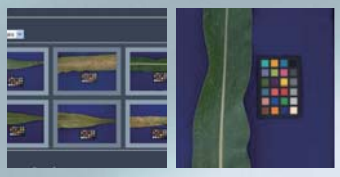
BioResource Now !
Issue Number 6 February 2010



Chi-Ren Shyu, Director Informatics Institute, University of Missouri

Hot News from Abroad No.28

Advanced Search Engine for Plant Phenotype Images

Reprinting and reduplication of any content of this newsletter is prohibited.
All the contents are protected by the Japanese copyright law and international regulations.

Download the PDF version of this newsletter at
http://www.shigen.nig.ac.jp/shigen/news/

Informatics Institute,
University of Missouri
Chi-Ren Shyu, Director

Advanced Search Engine for Plant Phenotype Images

Introduction

Phenotypes play a vital role in the majority of genomics research. Whether that be starting from a mutant phenotype and tracing the phenotype difference back to the cause at the genetic or epigenetic level, i.e. forward genetics; knocking out or over expressing a gene and examining the resultant phenotype to determine the biological function of that gene, i.e. reverse genetics; or just being able to look at the similarities and differences between groups of phenotypes, alleles, or germplasms, the ability to identify and quantify wild type and mutant phenotypes is of utmost importance to the scientific community. The difficulty in measuring a phenotype, however, is highly dependent on the phenotype itself. Many times, phenotypes are easy to identify or are directly measurable, e.g. survivorship, days until flowering, plant height. However, researchers are often interested in more complex traits that are challenging to measure, e.g. growth, disease resistance, subtle color changes. In these cases, domain experts will either make qualitative observations about a phenotype, in the form of descriptive narratives, or will assign quantitative values to a mutant phenotype, using their knowledge, perception, and memory in tandem with a scoring rubric. While these forms of phenotyping are useful, they are of limited accuracy due to the subjectivity and inconsistency of the human observer.

Capturing phenotypes in the form of digital imagery, which has become increasingly popular in recent years, presents an alternative approach to phenotype identification and quantification with the potential for more accurate phenotyping. If imaging is performed using a predefined protocol that facilitates the use of computer vision and image processing algorithms, images can be automatically processed so that interesting and distinguishing aspects of a phenotype are isolated and measured. With this technique, more difficult phenotypes can now be more accurately measured; for example, leaf growth can be assessed automatically simply by comparing the amount of plant matter present in “before” and “after” images.

In addition to the applications described above, imaging of phenotypes also gives rise to several other applications that can be useful to the genomic, biological, and agricultural communities. One class of applications involves making phenotype information, including both image content and any associated textual information, searchable with advanced query methods. We have developed such a retrieval system called VPhenoDBS for leaf phenotypes in *Zea mays*.

VPhenoDBS

The Visual Phenotype Database System (VPhenoDBS), which is located at http://PhenomicsWorld.org , contains two searchable datasets. The first is a set of leaf images of maize lesion mimic mutants taken using a fairly strict imaging protocol. Lesion mimic mutants are among the common mutant phenotypes in plants and are important in research because they may shed light on the mechanisms underlying apoptosis, or programmed cell death, as well as defense responses (Johal 2007). These mutations cause lesions to appear on the leaf in various colors, sizes, and shapes, and the expression of these mutations is highly dependent on genetic background as well as environmental conditions. Currently, this dataset contains about 730 images covering 16 lesion mimics. Figure 1 shows some examples of the variety of lesion mimics in this collection. Associated with this image collection are two advanced retrieval methods that allow for retrieval of images based on image content and semantics.

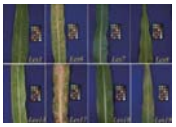


Figure 1: Examples of various les mutants

The first retrieval method available for lesion mimics is query by image example. This content-based image retrieval (CBIR) mechanism can be thought of as similar to a Google-type search, except that instead of typing the search query, the user submits an image of the phenotype of interest. The system then retrieves the most visually similar images from the database. To facilitate this retrieval, 176 phenotype measurements (Shyu 2007), including measurements related to lesion color, size, shape, and distribution, are automatically extracted from each image and organized into several high-dimensional indexing structures. The interface for the system is shown in Figure 2.

Figure 2: CBIR search interface for VPhenoDBS.

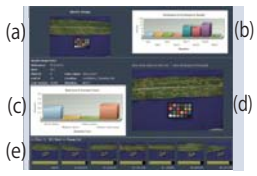


Figure 3: The CBIR results page for VPhenoDBS, which contains several areas of interest including the (a) query image, (b) a bar graph of the frequencies of various les mutants in the top results, (c) suggested semantic annotations for the current result image, (d) the current result, and (e) the top ranked results. The results page for the semantics search has a similar appearance.

Slider bars are available to give users the ability to emphasize various aspects of the phenotype as desired in the search. The results page, shown in Figure 3, displays the ranked results and any textual metadata known about those results. A bar chart is also provided (Figure 3b), which gives the frequency of each mutant within the top ranked results. If an unlabeled image were provided to the system, this graph might be used to indicate which mutants in the database are the most similar. In addition, the system can also perform automatic annotation of phenotype images with semantic terms commonly used by the biologists. All semantics with non-zero relevance to the image are displayed (Figure 3), along with their relevance values.

In addition to CBIR, VPhenoDBS also provides a semantic-based search mechanism (see Figure 4). An example semantic search could be “find maize phenotype images that have large necrotic regions on the leaves.” In this type of search, a user selects one or more semantics from a list of modeled terms. The system then searches for images that best match that semantics. It should be noted that these images do not have accompanying text; rather, the search is conducted by mapping semantic labels to image content itself using a mathematical model. The results screen for this retrieval method has a similar appearance to the CBIR results page.

Figure 4: Search interface for the semantic search in VPhenoDBS.

↳To the next page

Please refer to the following website for Figures(they have not been included here because of space constraints).
http://www.shigen.nig.ac.jp/shigen/news/n_letter/2010/newsletter_v6_n2En.html

The second dataset in VPhenoDBS is the set of all maize mutant images in MaizeGDB. These images were not taken using any defined protocol, and cannot be easily used in the type of searches just described. However, they do have accompanying annotations describing the phenotypes depicted. For this dataset, we provide a pure Google-style search of the image captions. The distinguishing feature of this retrieval method, however, is the inclusion of domain ontologies to improve the search. When a user submits a query, the system automatically links individual words or phrases from the query to the Gene Ontology and Plant Ontology. It then automatically performs query expansion by including synonyms, children, and parents of matched terms in the query, and these added terms are weighted as directed by the user. This type of query expansion has the potential to improve results in this type of text search. A screenshot of the search interface and results are shown in Figure 5. Terms that matched the Plant or Gene Ontology are color-coded for the user.

Figure 5: The text search page for VPhenoDBS. The query string is shown at the top of the page. Slider bars controlling weights of various types of terms and ontologies are shown in the middle, and the bottom contains the ranked results. The matched terms are color coded by type of term and ontology.

Making Your Phenotypes Searchable

The same techniques that were used to create VPhenoDBS and measure the characteristics of the lesion mimic mutants can be applied to nearly any visual phenotype. To do this, one must adhere to the following steps:

1. Determine a minimal standardized method for imaging your phenotype, such as background setting, color checker location, and camera/scanner selection.
2. Consult with computer vision and image processing researchers in our team to discuss the important aspects of the phenotypes as well as computer algorithms that can directly or indirectly measure those characteristics. Algorithm development could take anywhere from a few weeks to several months, depending on the features to be measured.
3. Photograph phenotypes and process images.
4. Build computer indexing structures to facilitate retrieval methods.

We are continually looking for collaborators wanting to utilize our tools for their phenotype image collections and assist us in our ongoing plant phenotype image research. If you are interested in making your phenotypes searchable, please contact us at ShyuC@missouri.edu.

Funding

This project is supported by the National Science Foundation grant number 0447794 and Shumaker Endowment for Bioinformatics.

Coauthors of This Article

Jason M. Green, Jaturon Harnsomburana, Thomza DeSouza, and Mary Schaeffer

References

[1] Johal, G. S. (2007) Disease Lesion Mimic Mutants of Maize. *APSnet*

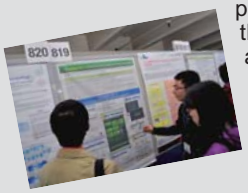
[2] Shyu, C. R., Green, J., Lun, D. P. K, Kazic, T., Schaeffer, M. and Coe, E. (2007) Image Analysis for Mapping Immeasurable Phenotypes in Maize, *IEEE Signal Processing Magazine*, Vol. 24, No. 3 , May 2007; 116-119

"Original article is written in English."

Participation in the Plant and Animal Genome (PAG) XVIII Conference

January 9-13, 2010 San Diego, California

It struck me during this year' s PAG that there was a significant increase of 2 – 3 times in the number of corporate booths compared to last year. Furthermore, the topic of NGS (Next Generation Sequencing) was brought up in many general poster presentations and I got the impression that the bioindustry is focused actively on the next-generation sequencing technology. One of the popular genomic browsers, GBrowse announced that it supports next generation sequencing data. Although it is now possible to produce nucleotide sequence data at a speed that was not possible before NGS technology, a company called LemnaTec has gone a step further than nucleotide sequences to introduce a device which uses RFID (Radio Frequency IDentification) and conveyor belt to automatically measure plant phenotype data at a high speed. I think that the speed of researches will accelerate due to the automation of experimental equipment and facilities and therefore how we handle the large amount of data will become more important in the future. (Shingo SAKANIWA)



I attended PAG after a 3 year absence and presented BRW(*1) on the third day of the PAG poster session. Currently, there are 4.5 million searchable resources in BRW and we provide several options for the search. During the poster presentation, I emphasized the architecture of the BRW search. I received many comments but most of the questions were about the resources in BRW. It was a good experience that I managed to answer those questions and also conveyed that basically these services are provided free of charge. Moreover, several people expressed interest in the technical implementation of BRW and I thought that it was good that I could explain about both the resources and also the technical side of BRW. Aside from this, the thriving corporate booths also leave a strong impression. They were mainly by exhibitors from the US but the number of Asian companies has increased compared to the last time I attended the conference. Interestingly, I thought this reflects the rapid progress that Asian has been making in recent years. (Tohru WATANABE)

(*1)BioResource World : a biological resources portal website published by the National Bioresource Project (NBRP)

I attended PAG after a 2 year absence and presented a poster about the major updates for Oryzabse. For the latest version 4, we plan to renew the website design and provide contents using Ajax and Flash technology. Among these features, I demonstrated the image gallery on my laptop and many people showed interest. In addition, I introduced Oryzabase at the "PLANT GENOME DATABASES Outreach CONSORTIUM" , a joint booth by 13 databases. Aside from being able to introduce Oryzabase, participating in this booth also gave me an opportunity to interact with other database developers. Although it was difficult for me to communicate in English, it was a very good experience to discuss technology and talk about the demands on database features. (Yuka TAKAHASHI)

"translated by Sharoh Yip"

BioResource Information

(NBRP) www.nbrp.jp/
(SHIGEN) www.shigen.nig.ac.jp/
(WGR) www.shigen.nig.ac.jp/wgr/
(JGR) www.shigen.nig.ac.jp/wgr/jgr/jgrUrlList.jsp

BioResource Now ! Issue Number 6 February 2010

Contact Address Center for Genetic Resource Information, National Institute of Genetics
1111 Yata, Mishima-shi, Shizuoka 411-8540, Japan
Tel.: 055-981-6885 (Yamazaki)
E-mail brnews@chanko.lab.nig.ac.jp

Editor's Note

It was a few years ago when I first met Dr. Chi-Ren Shyu and heard of his work on the phenotype image search engine at the group meeting of PAG. So I am very glad to learn of his success with "VPhenoDBS". Its interface is intuitive and search speed is comfortable. I am sure this will become a powerful tool for Plant Phenomics in the near future. I am also eager to apply the tool to various plant images of our databases. (Y.Y.)