# EFFECTS OF GENOTYPE, ENVIRONMENT AND GENOTYPE BY ENVIRONMENT INTERACTION ON SOYBEAN PROTEIN AND AMINO ACID CONTENT USING NEAR-INFRARED REFLECTANCE SPECTROSCOPY

**By**

**Da Shi**

A thesis submitted to The Faculty of Graduate Studies of

The University of Manitoba

In partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Food and Human Nutritional Sciences

University of Manitoba

Winnipeg, Manitoba, Canada

## ABSTRACT

Soybean [*Glycine max (L.) Merr.*] protein and amino acid contents are important for soybean quality assessment. It is desirable to cultivate high protein soybean to meet the nutritional requirements for both animal and human. The current study evaluates the effects of genotype, environment and genotype × environment interactions on soybean protein and amino acid concentrations. A relatively new method named near-infrared (NIR) spectroscopy was used to measure crude protein and amino acid contents in soybean grain. The predictive ability of NIR calibration models and factors that influence the performance of the NIR system were estimated. The effects of genotype, environment and genotype × environment interactions on soybean protein and amino acid contents were significant ($P < 0.05$). Among those factors, genotype explained the main part of variation for all traits. Protein and amino acids responded differently to various environments, but the favorable environments for soybean protein and amino acids accumulation were still unclear. The NIR calibration models for crude protein and most amino acids except for cysteine (Cys), methionine (Met) and tryptophan (Trp), showed acceptable coefficients of determination ($R^2_c$ = 0.605-0.952), while models for Cys, Met and Trp might be less accurate ($R^2_c$ = 0.498-0.667 for Cys, $R^2_c$ = 0.482-0.615 for Met and $R^2_c$ = 0.406-0.481 for Trp). The grinding process and lipid extraction improved the $R^2_c$ values of NIR calibration models for crude protein and most amino acid predictions. Strong correlations (R = 0.85-0.97) and no significant difference ($P > 0.05$) were found between crude protein and amino acid contents predicted by two different types of NIR spectroscopy instrument: PerkinElmer DA 7250 and PerkinElmer FT 9700. This

work has the potential to develop a faster way for measuring crude protein and amino acid contents in soybean grain and help soybean farmers to select the optimal soybean varieties based on different purposes.

## ACKNOWLEDGMENTS

First and foremost, I would extremely appreciate Dr. James House, the leader and my advisor in this project. I am grateful that Dr. House felt my passion in food and human nutritional sciences and provided me an unique opportunity to work and study alongside his research team. Through his patient guidance, encouragement and support, I would be able to complete this meaningful project.

I would also like to show my appreciation to Manitoba Pulse and Soybean Growers for supporting this research via direct funding and by the provision of soybean samples. I wish to acknowledge PerkinElmer company for providing us the indispensable near-infrared spectroscopy instruments used in this project.

Also, I would like to acknowledge my committee: Dr. Curtis Rempel and Dr. Yvonne Lawley for taking the time to give me useful suggestions during my master's study and during the review of my thesis.

I am very thankful for the assistance from Jason Neufeld, who taught me the skills related to analytical methods and helped with the data collection, and Shusheng Zhao, who assisted me in completing the sample analysis. Likewise, I am grateful to my other peer lab members: Adam Franczyk, Shengnan Li and Zhongyang Wan, and summer students: Cameron Dubois and Jennifer Nguyen.

Specially, I am greatly appreciative of the help from Jiayi Hang. We always discussed the principles of near-infrared, data processing and result interpretations. This made me learn a lot that I didn't grasp myself.

Finally, my deep thanks go to my beloved family for their unselfish love and confidence in my whole life.

## DEDICATION

This thesis is dedicated to every scientific worker, since we are all shining stars in a galaxy named science.

# TABLE OF CONTENT

**ABSTRACT**..... i
**ACKNOWLEDGMENTS**..... iii
**DEDICATION**..... iv
**TABLE OF CONTENT**..... v
**LIST OF TABLES**..... vii
**LIST OF FIGURES**..... ix
**LIST OF ABBREVIATIONS**..... x
**Chapter 1: Literature Review**..... 1
**1.1 Introduction**..... 1
**1.2 Soybean**..... 2
*1.2.1 Background*..... 2
*1.2.2 Chemical composition in soybean*..... 3
*1.2.3 Applications of soybean*..... 3
**1.3 Soy protein**..... 4
*1.3.1 Soy protein composition*..... 4
*1.3.2 Soy protein quality*..... 5
*1.3.3 Health benefits of soy protein*..... 10
**1.4 Soy amino acid**..... 11
*1.4.1 Soy amino acid composition*..... 11
*1.4.2 Amino acid composition and protein quality*..... 14
*1.4.3 Limiting amino acid in soybean*..... 16
1.5 **Factors affecting crude protein and amino acids concentrations in soybean**..... 17
*1.5.1 Genotype effect*..... 17
*1.5.2 Environmental Factors*..... 18
**1.6 Traditional analysis methods for protein and amino acid**..... 22
*1.6.1 Protein analysis*..... 23
*1.6.2 Amino acid analysis*..... 26
**1.7 Near-infrared technology**..... 29
*1.7.1 Background*..... 29
*1.7.2 Composition of different NIR spectroscopy instruments*..... 32
*1.7.3 Calibration of NIR system*..... 33
*1.7.4 Application of NIR for measuring crude protein and amino acid concentrations*..36
**1.8 Summary**..... 39
**1.9 Hypothesis and Objectives**..... 40
*1.9.1 Hypothesis*..... 40
*1.9.2 Objectives*..... 40
**Chapter 2 Estimation of Crude Protein and Amino Acid Contents in Whole, Ground and Defatted Ground Soybean by Different Types of Near-infrared (NIR) Reflectance Spectroscopy**..... 41
**2.1 Abstract**..... 41
**2.2 Introduction**..... 42
**2.3 Materials and methods**..... 44
*2.3.1 Sample acquisition and preparation*..... 44

*2.3.2 Protein and amino acid analysis* ..... 45
*2.3.3 NIR analysis* ..... 47
**2.4 Results and Discussion** ..... 48
*2.4.1Raw data for NIR calibration* ..... 48
*2.4.2 NIR spectral assignment* ..... 49
*2.4.3 NIR calibration and cross-validation* ..... 49
*2.4.4 Factors influencing NIR predictive ability* ..... 53
*2.4.5 Performance of different NIR spectroscopy instruments* ..... 54
**2.5 Conclusion** ..... 56
**Chapter 3 Effects of Genotype, Environment and Their Interaction on Protein and Amino Acid contents in Soybean** ..... 67
**3.1 Abstract** ..... 67
**3.2 Introduction** ..... 68
**3.3 Materials and Methods** ..... 70
*3.3.1 Plant materials* ..... 70
*3.3.2 Protein and amino acid analysis* ..... 71
*3.3.3 Statistical analysis* ..... 72
**3.4 Results and Discussion** ..... 73
*3.4.1 Analysis of Variance* ..... 73
*3.4.2 Stability Analysis* ..... 75
*3.4.3 Environment effects on soybean protein and amino acid contents* ..... 76
*3.4.4 Critical amino acid value and protein* ..... 77
**3.5 Conclusion** ..... 79
**Chapter 4 General Discussion** ..... 89
**Chapter 5 Future Directions** ..... 95
**Reference** ..... 96
**Appendix I** ..... 116
**Appendix II** ..... 117

## LIST OF TABLES

Table 1.1. Protein Digestibility Corrected Amino Acid Score values for various of plant and animal-based food proteins.....7

Table 1.2. Protein Efficiency Ratio values for various of plant and animal-based food proteins.....8

Table 1.3. Mean amino acid content (on dry weight basis%) of soybean from Kovalenko et al. (2006).....12

Table 1.4. Essential amino acids of different food and FAO/WHO recommended human amino acid requirements.....14

Table 1.5. Comparison of traditional analysis methods for protein and amino acid and NIR spectroscopy.....27

Table 1.6. Near infrared calibration statistics for soybean amino acid measurements from Kovalenko et al. (2006).....33

Table 2.1. Statistics for reference amino acid and crude protein concentrations (dry weight content%) of soybean used for NIR calibration.....49

Table 2.2. Calibration and cross-validation statistics in PerkinElmer DA 7250 NIR models for crude protein and amino acids measurement (dry basis%) in whole soybean.....50

Table 2.3. Calibration and cross-validation statistics in PerkinElmer DA 7250 NIR models for crude protein and amino acids measurement (as is basis%) in whole soybean.....51

Table 2.4. Calibration and cross-validation statistics in PerkinElmer DA 7250 NIR

models for crude protein and amino acids measurement (dry basis%) in ground soybean……..…...52
Table 2.5. Calibration and cross-validation statistics in PerkinElmer DA 7250 NIR models for crude protein and amino acids measurement (dry basis%) in defatted ground soybean.....53
Table 2.6. Calibration and cross-validation statistics in PerkinElmer FT 9700 NIR models for crude protein and amino acids measurement (dry basis%) in ground and defatted soybean.....54
Table 2.7. Statistics of agreement assessment between DA 7250 and FT 9700 NIR spectroscopy instruments.....55
Table 3.1. Temperature, rainfall, and solar radiation of the eight soybean growing environments in Manitoba, Canada during May to September.....68
Table 3.2. Mean squares from analysis of variance of protein and amino acids for soybean genotypes grown in four locations in Manitoba, Canada during 2018 and 2019…….....69
Table 3.3. Estimates of variance components for protein and amino acids of soybean genotypes in four locations in Manitoba, Canada during 2018 and 2019.....70
Table 3.4. Mean crude protein content, regression coefficient, standard error of coefficient and deviation from regression for the 23 soybean genotypes tested across eight environments in Manitoba, Canada.....71
Table 3.5. Mean crude protein and amino acid contents of soybean grown in distinct environments in Manitoba, Canada.…….72
Table 3.6. Estimates of Pearson correlation coefficients characterizing the relationship

between various soybean biochemical parameters across different environments in Manitoba, Canada…….....73

Table 3.7. Statistics for linear regression between Critical Amino Acid Value% and protein content for different environments.....74

## LIST OF FIGURES

Figure 2.1. NIR spectra for whole soybean from calibration set scanned by PerkinElmer DA 7250……...……...56

Figure 2.2. Standard Normal Variate and de-trending treated NIR spectra for whole soybean from calibration set scanned by PerkinElmer DA 7250……..……...57

Figure 2.3. $R^2$ for linear regression of amino acids to crude protein versus $R^2c$ for NIR calibration on whole soybean samples by PerkinElmer DA 7250.....58

Figure 3.1. Linear regression of critical amino acid value% versus crude protein content for soybean samples.....75

**LIST OF ABBREVIATIONS**

**AACC -** American Association of Cereal Chemists

**AAS** - Amino Acid Score

**ALA -** Alanine

**ANN -** Artificial Neural Networks

**ANOVA** - Analysis of Variance

**ARG -** Arginine

**AOAC -** Association of Official Agricultural Chemists

**ASP -** Aspartate

$\mathbf{b_i}$ **-** Regression Coefficient

**CAAV -** Critical Amino Acid Value

**CARS** - Competitive Adaptive Reweighted Sampling

**CYS -** Cysteine

**DOSC** - Direct Orthogonal Signal Correction

**DT -** De-trending

**FAO** - Food and Agriculture Organization

**FD -** First Derivative

**GA** - genetic algorithm

**GLU -** Glutamate

**GLY -** Glycine

**HIS -** Histidine

**ISO -** International Organization for Standardization

**ILE -** Isoleucine

**ISP** - Isolated Soy Protein

**LED -** Light Emitting Diodes

**LEU -** Leucine

**LYS -** Lysine

**MC** - mean centering

**MET -** Methionine

**MIR -** Mid Infrared

**MLR -** Multiple Linear Regression

**MSC -** Multiplicative Scatter Correction

**NIR -** Near Infrared

**NPCF** - nitrogen-to-protein conversion factor

**PCR -** principal component regression

**PDCAAS** - Protein Digestibility Corrected Amino Acid Score

**PER** - Protein Efficiency Ratio

**Phe -** Phenylalanine

**PLS -** partial least squares

**PRO -** Proline

**$R^2$ -** Coefficient of Determination

**REML** - Restricted Maximum Likelihood

**RER** - Range Error Ratio

**RPD -** Ratio of Performance of Deviation

**$S^2_{di}$** - Deviation from Regression

**SD** - Standard Deviation

**SECV -** Standard Error of Cross Validation

**SEP -** Standard Error of Prediction

**SER -** Serine

**SMLR** - Stepwise Multiple Linear Regression

**SNV -** Standard Normal Variate

**SPA** - Sive Projections Algorithm

**SPC** - Soy Protein Concentrate

**SVM -** Support Vector Machines

**THR -** Threonine

**TRP -** Tryptophan

**TYR -** Tyrosine

**UV -** Ultraviolet

**UVE** - Uninformative Variable Elimination

**VAL -** Valine

**WHO -** World Health Organization

# Chapter 1: Literature Review

## 1.1 Introduction

Current trends show that an increasing number of people from wealthy countries are trying to consume more plant-based foods. In a 2017 global consumer study, 39% of Americans and 43% of Canadians were attempting to put more plant-based foods into their diets (Doris, 2018). Also, the new Canada's Food Guide, published by the government of Canada, encourages people to consume more fruits and vegetables and to consume plant-based proteins, such as pulses and nuts, more often (Government of Canada, 2019). In addition, according to Young and Pellett (1994), most of the world's supply of edible protein (around 65%) came from plants.

Although meat is an excellent source of protein, issues regarding contamination of meat products with pathogenic bacteria are a concern in the minds of some consumers. Foodborne illness bacteria such as *Salmonella*, *E. coli* and *Listeria* are sometimes found in meat products, which may cause foodborne illness and food recalls (Canadian Food Inspection Agency, 2017). Antibiotic resistance can also be a reason why consumers seek to reduce meat consumption, as its emergence may be linked to increased antibiotic usage (WHO, 2017). Furthermore, the linkage of animal agriculture with greenhouse gas emissions and environmental impacts may also contribute to these consumer trends (Specht, 2018).

There are many reported benefits of consuming plant protein more often. For example, soy protein has been reported to reduce the risk of cardiovascular disease by lowering triglyceride levels in liver or blood, reducing total and LDL cholesterol levels, and increasing HDL cholesterol and the ratio of HDL/LDL cholesterol (Xiao,

2008). In addition, Kingman et al. (1993) reported putative blood pressure lowering effects of a lentil protein hydrolysate, which was related to cardiovascular health. Moreover, plant lectin, a unique protein group, was found to exert anticancer properties in human studies (De Mejía & Prisecaru, 2005).

Therefore, plant-based foods and proteins will likely continued to be important trends in the future as more and more scientists and industries are working to enhance plant-based food options for nutritional requirements of consumers. One such plant protein source is soybean.

## 1.2 Soybean

### *1.2.1 Background*

One crop that is important as an established source of plant-based protein, both for animal and human foods alike, is soybean. Soybean is, on worldwide basis, the most important leguminous plant whose grains contain high levels of oil and protein (Wee et al., 2018), and it is one of East Asia's important native crops (Qin et al., 2014). In recent times, an increasing number of food industries in different regions, including the Americas and Europe, have used soybean to make food products because of its nutritional and phytochemical characteristics (Qin et al., 2014). For instance, in Canada, soybean is now the 3rd largest field crop, on the basis of on-farm receipts. In 2017, about 29% of the total Canadian crop production was soybean (on weight basis) (Statistic Canada, 2017). In America, soybean occupied the largest crop area in 2018 (33% of total crop area), which had the similar acreage as corn (American Soybean Association, 2019). Soybean grain plays an important role in agricultural production,

human food security and international trade (Jiang, 2020). In 2019, the global production of soybean was 334M tonnes, and both US and Canada were among the top 10 soybean producers, with the production of 97M tonnes for America and 6M tonnes for Canada (FAO stat, 2019)

### *1.2.2 Chemical composition in soybean*

Soybean contain protein (40–50% in dry grain), lipids (20–30% in dry grain), and carbohydrates (26-30% in dry grain) (Gibbs et al., 2004). Soybean protein includes all essential amino acids and tends to supply these at a lower cost relative to other high-quality protein sources (Carrera et al., 2011). Additionally, soybean grains contain many various kinds of phytochemicals, such as isoflavones, tocopherols, saponins, and anthocyanins (Gibbs et al., 2004; Malenčić et al., 2007). These components in soybean grains determine the nutritional value and usage of soybean (Gibbs et al., 2004; Jiang, 2020).

### *1.2.3 Applications of soybean*

Traditionally, due to their high lipid content, the primary aim of growing soybean was to extract oil for food and feed, as well as industrial applications. Evidence shows that soybean oil is a renewable raw material that can be used in making a wide range of industrial products, including inks, plasticizers, paints and biodiesel (Cahoon, 2003; Jiang, 2020).

Additionally, their high protein content makes soybean a good material for feeding animals and for use in human food applications. Soybean can be a dietary

protein source for feeding Nile tilapia (Wee & Shu, 1989), there is a long history of its use in feeding livestock and poultry (Dei, 2011). As for human food, soybean can be used to produce tofu and soymilk, which are good sources of protein (Lusas & Riaz, 1995).

Other soy-based bioactive compounds, such as isoflavones, have been found to possess anticarcinogenic properties (Wang & Murphy, 1994), via their putative activity as antiestrogens (Adlercreutz et al., 1986), antioxidants (Naim et al., 1976) and tyrosine-protein kinase inhibitors (Akiyama et al., 1987). As such, soybean have tremendous potential to serve as a functional food/health food.

## 1.3 Soy protein

### *1.3.1 Soy protein composition*

Protein is the major component in dry soybean grains. Soy protein is mainly composed of 2 storage globulins (85-90%): 11S glycinin and 7S β-conglycinin (Gibbs et al., 2004; Torres, Torre-Villalvazo & Tovar, 2006). Other soy proteins include lipoxygenases, lectins, trypsin inhibitors and α-amylases (Liu, 1997). Glycinin has two subunits, namely acidic and basic subunits (Xiao, 2008), which don't include any carbohydrate (Gibbs et al., 2004). On the other side, β-conglycinin has α, α', and β subunits (Xiao, 2008), and it is a glycoprotein that contains around 4% carbohydrate (mostly mannose moieties) (Gibbs et al., 2004). Additionally, some other soy proteins may undergo phosphorylation and/or glycosylation (Gibbs et al., 2004). Glycinin and β-conglycinin were reported to have different properties in previous studies. Saio et al. (1975) reported that the heat stability of these two proteins are different. They also

found that these two proteins had different gel strengths (Saio et al., 1969). The abilities of these two proteins to react with metal irons are different as well (Briggs & Wolf, 1957). Furthermore, these two proteins have different abilities to bind some soy off-flavor compounds (Damodaran & Kinsella, 1981). As such, the relative distribution of these proteins within soy can ultimately affect its functional and nutritional properties. This may have implications for the quality of the soy protein as well.

### *1.3.2 Soy protein quality*

Protein quality measures the bioavailability of protein which is the percentage of the constituent amino acids that are absorbed and ultimately utilized by the human body for productive purposes (Leser, 2013). The evaluation of protein quality is based on the amino acid composition, the digestibility of the protein and the bioavailability of individual amino acids (Boye et al., 2012). Therefore, due to the fact that soy protein contains all amino acids essential to human nutrition, soy products almost have the same protein quality as animal-based protein sources but with less saturated fat and no cholesterol (Young, 1991).

Since 1951, the Food and Agriculture Organization (FAO) positioned methods for the evaluation of protein quality from different staple protein sources to satisfy human protein requirements (Leser, 2013). The Protein Digestibility Corrected Amino Acid Score (PDCAAS) was the method recommended by Food and Agriculture Organization/ World Health Organization (FAO/WHO) in 1991 as the best method available at that time to describe dietary protein quality (FAO/WHO, 1991). This is

the method used to assess protein quality, for the purposes of protein content claim substantiation, in the United States. The PDCAAS method is based on comparing the amino acid composition of a protein source to a reference value for 2-5 year old school children thus allowing the determination of an amino acid score (AAS),which reflects the lowest ratio of the essential amino acid provided by protein source in relation to the reference pattern (Hughes et al., 2011). The PDCAAS value is then calculated using the AAS by multiplying it by the true digestibility of the protein (Hughes et al., 2011), which is determined using a rat bioassay to assess thecproportion of food-based nitrogen that is absorbed (Wu et al., 1995). A PDCAAS value of 1.00 or 100% means that the protein source satisfies the needs for all of the essential amino acids for children aged two and above and adults when nutritionally appropriate amounts of protein are consumed (Hughes et al., 2011). Table 1.1 shows PDCAAS values for various plant and animal-based food proteins (FAO/WHO, 1991; Ahrens et al., 2005). Most plant proteins, such as kidney beans, rice and wheat, have lower PDCAAS values than animal proteins, but soybean has been reported to have similar PDCAAS to milk, egg white and beef. Isolated soy protein (ISP) was reported to have PDCAAS values ranged from 0.92 to 1.00, and the value for soy protein concentrate (SPC) was 0.99 (FAO/WHO, 1991; Sarwar, 1997).

In Canada, the protein efficiency ratio (PER), as applied within the protein rating system, is used for characterizing the protein quality in a given food, and is the approved system for protein content claim substantiation (Government of Canada, 2016). The PER method is a bioassay method involving rats with body weight gain recorded over a 28 day feeding periods (Government of Canada, 1981). Weight gain

and feed consumption are measured to calculate the PER value for a protein source, and the PER of casein (determined to be a high-quality protein source) is measured concurrently as a standardized approach (Government of Canada, 1981). The PER value is then used to calculate a protein rating to substantiate protein content claims on foods including “good source” of protein (protein rating ≥ 20) or “excellent source” of protein (protein rating ≥ 40) (Government of Canada, 1981). Table 1.2 shows PER values for various plant and animal-based food proteins (Canadian Food Inspection Agency, 2016). Similar to PDCAAS, the PER values for animal proteins are higher than that of plant proteins. However, the PER value for soybean is relatively high in plant proteins and close to that of animal proteins. As of December, 2020, Health Canada will now accept PER values calculated from PDCAAS (PDCAAS x 2.5) but protein rating values must still be calculated.

Table 1.1. Protein Digestibility Corrected Amino Acid Score values for various of plant and animal-based food proteins (FAO/WHO, 1991; Ahrens et al., 2005)

| Protein Source | PDCAAS |
|---|---|
| Milk (casein) | 1.00 |
| Egg white | 1.00 |
| Isolated soy protein | 1.00 |
| Beef | 0.92 |
| Kidney beans | 0.68 |
| Pinto beans | 0.63 |
| Rice | 0.53 |
| Whole wheat | 0.40 |
| Almonds | 0.23 |

Table 1.2. Protein Efficiency Ratio values for various of plant and animal-based food proteins (Canadian Food Inspection Agency, 2016)

| Protein Source | PER |
|---|---|
| Milk (casein) | 2.50 |
| Egg white | 3.00 |
| Beef | 2.70 |
| Soybean, heated | 2.30 |
| Pinto beans | 1.64 |
| Kidney beans | 1.55 |
| Rice | 1.50 |
| Whole wheat | 0.80 |
| Almond | 0.40 |

*1.3.3 Health benefits of soy protein*

Beyond the provision of amino acids for protein synthesis, soy protein has been reported to reduce blood cholesterol in humans. In 1967, a replacement of mixed proteins by ISP, at an intake of 100 g/d, was found to lower the average total cholesterol levels by > 2.59 mmol/L in hypercholesterolemic men (Hodges et al., 1967). This was the first time that people discovered the cholesterol-lowering effect of soy protein. More research was conducted to reveal the relationship between soy protein and blood cholesterol levels in the following years. In 1995, a meta-analysis summarized 38 randomized controlled clinical studies published from 1977 to 1994, and the results showed an average intake of 47 g/d of isolated or textured soy protein reduced total cholesterol by 9.3%, LDL-cholesterol by 12.9%, and triglycerides by 10.5% (Anderson et al., 1995). Furthermore, soy protein received health claims in several countries, such as the United States (in 1999), Japan (in 1996), Korean (in 2005), and Canada (in 2015) (Paul, 2005).

Some research has shown that the intake of soy protein or soy isoflavones reduced markers of cancer development and progression in prostate cells in both patients with prostate cancer (Kumar et al., 2004; Dalais et al., 2004) and men at high risk for developing advanced prostate cancer (Hamilton-Reeves et al., 2007). However, research also suggested the supplementation of soy protein or soy isoflavones couldn't affect the serum total or free prostate-specific antigen (a marker of prostate cancer) in healthy men subjects (Jenkins et al., 2003). Further research is needed to understand the relationship between soy protein or soy isoflavones intake and prostate cancer.

Research has also been conducted to study the potential association of soy protein with menopause symptoms and breast cancer, but the results didn't show a significant relationship (Xiao, 2008).

## 1.4 Soy amino acid

### *1.4.1 Soy amino acid composition*

In nature, there are hundreds of biologicals that can be chemically regarded as amino acids, but only 20 amino acids are incorporated into proteins. Additionally, other amino acids, including ornithine, citrulline, and g-aminobutyrate, play critical metabolic roles within the body (Watford & Wu, 2011). Of the 20 amino acids found in proteins, nine are considered essential or indispensable amino acids, including histidine (His), isoleucine (Ile), leucine (Leu), lysine (Lys), methionine (Met), phenylalanine (Phe), threonine (Thr), tryptophan (Trp) and valine (Val). This designation is due to the fact that the human body cannot synthesize these amino acids or cannot synthesize them in adequate amounts to maintain growth (Watford & Wu, 2011). In addition, tyrosine (Tyr) and cysteine (Cys) can only be made from their essential precursors: phenylalanine (Phe) and methionine (Met), respectively (Watford & Wu, 2011).

Soybean protein contains all of the essential amino acids (Carrera et al., 2011), and at levels that are nutritionally relevant. Kovalenko et al. (2006) tested the crude protein and amino acid concentrations in 673 soybean samples (Table 1.3). Based on their results, Glu and Asp had the highest value, with average contents (dry weight basis) of 7.66% and 4.79%, respectively; sulfur amino acid (Met and Cys) and Trp

were found to be in the lowest levels in soybean. Other research has assessed crude protein and amino acid contents in soybean and showed similar mean values as those reported by Kovalenko (Grieshop & Fahey, 2001; Karr-Lilienthal et al., 2004; Carrera et al., 2011).

Additionally, research by Karr-Lilienthal et al. (2004) assessed the crude protein and amino acid concentrations of soybean from the five leading soybean producing countries: Argentina, Brazil, China, India, and the United States. In the experiment, they picked one sample of high quality soybean from each country, but both a low and a high quality soybean sample from India (Karr-Lilienthal et al., 2004). From their results, they found that Chinese soybean had the highest crude protein content (44.9% in dry weight basis) while Argentinean soybean had the lowest crude protein content (32.6% in dry weight basis), and the amino acid concentrations of soybean varied from the countries as well (Karr-Lilienthal et al., 2004).

Table 1.3. Mean amino acid content (on dry weight basis%) of soybean from Kovalenko et al. (2006)

| Constituent | Min, % DB[a] | Mean, % DB | Max, % DB | SD[b] |
|---|---|---|---|---|
| Ala | 1.46 | 1.79 | 2.13 | 0.128 |
| Arg | 2.21 | 3.17 | 4.44 | 0.397 |
| Asp | 3.59 | 4.79 | 6.03 | 0.470 |
| Cys | 0.52 | 0.70 | 0.86 | 0.063 |
| Glu | 5.36 | 7.66 | 10.18 | 0.868 |
| Gly | 1.38 | 1.77 | 2.15 | 0.143 |
| His | 0.91 | 1.15 | 1.41 | 0.096 |
| Ile | 1.47 | 1.94 | 2.36 | 0.172 |
| Leu | 2.47 | 3.26 | 3.95 | 0.274 |
| Lys | 2.15 | 2.69 | 3.28 | 0.200 |
| Met | 0.48 | 0.61 | 0.76 | 0.048 |
| Phe | 1.54 | 2.16 | 2.68 | 0.207 |
| Pro | 1.46 | 2.04 | 2.65 | 0.225 |
| Ser | 1.43 | 1.92 | 2.58 | 0.209 |
| Thr | 1.29 | 1.62 | 1.96 | 0.117 |
| Trp | 0.32 | 0.50 | 0.66 | 0.064 |
| Tyr | 1.18 | 1.53 | 1.83 | 0.129 |
| Val | 1.51 | 2.06 | 2.54 | 0.186 |
| Cruder protein | 33.82 | 43.16 | 54.61 | 3.960 |

[a]Dry weight basis

[b]Standard deviation

*1.4.2 Amino acid composition and protein quality*

As mentioned before, the protein quality of a food is related to the concentration, digestion, absorption, and utilization of the constituent amino acids (Friedman & Brandon, 2001). The oxidation of amino acids increases sharply when the consumption is higher than the amount required for protein synthesis (Friedman, 1996). However, a high level of amino acid oxidation may also indicate amino acid deficiency, which is the condition of low plasma levels of some specific amino acids . For example, when any one of the indispensable amino acids is deficient for protein synthesis, oxidation of all other indispensable amino acids will increase (Elango et al., 2008). Therefore, plasma amino acid levels and the oxidation of amino acid are used to measure amino acid requirements (Block, 1989; Reichl, 1989). Establishing amino acids requirements is critical for understanding how insufficient amino acid consumption or unbalanced indispensable amino acid intake leads to reduced growth in children or a reduction of lean body mass in adults (Watford & Wu, 2011). Proteins with deficiencies of one or more amino acids are regarded as poor quality protein sources (Friedman & Brandon, 2001).

Amino acid concentrations of some foods (Sarwar et al., 1983) and the amino acid requirement recommended by FAO/WHO (1991) and FAO/WHO/UNU (2007) are listed in Table 1.4. The essential amino acid contents in soybean can generally meet the requirement estimates of most humans (over two years old).

Table 1.4. Essential amino acids of different food and FAO/WHO recommended human amino acid requirements (Sarwar et al., 1983; FAO/WHO, 1991; FAO/WHO/UNU, 2007)

| Amino acid (mg/g protein) | Casein | Beef | Soy protein | Wheat flour | FAO/WHO (1991) | | | | FAO/WHO/UNU (2007) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 1 yr old | 2-5 yr old | 10-12 yr old | Adult | 1-2 yr old | 3-10 yr old | 11-14 yr old | Adult |
| Thr | 46.4 | 42.1 | 38.4 | 29.3 | 43 | 34 | 28 | 9 | 27 | 25 | 25 | 23 |
| Cys+Met | 34.9 | 32.7 | 68.1 | 38.7 | 43 | 25 | 22 | 17 | 26 | 24 | 23 | 22 |
| Val | 68.5 | 45.4 | 49.1 | 42.7 | 55 | 35 | 25 | 13 | 42 | 40 | 40 | 39 |
| Ile | 53.6 | 41.8 | 47.1 | 33.4 | 46 | 28 | 28 | 13 | 31 | 31 | 30 | 30 |
| Leu | 101.6 | 77.5 | 85.1 | 68.5 | 93 | 66 | 44 | 19 | 63 | 61 | 60 | 59 |
| Tyr+Phe | 125.4 | 70.2 | 96.6 | 77.8 | 72 | 63 | 22 | 19 | 46 | 41 | 41 | 38 |
| His | 29.7 | 32.0 | 25.4 | 21.9 | 26 | 19 | 19 | 16 | 18 | 16 | 16 | 15 |
| Lys | 84.4 | 79.4 | 63.4 | 26.6 | 66 | 58 | 44 | 16 | 52 | 48 | 48 | 45 |
| Trp | 13.1 | 9.9 | 11.4 | 11.2 | 17 | 11 | 9 | 5 | 7.4 | 6.6 | 6.5 | 6.0 |

*1.4.3 Limiting amino acid in soybean*

While the overall content of indispensable amino acids in soybean are high for plant proteins, the content of certain amino acids in soybean can limit its nutritional value. The limiting amino acids are indispensable amino acids that are present in the lowest amount relative to requirement in a specific food or feed material. For example, tryptophan and lysine are nutritionally limiting in corn, while cereals in general are limiting in lysine (Friedman & Brandon, 2001).

Methionine is one of the limiting amino acids in soybean. The Met content in soybean meal is about 1.39g/100g crude protein (1.39g/16g N), which is much lower than that of cereal and meat proteins (Friedman & Brandon, 2001). In addition, during food processing and storage, Met and other amino acids may be chemically modified, including the oxidation of Met to methionine sulfoxide and methionine sulfone (Friedman & Brandon, 2001). Also, the protein-bound Met in soybean may be hard to digest, making it more difficult to absorb Met from soybean (Öste, 1991; Gumbmann et al., 1983). It is important to develop soybean with Met-rich proteins to solve this problem (Benito et al., 1999; Kho & Benito, 1988).

Table 1.4 shows that the lysine content of soybean is much higher than that of wheat flour and similar to that of animal proteins. Although the amount of Lys in soybean can match the pattern of requirements for most human life stages, efforts have been placed on developing soybean lines with high Lys content (Falco et al., 1995). A primary reason for these efforts relates to the potential for Maillard reactions to occur between the amide nitrogen of Lys and reduced carbohydrates, leading to the production of fructosyl-lysine and cross-linking of Lys to generate lysinoalanine at

high pH (Friedman & Brandon, 2001). Exposure to carbohydrates can lead to a reduction of Lys content up to 85% in soybean (Mao et al., 1993). This loss of nutritional quality is due to the lysine being unavailable for protein synthesis rather than the destruction of the amino acid (Gumbmann et al., 1983).

## 1.5 **Factors affecting crude protein and amino acids concentrations in soybean**

### *1.5.1 Genotype effect*

Many factors influence the crude protein content of soybean. Genotype or variety, the inherent property of soybean, has been confirmed to determine the protein profile of soybean.

As early as 1984, Murphy & Resurreccion tested the glycinin and β-conglycinin content, which are the two major protein fractions in soybean in ten different varieties of soybean. They used five Japanese varieties and five American varieties grown in the summers of 1980 and 1981 (Murphy & Resurreccion, 1984). The results showed a significant difference in glycinin concentration within different soybean genotypes where an American variety Vinton had the largest value of glycinin content in soy protein (55.8% of total protein) ($P < 0.05$) (Murphy & Resurreccion, 1984). Additionally, β-conglycinin content was found to vary from 16.8% to 20.9% among different soybean varieties and years (Murphy & Resurreccion, 1984).

Fehr et al. (2003) examined 14 soybean cultivars including general-use, large-seeded, large-seeded + high protein, small-seeded and lipoxygenase-free cultivars. From the results, they found that the IA2035 cultivar (a small seeded-cultivar) had the highest mean protein content (39.7% on a 13% moisture

basis) ($P < 0.05$) (Fehr et al., 2003).

Kumar et al. (2006) conducted an experiment to measure protein content in seven Indian soybean cultivars (Hara soya, JS335, KHSb2, Kalitur, NRC37, Pb1 and Shilajeet. The protein content of soybean was calculated as nitrogen content × 5.71 (the conversion factor) (Kumer et al., 2006). From the results, the maximum value of protein content was found in Pb1 soybean cultivar (mean value of 39.4% on dry weight basis) when averaged over four locations (Kumar et al., 2006).

Other studies have also reported varieties of soybean grown in separate environments had different protein contents (Vollmann et al., 2000; Yan et al., 2010).

*1.5.2 Environmental Factors*

The growing environment is another critical factor that can influence both soybean protein content and composition. For example, Yaklich et al. (2002) had compared long-term trends of protein content in soybean from the northern and southern regions of the United States and Canada over 51 years (Yaklich et al., 2002). The results showed that the mean protein content of soybean grown in southern regions (411g/kg on dry basis) was higher than those grown in northern regions (407g/kg on dry basis) ($P < 0.05$) (Yaklich et al., 2002).

Another research study reported the protein content of soybean grown in different latitudes of India (Kumar et al., 2006). Soybean were grown in Palampur (32 °N), Pantnagar (29 °N), Indore (22.2 °N) and Bangalore (12.6 °N) (Kumer et al., 2006). The results showed there was a significant negative correlation of latitude with soy protein content ($P < 0.01$) (Kumer et al., 2006).

Similar research has been conducted on Chinese soybean. Specifically, soybean were sourced from four different regions of China: the northeast spring planting sub-region, the northwest spring planting sub-region, Huang-Huai-Hai valleys summer planting region and south multiple cropping region (Qin et al., 2014). The mean protein contents of soybean in these four regions were significantly different ($P < 0.05$), with a highest mean protein level of 44.54% (on dry basis) in the south region and the lowest level of 39.45% (on dry basis) in the northeast region (Qin et al., 2014).

While the latitude of growing location can influence protein and amino acid composition of soybean, specific environmental factors, such as temperature, water availability and sunlight are confounded within this variable.

Wolf et al. (1982) used Fiskeby V soybean to reveal the effect of temperature on soybean protein concentration. Soybean were grown at day/night temperatures of 24/19 ℃ (8 hours/day) before grain growth; then, soybean were moved to five different day/night temperature environments until maturity: 18/13 ℃, 24/19 ℃, 27/22 ℃, 30/25 ℃ and 33/28 ℃ (Wolf et al., 1982). The results showed that the mean protein content of soybean didn't change when day/night temperatures were between 18/13 ℃ to 30/25 ℃ (protein content around 35% on as is basis), but the protein content increased significantly when the day/night temperature was 33/28 ℃ (protein content around 42% on as is basis) ($P < 0.05$) (Wolf et al., 1982).

Piper & Boote (1999) used soybean cultivar trial data from the uniform soybean tests and daily temperature data from Earth Info, Inc to conduct research. They built models to show the relationship between temperature and soy protein content, and

they found a quadratic regression was the most suitable model, with the adjusted $R^2$ value of 0.3337 ($R^2$ = 0.3198 for linear regression model), showed that the mean protein content decreased (from around 43% to 41% on dry basis) with increasing mean temperature between 12 to 20 ℃ and increased (From 41% to 43% on dry basis) with mean temperature over 25 ℃ (Piper & Boote, 1999).

Recent research involving field trials at four locations over three years was conducted, and they recorded T5 to T8, which was the average temperatures from developmental stages R5 to R8 (Mourtzinis et al., 2017). The results showed the range of T5 to T8 was 14 to 22 ℃, and a negative correlation was reported between T5 to T8 and protein content of soybean, which meant when the average ambient temperature increased from 14 to 22 ℃, the protein content of soybean decreased (Mourtzinis et al., 2017).

While a key environmental factor, the independent effects of temperature on soy protein content and composition remains ill-defined.

Water availability is another factor that can impact phenotypic expression in soybean. Dornbos & Mullen (1992), conducted experiments in 1985 and 1986 using the soybean cultivar “Gnome”, and they investigated three different drought stress levels (control, moderate and severe) achieved by an independent trickle irrigation system in an agronomy greenhouse, on protein content (Dornbos & Mullen, 1992). The results revealed that compared to control soybean (protein content with 40.3% in 1985 and 37.4% in 1986), grains in severe drought situations contained significantly higher protein (44.9% in 1985 and 42.4% in 1986) (Dornbos & Mullen, 1992). In contrary, oil content suffered a reduction under drought environments (Dornbos &

Mullen, 1992).

Boydak et al. (2002) also investigated the impact of water availability on soybean protein content, under controlled settings. These authors separated soybean into four groups and used sprinkle-type irrigation systems to irrigate soybean every 3$^{rd}$, 6$^{th}$, 9$^{th}$, or 12$^{th}$ (control) day after emergence (Boydak et al., 2002). From the results, the mean protein content of soybean in the "every 12$^{th}$ day" group (37.6%) was significantly lower than that of soybean in other treatment groups (38.60%, 38.58% and 38.45%, separately) ($P < 0.05$) while no significant difference was found in soy protein concentration in the other three groups (Boydak et al., 2002). These result paper suggest that less irrigation resulted in lower protein content in soybean, in contrast to the results published by Dornbos and Mullen (1992).

Carrera et al. (2009) also conducted an experiment to explore the effects of water deficit on soybean protein levels. These authors examined data from 82 soybean multi-environment trials conducted at the INTA Agricultural Experimental Stations across the Argentinean soybean growing regions (Carrera et al., 2009). Precipitation minus potential evapotranspiration during grain fill (from developmental stages R5 to R7) (pp-$PET_{R5R7}$) and reproductive period (from developmental stages R1 to R7) (pp-$PET_{R1R7}$)were recorded to represent water availability level (Carrera et al., 2009). The results indicated that diminution of pp-$PET_{R1R7}$ led to lower protein concentration in soybean, which agrees with the findings of Boydak et al. (2002), but in contrast to data of Dornbos and Mullen (1992). Similar to the independent effects of temperature, it is complex to understand the effects of water availability on the protein concentration of soybean.

Beyond total protein content, some research has also explored the influence of environmental factors on soybean amino acid contents. Wolf et al. (1982) reported that Met, the limiting amino acid in soybean, sharply increased in soybean grown at a warmer temperature. Karr-Lilienthal et al. (2005) found that mean Arg contents in soybean grown in northern America (3.77% on dry basis) was lower than that in southern America (3.96% on dry basis) ($P < 0.05$). The mean His contents followed similar trend to Arg (1.36% on dry basis in northern America and 1.43% on dry basis in southern America) ($P < 0.05$). The mean Met contents of soybean in northern America (0.78% on dry basis) were lower than those in central America (0.84 on dry basis) ($P < 0.05$) but similar to those in southern America (0.77% on dry basis). Moreover, Carrera et al. (2009) concluded that significant influences of environmental factors were found on amino acid concentrations in soybean during the grain filling period (from developmental stage R5 to R7) ($P < 0.05$). Moreover, Assefa et al. (2018) identified that all essential amino acids had negative correlations with latitude where correlation coefficients ranged from -0.05 for Val to -0.21 for Trp. Additionally, they observed positive correlations with longitude, where correlation coefficients ranged from 0.01 for Cys to 0.18 for Trp, with the exception of Val.

### 1.6 Traditional analysis methods for protein and amino acid

As protein and amino acids are important indices to determine the nutritional value of soybean, several conventional methods are used for measuring their protein and amino acid contents.

*1.6.1 Protein analysis*

The concept of "protein" was first put forward by Dutch scientist, Gerrit Jan Mulder (Moore et al., 2010). At that time, nitrogen content was considered to be a means to accurately measure the protein content in a food (Moore et al., 2010). The methods of measuring nitrogen content are still used today for determining the protein content of foods and food ingredients (Krul, 2019). Since nitrogen measurement is a proxy for total protein, a nitrogen-to-protein conversion factor (NPCF) is needed in the analysis (Krul, 2019). In the latter half of the 19$^{th}$ century, the NPCF was identified as 6.25 due to measurements that indicated that proteins contained 16% nitrogen by weight and the assumption that all nitrogen in food came from protein (Moore et al., 2010).

However, in the late 19$^{th}$ century, several studies pointed out that nonprotein nitrogenous substance were also found in food, which showed that the measurement of total protein based on total nitrogen content in food wasn't perfect (Moore et al., 2010). Additionally, in the early 20$^{th}$ century, Jones (1931) found that proteins from different food sources had various nitrogen contents because of their diverse amino acid profiles. Thus, 6.25 might not be an applicable NPCF for all proteins. Several studies have been conducted to determine the NPCF in soybean and soybean protein isolate. Tkachuk (1969) indicated the NPCFs for soybean meal and soybean protein isolate were 5.71 and 5.74, respectively. Sosulski & Holt (1980) reported a NPCF of 5.63 for a set of soybean grain sample. Recent work from Spriperm et al. (2011) calculated NPCFs for feedstuffs and determined a NPCF value of 5.64 for soybean

meal.

Although the optimal NPCFs for proteins in diverse foods are different, most trade and food regulatory agencies all over the world continue to use 6.25 as the NPCF for crude protein analysis (Krul, 2019). This may be due to the fact that no standard method for NPCF calculation has been established and the applicable NPCF values have been studied for limited protein sources (Krul, 2019).

For the measurement of total nitrogen with samples, two methods are widely used. These include the Kjeldahl and Dumas methods.

The Kjeldahl method was first proposed in 1883 by a Danish chemist, Johan G.C.T. Kjeldahl (Kjeldahl, 1883). For this method, concentrated sulfuric acid is used to oxidize samples with copper sulfate as the catalyst, and nitrogen is then released from proteins in the form of ammonium ions. The latter reacts with sulfuric acid to generate ammonium sulfate (Mihaljev et al., 2015). Then, the ammonium sulfate is mixed with sodium hydroxide, which converts the ammonium ion into ammonia gas, which is volatile (Mihaljev et al., 2015). The ammonia gas is thus vaporized and moved out of the solution. An excess of boric acid is used to absorb the condensed ammonia gas where the low pH environment converts ammonia gas back to the ammonium ion in solution with borate ion (Mihaljev et al., 2015). Sulfuric acid or hydrochloric acid is then used to titrate against the ammonium borate solution with an optimal indicator to calculate nitrogen content in food samples (Mihaljev et al., 2015). Finally, the nitrogen content is transformed to protein content by multiplying it by the appropriate conversion factor (usually 6.25) (Owusu-Apenten, 2002).

Many research papers have reported the use of the Kjeldahl method to analyze

crude protein content in soybean (Wolf et al., 1982; Boydak et al., 2002; Kumar et al., 2006; Qin et al., 2014). Additionally, the basic Kjeldahl procedure, has been adapted for semiautomation, automation and micro Kjeldahl methods, as positioned within AOAC methods 955.04, 976.06, 976.05, and 960.52, respectively (Official methods of analysis of AOAC International, n.d.).

While the Kjeldahl method has high precision and accuracy for the measurement of total nitrogen in biological samples (including foods), it does have some disadvantages. The main disadvantages of the Kjeldahl method includes the required usage of hazardous chemicals, including concentrated sulfuric acid under high temperature, sodium hydroxide and heavy metals as catalysts, in addition to it being a time-consuming process (Sader et al., 2004; Mihaljev et al., 2015). Given the reduced reliance on hazardous methods, the Dumas method has become the method of choice for the measurement of total nitrogen in samples, particular for food and feed products.

The Dumas method was first described by a French chemist, Jean-Baptiste Dumas, in 1826. For this method, samples are combusted at approximately 1000 ℃ in excess oxygen, and the organic compounds are oxidized to release gases ($O_2$, $CO_2$, $H_2O$, $N_2$ and $NO_x$), which pass through several traps (Saint-Denis & Goupy, 2004). The only gases retained by traps are nitrogen and nitrogen oxides, which are derived from bound nitrogen found in original sample, including food samples (Saint-Denis & Goupy, 2004). Subsequently, the gases are carried by $CO_2$ to the reduction zone where nitric oxides are converted to nitrogen by hot tungsten, and a thermoconductivity detector with an electronic flow controller is used to calculate the nitrogen

concentration in samples (Mihaljev et al., 2015). Finally, the protein content is acquired by multiplying the nitrogen content by the appropriate nitrogen conversion factor (usually 6.25).

The Dumas combustion method had been used in many studies to measure protein content in soybean (Bakalli et al., 2000; Fontaine et al., 2001; Kovalenko et al., 2006). Also, this method has been established by AOAC as approved method 992.23 (Tahir et al., 2011; Official methods of analysis of AOAC International, n.d.).

Similar to the Kjeldahl method, the Dumas method also has high precision and accuracy in protein measurement but takes much less time (around 5 minutes per measurement) without the need to use toxic chemicals or catalysts (Saint-Denis & Goupy, 2004; Mihaljev et al., 2015). The main disadvantages of the Dumas method include its high initial cost for the equipment(Mihaljev et al., 2015) and the requirement for regular maintenance and a dedicated technician. Additionally, the nitrogen content of samples acquired from Dumas method may be higher than that from Kjeldahl method since Dumas method measures total nitrogen while the Kjeldahl method only measures organic nitrogen (Thompson et al., 2002; Nielsen, 2010).

### *1.6.2 Amino acid analysis*

Although protein content is important for assessing soybean quality, as mentioned before, the amino acid content of a protein is a more valuable indicator of the nutritional quality of soybean.

The basic aim of amino acid analysis methods is to break the peptide bond that

binds amino acids in proteins through the use of either acid or alkaline hydrolysis, and the subsequent measurement of the free amino acids via separation and detection using chromatographic techniques (Rutherfurd & Gilani, 2009; Nielsen, 2010).

Standard approaches for amino acid analysis include three parts: 1) Regular acid hydrolysis conditions (measurement of amino acids except for sulfur amino acids and tryptophan); 2) Pre-oxidation followed by acid hydrolysis (measurement of sulfur amino acids); and 3) Alkaline hydrolysis (for tryptophan analysis).

For both the regular and pre-oxidation acid hydrolysis techniques, 6 N hydrochloric acid is used to hydrolyze protein at 110 ℃ for 20 to 24 hours (Blackburn, 1968). During this process, the amino acids asparagine and glutamine are deaminated, yielding Asp and Glu, respectively (Fountoulakis & Lahm, 1998). It is important to keep an oxygen-free environment during hydrolysis to enhance the recovery of several amino acids (Rutherfurd & Gilani, 2009). To achieve this condition, sparging with an inert gas, sealing hydrolysis tubes under vacuum or a combination of these two methods are always used (Yamada et al., 1991; Weiss et al., 1998). However, it is difficult to remove all oxygen from the tube exactly; Cys and Met can be oxidized to cysteic acid and methionine sulfoxide or methionine sulfone, respectively (Rutherfurd & Gilani, 2009). To address this issue of variable oxidation of sulphur amino acids, a pre-oxidation step is employed to completely oxidize Cys to cysteic acid and Met to methionine sulfone. This process employs performic acid, prepared by combining one part 30% hydrogen peroxide and nine-parts 88% formic acid (Spindler et al., 1984; Cooper et al., 2001). Sodium metabisulfite or hydrobromic acid is used to reduce performic acid to formic acid after the oxidation step (Rutherfurd & Gilani, 2009),

after which the samples are subjected to the regular acid hydrolysis step. After hydrolysis, samples are adjusted to suitable pH and sent to chromatographic techniques for separation and detection (Rutherfurd & Gilani, 2009). Recently, ion-exchange chromatography with post-derivatization and reverse-phase chromatography with pre-column derivatization are most generally used (Rutherfurd & Gilani, 2009).

For tryptophan analysis, sodium hydroxide, barium hydroxide or lithium hydroxide is used for hydrolysis (Delhaye & Landry, 1986), as Trp is destroyed during acid hydrolysis, particularly in the presence of > 5% carbohydrate (Finley, 1985). However, alkaline hydrolysis destroys Thr, Arg and Cys, and other amino acids are racemized under these conditions (Fountoulakis & Lahm, 1998). Alkali is added to samples and the samples heated at 110 ℃ for 20 hours of hydrolysis with the addition of α-methyltryptophan or 5-methyltryptophan as an internal standard (Fontaine et al., 2001; D'Mello, 2003). Chromatographic techniques are used for separation and detection after hydrolysis, and reverse-phase chromatography with ultraviolet (UV) or fluorescence is recommended by many authors (Delhaye & Landry, 1992; D'Mello, 2003).

Many papers reported the use of these approaches to measure amino acid concentrations in soybean grains (Fontaine et al., 2001; Karr-Lilienthal et al., 2005; Kovalenko, Rippke & Hurburgh, 2006; Carrera et al., 2011). Also, these methods have been established by AOAC methods 982.30 and 985.28, and by ISO in method 13904:2005(E) (Official methods of analysis of AOAC International, n.d.; ISO, 2005).

The main disadvantages of these methods include the facts that they are time-consuming, have high labor costs when a large number of samples are analyzed, and pose chemical hazards caused by the use of highly concentrated acid or alkali at high temperature (González-Martín et al., 2006; Kovalenko et al., 2006). As such, alternative methods need to be positioned to assist in measuring the amino acid content of protein sources. Once such method is Near-Infrared Technology.

## 1.7 Near-infrared technology

### *1.7.1 Background*

Near-infrared (NIR) light, electromagnetic energy with a wavelength range between 750 and 2500 nm, is widely used for the quantitative analysis of foods (Nielsen, 2010; Agelet & Hurburgh, 2010). When samples are explosed to light of specific wavelengths, the light energy is absorbed for vibrations of molecular and bonds, including C-H (methyl, aromatic, carbonyl, etc.), N-H (amides and amine salt), O-H (alcohols and water), S-H, C=O groups, etc. (Agelet & Hurburgh, 2010). A portion of the light is absorbed by the substance, and other light penetrates (transmission) or retraces back (diffuse reflectance), which is received by the detector of the equipment (Nielsen, 2010; Mihaljev et al., 2015). Transmission and diffuse reflectance are two modes of the NIR system. Diffuse reflectance mode allows the instrument to measure thicker and denser samples but is also greatly affected by density or packing, particle size, and absorption of samples (Berntsson et al., 1998). Although the light absorbed by the substance cannot be measured directly, unabsorbed light could be acquired directly by the detector and used for absorbance calculation

(Agelet & Hurburgh, 2010; Mihaljev et al., 2015). Finally, the digital signal processor with a computer calculates the final results of predicted component concentration in samples (Mihaljev et al., 2015).

NIR technology is a rapid and non-destructive method for measuring chemical components and nutrients in food materials with little sample preparation (Cen & He, 2007; Mihaljev et al., 2015). The automatic reading and recording functions make it simple to operate, and it allows measurement for multiple constituents of samples at the same time (Osborne, 2006). Additionally, the NIR instrument has no need for hazardous chemicals, which is safe and environmentally friendly (Wu et al., 2002). The major disadvantages include the high initial cost for the instrument and extensive calibration requirements (Blanco & Villarroya, 2002; Mihaljev et al., 2015). A summary of comparison of traditional analysis methods for protein and amino acid and NIR spectroscopy analysis is shown in Table 1.5.

Table 1.5. Comparison of traditional analysis methods for protein and amino acid and NIR spectroscopy

| | Traditional analysis method | | | NIR technique |
|---|---|---|---|---|
| | Kjeldahl | Dumas | Amino acid analysis | |
| Advantage | •High precision and accuracy<br>•Relatively cheap | •High precision and accuracy<br>•Relatively fast<br>•No need for toxic chemicals | •High precision and accuracy | •Time saving<br>•Less sample preparation<br>•Non-destructive<br>•Environmentally friendly<br>•Measurement of multiple traits |
| Disadvantage | •Chemical hazard<br>•Time consuming | •High initial cost<br>•Requirement of regular maintenance<br>•Measurement of non-protein nitrogen | •Time consuming<br>•High labor cost<br>•Chemical hazard | •High initial cost<br>•Further calibration requirement |

*1.7.2 Composition of different NIR spectroscopy instruments*

A commercial NIR spectrophotometer consists of five basic parts: (1) Sample compartment, (2) Light source, (3) Wavelength selection system, (4) detector/s, and (5) signal processor or computer (Agelet & Hurburgh, 2010).

The sample-compartment for the reflectance equipment is usually open sample cups or sample cells confined by silica or quartz, allowing NIR light to pass through (Agelet & Hurburgh, 2010). Transmission instruments may also work with confined sample cells with specific pathlengths, which depend on the analyzed product (Lipp, 1992).

Light-emitting diodes (LED) and tungsten halogen lamps are commonly used as light sources in NIR technology. LEDs have relatively broadband wavelength emission, low power requirement, low price, small size and long shelf life (Stark & Luchter, 2005; Agelet & Hurburgh, 2010). Tungsten halogen lamp emits light with a wavelength range from 320 to 2500 nm, and halogen gas can recycle the evaporated tungsten, which prolongs the shelf life of the lamp (Stark & Luchter, 2005; Agelet & Hurburgh, 2010).

The wavelength selection (also called “feature selection” or “variable selection”) filters polychromic light beams to receive discrete wavelength data (Agelet & Hurburgh, 2010; Chen & Wang, 2019). The is due to the fact that not all wavelengths have the same importance for qualitative or quantitative analysis; most representative wavelengths should be chosen for analytical purposes (Agelet & Hurburgh, 2010; Chen & Wang, 2019). The common approaches for wavelength selection include filters (narrow bandpass interference filter, acousto-optic tunable filter and liquid

crystal tunable filter), grating (monochromators and diode-arrays spectrograph) and interferometer (Michelson interferometer and crystal polarisation interferometer) (Stark & Luchter, 2005).

Detectors convert the light energy information to an electric analog signal, which is amplified and transformed to digital signals for later processing by a computer (Agelet & Hurburgh, 2010). The material of the detector depends on the NIR region for analysis. Silicon detectors are usually used for wavelengths ranging from 400 to 1100 nm (Stark & Luchter, 2005). Lead Sulfide (PbS) or Indium Gallium arsenide (InGaAs) detectors cover higher wavelength regions than Si detectors (Agelet & Hurburgh, 2010).

### *1.7.3 Calibration of NIR system*

Since NIR has the least wavelengths in the infrared region, it is the most energetic (Agelet & Hurburgh, 2010). The high energy of NIR causes up to four overtones and combination bands from molecular bonds (Agelet & Hurburgh, 2010). Therefore, the NIR spectra are made up of overlapped overtones and combination bands, which show broader peaks than mid-infrared (MIR) spectra (Dryden, 2003). These broad peaks in NIR spectra cannot be represented to any specific chemical compounds in samples (Agelet & Hurburgh, 2010). Also, physical, chemical and structural variables affect the result of NIR spectra (Blanco & Villarroya, 2002).

Therefore, several steps are required to combine spectra data with reference data acquired from traditional analysis methods and this is accomplished during calibration processing.

Before calibration, a calibration set is required. The desired calibration set should represent the chemical, spectral, and physical characteristics of the analyzed population and avoid future extrapolations when measuring new sample sets (Fearn, 2005).

Also, spectra data require a pre-treatment due to the spectra of solid samples being affected by their physical properties (Blanco & Villarroya, 2002). Mathematical pre-treatment of spectra minimizes irrelevant information (noise or background) and maximizes signal from the chemical information, which helps to develop simple and robust calibration models (Blanco & Villarroya, 2002; Agelet & Hurburgh, 2010). Common approaches for spectra pre-treatment include normalization (Griffiths, 1995), derivatives (Agelet & Hurburgh, 2010), multiplicative scatter correction (MSC) (Geladi et al., 1985), standard normal variate (SNV) (Barnes et al., 1989), de-trending (DT) (Barnes et al., 1989) or a combination of the above-mentioned approaches.

NIR calibration models correlate raw or pre-treated spectra data with one or more chemical-physical properties of samples, which requires the use of multivariate-analysis methods to achieve this aim (Blanco & Villarroya, 2002; Agelet & Hurburgh, 2010). For quantitative analysis, the relationship between spectral data and reference values is usually linear due to Beer's law (Agelet & Hurburgh, 2010). Based on this assumption, multiple linear regression (MLR), principal component regression (PCR) and partial least squares (PLS) methods are most often used (Blanco & Villarroya, 2002; Agelet & Hurburgh, 2010). Of the methods cited, PLS is preferred among them because of the faster algorithm, higher precision of models and harmonious calibration models (Kalivas & Gemperline, 2006). When the spectral data

and target property of the samples are not related in a linear fashion, artificial neural networks (ANN) (Zupan & Gasteiger, 1993), non-linear PLS (Blanco & Villarroya, 2002) and support vector machines (SVM) (Agelet & Hurburgh, 2010) are used for non-linear NIR calibration.

Adequate validation is a necessary part of examining the predictive ability of NIR models after calibration. The ideal validation involves samples that aren't used for calibration of NIR models, a process called independent validation (Agelet & Hurburgh, 2010). Since this validation isn't always available, another accepted validation method that can basically assess the performance of calibration models is called cross-validation, which is a method to exclude a single sample (full cross-validation) or a group of samples (k-fold cross-validation) and establish a calibration model using the remaining samples (Agelet & Hurburgh, 2010). The calibration model from cross-validation is then validated with the excluded samples (Agelet & Hurburgh, 2010). However, the calibration model developed with whole sample set isn't tested. Thus, any statistics from cross-validation cannot be directly compared or interpreted with statistics from the true validation of the final calibration model with external samples (Agelet & Hurburgh, 2010).

Statistics are used to support the results of calibration and validation approaches, and these include the coefficient of determination ($R^2$), standard error of prediction (SEP), standard error of cross-validation (SECV), ratio of performance of deviation (RPD), etc. $R^2$ indicates the explained variance between reference and predicted values versus the total variance, and the standard error (SEP or SECV) determine the precision of the calibration (Agelet & Hurburgh, 2010). RPD is calculated by the

formular of standard deviation (SD)/ (SEP or SECV), which is used to evaluate the validation results (Li et al., 2011)

*1.7.4 Application of NIR for measuring crude protein and amino acid concentrations*

In 1978, Rubenthaler & Bruinsma calibrated a NIR model for estimating lysine in cereals with a $R^2$ value of 0.96 for wheat NIR calibration, which was the first successful example of NIR calibration for measuring amino acids. In the following years, Williams et al. (1984) used NIR technology to measure amino acid composition in ground wheat and barley. Fontaine et al. (2001) tested the performance of NIR in protein and amino acid measurement in soy, rapeseed, sunflower and peas meal. Wu et al. (2002) reported the use of NIR spectroscopy on estimating amino acid content in milled rice. Kovalenko et al. (2006) designed an experiment to correlate NIR spectra data to soybean protein and amino acid content, and the results are summarized in Table 1.6.

A recent study included the use of NIR spectroscopy to measure moisture, fat and protein content in soybean meal (Zhu et al., 2018). They used 216 samples to build the NIR calibration models and 54 samples to validate the models, resulting in $R^2$ of external validation for moisture, crude fat and protein to be 0.966, 0.958 and 0.958, respectively (Zhu et al., 2018). Wee et al. (2018) used Fourier NIR system to measure protein, lipid and sucrose contents in wild soybean meals (particle size around 0.5mm). They used 50 samples to make calibration models and 26 to 38 samples to test the accuracy, resulting in $R^2$ of calibration to be 0.985, 0.944 and 0.871 and $R^2$ of validation to be 0.884, 0.832 and 0.884 for protein, lipid and sucrose,

respectively (Wee et al., 2018). Xu et al. (2020) used 167 whole soybean to build NIR calibration model for measuring water soluble protein with 49 external samples as the validation set. They got the highest $R^2$ of cross-validation of 0.831 when the combination of multiplicative scatter correction (MSC) and Savitsky-Golay transformation (2nd derivative) was used for spectral pretreatment (Xu et al., 2020).

Additionally, the American Association of Cereal Chemists (AACC) has approved measuring protein and lipid concentration in whole and ground soybean, and protein content in wheat and small grains by NIR system (AACC, 1999).

Table 1.6. Near infrared calibration statistics for soybean amino acid measurements from Kovalenko et al. (2006)

| $R^2$ range | Component | Availability |
|---|---|---|
| 0-0.25 | tryptophan | unusable models |
| 0.26-0.49 | cysteine | poor correlation models |
| 0.50-0.64 | methionine and serine | models usable for rough sample screening |
| 0.66-0.81 | alanine, glutamic acid, isoleucine, threonine and valine | models usable for sample screening |
| 0.83-0.90 | arginine, aspartic acid, glycine, histidine, leucine, lysine, phenylalanine and tyrosine | models “usable with caution for most applications” |

## 1.8 Summary

In summary, proteins and amino acid contents are important indicators for determining soybean quality and nutritional value. Selecting and cultivating high protein soybean with stable performance is required for some purposes. Previous studies indicated that factors affecting soybean compositions include soybean cultivars and the growing environment, as well as the interaction of genotype by environment. Thus, it is necessary to explore the effects of genotype, environments and genotype by environment interaction on soybean protein and amino acid content. In order to achieve a high efficiency of analyze and selection, near-infrared spectroscopy as a method with high speed, low cost and many successful cases for measuring protein and amino acid is applicable for this achievement.

## 1.9 Hypothesis and Objectives

### *1.9.1 Hypothesis*

$H_{a1}$: NIR spectroscopy can measure crude protein and most amino acid with acceptable accuracy.

$H_{a2}$: Genotype, environment and genotype by environment interactions affect soybean protein and amino acid contents.

### *1.9.2 Objectives*

1. To develop NIR calibration models for crude protein and amino acid measurements in soybean and evaluate their performances.
2. To explore the impacts of particle size, fat and types of NIR spectroscopy instruments on the predictive ability of NIR calibration models.
3. To explore the effects of genotype, growing environment and their interactions on soybean protein and amino acid contents.
4. To evaluate phenotypic stability of soybean genotypes across various environments.

# Chapter 2 Estimation of Crude Protein and Amino Acid Contents in Whole, Ground and Defatted Ground Soybean by Different Types of Near-infrared (NIR) Reflectance Spectroscopy

## 2.1 Abstract

This study was designed to exam the ability of near-infrared (NIR) spectroscopy to determine amino acid and crude protein concentrations in soybean and explore the effects of particle size, fat content and types of NIR spectroscopy instruments on the predictive ability of the NIR models. Whole, ground and defatted ground soybean were scanned by diode array NIR analyzer and defatted ground soybean were scanned by Fourier transformed NIR analyzer. Samples were analyzed for crude protein and amino acid contents using reference wet chemistry methods. The NIR calibration models of crude protein and most amino acids except for CYS, MET and TRP showed acceptable coefficient of determinations ($R^2_c$ = 0.605-0.952), while the NIR models for CYS, MET and TRP were less accurate ($R^2_c$ = 0.498-0.667 for CYS and $R^2_c$ = 0.482-0.615 for MET and $R^2_c$ = 0.406-0.481 for TRP), which might require further calibration in the future study. The $R^2_c$ values in NIR calibration were found to relate to the correlation between amino acids and crude protein. The grinding process and lipid extraction seemed to improve the NIR spectroscopy in crude protein and amino acids prediction. PerkinElmer FT 9700 NIR spectroscopy instrument was found to perform similar to PerkinElmer DA 7250. No significant differences were found between crude protein and amino acid contents predicted by these two NIR spectroscopy instruments ($P > 0.05$). Thus, NIR spectroscopy has potential to measure crude protein and most amino acid concentrations with an acceptable accuracy.

## 2.2 Introduction

Soybean [*Glycine max (L.) Merr.*] is a leguminous plant with high protein (around 40% in dry grain) and lipid (around 20% in dry grain) contents (Jiang, 2020). It is a major crop that plays an important role in agricultural production, human food security and international trade (Jiang, 2020). As the most essential grain composition and nutritional quality traits in soybean, soybean protein and oil have a variety of applications. For instance, soy protein is used in poultry and livestock feeding (Dei, 2011) and human food applications including tofu and soymilk (Lusas & Riaz, 1995).

In recent years, plant-based foods have become more popular globally. In 2017, a global consumer study showed that 39% of Americans and 43% of Canadians were planning to consume more plant-based foods (Doris, 2018). Moreover, the new Canadian food guide issued by Health Canada encourages people to consume more vegetables and plant-based protein foods, such as pulses and nuts (Government of Canada, 2019). Thus, as the main source of plant protein (Kovalenko, Rippke & Hurburgh, 2006), soybean has potential in food market. Except protein, the amino acid content of a protein is a more valuable indicator of the nutritional quality of soybean. Additionally, with respect to livestock feeding applications, modern animal feed formulation is based on the amino acid content in the feed ingredients (Kovalenko, Rippke & Hurburgh, 2006). Therefore, it is necessary to determine protein and amino acid contents in soybean.

Traditionally, the concentrations of protein and amino acids have been determined through the use of wet chemistry approaches, such as the Kjeldahl and Dumas methods for protein, and acidic or alkaline hydrolysis methods for the ultimate measurement of amino acid

composition. The conventional approaches can measure protein and amino acid concentration accurately and precisely. However, these methods require a long experimental period and high labor cost (Mihaljev et al., 2015; Lee et al., 2013; Font, del Río-Celestino & de Haro-Bailón, 2006). Also, the high concentration of acid and alkali during the test could cause chemical hazards and residues (Santos et al., 2018). In addition, traditional methods cannot measure multiple constituent contents of samples at the same time (Singh et al., 2018), and they are destructive methods thus negating the ability to use the whole grains for subsequent planting for varietal selection (Jiang, 2020). Therefore, a rapid and economical technique is required for soybean protein and amino acid measurements.

Near infrared (NIR) spectroscopy is a rapid and non-destructive approach for measuring chemical compounds and nutrients in food and crops with little sample preparation (Cen & He, 2007; Mihaljev et al., 2015). It is easy to operate and store this technology due to its automatic reading and recording functions with less space requirements. Furthermore, it allows for the measurement of multiple constituents of samples simultaneously (Osborne, 2006). Its major disadvantages are the high initial cost for the equipment and the need for extensive calibration (Blanco & Villarroya, 2002; Mihaljev et al., 2015).

In 1978, Rubenthaler & Bruinsma calibrated the NIR calibration model for estimating lysine in cereals with a coefficient of determination of 0.96 for wheat NIR calibration. This was the first successful example of NIR calibration for an amino acid, and more research was conducted with NIR to predict protein and amino acid contents in feedstuffs in subsequent studies. Wu, Shi & Zhang (2002) reported the use of NIR spectroscopy on estimating amino acid composition of milled rice, with high coefficients of determination for most amino acids (0.85-0.98) except cysteine, methionine and histidine. Williams et al. (1984) used NIR

technology to predict the amino acid concentrations in ground wheat and barley. Kovalenko, Rippke, & Hurburgh (2006) designed an experiment to correlate NIR spectral data with soybean protein and amino acid contents. Additionally, NIR technology for measuring protein and lipid concentration in whole and ground soybean, and protein content in wheat and small grains has been approved by American Association of Cereal Chemists (AACC) (AACC, 1999).

NIR spectroscopy is widely used to predict amino acid concentrations in grain plants with high accuracy (Mihaljev et al., 2015). However, certain factors may influence the accuracy and precision of NIR technology. For example, the size and shape of the sample particles may affect the amount radiation that absorbed and reflected by sample surface (Matern & Naes, 2001). In addition, Kovalenko, Rippke, & Hurburgh (2006) reported that ground samples may enhance the ability of NIR technology to predict amino acid contents by comparing their results with previous studies. Nevertheless, there is limited research exploring the effects of particle size of samples on NIR spectroscopy performance. Also, it is unknown whether oil, one of the major constituents in soybean, and types of NIR spectroscopy instruments, affect the protein and amino acid measurement by NIR technology.

Therefore, the objectives of this experiment were to (i) establish NIR calibration models for crude protein and amino acid predictions in soybean and evaluate their performances; (ii) explore the impacts of particle size, fat and types of NIR spectroscopy instruments on the predictive ability of NIR calibration models.

## 2.3 Materials and methods

### *2.3.1 Sample acquisition and preparation*

Approximately 4700 whole soybean samples were supplied by Manitoba Pulse and Soybean Growers Association. Soybean were grown in 13 different locations in Manitoba during 2018 and 2019. All whole soybean samples were analyzed via NIR spectroscopy on a PerkinElmer DA 7250 diode array NIR system (PerkinElmer Health Sciences Canada Inc., Winnipeg, MB, Canada). Total analyses covered moisture (%), protein (%), amino acid (%), and fat (%), which were determined using calibration equations established previously within our laboratory. Following analysis, a sub-sample of 360 was drawn from the full lot, as follows: 1) The sampling was based on the predicted protein content of samples; 2) Samples were divided into quartiles of predicted crude protein content, covering 0-24.9, 25-49.9, 50-74.9, and 75-100% relative to the median. Samples (90) were randomly drawn from each quartile.

The sub-selected samples were ground through a PerkinElmer LM-3610 grinder (PerkinElmer Health Sciences Canada Inc., Winnipeg, MB, Canada) into a 1.0 mm particle size. Following grinding, the samples were defatted following soxhlet extraction via hexane (De Castro & Priego-Capote, 2010) for 16 hours and dried in a fume hood (24 hours) in advance of chemical analyses. Defatted soybean meals were ground through a Retsch ZM-200 grinder (Retsch, Haan, Germany) into a smaller particle size powder (around 0.75 mm) and stored at -20 ℃ prior to analysis.

*2.3.2 Protein and amino acid analysis*

The crude protein content of the samples was determined by multiplying the nitrogen content by 6.25, where the nitrogen content was determined by combustion (Dumas method, AOAC 990.03). Amino acid analysis was measured using official methods AOAC 982.30 and

985.28 (AOAC, 1995) and ISO 13904:2005(E) (ISO, 2005).

A total of 18 amino acids including alanine (Ala), arginine (Arg), aspartic acid (Asp), cysteine (Cys), glutamic acid (Glu), glycine (Gly), histidine (His), isoleucine (Ile), leucine (Leu), lysine (Lys), methionine (Met), phenylalanine (Phe), proline (Pro), serine (Ser), threonine (Thr), tryptophan (Trp), tyrosine (Tyr), and valine (Val) were measured by official methods. For most amino acids except the sulfur amino acids (Met and Cys) and Trp, 25-35 mg of sample was weighed with 4 ml 6 N HCl in tubes with screw caps. The hydrolysis process was conducted under low oxygen conditions achieved by flushing the tubes with nitrogen gas, and hydrolysis was conducted for 24 hours at 110 ℃. Norvaline (in HCl) was added as the internal standard. Afterwards, sodium hydroxide was added to neutralize samples to reach a pH of 5.5-6.0, the samples were diluted to 50 ml with deionized water. A reverse phase ultra-performance liquid chromatography (UPLC) with the AccQ•TagTM Ultra Derivatization Kit as a pre-column derivatization was used for amino acid separation, and each amino acid was then detected by ultraviolet spectroscopy at 260 nm.

The measurement of sulfur amino acids (Met and Cys) was performed in a similar manner with the addition of a pre-oxidation step. In brief, 25-35 mg of sample was weighed and put in the fridge with 2 ml performic acid overnight. During this period, Met and Cys were oxidized by performic acid to form methionine sulfone and cysteic acid, separately. After oxidation, sodium metabisulfite was added in samples to stop oxidative reaction, and 2 ml 6 N HCl was added with norvaline to hydrolyze protein from sample for over 18 hours at 110 ℃. The remaining steps were the same as in the previous paragraph except for the detection method (fluorescence with an excitation wavelength of 266 nm and an emission wavelength of 473 nm).

The Trp content of the samples was analyzed by alkaline hydrolysis. 25-35 mg of sample was mixed with 8.4 g barium hydroxide and hydrolyzed for 20 hours at 110 ℃ in autoclave with about 1.5 times normal atmospheric pressure. The hydrolysate of sample was then mixed with 5 ml of phosphoric acid and 8.4 ml HCl and adjusted to a pH of 2.95-3.20. Deionized water and 20 ml methanol were added to dilute the sample to 100 ml. Consequently, samples were injected into UPLC with the buffer consisted of 0.3% glacial acetic acid and 0.05% 1,1,1-trichloro-2-methyl-2-propanol. The detection was through fluorescence with an excitation wavelength of 280 nm and an emission wavelength of 356 nm.

*2.3.3 NIR analysis*

Whole, ground and ground + defatted samples were scanned by PerkinElmer DA 7250 diode array NIR analyzer with spinning sample system and reflectance detector. Each whole soybean sample was fitted in a 150 mm (400 ml) or 105 mm (150 ml) diameter cup, and each ground sample was loaded and scanned twice, and an average reflectance was recorded. The reflectance spectra (log1/R) were recorded from 950 to 1650 nm in 5 nm intervals. Additionally, defatted ground soybean samples were scanned on a PerkinElmer FT 9700 Fourier transform NIR analyzer, using a spinning system and reflectance recorder. This analyzer recorded spectra from 14304 to 3856 $cm^{-1}$ (around 699 to 2593 nm) in 8 $cm^{-1}$ intervals.

NIR calibration models were developed using The Unscrambler® X version 10.3 software (CAMO Software, Oslo, Norway). Partial least square (PLS) regression was conducted using the reflectance spectra and wet chemistry analysis of protein and amino acid concentrations. Spectral data from the DA 7250 was between 950 to 1650 nm in 5 nm steps

(total 141 points ) and, from the FT 9700, between was from 14304 to 3856 $cm^{-1}$ with an interval of 8 $cm^{-1}$ (total 1307 points). The spectral data had been pre-treated with the scatter correction processes (SNV and de-trending) to reduce the particle size effect (Fontaine, Hörr, & Schirmer, 2001). The maximum number of factors for PLS analysis was set to 20, and the number of cross-validation groups was set to 20 with 18 random samples in each group. The predictive ability of models was evaluated using parameters including standard error of calibration (SEC), coefficient of determination in calibration ($R^2_c$), standard error of cross-validation (SECV) and coefficient of determination in cross-validation ($R^2_{cv}$), which were measured by the software. The optimal number of factors for each calibration model was selected when SECV was minimum. The ratio of the standard deviation of reference values to SECV (RPD) was also calculated for model evaluation (Williams & Norris, 2001).

## 2.4 Results and Discussion

### *2.4.1Raw data for NIR calibration*

The content of crude protein and 18 amino acids in soybean from the calibration set were measured by the reference methods, and the concentrations of these traits are summarized in Table 2.1. Crude protein and amino acid contents were expressed on a dry weight basis, and amino acids were measured as free (hydrated) amino acids. Table 2.1 shows a wide range in crude protein and amino acid contents, which is essential for NIR calibration model development (Wang et al., 2013). The broad range of concentrations of these compounds might be explained with the soybean samples that were composed of various genotypes and planted in different environments.

The coefficient of determination ($R^2$) described the correlation of crude protein content

with amino acid contents in soybean (Table 2.1). The higher $R^2$ value indicated that amino acid content was more predictable by crude protein content. Val had the highest $R^2$ (0.72), which was suggested to correlate with crude protein relatively well. CYS was the only amino acid with no significant correlation to crude protein ($P > 0.05$). Slope and intercept are also given in Table 2.1.

*2.4.2 NIR spectral assignment*

Figure 2.1 shows the raw NIR spectra data of whole soybean samples from the calibration set scanned with the DA 7250, with the spectra range between 950-1650 nm. This range of NIR spectra contains information of C-H, N-H, O-H and C=O (Williams et al., 2019), which refer to oil, protein, water and carbohydrate contents of soybean. Three peaks could be found from the spectra when wavelengths are approximately 1000, 1200 and 1450 nm. After processed by SNV and de-trending, the spectra became more concentrated and the peaks turned sharper (Figure 2.2). The flat peak at 980-1000 nm might be related to the O-H and N-H functional groups; the peak near 1200 nm was possible to be explained by the second overtone of C-H stretching; and the peak around 1450 nm corresponded to the first overtone of O-H and N-H stretching (Hourant et al., 2000; Wang et al., 2013; García-Sánchez et al., 2017; Williams et al., 2019).

*2.4.3 NIR calibration and cross-validation*

Table 2.2 - 2.6 summarize the statistics obtained from NIR calibration and cross-validation including the information of samples that remained for calibration and parameters for model evaluation. Statistics of models for whole soybean prediction by the DA

7250 are summarized in Table 2.2 and Table 2.3. Table 2.4 and Table 2.5 show the results of models for ground soybean (full fat and defatted) measurement with the DA 7250. Table 2.6 includes the results of models for defatted soybean meal prediction developed with the FT 9700.

Based on Table 2.2 - 2.6, calibration models for crude protein had the highest $R^2_c$ (0.894-0.952), $R^2_{cv}$ (0.852-0.934) and RPD (2.68-3.89) value, showing that these equations accurately predicted the crude protein content of the calibration set. The higher RPD value indicates the better predictive ability of a calibration model (Wang et al., 2013). RPD values of 1.40 or less describe a relatively poor calibration model (Saeys et al., 2005; Kovalenko et al., 2006). Models with RPD values from 1.41 to 1.70 are usable for rough sample screening and from 1.71 to 2.40 are usable for quantitative ratings. When RPD values are over 2.50, models are regarded as good or excellent.

The evaluation parameters of calibration models for amino acids varied in a wide range (Table 2.2 – 2.6). The $R^2_c$ values for various amino acids ranged from 0.406 to 0.880 and RPD values ranged from 1.22 to 2.45. The RPD values suggested most calibration models are usable for sample screening (RPD > 1.40), which indicated the potential of NIR spectroscopy to screen out soybean with high amino acid contents. ARG, GLU, PRO, SER and VAL had relatively high $R^2_c$, $R^2_{cv}$ and RPD values ($0.795 < R^2_c < 0.881$, $0.750 < R^2_{cv} < 0.836$ and $1.98 < RPD < 2.46$), showing that NIR calibration model could predict the contents of these amino acids more accurately than other amino acids. Models for ASP, ILE, LEU and PHE had $R^2_c$ values over 0.764 and RPD values over 1.89, which suggested the calibration models for measurement of these amino acids were feasible but with less accuracy. ALA, GLY, HIS, LYS, THR and TYR had relatively low $R^2_c$ with values ranged from 0.605 to 0.758 and RPD

ranged from 1.51 to 1.85. This indicated the NIR calibration models were able to distinguish samples that had high or low contents of these amino acids but with relatively low accuracy. CYS, MET and TRP had the lowest RPD values as well as $R^2_c$ and $R^2_{cv}$. The RPD values for TRP ranged from 1.22 to 1.35 and $R^2_c$ ranged from 0.406 to 0.481, indicating the calibration models for this amino acid were poor and might not be used for measurement. CYS and MET had RPD values ranged from 1.34 to 1.48 and 1.36 to 1.47, separately, showing on both sides of 1.40, which suggested that some of the calibration models could roughly predict CYS and MET concentrations in soybean but with low accuracy.

Compared to previous studies, Kovalenko et al. (2006) found low RPD values for CYS (< 1.26), MET (< 1.55) and TRP (< 1.10), and Fontaine et al. (2001) showed relatively low $R^2_c$ and $R^2_{cv}$ of calibration models for CYS and MET, which were consistent with the results in this study. It seems possible that these results are due to the limited concentrations of sulfur amino acids and TRP in soybean (Wilcox, 1987). The low TRP and sulfur amino acid contents in soybean and the small range of variance might explain the low $R^2_c$ values for calibration models (Wu et al., 2002). Low variation leads to a small S.D. for the calibration set, which reduces RPD values as well.

The $R^2_c$ and $R^2_{cv}$ values for both whole and ground soybean calibrations in this experiment were lower than those previously reported by Fontaine et al. (2001) who found high correlations between protein and amino acids ($R^2$ ranged from 0.81 to 0.98). A possible explanation for this might be the calibration set in this study was composed of various genotypes planted in different environments. This might lead to lower correlations between protein and amino acids (Table 2.1) since each amino acid responded differently to environmental alteration (Carrera et al., 2011). Additionally, protein profiles of soybean vary

from different genotypes. Compared with results reported by Pazdernik et al. (1997), whose calibration set was established from different soybean lines, most $R^2_c$ and $R^2_{cv}$ in this study were higher for both whole and ground soybean calibrations. This might be explained with a relatively larger calibration set used in this experiment (360 samples in this study and 90 samples for Pazdernik et al.).

Previous studies suggested the predictive ability of NIR models was related to the correlation between the certain amino acid and crude protein (Kovalenko et al., 2006; Fontaine et al., 2001). Kovalenko et al. (2006) concluded that if a specific amino acid content of soybean could be predicted by crude protein content, it could be measured using NIR spectroscopy. On the contrary, Rubenthaler & Bruinsma (1978) found that the prediction of amino acids by NIR spectroscopy was independent of crude protein. Figure 2.3 shows the $R^2_{cp}$ from Table 2.1 for linear regression of amino acids to crude protein versus the $R^2_c$ from Table 2.2 for NIR calibration models on whole soybean samples in dry weight content using PerkinElemer DA 7250. In addition, the slope value was tested using PROC REG in SAS 9.4 to confirm its significance ($P < 0.01$). Based on Figure 2.3, $R^2_{cp}$ and $R^2_c$ had a positive linear relationship with the $R^2$ value of 0.39. However, due to the samples from the calibration set included diverse genotypes planted in different environments, the correlations of crude protein to amino acids were relatively poor, with $R^2_{cp}$ values ranged from 0 to 0.56 for most amino acids, which were lower than $R^2_c$ from NIR calibration. This was in agreement with Wang et al. (2013), who concluded that NIR spectroscopy explained more variance than crude protein regression. Therefore, it was suggested that amino acids with higher correlations to crude protein might be predicted by NIR spectroscopy with great accuracy.

*2.4.4 Factors influencing NIR predictive ability*

Table 2.2 and Table 2.3 show the calibration and cross-validation statistics for NIR calibration models for whole soybean samples. ARG, ASP, GLU, HIS, ILE, LEU, LYS, PRO, THR and TYR in dry weight basis had higher $R^2_c$, $R^2_{cv}$ values than those amino acids in as received basis while an inverse relationship was found for ALA, CYS, PHE, SER, TYR and crude protein. Except for LYS, the differences of evaluation parameters between the two groups were small, with $\Delta R^2_c$ values ranged from 0 to 0.019 and $\Delta R^2_{cv}$ ranged from 0.001 to 0.021. However, the calibration models based on dry weight content had higher RPD values than that based on as is content for all amino acids. This might suggest that the NIR calibration models could measure amino acids in dry weight content more accurately than those presented in "as received/fresh weight" basis.

Table 2.4 summarizes the evaluation parameters of calibration models for ground soybean samples. Compared to results from Table 2.2, most calibration models for ground samples had higher $R^2_c$ values than those for whole samples, and all amino acids except ARG for ground samples had higher $R^2_{cv}$ and lower SECV. Additionally, the RPD values for ground samples were higher than those for whole samples except ARG and SER. These results might indicate the grinding process improved the predictive ability of NIR calibration models in crude protein and amino acids content prediction in soybean, which was in agreement with previous reports (Pazdernik et al., 1997; Kovalenko et al., 2006).

Table 2.5 includes the NIR calibration and cross-validation statistics for defatted soybean meal by the DA 7250. Models for defatted soybean meal had higher $R^2_c$, $R^2_{cv}$ and lower SECV values than those for full-fat soybean meal for crude protein and most amino acids. The RPD values for defatted soybean meal were also higher than those for full-fat soybean meal except

HIS, ILE, PHE and TYR. These might suggest that the lipid extraction increased the accuracy of NIR spectroscopy with respect to crude protein and amino acids measurement.

Although grinding and lipid extraction have the potential to improve the predictive performance of NIR calibration models, the cost of preparation also increases, which contradicts the advantages of NIR spectroscopy, namely its ability to measure analytes in a non-destructive manner with minimal preparation. Therefore, it is important to balance the cost of analysis and accuracy when using NIR spectroscopy as a sample screening tool.

*2.4.5 Performance of different NIR spectroscopy instruments*

Table 2.6 shows the NIR calibration and cross-validation results for defatted soybean meal as assessed with FT 9700. Compared with results from Table 2.5, models for FT 9700 NIR spectroscopy instrument had lower $R^2_c$ values than those for DA 7250 for crude protein and most amino acids. The $R^2_{cv}$ and RPD values for FT 9700 seemed better: LEU, LYS, PHE, PRO, THR, TRP and TYR for FT 9700 showed higher $R^2_{cv}$ and RPD than those determined with the DA 7250. Overall, it might suggest that the DA 7250 had a better predictive ability for assessing crude protein and amino acid contents. However, as models developed with the DA 7250 had higher PLS factors, the calibration fitted the samples in the calibration set splendidly, but the prediction of samples derived from outside the calibration set might be poorer (Baianu et al., 2004). Further validation using an external soybean set may be required to compare the predictive abilities of models for the DA 7250 and FT 9700 NIR spectroscopy.

Crude protein and amino acid contents of soybean samples in the calibration set were analyzed using NIR calibration models based on both NIR spectroscopy instruments to determine the agreement between them. The agreement assessment was based on the method

reported by Bland & Altman (1986).

Table 2.7 shows the agreement assessment statistics, including the correlation coefficient (R), mean difference, standard deviation of differences, upper limit, lower limit, number of samples outside the limits and P value for paired student t-test. The R values between the two NIR instruments ranged from 0.85 to 0.97 ($P < 0.001$). For crude protein and most amino acids, the R values reached high levels (> 0.90), indicating a high strength of the relation between the two instruments. Mean differences were the average values of crude protein and amino acids differences between two NIR spectroscopy instruments for each sample. Differences for crude protein and most amino acids were tested to follow a Normal distribution using the Shapiro-Wilk test. Distributions of differences for other amino acids were regarded as being normally distributed since their W values were above 0.973 in the test. Upper and lower limits were set to a mean difference plus or minus two times the standard deviation of difference. Most differences (approximately 95%) would be expected to lie between the limits, and differences outside the limits were unacceptable for clinical purposes (Bland & Altman, 1986).

Based on Table 2.7, crude protein and amino acids had low values of mean difference (-0.002 to 0.007 for amino acids and 0.030 for crude protein). The number of samples outside the limits ranged from 12 to 22, close to 18 (5% of the sample size). Since the differences for crude protein and amino acids were Normally or likely Normally distributed, the number of outside samples might be accepted. The paired student t-test indicated no significant differences between the prediction of crude protein and amino acid contents by DA 7250 and FT 9700 ($P > 0.05$). However, TYR had the lowest P-value and highest number of samples outside the limits. These might suggest FT 9700 agreed well with DA 7250 in crude protein

and amino acids measurement, but more samples were required to confirm it in future studies.

## 2.5 Conclusion

This study evaluated the predictive ability of using NIR spectroscopy in crude protein and amino acids measurement in soybean. NIR calibration models could predict crude protein and most amino acid contents with acceptable accuracy. Models for CYS, MET and TRP analysis in soybean required further calibration to increase the utility. The R2c values in NIR calibration models were found to relate to the correlation between amino acids and crude protein. However, as R2cp values for amino acids were lower than R2c for those from NIR calibration models, the NIR spectroscopy was believed to explain more variance than crude protein regression. Grinding and lipid extraction were found to likely improve the NIR spectroscopy in crude protein and amino acids prediction, but the increased cost of preparation requires consideration. Data derived with the FT 9700 NIR spectroscopy instrument was suggested to agreed well with DA 7250, but more samples might be analyzed using them to confirm the relationship in future research.

Table 2.1. Statistics for reference amino acid and crude protein concentrations (dry weight content%) of soybean used for NIR calibration

| | Min | Mean | Max | S.D.[a] | CV[b] % | Linear regression of amino acids to crude protein | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | intercept | slope | $R^2_{cp}$[c] |
| ALA | 1.08 | 1.58 | 1.86 | 0.116 | 7.35 | 0.763 | 0.020 | 0.43 |
| ARG | 1.80 | 2.72 | 3.98 | 0.320 | 11.75 | 0.148 | 0.063 | 0.56 |
| ASP | 2.90 | 4.33 | 5.43 | 0.393 | 9.07 | 1.551 | 0.068 | 0.32 |
| CYS | 0.40 | 0.57 | 0.81 | 0.058 | 10.09 | 0.532 | 0.001[ns] | 0.00 |
| GLU | 4.60 | 6.78 | 8.61 | 0.652 | 9.62 | 1.844 | 0.121 | 0.50 |
| GLY | 1.07 | 1.64 | 2.26 | 0.145 | 8.84 | 0.644 | 0.024 | 0.41 |
| HIS | 0.59 | 0.85 | 1.41 | 0.099 | 11.69 | 0.393 | 0.011 | 0.18 |
| ILE | 1.16 | 1.72 | 2.11 | 0.140 | 8.15 | 0.618 | 0.027 | 0.54 |
| LEU | 2.02 | 2.90 | 3.54 | 0.246 | 8.48 | 1.330 | 0.038 | 0.35 |
| LYS | 1.67 | 2.44 | 2.91 | 0.195 | 7.98 | 1.149 | 0.032 | 0.38 |
| MET | 0.36 | 0.57 | 0.76 | 0.047 | 8.31 | 0.241 | 0.008 | 0.42 |
| PHE | 1.32 | 1.92 | 2.81 | 0.184 | 9.58 | 0.714 | 0.029 | 0.37 |
| PRO | 1.35 | 1.92 | 2.40 | 0.175 | 9.11 | 0.773 | 0.028 | 0.37 |
| SER | 1.37 | 1.97 | 2.57 | 0.176 | 8.92 | 0.955 | 0.025 | 0.29 |
| THR | 1.08 | 1.51 | 1.92 | 0.113 | 7.47 | 0.871 | 0.016 | 0.28 |
| TRP | 0.35 | 0.51 | 0.68 | 0.049 | 9.53 | 0.359 | 0.004 | 0.08 |
| TYR | 0.95 | 1.38 | 2.04 | 0.125 | 9.07 | 0.605 | 0.019 | 0.33 |
| VAL | 1.17 | 1.79 | 2.24 | 0.156 | 8.72 | 0.372 | 0.035 | 0.72 |
| Crude Protein | 28.75 | 40.87 | 51.60 | 3.807 | 9.32 | | | |

[a]Standard deviation

[b]Coefficient of variation

[c]Coefficient of determination in regression of amino acids to crude protein

ns = non-significant

Table 2.2. Calibration and cross-validation statistics in PerkinElmer DA 7250 NIR models for crude protein and amino acids measurement (dry basis%) in whole soybean

| | | | | | | Calibration | | Cross-validation | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Compound | Mean | S.D. | Min | Max | N | SEC | $R^2_c$ | SECV | $R^2_{cv}$ | RPD | Factor |
| ALA | 1.59 | 0.106 | 1.23 | 1.86 | 353 | 0.058 | 0.702 | 0.062 | 0.652 | 1.68 | 13 |
| ARG | 2.72 | 0.307 | 1.89 | 3.93 | 354 | 0.129 | 0.824 | 0.141 | 0.790 | 2.07 | 13 |
| ASP | 4.34 | 0.375 | 3.09 | 5.43 | 353 | 0.174 | 0.785 | 0.190 | 0.744 | 1.91 | 13 |
| CYS | 0.57 | 0.055 | 0.43 | 0.74 | 355 | 0.034 | 0.611 | 0.039 | 0.484 | 1.41 | 15 |
| GLU | 6.79 | 0.627 | 4.60 | 8.61 | 354 | 0.260 | 0.828 | 0.284 | 0.796 | 2.12 | 14 |
| GLY | 1.64 | 0.131 | 1.17 | 1.97 | 352 | 0.067 | 0.734 | 0.074 | 0.679 | 1.65 | 14 |
| HIS | 0.84 | 0.086 | 0.62 | 1.08 | 354 | 0.050 | 0.656 | 0.054 | 0.606 | 1.61 | 12 |
| ILE | 1.73 | 0.130 | 1.28 | 2.11 | 354 | 0.061 | 0.778 | 0.068 | 0.729 | 1.86 | 14 |
| LEU | 2.90 | 0.233 | 2.12 | 3.54 | 353 | 0.105 | 0.798 | 0.115 | 0.757 | 1.99 | 14 |
| LYS | 2.45 | 0.183 | 1.81 | 2.91 | 354 | 0.110 | 0.640 | 0.119 | 0.583 | 1.54 | 13 |
| MET | 0.57 | 0.043 | 0.42 | 0.69 | 355 | 0.027 | 0.594 | 0.031 | 0.476 | 1.35 | 15 |
| PHE | 1.92 | 0.170 | 1.37 | 2.32 | 353 | 0.080 | 0.778 | 0.088 | 0.730 | 1.89 | 14 |
| PRO | 1.93 | 0.167 | 1.41 | 2.40 | 352 | 0.067 | 0.840 | 0.073 | 0.809 | 2.08 | 14 |
| SER | 1.97 | 0.166 | 1.44 | 2.43 | 352 | 0.072 | 0.811 | 0.079 | 0.771 | 2.05 | 14 |
| THR | 1.52 | 0.103 | 1.21 | 1.79 | 353 | 0.054 | 0.729 | 0.059 | 0.672 | 1.73 | 14 |
| TRP | 0.51 | 0.048 | 0.36 | 0.68 | 356 | 0.037 | 0.406 | 0.039 | 0.350 | 1.23 | 9 |
| TYR | 1.38 | 0.113 | 1.01 | 1.68 | 354 | 0.068 | 0.639 | 0.074 | 0.576 | 1.58 | 12 |
| VAL | 1.80 | 0.142 | 1.33 | 2.24 | 352 | 0.061 | 0.814 | 0.070 | 0.761 | 1.99 | 16 |
| Crude protein | 40.90 | 3.704 | 28.85 | 51.60 | 356 | 1.171 | 0.900 | 1.428 | 0.852 | 2.68 | 19 |

S.D.: standard deviation; N: number of samples used for calibration; SEC: standard error of calibration; $R^2_c$: coefficient of determination in calibration; SECV: standard error of cross-validation; $R^2_{cv}$: coefficient of determination in cross-validation; RPD: residual predictive deviation.

Table 2.3. Calibration and cross-validation statistics in PerkinElmer DA 7250 NIR models for crude protein and amino acids measurement (as is basis%) in whole soybean

| | | | | | | Calibration | | Cross-validation | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Compound | Mean | S.D. | Min | Max | N | SEC | $R^2_c$ | SECV | $R^2_{cv}$ | RPD | Factor |
| ALA | 1.45 | 0.094 | 1.14 | 1.69 | 350 | 0.051 | 0.711 | 0.056 | 0.656 | 1.68 | 13 |
| ARG | 2.49 | 0.273 | 1.74 | 3.46 | 354 | 0.116 | 0.819 | 0.128 | 0.784 | 2.13 | 13 |
| ASP | 3.97 | 0.340 | 2.90 | 5.02 | 353 | 0.158 | 0.783 | 0.174 | 0.739 | 1.96 | 14 |
| CYS | 0.52 | 0.051 | 0.40 | 0.68 | 355 | 0.032 | 0.617 | 0.037 | 0.494 | 1.38 | 15 |
| GLU | 6.21 | 0.568 | 4.29 | 7.95 | 354 | 0.243 | 0.816 | 0.268 | 0.776 | 2.12 | 14 |
| GLY | 1.50 | 0.117 | 1.09 | 1.83 | 351 | 0.060 | 0.737 | 0.067 | 0.676 | 1.75 | 14 |
| HIS | 0.77 | 0.079 | 0.58 | 0.97 | 354 | 0.047 | 0.637 | 0.051 | 0.576 | 1.55 | 12 |
| ILE | 1.58 | 0.116 | 1.19 | 1.90 | 353 | 0.056 | 0.765 | 0.061 | 0.727 | 1.90 | 12 |
| LEU | 2.66 | 0.212 | 1.99 | 3.26 | 353 | 0.097 | 0.793 | 0.108 | 0.746 | 1.97 | 14 |
| LYS | 2.24 | 0.166 | 1.68 | 2.68 | 354 | 0.104 | 0.605 | 0.110 | 0.556 | 1.51 | 10 |
| MET | 0.52 | 0.038 | 0.39 | 0.64 | 355 | 0.024 | 0.595 | 0.028 | 0.467 | 1.36 | 16 |
| PHE | 1.75 | 0.156 | 1.25 | 2.13 | 353 | 0.073 | 0.779 | 0.081 | 0.734 | 1.92 | 14 |
| PRO | 1.76 | 0.153 | 1.29 | 2.21 | 353 | 0.063 | 0.831 | 0.071 | 0.788 | 2.15 | 15 |
| SER | 1.81 | 0.152 | 1.35 | 2.22 | 351 | 0.064 | 0.820 | 0.073 | 0.773 | 2.08 | 15 |
| THR | 1.39 | 0.095 | 1.12 | 1.66 | 353 | 0.049 | 0.729 | 0.055 | 0.661 | 1.73 | 15 |
| TRP | 0.47 | 0.043 | 0.35 | 0.61 | 354 | 0.033 | 0.410 | 0.035 | 0.349 | 1.22 | 9 |
| TYR | 1.26 | 0.103 | 0.94 | 1.54 | 353 | 0.061 | 0.642 | 0.067 | 0.578 | 1.53 | 13 |
| VAL | 1.65 | 0.126 | 1.23 | 2.02 | 351 | 0.055 | 0.808 | 0.063 | 0.750 | 1.99 | 16 |
| Crude protein | 37.41 | 3.160 | 26.87 | 46.38 | 357 | 0.938 | 0.912 | 1.140 | 0.871 | 2.77 | 18 |

S.D.: standard deviation; N: number of samples used for calibration; SEC: standard error of calibration; $R^2_c$: coefficient of determination in calibration; SECV: standard error of cross-validation; $R^2_{cv}$: coefficient of determination in cross-validation; RPD: residual predictive deviation.

Table 2.4. Calibration and cross-validation statistics in PerkinElmer DA 7250 NIR models for crude protein and amino acids measurement (dry basis%) in ground soybean

| | | | | | | Calibration | | Cross-validation | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Compound | Mean | S.D. | Min | Max | N | SEC | $R^2_c$ | SECV | $R^2_{cv}$ | RPD | Factor |
| ALA | 1.59 | 0.106 | 1.23 | 1.86 | 352 | 0.057 | 0.707 | 0.061 | 0.664 | 1.73 | 11 |
| ARG | 2.72 | 0.304 | 1.80 | 3.74 | 355 | 0.133 | 0.810 | 0.145 | 0.774 | 2.10 | 11 |
| ASP | 4.34 | 0.377 | 3.09 | 5.43 | 354 | 0.174 | 0.786 | 0.186 | 0.759 | 2.02 | 11 |
| CYS | 0.57 | 0.055 | 0.43 | 0.74 | 355 | 0.036 | 0.561 | 0.039 | 0.486 | 1.41 | 12 |
| GLU | 6.79 | 0.627 | 4.60 | 8.61 | 354 | 0.262 | 0.826 | 0.281 | 0.800 | 2.23 | 11 |
| GLY | 1.64 | 0.133 | 1.08 | 1.97 | 352 | 0.069 | 0.732 | 0.074 | 0.692 | 1.79 | 11 |
| HIS | 0.84 | 0.086 | 0.61 | 1.08 | 353 | 0.046 | 0.712 | 0.050 | 0.660 | 1.72 | 12 |
| ILE | 1.73 | 0.130 | 1.28 | 2.11 | 351 | 0.056 | 0.810 | 0.061 | 0.779 | 2.13 | 12 |
| LEU | 2.90 | 0.235 | 2.12 | 3.54 | 354 | 0.106 | 0.794 | 0.115 | 0.764 | 2.04 | 12 |
| LYS | 2.45 | 0.187 | 1.71 | 2.91 | 355 | 0.109 | 0.658 | 0.118 | 0.606 | 1.58 | 11 |
| MET | 0.57 | 0.043 | 0.42 | 0.69 | 355 | 0.026 | 0.615 | 0.029 | 0.542 | 1.47 | 13 |
| PHE | 1.92 | 0.169 | 1.37 | 2.32 | 352 | 0.075 | 0.802 | 0.081 | 0.770 | 2.09 | 11 |
| PRO | 1.92 | 0.169 | 1.38 | 2.40 | 353 | 0.068 | 0.839 | 0.073 | 0.814 | 2.31 | 10 |
| SER | 1.97 | 0.165 | 1.44 | 2.43 | 353 | 0.075 | 0.796 | 0.079 | 0.775 | 2.09 | 9 |
| THR | 1.52 | 0.104 | 1.20 | 1.78 | 354 | 0.055 | 0.718 | 0.059 | 0.680 | 1.76 | 10 |
| TRP | 0.51 | 0.047 | 0.35 | 0.66 | 355 | 0.035 | 0.457 | 0.037 | 0.398 | 1.27 | 9 |
| TYR | 1.38 | 0.113 | 1.01 | 1.68 | 353 | 0.066 | 0.656 | 0.069 | 0.629 | 1.64 | 8 |
| VAL | 1.80 | 0.147 | 1.24 | 2.24 | 353 | 0.061 | 0.827 | 0.065 | 0.802 | 2.26 | 12 |
| Crude protein | 40.94 | 3.728 | 28.85 | 51.60 | 357 | 0.832 | 0.950 | 0.969 | 0.933 | 3.85 | 17 |

S.D.: standard deviation; N: number of samples used for calibration; SEC: standard error of calibration; $R^2_c$: coefficient of determination in calibration; SECV: standard error of cross-validation; $R^2_{cv}$: coefficient of determination in cross-validation; RPD: residual predictive deviation.

Table 2.5. Calibration and cross-validation statistics in PerkinElmer DA 7250 NIR models for crude protein and amino acids measurement (dry basis%) in ground and defatted soybean

| | | | | | | Calibration | | Cross-validation | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Compound | Mean | S.D. | Min | Max | N | SEC | $R^2_c$ | SECV | $R^2_{cv}$ | RPD | Factor |
| ALA | 1.59 | 0.108 | 1.14 | 1.86 | 352 | 0.049 | 0.791 | 0.057 | 0.715 | 1.89 | 14 |
| ARG | 2.72 | 0.306 | 1.89 | 3.93 | 354 | 0.121 | 0.843 | 0.145 | 0.776 | 2.11 | 15 |
| ASP | 4.34 | 0.373 | 3.07 | 5.43 | 351 | 0.150 | 0.839 | 0.173 | 0.786 | 2.16 | 14 |
| CYS | 0.57 | 0.055 | 0.43 | 0.74 | 355 | 0.032 | 0.667 | 0.037 | 0.553 | 1.48 | 15 |
| GLU | 6.80 | 0.624 | 4.60 | 8.61 | 352 | 0.218 | 0.879 | 0.258 | 0.830 | 2.42 | 15 |
| GLY | 1.64 | 0.129 | 1.17 | 1.96 | 351 | 0.064 | 0.757 | 0.070 | 0.710 | 1.85 | 11 |
| HIS | 0.84 | 0.087 | 0.59 | 1.08 | 353 | 0.043 | 0.758 | 0.052 | 0.651 | 1.67 | 15 |
| ILE | 1.73 | 0.129 | 1.28 | 2.11 | 352 | 0.056 | 0.809 | 0.061 | 0.780 | 2.12 | 10 |
| LEU | 2.90 | 0.233 | 2.14 | 3.54 | 353 | 0.091 | 0.847 | 0.112 | 0.772 | 2.08 | 15 |
| LYS | 2.45 | 0.182 | 1.81 | 2.91 | 354 | 0.106 | 0.661 | 0.115 | 0.602 | 1.58 | 11 |
| MET | 0.57 | 0.043 | 0.42 | 0.69 | 354 | 0.027 | 0.611 | 0.029 | 0.550 | 1.47 | 10 |
| PHE | 1.92 | 0.169 | 1.35 | 2.32 | 352 | 0.078 | 0.790 | 0.085 | 0.748 | 1.99 | 11 |
| PRO | 1.93 | 0.168 | 1.40 | 2.40 | 352 | 0.058 | 0.880 | 0.070 | 0.828 | 2.40 | 15 |
| SER | 1.97 | 0.165 | 1.44 | 2.43 | 352 | 0.069 | 0.827 | 0.075 | 0.794 | 2.20 | 11 |
| THR | 1.52 | 0.104 | 1.21 | 1.79 | 354 | 0.053 | 0.741 | 0.057 | 0.695 | 1.82 | 11 |
| TRP | 0.51 | 0.047 | 0.36 | 0.66 | 356 | 0.035 | 0.463 | 0.036 | 0.415 | 1.31 | 8 |
| TYR | 1.38 | 0.113 | 1.01 | 1.68 | 353 | 0.069 | 0.631 | 0.072 | 0.592 | 1.57 | 8 |
| VAL | 1.80 | 0.145 | 1.33 | 2.24 | 352 | 0.053 | 0.866 | 0.063 | 0.811 | 2.30 | 15 |
| Crude protein | 40.94 | 3.719 | 28.85 | 51.60 | 357 | 0.811 | 0.952 | 0.957 | 0.934 | 3.89 | 15 |

S.D.: standard deviation; N: number of samples used for calibration; SEC: standard error of calibration; $R^2_c$: coefficient of determination in calibration; SECV: standard error of cross-validation; $R^2_{cv}$: coefficient of determination in cross-validation; RPD: residual predictive deviation.

Table 2.6. Calibration and cross-validation statistics in PerkinElmer FT 9700 NIR models for crude protein and amino acids measurement (dry basis%) in ground and defatted soybean

| | | | | | | Calibration | | Cross-validation | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Compound | Mean | S.D. | Min | Max | N | SEC | $R^2_c$ | SECV | $R^2_{cv}$ | RPD | Factor |
| ALA | 1.59 | 0.108 | 1.12 | 1.86 | 352 | 0.057 | 0.719 | 0.060 | 0.696 | 1.80 | 5 |
| ARG | 2.72 | 0.310 | 1.80 | 3.93 | 354 | 0.123 | 0.842 | 0.152 | 0.761 | 2.04 | 9 |
| ASP | 4.34 | 0.377 | 3.09 | 5.43 | 352 | 0.169 | 0.800 | 0.179 | 0.777 | 2.11 | 6 |
| CYS | 0.57 | 0.055 | 0.43 | 0.74 | 355 | 0.039 | 0.498 | 0.041 | 0.451 | 1.34 | 6 |
| GLU | 6.80 | 0.632 | 4.60 | 8.61 | 354 | 0.256 | 0.836 | 0.276 | 0.811 | 2.29 | 7 |
| GLY | 1.64 | 0.133 | 1.08 | 1.96 | 352 | 0.067 | 0.743 | 0.074 | 0.692 | 1.79 | 7 |
| HIS | 0.84 | 0.086 | 0.61 | 1.08 | 353 | 0.053 | 0.621 | 0.056 | 0.579 | 1.54 | 6 |
| ILE | 1.73 | 0.132 | 1.18 | 2.11 | 353 | 0.060 | 0.795 | 0.063 | 0.773 | 2.10 | 6 |
| LEU | 2.90 | 0.235 | 2.12 | 3.54 | 353 | 0.104 | 0.802 | 0.111 | 0.776 | 2.11 | 7 |
| LYS | 2.45 | 0.185 | 1.71 | 2.91 | 354 | 0.109 | 0.652 | 0.116 | 0.611 | 1.59 | 6 |
| MET | 0.57 | 0.043 | 0.42 | 0.69 | 355 | 0.031 | 0.482 | 0.031 | 0.467 | 1.38 | 3 |
| PHE | 1.92 | 0.171 | 1.35 | 2.32 | 352 | 0.081 | 0.778 | 0.086 | 0.748 | 1.99 | 7 |
| PRO | 1.93 | 0.169 | 1.38 | 2.40 | 352 | 0.064 | 0.854 | 0.069 | 0.836 | 2.45 | 7 |
| SER | 1.97 | 0.166 | 1.44 | 2.43 | 352 | 0.071 | 0.818 | 0.076 | 0.791 | 2.18 | 7 |
| THR | 1.52 | 0.105 | 1.09 | 1.78 | 353 | 0.052 | 0.752 | 0.057 | 0.712 | 1.84 | 7 |
| TRP | 0.51 | 0.046 | 0.38 | 0.63 | 353 | 0.033 | 0.481 | 0.034 | 0.445 | 1.35 | 5 |
| TYR | 1.38 | 0.115 | 0.95 | 1.68 | 353 | 0.070 | 0.632 | 0.073 | 0.593 | 1.57 | 6 |
| VAL | 1.80 | 0.150 | 1.24 | 2.24 | 352 | 0.064 | 0.815 | 0.068 | 0.797 | 2.20 | 6 |
| Crude protein | 40.90 | 3.771 | 28.75 | 51.60 | 357 | 1.225 | 0.894 | 1.336 | 0.875 | 2.82 | 8 |

S.D.: standard deviation; N: number of samples used for calibration; SEC: standard error of calibration; $R^2_c$: coefficient of determination in calibration; SECV: standard error of cross-validation; $R^2_{cv}$: coefficient of determination in cross-validation; RPD: residual predictive deviation

Table 2.7. Statistics of agreement assessment between DA 7250 and FT 9700 NIR spectroscopy instruments

| Compound | R | Mean | S.D. | Upper limit | Lower limit | N | P |
|---|---|---|---|---|---|---|---|
| ALA | 0.93 | 0.000* | 0.035 | 0.070 | -0.069 | 17 | 0.848 |
| ARG | 0.93 | 0.004* | 0.103 | 0.211 | -0.203 | 16 | 0.476 |
| ASP | 0.95 | 0.003 | 0.107 | 0.217 | -0.212 | 17 | 0.642 |
| CYS | 0.85 | 0.000* | 0.024 | 0.048 | -0.047 | 19 | 0.848 |
| GLU | 0.96 | 0.007 | 0.174 | 0.355 | -0.341 | 15 | 0.436 |
| GLY | 0.96 | 0.000* | 0.032 | 0.063 | -0.064 | 16 | 0.788 |
| HIS | 0.88 | 0.001* | 0.035 | 0.071 | -0.070 | 17 | 0.786 |
| ILE | 0.96 | 0.001* | 0.031 | 0.062 | -0.061 | 21 | 0.647 |
| LEU | 0.93 | 0.000 | 0.077 | 0.153 | -0.154 | 18 | 0.984 |
| LYS | 0.95 | 0.000* | 0.046 | 0.092 | -0.092 | 21 | 0.924 |
| MET | 0.90 | 0.000 | 0.015 | 0.029 | -0.030 | 19 | 0.588 |
| PHE | 0.96 | 0.000 | 0.041 | 0.081 | -0.081 | 15 | 0.959 |
| PRO | 0.96 | -0.002 | 0.044 | 0.087 | -0.090 | 16 | 0.494 |
| SER | 0.97 | 0.002 | 0.036 | 0.073 | -0.070 | 16 | 0.369 |
| THR | 0.96 | 0.001 | 0.027 | 0.054 | -0.052 | 17 | 0.613 |
| TRP | 0.92 | 0.000* | 0.012 | 0.024 | -0.024 | 12 | 0.633 |
| TYR | 0.96 | 0.002 | 0.026 | 0.055 | -0.051 | 22 | 0.175 |
| VAL | 0.94 | 0.000* | 0.047 | 0.095 | -0.095 | 18 | 0.965 |
| Crude protein | 0.96 | 0.030* | 1.058 | 2.146 | -2.087 | 13 | 0.594 |

*Difference follows Normal distribution

R: correlation coefficient; Mean: mean difference; S.D.: standard deviation of difference; N: number of samples outside the limits; P: P value for paired student t-test

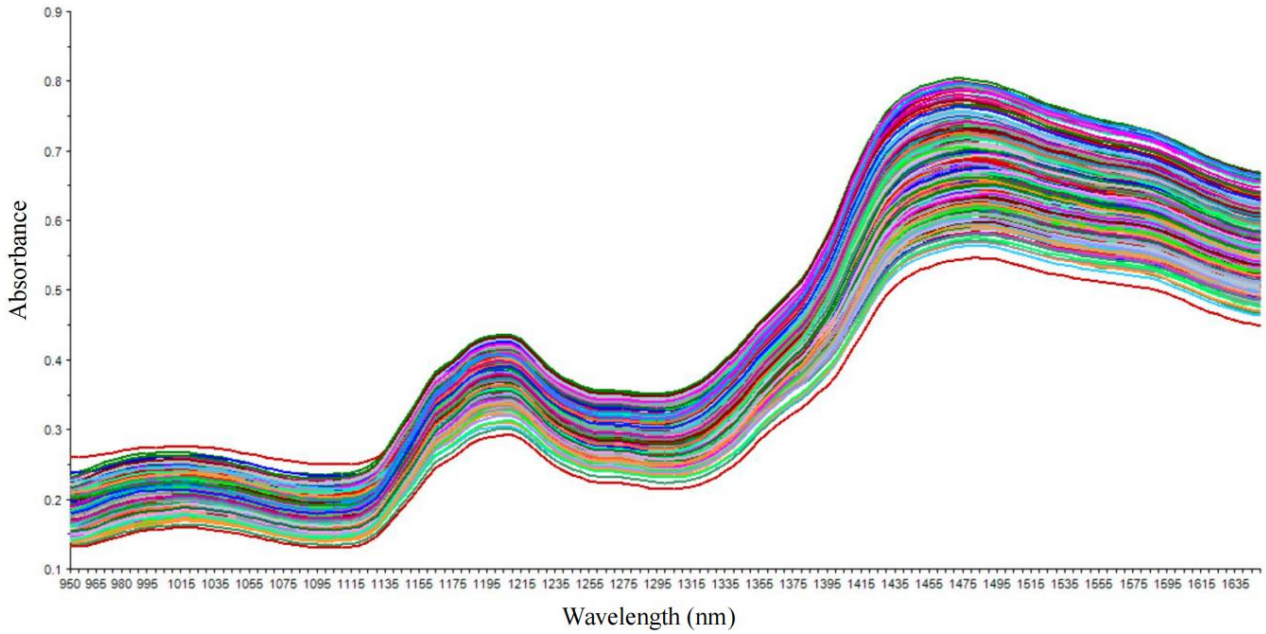


Figure 2.1. NIR spectra for whole soybean from calibration set scanned by PerkinElmer DA 7250

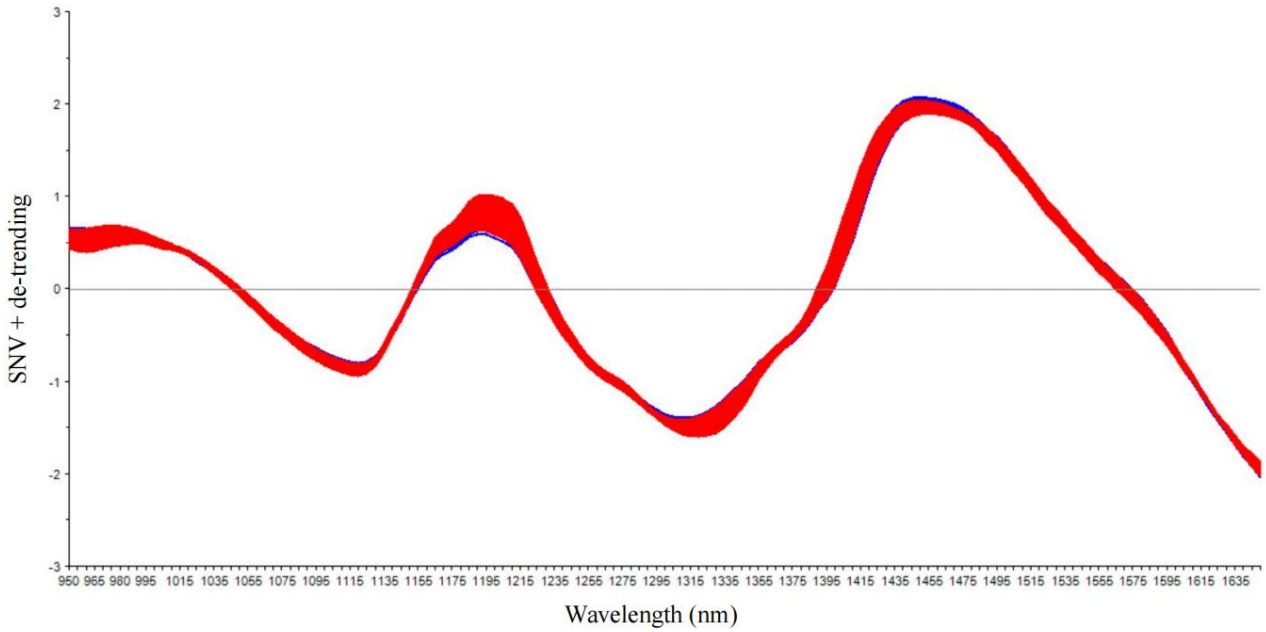



Figure 2.2. Standard Normal Variate and de-trending treated NIR spectra for whole soybean from calibration set scanned by PerkinElmer DA 7250

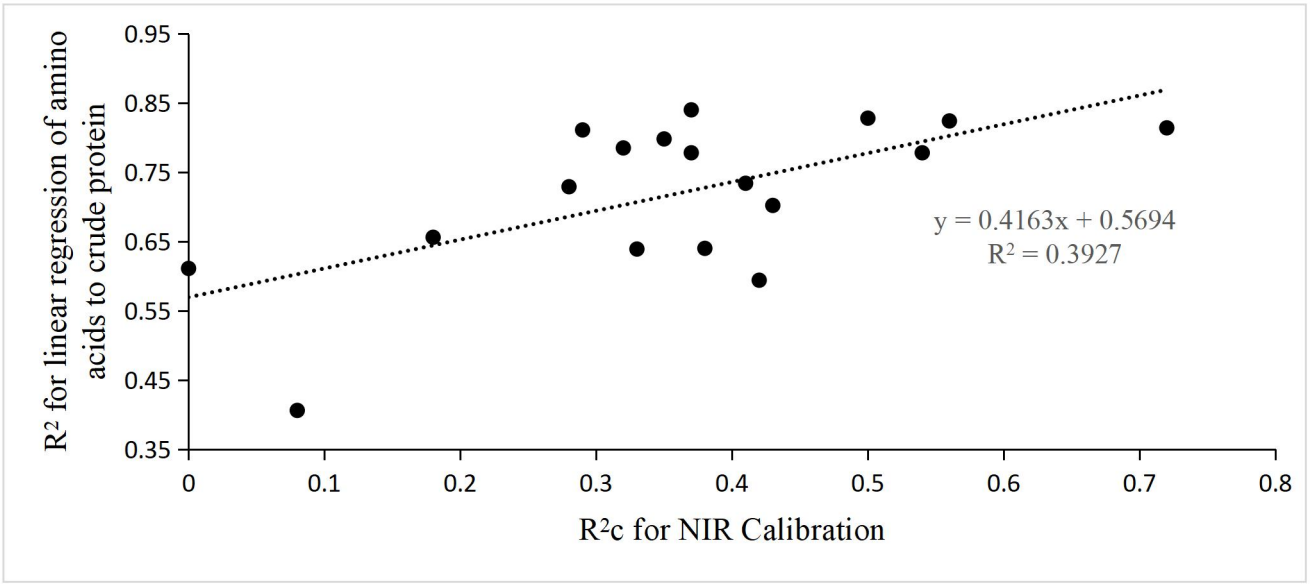



Fig 2.3. $R^2$ for linear regression of amino acids to crude protein versus $R^2c$ for NIR calibration on whole soybean samples by PerkinElmer DA 7250

# Chapter 3 Effects of Genotype, Environment and Their Interaction on Protein and Amino Acid contents in Soybean

## 3.1 Abstract

Soybean [*Glycine max (L.) Merr.*] is an important source of protein and oil. Genotypes, environment and the interaction of genotype × environment can influence protein composition. The main objectives of this study were to i) study the influence of genotype, environment and their interaction on soybean protein and amino acids content and ii) evaluate the stability of soybean genotype across various environments. Twenty-three soybean genotypes were grown at four locations in Manitoba, Canada for two years (2018 and 2019). Soybean grain were analyzed for protein and amino acids (nine essential amino acids and cysteine). The effects of genotype, environment and genotype × environment interactions on all traits were significant ($P < 0.05$). Genotype and environments explained the main part of variation for all traits. G13 and G15 cultivars performed better in favorable environments (bi > 1), and G22 cultivar showed greater resistance to environmental change. Protein and amino acids responded differently to various environments, but the optimal environments for greater soybean protein and amino acid remains to be established. Critical amino acid value (CAAV) was negatively linearly correlated with protein content in soybean (k = -0.17). This work provided possibility for marketing low protein soybean based on its critical amino acid value. Overall, genotype, environment and genotype × environment interactions affect soybean protein and amino acid concentrations, and

the relationship of CAAV with protein content and the variation of CAAV within soybean genotypes vary from different environments.

### 3.2 Introduction

Soybean [*Glycine max (L.) Merr.*] is an important crop worldwide containing high quality protein and oil content (Arslanoglu et al., 2011). It has the highest protein content (40%) among all food crops, and its oil content (20%) is second only to groundnut within the legumes (Gurmu et al., 2009). Furthermore, soybean is rich in minerals, especially calcium, phosphorus and iron and soy-based bioactives including isoflavones (Wang & Murphy, 1994; Slinkard & Knott, 1995; Ogoke et al., 2003). Consumption of soybean can reduce the risk of several chronic diseases such as cancer, diabetes, obesity, etc. and has been reported to improve bone health (Natarajan et al., 2016).

Currently, an increasing number of people from wealthy countries are trying to consume more plant-based foods. In 2017, a global consumer study showed that 39% of Americans and 43% of Canadians were planning to consume more plant-based foods (Doris, 2018). Moreover, the new Canada's Food Guide, published by Health Canada, encourages people to consume more fruits and vegetables as well as plant-based proteins, such as pulses and nuts, more often (Government of Canada, 2019). For food purposes, it is desirable to use soybean containing a higher protein content, low oil content, lighter grain coat and a clear hilum (Liu et al., 1995). This has increased the demand of soybean for soy-based foods, emphasizing the need for high protein soybean cultivars.

It has been reported that soybean protein content is affected by the growing environment and by genotype. Soybean composition can be impacted by environmental factors including temperature, humidity, insect/pathogen infection, water stress, and soil nutrients (El-Din et al., 2002; Krishnan et al., 2007; Stevenson et al., 2012). Qin et al. (2014) identified that soybean grown in Southern China had higher protein and greater amino acid contents as compared to soybean grown in Northern China. Kumar et al. (2006) studied soybean from India and identified a negative correlation between latitude and protein. Except for latitude, Boydak et al. (2002) found that less irrigation resulted in lower protein content (37.6% vs. 38.6% in soybean under more irrigation) in soybean. Also, Carrera et al. (2009) concluded that diminution of precipitation minus potential evapotranspiration during the reproductive period (pp-PETR1R7) occurs with decreased protein content in soybean.

Temperature effects on protein concentration during soybean growth has been established as well. Mourtzinis et al. (2017) noted increased average ambient temperatures between developmental stages R5 and R8 reduced protein in soybean. Piper & Boote (1999) pointed out the relationship between temperature and soy protein content followed a quadratic response.

The environmental effects on protein content vary with different soybean genotypes. The interaction between genetic structure and growing environment is known as genotype × environment interaction, which results in significant differences in the performance of genotypes when evaluated in different locations (Gauch & Zobel, 1997; Arslanoglu et al., 2011). Arslanoglu et al. (2011) reported a significant effect of genotype × environment interaction on protein content in soybean. Natarajan

et al. (2016) grew 27 soybean varieties at two distinct time periods and investigated genotype ×sowing date interactions impact soybean grain protein. A relatively recent study tested two recombinant inbred line soybean populations under five environmental conditions and found significant genotype × environment effects for both grain protein concentration and yield (Whaley & Eskandari, 2019).

Soybean grain amino acid content is another main factor that determines soybean quality (Bellaloui et al., 2009). Soybean protein contains of all of the essential amino acids (Carrera et al., 2011), and the sum of essential amino acids and conditionally amino acids constitute approximately 20% of the soybean grain protein (Tessari et al., 2016). Previous studies investigated the environmental effects on soybean protein composition (Wolf et al., 1982; Karr-Lilienthal et al., 2005; Carrera et al., 2011; Assefa et al., 2018). However, there is limited research exploring how genotype × environment interaction affects soybean amino acid concentrations. The study of soybean amino acid composition is essential for both nutritional purposes and consumer acceptance of soy food products (Gao et al., 2011).

The objectives of this study were to i) study the effects of genotype, growing environment and their interactions on soybean protein and amino acid contents and ii) evaluate stability of soybean grain protein of genotypes across environments.

## 3.3 Materials and Methods

### *3.3.1 Plant materials*

Soybean grain samples were collected and obtained from Manitoba Pulse and Soybean Growers Association. All samples were stored under cold condition (-20℃)

before analysis Twenty-three soybean genotypes including conventional (CN) and glyphosate-tolerant (GT) groups were used for the study. The first group, conventional soybean, consisted of 8 genotypes: AAC Halli, DH401, Meteor, OAC Prudence, OT 16-02, SC10-11.97, SVX17T000S1 and SVX17T0S12. The second group, glyphosate-tolerant soybean, which is tollerant against Monsanto's Roundup herbicide, consisted of 15 genotypes: Akras R2, DKB003-29, Foote R2, LS 001XT, Mahony R2, Nocoma R2, Notus R2, P005A27X, P006A37X, Prince R2X, PS 0044 XRN, S0009-M2, S006-W5, S007-Y4 and Torro R2 (For some reasons, the name of each variety will be anonymous and replaced with G1-G23). Each Soybean genotype was planted in four locations, Arborg (Abg), Carman (Car), Morris (Mor) and Ste Adolphe (Sta) in Manitoba during 2018 and 2019. The coordinate information and environment data (temperature, rainfall and solar radiation) during May to September for each location are summarized in Table 3.1. Environment data were provided by Manitoba Agriculture and Resource Development. A randomized complete block design (RCBD) with three replications was used to conduct the experiments. The soil type in Abg, Mor and Sta was clay and Car was clay-loam. Plot size for Car and Mor was 11.25 $m^2$ in 2018 and 9.20 $m^2$ with row spacing of in 2019, with row spacing of 9 cm for both cropping years. For Abg, the plot size in 2018 was 10.96 $m^2$ and 2019 was 12.16 $m^2$, with spacing row of 9 cm for both years. Sta had the plot size of 9.00 $m^2$ and spacing row of 7 cm for both years.

*3.3.2 Protein and amino acid analysis*

Protein and amino acid content of soybean grain were measured using the

PerkinElmer DA 7250 diode array near infrared (NIR) system that was previously calibrated by our group. The NIR calibration models were made using 360 soybean samples planted in 13 locations from Manitoba during 2018 and 2019. Soybean grain compositions for NIR calibration set were analyzed by the official reference method. Crude protein content was determined using Dumas combustion approach (Dumas method, AOAC 990.03), and amino acid content was measured by official methods AOAC 982.30 E(a,b) Ch. 45.3.05 (AOAC, 1995) and ISO 13904:2005(E) (ISO, 2005). Spectra and reference data were correlated by partial least square regression to establish NIR calibration models, which were used for protein and amino acid profile analysis in the current experiment. Full details on the analytical methods are presented in Section 2.3.2.

All 552 soybean grain samples were scanned with the NIR instrument using a 150 mm (400 ml) diameter cup, and crude protein and amino acid contents (on dry weight basis) for each sample were read directly from the NIR analyzer.

*3.3.3 Statistical analysis*

An analysis of variance (ANOVA) was conducted for protein and ten amino acids (nine essential amino acids + cysteine) using PROC MIXED in SAS 9.4. Genotype was considered a fixed effect, whereas year, locations and blocks were considered random effects. The statistical model used for the ANOVA was:

$$Y_{ijkl} = \mu + G_i + L_j + Y_k + GL_{ij} + GY_{ik} + LY_{jk} + GLY_{ijk} + B_{l(jk)} + \varepsilon_{ijkl}$$

Where, $Y_{ijkl}$ = observed protein or amino acid value of genotype i in block l of location j in year k, $\mu$ = grand mean, $G_i$ = genotype effect, $L_j$ = location effect, $Y_k$ =

year effect, $GL_{ij}$ = the interaction effect of genotype i with location j, $GY_{ik}$ = the interaction effect of genotype i with year k, $LY_{jk}$ = the interaction effect of location j with year k, $GLY_{ijk}$ = the interaction effect of genotype i with location j and year k, $B_{l(jk)}$ = the effect of block l in location j from year k, $\varepsilon_{ijkl}$ = error or residual effect of genotype i in block l of location j in year k.

Variance components of genotype, location, year and their interactions for each of the ten amino acids and crude protein were estimated by Restricted maximum likelihood (REML) method in PROC MIXED. Least-squares means of protein and amino acid contents for each soybean genotype and environment were obtained from PROC MIXED, and means were compared by using Tukey's test at a significant level of 0.05. Pearson correlation coefficients between soybean protein and amino acid contents across environments were generated using least-squares means. For stability analysis, parameters were estimated by linear regression analysis, according to Eberhart and Russell (1966), where the mean protein content of each soybean genotype under individual environment was set as the dependent variable and the environmental index was the independent variable. The environmental index for each environment was calculated as the mean protein content of all genotypes at that environment minus the grand mean.

## 3.4 Results and Discussion

### *3.4.1 Analysis of Variance*

The results of the ANOVA indicated that the effects of genotype and environment (location, year and location × year) were significant at the $P<0.01$ level

for both protein and amino acid contents in soybean grains (Table 3.2). Based on the results presented in Table 3.2, genotype × environment (including genotype × location, genotype × year and genotype × location × year) interaction effects were found to significantly affect agronomic traits in soybean, which agreed with results reported in previous studies (Yan et al., 2010; Arslanoglu et al., 2011; Whale & Eskandari, 2019). Nevertheless, genotype × location and genotype × year interactions were not significant for all amino acids. The genotype × location interaction was significant ($P<0.05$) for Lys, Thr, Cys, Trp, Phe, Val, Leu, Ile and His while genotype × year interaction was only found to be significant for protein, Met, Cys, Trp and Val ($P<0.05$). Genotype × location × year effect was significant for soybean protein and amino acids concentration at different probability levels (0.05 or 0.01).

Despite genotype, environment and genotype × environment interaction showing significant effects on soybean protein and amino acid contents, the variance components in Table 3.3 were more informative to explain the relative importance of each effect (Lee et al., 2002). Based on Table 3.3, genotype had the highest variance component for soybean protein and amino acids, and was the predominant source of variance. The environment effect for soybean protein and amino acids were relatively less important than the genotype effect but more important than the genotype × environment effect. Year or location × year effect dominated the main variance among environment effects. For protein, Cys, Val and His, the year effect was more important than location × year interaction, while an inverse relationship was observed for the rest of the amino acids. Within the genotype × environment interaction, the genotype × location × year interaction was the predominant source, which had a higher variance

component than genotype × location and genotype × year. Location and genotype × location effects showed little importance for protein and any of the amino acids.

*3.4.2 Stability Analysis*

Stability demonstrates the degree to which the performance of a genotype is similar to the estimated or predicted level (Becker & Leon, 1988). According to Eberhart and Russell (1966), the regression coefficient ($b_i$) shows the linear response of a genotype under various environments and the deviation from regression ($S^2_{di}$) illustrates the consistency of performance of that genotype. The average stability under all environments occurs when $b_i$ is approximately equal to 1.0 with $S^2_{di}$ of zero. Based on the analysis, genotypes with higher coefficients ($b_i > 1$) had higher sensitivities to environmental change and greater performance in optimal environments but worse adaptability to undesired environments. Genotypes with lower coefficients ($b_i < 1$) describe greater resistance to environmental change, which perform better in unfavorable environments but worse than average in desirable environments.

In this study, regression coefficients ranged from 0.30 to 1.27 for protein content (Table 3.4), which indicated that genotypes had different responses to environmental changes. However, for most soybean genotypes, the coefficient value had no significant difference from a value of 1.0. Therefore, these soybean genotypes were suggested to have average stability for protein concentration. G13 and G15 soybean genotypes had regression coefficients significantly higher than 1.0, which were more sensitive to environmental changes and suggested to be planed under desirable

environments only. G22 cultivar had a significantly lower regression coefficient ($b_i$ = 0.30) and specific adaptability to unfavorable environments. G3, G8 and G20 cultivars had regression coefficients not significantly different from 1.0 with higher protein content than other soybean genotypes, which might be considered superior for high protein soybean cultivation and widely adapted indifferent environments.

*3.4.3 Environment effects on soybean protein and amino acid contents*

Growing environment resulted in significant differences in protein and amino acid contents in soybean grains ($P < 0.01$). Based on Table 3.5, soybean planted in different locations exhibited no significant difference for protein content ($P > 0.05$) in 2018. For Arborg, Morris and Ste Adolphe, soybean had a significantly higher protein content in 2019 ($P < 0.05$). Morris and Ste Adolphe had higher precipitation in 2019 than 2018 (Table 3.1), suggesting that higher water availability might lead to higher protein content in soybean, which was in agreement with earlier reports (Boydak et al., 2002; Carrera et al., 2009). Moreover, the mean temperature during May to Septemberin 2019 was lower than in 2018 for all four locations. Previous studies found that soybean protein content decreased with an increase of mean temperature particularly at low levels (less than 25 ℃) (Piper & Boote, 1999; Mourtzinis et al., 2017), which could also explain the higher protein content found in soybean from Arborg, Morris and Ste Adolphe in 2019. However, although Carman had a lower mean temperature and higher precipitation in 2019 (16.0 ℃ and 355.0 mm) than in 2018 (17.2 ℃ and 253.0 mm), soybean from Carman exhibited no significant difference for protein content in these two years ($P > 0.05$). The clear effects of

different environmental factors were not analyzed due to the environment data obtained in this study were not exactly during the cropping season. The future work might focus on revealing the specific effects of temperature, precipitation, solar radiation, etc.

The Pearson correlation coefficients in Table 3.6 described the connection between each soybean biochemical parameter derived across all environments. The high correlation coefficients indicated that the changes observed in soybean phenotypic traits had a similar tendency across the different environments. Table 3.6 shows that Met and Val had high coefficients (0.83 and 0.97) and significant correlations with protein ($P < 0.05$), which indicated that the environments that were favorable for soybean protein accumulation might also be desirable for producing Met and Val. Nevertheless, other essential amino acids and Cys had non-significant correlations with protein. These data may suggest that the effects of environments for soybean protein content were different from that for Cys, His, Ile, Leu, Lys, Phe, Thr and Trp content. Moreover, the correlations between various amino acids were different. Therefore, each amino acid appears to respond differently to environmental alterations, which was in agreement with an earlier report (Carrera et al., 2011).

### *3.4.4 Critical amino acid value and protein*

The Critical Amino Acid Value (CAAV) is the sum of four essential amino acids (Lys, Thr, Trp and Met) and Cys as a percent of crude protein. It is a protein quality evaluator, allowing buyers to assess the natural balance and gross limiting essential amino acid levels in the soybean, with particular importance for livestock feeding

purposes. In some cases, such as for the starter rations of swine and broiler, CAAV might be more consistent than crude protein as the selection tool for feed ingredients.

CAAV was found to be negatively linearly correlated with protein content (Figure 1). The result indicates that, in soybean with high protein content, less of the protein consists of critical amino acids, suggesting that the increased protein in soybean grain after a certain stage is mainly composed of other amino acids. However, the coefficient of determination ($R^2$) of the model in Figure 1 was 0.53, which was lower than the $R^2$ of models made for the individual environment (Table 3.7). This suggested that the relationship between CAAV and protein content in different soybean genotypes were similar in the same environment, and it varied between different environments.

Based on Table 3.7, soybean planted in Car2018 and Mor2018 had the highest mean CAAV (14.58% and 14.71%), which could be explained by high critical amino acid content and relatively low protein concentration in these soybean (Table 3.5). Soybean from Abg2019 had the lowest mean CAAV (12.83%) because of the highest protein content but almost as high in critical amino acids level as the soybean from Car2018 and Mor2018. This phenomenon also confirms the findings from Figure 1. However, soybean from Car2019 had comparatively low protein content (mean value of 40.32%) but low CAAV (13.42%), and the same CAAV was found in soybean from Mor2019 with higher protein content (mean value of 43.98%), which indicated that environments altered the relationship between CAAV and protein content in soybean grains. The reason for the latter observation might be due to the fact that the impacts of environment on soybean protein and specific amino acid contents were different, as

previously discussed.

The slope value from Table 3.7 shows the variance of CAAV within different soybean genotypes under each environment. A higher absolute value of slope indicated a more considerable difference of CAAV between soybean with various protein content. Soybean from Abg2019 had the minimum absolute value of slope (0.055), which demonstrated slight variation in CAAV between these soybean resulting in a small range (12.46-13.18%). Although soybean from Car2018 and Mor2018 had the highest mean CAAV, these two groups also had the highest absolute slope value (0.164 and 0.146) and a large range of CAAV (13.40-15.28% and 13.90-15.30%).

## 3.5 Conclusion

Genotype × environment interactions had a significant impact on soybean protein and amino acid concentrations. Genotype and environments dominated the largest variation of protein and amino acid contents in soybean based on the variance component values. Most soybean genotypes in this study performed with average stability in relation to protein accumulation across various environments. Soybean genotypes with high protein content and average stability might be considered superior for high protein soybean cultivation purpose. Protein and amino acids were found to respond differently to various environments, but the favorable and undesirable environments for soybean protein and amino acid accumulation were not clear in this study. CAAV was negatively correlated with protein content, which suggested that the protein produced in soybean grains after a certain stage was mainly

composed of other amino acids. Environments affected the relation of CAAV to protein content and the variation of CAAV within soybean genotypes. This work would help Manitoba soybean farmers to select optimal soybean varieties for various purposes and prove the feasibility of marketing low protein soybean based on their essential amino acid contents.

Table 3.1. Temperature, rainfall, and solar radiation of the eight soybean growing environments in Manitoba, Canada during May to September

| Location | Coordinate | Year | Temperature (℃) | | | Rainfall (mm) | Solar radiation (mega joules) |
|---|---|---|---|---|---|---|---|
| | | | Min | Mean | Max | | |
| Arborg | 50°54′27.1″N 97°13′5.4″W | 2018 | -7.1 | 16.2 | 36.5 | 249.3 | 2920.5 |
| | | 2019 | -5.5 | 15.2 | 35.9 | 263.0 | 3059.8 |
| Carman | 49°29′57″N 98°0′3″W | 2018 | -4.9 | 17.2 | 39.4 | 253.0 | 3041.1 |
| | | 2019 | -5.8 | 16.0 | 36.9 | 355.0 | 3007.8 |
| Morris | 49°21'10.79"N 97°21'32.39"W | 2018 | -3.9 | 17.4 | 37.5 | 244.0 | 2821.7 |
| | | 2019 | -5.8 | 16.0 | 36.8 | 430.0 | 2801.1 |
| Ste Adolphe | 49°40′31″N 97°06′35″W | 2018 | -4.0 | 17.2 | 35.7 | 275.0 | 2971.5 |
| | | 2019 | -4.5 | 16.0 | 37.5 | 406.0 | 3022.2 |

Table 3.2. Mean squares from analysis of variance of protein and amino acids for soybean genotypes grown in four locations in Manitoba, Canada during 2018 and 2019

| Source | df | Protein | Lys | Thr | Met | Cys | Trp | Phe | Val | Leu | Ile | His |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Year (Y) | 1 | 885.02** | 0.026** | 0.547** | 0.009** | 0.494** | 0.046** | 0.662** | 1.031** | 1.302** | 0.030** | 0.402** |
| Location (L) | 3 | 117.27** | 0.160** | 0.069** | 0.004** | 0.007** | 0.009** | 0.262** | 0.157** | 0.406** | 0.156** | 0.013** |
| L×Y | 3 | 175.97** | 0.757** | 0.202** | 0.022** | 0.026** | 0.017** | 0.554** | 0.322** | 1.538** | 0.269** | 0.101** |
| Replicate (L Y ) | 16 | 6.55** | 0.009** | 0.003** | 0.000* | 0.000 | 0.000 | 0.009** | 0.007** | 0.018** | 0.005** | 0.003** |
| Genotype (G) | 22 | 145.55** | 0.267** | 0.085** | 0.015** | 0.016** | 0.010** | 0.236** | 0.190** | 0.448** | 0.159** | 0.056** |
| G×Y | 22 | 3.46* | 0.005 | 0.002 | 0.000* | 0.001** | 0.000** | 0.006 | 0.005** | 0.010 | 0.003 | 0.001 |
| G×L | 66 | 2.41 | 0.006** | 0.002** | 0.000 | 0.000** | 0.000** | 0.006** | 0.004* | 0.011** | 0.004** | 0.001** |
| G×L×Y | 66 | 2.85* | 0.006** | 0.002** | 0.000* | 0.000** | 0.000** | 0.008** | 0.004** | 0.017** | 0.004** | 0.001** |
| Residual | 330 | 1.85 | 0.003 | 0.001 | 0.000 | 0.000 | 0.000 | 0.004 | 0.002 | 0.007 | 0.002 | 0.001 |

df: degree of freedom; Lys: lysine; Thr: threonine; Met: methionine; Cys: cysteine; Trp: tryptophan; Phe: phenylalanine; Val: valine; Leu: leucine: Ile: isoleucine; His: histidine

*Significant at the 0.05 probability level.

**Significant at the 0.01 probability level.

Table 3.3. Estimates of variance components for protein and amino acids of soybean genotypes in four locations in Manitoba, Canada during 2018 and 2019

| Environment | Protein | Lys | Thr | Met | Cys | Trp | Phe | Val | Leu | Ile | His |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Year (Y) | 3.190 | 0.0000 | 0.0018 | 0.0000 | 0.0019 | 0.0001 | 0.0012 | 0.0035 | 0.0017 | 0.0000 | 0.0015 |
| Location (L) | 0.000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| L×Y | 2.045 | 0.0058 | 0.0020 | 0.0002 | 0.0003 | 0.0002 | 0.0059 | 0.0034 | 0.0141 | 0.0027 | 0.0008 |
| Genotype (G) | 6.185 | 0.0117 | 0.0037 | 0.0006 | 0.0007 | 0.0004 | 0.0102 | 0.0082 | 0.0193 | 0.0069 | 0.0024 |
| G×Y | 0.064 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0000 | 0.0000 | 0.0000 |
| G×L | 0.000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| G×L×Y | 0.272 | 0.0008 | 0.0003 | 0.0000 | 0.0001 | 0.0000 | 0.0011 | 0.0005 | 0.0020 | 0.0005 | 0.0002 |
| Residual | 1.847 | 0.0033 | 0.0012 | 0.0002 | 0.0002 | 0.0001 | 0.0038 | 0.0024 | 0.0072 | 0.0020 | 0.0008 |

Lys: lysine; Thr: threonine; Met: methionine; Cys: cysteine; Trp: tryptophan; Phe: phenylalanine; Val: valine; Leu: leucine: Ile: isoleucine; His: histidine

Table 3.4. Mean crude protein content, regression coefficient, standard error of coefficient and deviation from regression for the 23 soybean genotypes tested across eight environments in Manitoba, Canada

| No. | Variety Group | Mean | $b_i$ | SE | $S^2_{di}$ |
|---|---|---|---|---|---|
| 1 | CN | 41.67 | 0.84 | 0.15 | 0.78 |
| 2 | GT | 39.96 | 0.75 | 0.20 | 1.07 |
| 3 | CN | 47.62 | 0.79 | 0.19 | 1.01 |
| 4 | GT | 39.94 | 1.27 | 0.15 | 0.81 |
| 5 | GT | 39.19 | 1.04 | 0.15 | 0.78 |
| 6 | GT | 40.19 | 1.24 | 0.17 | 0.89 |
| 7 | GT | 40.13 | 1.05 | 0.13 | 0.69 |
| 8 | CN | 47.52 | 0.80 | 0.10 | 0.51 |
| 9 | GT | 41.45 | 1.22 | 0.16 | 0.82 |
| 10 | GT | 40.47 | 0.99 | 0.12 | 0.62 |
| 11 | CN | 42.12 | 1.26 | 0.18 | 0.93 |
| 12 | CN | 40.29 | 0.84 | 0.14 | 0.73 |
| 13 | GT | 40.44 | 1.25* | 0.10 | 0.54 |
| 14 | GT | 39.78 | 0.81 | 0.29 | 1.55 |
| 15 | GT | 40.22 | 1.23* | 0.08 | 0.41 |
| 16 | GT | 39.79 | 1.25 | 0.16 | 0.83 |
| 17 | GT | 41.17 | 0.96 | 0.21 | 1.12 |
| 18 | GT | 40.77 | 0.95 | 0.15 | 0.79 |
| 19 | GT | 40.19 | 0.88 | 0.15 | 0.79 |
| 20 | CN | 47.16 | 0.89 | 0.15 | 0.82 |
| 21 | CN | 43.47 | 1.13 | 0.29 | 1.54 |
| 22 | CN | 41.87 | 0.30** | 0.10 | 0.55 |
| 23 | GT | 43.03 | 1.25 | 0.24 | 1.27 |

$b_i$: regression coefficient; SE: standard error; $S^2_{di}$: deviation from regression; CN: conventional soybean variety; GT: glyphosate-tolerant soybean variety

*Significant at the 0.05 probability level.

**Significant at the 0.01 probability level.

Table 3.5. Mean crude protein and amino acid contents of soybean grown in distinct environments in Manitoba, Canada.

| Environment[1] | Protein | Lys | Thr | Met | Cys | Trp | Phe | Val | Leu | Ile | His |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Abg 2018 | 39.78$^{c}$ | 2.42$^{c}$ | 1.51$^{b}$ | 0.56$^{c}$ | 0.59$^{b}$ | 0.51$^{c}$ | 1.92$^{c}$ | 1.74$^{d}$ | 2.88$^{d}$ | 1.72$^{c}$ | 0.86$^{cd}$ |
| Abg 2019 | 45.03$^{a}$ | 2.56$^{a}$ | 1.54$^{b}$ | 0.60$^{a}$ | 0.56$^{c}$ | 0.52$^{b}$ | 1.98$^{ab}$ | 1.92$^{ab}$ | 2.99$^{bc}$ | 1.80$^{a}$ | 0.88$^{bc}$ |
| Car 2018 | 40.55$^{c}$ | 2.55$^{a}$ | 1.60$^{a}$ | 0.59$^{a}$ | 0.62$^{a}$ | 0.54$^{a}$ | 2.02$^{a}$ | 1.80$^{c}$ | 3.08$^{a}$ | 1.78$^{ab}$ | 0.90$^{ab}$ |
| Car 2019 | 40.32$^{c}$ | 2.35$^{d}$ | 1.44$^{c}$ | 0.57$^{bc}$ | 0.56$^{c}$ | 0.49$^{e}$ | 1.77$^{d}$ | 1.76$^{cd}$ | 2.70$^{e}$ | 1.64$^{d}$ | 0.80$^{e}$ |
| Mor 2018 | 40.14$^{c}$ | 2.56$^{a}$ | 1.61$^{a}$ | 0.57$^{b}$ | 0.63$^{a}$ | 0.52$^{b}$ | 2.03$^{a}$ | 1.78$^{cd}$ | 3.07$^{ab}$ | 1.79$^{ab}$ | 0.91$^{a}$ |
| Mor 2019 | 42.80$^{b}$ | 2.49$^{b}$ | 1.52$^{b}$ | 0.58$^{b}$ | 0.55$^{d}$ | 0.50$^{d}$ | 1.94$^{bc}$ | 1.87$^{b}$ | 2.93$^{cd}$ | 1.75$^{bc}$ | 0.84$^{d}$ |
| Sta 2018 | 40.75$^{c}$ | 2.49$^{b}$ | 1.59$^{a}$ | 0.58$^{b}$ | 0.63$^{a}$ | 0.54$^{a}$ | 2.01$^{a}$ | 1.79$^{c}$ | 3.01$^{ab}$ | 1.78$^{ab}$ | 0.91$^{ab}$ |
| Sta 2019 | 43.98$^{ab}$ | 2.56$^{a}$ | 1.54$^{b}$ | 0.60$^{a}$ | 0.55$^{d}$ | 0.52$^{b}$ | 1.98$^{ab}$ | 1.93$^{a}$ | 3.01$^{ab}$ | 1.80$^{a}$ | 0.83$^{d}$ |
| SEM | 0.32 | 0.012 | 0.007 | 0.002 | 0.002 | 0.002 | 0.012 | 0.010 | 0.017 | 0.009 | 0.006 |

Lys: lysine; Thr: threonine; Met: methionine; Cys: cysteine; Trp: tryptophan; Phe: phenylalanine; Val: valine; Leu: leucine: Ile: isoleucine; His: histidine; Abg: Arborg; Car: Carman; Mor: Morris; Sta: St. Adolphe; SEM: standard error of mean

[1]Groups with with same letter have no significant difference at 0.05 probability level

Table 3.6. Estimates of Pearson correlation coefficients characterizing the relationship between various soybean biochemical parameters across different environments in Manitoba, Canada

| | Cys | His | Ile | Leu | Lys | Met | Phe | Thr | Trp | Val |
|---|---|---|---|---|---|---|---|---|---|---|
| Protein | -0.66 | -0.20 | 0.48 | 0.19 | 0.50 | 0.83* | 0.17 | -0.07 | -0.02 | 0.97** |
| Cys | | 0.84** | 0.27 | 0.52 | 0.21 | -0.32 | 0.55 | 0.74* | 0.69 | -0.59 |
| His | | | 0.67 | 0.80* | 0.60 | 0.06 | 0.84 | 0.91** | 0.83* | -0.15 |
| Ile | | | | 0.94** | 0.96** | 0.63 | 0.94** | 0.82* | 0.74* | 0.58 |
| Leu | | | | | 0.93** | 0.47 | 0.99** | 0.95** | 0.83* | 0.33 |
| Lys | | | | | | 0.68 | 0.89** | 0.80* | 0.65 | 0.62 |
| Met | | | | | | | 0.40 | 0.25 | 0.41 | 0.88** |
| Phe | | | | | | | | 0.95** | 0.83* | 0.29 |
| Thr | | | | | | | | | 0.87** | 0.06 |
| Trp | | | | | | | | | | 0.08 |

Lys: lysine; Thr: threonine; Met: methionine; Cys: cysteine; Trp: tryptophan; Phe: phenylalanine; Val: valine; Leu: leucine: Ile: isoleucine; His: histidine

*Significant at the 0.05 probability level.

**Significant at the 0.01 probability level.

Table 3.7. Statistics for linear regression between Critical Amino Acid Value% and protein content of soybean grown in different environments in Manitoba, Canada

| Environment[1] | Protein Range | | Slope | $R^2$ | CAAV% | | |
|---|---|---|---|---|---|---|---|
| | Min protein% | Max protein% | | | Min CAAV% | Mean CAAV% | Max CAAV% |
| Abg 2018 | 36.09 | 46.91 | -0.086 | 0.80 | 13.47 | 14.09$^{c}$ | 14.54 |
| Abg 2019 | 42.36 | 50.52 | -0.055 | 0.57 | 12.46 | 12.83$^{e}$ | 13.18 |
| Car 2018 | 37.33 | 47.98 | -0.164 | 0.95 | 13.40 | 14.58$^{ab}$ | 15.28 |
| Car 2019 | 37.59 | 46.51 | -0.088 | 0.64 | 12.86 | 13.42$^{d}$ | 13.72 |
| Mor 2018 | 37.80 | 45.65 | -0.146 | 0.82 | 13.90 | 14.71$^{a}$ | 15.30 |
| Mor 2019 | 39.44 | 48.65 | -0.074 | 0.63 | 12.70 | 13.18$^{d}$ | 13.54 |
| Sta 2018 | 37.61 | 47.30 | -0.133 | 0.83 | 13.48 | 14.31$^{bc}$ | 14.91 |
| Sta 2019 | 38.93 | 50.15 | -0.103 | 0.87 | 12.51 | 13.13$^{d}$ | 13.96 |

CAAV: Critical Amino Acid Value; Abg: Arborg; Car: Carman; Mor: Morris; Sta: St. Adolphe

[1]Groups with with same letter have no significant difference at 0.05 probability level

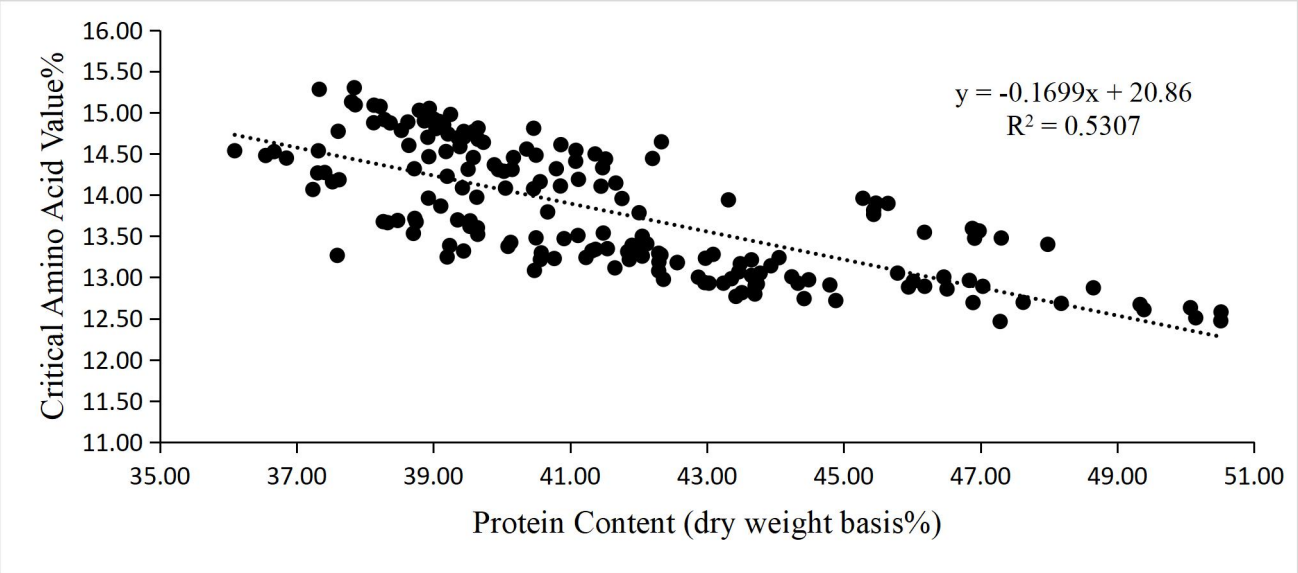


Figure 3.1. Linear regression of critical amino acid value% versus crude protein content for soybean samples

# Chapter 4 General Discussion

In this thesis, the main objectives were to i) establish NIR calibration models for measuring crude protein and amino acid contents in soybean and evaluate the predictive abilities of NIR calibration models; and ii) study the effects of genotype, environment and genotype × environment interaction on soybean protein and amino acid contents.

The results from this thesis indicated that NIR spectroscopy could be able to distinguish soybean with high and low contents of protein and most amino acids, which was based on the ratio of prediction to deviation or relative predictive determinant (RPD) and the determination coefficient of calibration values ($R^2_c$). According to Williams & Norris (2001), $R^2$ values between 0.66 and 0.81 indicated rough quantitative predictions, while a value for $R^2$ between 0.82 and 0.90 showed a good prediction. Excellent prediction models were thought when $R^2$ values above 0.91.

RPD is another NIR model evaluator, which is the ratio of standard deviation (SD) of the reference values in the validation set to the standard error of cross validation (SECV) or the standard error of prediction (SEP) (Williams & Norris, 2001). RPD represents the ratio of natural variation in the samples to the size of likely prediction errors (Fearn, 2002). Thus, the measurement with a low RPD might not discriminate between samples. Different but similar guidelines were used in previous studies to describe the performance of calibration models. Usually models with RPD values less than 1.0 to 1.5 are regarded as unusable model, while values over 1.4 or 1.5 indicate that models are able for screening purpose (Saeys et al., 2005; Kovalenko

et al., 2006; Malley et al., 2005). For RPD values over 2.4 or 2.5, models are classified as good or successful models, which usable with caution for most applications (Saeys et al., 2005; Kovalenko et al., 2006; Malley et al., 2005). The range error ratio (RER) is calculated using range of the reference values in the validation set to divided by SECV or SEP. It is an index similar to RPD. The relation between RPD and RER is based on the distribution of reference samples, and there is no clear conversion coefficient between these two indexes (Fearn, 2002).

The spectral pre-treatment in this thesis is the combination of de-trending (DT) and standard normal variate (SNV), which is a commonly used preprocessing approach. The aim of spectral pre-treatment is to minimizes irrelevant information (noise or background) due to overlapping absorption bands and light scattering effects (Blanco & Villarroya, 2002; Agelet & Hurburgh, 2010; Azzouz et al, 2003). However, pre-treatment methods should be used with caution since they sometimes also reduce the useful information from the spectra related to component concentrations (Azzouz et al, 2003). It is suggested to use different pre-treatment and their combinations in model calibrations and identify the better preprocessing methods based on the predictive ability of model. Other common pre-treatment approaches include derivatives, multiplicative scatter correction (MSC) (Chen et al., 2004).

Liu et al. (2019) reported the use of six common spectral pre-treatment in NIR model calibration. The results in their study showed NIR calibration for non-processing group had $R^2_{cv}$ of 0.72 and RPD value of 1.87. Groups had higher $R^2_{cv}$ and RPD values when pre-treated using first derivative (FD) and SNV had higher $R^2_{cv}$ and RPD values but lower values with pre-treatment by mean centering (MC), MSC

and logarithmic function (Liu et al., 2019). Chen et al. (2004) evaluated different pre-treatment methods in NIR calibration for poultry manure. In their results, NIR calibration models for total nitrogen without preprocessing had a determination coefficient on validation ($R^2_v$) value of 0.67 and RPD value of 1.72. Models with preprocessing higher $R^2_v$ and RPD values, while model pre-treated with direct orthogonal signal correction (DOSC) had the highest $R^2_v$ and RPD value (Chen et al., 2004). It is suggested to compare different pre-treatment methods and their combinations and figure out the optimal preprocessing approach for our lab.

In this thesis, two types of NIR spectroscopy were used. The wavelength range for each instrument was different. The wavelength range for Perkin Elmer DA 7250 is from 950 to 1650 nm, while Perkin Elmer FT 9700 has a wider range between 699 to 2593 nm. A broader range of wavelength contain more information of interest. However, the larger number of wavelengths also include noninformation wavelengths, which would prevent the acquisition of high-quality quantitative models (Chen et al., 2013). Therefore, it is suggested to eliminate the unhelpful wavelengths in calibration to improve the predictive ability and reduce the complexity of the model (Wang et al., 2020). Moreover, the calibration based on optimal combinations of sensitive wavelengths might be better than using whole NIR spectra (Qing et al., 2007). The sensitive wavelengths are usually selected based on the peak absorbance of the component of interest and the component whose content is highly correlated to the component of interest (Xiaobo et al., 2010). Several wavelength-selection methods based on PLS include sive projections algorithm (SPA) (Xiaobo et al., 2010), uninformative variable elimination (UVE) (Centner et al., 1996), competitive adaptive

reweighted sampling (CARS) (Li et al., 2009), genetic algorithm (GA) (Qing et al., 2007), etc.

Qing et al. (2007) reported the use of GA and correlation coefficient method in sensitive wavelength selection for fruit quality sensing. Based on their results, models obtained from combinations of sensitive wavelengths had similar correlation coefficient (R) and SECV values. Xu et al. (2012) compared four different variable selection methods: stepwise multiple linear regression (SMLR), genetic algorithm-partial least squares regression (GA-PLS), interval PLS (iPLS), and successive projection algorithm-multiple linear regression combined with GA (GA-SPA-MLR). They concluded GA-PLS and GA-SPA-MLR had a better performance than SMLR and iPLS on measuring sugar content in pear. They found GA-PLS had a relatively higher $R^2_c$ and lower RMSEC values than those for GA-SPA-MLR, but latter was developed with a smaller number of wavelengths, which was believed to fit industrial application due to its simple algorithm. In this thesis, FT 9700 had relatively lower $R^2_c$ and RPD values than those for DA 7250 on measuring crude protein and most amino acid contents, which might be partially due to the broader range of wavelengths used in FT 9700. It is highly recommended to test different wavelength selection methods and select optimal combinations of wavelengths for future calibration using FT 9700 NIR spectroscopy instrument.

Partial least square (PLS) regression was the calibration method used for developing models in this thesis. Among linear regression methods, PLS has advantages including faster algorithm, higher precision and harmonious calibration models (Kalivas & Gemperline, 2006). Non-linear regression methods, such as

artificial neural networks (ANN) (Zupan & Gasteiger, 1993), non-linear PLS (Blanco & Villarroya, 2002) and support vector machines (SVM) (Agelet & Hurburgh, 2010) can also be used for NIR calibration model development.

Kovalenko et al. (2006) used three regression methods: PLS, ANN and SVM to develop NIR calibration models for measuring crude protein and amino acid contents in soybean. They concluded that PLS and SVM had better prediction performance than ANN. Moreover, they found calibration models developed using PLS for amino acids had negative bias values, while positive bias was found in models developed using ANN and SVM. This indicated that NIR spectroscopy with linear calibration models might overpredict amino acid contents, while amino acid contents might be underestimated using spectrometer with non-linear calibrations (Kovalenko et al., 2006). Dr. David Honigs from Perkin Elmer company has developed a non-linear regression method named “Honigs regression”, which is now being tested to develop NIR calibration model. It has potential to use non-linear regressions to calibrate NIR spectroscopy in the future research.

Soybean protein content is considered as an important index for determining benefits for soybean farmers. Grain yield is another agronomic trait closely related to the benefits of soybean. It is also important to search a genotype having high yield over a wide array of environments since the superior cultivars should have desired performance under ideal environments, and also produce acceptable yields in unfavorable growing conditions (Gurmu et al., 2009). Several previous studies had indicated that there was a significant genotype × environment interaction effect on soybean grain yields (Gurmu et al., 2009; Yan et al., 2010; Whaley & Eskandari,

2019). Grain yield was reported to negatively correlate with protein concentration in soybean (Shannon et al., 1972; Yin & Vyn, 2005; Kim et al., 2016), showing the difficulty of high yielding high-protein soybean cultivation. However, some studies found that the negative relationship could be moderated in specific genotypes and environments (Wilcox & Cavins, 1995; Panthee & Pantalone, 2006; Mian et al., 2017). Therefore, genotype × environment interaction on grain yield and relationship between yield and protein content for Manitoba soybean could be conducted in the future to help select soybean genotypes having a high protein content with appreciable production across various environments.

# Chapter 5 Future Directions

Future research efforts should focus on:

1. Comparing the prediction performance of NIR calibration models with different spectral treatment methods and selecting the most suitable preprocessing approach for future calibration models development.
2. Selecting the sensitive wavelengths and evaluating the predictive ability of models made using the selected wavelengths. Also, comparing the performance of models made using the entire wavelengths and new models.
3. Testing the feasibility of using other regression methods (linear or non-linear) for developing NIR calibration models and evaluating the performance of the models.
4. Recording the temperature data during the specific cropping season. Using multiple linear regression model to identify how different environmental factors affect soybean protein and amino acid contents.
5. Exploring the genotype × environment interaction effect on soybean grain yield and the relationship between soybean protein content and grain yield. Also, identifying the soybean genotypes with stable high protein content and acceptable yield.

**Reference**

AACC. (1999). *Approved Methods of the American Association of Cereal Chemists*. AACC International: St. Paul.

Adlercreutz, H., Fotsis, T., Bannwart, C., Wähälä, K., Mäkelä, T., Brunow, G., & Hase, T. (1986). Determination of urinary lignans and phytoestrogen metabolites, potential antiestrogens and anticarcinogens, in urine of women on various habitual diets. *Journal of steroid biochemistry, 25*(5), 791-797.

Agelet, L. E., & Hurburgh Jr, C. R. (2010). A tutorial on near infrared spectroscopy and its calibration. *Critical Reviews in Analytical Chemistry, 40*(4), 246-260.

Ahrens, S., Venkatachalam, M., Mistry, A. M., Lapsley, K., & Sathe, S. K. (2005). Almond (Prunus dulcis L.) protein quality. *Plant Foods for Human Nutrition*, 60(3), 123-128.

Akiyama, T., Ishida, J., Nakagawa, S., Ogawara, H., Watanabe, S.-i., Itoh, N., . . . Fukami, Y. (1987). Genistein, a specific inhibitor of tyrosine-specific protein kinases. *Journal of Biological Chemistry, 262*(12), 5592-5595.

American Soybean Association. (2019). 2019 Soystats. Retrieved from https://soygrowers.com/wp-content/uploads/2019/10/Soy-Stats-2019_FNL-Web.pdf

Anderson, J. W., Johnstone, B. M., & Cook-Newell, M. E. (1995). Meta-analysis of the effects of soy protein intake on serum lipids. *New England Journal of Medicine*, 333(5), 276-282.

Assefa, Y., Bajjalieh, N., Archontoulis, S., Casteel, S., Davidson, D., Kovács, P., ... & Ciampitti, I. A. (2018). Spatial characterization of soybean yield and quality (amino acids, oil, and protein) for United States. *Scientific reports, 8*(1), 1-11.

Azzouz, T., Puigdoménech, A., Aragay, M., & Tauler, R. (2003). Comparison between different data pre-treatment methods in the analysis of forage samples using near-infrared diffuse reflectance spectroscopy and partial least-squares multivariate calibration method. *Analytica Chimica Acta, 484*(1), 121-134.

Bakalli, R. I., Pesti, G. M., & Etheridge, R. D. (2000). Comparison of a commercial near-infrared reflectance spectroscope and standard chemical assay procedures for analyzing feed ingredients: Influence of grinding methods. *Journal of Applied Poultry Research, 9*(2), 204-213.

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied spectroscopy, 43*(5), 772-777.

Benito, O., Galvez, A. F., Revilleza, M. J., & Krenz, D. C. (1999). Molecular strategies to improve the nutritional quality of legume proteins. In *Chemicals via higher plant bioengineering* (pp. 117-126). Springer, Boston, MA.

Berntsson, O., Danielsson, L. G., & Folestad, S. (1998). Estimation of effective sample size when analysing powders with diffuse reflectance near-infrared spectrometry. *Analytica chimica acta, 364*(1-3), 243-251.

Blackburn, S. (1968). Amino acid determination. Methods and techniques. *Amino acid determination. Methods and techniques.*

Blanco, M., & Villarroya, I. N. I. R. (2002). NIR spectroscopy: a rapid-response analytical tool. *TrAC Trends in Analytical Chemistry, 21*(4), 240-250.

Block, K. P. (1989). Interactions among leucine, isoleucine, and valine with special reference to the branched-chain amino acid antagonism. *Absorption and Utilization of Amino acids, 1,* 229-244.

Boydak, E., Alpaslan, M., Hayta, M., GERCek, S. I. N. A. N., & Simsek, M. (2002). Seed composition of soybeans grown in the Harran region of Turkey as affected by row spacing and irrigation. *Journal of agricultural and food chemistry, 50*(16), 4718-4720.

Boye, J., Wijesinha-Bettoni, R., & Burlingame, B. (2012). Protein quality evaluation twenty years after the introduction of the protein digestibility corrected amino acid score method. *British Journal of Nutrition*, 108(S2), S183-S211.

Briggs, D. R., & Wolf, W. J. (1957). Studies on the cold-insoluble fraction of the water-extractable soybean proteins. I. Polymerization of the 11 S component through reactions of sulfhydryl groups to form disulfide bonds. *Archives of biochemistry and biophysics*, 72(1), 127-144.

Cahoon, E. B. (2003). Genetic enhancement of soybean oil for industrial uses: prospects and challenges.

Cai, W., Li, Y., & Shao, X. (2008). A variable selection method based on uninformative variable elimination for multivariate calibration of

near-infrared spectra. *Chemometrics and intelligent laboratory systems, 90*(2), 188-194.

Carrera, C., Martínez, M. J., Dardanelli, J., & Balzarini, M. (2009). Water deficit effect on the relationship between temperature during the seed fill period and soybean seed oil and protein concentrations. *Crop Science, 49*(3), 990-998.

Carrera, C. S., Reynoso, C. M., Funes, G. J., Martínez, M. J., Dardanelli, J., & Resnik, S. L. (2011). Amino acid composition of soybean seeds as affected by climatic variables. *Pesquisa Agropecuaria Brasileira*, 46(12), 1579-1587.

Canadian Food Inspection Agency. (2016). Food labeling for industry: elements within the nutrition facts table—protein. Retrieved from http://www.inspection.gc.ca/food/labelling/food-labelling-for-industry/nutrition-labelling/elements-within-the-nutrition-facts-table/eng/1389206763218/1389206811747?chap¼7.

Canadian Food Inspection Agency. (2017). Causes of food poisoning. Retrieved from http://inspection.gc.ca/food/information-for-consumers/fact-sheets-and-infographics/food-poisoning/eng/1331151916451/1331152055552.

Cen, H., & He, Y. (2007). Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends in Food Science & Technology, 18*(2), 72-83.

Centner, V., Massart, D. L., de Noord, O. E., de Jong, S., Vandeginste, B. M., & Sterna, C. (1996). Elimination of uninformative variables for multivariate calibration. *Analytical chemistry, 68*(21), 3851-3858.

Chen, L. J., Xing, L., & Han, L. J. (2010). Influence of data preprocessing on the quantitative determination of nutrient content in poultry manure by near infrared spectroscopy. *Journal of environmental quality, 39*(5), 1841-1847.

Chen, M., Khare, S., Huang, B., Zhang, H., Lau, E., & Feng, E. (2013). Recursive wavelength-selection strategy to update near-infrared spectroscopy model with an industrial application. *Industrial & Engineering Chemistry Research, 52*(23), 7886-7895.

Chen, Y., & Wang, Z. (2019). Wavelength selection for NIR spectroscopy based on the binary dragonfly algorithm. *Molecules, 24*(3), 421.

Cooper, C., Packer, N., & Williams, K. (Eds.). (2001). *Amino acid analysis protocols* (Vol. 159). Springer Science & Business Media.

Dalais, F. S., Meliala, A., Wattanapenpaiboon, N., Frydenberg, M., Suter, D. A., Thomson, W. K., & Wahlqvist, M. L. (2004). Effects of a diet rich in phytoestrogens on prostate-specific antigen and sex hormones in men diagnosed with prostate cancer. *Urology, 64*(3), 510-515.

Damodaran, S., & Kinsella, J. E. (1981). Interaction of carbonyls with soy protein: Conformational effects. *Journal of Agricultural and Food Chemistry*, 29(6), 1253-1257.

De Mejía, E. G., & Prisecaru, V. I. (2005). Lectins as bioactive plant proteins: a potential in cancer treatment. *Critical reviews in food science and nutrition*, 45(6), 425-445.

Dei, H. (2011). Soybean as a feed ingredient for livestock and poultry *Recent trends*

*for e nhancing the diversity and quality of soybean products*: InTech.

Delhaye, S., & Landry, J. (1986). High-performance liquid chromatography and ultraviolet spectrophotometry for quantitation of tryptophan in barytic hydrolysates. *Analytical biochemistry, 159*(1), 175-178.

Delhaye, S., & Landry, J. (1992). Determination of tryptophan in pure proteins and plant material by three methods. *Analyst, 117*(12), 1875-1877.

D'Mello, J. F. (2003). *Amino acids in animal nutrition* (No. Ed. 2). CABI publishing.

Doris C. (2018). Mimicking meat, seafood, and dairy. *Food Technology Magazine, 72*(5).

Dornbos, D. L., & Mullen, R. E. (1992). Soybean seed protein and oil contents and fatty acid composition adjustments by drought and temperature. *Journal of the American Oil Chemists Society, 69*(3), 228-231.

Dryden, G. (2003). Near infrared spectroscopy: Applications in deer nutrition. *RIRDC Pub*, (W03/007).

Elango, R., Ball, R. O., & Pencharz, P. B. (2008). Indicator amino acid oxidation: concept and application. *The Journal of nutrition, 138*(2), 243-246.

FAO/WHO. (1991). P*rotein Quality Evaluation: Report of the Joint FAO/WHO Expert Consultation, Bethesda, Md., USA 4-8 December 1989* (Vol. 51). Food & Agriculture Org..

FAO/WHO/UNU. (2007). *Protein and amino acid requirement in human nutrition: report of a joint FAO/WHO/UNU Expert Consultation* (Vol. 935). World Health Organization..

FAO stat. (2019). Production quantities of soybeans by country. Retrieved from

http://www.fao.org/faostat/en/#data/QC/visualize

Falco, S. C., Guida, T., Locke, M., Mauvais, J., Sanders, C., Ward, R. T., & Webber, P. (1995). Transgenic canola and soybean seeds with increased lysine. *Bio/technology, 13*(6), 577-582.

Fearn, T. (2002). Assessing calibrations: sep, rpd, rer and r2. *NIR news, 13*(6), 12-13.

Fearn, T. (2005). Chemometrics: an enabling tool for NIR. *NIR news, 16*(7), 17-19.

Fehr, W. R., Hoeck, J. A., Johnson, S. L., Murphy, P. A., Nott, J. D., Padilla, G. I., & Welke, G. A. (2003). Genotype and environment influence on protein components of soybean. *Crop Science, 43*(2), 511-514.

Finley, J. W. (1985). Reducing variability in amino acid analysis. *Digestibility and amino acid availability in cereals and oilseeds*, 15-30.

Fontaine, J., Hörr, J., & Schirmer, B. (2001). Near-infrared reflectance spectroscopy enables the fast and accurate prediction of the essential amino acid contents in soy, rapeseed meal, sunflower meal, peas, fishmeal, meat meal products, and poultry meal. *Journal of Agricultural and Food Chemistry, 49*(1), 57-66.

Fountoulakis, M., & Lahm, H. W. (1998). Hydrolysis and amino acid composition analysis of proteins. *Journal of chromatography A, 826*(2), 109-134.

Friedman, M. (1996). Nutritional value of proteins from different food sources. A review. *Journal of Agricultural and Food Chemistry, 44*(1), 6-29.

Friedman, M., & Brandon, D. L. (2001). Nutritional and health benefits of soy proteins. *Journal of Agricultural and Food Chemistry, 49*(3), 1069-1086.

García-Sánchez, F., Galvez-Sola, L., Martínez-Nicolás, J. J., Muelas-Domingo, R., & Nieves, M. (2017). Using near-infrared spectroscopy in agricultural

systems. *Developments in near-infrared spectroscopy, 1*, 97-127.

Gibbs, B. F., Zougman, A., Masse, R., & Mulligan, C. (2004). Production and characterization of bioactive peptides from soy hydrolysate and soy-fermented food. *Food research international, 37*(2), 123-131

Geladi, P., MacDougall, D., & Martens, H. (1985). Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied spectroscopy, 39*(3), 491-500.

González-Martín, I., Álvarez-García, N., & González-Cabrera, J. (2006). Near-infrared spectroscopy (NIRS) with a fibre-optic probe for the prediction of the amino acid composition in animal feeds. *Talanta, 69*(3), 706–710.

Government of Canada. (1981). Method FO-1: Determination of Protein Rating. Ottawa, Canada: Health Protection Branch.

Government of Canada. (2016). Food and Drug Regulations: Nutrient Content Claims, B.01.500. Canada: Ottawa.

Government of Canada. (2019). Canada's food guide. Retrieved from https://food-guide.canada.ca/en/.

Griffiths, P. R. (1995). Practical consequences of math pre-treatment of near infrared reflectance data: log (1/R) vs F (R). *Journal of Near Infrared Spectroscopy, 3*(1), 60-62.

Gumbmann, M. R., Friedman, M., & Smith, G. A. (1983). The nutritional values and digestibilities of heat damaged casein and casein-carbohydrate mixtures. *Nutrition reports international, 28*(2), 355-358.

Gurmu, F., Mohammed, H., & Alemaw, G. (2009). Genotype x environment

interactions and stability of soybean for grain yield and nutrition quality. *African Crop Science Journal, 17*(2).

Hamilton-Reeves, J. M., Rebello, S. A., Thomas, W., Slaton, J. W., & Kurzer, M. S. (2007). Isoflavone-rich soy protein isolate suppresses androgen receptor expression without altering estrogen receptor-β expression or serum hormonal profiles in men at high risk of prostate cancer. *The Journal of nutrition*, 137(7), 1769-1775.

Hodges, R. E., Krehl, W. A., Stone, D. B., & Lopez, A. (1967). Dietary carbohydrates and low chole-sterol diets: effects on serum lipids of man. *American Journal of Clinical Nutrition, 20*, 198-208.

Hourant, P., Baeten, V., Morales, M. T., Meurens, M., & Aparicio, R. (2000). Oil and fat classification by selected bands of near-infrared spectroscopy. *Applied spectroscopy, 54*(8), 1168-1174.

Hughes, G. J., Ryan, D. J., Mukherjea, R., & Schasteen, C. S. (2011). Protein digestibility-corrected amino acid scores (PDCAAS) for soy protein isolates and concentrate: Criteria for evaluation. *Journal of agricultural and food chemistry*, 59(23), 12707-12712.

ISO. 2005. Animal feeding stuffs–Determination of tryptophan content. ISO 13904: 2005. 1st ed, Geneva.

Jenkins, D. J., Kendall, C. W., D'Costa, M. A., Jackson, C. J., Vidgen, E., Singer, W., ... & Fleshner, N. (2003). Soy consumption and phytoestrogens: effect on serum prostate specific antigen when blood lipids and oxidized low-density lipoprotein are reduced in hyperlipidemic men. *The Journal of urology, 169*(2), 507-511.

Jiang, G. L. (2020). Comparison and Application of Non-Destructive NIR Evaluations of Seed Protein and Oil Content in Soybean Breeding. *Agronomy*, 10(1), 77.

Jones, D. B. (1931). *Factors for converting percentages of nitrogen in foods and feeds into percentages of proteins (No. 183)*. US Department of Agriculture.

Kalivas, J. H., & Gemperline, P. J. (2006). Calibration. *Practical Guide to Chemometrics*, 105-166. CRC press.

Karr-Lilienthal, L. K., Grieshop, C. M., Merchen, N. R., Mahan, D. C., & Fahey, G. C. (2004). Chemical composition and protein quality comparisons of soybeans and soybean meals from five leading soybean-producing countries. *Journal of Agricultural and Food Chemistry, 52*(20), 6193-6199.

Karr-Lilienthal, L. K., Grieshop, C. M., Spears, J. K., & Fahey, G. C. (2005). Amino acid, carbohydrate, and fat composition of soybean meals prepared at 55 commercial US soybean processing plants. *Journal of agricultural and food chemistry, 53*(6), 2146-2150.

Kho, C. J., & Benito, O. (1988). Identification and isolation of methionine-cysteine rich proteins in soybean seed. *Plant Foods for Human Nutrition, 38*(4), 287-296.

Kim, M., Schultz, S., Nelson, R. L., & Diers, B. W. (2016). Identification and fine mapping of a soybean seed protein QTL from PI 407788A on chromosome 15. *Crop Science, 56*(1), 219-225.

Kingman, S. M., Walker, A. F., Low, A. G., Sambrook, I. E., Owen, R. W., & Cole, T.

J. (1993). Comparative effects of four legume species on plasma lipids and faecal steroid excretion in hypercholesterolaemic pigs. *British Journal of Nutrition, 69*(2), 409-421.

Kjeldahl, J. G. C. T. (1883). Neue methode zur bestimmung des stickstoffs in organischen körpern. *Zeitschrift für analytische Chemie, 22*(1), 366-382.

Kovalenko, I. V., Rippke, G. R., & Hurburgh, C. R. (2006). Determination of amino acid composition of soybeans (Glycine max) by near-infrared spectroscopy. *Journal of agricultural and food chemistry, 54*(10), 3485-3491.

Krul, E. S. (2019). Calculation of nitrogen‐to‐protein conversion factors: A review with a focus on soy protein. *Journal of the American Oil Chemists' Society, 96*(4), 339-364.

Kumar, N. B., Cantor, A., Allen, K., Riccardi, D., Besterman-Dahan, K., Seigne, J., ... & Pow-Sang, J. (2004). The specific role of isoflavones in reducing prostate cancer risk. *The Prostate*, 59(2), 141-147.

Kumar, V., Rani, A., Solanki, S., & Hussain, S. (2006). Influence of growing environment on the biochemical composition and physical characteristics of soybean seed. *Journal of Food Composition and Analysis, 19*(2-3), 188-195.

Leser, S. (2013). The 2013 FAO report on dietary protein quality evaluation in human nutrition: Recommendations and implications. *Nutrition Bulletin*, 38(4), 421-428.

Li, H., Liang, Y., Xu, Q., & Cao, D. (2009). Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Analytica chimica acta, 648*(1), 77-84.

Li, G., Wang, R., Quampah, A. J., Rong, Z., Shi, C., & Wu, J. (2011). Calibration and prediction of amino acids in stevia leaf powder using near infrared reflectance spectroscopy. *Journal of agricultural and food chemistry, 59*(24), 13065-13071.

Lipp, E. D. (1992). Near-infrared spectroscopy of silicon-containing materials. *Applied Spectroscopy Reviews, 27*(4), 385-408.

Liu, K. (1997). *Soybeans, technology, and utilization*. New York: Chapman & Hall.

Liu, Y., Liu, Y., Chen, Y., Zhang, Y., Shi, T., Wang, J., ... & Fei, T. (2019). The influence of spectral pretreatment on the selection of representative calibration samples for soil organic matter estimation using Vis-NIR reflectance spectroscopy. *Remote Sensing, 11*(4), 450.

Lusas, E. W., & Riaz, M. N. (1995). Soy protein products: processing and use. *The Journal of nutrition, 125*(suppl_3), 573S-580S.

Malenčić, D., Popović, M., & Miladinović, J. (2007). Phenolic content and antioxidant properties of soybean (Glycine max (L.) Merr.) seeds. *Molecules, 12*(3), 576-581.

Malley, D. F., McClure, C., Martin, P. D., Buckley, K., & McCaughey, W. P. (2005). Compositional analysis of cattle manure during composting using a field‐portable near‐infrared spectrometer. *Communications in Soil Science and Plant Analysis, 36*(4-6), 455-475.

Mao, L. C., Lee, K. H., & Erbersdobler, H. F. (1993). Effects of heat treatment on lysine in soya protein. *Journal of the Science of Food and Agriculture, 62*(3), 307-309.

Mihaljev, Ž. A., Jakšić, S. M., Prica, N. B., Ćupić, Ž. N., & Živkov-Baloš, M. M. (2015). Comparison of the Kjeldahl method, Dumas method and NIR method for total nitrogen determination in meat and meat products. *gas, 2, 7*.

Mian, M. R., McHale, L., Li, Z., & Dorrance, A. E. (2017). Registration of 'Highpro1'soybean with high protein and high yield developed from a North× South cross. *Journal of Plant Registrations, 11*(1), 51-54.

Moore, J. C., DeVries, J. W., Lipp, M., Griffiths, J. C., & Abernethy, D. R. (2010). Total protein methods and their potential utility to reduce the risk of food protein adulteration. *Comprehensive Reviews in Food Science and Food Safety, 9*(4), 330-357.

Mourtzinis, S., Gaspar, A. P., Naeve, S. L., & Conley, S. P. (2017). Planting date, maturity, and temperature effects on soybean seed yield and composition. *Agronomy Journal, 109*(5), 2040-2049.

Murphy, P. A., & Resurreccion, A. P. (1984). Varietal and environmental differences in soybean glycinin and. beta.-conglycinin content. *Journal of Agricultural and Food Chemistry, 32*(4), 911-915.

Naim, M., Gestetner, B., Bondi, A., & Birk, Y. (1976). Antioxidative and antihemolytic activities of soybean isoflavones. *Journal of agricultural and food chemistry, 24*(6), 1174-1177.

Nielsen, S. (2010). *Food Analysis* (4th ed.). https://doi.org/10.1007/978-1-4419-1478-1

Official methods of analysis of AOAC International (CD-ROM). (n.d.). Gaithersburg, Md: AOAC International. Retrieved from http://www.eoma.aoac.org/

Öste, R. E. (1991). Digestibility of processed food protein. In *Nutritional and toxicological consequences of food processing* (pp. 371-388). Springer, Boston, MA.

Owusu-Apenten, R. (2002). *Food protein analysis: quantitative effects on processing* (Vol. 118). CRC press.

Paul G. (2007). Soy protein label claims: where regulatory and marketing meet. 5th Southeast Asia Soyfood Seminar & Trade Show: Science to Market－Opportunities in Asia, Bangkok, Thailand. March 6－8.

Panthee, D. R., & Pantalone, V. R. (2006). Registration of soybean germplasm lines TN03-350 and TN04-5321 with improved protein concentration and quality. *Crop science, 46*(5), 2328.

Piper, E. L., & Boote, K. I. (1999). Temperature and cultivar effects on soybean seed oil and protein concentrations. *Journal of the American Oil Chemists' Society, 76*(10), 1233-1241.

Qin, P., Song, W., Yang, X., Sun, S., Zhou, X., Yang, R., ... & Ren, G. (2014). Regional distribution of protein and oil compositions of soybean cultivars in China. *Crop Science, 54*(3), 1139-1146.

Qing, Z., Ji, B., & Zude, M. (2007). Wavelength selection for predicting physicochemical properties of apple fruit based on near‐infrared spectroscopy. *Journal of food quality, 30*(4), 511-526.

Reichl, J. R. (1989). Absorption and metabolism of amino acids studied in vitro, in vivo, and with computer simulations. *Absorption and Utilization of Amino Acids, 1,* 93-156.

Rubenthaler, G. L., & Bruinsma, B. L. (1978). Lysine Estimation in Cereals by Near-Infrared Reflectance 1. *Crop Science, 18*(6), 1039-1042.

Rutherfurd, S. M., & Gilani, G. S. (2009). Amino acid analysis. *Current protocols in protein science, 58*(1), 11-9.

Sader, A. P. O., Oliveira, S. G., & Berchielli, T. T. (2004). Application of Kjeldahl and Dumas combustion methods for nitrogen analysis. *Archives of Veterinary Science, 9*(2).

Saint-Denis, T., & Goupy, J. (2004). Optimization of a nitrogen analyser based on the Dumas method. *Analytica Chimica Acta, 515*(1), 191-198.

Saio, K., Kamiya, M., & Watanabe, T. (1969). Food processing characteristics of soybean 11s and 7s proteins: Part i. Effect of difference of protein components among soybean varieties on formation of tofu-gel. *Agricultural and Biological Chemistry*, 33(9), 1301-1308.

Saio, K., Terashima, M., & Watanabe, T. (1975). Food use of soybean 7S and 11S proteins heat denaturation of soybean proteins at high temperature. *Journal of Food Science*, 40(3), 537-538.

Sarwar, G., Christensen, D. A., Finlayson, A. J., Friedman, M., Hackler, L. R., Mackenzie, S. L., ... & Tkachuk, R. (1983). Inter-and intra-laboratory variation in amino acid analysis of food proteins. *Journal of Food Science, 48*(2), 526-531.

Sarwar, G. (1997). The protein digestibility–corrected amino acid score method overestimates quality of proteins containing antinutritional factors and of poorly digestible proteins supplemented with limiting amino acids in rats. *The Journal of nutrition*, 127(5), 758-764.

Shannon, J. G., Wilcox, J. R., & Probst, A. H. (1972). Estimated Gains from Selection for Protein and Yield in the F4 Generation of Six Soybean Populations. *Crop Science, 12*(6), 824-826.

Sosulski, F. W., & Holt, N. W. (1980). Amino acid composition and nitrogen-to-protein factors for grain legumes. *Canadian Journal of Plant Science, 60*(4), 1327-1331.

Specht L. (2018). Is the future of meat animal-free? *Food Technology Magazine, 72*(1).

Spindler, M., Stadler, R., & Tanner, H. (1984). Amino acid analysis of feedstuffs: determination of methionine and cystine after oxidation with performic acid and hydrolysis. *Journal of agricultural and food chemistry, 32*(6), 1366-1371.

Sriperm, N., Pesti, G. M., & Tillman, P. B. (2011). Evaluation of the fixed nitrogen‐to‐protein (N: P) conversion factor (6.25) versus ingredient specific N:

P conversion factors in feedstuffs. *Journal of the Science of Food and Agriculture, 91*(7), 1182-1186.

Stark, E., & Luchter, K. (2005). NIR instrumentation technology. *NIR news, 16*(7), 13-16.

Statistic Canada. (2017). Table 001-0017-Estimate areas, yield, production and average farm price of principal field crops, in metric units, annual.

Tahir, M., Lindeboom, N., Baga, M., Vandenberg, A., & Chibbar, R. (2011). Composition and correlation between major seed constituents in selected lentil (Lens culinaris. Medik) genotypes. *Canadian Journal of Plant Science*, *91*(5), 825–835.

Thompson, M., Owen, L., Wilkinson, K., Wood, R., & Damant, A. (2002). A comparison of the Kjeldahl and Dumas methods for the determination of protein in foods, using data from a proficiency testing scheme. *Analyst, 127*(12), 1666-1668.

Tkachuk, R. (1969). Nitrogen-to-protein conversion factors for cereals and oilseed meals. *Cereal Chemistry, 46*, 419-424.

Torres, N., Torre-Villalvazo, I., & Tovar, A. R. (2006). Regulation of lipid metabolism by soy protein and its implication in diseases mediated by lipid disorders. *The Journal of nutritional biochemistry, 17*(6), 365-373.

Vollmann, J., Fritz, C. N., Wagentristl, H., & Ruckenbauer, P. (2000). Environmental and genetic variation of soybean seed protein content under Central European growing conditions. *Journal of the Science of Food and Agriculture, 80*(9), 1300-1306.

Watford, M., & Wu, G. (2011). *Protein. Advances in Nutrition, 2*(1), 62-63.

Wang, L., Wang, Q., Liu, H., Liu, L., & Du, Y. (2013). Determining the contents of protein and amino acids in peanuts using near-infrared reflectance spectroscopy. *Journal of the Science of Food and Agriculture, 93*(1), 118-124.

Wang, K., Du, W., & Long, J. (2020). Near-infrared wavelength-selection method based on joint mutual information and weighted bootstrap sampling. *IEEE Transactions on Industrial Informatics, 16*(9), 5884-5894.

Wee, K. L., & Shu, S.-W. (1989). The nutritive value of boiled full-fat soybean in pelleted feed for Nile tilapia. *Aquaculture, 81*(3-4), 303-314.

Wee, C. D., Hashiguchi, M., Ishigaki, G., Muguerza, M., Oba, C., Abe, J., ... & Akashi, R. (2018). Evaluation of seed components of wild soybean (Glycine soja) collected in Japan using near-infrared reflectance spectroscopy. *Plant Genetic Resources*, 16(2), 94-102.

Weiss, M., Manneberg, M., Juranville, J. F., Lahm, H. W., & Fountoulakis, M. (1998). Effect of the hydrolysis method on the determination of the amino acid composition of proteins. *Journal of chromatography A, 795*(2), 263-275.

WHO. (2017). "Stop using antibiotics in healthy animals to prevent the spread of antibiotic resistance." Press release, Nov 7. World Health Organization, Geneva. who. int.

Williams, P. C., Preston, K. R., Norris, K. H., & Starkey, P. M. (1984). Determination of amino acids in wheat and barley by near-infrared reflectance spectroscopy. *Journal of Food Science, 49*(1), 17-20.

Williams, P., & Norris, K. (2001). *Near-infrared technology in the agricultural and food industries*, 2nd ed. American Association of Cereal Chemists, Inc..

Williams, P., Manley, M., & Antoniszyn, J. (2019). *Near infrared technology: getting the best out of light*. African Sun Media.

Wilcox, J. R. (1987). Soybeans: Improvement, production, and uses. 646-686.

Madison, Wisconsin: American Society of Agronomy.

Wilcox, J. R., & Cavins, J. F. (1995). Backcrossing high seed protein to a soybean cultivar. *Crop Science, 35*(4), 1036-1041.

Wolf, R. B., Cavins, J. F., Kleiman, R., & Black, L. T. (1982). Effect of temperature on soybean seed constituents: oil, protein, moisture, fatty acids, amino acids and sugars. J*ournal of the American Oil Chemists' Society, 59*(5), 230-232.

Wu, W., Williams, W. P., Kunkel, M. E., Acton, J. C., Huang, Y., Wardlaw, F. B., & Grimes, L. W. (1995). True protein digestibility and digestibility-corrected amino acid score of red kidney beans (Phaseolus vulgaris L.). *Journal of agricultural and food chemistry, 43*(5), 1295-1298.

Wu, J. G., Shi, C., & Zhang, X. (2002). Estimating the amino acid composition in milled rice by near-infrared reflectance spectroscopy. *Field Crops Research, 75*(1), 1-7.

Xiao, C. W. (2008). Health effects of soy protein and isoflavones in humans. *The Journal of nutrition*, 138(6), 1244S-1249S.

Xiaobo, Z., Jiewen, Z., Povey, M. J., Holmes, M., & Hanpin, M. (2010). Variables selection methods in near-infrared spectroscopy. *Analytica chimica acta, 667*(1-2), 14-32.

Xu, H., Qi, B., Sun, T., Fu, X., & Ying, Y. (2012). Variable selection in visible and near-infrared spectra: Application to on-line determination of sugar content in pears. *Journal of Food Engineering, 109*(1), 142-147.

Xu, R., Hu, W., Zhou, Y., Zhang, X., Xu, S., Guo, Q., ... & Wang, J. (2020). Use of

near-infrared spectroscopy for the rapid evaluation of soybean [Glycine max (L.) Merri.] water soluble protein content. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 224*, 117400.

Yaklich, R. W., Vinyard, B., Camp, M., & Douglass, S. (2002). Analysis of seed protein and oil from soybean northern and southern region uniform tests. *Crop Science, 42*(5), 1504-1515.

Yamada, H., Moriya, H., & Tsugita, A. (1991). Development of an acid hydrolysis method with high recoveries of tryptophan and cysteine for microquantities of protein. *Analytical biochemistry, 198*(1), 1-5.

Yan, Z., Lauer, J. G., Borges, R., & De Leon, N. (2010). Effects of genotype× environment interaction on agronomic traits in soybean. *Crop Science, 50*(2), 696-702.

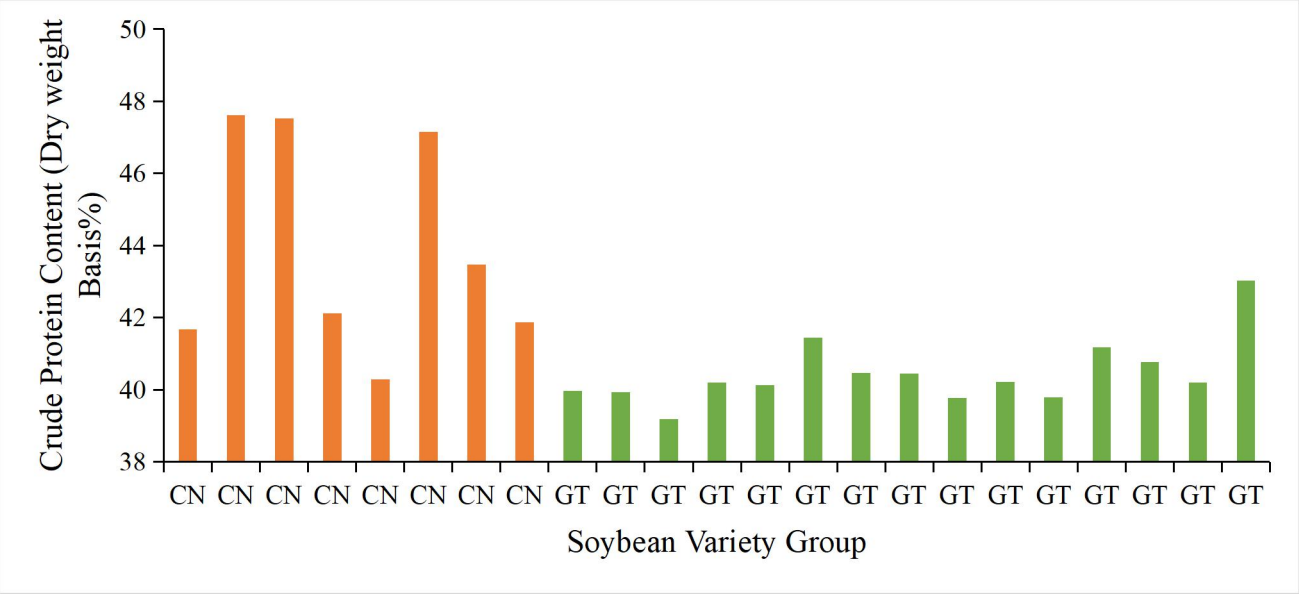
Yin, X., & Vyn, T. J. (2005). Relationships of isoflavone, oil, and protein in seed with yield of soybean. *Agronomy Journal, 97*(5), 1314-1321.

Young, V. R. (1991). Soy protein in relation to human protein and amino acid nutrition. *Journal of the American Dietetic Association*, 91(7), 828-835.

Zhu, Z., Chen, S., Wu, X., Xing, C., & Yuan, J. (2018). Determination of soybean routine quality parameters using near‐infrared spectroscopy. *Food science & nutrition, 6*(4), 1109-1118.

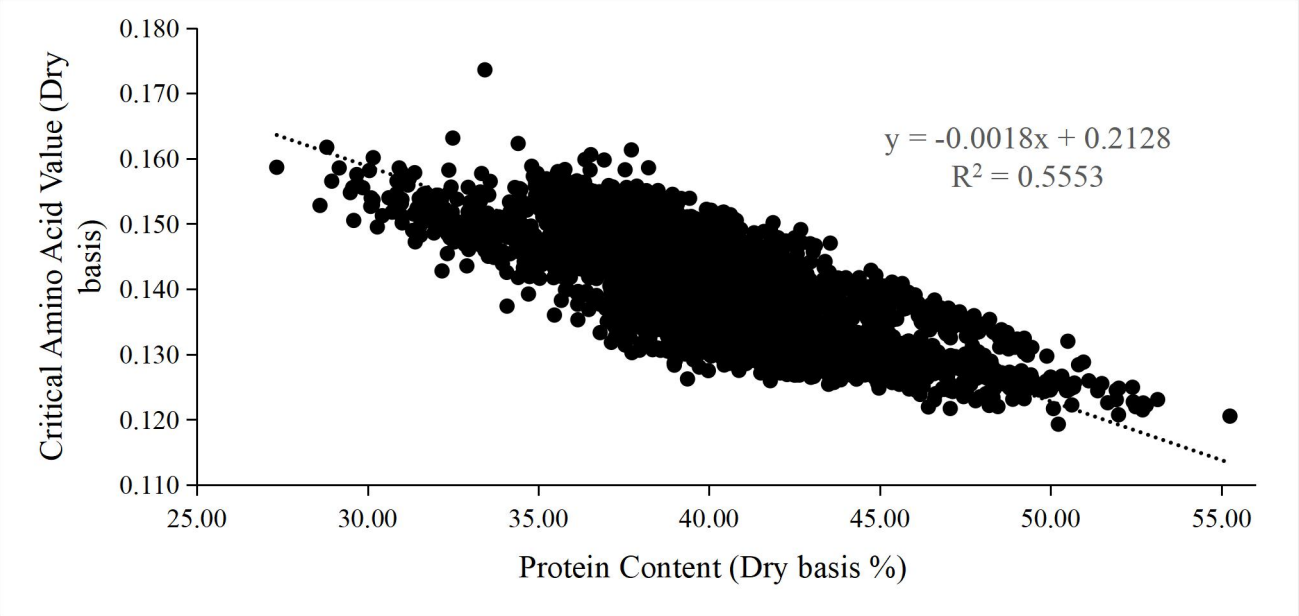
Zupan, J., & Gasteiger, J. (1993). Neural networks for chemists: an introduction. John Wiley & Sons, Inc..

**Appendix I**




Crude protein of soybean from conventional (CN) and glyphosate-tolerant (GT) groups

**Appendix II**




Linear regression of critical amino acid value% versus crude protein content for 4750 soybean samples from Manitoba during 2018 and 2019 cropping years