Extracting Numerical Information about Corn Composition from Texts

Nicholas PIPPENGER
Acxiom Corporation, Conway, AR 72032 USA

Richard S. SEGALL
Department of Computer & Information Technology, College of Business
Arkansas State University, State University, AR 72467-0130 USA

Daniel BERLEANT
Department of Information Science, University of Arkansas at Little Rock,
Little Rock, AR 72204 USA

Kellye A. EVERSOLE and Robert A. MUSTELL
Infinite Eversole Strategic Crop Services, LLC, Jonesboro, AR 72404 USA

Deborah VICUNA-REQUESENS
721 Ashurst, Philadelphia, PA 19083

Elizabeth E. HOOD
College of Agriculture and Technology, Arkansas State University, State University, AR 72467
Infinite Eversole Strategic Crop Services, LLC, Jonesboro, AR 72404 USA

ABSTRACT

The objective of this paper is to evaluate information quality processes and text mining methods that can be utilized to improve the extraction of numerical information from scientific articles related to the commodity agricultural crop corn. Specifically, this paper discusses the information extraction of crude protein content of corn.

Keywords: Text mining, Corn, Filtering

1. INTRODUCTION

Information about "crude protein" of corn components was extracted with four separate data identification and extraction methods. The first method made use of keyword based filtering to identify short passages that either contained "crude protein" corn substances being evaluated or a combination of one of the keywords being evaluated as well as the keyword "corn." The second method used relevancy filtering to identify short passages that contained the term "corn" in the title of the journal article from which they came. The third method used metadata filtering to identify short passages that corresponded to journal articles that were cited in at least 10 other journal articles. The fourth method used distance filtering to identify records containing short passages in which the number of characters between the term "corn" and the term "protein" was at most a specified maximum value.

Additionally, another variation of distance filtering was used to identify records containing short passages in which the maximum number of characters between the term "corn" and the word "percent" or "percentage" or the percent sign at most a specified maximum value. Generally speaking, the most successful method of the four that were tested was the first variation of the distance filtering method which identified records containing short passages with a maximum character distance of 500 between the terms "corn" and "protein". In terms of information quality related issues, the two primary ones were related to the data quality dimensions of accessibility and the amount of data. The accessibility of the dataset being evaluated was problematic from an input/output throughput perspective due to the large size of the dataset being evaluated (50,201,283 records, each containing a passage from a scientific paper). As a result, it was common for SQL queries to run for an extended period of time even when table indices for record identification were implemented.

2. BACKGROUND

The motivation for initiating this research was the large amount of information required in petitions for achieving non-regulated status of genetically engineered (GE) crops. This has become a significant burden in both the United States (USA) and Canada (McHughen and Smyth, 2008; Smyth and McHughen, 2008). This is in part because the concept of "substantial equivalence" is vague. Proving substantial equivalence between a GE crop and the non-GE crop is difficult without guidance as to what the non-GE baseline is to which one is comparing data. It is important to standardize this concept as much as possible to ensure confidence in the comparison.

Thus, a suite of technologies should be established that are standard in the industry and informed by scientific merit (Shewry et al., 2007). Shewry et al. surveyed genetically modified (GM) and non-GM wheat varieties in field trials and showed that no significant differences were attributable to the biotechnology used for preparing the new lines. While in this case the GM crop, wheat, met the criterion of substantial equivalence, the conclusion was reached only after many data were laboriously collected. Standardization of this collection process would contribute significantly to acceptance of biotechnology-derived crops.

We have chosen to investigate the commodity crop of corn because of the plentiful amount of data about it in the scientific literature. The research plan was to mine the literature for criteria important for deregulation, for example, composition of seed, composition of leaves, growth in a variety of environments, plant growth in multiple environments, etc. The best chance to find available data is for a commodity crop, such as corn (Hood et al., 2007) or rice (Chawla et al., 2006). A number of transgenic corn events and progeny of this crop are being grown in a variety of environments and thus seed and plants will be available to compare to published field-derived data about non-GE crops. Analyses of these transgenic lines needs to include for example their seed protein, carbohydrate, and oil content. In addition, two dimensional gel analysis needs to be used to determine protein variation in samples grown in two environments. In future work, we will collaborate with a breeder to begin gathering data on field performance.

Preliminary research performed by the co-authors of this paper have been presented in Berleant et al. (2010), Vicuna-Requesens et al. (2010a, 2010b), Hood et al. (2011), Pippinger (2014), and Pamarthi (2010).

The dataset utilized for this project consists of 38,343 digital scientific articles that were identified by searching for articles that were identified by the keyword "corn." As a result, the articles primarily come from journals with a biological focus. The two sources of the articles were *PubMed* and *ScienceDirec*t, with the latter being the primary source of the articles. The specific numerical information queried for the particular study described herein was related to the composition of corn and the percentage of crude protein.

The accuracy and correctness of experiments conducted are directly related to data quality dimensions and those that were particularly relevant in this research include accessibility, completeness, and amount of data, relevancy, and believability as discussed below:

- The dimension of accessibility refers to the extent to which the article data can be stored, maintained, and retrieved from the relational MySQL database server.
- The dimension of completeness refers to the extent to which the articles loaded into the relational database are able to return adequate results for representing the corn substances queried. The amount of data dimension refers to the extent to which the volume of records in

the relational database is sufficient enough to provide adequate responses to the SQL queries.

- The data quality dimension of relevancy refers to the extent to which the data repository is appropriate for the purpose of returning meaningful results to the different corn queries
- The believability dimension refers to the extent to which the data results seem to be unbiased and objective.

3. EXTRACTION OF NUMERICAL INFORMATION

The majority of the issues encountered while attempting to identify and extract specific numerical information from the corn related digital article dataset were data quality-centric. A MySQL database server was used to store processed data. The primary advantage of using a relational database server to evaluate the repository is the ability to use SQL queries to retrieve data from the repository. In addition to being a standardized easy-to-read querying language, there is also a significant amount of SQL documentation that is readily available. After the dataset was successfully loaded into an initial staging table in the MySQL database, it was possible to evaluate the data.

3.1 Methodology

In order to evaluate the dataset of 38,343 scientific articles, the digital articles were loaded to an initial staging table in a MySQL database. The total record count of the initial staging table (all_sentence_stage_700) was 50,201,283 records. Each record contained a string of text beginning at a sentence boundary and potentially containing multiple sentences of text. The data model for this table was relatively simple and only had three fields. The three fields are file_name, sentence, and sentID. Several data quality processes were performed on this dataset to prepare the articles for uploading into the MySQL staging table.

The first data quality improvement task that was performed on the original source data was identifying and removing HTML tags from the source articles, which were stored in HTML format. In order to accomplish this task a short Python program evaluated each of the HTML article files and generated corresponding flat files that did not contain the HTML tags.

The next data quality process that was performed was loading the modified article files into the initial MySQL staging table. The primary key for the initial staging table was the sentID field, a numerical ID that auto-increments with each record added to the table. The initial approach to loading the article short passages involved enforcing a constraint on the maximum passage length. This constraint was needed because our version of MySQL only supported indexing of fields with a maximum length of approximately 750 characters.

The sentence field was populated by generating a set of intermediate files that had a one-to-one correspondence with the 38,343 modified digital articles. This new set of flat files was generated by a python script that read the first 600 bytes of a given article file and then continued to read the next byte until a space character was encountered. Once a space character was detected the python script wrote out the stored sequence of characters to a text file followed by the newline character. The advantage of using this technique was that it limited the maximum length of the sentence field. The maximum short passage length in the 50,201,282 record dataset using this approach was 708 characters. This is below 750, enabling us to use MySQL to generate an index on the sentence field within the database to improve performance. Another advantage to this approach was that it resulted in a more uniform sentence field within the MySQL table. The maximum short passage length of 708 characters also aided the process of manually auditing the data. After the final set of article files was generated with the defined maximum short passage length value, the Python program loaded the short passages within the articles into the MySQL staging table using the “MySQLdb” python module.

3.2 Findings and Results

The purpose of the initial article scoring algorithm was to numerically rank article short passages that contained potentially relevant keywords. The short passages were identified by querying the full staging table and selecting records that contained one or more keywords. The list of keywords that were included in the selection query are listed below. This list of keywords is not case sensitive. The record count of the table containing the selected records is 1,685,081, which represents 3.36% of the total dataset.

3.3 Crude Protein Percentage of Corn

The crude protein content of corn refers to the amount of protein present when measured, for example, based on nitrogen content. Figure 1A is a histogram representing the crude protein content of corn. It was generated by identifying and parsing short passages that contained the terms "percent crude protein" or "% crude protein" as well as a numerical value. A bin size of 1 was used in the generation of the histogram. Numerical values after the decimal point in non-integer values were truncated. As a result, non-integer values such as 2.4 or 2.7 would be associated with the bin for 2–3 in the histogram. The expected percentage of crude protein in corn can vary depending on the type of corn and its growth conditions.

Typical crude protein percentages reported for a mean of 20 high protein cultivars are 10.4–11.6% (De Geus et al., 2008). A selection of commercialized ZP lines (504su, 531su, 74, 611k, Rumenka, 434, and 633) ranged from 10.13–13.27% protein (Zilic et al., 2011). Drinic et al. (2014) also investigated ZP lines, reporting a range of 9.85–12.84% protein over a set of 9 inbred lines (ZPL1 through ZPL9) and a narrower range of 9.81–11.42% protein over 8 hybrids. Application of nitrogen fertilizer was reported to increase protein content in two "popular varieties of Pioneer corn," the names of which the authors did not release citing lack of company permission. Their mean protein content under various experimental conditions of fertilizer application ranged from 7.1–9.9%, with minimums and maximums ranging from 5.7–11.0% (Singh et al., 2005).

The present histogram results show extraction from text passages of putative corn crude protein percentages that were often compatible with these data, yet also often tended high.

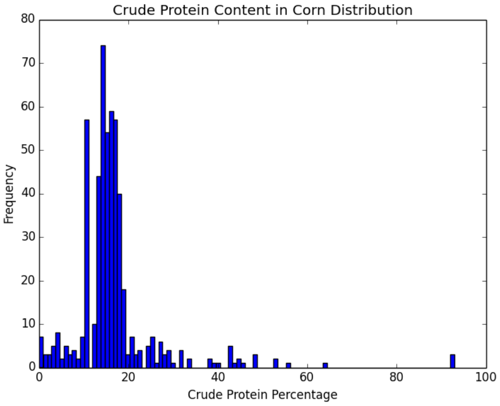

Figure 1A. Short passages containing the Term "Crude Protein"

The histogram illustrated in Figure 1A above was created from the file of 11,578 records each containing a short passage, filtered to use those containing the term "crude protein." The first major spike on the histogram occurs at 11 percent with 57 occurrences. There is also a spike at 13 percent with 44 occurrences. There are 74 occurrences with a corn crude protein of 14 percent. There are also spikes at 15 percent with 54 occurrences, 16 percent with 59 occurrences, 17 percent with 57 occurrences, and 18 percent with 38 occurrences. Beginning with 19 percent there is a drop off in the histogram with only 18 occurrences. There are only 3 occurrences at 20 percent and 7 occurrences at 21 percent. Beyond twenty-one, there are no occurrences greater than 7 for a given percentage.

After manually inspecting several of the short passages that produced the histogram in Figure 1A, it was apparent that in many cases the crude protein percentage in the short passage was not describing the crude protein content of corn. To improve the results, we created another MySQL table which was a subset of the original table and that required that each short passage also contain the term "corn." The modified dataset contained a total of 1,592 short passages. Of those 1,592 records only the ones that contained the term "crude protein" and "corn" were filtered and used to produce the following histogram as illustrated below in Figure 2A.

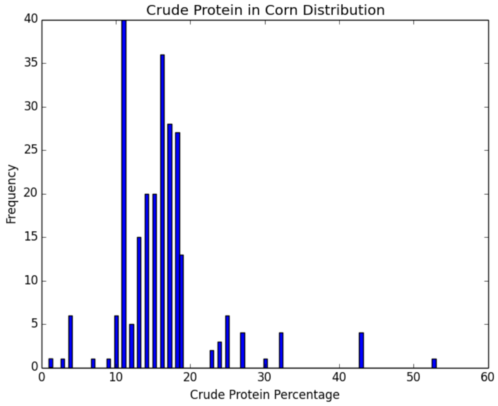

Figure 2A Short passages Containing Term "Crude Protein" and "Corn"

The results illustrated in this histogram of Figure 2A are more consistent with the expected corn crude protein percentage. The largest spike occurs at 11 percent with 40 occurrences. This is an improved result.

As an alternative to requiring that the term "corn" be present in each short passage, we created a third table that is a subset of the dataset used in Figure 1A in which the term "corn" was required to be present in the journal article title rather than the short passage. This resulted in Figure 3A from the creation of a new MySQL table containing 1,359 records (sentence_rank_table_1a_crude_protein_art). Of those 1,359 records only the ones that contained the term "crude protein" and journal article title with term 'corn' were filtered and used to produce the histogram as shown in Figure 3A. In this histogram of Figure 3A, the largest spike is at 15% with 32 occurrences. Requiring that the term "corn" be present in the journal article title rather than in the short passage did not improve the query results. Instead, the previous method of requiring that the term corn be present in the actual short passage generated better results.

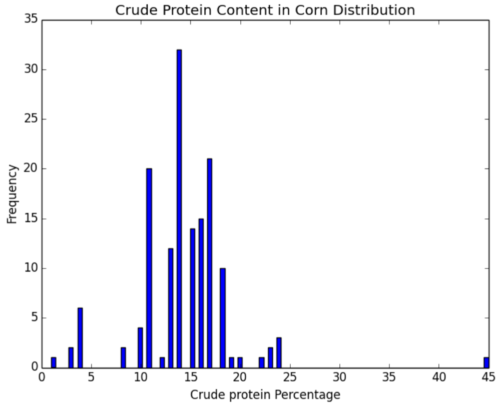

Figure 3A: Short passages Containing Term "Crude Protein" Where Journal Article Title Contains Term "Corn"

Figure 4A was generated by creating a subset of the table used to produce Figure 1A in which each short passage evaluated was required to come from an article that was cited at least 10 times in the literature. This was an attempt of using metadata filtering to improve the crude protein results. The highest number of hits was at 17% protein. This resulted in the generation of a new MySQL table (sentence_rank_table_1a_crude_protein_cite) consisting of a record count of 7,376 records.

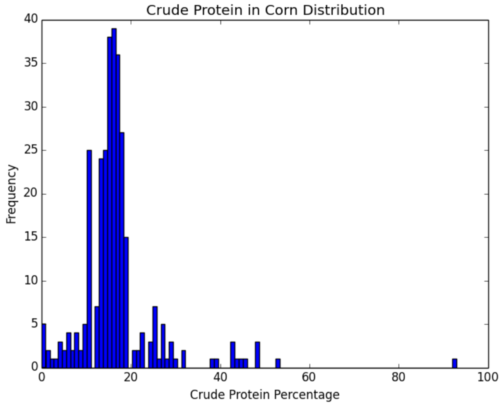

Figure 4A: Short passages Containing Term "Crude Protein" Where Journal Article has at Least 10 Citations

A fifth crude protein dataset was generated by modifying the MySQL table creation statement used to create the dataset used in Figure 1A by requiring that each article contain both the key word "corn" as well as the term "crude protein." Also, the maximum number of characters distance between the beginning of one term and the beginning of the second term had to be equal to or less than 500 characters. This resulted in the creation of Figure 5A from short passages filtered from a new MySQL table containing 1,546 records (sentence_rank_table_1a_crude_protein_dist1_500_5a). In this table, the largest spike is at 11% with 40 occurrences. There is also a spike at 16% with 35 occurrences.

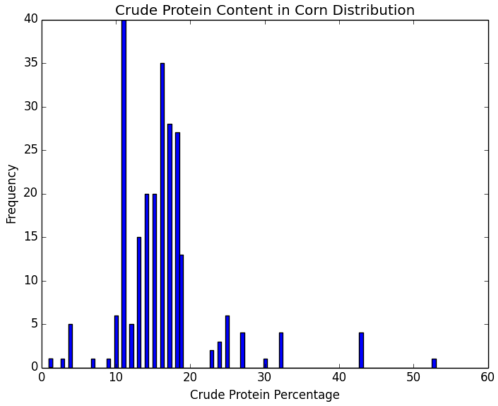

Figure 5A: Distance between Terms "Crude Protein" and "Corn" is at most 500 Characters

A sixth crude protein dataset was generated by modifying the MySQL table creation statement used to create the dataset used in generating Figure 1A by requiring that each short passage contain both the term "crude protein" as well as the word "percent" or "percentage" or the percent sign. Also, the maximum character distance between the two terms had to be equal to or less than 500 characters. The character distance was determined by computing the absolute value of the difference between the numerical position of the first letter of the term "percent" or "percentage" or the percent sign measured from the beginning of the short passage and the numerical position of the first letter of the term "crude protein" measured from the beginning of the short passage. This resulted in the creation of Figure 6A using short passages filtered from a new MySQL table containing 11,578 records (sentence_rank_table_1a_crude_protein_dist2). In this table, the largest spike is at 14% with 74 occurrences. There is also a spike at 11% with 57 occurrences.

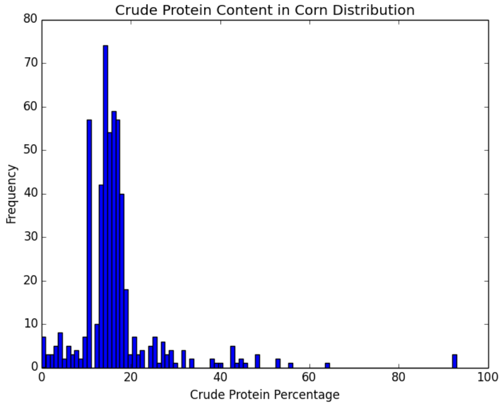

Figure 6A: Distance between Terms "Crude Protein and "Percent/%" is at most 500 Characters"

A seventh crude protein dataset was generated by modifying the MySQL table creation statement used to create the dataset from which short passages were filtered for Figure 1A by requiring that each short passage contain both the term "crude protein" as well as the word "percent" or "percentage" or the percent sign. Also, the maximum character distance between the two terms had to be equal to or less than 200 characters. The character distance was determined by computing the absolute value of the difference between the numerical position of the first letter of the term "percent" or "percentage" or the percent sign measured from the beginning of the short passage and the numerical position of the first letter of the term "crude protein" measured from the beginning of the short passage. This resulted in Figure 7A, based on short passages filtered from a new MySQL table containing 7,393 records (sentence_rank_table_1a_crude_protein_dist2_200).

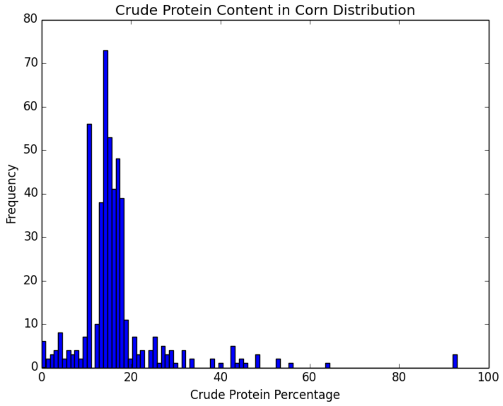

Figure 7A: Distance between Terms "Crude Protein and "Percent/%" is at most 200 Characters"

In Figure 7A, the largest spike is at 14% with 73 occurrences. There is also a spike at 11% with 56 occurrences.

An eighth crude protein dataset was generated by modifying the MySQL table creation statement used to create the dataset from which short passages were filtered for use in Figure 1A by requiring that each short passage contain both the term "crude protein" as well as the word "percent" or "percentage" or the percent sign. Also, the maximum character distance between the two terms was considerably lowered, to less than or equal to 15 characters. This character distance was determined by computing the absolute value of the difference between the numerical position of the first letter of the term "percent" or "percentage" or the percent sign measured from the beginning of the short passage and the numerical position of the first letter of the term "crude protein" measured from the beginning of the short passage. This resulted in the creation of Figure 8A based on short passages filtered from a new MySQL table containing 2,158 records (sentence_rank_table_1a_crude_protein_dist2_15). In this table, the largest spike is at 14% with 66 occurrences. There is also a spike at 11% with 40 occurrences.

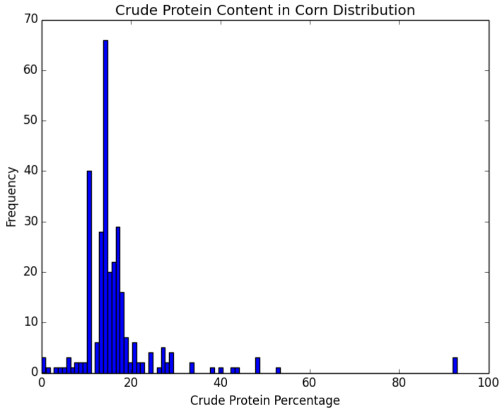

Figure 8A: Distance between Terms "Crude Protein and "Percent/%" is at most 15 Characters"

4. CONCLUSIONS AND FUTURE DIRECTIONS

In terms of the results for the term "crude protein," the data identification and extraction method that produced the best results was the first variation of the distance filtering method in which the maximum number of characters between the terms "corn" and "crude protein" were limited to a maximum of 500 characters. The results for this are illustrated in histogram 5A. The greatest spike in the histogram is at 11% with 40 occurrences which is within the expected range for the crude protein content of corn. Future directions for this research include performing the same four data identification and extraction methods for six other corn components. These are dry matter, ash, crude fiber, starch, crude fat, and sulfur.

5. ACKNOWLEDGEMENTS

The authors of this paper acknowledge the support of this research by Arkansas Science and Technology Authority (ASTA) under National Science Foundation (NSF) Grant Award #EPS-0701890, Subgrant P3-209, and by a subcontract from Infinite Eversole Strategic Crop Services, LLC, USDA SBIR 2009-33610-19721.

REFERENCES

[1.] Atiff, S., and Segall, R. (2010), Use of SAS Text Miner in Bioinformatics, ***Poster Presentation at MidSouth Computational Biology and Informatics Society (MCBIOS) Annual Conference,*** Arkansas State University Jonesboro, AR, February 19-20, 2010.

[2.] Berleant, D., Segall, R., Hood, E., Eversole, K., Vicuna, D., Atiff, S., Biedenbender, C., Pamarthi, J., and Pippenger, N. (2010), Enabling Crop Deregulation with Software: a Prototype, ***Presentation at Arkansas Plant-Powered Prodcution (P3) Symposium***, Winthrop Rockefeller Institute, Pete Jean Mountain, AR, August 17, 2010.

[3.] Chawla, R., Ariza-Nieto, M., Wilson, A. J., Moore, S. K., and Srivastava, V. (2006), Transgene Expression Produced by Biolostic-Mediated, Site-Specific Gene Integration Is Consistently Inherited by the Subsequent Generations, ***Plant Biotechnology Journal*** 5, 209-218.

[4.] De Geus, Y. N., Goggi, A. S., and Pollack, L. M. (2008), Seed Quality of High Protein Corn Lines in Low Input and Conventional Farming Systems, ***Agronomy for Sustainable Development*** 28:541-550.

[5.] Drinic, S. M., Dragicevic, V., Zilic, S., Basic, Z., and Kovacevic, D. (2014), Varijability [sic] of Tocopherol and b-Carotine Contents in Maize Genotypes, ***Journal of International Scientific Publications: Ariculture and Food*** 2:192-198, http://www.scientific-publications.net/en/article/1000026/.

[6.] Hood, E. E., Love, R., Lane, J., Bray, J., Clough, R., Pappu, K., Drees, C., Hood, K. R., Yoon, S., Ahmad, A., and Howard, J. A. (2007), Subcellular Targeting Is a Key Condition for High Level Accumulation of Cellulase Protein In Transgenic Maize Seed, ***Plant Biotechnology Journal*** 5, 709-719.

[7.] Hood, E. E., Eversole, K. A., Berleant, J. D., Segall, R. S., Mustell, R. A., and Requesens, D. V., Method and System for Data Collection and Analysis to Assist in Facilitating Regulatory Approval of a Product, ***US Patent Application Publication [US 2011/0224933 A1]***, filed February 1, 2011, published Sept 15, 2011.

[8.] McHughen, A., and Smyth, S. (2008), US Regulatory System for Genetically Modified [Genetically Modified Organism (GMO), rDNA or Transgenic] Crop Cultivars, ***Plant Biotechnology Journal*** 6, 2-12.

[9.] Pararthi, Jagadish (2010), Extracting Properties of Crops from Web Data for Deregulation Using ProExTrac, ***Master's Thesis, Department of Computer Science, University of Arkansas at Little Rock, AR USA***.

[10.] Pippinger, N. (2014), Information Quality Processes and Methods to Improve Extraction of Numerical Information from Unstructured Text, ***Master's Project, Program in Information Quality, University of Arkansas at Little Rock, AR USA.***

[11.] Shewry, P. R., Baudo, M., Lovegrove, A., Powers, S., Napier, J. A., Ward, J. L., Baker, J.M., and Beale, M. H. (2007), Are GM and Conventionally Bred Cereals Really Different? ***Trends in Food Science and Technology*** 18, 201-209.

[12.] Singh, M., Paulsen, M. R., Tian, L., and Yao, H. (2005), Site-Specific Study of Corn Protein, Oil, and Extractable Starch Variability Using NIT Spectroscopy, ***Applied Engineering in Agriculture,*** 21(2):239-251.

[13.] Smyth, S., and McHughen, A. (2008), Regulating Innovative Crop Technologies in Canada: The Case of Regulating Genetically Modified Crops, ***Plant Biotechnology Journal*** 6, 213-225

[14.] Vicuna-Requesens, D. V., Eversole, K. A., Mustell, R. A., Segall, R. S., Berleant, D., and Hood, E. (2010a), Establishing a Baseline Database to Demonstrate Substantial Equivalence of GE and Non-GE Crops Through Data Mining and Text Mining, ***Poster #36, Arkansas Plant-Powered Production (P3) Symposium,*** Winthrop Rockefeller Institute, Pete Jean Mountain, AR, August 15-17, 2010.

[15.] Vicuna-Requesens, D. V., Eversole, K. A., Mustell, R. A., Segall, R. S., Berleant, D., and Hood, E. (2010b), Establishing a Baseline Database to demonstrate Substantial Equivalence of GE and Non-GE Crops through Data Mining and Text Mining, Abstract #P01013, ***Plant Biology 2010: Joint Annual Meeting of the American Society of Plant Biologists (ASPB)***, Montreal, Quebec, Canada, July 31-August 4, 2010.

[16.] Zilic, S., Milasinovic, M., Terzic, D., Barac, M., and Ignjatovic-Micic, D. (2011), Grain Characteristics and Composition of Maize Specialty Hybrids, ***Spanish Journal of Agricultural Research*** 9(1):230-241.