al. (2011); Pomeranz (1988); Wieser (2007).

In addition to multiple genes for each of the prolamins, there is allelic variation among cultivars, which has been shown to impact wheat quality. The most significant contribution to wheat quality has been attributed to allelic variation among the high molecular weight glutenin genes although multiple alleles for the low molecular weight glutenins and gliadins also exist (Pirozi et al. 2008).

In their study on the effect of allelic variation in bread making Payne et al. (1987) found alleles on the 1D chromosome correspond closely with baking quality. Payne et al. (1984) had already begun studying allelic variation at an earlier time and developed nomenclature to identify the various high molecular weight glutenins and their alleles. This convention assigns a number to each high molecular weight glutenin based on SDS band mobility, such as with HMW-GS 1Dx5 (Payne 1987; Payne et al. 1984). Multiple variants have since been discovered and a recent screening of US hard red winter wheat germplasm has revealed significant allelic variation for the HMW-GS (Shan et al. 2007). An example of the SDS-PAGE banding pattern for allelic variation is shown in Figure 19 (Gao et al. 2010).

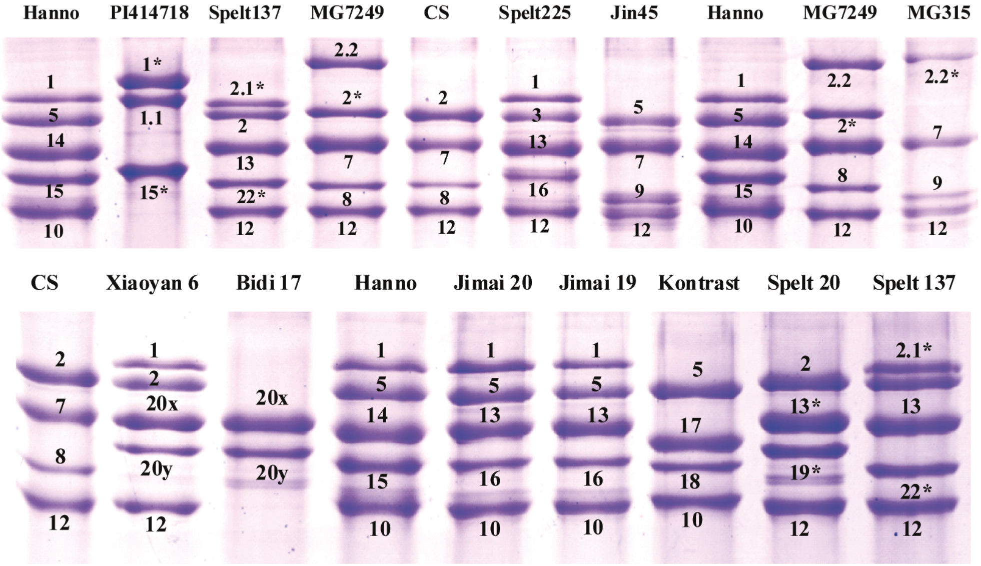


Figure 19 SDS-PAGE for high molecular weight glutenins from wheat and related species. SDS-PAGE banding patterns correspond to glutenin proteins from allelic variants for high molecular weight glutenin genes. The numbering system was devised to identify bands by mobility and distinguish differences between cultivars. Researchers have used this technique to identify allelic variation between wheat cultivars and also between related species (Gao et al. 2010).

Considering the large number of genes each with allelic variants, one can conceive a vast number of combinations with only partially known affects on finished product quality. With that being said, researchers like Payne *et al*. have observed measurable differences in baking quality from wheat cultivars with certain allelic combinations of the 1Dx high molecular weight glutenin gene.

Sequences are available for most of the known variants, although some have only been identified by SDS-PAGE. Comparison of the sequences for individual gene variants reveal only slight differences between variants of a each gene. A remarkable feature of glutenins and gliadins is the presence of long spans of repeating peptide sequences. The peptide-repeat units are neither perfect in sequence or in periodicity and some repeats apparently overlap. The most common repeat is the tripeptide repeat GQQ, which occurs over 61 times in the central repeat domain of HMW-GS 1Dx5. Many of these occurrences are within the hexapeptide repeat PGQGQQ (51 occurrences) to form a nonapeptide repeat GQQPGQGQQ (31 occurrences).

Two other frequent repeats are a di-tyrosine containing sequence GYYPTSPQQ (16 occurrences) and a serine containing sequence SGQGQQ (10 occurrences). Slight variations exist for each of these consensus repeats. It is unclear what effect these variations have on the protein secondary structure or function. The sequence for HMW-GS 1Dx5 is shown with annotated repeats in Figure 20. The protein secondary structure was predicted using the EMBOSS garnier tool (Rice et al. 2000) the N- and C-terminal regions are predicted to be helix, and the central domain is predicted to contain repeated turns, β-sheet and extended coil.

It is difficult to say what constitutes a repeating unit, since it is possible to build repeats having variable length with the largest single unit being approximately one half of the central domain repeated once. If the repeating sequences were assumed to produce repeating structural features, then the repeat unit would more accurately be defined by those structures. Unfortunately, isolating gluten proteins from wheat flour requires reducing agents and strong denaturants therefore the native structures remain unsolved.

Researchers working to develop a macro scale description of wheat flour properties found a strong correlation between the percent of un-extractable, polymeric protein and gluten strength (Gupta et al. 1993). Other evidence supports this finding, including a study of near isogenic lines differing only in the cysteine content of the high molecular weight glutenin protein. In this work Pirozi et al. (2008) found high cysteine variants yielded a higher percent of un-extractable polymeric protein and produced superior baking performance.

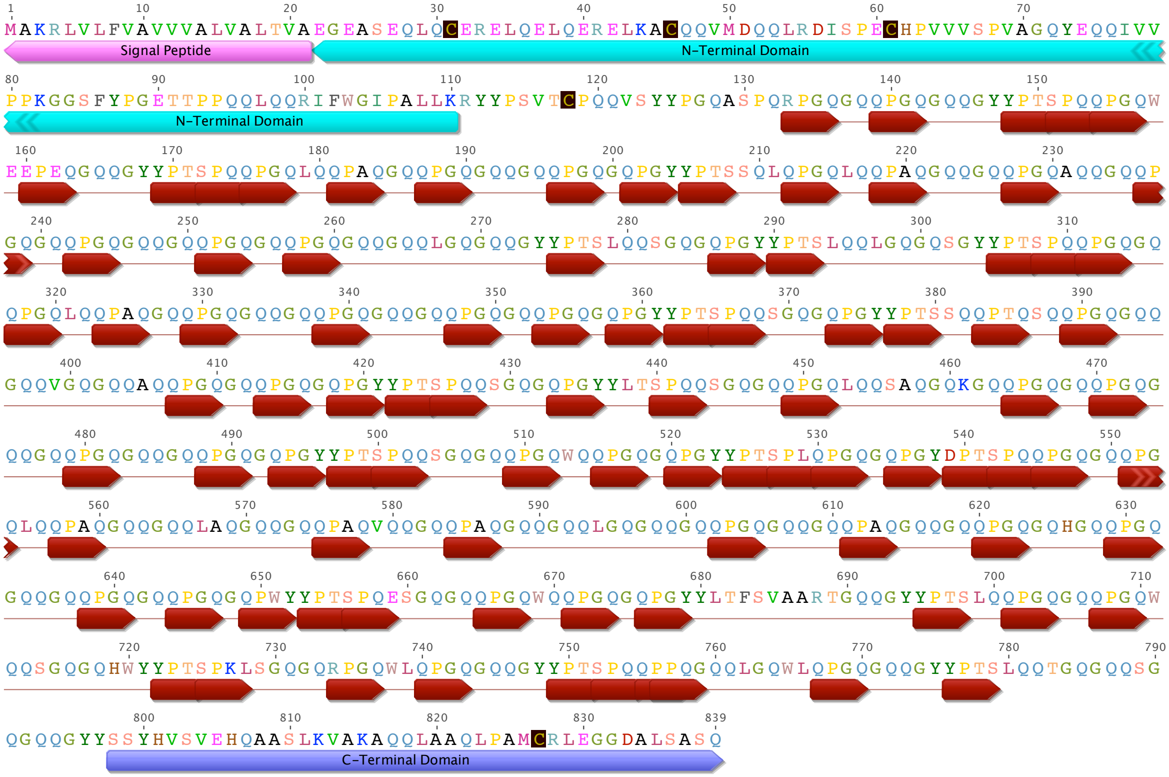

1 10 20 30 40 50 60 70
MAKRLVLFVAVVVALVALTVAEGEASEQLQCERELQELQERELKACQQVMDQQLRDISPECHPVVVSPVAGQYEQQIVV
Signal Peptide
N-Terminal Domain
80 90 100 110 120 130 140 150
PPKGGSFYPGETTPPQQLQQRIFWGIPALLKRYYPSVTCPQQVSYYPGQASPQRPGQGQQPGQGQQGYYPTSPQQPGQW
N-Terminal Domain
160 170 180 190 200 210 220 230
EEPEQGQQGYYPTSPQQPGQLQQPAQGQQPGQGQQGQQPGQGQPGYYPTSSQLQPGQLQQPAQGQQGQQPGQAQQGQQP
240 250 260 270 280 290 300 310
GQGQQPGQGQQGQQPGQGQQPGQGQQGQQLGQGQQGYYPTSLQQSGQGQPGYYPTSLQQLGQGQSGYYPTSPQQPGQGQ
320 330 340 350 360 370 380 390
QPGQLQQPAQGQQPGQGQQGQQPGQGQQGQQPGQGQQPGQGQPGYYPTSPQQSGQGQPGYYPTSSQQPTQSQQPGQGQQ
400 410 420 430 440 450 460 470
GQQVGQGQQAQQPGQGQQPGQGQPGYYPTSPQQSGQGQPGYYLTSPQQSGQGQQPGQLQQSAQGQKGQQPGQGQQPGQG
480 490 500 510 520 530 540 550
QQGQQPGQGQQGQQPGQGQPGYYPTSPQQSGQGQQPGQWQQPGQGQPGYYPTSPLQPGQGQPGYDPTSPQQPGQGQQPG
560 570 580 590 600 610 620 630
QLQQPAQGQQGQQLAQGQQGQQPAQVQQGQQPAQGQQGQQLGQGQQGQQPGQGQQGQQPAQGQQGQQPGQGQHGQQPGQ
640 650 660 670 680 690 700 710
GQQGQQPGQGQQPGQGQPWYYPTSPQESGQGQQPGQWQQPGQGQPGYYLTFSVAARTGQQGYYPTSLQQPGQGQQPGQW
720 730 740 750 760 770 780 790
QQSGQGQHWYYPTSPKLSGQGQRPGQWLQPGQGQQGYYPTSPQQPPQGQQLGQWLQPGQGQQGYYPTSLQQTGQGQQSG
800 810 820 830 839
QGQQGYYSSYHVSVEHQAASLKVAKAQQLAAQLPAMCRLEGGDALSASQ
C-Terminal Domain

Figure 20 Amino Acid Sequence for HMW-GS 1Dx5. Amino acid sequence for high molecular weight glutenin 1Dx5. The signal peptide (residues 1-21) is followed by the predominantly helix N-terminal domain (22-110). The large central domain (111-687) is comprised of repeating peptide units with the consensus sequences XPXX (red). The repeat domain is predicted to be repeated turns, β-sheet and extended coil. The C-terminal domain makes up the remaining residues (798-839) and is predicted to be helix in structure using methods from EMBOSS Rice et al. (2000)

These findings point back to the original work by Payne *et al*. (1987) who studied the effect of allelic variation on bread baking. Their work revealed superior cultivars contained the 1Dx5 + 1Dy10 pair and inferior cultivars contained 1Dx2 + 1Dy12 subunit pairs (Payne 1987). When comparing the amino acid sequences in Figure 21, only slight differences are observed although the effects of these variations on protein structure and function are unknown. The extra cysteine residue in 1Dx5 compared to 1Dx2 has been suggested as a possible cause for improved dough functionality in cultivars carrying the 1Dx5 gene (Alberti et al. 2002; Bonomi et al. 2013; Li et al. 2006).

.

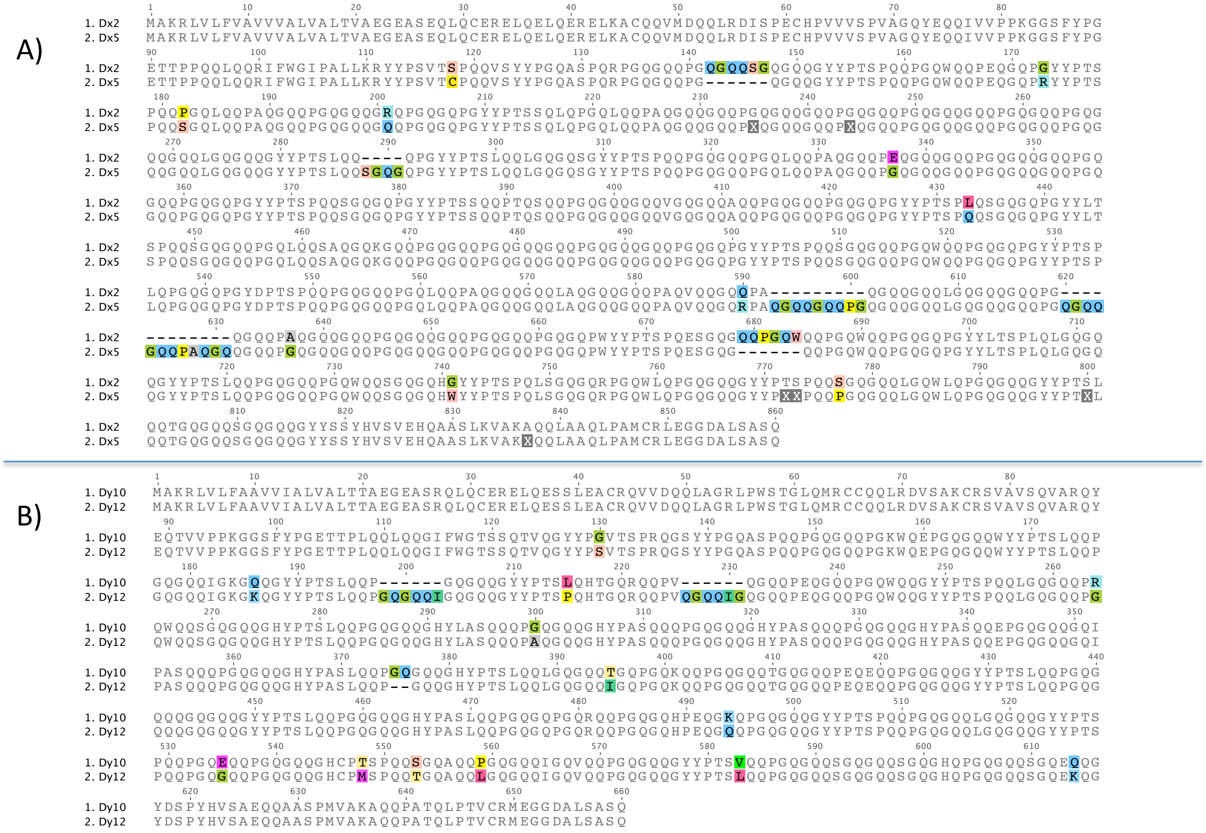


Figure 21 Amino acid sequence alignment of high molecular weight glutenin variants 1Dx5 with 1Dx2 and corresponding y-type variants 1Dy10 with 1Dy12. The high molecular weight glutenin allele pair 1Dx5+1Dy10 has been shown to produce superior baking performance compared to1Dx2+1Dy12 Payne (1987); Tatham et al. (2000). Sequence alignments for A) 1Dx5 (superior) and 1Dx2 (inferior) showing sequence variations (highlighted). B) Sequence alignment for the corresponding y-type pair 1Dy10 (superior) and 1Dy12 (inferior).

Pirozi et al. (2008) have shown disulfide cross-linking plays a significant role in gluten elasticity, however cross-linking alone cannot be the basis of viscoelasticity since other cereals such as sorghum have cross-linked storage proteins and lacks extensibility. Therefore, other sequence features, presumably including the repeat sequences, must play a role in the elastic behavior of wheat gluten.

Tatham and Shewry (1984) proposed molecular models for gluten elasticity including models relating gluten elasticity that of elastin (Tatham & Shewry 1984; Tatham et al. 1990). These models are based on secondary structure predictions indicating repeating β-turns in the central domain of glutenin. Tatham et al. (1985) hypothesized that repeating β-turns should translate to a “β-spiral” structure which is responsible for gluten elasticity (Tatham et al. 1990;

1985). Their hypothesis was based on experimental evidence collected using glutenin from durum wheat cultivar Bidi 17. In this study high molecular weight glutenins were extracted using 0.5% SDS plus 1% β-mercaptoethanol. After precipitating the proteins in 70% ethanol, the pellet was re-suspended in 8M urea plus 1% β-mercaptoethanol then alkylated using 4-vinylpyridine. Hydrodynamic studies of the purified glutenin indicated a rod shaped structure with approximate dimensions 18Å × 500Å and circular dichroism indicated β-turn in 50%(v/v) propan-1-ol (Field et al. 1987).

The findings of Tatham et al. (1985) were extended by the same group using scanning tunneling microscope images. Which detailed fibrillar structure with 19.5Å diameter, containing a repeating structure with a pitch of 14.9Å (Miles et al. 1991). Matsushima et al. (1992) studied high molecular weight glutenin using small angle x-ray scattering and also calculated a rod-like structure of somewhat larger dimensions of 64Å × 690Å. McIntire et al. (2005) attempted to determine the structure of high molecular weight glutenin using atomic force microscopy and their results suggested a rod-like shape resulting from poly-proline II helix having 7Å × 3000Å dimensions.

Each of these studies and others (Alberti et al. 2002; Li et al. 2006; Mackintosh et al. 2008) all used extraction protocols similar to that published by Tatham et al. (2000) which use strong denaturants, reducing agents, alkylation followed by precipitation and solubilizing in 6-8M urea. Tertiary structure is clearly destroyed using these techniques and therefore it is unclear whether such studies are relevant to the native structure of the protein.

Blanch et al. (2003) obtained a soluble 30kDa fragment of intact high molecular weight glutenin 1Dx5 (residues 147-440) using trypsin. However, they purified the fragment with RP-HPLC using a C18 column, which is known to disrupt hydrophobic interactions and unfold proteins during elution. This reductive approach is similar in principle to studies of synthetic peptides derived from the repeating sequences of glutenins and gliadins.

Using synthetic peptides corresponding to glutenin repeat units, Tatham et al. (1990) collected CD and FTIR spectra, which were consistent with β-turn and β-sheet secondary structure. In this work the researchers used peptides containing of sequences GQQPGQG, GQPGYYPTSP and GQQGYYPTSP. Each of these peptides contained unblocked N- and C-termini and was studied by CD and FTIR in 100% water, 100% TFE and 50:50 (v/v) TFE in water. They found that GQQPGQG was structured in 100%TFE, but produced CD spectra similar to random coil in solutions below 50% TFE (Tatham et al. 1990). Using software the β-turn was predicted to be a type II β-turn across QPGQ in peptide 1. Several turn types were predicted for the other peptides. This work led to later studies of cyclic peptides containing the same sequences (van Dijk et al. 1997b).

van Dijk et al. (1997b) synthesized cyclic peptides of the sequences cyclo-[PGQGQQPGQGQQ], cyclo-[GYYPTSPQQGA], and cyclo-[PGQGQQGYYPTSPQQ]. In addition to these they also studied unblocked linear peptides $(PGQGQQ)_{n=1,3,5}$. In this work all peptides were analyzed in 20mM NaPi pH 6.0, or in unbuffered D_2O. The CD and FTIR spectra for cyclic peptide 1 indicated a β-turn, however the NMR data were ambiguous and only sequential assignments were made (van Dijk et al. 1997b).

The single linear unblocked peptide PGQGQQ produced CD spectra in 20mM NaPi indicative of random coil, however after extending the length to $(PGQGQQ)_{n=3,5}$ the spectra showed some β-turn character similar to the cyclic version in the same solvent. The NMR spectra of the linear peptides did not provide any useful structural information (van Dijk et al. 1997b).

Van Dijk et al. (1997b) also collected FTIR spectra of the linear peptides, which they interpreted to indicate β-turn in D_2O, whereas Tatham et al. (1990) concluded a random coil structure in water for identical linear peptides. In contrast to Tatham et al., van Dijk *et al.* (1997b) did not study the peptides in TFE, which was shown to produce β-turn in the earlier studies. Had they used TFE in their NMR studies van Dijk et al. (1997a) may have discovered important structural information possibly leading to different conclusions regarding the peptide structure in water.

In the following pages I will describe my work on the glutenin repeat peptide unit GQQPGQG in the blocked form Ac-GQQPGQG-Am later referred to as GQQPGQG. Unlike previous studies of the glutenin repeat peptides, this work resulted in a three-dimensional structure of the peptide in 100% TFE. Using this structure as a starting point for molecular dynamics simulations, I observed inter-conversion between folded and unfolded structures, plus flexibility in the glutamine-3 side-chain positioning within the β-turn structure. My research identifies a type II β-turn across the residues QPGQ, which is stabilized by a hydrogen bond between Q6 amide proton, and Q3 carbonyl oxygen. I found remarkable similarity between GQQPGQG compared to proteins with known three-dimensional structure containing the exact or very similar sequence. In nearly all examples, the peptide forms a β-turn at the protein surface.

Several studies have shown a significant contribution from β-sheet to the structure of the central repetitive domain in glutenin Feeney et al. (2003); van Dijk et al. (1997a). Despite these findings, the widely accepted model for high molecular weight glutenin is a regularly repeating β-spiral composed of the various peptide repeat units.

Materials and Methods

Synthetic Heptapeptide

Two synthetic heptapeptides purchased from Peptide 2.0 (Peptide 2.0, Chantilly, VA) peptide 1, GQQPGQG was purchased with unblocked termini (i.e. charged end-groups) and peptide 2 was purchased with N-terminal acetyl capping (CH_3CO) and C-terminal amide capping (NH_2). Peptide 2 was ^{15}N labeled at glycine-5 to help identify ambiguous glycine resonances in the NMR spectra.

NMR Spectroscopy

NMR[xiii] spectra were recorded at 5°C and 25°C using a Varian System 500MHz NMR spectrometer equipped with a 5mm triple resonance cryogenic probe and z-axis pulse field gradient. I used solvent conditions matching previously published work where repeat peptides were studied in 100% TFE[xiv] and 100% H_2O Tatham et al. (1989). In each experiment lyophilized peptide was dissolved in 0.5mL of solvent to a final concentration of 3.0mM. To help suppress NMR signal from the solvent I used deuterated TFE for structures solved in TFE, and 90% H_2O, 10% D_2O (v/v) for NMR experiments in water.

I repeated a series of two-dimensional ^1H-^1H NMR experiments namely TOCSY[xv], NOESY[xvi] and COSY[xvii] at each solvent condition and temperature setting. The two-dimensional TOCSY spectra are important when identifying proton resonances connected through up to five bonds. To help uncover overlapping resonances I used two-dimensional ^1H-^1H COSY spectra, which reveal proton resonances separated by a maximum of three bonds Roberts & Lian (2011). The final two-dimensional ^1H-^1H experiment called NOESY, is used to determine through space proximity between proton resonances. Since through-space signal intensity depends on a short mixing time after the magnetic pulse, we performed three separate NOESY experiments at 200ms, 300ms and 500ms mixing times.

After studying the two-dimensional ^1H-^1H spectra it became clear our peptide was present with at least two conformations in 100% TFE. This created ambiguous resonance assignment for glycine residues 1, 5 and 7. Therefore we repeated the experiments above on peptide-2, which was isotopically labeled with ^{15}N in glycine-5.

In addition to TOCSY, COSY and NOESY spectra at 5°C and 25°C in both TFE and water; we also collected ^{15}N-^{1}H HSQC[xviii] spectra and the 1D ^{1}H-^{1}H spectrum. HSQC reveals through-bond correlation between ^{15}N and ^{1}H resonances. Labeling the NH group of glycine-5 results in a signal intensity of ^{15}N much greater than the signal from ^{15}N in the other glycine residues. This allowed for unambiguous identification of each glycine in the peptide.

Resonance Assignment

The NMR data were transformed from varian to *.ucsf* format using NMRPipe (Delaglio et al. 1995) and resonances were assigned using SPARKY v3.115 (Goddard & Kneller 2008).

Structure Prediction and Validation

I used the Crystallography and NMR System package CNSsolve *v1.3* (Brünger et al. 1998) to solve the three-dimensional structure for the end-capped peptide Ac-GQQPGQG-Am in 100%TFE at 5°C. The inter-proton distances were estimated from NOE resonance intensities ranked as Strong (1.8-2.5 Å) Medium (1.8-3.5 Å) and Weak (1.8-5.0 Å). 24 candidate structures were chosen from 30 low energy structures generated through simulated annealing and minimization. The quality of each low-energy structure was assessed using AQUA and PROCHECK-NMR (Laskowski et al. 1996). All three-dimensional structures were visualized using PyMol *v1.2* (DeLano Scientific LLC, Portland OR).

CD Spectroscopy

I used a JASCO J-815 series CD[xix] spectrometer to study the secondary structure of the end-capped peptide in varying TFE-H_2O solutions ranging from 100% TFE-0% H_2O (v/v) to 0%TFE-100%H2O while maintaining constant peptide concentration. I collected five consecutive CD spectra for each sample and solvent blank using a quartz cuvette with 3.0 mm path-length and total volume of 300µL. The CD spectra were accumulated at a rate of 50nm/min with absorbance measured from 260-190 nm in 1nm increments. The spectra were analyzed using Jasco Spectra Manager software.

Results

CD Experiments

I used CD spectroscopy to study the conformational effects of TFE on GQQPGQG by collecting spectra in varying concentrations of TFE in water from 0%-100% TFE. The spectra shown in Figure 22 indicate β-turn character in TFE concentrations greater than 50%. In concentrations below 50% TFE the spectra are more representative of random coil secondary structure. The strongest signal indicating β-turn was observed in 100% TFE, which led to our study of GQQPGQG in 100% TFE using NMR.

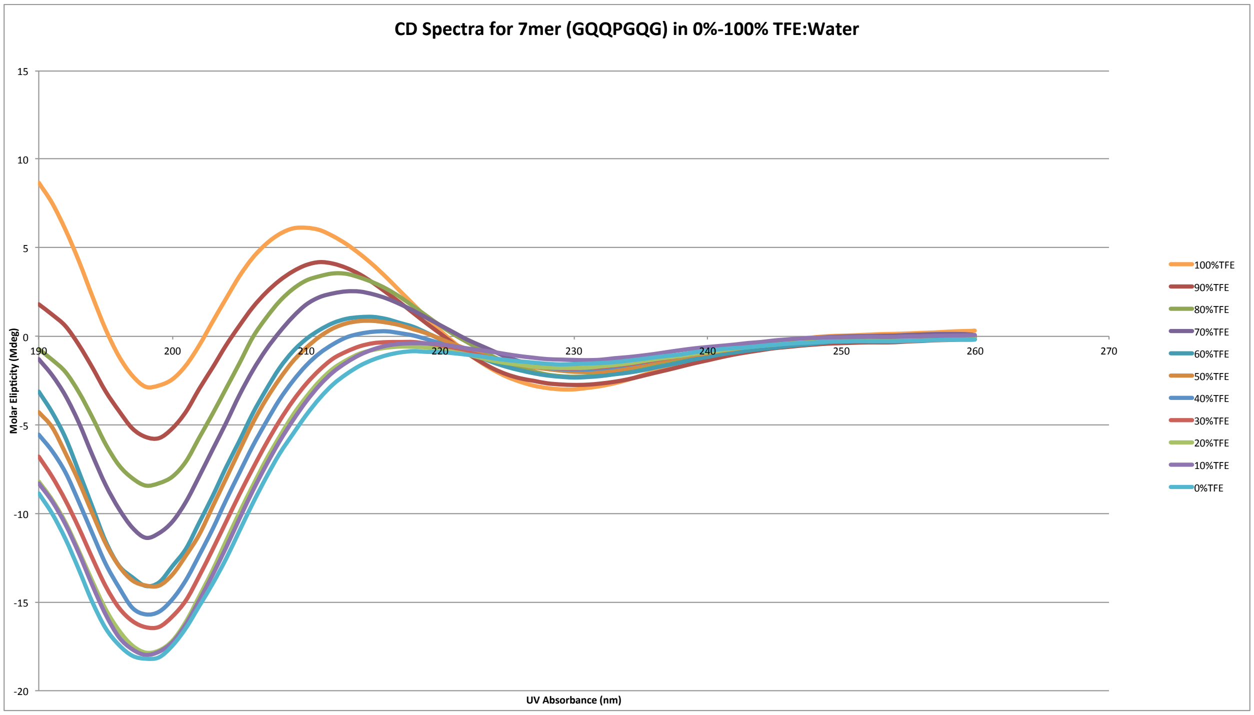


Figure 22 CD spectra showing secondary structure changes for CH_3CO-GQQPGQG-NH_2. CD spectra for CH_3CO-GQQPGQG-NH_2 in varying concentrations of TFE in water ranging from 100% TFE to 100% water. The spectra indicate turn character above 50% TFE and random coil below 50% TFE.

NMR Experiments

The three-dimensional structure of GQQPGQG was determined to be predominantly Type II β-turn in 100% TFE at 5°C using ^{1}H-NMR. Solving the three dimensional structure began by identifying proton resonances for each amino acid residue in the peptide. Figure 23 shows the TOCSY spectra indicating through-bond connectivity between H_N and $H_{C\alpha}$ protons also known as the fingerprint region. To unambiguously assign the glycine-5 resonances, the peptide was isotopically labeled with ^{15}N at the glycine-5 backbone nitrogen. The effect of labeling with ^{15}N resulted in splitting of the H_N proton resonances and a double set of resonances for glycine-5 in the fingerprint region.

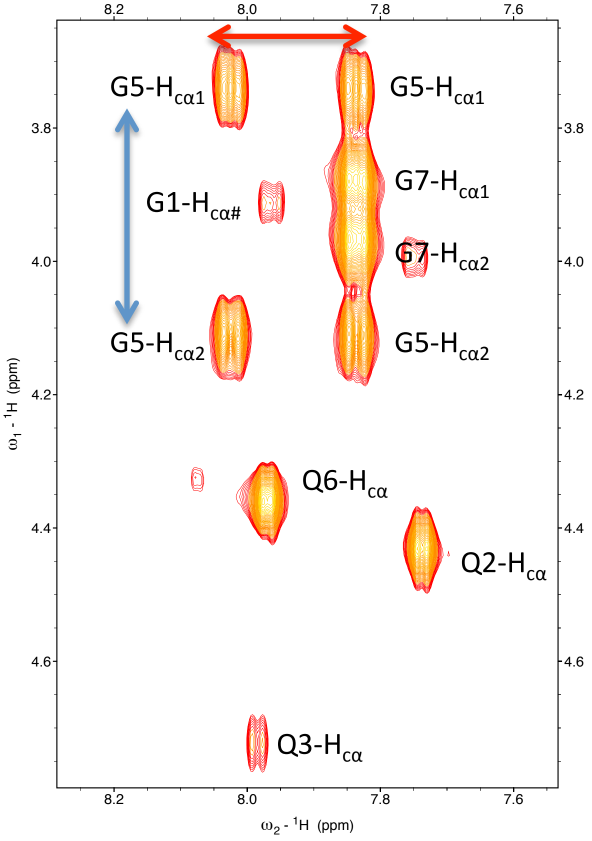


Figure 23 1H-NMR TOCSY spectra showing the fingerprint region for GQQPGQG. The fingerprint region shows through-bond connectivity between H_N and $H_{C\alpha}$ resonances. Glycine-5 was uniquely identified by a double set of resonances resulting from ^{15}N labeling of the backbone nitrogen (shown by red arrow). The large separation (shown by blue arrow) between $H_{C\alpha 1}$ and $H_{C\alpha 2}$ protons for Glycine-5 is indication this residue is part of a stable folded structure. The single proton resonance at Glycine-1 indicates flexibility at this residue.

The TOCSY spectrum generally allows identification of amino acids, although not their position in the peptide sequence. I used a NOESY$_{500}$ spectrum along with the TOCSY spectrum to make sequence specific identification. The NOESY spectrum shown in Figure 24 provides information regarding through-space connectivity, allowing sequential assignment in the fingerprint region. Since proline does not have an amide proton it is not visible in the fingerprint region and results in a break of the sequential resonance assignment. Proline is identifiable by strong NOE cross-peaks from glycine-5 H_N to proline-4 $H_{C\alpha}$. Glutamine-3 is unique in the sequence being the only resonance showing glutamine-to-glutamine sequential NOE cross-peaks.

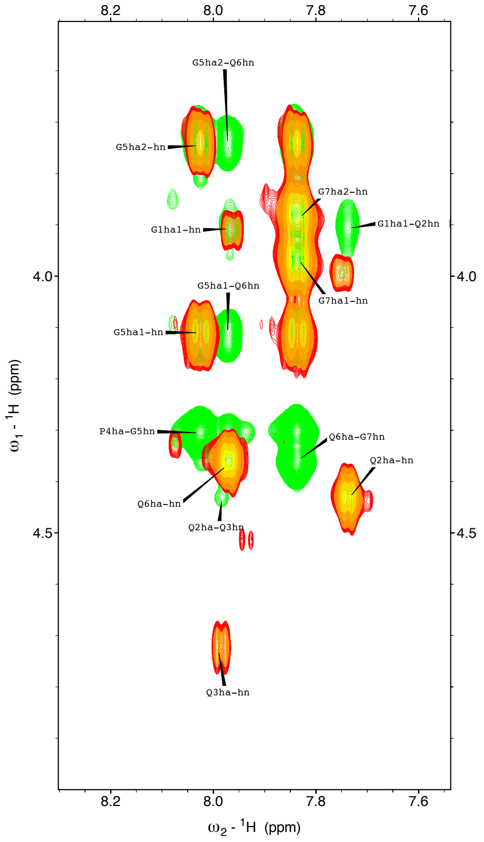


Figure 24 NOESY and TOCSY spectra in the fingerprint region of GQQPGQG. ^{1}H-NMR TOCSY spectra (red-yellow) superimposed onto the ^{1}H-NMR NOESY500 spectra (green). Together these spectra indicate connectivity between the backbone H_N proton and $H_{C\alpha}$ proton within a residue (TOCSY) and connectivity between the H_N proton and $H_{C\alpha}$ proton in the preceding residue (NOESY). From these spectra sequence specific assignment is possible.

Uniquely identifying backbone H_N and $H_{C\alpha}$ resonances in the previous step is required for identifying the side-chain resonances $H_{C\beta}$, $_{C\gamma}$ and $H_{C\delta}$. The side-chain protons are observed in the CH_2 region of the TOCSY spectrum shown in Figure 25. The glutamine and proline side-chains are the only residues in my peptide with resonance in this region. Glutamine-3 and Glutamine-2 both have clear separation of $H_{\beta 1}$ and $H_{\beta 2}$ protons although display overlapping $H_{\gamma 1/2}$ protons. Glutamine-6 is less resolved with slight overlapping between $H_{\beta 1/2}$ and $H_{\gamma 1/2}$ protons. Proline is unique in this peptide having a ring structure that results in connectivities through-bond to both $H_{C\alpha}$ and $H_{\delta 1/2}$ protons.

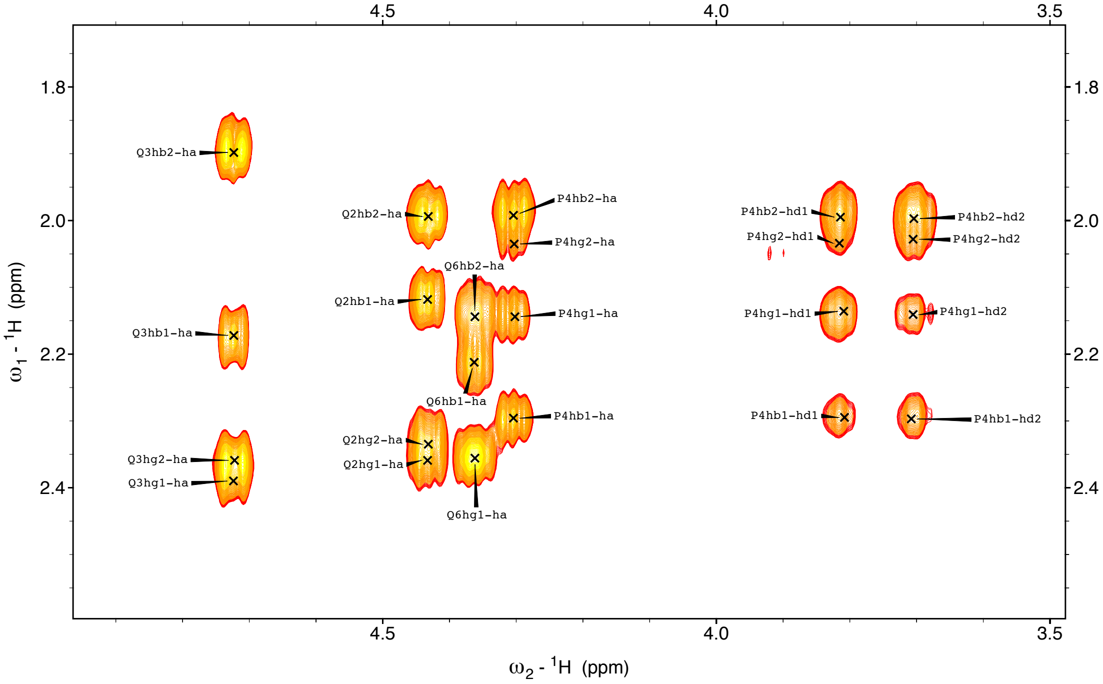


Figure 25 TOCSY spectrum showing the side-chain resonances for GQQPGQG. This region of the TOCSY spectrum indicates resonances connected through-bond to the Cα proton. The TOCSY NMR experiment can detect proton connectivity through up to five bonds, although not through carbonyl (C=O) or amide (N-H) bonds.

I used a COSY spectrum along with the TOCSY spectrum to confirm the side-chain assignments between Hβ and Hγ protons. The COSY NMR experiment shown in Figure 26 is similar to the TOCSY experiment, however the observed connectivity only extends through three bonds.

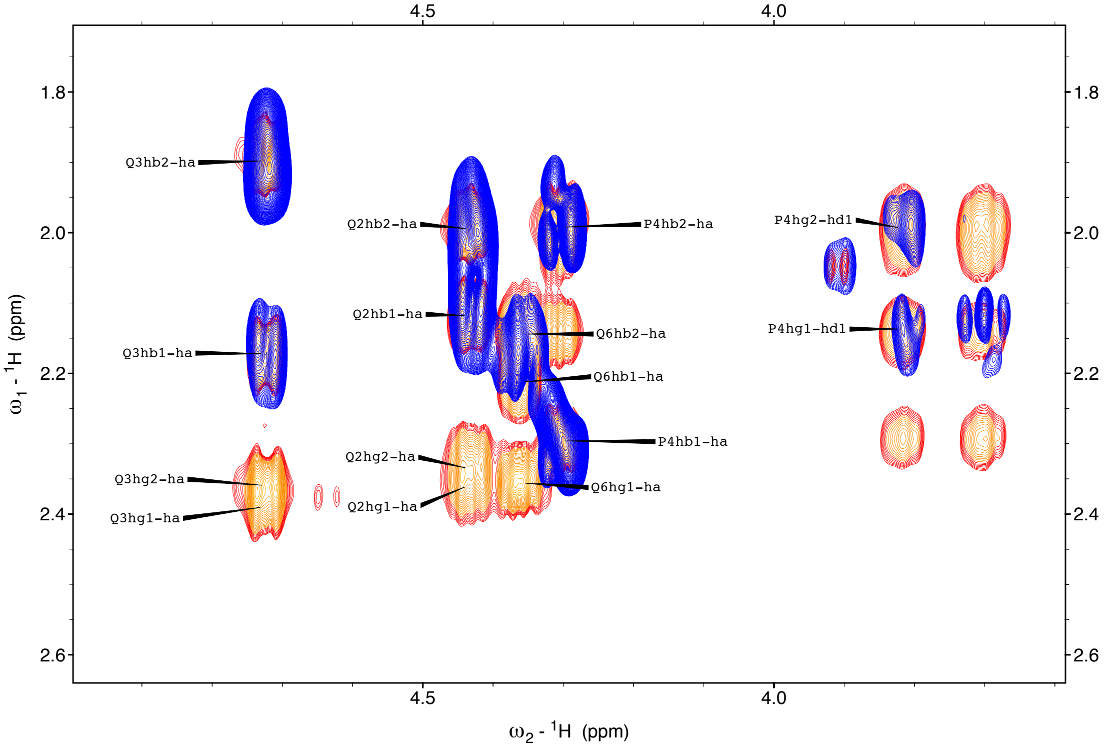


Figure 26 COSY and TOCSY spectra shown for the $H_{C\alpha}$ – H_{CH2} region. The COSY spectra shown in blue indicate resonances connected through a maximum of 3 bonds. In the $H_{C\alpha}$ - $H_{C\beta/\gamma}$ region the COSY spectra distinguishes H_β resonances from more distant H_γ resonances, which are 4 bond lengths from the $H_{C\alpha}$ proton and do not appear in the COSY spectrum.

Because the COSY experiment is limited to three bonds it can only reveal connections between $H_{C\alpha}$ - $H_{C\beta}$ protons in this region of the spectrum. To distinguish between $H_{\beta1}$ and $H_{\beta2}$ protons, I used tables of standard amino acid chemical shifts. It was unnecessary to use more complicated COSY experiments, which can be used to measure the J-coupling constants and determine the exact proton orientation.

After identifying each backbone and side-chain resonance I used the NOESY spectrum to identify through space connectivity between these resonances. Figure 27 shows the overlaid TOCSY and NOESY spectra for the $H_{C\alpha}$ - $H_{C\beta/\gamma}$ region. This region provides important structural information for our peptide because the glycine residues (Gly-1, Gly-5 and Gly-7) only contain $H_{C\alpha}$ resonances in their side-chains. Proline-4 also provides useful information due to the cyclic side-chain, which is constrained relative to the backbone movement. The side-chain of Q3 shows strong NOEs to P4-$H_{\delta1/2}$, which indicates the close proximity and stable conformation of

Q3 relative to P4. This is also true for Glycine-5 $H_{C\alpha, \text{which}}$ is stabilized relative to Proline-4 by the β-turn structure.

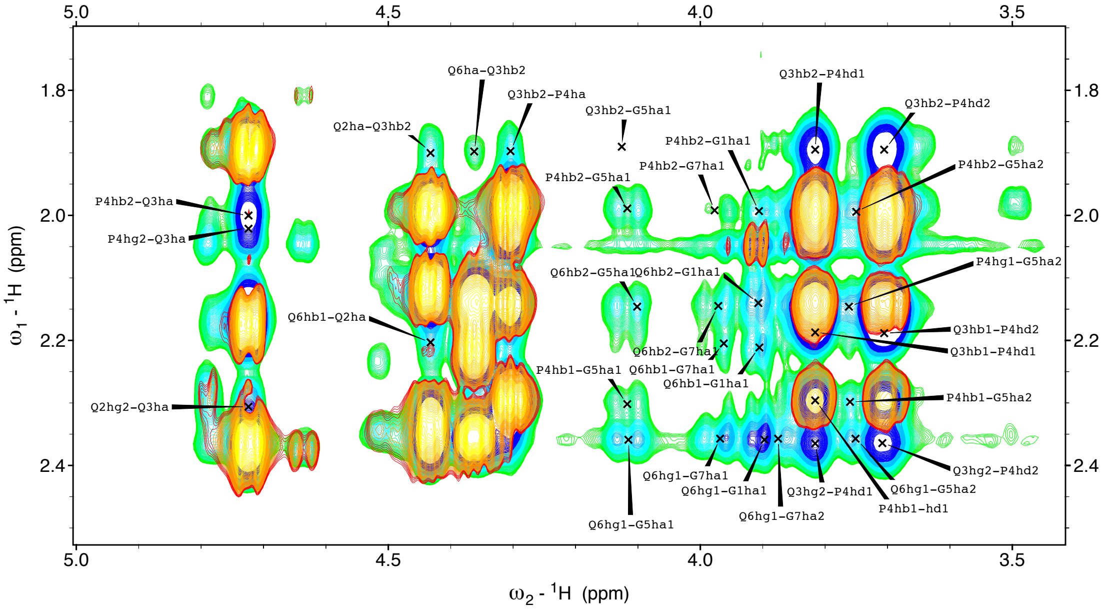


Figure 27 NOESY and TOCSY spectrum showing the $H_{C\alpha}$-$H_{C\beta,C\gamma}$ region for GQQPGQG. The NOE cross-peaks shown in these spectra indicate side-chain orientation in reference to the backbone. The resonances observed here are from the side-chain protons of glutamine and proline (TOCSY spectrum in yellow). The structural information is revealed by their through-space proximity (NOE in green/blue).

The $H_{C\alpha}$ - $H_{C\alpha}$ region of the TOCSY and NOESY spectra can be seen in Figure 28 and Figure 29. Proline is one amino acid, which can adopt either cis- or trans-proline conformation and together these spectra provide information about the conformation of proline in my peptide. I observed characteristic resonances, which indicate trans-proline, such as strong NOE cross-peaks between Q3-$H_{C\alpha}$ and P4-$H_{C\delta1,C\delta2}$. If Proline-4 were in the cis conformation I would have expected to see a strong NOE cross-peak between Q3-$H_{C\alpha}$ and P4-$H_{C\alpha}$. There is only a weak NOE cross-peak observed in this region and no other NOEs were observed that indicated a cis-proline conformation. The NOE cross-peaks observed in the spectra already mentioned along with other regions of the spectra shown in appendix Figure A. 2 and Figure A. 3 were used to solve the three-dimensional structure for GQQPGQG.

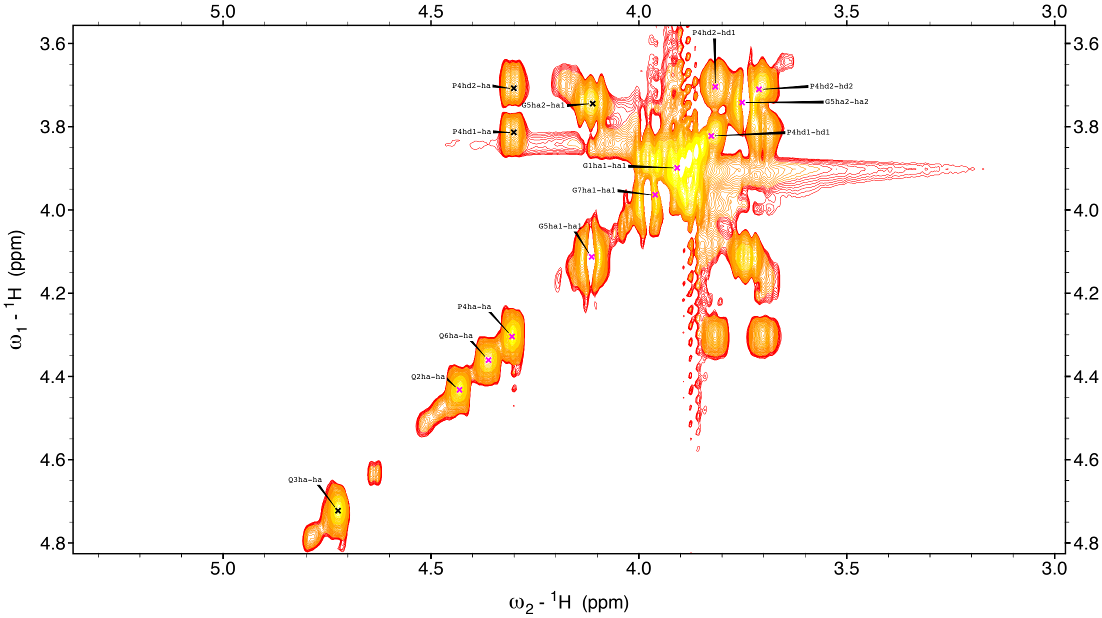


Figure 28 TOCSY spectrum showing the $H_{C\alpha}$-$H_{C\alpha}$ region for GQQPGQG. This region identifies Glycine-5 with two $H_{C\alpha}$ resonances (4.11 and 3.74ppm) and Proline-4 with two $H_{C\delta}$ protons (3.81 and 3.70ppm). Glycine-7 and Glycine -1 are difficult to distinguish from the solvent peak for TFE (3.88ppm)

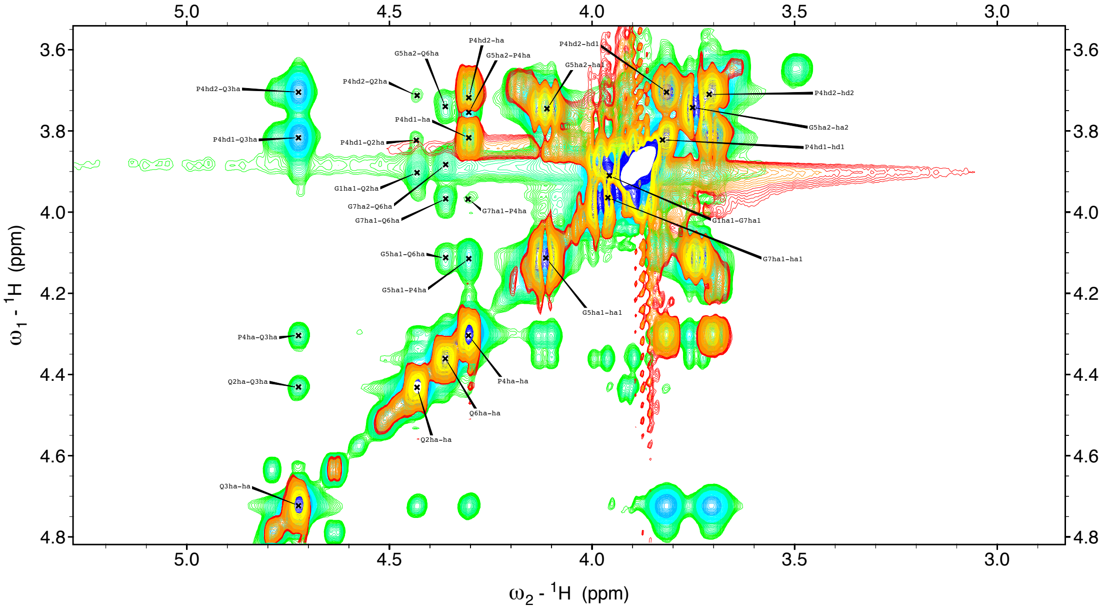


Figure 29 TOCSY and NOESY spectra showing the $H_{C\alpha}$-$H_{C\alpha}$ region for GQQPGQG. The through-space NOE connectivities in this region are important for determining backbone conformation especially related to proline cis-trans isomerization. Trans-proline is identified by characteristically strong NOE cross-peaks between P4-H_{δ} and Q3-$H_{C\alpha}$ protons as observed in these spectra. Cis-proline, which is less abundant in nature, is typically identified by a strong NOE cross-peak between proline H_{α} and the preceding residue Q3-H_{α}. A weak NOE resonance is observed indicating cis-proline conformation, which is not supported by other NOEs or the model.

The NOESY experiment produces cross-peaks between resonances within ~5.0Å distance through space. Signal intensity provides a measure of proximity and allows for classification into one of four groups; strong 1.8Å-2.8Å, medium 1.8Å-3.8Å, weak 1.8Å-5.0Å, and very weak 1.8Å-5.5Å. The distance restraints were internally calibrated using proton resonances in proline, which are distance constrained by the ring structure.

After the initial structure was solved, the resonance assignments and NOE distance constraints were refined to satisfy all NOE and dihedral angle violations. In total 139 resonances were identified, which were distributed between intra-residue, sequential, medium range and long-range constraints.

To calculate the three-dimensional structure of GQQPGQG using NMR data, I supplied the NOE distance constraints to the software package CNSsolve v1.3 Brünger et al. (1998). Using the constraints plus simulated annealing and minimization I generated 30 trial structures from which 24 low energy structures were chosen for further analysis. A summary of NOE constraints can be seen in Table 4 below. A table of chemical shift values and NOEs can be seen in Appendix Table A. 1 and Table A. 2.

Table 4 Summary of NOE constraints for the final 24 low energy structures.

Summary of NOE Constraints 24 Low Energy Structures

Total Number of Constraints	**139**	
Intra-Residue Constraints (I=J)	54	
Sequential Constraints (I-J)=1	51	
Backbone-Backbone		7
Backbone-Side Chain		1
Side Chain-Side Chain		43
Medium Range Constraints 1< (I-J) <5	26	
Backbone-Backbone		6
Backbone-Side Chain		2
Side Chain-Side Chain		18
Long Range Constraints (I-J) >= 5	8	
RMSDs to Mean Structure (Å)		
Backbone	0.375 Å	
All atoms	1.117 Å	
Percentage of residues in regions of ϕ-ψ space		
Core	86.0%	
Allowed	14.0%	

The 24 structures shown superimposed in Figure 30 have a backbone RMSD of 0.375Å a low value. The backbone RMSD is a measure of the goodness-of-fit for the backbone atoms in a group of structures with identical residues. The low RMSD value provides a measure of confidence that the structures are correct, since there are few alternate solutions for a low energy structure given our experimental restraints.

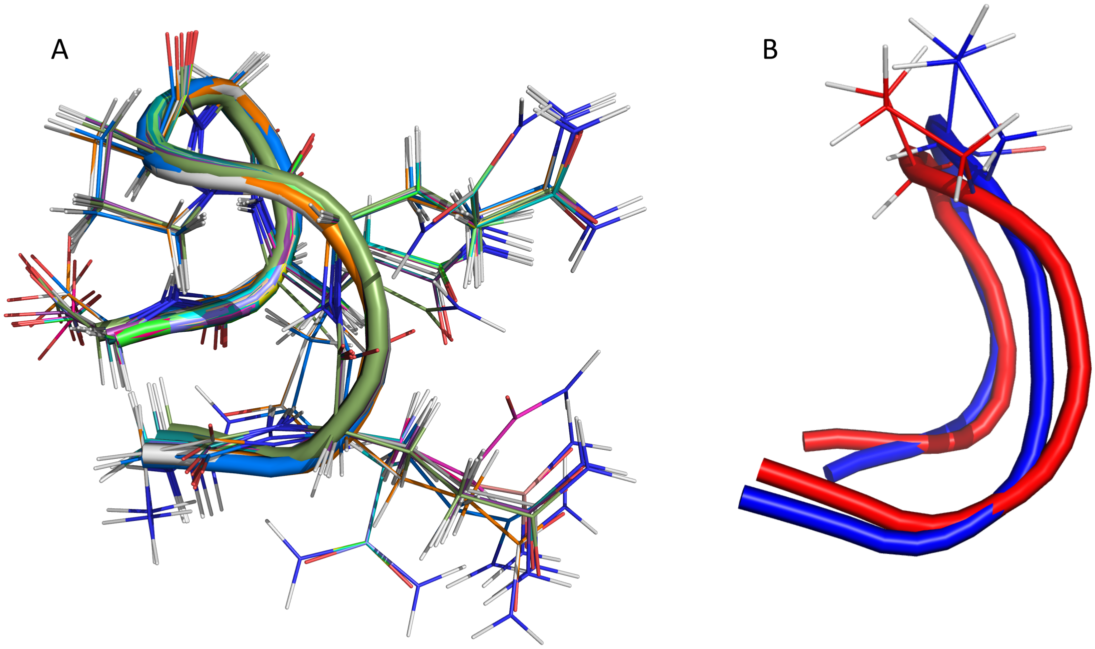


Figure 30 Superimposition of 24 low energy structures (A) generated using NOE distance constraints with simulated annealing and minimization. The backbone RMSD for these structures was calculated to be 0.375Å and the all atom RMSD = 1.117Å this measure of fit suggests there were few other solutions for soling the low-energy structure given the NMR constraints used to generate those structures. The average structure (B) in red was further minimized using free MD to achieve the final lowest energy structure in blue.

Using an average structure generated from the 24 low energy structures I did a final step of refinement using free molecular dynamics in a water box (1.0ps). This step was necessary since CNSsolve uses NOE constraints in the minimization stage, and those constraints are lifted during the MD minimization step. The final and lowest energy structure is shown in Figure 30, and is the structure used for the remaining checks of structure quality.

Using the software package Procheck NMR Laskowski et al. (1996) I performed checks of structure quality for each of the 24 structures and the average structure. These checks include Ramachandran plots and checks for NOE restraint violations, of which none were found.

Using the average structure I generated a set of Ramachandran plots based on work by Lovell et al. (2003). In their work they used a large set of high-resolution structures to improve upon the Ramachandran plot creating a subset of allowable phi and psi regions for proline, glycine, and residues preceding proline. Their premise is that proline is constrained to a narrower range of phi and psi angles, which also narrows the allowable regions for residues preceding proline. Conversely, glycine with no side-chain can accommodate a wider range of phi and psi angles resulting in a relaxed allowable region compared to the other standard amino acids. Similar to the traditional Ramachandran plot, N-terminal and C-terminal glycine residues are excluded. The Ramachandran plots shown in Figure 31 place residues Q2, P4, G5 and Q6 in the most favorable regions and Q3 in the allowable phi psi region.

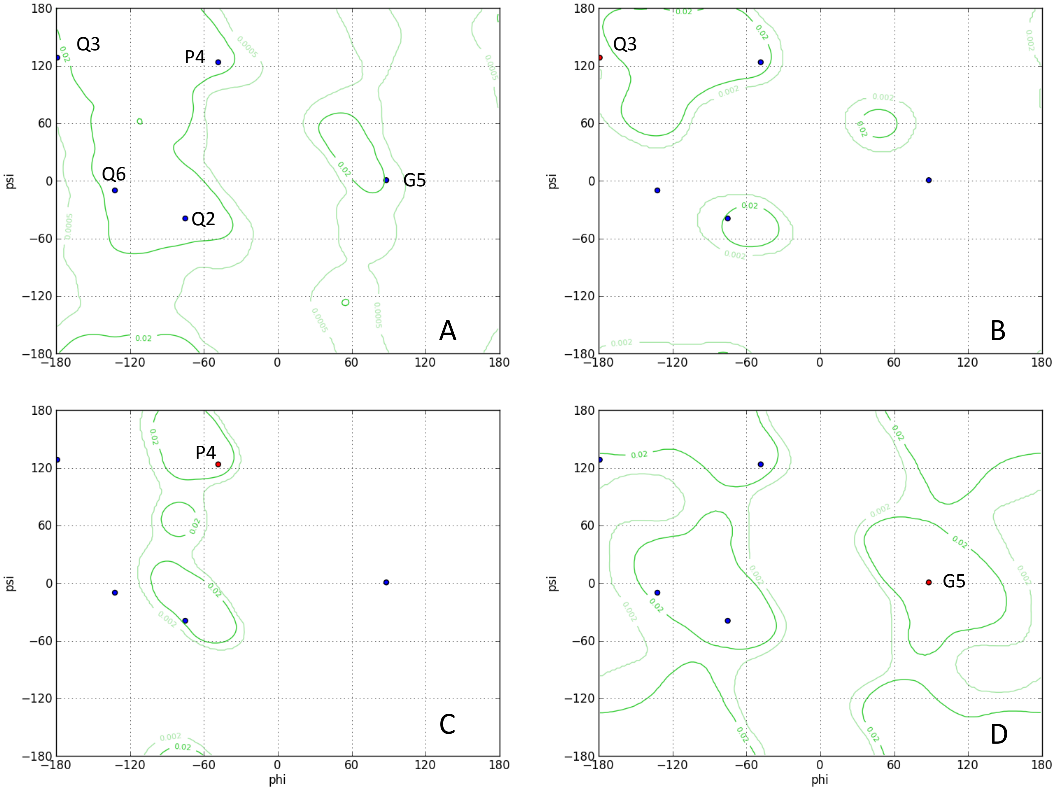


Figure 31 Ramachandran plots for QQPGQ. Four Ramachandran plots are needed to best describe the peptide. (A), general case describing most standard amino acid residues, (B) the allowable regions for amino acids preceding proline with Q3 shown in red, (C) the allowable region for proline with P4 shown in red and (D) allowable regions for glycine (other than N-, and C-terminal glycine) with G5 shown in red. Only Q2,Q3, P4, G5 and Q6 from the lowest energy average structure are considered in these plots.

I determined the turn to be a type II β-turn based on nomenclature by Hutchinson & Thornton (1994) and using the measured phi and psi angles across QPGQ shown in Table 5. This finding is supported by the chemical shift index and related NOEs shown in Figure 32.

Table 5 Phi-Psi angle assignment for the minimized average structure GQQPGQG. The assigned dihedral angles are predicted to form a type-II β-βturn across QPGQ based on the method of Hutchinson & Thornton (1994).

Residue	ϕ-Phi	ψ-Psi	Type-II β-Turn Predicted ϕ,ψ
Gly-1	-	-	
Gln-2	-73.99	-37.34	
Gln-3	-177.69	127.53	i
Pro-4	-47.71	124.79	i+1 -60_{ϕ}, 131_{ψ}
Gly-5	89.36	0	i+2 84_{ϕ}, 0_{ψ}
Gln-6	-131.52	-10.02	i+3
Gly-7	-	-	

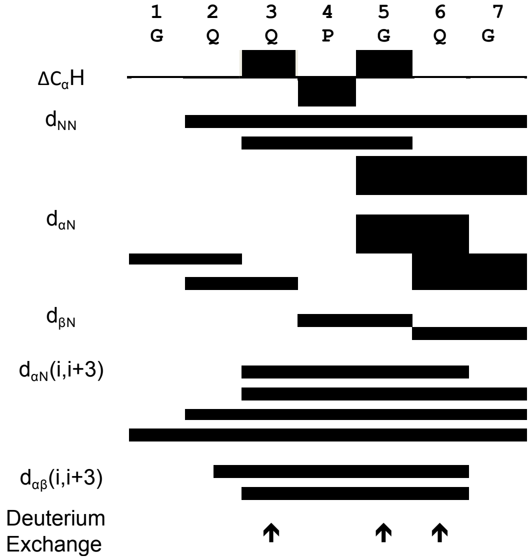


Figure 32 Schematic diagram of sequential and medium range NOE connectivities plus H_N-deuterium exchange. The chemical shift index (CSI) $\Delta H_{C\alpha}$ (top) indicates a turn across QPGQ. CSI values for G1, Q2, Q6 and G7 were less than 0.1ppm from the random coil $H_{C\alpha}$ and were given a CSI value of 0. Sequential and medium range NOE distances are shown (lower graphs) with NOE intensity indicated by bar thickness. Hydrogen-Deuterium exchange rate indicates hydrogen bonding by Q3-H_N, Q6-H_N and G5-H_N (bottom graph)

Hydrogen Deuterium Exchange

One characteristic of a type-II β-turn is a hydrogen bond between residues in positions 1 and 4 of the turn. Using a hydrogen deuterium exchange experiment (Figure 33) I confirmed Q6-H_N was involved in a hydrogen bond. During this experiment multiple 1D-^{1}H-NMR spectra are collected over a 24-hour time-span. Labile protons rapidly exchange with deuterium present in the solvent (deuterated-TFE plus 1% D_2O v/v) and protons involved in hydrogen bonds exchange more slowly. As hydrogen atoms are exchanged with deuterium the hydrogen signal diminishes until eventually hydrogen and deuterium within the peptide are at equilibrium.

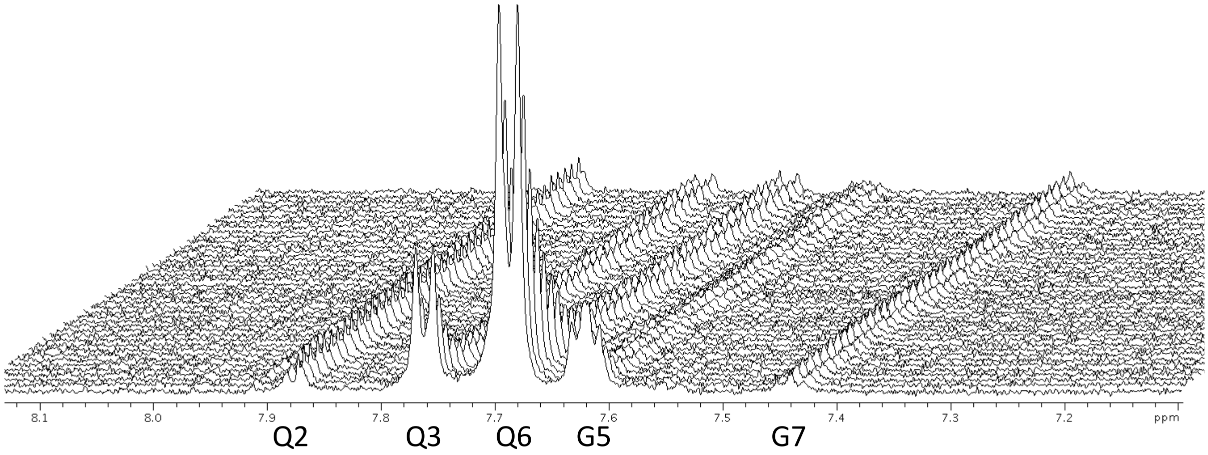


Figure 33 Hydrogen – deuterium exchange experiment indicating backbone H_N hydrogen bonding. In this experiment a series of one-dimensional ^{1}H-NMR spectra are collected over a 24 hour time-span. The proton signal intensities diminish as they exchange with deuterium present from added D_2O. Those resonances involved in hydrogen bonds exchange more slowly than non-bonded labile protons.

Observing a hydrogen bond for Q6-H_N is support for our solved three-dimensional structure showing a β-turn . It is expected that a β-turn should form a 10-member ring through the hydrogen bond, which is shown in Figure 34. As part of this 10-member ring we would also expect to see slightly weaker hydrogen bond character for Q2 and G5 H_N protons, which is also observed in the hydrogen deuterium exchange experiment.

Other indications of a turn are observed in the NOESY spectrum shown in Appendix Figure A. 2, where G5-H_N and Q6-H_N show through space connectivity indicating a bent backbone conformation.

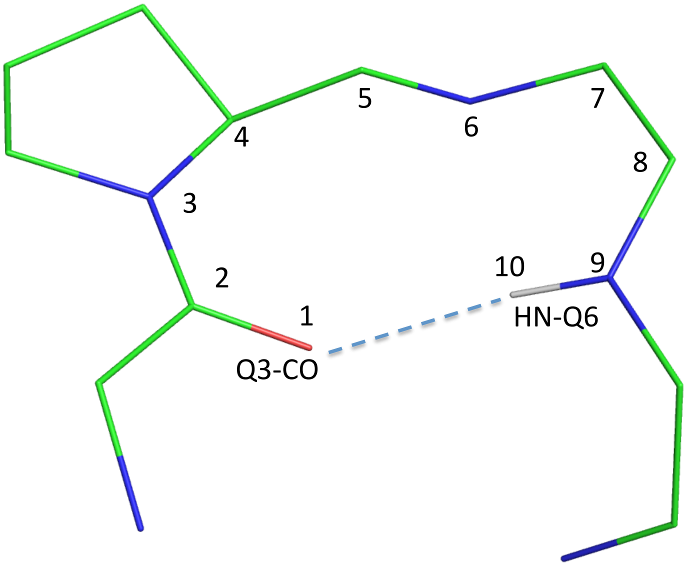


Figure 34 Backbone structure of GQQPGQG showing hydrogen bond between Q6-H$_N$ and Q3-CO. The hydrogen bond, which stabilizes the β-turn structure forms a characteristic 10-member ring.

Proteins of Known Structure

To understand how this 4-residue sequence may be found in nature I did a search of the Protein Data Bank for proteins of known structure containing the exact sequence GQQPGQG. No exact matches were found however 58 contained the shortened sequence QPGQ. I examined the structures for each protein and found that 44 were resolved in the region covering XXQPGQX.

The proteins were unrelated in structure and function however, in 100% of the proteins these residues (XXQPGQXX) are found as a link between regularly folded structures. In 33 of the 44 proteins, the sequence was found as a β-turn highly similar to the minimized structure of GQQPGQG in 100%TFE. In the remaining 11 proteins the sequence was either a partial turn or extended structure (not shown).

For the all of these proteins the turn was exposed at the surface of the protein, and in at least 6 the turn was annotated as being at or adjacent to a dimer interface.

From each of the 33 structures I extracted the 7 amino acid residues covering the β-turn sequence XXQPGQX. Since none of these proteins contained the exact sequence of my 7-residue peptide, I confined the analysis to the region covering QPGQ. When comparing the 33 regions to the same residues in GQQPGQG the backbone RMSD = 0.802Å. When the entire group of 34 is compared to an average, the backbone RMSD then equals 0.50Å. Figure 35 shows the backbone of these 33 structures overlaid with the structure of the full-length peptide GQQPGQG.

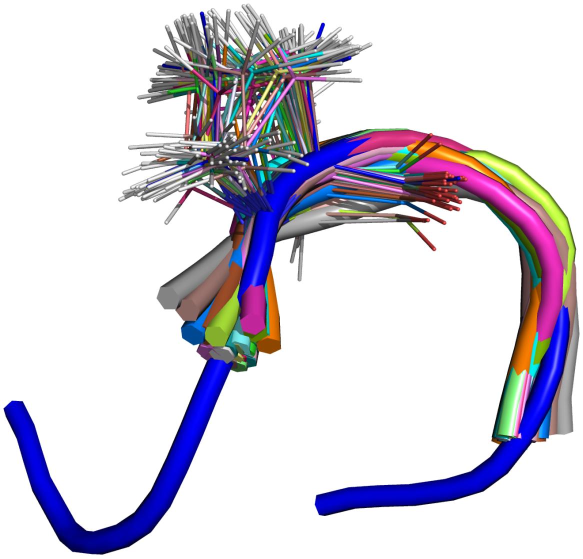


Figure 35 Structure alignment for 33 protein fragments with exact sequence QPGQ aligned with the full length 7mer GQQPGQG (shown in blue). The backbone RMSD between these 33 structures and the 7mer is .8022Å.

Molecular Dynamics

There were two observations from NMR experiments that could not be explained from the three-dimensional structure. In experiments conducted with 100% TFE, there were a second incomplete set of resonances within each spectrum. Due to their weak signal strength it was impossible to solve a second structure using these resonances and therefore it is unclear if a second stable conformation exists.

My second observation from NMR involved spectra collected in 100% water. The spectra were consistent with random coil structure predicted by CD and only sequential NOE assignments could be made. However, similar to the spectra in TFE additional resonances were observed in water. The observed $H_{C\alpha}$ chemical shift values were not consistent with those expected for random coil, and therefore the structure in water remains unclear.

To help understand the possible conformational changes in both 100% TFE and 100% water, I used molecular dynamics to simulate 50ns of molecular movement in both solvents.

Using this approach I measured the peptide in TFE having two states which both maintained the turn conformation. Each state shown in Figure 36 involves the side chain of Q2, where in state-1 it is aligned with the side chains of Q3 and Q6, while in state-2 it is found mostly directed away from Q3. In water the simulation modeled a transition from turn to extended then regained the turn conformation. The transition from turn to random coil happens after the peptide adopts state-2 when Q2 side-chain is positioned away from Q3 and Q6.

I observed the peptide transition through both states and into the extended conformation in both solvent conditions. Both simulations were started using the NMR structure solved in 100%TFE because the structure in water was not solved. During simulations for both solvent conditions, I observed the peptide regain the turn conformation from the extended form indicating the starting structure did not impose structure on the simulation.

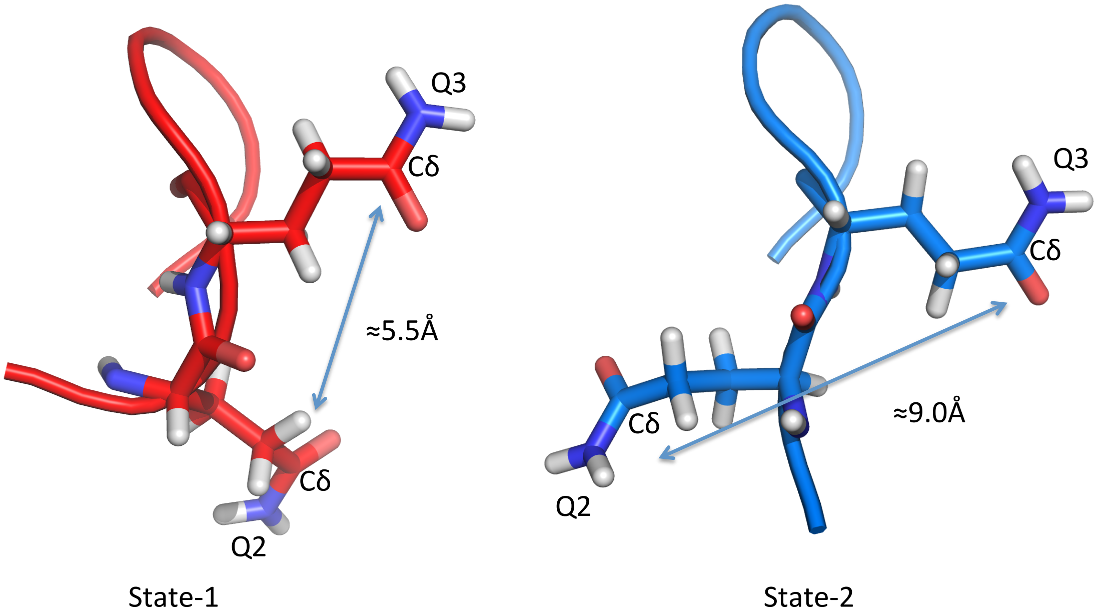


Figure 36 Representative structures taken during MD simulation in 100%TFE (left side in red) and in water (right side in blue). Each state is observed in both solvent conditions, however state-1 predominates in TFE and state-2 predominates in water. The states are defined by side chain orientation of glutamine 2. In state-1 the Q2 side-chain is aligned with the side-chain of Q3 and the radial distance between Cδ carbon is at it's least. In state-2 the Q2 side-chain is positioned away from the side-chain of Q3 resulting in the greatest distance between Cδ atoms. The peptide transitions to an extended conformation through state-2 and the β-turn conformation is stabilized by state-1.

I determined the radial pair distribution of the distance between δ-carbon atoms in Q2 and Q3 over the ~50ns simulation. This approach measures the distance between two selected atoms at each time-step (2.0fs/ts = 25 million observations), then builds a probability distribution of the observed distances. The plot of this distribution can be seen in Figure 37 and clearly shows the glutamine side-chains of Q2 and Q3 with closer proximity (Å) in TFE than when modeled in water over the same timespan.

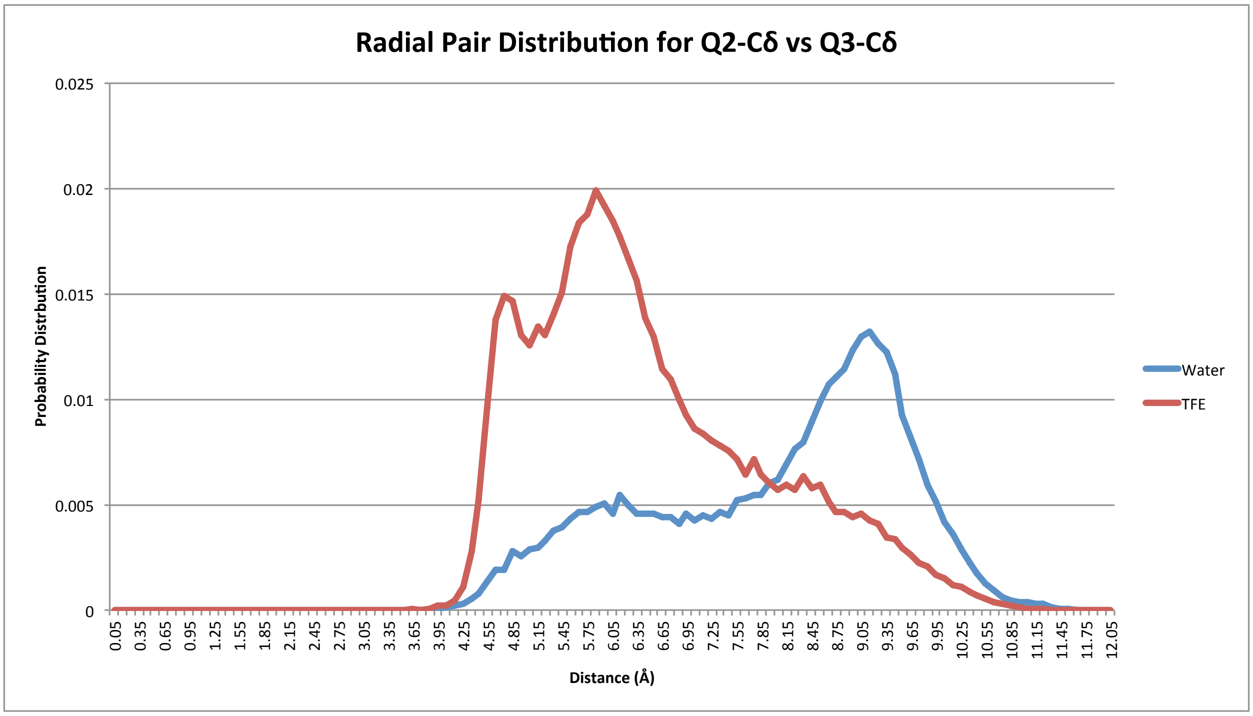


Figure 37 Plot of the radial pair distribution of distance between Cδ carbon atoms of Q2 and Q3 during ~50ns molecular dynamics simulation in 100% water (blue) and 100%TFE (red). State-1 predominates in TFE (see **Figure 36**) stabilizing the type II β-turn , while state-2 (see **Figure 36**) predominates in water.

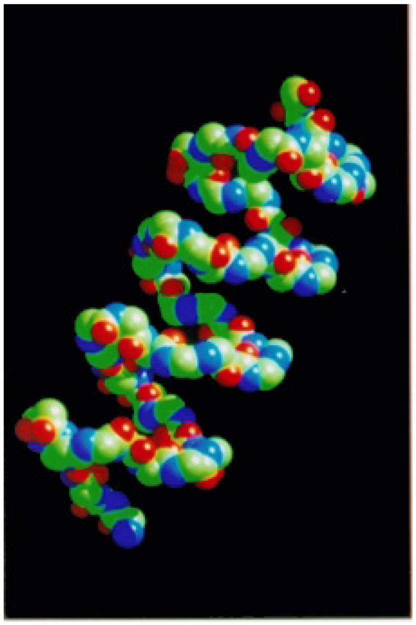
Discussion

My work represents the first known structure for any portion of the gluten forming proteins. The structure solved in 100% TFE was confirmed through experimental measures and by simulated molecular dynamics. I validated the use of TFE for peptide NMR by careful comparison with proteins of known structure finding near perfect agreement across QPGQ.

Researchers have tried to describe to molecular nature of gluten elasticity for many decades. In studies using native glutenin, the extractions are made from wheat flour under denaturing conditions. It is unclear how these structural studies relate to the native structure of glutenin, though studying the intact protein by other means has yet to be demonstrated.

My work builds upon a reductionist approach to study the smallest relevant structural unit of the folded protein that can itself fold. We have demonstrated the 7-mer QGGPGQG can fold into a stable type-II β-turn in100% TFE. Though it isn't clear whether the 7mer structure is itself repeated in the native protein.

Parchment et al. (2001) have proposed a model of glutenin, which is described as a continuous β-spiral structure shown in Figure 38 , which is based on repeating beta-turns.



.

Figure 38 Three-dimensional model for glutenin 1Dx5. proposed by Parchment et al. (2001) based on the repeating sequence GQQPGQG and assuming perfectly repeating type-II β-turn s Parchment et al. (2001)

In their model of glutenin Parchment et al. (2001) assume the repeated sequences translate to repeated turns and hence the spiral structure. I believe glutenin does contain repeating structures, which likely confer elasticity, although the length and composition of those repeats are still unknown. As stated earlier, if the repeating sequences are assumed to produce repeating structural features, then the repeat unit should be defined by those structures.

The earlier models of glutenin relied heavily on secondary structure prediction tools. These predictions are based largely on proteins of known structure, however there are very few examples of multiple tandem repeats in nature. Therefore it is unclear whether secondary structure predictions are valid in the context of sequence repeats from glutenin.

In my molecular dynamics study of the 7mer GQQPGQG I found glutamine-2 to be an important and flexible residue in the peptide. Q2 precedes the type-II β-turn across QPGQ and may allow some conformational flexibility before the more stable β-turn region.

There were not enough NOE resonances to solve the three-dimensional structure in water. However evidence from CD and NMR experiments in 100% water indicate an extended conformation. In my simulations of the peptide in water, the structure was not entirely random coil, rather it underwent fast conversion between the beta-turn structure and extended conformation. It is possible the NMR time-scale is too long (milliseconds) to observe conformational changes that are modeled on a nanosecond time-scale.

Future Work

I plan to solve the structure for QGQQPGQGQQPGQGQQ an offset double repeat 16-mer in 100% TFE to determine the structure for a single tandem repeat. Solving this structure will allow us to test the secondary structure predictions for tandem β-turns, and inform past and future models of the high molecular weight glutenin.

Blanch et al. (2003) produced water soluble fragments of high molecular weight glutenin by cleavage with trypsin. This is a promising technique and could lead to a solution structure using NMR. I plan to create lager fragments of the high molecular weight glutenin using recombinant expression as well as tryptic digests for study by NMR. Past researchers have used atomic force microscopy imaging to study the topography of high molecular weight glutenin. I plan to study larger fragments and intact glutenin using atomic force microscopy under the force extension mode. This type of experiment can be used to study the elastic properties of glutenin under denaturing and non-denaturing conditions.

Synthetic β-turns are an area of current interest for their potential anti-microbial properties. GQQPGQG may be important for its ability to form a stable beta-turn within a short sequence. The anti-microbial properties of GQQPGQG are unknown so I will study this peptide and variants for their turn forming and antimicrobial properties.

Postlude

In my work I studied two important cereal proteins using a reductionist approach. I studied the sequence repeat GQQPGQG regarded as the primary structural unit in glutenin conferring elastic properties to wheat dough. And I studied β-kafirin, a single member among a large group of proteins, the kafirins, known to have poor digestibility after cooking. Both studies although different were devised under a philosophy I have been shaping over the past five years. A philosophy that will guide the rest of my career as a scientist. And that is the greatest opportunities lie at the boundary between disciplines. Early on, I realized the great potential from combining of grain processing and biochemistry, and now I have shown a way to approach real problems in cereal science using biochemical strategies.

We will use biochemistry in our efforts to understand the effects of cereal storage proteins on human health and disease. This is especially important in the area of sorghum protein digestibility and gluten toxicity. We will solve these problems by isolating parts of a complex mixture, by studying their characteristics and measuring (or modeling) their interactions. The result will be a safer and more nutritious food supply.

Our work ahead is challenging. As biochemists we study biological systems and the pathways that define them. We try to relate the structure of proteins, peptides and metabolites to their function within a living organism. This line of thinking can be applied to cereal science, though the system is man-made. Our man-made system is a group of ingredients and the environment contains non-natural mechanical forces, high temperatures and pressure. Structure and function are no longer biochemical terms rather they describe the effect of a protein or small molecule on the outcome of a product (e.g. color, texture and flavor). However, pathways to these outcomes also exist.

Every seed is a living organism, and the cellular machinery that exists in the seed is no less powerful than the machinery in our own factories. I plan to continue studying the biochemistry of cereal grains and hope to use my knowledge, and the tools of biochemistry to solve the problems that face our industry today.

Appendix A - In-Source-Decay Spectra from β-Kafirin

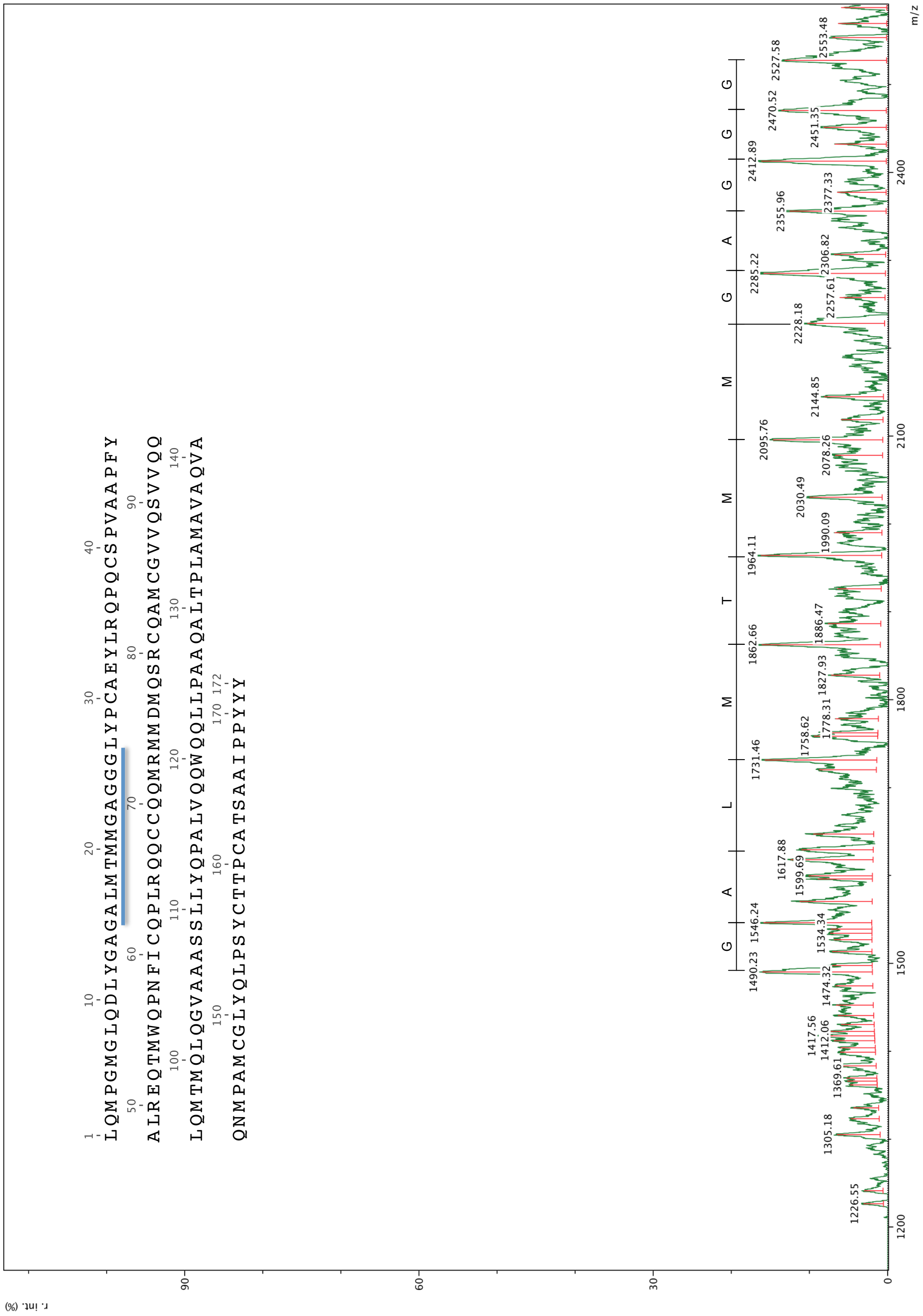
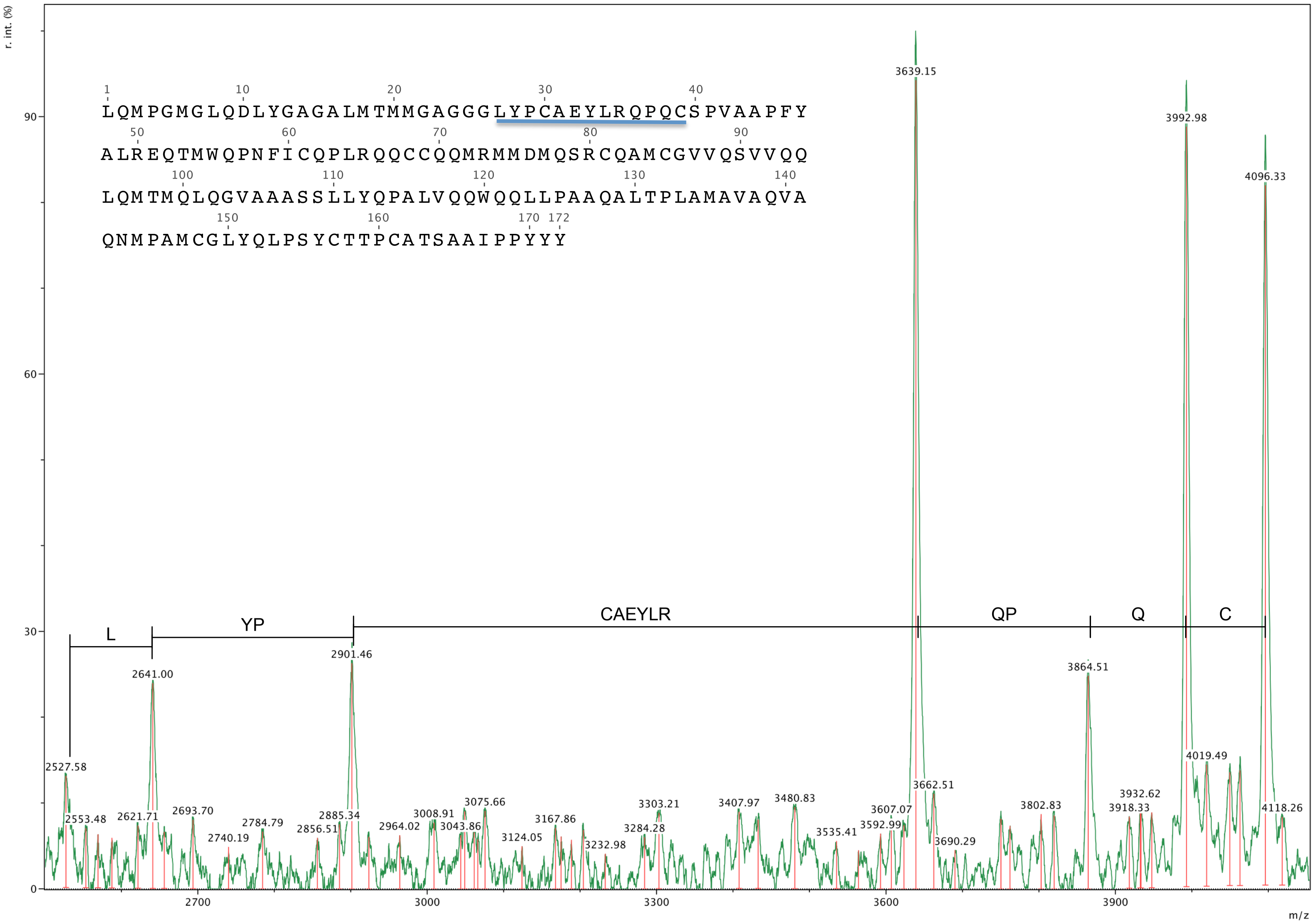
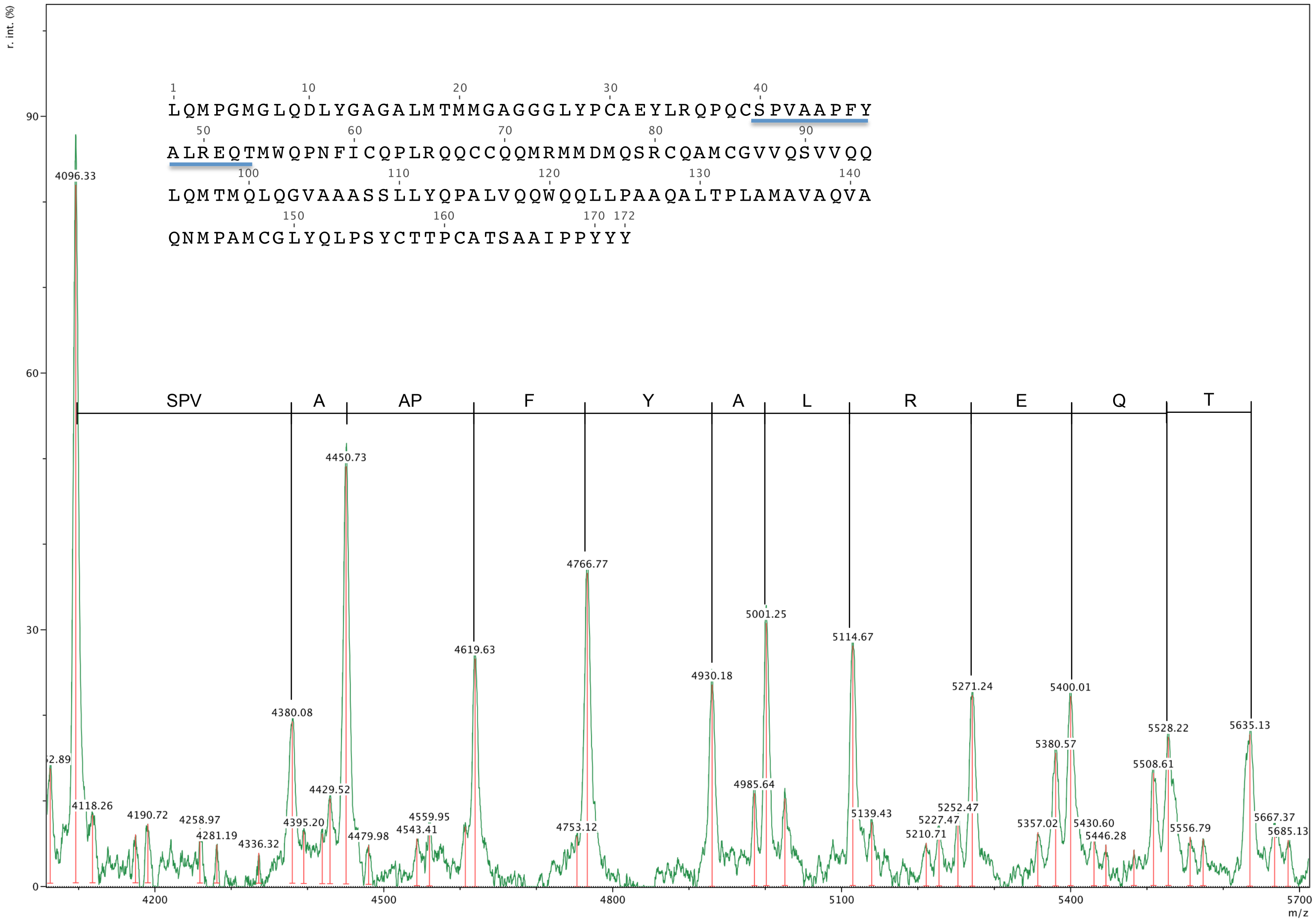
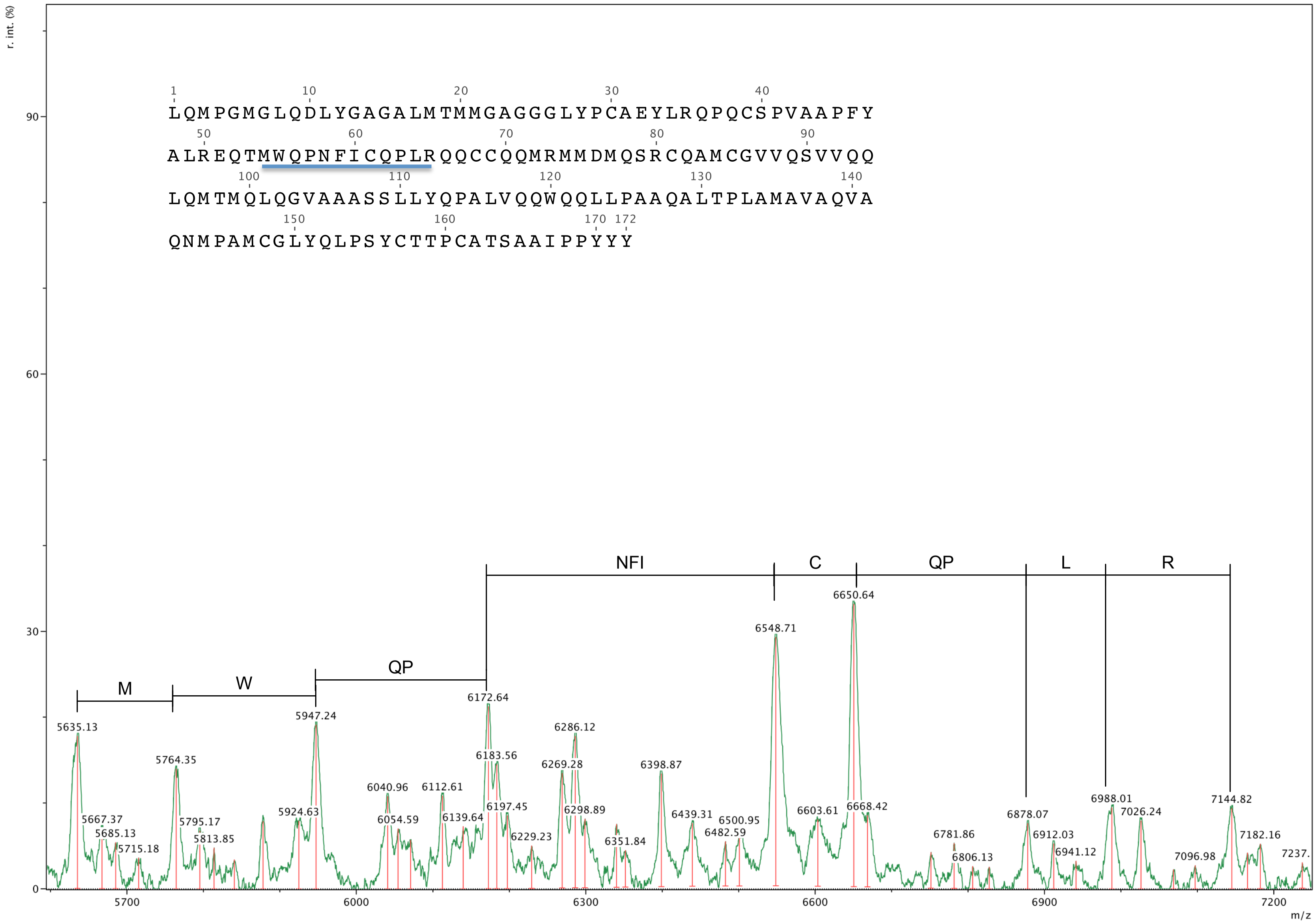
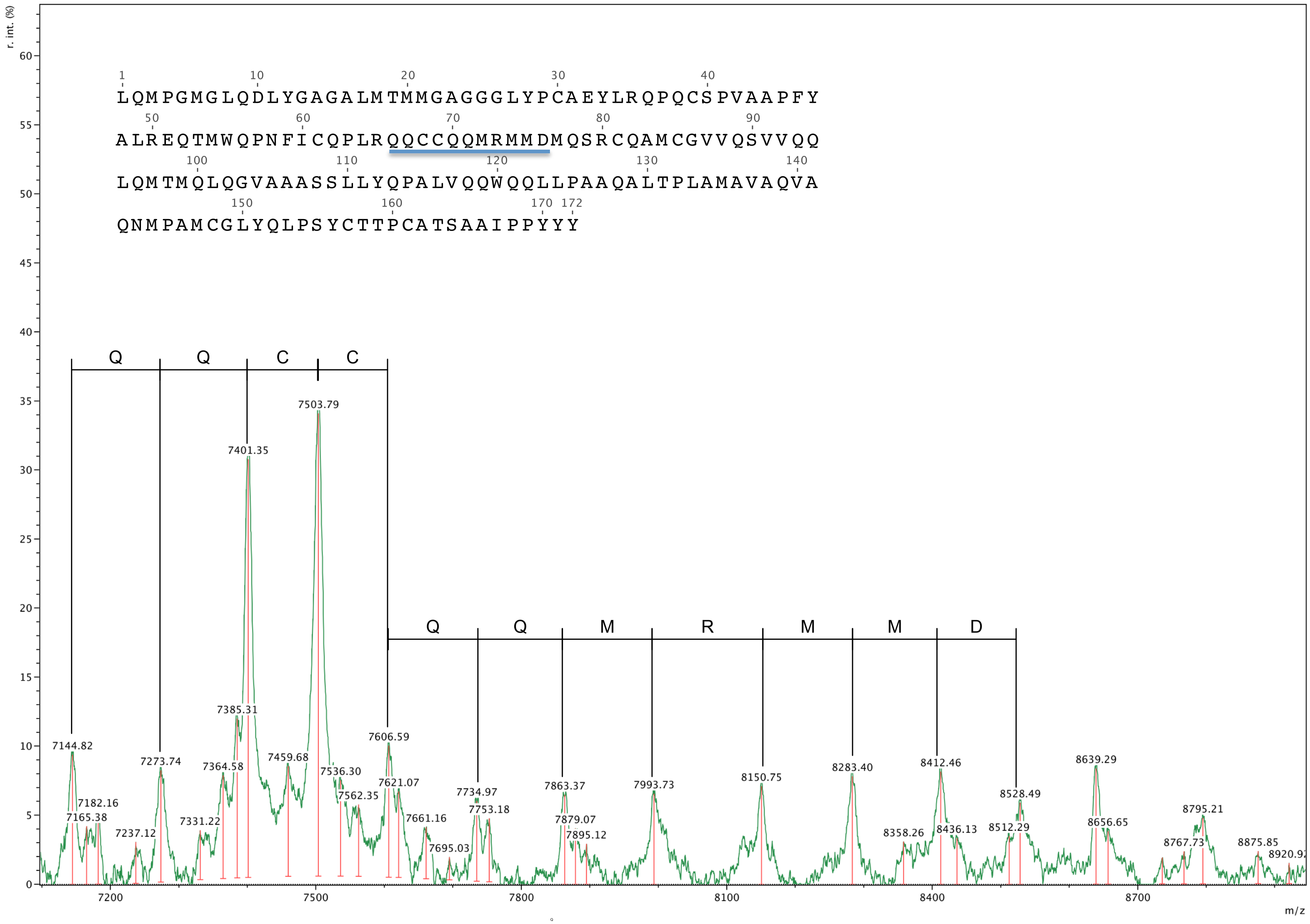


Figure A.1 In source decay sequencing for β-kafirin using MALDI-TOF MS with matrix 1,5-DAN


r. int. (%)
0
30
60
90
2700
3000
3300
3600
3900
m/z
1 10 20 30 40
LQMPGMGLQDLYGAGALMTMMGAGGGLYPCAEYLRQPQCSPVAAPFY
50 60 70 80 90
ALREQTMWQPNFICQPLRQQCCQQMRMMDMQSRCQAMCGVVQSVVQQ
100 110 120 130 140
LQMTMQLQGVAAASSLLYQPALVQQWQQLLPAAQALTPLAMAVAQVA
150 160 170 172
QNMPAMCGLYQLPSYCTTPCATSAAIPPYYY
L
YP
CAEYLR
QP
Q
C
2527.58
2553.48
2621.71
2641.00
2693.70
2740.19
2784.79
2856.51
2885.34
2901.46
2964.02
3008.91
3043.86
3075.66
3124.05
3167.86
3232.98
3284.28
3303.21
3407.97
3480.83
3535.41
3592.99
3607.07
3639.15
3662.51
3690.29
3802.83
3864.51
3918.33
3932.62
3992.98
4019.49
4096.33
4118.26


r. int. (%)
90
60
30
0
1 10 20 30 40
LQMPGMGLQDLYGAGALMTMMGAGGGLYPCAEYLRQPQCSPVAAPFY
50 60 70 80 90
ALREQTMWQPNFICQPLRQQCCQQMRMMDMQSRCQAMCGVVQSVVQQ
100 110 120 130 140
LQMTMQLQGVAAASSLLYQPALVQQWQQLLPAAQALTPLAMAVAQVA
150 160 170 172
QNMPAMCGLYQLPSYCTTPCATSAAIPPYYY
SPV
A
AP
F
Y
A
L
R
E
Q
T
4096.33
2.89
4118.26
4190.72
4258.97
4281.19
4336.32
4380.08
4395.20
4429.52
4450.73
4479.98
4543.41
4559.95
4619.63
4753.12
4766.77
4930.18
4985.64
5001.25
5114.67
5139.43
5210.71
5227.47
5252.47
5271.24
5357.02
5380.57
5400.01
5430.60
5446.28
5508.61
5528.22
5556.79
5635.13
5667.37
5685.13
4200
4500
4800
5100
5400
5700
m/z


r. int. (%)
90
60
30
0
1
10
20
30
40
LQMPGMGLQDLYGAGALMTMMGAGGGLYPCAEYLRQPQCSPVAAPFY
50
60
70
80
90
ALREQTMWQPNFICQPLRQQCCQQMRMMDMQSRCQAMCGVVQSVVQQ
100
110
120
130
140
LQMTMQLQGVAAASSLLYQPALVQQWQQLLPAAQALTPLAMAVAQVA
150
160
170
172
QNMPAMCGLYQLPSYCTTPCATSAAIPPYYY
M
W
QP
NFI
C
QP
L
R
5635.13
5667.37
5685.13
5715.18
5764.35
5795.17
5813.85
5924.63
5947.24
6040.96
6054.59
6112.61
6139.64
6172.64
6183.56
6197.45
6229.23
6269.28
6286.12
6298.89
6351.84
6398.87
6439.31
6482.59
6500.95
6548.71
6603.61
6650.64
6668.42
6781.86
6806.13
6878.07
6912.03
6941.12
6988.01
7026.24
7096.98
7144.82
7182.16
7237.
5700
6000
6300
6600
6900
7200
m/z


r. int. (%)
60
55
50
45
40
35
30
25
20
15
10
5
0
1 10 20 30 40
LQMPGMGLQDLYGAGALMTMMGAGGGLYPCAEYLRQPQCSPVAAPFY
50 60 70 80 90
ALREQTMWQPNFICQPLRQQCCQQMRMMDMQSRCQAMCGVVQSVVQQ
100 110 120 130 140
LQMTMQLQGVAAASSLLYQPALVQQWQQLLPAAQALTPLAMAVAQVA
150 160 170 172
QNMPAMCGLYQLPSYCTTPCATSAAIPPYYY
Q
Q
C
C
Q
Q
M
R
M
M
D
7144.82
7165.38
7182.16
7237.12
7273.74
7331.22
7364.58
7385.31
7401.35
7459.68
7503.79
7536.30
7562.35
7606.59
7621.07
7661.16
7695.03
7734.97
7753.18
7863.37
7879.07
7895.12
7993.73
8150.75
8283.40
8358.26
8412.46
8436.13
8512.29
8528.49
8639.29
8656.65
8767.73
8795.21
8875.85
8920.92
7200
7500
7800
8100
8400
8700
m/z

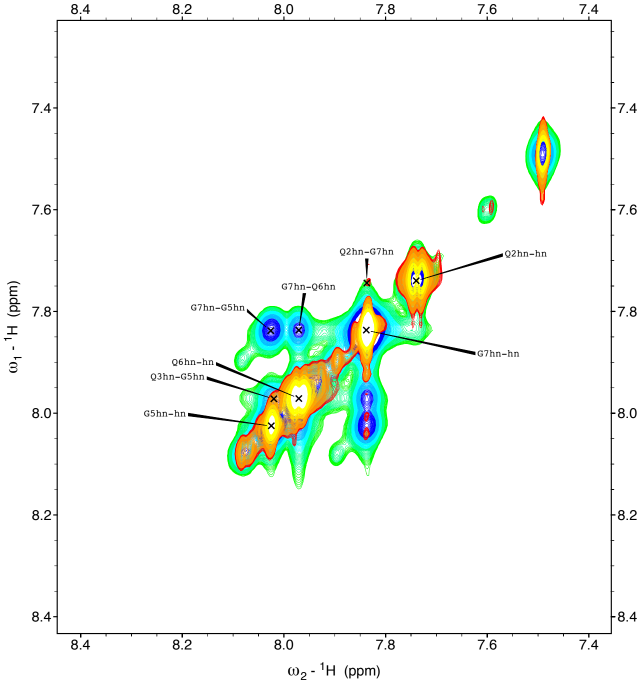


Figure A. 2 NOESY and TOCSY spectra showing H_N-H_N region for GQQPGQG. Through space connectivities between backbone H_N resonances indicate a bend or turn in the backbone structure. The strong NOE cross-peak between G7- H_N and Q6-H_N results from Glycine-7 bending in an apparent turn, which can not be characterized based on being the terminal residue.

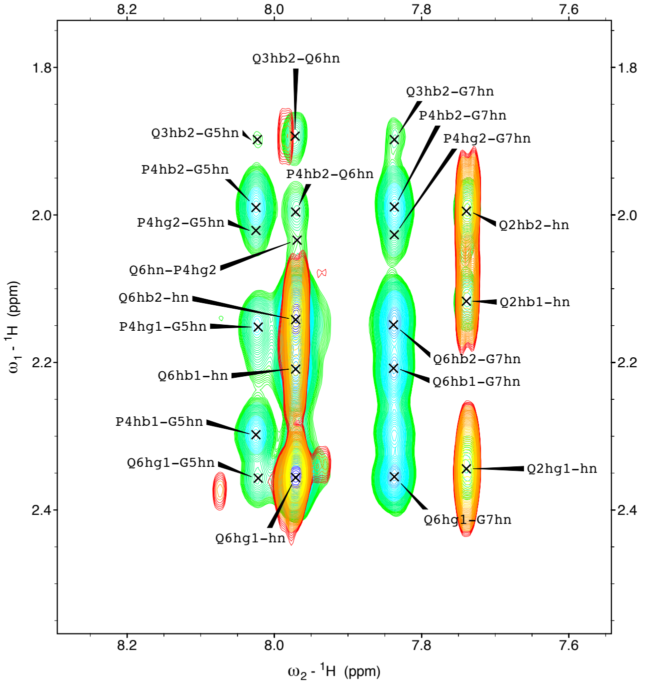


Figure A. 3 NOESY and TOCSY spectra showing H_N- $H_{C\beta}$ and $H_{C\gamma}$ region for GQQPGQG. NOE cross-peaks in this region indicate side-chain positions in reference to the backbone H_N protons. This is especially true for Proline-4 with a compact side-chain.

Table A. 1 Table of chemical shift values.

^{1}H-Chemical Shift Table GQQPGQG									
	H_N	$H_{C\alpha1}$	$H_{C\alpha2}$	$H_{C\beta1}$	$H_{C\beta2}$	$H_{C\gamma1}$	$H_{C\gamma2}$	$H_{C\delta1}$	$H_{C\delta2}$
G1	7.96	3.908	-	-	-	-	-	-	-
Q2	7.74	4.43	-	2.118	2.117	2.34	2.33	-	-
Q3	7.986	4.72	-	2.17	1.89	2.39	2.35	-	-
P4	-	4.3	-	2.29	1.98	2.14	2.03	3.81	3.7
G5	8.025	4.11	3.74	-	-	-	-	-	-
Q6	7.971	4.36	-	2.21	2.14	2.35	-	-	-
G7	7.837	3.95	3.89	-	-	-	-	-	-

Table A. 2 NOE resonance assignment for GQQPGQG

Resonance 1	*Chemical Shift 1 (ppm)*	*Resonance 2*	*Chemical Shift 2 (ppm)*
G1-ha1	3.91	G7-ha1	3.957
Q2-hn	7.744	G7-hn	7.837
Q6-hb1	2.211	G1-ha1	3.905
Q6-hb2	2.14	G1-ha1	3.907
Q6-hg1	2.359	G1-ha1	3.897
G7-ha1	3.958	G1-hn	7.966
G7-ha1	3.95	Q2-hn	7.738
G7-ha2	3.878	Q2-hn	7.738
Q3-ha	4.725	G5-hn	8.023
Q3-ha	4.726	Q6-hn	7.963
Q3-hb2	1.89	G5-ha1	4.126
Q3-hb2	1.898	G5-hn	8.023
Q3-hb2	1.898	G7-hn	7.837
Q3-hb2	1.895	Q6-hg1	2.365
Q3-hb2	1.893	Q6-hn	7.972
Q3-hn	7.972	G5-hn	8.02
P4-ha	4.309	Q6-hn	7.971
P4-hb2	1.993	G1-ha1	3.906
P4-hb2	1.992	G7-ha1	3.977
P4-hb2	1.989	G7-hn	7.837
P4-hb2	1.998	Q6-hg1	2.337
P4-hb2	1.996	Q6-hn	7.971
P4-hd1	3.818	G7-hn	7.837
P4-hd1	3.823	Q2-ha	4.434
P4-hd1	3.804	Q6-hn	7.972
P4-hd2	3.713	Q2-ha	4.432
P4-hg2	2.027	G7-hn	7.837
P4-hg2	2.015	Q6-hg1	2.342
Q6-ha	1.898	Q3-hb2	4.362

Q6-hb1	2.203	Q2-ha	4.432
Q6-hn	2.034	P4-hg2	7.969
G7-ha1	3.968	P4-ha	4.306
G7-hn	7.838	G5-hn	8.026
G7-hn	4.723	Q3-ha	7.833
G1-ha1	3.903	Q2-ha	4.432
G1-ha1	3.906	Q2-hn	7.739
Q2-ha	4.431	Q3-ha	4.724
Q2-ha	1.9	Q3-hb2	4.432
Q2-ha	4.43	Q3-hn	7.983
Q2-hg2	2.306	Q3-ha	4.724
Q3-hb1	2.187	P4-hd1	3.816
Q3-hb1	2.188	P4-hd2	3.705
Q3-hb2	1.897	P4-ha	4.304
Q3-hb2	1.895	P4-hd1	3.816
Q3-hb2	1.895	P4-hd2	3.705
Q3-hg2	2.365	P4-hd1	3.816
Q3-hg2	2.364	P4-hd2	3.708
P4-ha	4.305	G5-hn	8.025
P4-ha	4.304	Q3-ha	4.724
P4-hb1	2.302	G5-ha1	4.117
P4-hb1	2.298	G5-ha2	3.76
P4-hb1	2.298	G5-hn	8.025
P4-hb2	1.989	G5-ha1	4.117
P4-hb2	1.994	G5-ha2	3.75
P4-hb2	1.99	G5-hn	8.025
P4-hb2	2	Q3-ha	4.724
P4-hd1	3.803	G5-hn	8.025
P4-hd1	3.817	Q3-ha	4.724
P4-hd2	3.704	Q3-ha	4.724
P4-hg1	2.146	G5-ha2	3.762
P4-hg1	2.152	G5-hn	8.022
P4-hg2	2.021	G5-hn	8.025
P4-hg2	2.021	Q3-ha	4.724
G5-ha1	4.115	P4-ha	4.304
G5-ha1	4.112	Q6-ha	4.361
G5-ha1	4.114	Q6-hn	7.972
G5-ha2	3.754	P4-ha	4.304
G5-ha2	3.74	Q6-ha	4.362
G5-ha2	3.746	Q6-hn	7.972
Q6-ha	4.353	G5-hn	8.02
Q6-ha	4.36	G7-hn	7.837

Q6-hb1	2.205	G7-ha1	3.962
Q6-hb1	2.208	G7-hn	7.838
Q6-hb2	2.146	G5-ha1	4.101
Q6-hb2	2.145	G7-ha1	3.971
Q6-hb2	2.149	G7-hn	7.838
Q6-hg1	2.359	G5-ha1	4.115
Q6-hg1	2.357	G5-ha2	3.751
Q6-hg1	2.357	G5-hn	8.022
Q6-hg1	2.357	G7-ha1	3.968
Q6-hg1	2.357	G7-ha2	3.875
Q6-hg1	2.355	G7-hn	7.837
G7-ha1	3.967	Q6-ha	4.361
G7-ha2	3.883	Q6-ha	4.361
G7-hn	7.837	Q6-hn	7.971

References

Alberti, E., Gilbert, S., Tatham, A. S., & Shewry, P. R. (2002). Study of wheat high molecular weight 1Dx5 subunit by 13C and 1H solid-state NMR. II. Roles of nonrepetitive terminal domains and length of repetitive domain, *Biopolymers,* 65(2),158-168.

Anderson, O. D., Gu, Y. Q., Kong, X., Lazo, G. R., & Wu, J. (2009). The wheat omega-gliadin genes: structure and EST analysis. *Functional & Integrative Genomics*, *9*(3), 397–410.

Anderson, O. D., Litts, J. C., & Greene, F. C. (1997). The α-gliadin gene family. I. Characterization of ten new wheat α-gliadin genomic clones, evidence for limited sequence conservation of flanking DNA, and southern analysis of the gene family, *Theoretical and Applied Genetics,* 95(1-2), 50-58.

Axtell, J. D., Kirleis, A. W., Hassen, M. M., D'Croz Mason, N., Mertz, E. T., & Munck, L. (1981). Digestibility of sorghum proteins. *Proceedings of the National Academy of Sciences*, *78*(3), 1333–1335.

Bean, S. R., Ioerger, B. P., & Blackwell, D. L. (2011). Separation of kafirins on surface porous reversed-phase high-performance liquid chromatography columns. *Journal Of Agricultural And Food Chemistry*, *59*(1), 85–91.

Behnke, C. A., Yee, V. C., Trong, I. L., Pedersen, L. C., Stenkamp, R. E., Kim, S. S., et al. (1998). Structural determinants of the bifunctional corn Hageman factor inhibitor: x-ray crystal structure at 1.95 A resolution. *Biochemistry*, *37*(44), 15277–15288.

Belton, P. S., Delgadillo, I., Halford, N. G., & Shewry, P. R. (2006). Kafirin structure and functionality. *Journal of Cereal Science*, *44*(3), 272–286.

Blackwell, D. L., & Bean, S. R. (2012). Separation of alcohol soluble sorghum proteins using non-porous cation-exchange columns. *Journal of chromatography. A*, *1230*, 48–53.

Blanch, E. W., Kasarda, D. D., Hecht, L., Nielsen, K., & Barron, L. D. (2003). New Insight into the Solution Structures of Wheat Gluten Proteins from Raman Optical Activity. *Biochemistry*, *42*(19), 5665–5673.

Bonomi, F., Iametti, S., Mamone, G., & Ferranti, P. (2013). The Performing Protein: Beyond Wheat Proteomics? *Cereal Chemistry*, *90*(4), 358–366.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., et al. (1998). Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta crystallographica Section D, Biological crystallography*, *54*(5), 905–921.

Chapman, J. R. (2000). *Mass Spectrometry of Proteins and Peptides*. Springer.

Clark, K. (1991, November 19). *Sequence Alignments without the use of Arbitrary Parameters*. (G. R. Reeck, Ed.).

de Mesa-Stonestreet, N. J., Alavi, S., & Bean, S. R. (2010). Sorghum proteins: the concentration, isolation, modification, and food applications of kafirins. *Journal of Food Science*, *75*(5), R90–R104.

Delaglio, F., Grzesiek, S., Vuister, G., Zhu, G., Pfeifer, J., & Bax, A. (1995). NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *Journal of Biomolecular NMR*, *6*(3), 277–293.

Dupont, F. M., Vensel, W. H., Tanaka, C. K., Hurkman, W. J., & Altenbach, S. B. (2011). Deciphering the complexities of the wheat flour proteome using quantitative two-dimensional electrophoresis, three proteases and tandem mass spectrometry. *Proteome Science*, *9*(1), 10.

Feeney, K., Wellner, N., Gilbert, S., Halford, N., Tatham, A. S., Shewry, P. R., & Belton, P. S. (2003). Molecular structures and interactions of repetitive peptides based on wheat glutenin subunits depend on chain length. *Biopolymers*, *72*(2), 123–131.

Field, J. M., Tatham, A. S., & Shewry, P. R. (1987). The structure of a high-Mr subunit of durum-wheat (Triticum durum) gluten. *Biochemical Journal*, *247*(1), 215.

Fukuyama, Y., Iwamoto, S., & Tanaka, K. (2006). Rapid sequencing and disulfide mapping of peptides containing disulfide bonds by using 1,5-diaminonaphthalene as a reductive matrix. *Journal of Mass Spectrometry*, *41*(2), 191–201.

Gao, L., Ma, W., Chen, J., Wang, K., Li, J., Wang, S., et al. (2010). Characterization and Comparative Analysis of Wheat High Molecular Weight Glutenin Subunits by SDS-PAGE, RP-HPLC, HPCE, and MALDI-TOF-MS. *Journal Of Agricultural And Food Chemistry*, *58*(5), 2777–2786.

Gianibelli, M., Larroque, O., MacRitchie, F., & Wrigley, C. (2001). Biochemical, genetic, and molecular characterization of wheat glutenin and its component subunits. *Cereal Chemistry*, *78*(6), 635–646.

Goddard, T. D., & Kneller, D. G. (2008, May 30). SPARKY 3. San Francisco.

Gu, Y. Q., Crossman, C., Kong, X., Luo, M., You, F. M., Coleman-Derr, D., et al. (2004). Genomic organization of the complex alpha-gliadin gene loci in wheat. *Theoretical and Applied Genetics*, *109*(3), 648–657.

Gupta, R. B., Khan, K., & MacRitchie, F. (1993). Biochemical Basis of Flour Properties in Bread Wheats. I. Effects of Variation in the Quantity and Size Distribution of Polymeric Protein. *Journal of Cereal Science*, *18*(1), 23–41.

Hamaker, B. R., Kirleis, A. W., Butler, L. G., Axtell, J. D., & Mertz, E. T. (1987). Improving the in vitro protein digestibility of sorghum with reducing agents. *Proceedings of the National Academy of Sciences of the United States of America*, *84*(3), 626–628.

Hamaker, B. R., Mohamed, A. A., Habben, J. E., Huang, C. P., & Larkins, B. A. (1995). Efficient Procedure for Extracting Maize and Sorghum Kernel Proteins Reveals Higher Prolamin Contents Than the Conventional Method. *Cereal Chemistry*, *72*(6), 583–588.

Hutchinson, E. G., & Thornton, J. M. (1994). A revised set of potentials for beta-turn formation in proteins. *Protein Science*, *3*(12), 2207–2216.

Izquierdo, L., & Godwin, I. D. (2005). Molecular Characterization of a Novel Methionine-Rich δ-Kafirin Seed Storage Protein Gene in Sorghum (Sorghum bicolorL.). *Cereal Chemistry*, *82*(6), 706–710.

Johns, C. O., & Brewster, J. F. (1916). Kafirin an Alcohol-Soluble Protein from Kafir, Andropogon Sorghum. *The Journal of Biological Chemistry*, *28*, 59–65.

Khan, K., & Shewry, P. R. (2009). *Wheat*. American Association of Cereal Chemists.

Klibanov, A. M. (2001). Improving enzymes by using them in organic solvents. *Nature*, *409*, 241–246.

Krishnan, H. B., White, J. A., & Pueppke, S. G. (1989). Immunocytochemical analysis of protein body formation in seeds of Sorghum bicolor. *Canadian Journal of Botany*, *67*(10), 2850–2856.

Kumar, T., Dweikat, I., Sato, S., Ge, Z., Nersesian, N., Chen, H., et al. (2012). Modulation of kernel storage proteins in grain sorghum (Sorghum bicolor (L.) Moench). *Plant Biotechnology Journal*, *10*(5), 533–544.

Kurien, P. P., Narayanarao, M., Swaminathan, M., & Subrahmanyan, V. (1960). The metabolism of nitrogen, calcium and phosphorus in undernourished children. *British Journal of Nutrition*, *14*(03), 339–345.

Larkins, B. A., & Hurkman, W. J. (1978). Synthesis and deposition of zein in protein bodies of maize endosperm. *Plant Physiology*, *62*(2), 256–263.

Laskowski, R., Rullmann, J. A., MacArthur, M., Kaptein, R., & Thornton, J. (1996). AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *Journal of Biomolecular NMR*, *8*(4).

Lending, C. R., & Larkins, B. A. (1989). Changes in the zein composition of protein bodies during maize endosperm development. *Plant Cell*, *1*(10), 1011–1023.

Li, W., Dobraszczyk, B. J., Dias, A., & Gil, A. M. (2006). Polymer Conformation Structure of Wheat Proteins and Gluten Subfractions Revealed by ATR-FTIR. *Cereal Chemistry*, *83*(4), 407–410.

Lovell, S. C., Davis, I. W., Arendall, W. B., III, de Bakker, P. I. W., Word, J. M., Prisant, M. G., et al. (2003). Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. *Proteins-Structure Function And Bioinformatics*, *50*(3), 437–450.

Mackintosh, S. H., Meade, S. J., Healy, J. P., Sutton, K. H., Larsen, N. G., Squires, A. M., & Gerrard, J. A. (2008). Wheat glutenin proteins assemble into a nanostructure with unusual structural features. *Journal of Cereal Science*, *49*(1), 157–162.

MacLean, W. C., Lopez de Romaña, G., Placko, R. P., & Graham, G. G. (1981). Protein quality and digestibility of sorghum in preschool children: balance studies and plasma free amino acids. *The Journal of nutrition*, *111*(11), 1928–1936.

Matsushima, N., Danno, G., Sasaki, N., & Izumi, Y. (1992). Small-angle X-ray scattering study by synchrotron orbital radiation reveals that high molecular weight subunit of glutenin is a very anisotropic molecule. *Biochemical and biophysical research communications*, *186*(2), 1057–1064.

Mazhar, H., & Chandrashekar, A. (1993). Differences in Kafirin Composition During Endosperm Development and Germination in Sorghum Cultivars of Varying Hardness. *Cereal Chemistry*, *70*(6), 667–671.

McIntire, T. M., Lew, E. J. L., Adalsteins, A. E., Blechl, A., Anderson, O. D., Brant, D. A., & Kasarda, D. D. (2005). Atomic force microscopy of a hybrid high-molecular-weight glutenin subunit from a transgenic hexaploid wheat. *Biopolymers*, *78*(2), 53–61.

McIntosh, R. A., Yamazaki, Y., & Devos, K. M. (2003). Catalogue of gene symbols for wheat. *Tenth International Wheat Genetics Symposium.*

McLachlan, A. D. (1971). Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551 *Journal of molecular biology*, *61*(2), 409–424.

Mertz, E. T., Hassen, M. M., Cairns-Whittern, C., Kirleis, A. W., Tu, L., & Axtell, J. D. (1984). Pepsin digestibility of proteins in sorghum and other major cereals. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(1), 1–2.

Miles, M. J., Carr, H. J., McMaster, T. C., I'Anson, K. J., Belton, P. S., Morris, V. J., et al. (1991). Scanning tunneling microscopy of a wheat seed storage protein reveals details of an unusual supersecondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(1), 68–71.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443–453.

Nour, El, I. N. A., Peruffo, A. D., & Curioni, A. (1998). Characterisation of Sorghum Kafirins in Relation to their Cross-linking Behaviour. *Journal of Cereal Science*, *28*(2), 197–207.

Oria, M. P., Hamaker, B. R., & Shull, J. M. (1995a). Resistance of Sorghum .alpha.-, .beta.-, and .gamma.-Kafirins to Pepsin Digestion. *Journal Of Agricultural And Food Chemistry*, *43*(8), 2148–2153.

Oria, M. P., Hamaker, B. R., Axtell, J. D., & Huang, C. P. (2000). A highly digestible sorghum mutant cultivar exhibits a unique folded structure of endosperm protein bodies. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(10), 5065–5070.

Oria, M. P., Hamaker, B., & Shull, J. M. (1995b). In-Vitro Protein Digestibility of Developing and Mature Sorghum Grain in Relation to Alpha-Kafirin, Beta-Kafirin, and Gamma-Kafirin Disulfide Cross-Linking. *Journal of Cereal Science*, *22*(1), 85–93.

Osborne, T. B. (1909). *The vegetable proteins*.

Parchment, O., Shewry, P. R., Tatham, A. S., & Osguthorpe, D. (2001). Molecular modeling of unusual spiral structure in elastomeric wheat seed protein. *Cereal Chemistry*, *78*(6), 658–662.

Park, S.-H., & Bean, S. R. (2003). Investigation and Optimization of the Factors Influencing Sorghum Protein Extraction. *Journal Of Agricultural And Food Chemistry*, *51*(24), 7050–7054.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The Sorghum bicolor genome and the diversification of grasses. *Nature*, *457*(7229), 551–556.

Payne, P. (1987). Genetics of wheat storage proteins and the effect of allelic variation on bread-making quality. *Annual Review of Plant Physiology*, *38*(1), 141–153.

Payne, P., Holt, L., Jackson, E., Law, C., & Damania, A. (1984). Wheat Storage Proteins: Their Genetics and Their Potential for Manipulation by Plant Breeding [and Discussion]. *Philosophical Transactions of the Royal Society of London. Series BBiological Sciences*, *304*(1120), 359–371.

Pirozi, M. R., Margiotta, B., Lafiandra, D., & MacRitchie, F. (2008). Composition of polymeric proteins and bread-making quality of wheat lines with allelic HMW-GS differing in number of cysteines. *Journal of Cereal Science*, *48*(1), 117–122.

Pomeranz, Y. (1988). *Wheat*. American Association of Cereal Chemists.

Qi, P.-F., Wei, Y.-M., Ouellet, T., Chen, Q., Tan, X., & Zheng, Y.-L. (2009). The gamma-gliadin multigene family in common wheat (Triticum aestivum) and its closely related species. *BMC genomics*, *10*, 168.

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics*, *16*(6), 276–277.

Roberts, G., & Lian, L.-Y. (2011). *Protein NMR Spectroscopy*. John Wiley & Sons.

Sabelli, P. A., & Larkins, B. A. (2009). The Development of Endosperm in Grasses. *Plant Physiology*, *149*(1), 14–26.

Sankoff, D., & Cedergren, R. J. (1983). Time warps, string edits, and macromolecules: The theory and practice of sequence comparison. *Reading: Addison-Wesley Publication, 1983, edited by Sankoff, David; Kruskal, Joseph B.*

Shan, X., Clayshulte, S. R., Haley, S. D., & Byrne, P. F. (2007). Variation for glutenin and waxy alleles in the US hard winter wheat germplasm. *Journal of Cereal Science*, *45*(2), 199–208.

Shewry, P. R. (1990). The prolamin storage proteins of cereal seeds: structure and evolution. *Biochemical Journal*, *267*(1), 1.

Shewry, P. R. (2009). Wheat. *Journal of Experimental Botany*, *60*(6), 1537–1553.

Shewry, P. R., Halford, N. G., & Lafiandra, D. (2003). Genetics of Wheat Gluten Proteins. In *Advances in genetics* (Vol. 49, pp. 111–184). Elsevier.

Shull, J. M., & Watterson, J. (1991). Proposed nomenclature for the alcohol-soluble proteins (kafirins) of Sorghum bicolor (L. Moench) based on molecular weight, solubility, and structure. *Journal Of Agricultural And Food Chemistry*, *39*, 83–87.

Shull, J. M., Watterson, J. J., & Kirleis, A. W. (1992). Purification and immunocytochemical localization of kafirins inSorghum bicolor (L. Moench) endosperm. *Protoplasma*, *171*(1-2), 64–74.

Skylas, D., Van Dyk, D., & Wrigley, C. (2005). Proteomics of wheat grain. *Journal of Cereal Science*, *41*(2), 165–179.

Song, R. (2004). Expression of the sorghum 10-member kafirin gene cluster in maize endosperm. *Nucleic Acids Research*, *32*(22), e189–e189.

Sreerama, N., & Woody, R. W. (2000). Estimation of Protein Secondary Structure from Circular Dichroism Spectra: Comparison of CONTIN, SELCON, and CDSSTR Methods with an Expanded Reference Set. *Analytical Biochemistry*, *287*(2), 252–260.

Tatham, A. S., & Shewry, P. R. (1984). Wheat gluten elasticity: a similar molecular basis to elastin? *FEBS letters*, *177*(2), 205–208.

Tatham, A. S., Drake, A. F., & Shewry, P. R. (1989). Conformational studies of a synthetic peptide corresponding to the repeat motif of C hordein. *The Biochemical journal*, *259*(2), 471–476.

Tatham, A. S., Drake, A. F., & Shewry, P. R. (1990). Conformational studies of synthetic peptides corresponding to the repetitive regions of the high molecular weight (HMW) glutenin subunits of wheat. *Journal of Cereal Science*, *11*(3), 189–200.

Tatham, A. S., Gilbert, S. M., Fido, R. J., & Shewry, P. R. (2000). Extraction, Separation, and Purification of Wheat Gluten Proteins and Related Proteins of Barley, Rye, and Oats. In *Celiac Disease* (Vol. 41, pp. 055–073). New Jersey: Humana Press.

Tatham, A. S., Miflin, B., & Shewry, P. R. (1985). The beta-turn conformation in wheat gluten proteins: Relationship to gluten elasticity. *Cereal Chemistry*, *62*(405-412).

Taylor, J. R. N., Novellie, L., & Liebenberg, N. (1984). Sorghum protein body composition and ultrastructure. *Cereal Chemistry*, *61*(1), 69–73.

van Dijk, A. A., van Wijk, L. L., Van Swieten, E., Robillard, G. T., De Boef, E., Bekkers, A., & Hamer, R. J. (1997a). Structure characterization of the central repetitive domain of high molecular weight gluten proteins. II. Characterization in solution and in the dry state. *Protein Science*, *6*(3), 649–656.

van Dijk, A. A., van Wijk, L. L., van Vliet, A., Haris, P., Van Swieten, E., Tesser, G. I., & Robillard, G. T. (1997b). Structure characterization of the central repetitive domain of high molecular weight gluten proteins. I. Model studies using cyclic and linear peptides. *Protein science*, *6*(3), 637–648.

Watterson, J. J., Shull, J. M., & Kirleis, A. W. (1993). Quantitation of α-, β-, and γ-kafirins in vitreous and opaque endosperm of Sorghum bicolor. *Cereal Chemistry*.

Weaver, C. A., Hamaker, B. R., & Axtell, J. D. (1998). Discovery of Grain Sorghum Germ Plasm with High Uncooked and Cooked In Vitro Protein Digestibilities 1. *Cereal Chemistry*, *75*(5), 665–670.

Wieser, H. (2007). Chemistry of gluten proteins. *Food microbiology*, *24*(2), 115–119.

Winn, J. A., Mason, R. E., Robbins, A. L., Rooney, W. L., & Hays, D. B. (2009). QTL mapping of a high protein digestibility trait in Sorghum bicolor. *International journal of plant genomics*, *2009*, 471853.

Wrigley, C. W., Békés, F., & Bushuk, W. (2006). *Gliadin and glutenin*. Amer Assn of Cereal Chemists.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC bioinformatics*.

Abbreviations Used

i Sodium Dodecyl Sulfate

ii Tris(2-carboxyethyl)phosphine hydrochloride

iii Dithiothreitol

iv Reversed Phase- High Performance Liquid Chromatography

v Matrix Assisted Laser Desorption Ionization- Time Of Flight/ Time Of Flight Mass Spectrometer

vi 3,5-Dimethoxy-4-hydroxycinnamic acid

vii α -cyano-4-hydroxycinnamic acid

viii 2,5-Dihydroxybenzoic acid

ix 1,5-Diaminonapthlene

[x] Circular Dichroism

xi Sinapinic acid

xii Corn Hageman Factor Inhibitor

[xiii] Nuclear Magnetic Resonance

[xiv] Trifluoroethanol

[xv] Total Correlation Spectroscopy

[xvi] Nuclear Overhauser Effect Spectroscopy

[xvii] Correlation Spectroscopy

[xviii] Heteronuclear Single Quantum Coherence Spectroscopy

[xix] Circular Dichroism