Data Visualization and Analytical Tools at Many Levels of Plant Biology are Available at the Bio-Analytic Resource

Stephanie Anderson*

Managing Editor, Journal of Natural Product and Plant Resources, United Kingdom

***Corresponding Author:** Stephanie Anderson, Managing Editor, Journal of Natural Product and Plant Resources, UK*

E-Mail: plantresourceseclinicaljournal@gmail.com

ABSTRACT

Large data sets from roughly 15 distinct plant species can be accessed through the Bio-Analytic Resource for Plant Biology, with a focus on promoter, transcriptome, and protein-protein interaction information. It comprises of several databases with relevant information that its curators have added, data visualization tools to show the results of these databases' searches, and visual analytic tools to find, for example, interesting patterns of gene expression based on publically available data. We briefly discuss a few of these tools and the situations in which plant researchers might find them useful.

Keywords: *Arabidopsis thaliana,* Agriculture, Natural variation, Cis-regulatory elements, Transcriptomics, Protein-Protein interactions, Protein-DNA interactions

INTRODUCTION

The amount of data being gathered by researchers from almost every level of plant biology is unparalleled. Today, most journals and funding organizations demand that all data disclosed in published research be made freely accessible. Becausethey enable the generation and testing of hypotheses in silico prior to their validation in the lab, as a complement or alternative to conventional genetic screens or molecular techniques, these open data have revolutionized biology research workflows. Large data sets are often produced and maintained by the research labs that produce them. They are then placed, upon publication, in infrastructure-tier databases, like as those administered by the National Center for Biotechnology Information, NCBI, or the European Bioinformatics Institute, EBI. These archives make them simple to download, but manipulating, deciphering, and analyzing them requires some knowledge of bioinformatics and data science. Researchers can find answers to many issues more quickly with the aid of web-based data visualization tools than they would be able to if they had to collect and analyses the raw data themselves. The Bio-Analytic Resource for Plant Biology provides a variety of large data sets covering various levels of plant biology, along with data visualization tools and database web services for exploration. It compiles publicly available data from numerous sources and offers simple tools for accessing and analyzing them without the need to download anything or create custom programmes. The E-Plant Molecule Viewer, for example, paints Pfam domains, CDD, protein motifs, and polymorphisms producing non-synonymous alterations from the 1001 Proteomes directly onto Phyre2-predicted, tertiary structures. The CDD "DNA binding site" motif and surrounding non-synonymous alterations that were mapped onto the partial structure of the ABI3 transcription factor may have an impact on DNA binding in the ecotypes where those polymorphisms exist. The BAR has been accessible online since 2005 and saw 60,000 usages on average monthly in 2015. About 2300 publications that have been published have referenced one or more of our visual analysis tools. The following sections include descriptions of some of the BAR's more well-liked tools and how one may use them for their own study.

Web services and data

The BAR currently contains 145.2 million measurements of gene expression across tens of thousands of genes from several agriculturally significant plant species, the model plants Arabidopsis thaliana, and moss; documented subcellular localizations for more than 9300 proteins in Arabidopsis from the SUBA3 database; 70,944 predicted and 36,306 documented interactions between proteins in Arabidopsis; and 123,484 SNP polymorphisms across 96 Arabidopsis. Along with 885 experimentally discovered protein tertiary structures from the Protein Data Bank, protein tertiary structures make up about 84% of the Arabidopsis proteome. A large number of these data sets are reachable through online services that produce JSON-formatted output.

Analytical and data visualization tools

E-plant: By combining various data visualization techniques in a zoom able user interface, E-Plant enables researchers to see the relationships between DNA sequences, natural variation (polymorphisms), molecular structures, protein-protein interactions, and gene expression patterns. To download the most recent genome, interactome, and transcriptome data for any number of relevant genes or gene products, E-Plant connects to a number of publically accessible web services. A collection of visualization tools that are organized conceptually in a hierarchy from large to small are used to display data. Links between the various viewpoints highlight the relationships between various levels of study. It is based on a previous version of Fucile. E-Plant. And were created using a user-centered design approach and an agile software development process that includes multiple rounds of user testing. It was explained that the ePlant shows all the expression levels for all the samples of all the loaded genes. From here, visitors may choose any gene they want to look at and zoom in and out to view data at different scales, from kilometers to nanometers.

E-FP browsers

Arabidopsis: Using publicly available information from 64 articles and hundreds of experiments from the At Gene Express project, create 17 "electronic fluorescent pictographic" illustrations of a gene of interest's pattern of expression. These data sets contain roughly 75 million measures of expression. Some views, like the Lateral Root Initiation view, have received community support and developed into their own entities. Contributors work with the BAR curators to add new data sets when they are made available.

Additional E-FP browsers

To develop eFP Browsers for poplar, maize, soybean, tomato, potato, grape, Medicago, truncatula, Camellia sativa, rice, barley, triticale, the moss Physcomitrella patens, and most recently for peanut, we collaborated with various plant laboratories. Both the read map patterns and the expression level summaries are presented here in a conveniently sortable table. For the Araport 11 release, the current version uses more than 100GB of RNA-seq data to re-annotate the Arabidopsis genome.

Other techniques for protein and gene expression

Viewer for Arabidopsis interactions: A database of 70,944 predicted and 36,352 experimentally validated Arabidopsis protein-protein interactions is accessible through the BAR's Arabidopsis Interactions Viewer. Since its introduction in 2007, it has undergone constant development and updating. One of the most recent additions is the inclusion of almost 40,000 protein-DNA interactions, which were either discovered using the yeast one hybrid system or predicted using the mapping tool FIMO, along with hundreds of transcription factor binding specificities from JASPAR and Weirauch. Additionally, via PSICQUIC online services, it is now feasible to query additional protein-protein interaction databases, such as Bio GRID and Intact, using the AIV. A Rice Interactions Viewer is also available, allowing users to explore the 37,472 predicted and 430 experimentally validated rice interacting proteins.

Cistome

With the use of various existing Cis-element prediction systems, a new version called Promoter, or both, Cistome is a versatile tool that can forecast novel Cis-elements in the promoters of co-expressed genes. Alternatively, it can be used to map well-known Cis-elements from Weirauch or place jasper. Cistome can also be used to map transcription factor binding sites onto the promoters of a predetermined set of AGI IDs based on consensus sequences or position-specific score matrices. Sequence logos for the elements identified by the mapping are included in the result.

Genomic slider

Gene Slider presents a single long sequence logo that can be zoomed in and out of, from an overview of the entire sequence length down to a few residues, to assist in examining the conservation between aligned DNA or protein sequences. The Gene Slider version on the BAR loads and shows data for more than 90,000 conserved non-coding areas throughout the Brassicaceae indexed, Arabidopsis sequence, in addition to showing user-supplied FASTA alignment files. Additionally, it shows transcription factor binding sites from JASPAR and Weirauch, making it simple to spot regions that may be transcription factor binding sites and are conserved across several species.

Angler expression

By calculating the Pearson correlation coefficients for all gene expression vectors in comparison to an expression pattern defined with a graphical input tool, or to an expression pattern related to an AGI ID or gene name that is specified, this tool or an earlier legacy version identifies Arabidopsis genes with similar expression patterns. It generates a set of eFP images showing the data for each gene's expression that satisfies a correlational cut-off criterion or encompasses a predetermined number of hits.

CONCLUSION

It is concluded that, the BAR is an initiative that was created in collaboration and shows the effectiveness of a web service strategy for integrating and visualizing data from various sources. As a developer of visualization "apps" and a user of their web services for a number of BAR hosted products, it is deeply intertwined with Araport. More integration with other plant portals is being worked on;s eFP photos are already made available on SoyKB and the gene pages of Maize DB. A Genome Canada funding has financed more ePlants, allowing researchers to explore various levels of information about any gene of interest across numerous species. This study crosses disciplinary boundaries by fusing software engineering, user experience design, and best practises for data visualization. To complement our dedication to user-centered design, we have established an agile development methodology that incorporates numerous rounds of user testing.