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 used to improve the extraction of numerical information from scientific articles about the commodity agricultural crop corn. Specifically, this paper focuses on extraction of crude protein content of corn, an important special case illuminating the general problem.

Keywords: Text mining, Corn, Filtering

1. INTRODUCTION

Information about crude protein, an important corn component, was extracted by four separate data identification and extraction methods. The first method used keyword-based filtering to identify short passages that contained the phrase "crude protein" and the keyword "corn." The second method used relevancy filtering to identify short passages from journal articles whose titles contained the term "corn." The third method used metadata filtering to identify those short passages that were in journal articles that had been cited in at least 10 other journal articles. The fourth method used distance filtering to identify short passages in which the number of characters between key terms was at most a specified maximum value.

Generally speaking, the most successful method of those tested was the distance filtering method, in which short passages were restricted to those with a maximum distance of 500 characters between the terms "corn" and the phrase "crude protein."

With respect to information quality, we discuss several dimensions, especially the two data quality dimensions of accessibility and amount of data.

2. BACKGROUND

The motivation for initiating this research was the large amount of information required in petitions for achieving non-regulated status for

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 required, but challenging because of the difficulty of appropriately characterizing the non-GE baseline to which GE crop data is compared.

It is important to standardize the baseline 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 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 to deregulation decisions, for example composition of seed, composition of leaves, growth in a variety of environments, plant growth in multiple environments, etc. The best chance for finding sufficient 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 contents. In addition, two dimensional gel analysis needs to be used to determine protein variation in samples grown across environments.

Preliminary research for this paper was reported by Berleant et al. (2010), Vicuna-Requesens et al. (2010a, 2010b), Hood et al. (2011), Pippinger (2014), and Pamarthi (2010).

2.1 Data Quality Issues

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 here 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 regarding 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 to provide adequate responses to the SQL queries.
- The dimension of relevancy refers to the extent to which the data repository is appropriate for the purpose, in this case 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. EXTRACTING NUMERICAL INFORMATION

The dataset utilized for the project consisted of 38,343 digital scientific articles that were obtained by searching for articles based on the keyword "corn." As a result, the articles mainly come from journals with a biological focus. The two sources of the articles were *PubMed* and *ScienceDirect*, with the latter being the primary source of the articles. The specific numerical information of interest in this study was the

percentage of crude protein in the composition of corn.

A MySQL database was used to store processed data. The primary advantage of using this database to evaluate the repository is the ability to use SQL queries to retrieve data from the repository. SQL is a standard query language and there is a significant amount of SQL documentation readily available to support its use.

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. 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.

After the dataset was successfully loaded into the MySQL database, it was possible to evaluate the data.

3.1 Methodology

In order to evaluate the 38,343 scientific articles, the digital articles were loaded into 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 up to a maximum of 708 characters. The data model for this table was relatively simple and only had three fields. The three fields were file_name, sentence, and sentID. 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.

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 string-containing fields limited to a maximum length.

The sentence field was populated by generating a set of intermediate files with a one-to-one correspondence to 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 was below the length limit required for 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 examining the data. After the final set of article files was generated, 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 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 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 in Corn

The crude protein content of corn refers to the amount of protein present as estimated from its nitrogen content. Figure 1 is a histogram representing the crude protein content of corn. It was generated by identifying and parsing short passages that contained both the phrase "crude protein" and 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 are as follows. For a mean of 20 high protein cultivars, 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). Based on these sources, we consider histogram bars associated with bins for 7–8% through 13–14% to be valid results in our analyses below that estimate the information retrieval precision of the data behind the histogram figures.

Figure 1, based on extraction from text passages of putative corn crude protein percentages, shows text mining results often compatible with chemical analysis data, yet also often tending high.

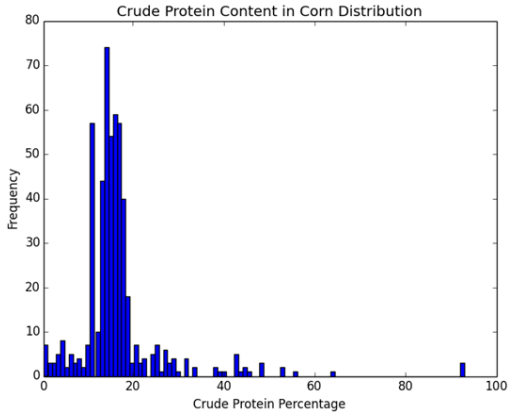

Figure 1. Short Passages Containing the Phrase "Crude Protein"

The first major spike on the histogram occurs at 11–12 percent with 57 occurrences. There is also a spike at 13–14 percent with 44 occurrences. There are 74 occurrences with a corn crude protein of 14–15 percent. There are also spikes at 15–16 percent with 54 occurrences, 16–17 percent with 59 occurrences, 17–18 percent with 57 occurrences, and 18–19 percent with 38 occurrences. Beginning with 19–20 percent there is a drop off in the histogram with only 18 occurrences. The information retrieval precision represented by Figure 1 is 24.7%.

After manually inspecting several of the short passages that produced the histogram in Figure 1, 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 to also contain the term "corn." This was used to produce Figure 2.

The results illustrated in the histogram of Figure 2 are more consistent with the expected corn crude protein percentage. The largest spike occurs at 11–12 percent with 40 occurrences. This is an improved result. The information retrieval precision represented by Figure 2 is 27.8%.

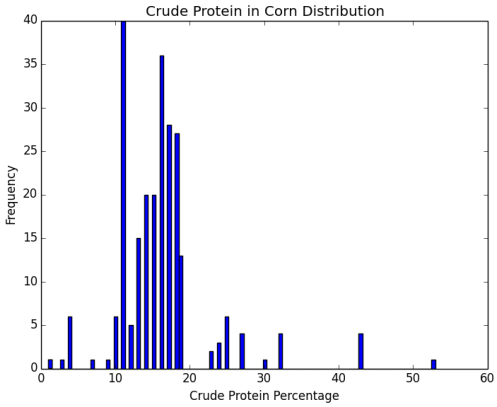

Figure 2: Short Passages that Also Contained the Term "Corn"

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 1 in which the term "corn" was required to be present in the journal article title rather than the short passage. This resulted in Figure 3. In the histogram of Figure 3, the largest spike is at 15–16% 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, as indicated by the observation that the information retrieval precision represented by Figure 3 is 26.2%, little changed.

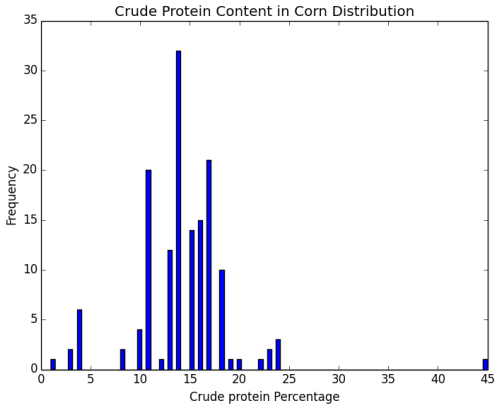

Figure 3: Short Passages Containing the Phrase "Crude Protein" Where Article Title Contains the Term "Corn"

Figure 4 was generated by creating a subset of the table used to produce Figure 1 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 17–18% protein, which is out of range. Bolstering that observation is that the information retrieval precision was 23.5%, a bit lower than for the previously discussed histograms. This method therefore did not produce quite as good results.

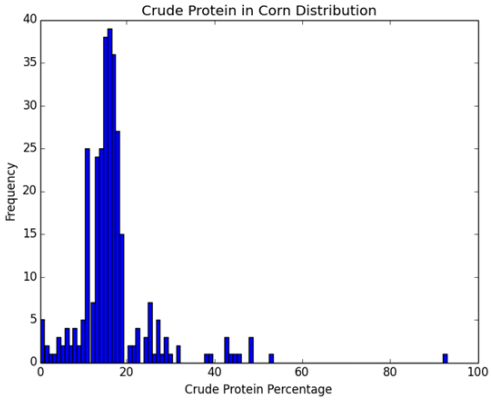

Figure 4: Short Passages Containing the Phrase "Crude Protein" Where Journal Article Had at Least 10 Citations

A fifth crude protein dataset was generated by modifying the MySQL table creation statement used to create the dataset used in Figure 1 by requiring that each article contain both the key term "corn" as well as the phrase "crude protein" and, in addition, the maximum number of characters between the beginning of one term and the beginning of the other term was 500 characters or less. This resulted in the creation of Figure 5 from short passages filtered from the new MySQL table. This is nearly but not exactly the same as Figure 2, indicating that the 500 character restriction was quite weak. In this table, the largest spike is at 11–12% with 40 occurrences. There is also a spike out of range at 16–17% with 35 occurrences. Nevertheless the information retrieval precision was the best obtained so far, at 28.0%, though the improvement over Figure 2 is tiny.

A sixth crude protein dataset was generated by modifying the MySQL table creation statement used to create the dataset used in generating Figure 1 by requiring that each short passage contain both the phrase "crude protein" as well as the word "percent" or "percentage" or

the percent sign "%." Also, the maximum distance between their beginnings had to be 500 characters or less. The distance was determined by computing the absolute value of the difference between the position of the first letter of the term "percent," "percentage," or "%" and the 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 6. In this table, the largest spike is out of range at 14–15% with 74 occurrences. There is also an in-range spike at 11–12% with 57 occurrences. Overall the information retrieval precision was 24.4%.

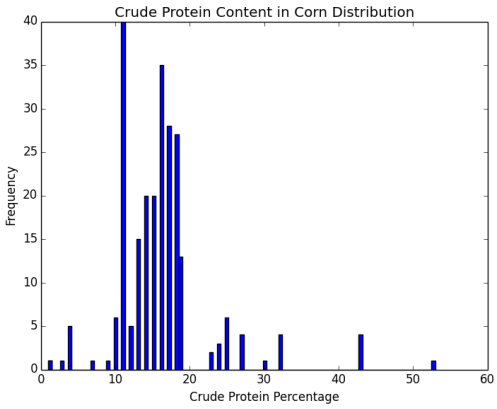

Figure 5: Distance Between Phrase "Crude Protein" and Term "Corn" Is at Most 500 Characters

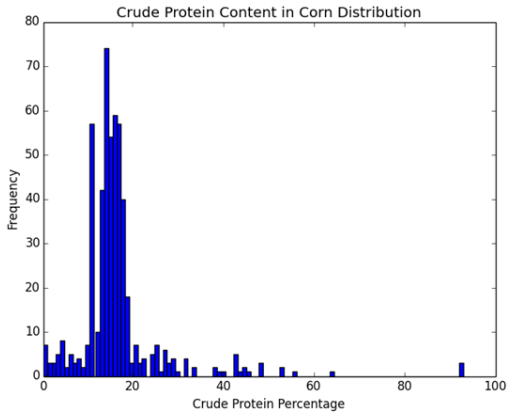

Figure 6: Distance Between "Crude Protein and "Percent," "Percentage, or "%" Is at Most 500 Characters

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 1 by requiring that each short passage contain both the phrase "crude protein" as well as the word "percent," "percentage," or "%," as also the maximum distance between them had to be 200 characters or less. This resulted in Figure 7.

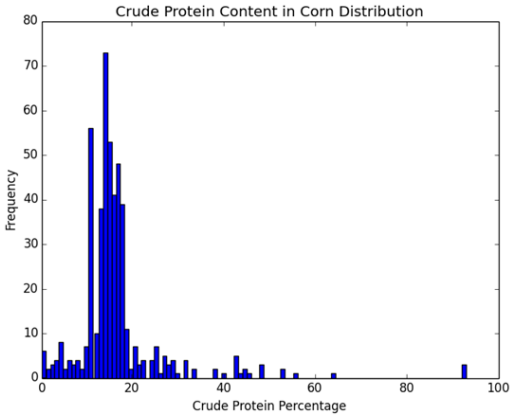

Figure 7: Distance Between Phrase "Crude Protein and Term "Percent," "Percentage," or "%" Is at Most 200 Characters

In Figure 7, the largest spike is at 14–15% with 73 occurrences. There is also a spike at 11–12% with 56 occurrences. The information retrieval precision of 25.7%.

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 1, by requiring that each short passage contain both the phrase "crude protein" as well as the word "percent," "percentage," or "%," and also the maximum distance between their beginnings was considerably lowered to 15 characters or less. This resulted in the creation of Figure 8. In this table, the largest spike is at 14–15% with 66 occurrences. There is also a spike at 11–12% with 40 occurrences. The precision was the lowest of all the histograms shown here, at 18.9%.

4. CONCLUSIONS AND FUTURE DIRECTIONS

For the phrase "crude protein," data identification and extraction seemed to produce the best results when the term "corn" was required to be present in proximity to the phrase "crude protein" (Figure 2). Adding distance filtering in which the maximum number of characters between the terms "corn" and "crude

protein" was limited to 500 characters made no significant difference (Figure 5). The highest spike in these histograms occurs at 11–12% 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.

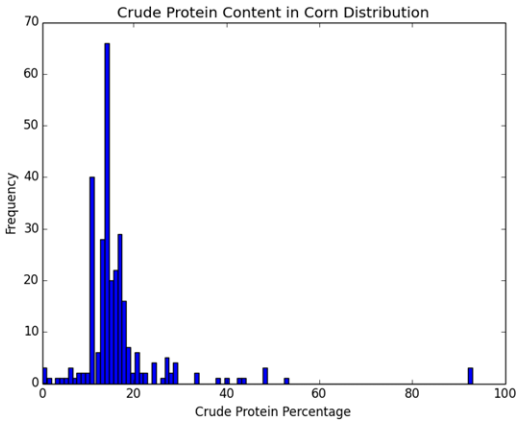

Figure 8: Distance Between Phrase "Crude Protein" and Term "Percent," "Percentage," or "%" Is at Most 15 Characters

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. Supported in part by grants from NCRR (5P20RR016460-11) and NIGMS (8 P20 GM103429-11) at NIH.

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 Biolistic-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.] Pamarthi, J. (2010), Extracting Properties of Crops from Web Data for Deregulation Using ProExTrac, ***Master's Thesis, Dept. 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.