

# Methods for the Automated Annotation of Maize Genes

Josh Kangas

April 28, 2008

## **Abstract**

With the advent and improvement of microarray technology has come a dramatic increase in the raw data available for analysis. The difficulty comes in extracting useful information from this mass of data. Biologists have undertaken efforts to annotate genes which are spotted onto microarray chips. The result of this effort has been the development of a large set of high quality annotations. Various constraints, such as time and funding, have kept these annotations from thoroughly covering the entire scope of experiments allowed by the microarray technology. In previous work, a system was developed which, given a gene sequence, was able to provide a functional annotation in the form of a value of membership for each category in the set of functional categories used in this project. The mechanism for the determination of any category classification was a back-propagation neural network which is a form of artificial neural network. The results of the artificial neural network annotations are from inputs which are based on the presence or absence of keywords in abstracts of related publications accessed from a BLASTx query. As an extension of these prior efforts, proposed is the development of an array of software tools to enable biologists to access the neural network system through a web interface for annotation of unique sets of sequences. Also the website will provide the users with a set of tools for the analysis and visualization of clustered automated annotation results. Additionally, the goal of the research team is to complete and submit a manuscript to *Bioinformatics* for peer review.

## **1 Introduction**

A group of cells in a plant called the shoot apical meristem (SAM), is ultimately responsible for the development of all of the above ground tissues in a mature plant [2]. Because the SAM is so important in plant growth and development biologists are interested in advancing their knowledge of the function of the cells composing this tissue. The National Science Foundation (NSF) has funded the

Maize Shoot Apical Meristem Project<sup>1</sup> (NSF Award No. 0321595<sup>2</sup>). The aim of this project is to identify, analyze and categorize the genes involved in meristem function and early stages of leaf development. Laser Capture Microdissection (LCM) is used to procure pure cells from different regions of the maize SAM for use in microarray hybridization studies. This allows the direct comparison of gene expression patterns in regions of the SAM and in developmental mutants. A comparative analysis of patterns in gene expression levels in meristems and leaf primordia helps provide clues to understanding the roles that different genes play in plant development [7].

The microarrays used in this project are populated primarily with maize cDNA sequences obtained from research labs in the U.S.; most of these cDNAs were generated during expressed sequence tag (EST) experiments and have not been characterized. The ESTs can be manually annotated and assigned to one or more functional categories using various research tools including primary literature, databases, and Internet sites [3]. These categories include cytoskeletal structure, transport, metabolism, transcription, translation and others. Some categories are large and need to be divided into subcategories; for example metabolism includes amino acid, sugar, lipid, energy, and nucleotide metabolism.

Since each of the four microarray chips used in this study has 15,000 to 20,000 gene ESTs, it is impractical to manually annotate each gene located on the chips. For the SAM Project, only genes which are significantly differentially regulated are annotated manually. The result of this effort is that approximately 6,500 of 30,000 unique genes on four chips have been annotated. The result of this effort is a set of quality annotations which can be used as a basis for any method of automatic annotation. The goal for automated annotation is to take the remaining unannotated genes and assign a functional category. The value of automatic annotation comes from the availability of a larger set of genes which on some level have a functional category assigned to them. Many of the genes that are computationally annotated are valuable to understand because they may interact with some of the proteins that are differentially regulated in mutants or in different regions of the SAM. This can identify many more genes that can be further studied.

## 2 Status of Previous Goals and Objectives

The goal of this project was to develop a set of tools to automatically annotate gene sequences in order to assist researchers in their explorations of the resulting data from microarray experiments. The following goals were set prior to the proposed work:

---

<sup>1</sup><http://maize-meristems.plantgenomics.iastate.edu/>

<sup>2</sup><http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0321595>

1. The name, functional category, and GO category will be determined for each EST using a computational method based on BPNNs.
2. The results of these methods will be checked against the annotations performed manually.
3. A 90% accuracy rate should be attained based on only using a portion of the total annotated genes. Additionally, false positive results are to be minimized.
4. The results will be displayed and stored in a meaningful method for public access and use.

A method was developed and implemented which yields a measure of membership across functional categories. Efforts to develop methods to accurately determine the name or GO category for a sequence were minimal and unsuccessful. Work on these remaining two areas will not continue on this project at this time. The functional categories used for annotation in this context were developed for this project specifically. They are, however, consistent with functional categories used in other projects of this nature.

A double blind test was used to determine the accuracy of the artificial neural networks. When sorting categories based on the top hits, the neural networks yielded the results shown in Table 1.

Top Hits	Categories Better Than Random (Out of 22)
1	13
2	14
3	15
4	17
5	19
6	20

Table 1: Number of category annotations exceeding random.

Subsequently, it was determined that double-blind testing was no longer necessary as there are better methods for analyzing results from the neural networks. After analyzing the outputs to determine if a cutoff value for activation would give resolution to the problem of categorization, it was found that the necessary cutoff varied greatly between categories. Using McNemar's marginal homogeneity test, we looked for the lowest cutoffs which yielded a  $p$ -value less than 0.5. The abbreviated results are shown in Table 2.

Additionally, work was done to shed some light on the actual membership values for each category. It was determined that from category to category the accuracy of the system varied greatly, so a single metric across categories is difficult to determine. From this analysis it is clear that a 90% accuracy rate will be difficult

Category Name	Cutoff
Cell Division	0.7
Cytoskeletal	0.75
Signal Transduction	0.95
Translation	0.85
Vesicle Trafficking	0.9
Defense	0.55
Photosynthesis-related	0.7
ATPase	0.3
Stress-related	0.4

Table 2: Categories required varying cutoffs for homogeneity with manual annotations indicating the need for category specific analyses.

to achieve. The inherent lack of accuracy resulting from the use of vocabulary words has necessitated the the reduction in the expected accuracy of the system. However, the system can still be useful for the reduction of the search field for genes of specific function.

A website was developed which allows users to query and view the results of the automated functional annotations used for testing purposes. There was limited functionality on the site and it mostly served as a proof of concept. Work will continue in this area.

### 3 Current Goals and Objectives

In order to assist researchers in their explorations of data from microarray hybridization experiments, the following goals are proposed in priority order:

1. Prepare a manuscript detailing the efforts for submission to *Bioinformatics*.
2. Fully annotate the sequences from the microarray chips and make the results publicly available.
3. Design and launch a website for the display and exploration of automated annotation information.
4. Allow users to enter unique sequences for annotation using the neural network system.
5. Develop a web interface to display the results of hierarchical clustering analyses.

## 4 Methodology

Prior to completing the manuscript, statistical tests must be run in order to determine the accuracy of the automated annotation system. In order to validate our results,  $t$ -tests will be run to determine that our accuracy is significantly better than random guesses.

After that determination, neural network configurations will be selected for each functional category. In order to test that the system is indeed accurate, known manual annotations will be split into four equal-sized groups. Training will restart using the first and second group for training data, the third group for validation, and the fourth group for testing. This test will be performed using all permutations of this data set and compare accuracy across each neural network configuration.

The usability of the system depends on the accuracy with which it functionally annotates sequences it has never experienced. To test this accuracy, approximately 100 maize genes of unknown function will be selected. Approximately half of these will be genes which are differentially regulated in the maize SAM. The other half will be randomly selected maize genes. These genes will be manually annotated by the biologists on the team. These genes will concurrently be automatically annotated using the ANN system and the two sets of results will be compared. This will give a true picture of the accuracy achieved by the system in the context of normal use.

To gather further information regarding the usefulness of the project, annotated sequences from the SAMs of other plants, such as *Arabidopsis thaliana*, may be gathered and tested with the neural network system in order to determine how biased the system is to the specific methods or species for annotation used for maize genes in this project.

The various features of the web portion of the project will be implemented using Python common gateway interface. MySQL will also be used for the storage and retrieval of relevant information. In order to perform the clustering analyses, we will use R to take advantage of the speed and efficiency of that statistical package. In order to interface with R, we will use the RPy package.

## 5 Significance and Evaluation

The overall goal of this project is to provide researchers with a tool which will allow for a quick method to determine the functionality of a gene. Through efficient determination of the functional category of genes, researchers will be able to more efficiently study genetically regulated biological processes. The proposed portion of the project is focused on making the result of the previous efforts available for others. Success of this project is based around the completion of the outlined goals.

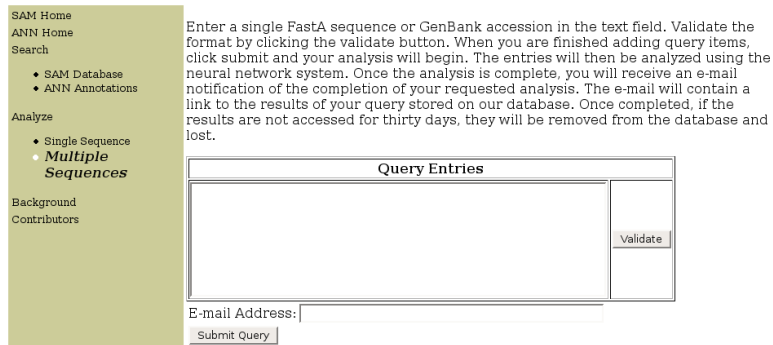


Figure 1: The website will be designed to allow queries which will be validated for accuracy. The resulting information will be stored on the server for future retrieval.

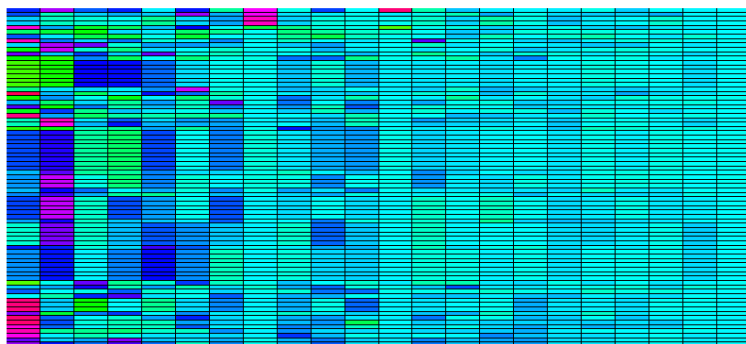


Figure 2: An example of clustering will be shown and clustered on the website. In this example, a row represents an accession and a column represents a functional category.

## 6 Conclusion

A user-friendly web interface will be designed and implemented to provide researchers with an automated method for annotation based on literature information. Although the system will be immediately designed for use with *Zea mays* SAM biology, there will be portions of the proposed work that will be easily expandable to fit into additional areas of biological inquiry.

## References

- [1] Miguel A. Andrade, Nigel P. Brown, Christophe Leroy, Sebastian Hoersch, Antoine de Daruvar, Christian Reich, Angelo Franchini, Javier Tamames, Alfonso Valencia, Christos Ouzounis, and Chris Sander. Automated genome sequence analysis and annotation. *Bioinformatics*, 15(5):391–412, 1999.
- [2] M. K. Barton and R. S. Poethig. Formation of the shoot apical meristem in *Arabidopsis thaliana*: an analysis of development in the wild type and in the shoot meristemless mutant. *Development*, 119(3):823–831, 1993.
- [3] Brent Buckner, Jon Beck, Katy Browning, Eneida Hoxha, Lisa Grantham, Zhian Kamvar, Ashley Lough, Olga Nikolova, Patrick Schnable, Michael Scanlon, and Diane Janick-Buckner. Involving undergraduates in the annotation and analysis of global gene expression studies: creation of a maize shoot apical meristem expression database. *Genetics*, 176(2), June 2007.
- [4] Val Curwen, Eduardo Eyras, T. Daniel Andrews, Laura Clarke, Emmanuel Mongin, Steven M.J. Searle, and Michele Clamp. The ensembl automatic gene annotation system. *Genome Research*, 14:942–950, 2004.
- [5] Steffen Hennig, Detlef Groth, and Hans Lehrach. Automated gene ontology annotation for anonymous sequence data. *Nucleic Acids Research*, 31(13):3712–3715, 2003.
- [6] John Moody. Prediction risk and architecture selection for neural networks. *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*, 1994.
- [7] Xiaolan Zhang, Shahinez Madi, Lisa Borsuk, Dan Nettleton, Robert J. Elshire, Brent Buckner, Diane Janick-Buckner, Jon Beck, Marja Timmermans, Patrick S. Schnable, and Michael J. Scanlon. Laser microdissection of narrow sheath mutant maize uncovers novel gene expression in the shoot apical meristem. *PLoS Genetics*, 3(6):e101, 2007.