Plant Co-expression Annotation Resource 2: a web tool for annotation inference in unknown function proteins related to abiotic stress in plants

Marcos José Andrade Viana[1,3,*], Adhemar Zerlotini[2], Maurício de Alvarenga Mudadu[2]

[1]Embrapa Milho e Sorgo, Sete Lagoas, MG, Brazil; [2]Embrapa Informática Agropecuária, Campinas, SP, Brazil; [3] Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais. * marcos.viana@embrapa.br

Introduction

The development of genetically modified (GM) crops includes the discovery of candidate genes through bioinformatics analysis, using genomic data, gene expression, among others. Proteins of unknown function (PUFs) are interesting targets for pipelines of GM crops due to the novelty associated and also to avoid copyright protections. One method to infer the possible function of PUFs is to relate them to factors of interest, such as abiotic stresses, using orthology and coexpression networks, applying the guilt by association approach.

Objective

The objective of this work is the discovery of PUFs involved in responses to abiotic stresses in plants for the development of genetically modified plants tolerant to climate change.

Methods

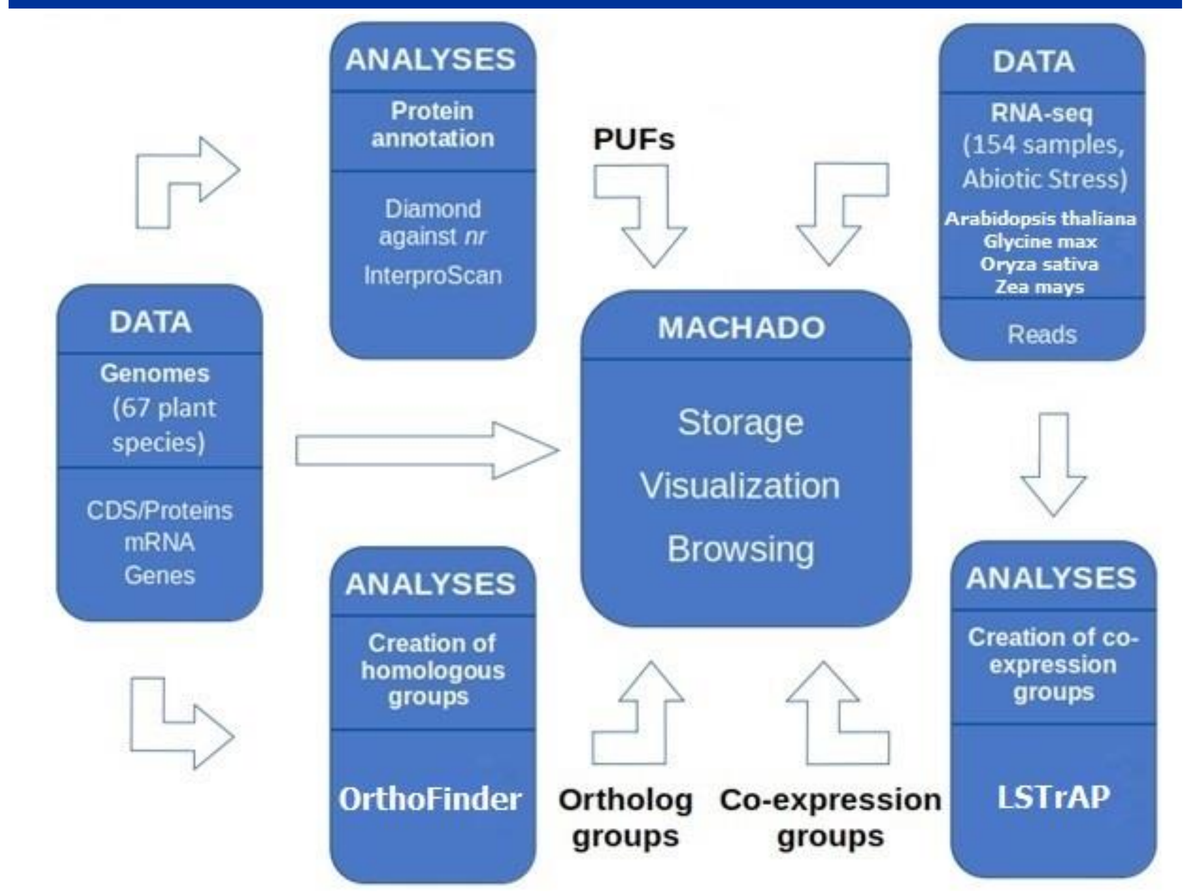

Figure 1 – Workflow Plantannot2

To run all the processing, analysis and data storage of this project, we used the computational infrastructure provided by Embrapa Informática Agropecuária's Laboratório Multiusuário de Bioinformática (LMB). All data were stored using python scripts made available https://github.com/lmb-embrapa/machado.The softwares and datasets (Figure 1) used to perform the analyses were:

Machado is a framework to store, visualize, and search biological data. It uses the Chado schema from GMOD - Generic Model Organism Database to store the data in a PosgreSQL database.

Downloaded 67 plants genome data (most of the Phytozome, 96%), including 5 species with important characteristics of tolerance to abiotic stresses (*Boea hygrométrica*, *Pearl millet, Oropetium Thomaeum, Populos simonii and Sorghum bicolor)*.

For all genomes downloaded, we look for similarities of sequence using Diamond BLAST and InterProScan looking for conserved domains on all proteins, to characterize the PUFs.

Identification of ortholog groups with OrthoFinder software.

Expression profile RNA-seq data for abiotic stress experiments on plants was downloaded from the NCBI GEO website.

Co-expression networks will be created using LSTrAP software with the transcriptome data downloaded.

Figure 2 shows the inferencing algorithm to PUFs annotation. Using the guilt-by-association approach, known function proteins that belong to ortholog groups can serve as a proxy for annotating PUFs (No match in Diamond and No match in Interproscan), as we only want PUFs related to abiotic stresses we only recover ortholog groups who have at least one protein whose mRNA belongs to a gene co-expression cluster.

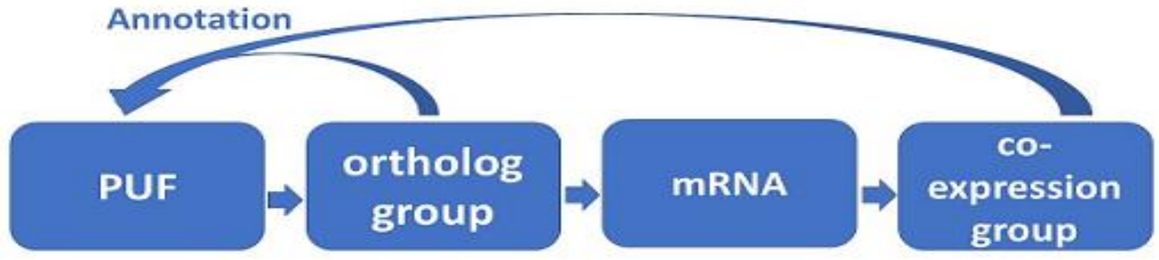

Figure 2 – Plant Co-expression Annotation Resource Algorithm

Results and Discussion

The "Plant Co-expression Annotation Resource 2" web interface was developed in python using Django framework to provide PUF queries and navigation. The resource provides analyzes such as comparative functional annotation research, expression values, biological paths and ontologies.

After analyzing and processing genomic data from 67 plants, we stored 2,136,336 genes and 2,714,161 mRNA in the Chado database, along with their translated proteins. We recovered 78,416 PFDs with Diamond and Interproscan analyzes, created 91,172 ortholog groups. With data from 154 RNA-seq samples related to various abiotic stresses for the organisms *Glycine max* (GMA), *Zea mays* (ZMA), *Arabidopsis thaliana* (ATH) and *Oryza sativa* (OSA) we created 1,975 co-expression clusters. Using our gold standard to search for PFD annotations, which is quite strict, and retrieves PUFs that belong to an ortholog group that also contains some proteins whose mRNA belongs to a co-expression cluster, we recovered 4,673 PFDs. We conducted a literature search on the proteins that belong to the orthologs groups, for all the PUFs that belong to the species Pearl millet, Populos simonii, Oropethium thomaeum e Boea hygrometrica, all known to be tolerant to abiotic stress (517 PUFs). We found studies related to abiotic stresses, on average, for 67.5% of PUFs. A webserver https://www.machado.cnptia.embrapa.br/plantannot2 is freely available and provides indexed queries of PUFs possibly associated with abiotic stresses.

In figure 3 we represent a search for PUF using the gold standard, but it is important to note that we can use other searches for PUFs related to responses to abiotic stresses mainly using the textual search provided by the tool, for example an interesting search would be to try to search all the proteins that matched in Diamond and did not match in Interposcan, which belong to an ortholog group and that at least one protein in that group belongs to a co-expression cluster and using the textual search for the word "unknown", that way we were able to recover 2447 PUFs with the little informative annotation "unknown".

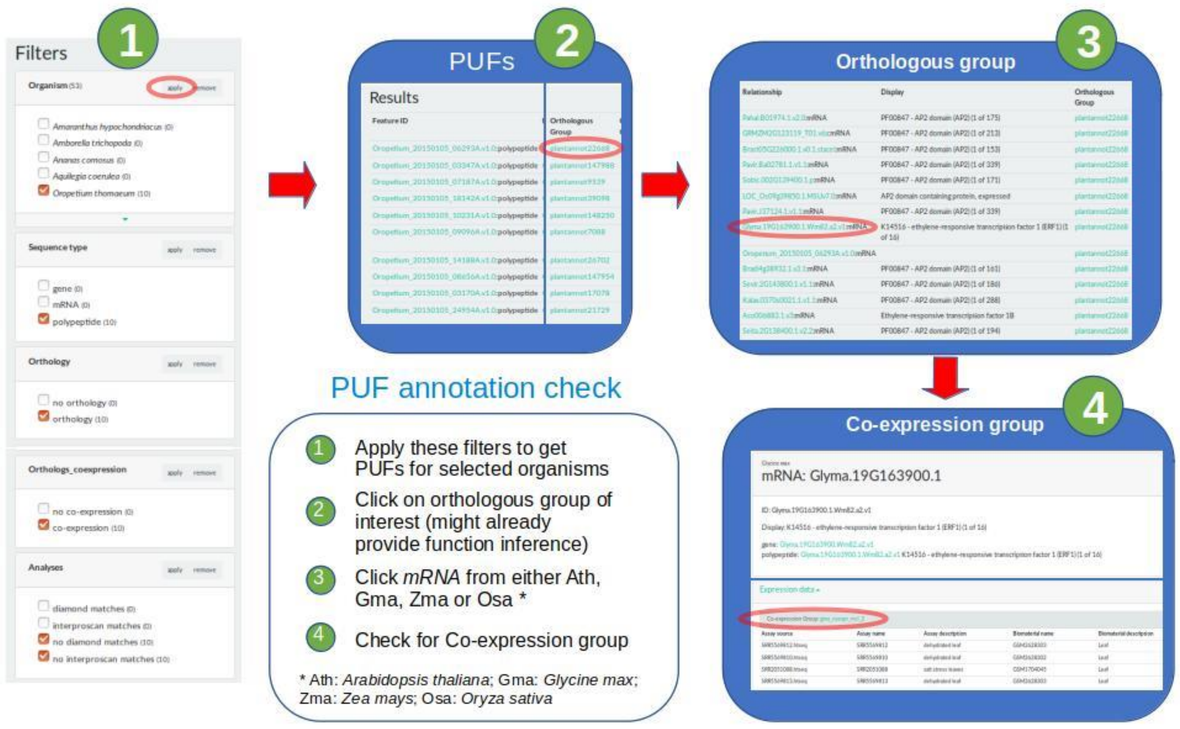

Figure 3 – Protocol to check PUF annotation.

Conclusion

We believe that our resource can be valuable in finding interesting targets to be used as proof of concept in breeding pipelines and thus contributing to increase the average productivity of national agriculture and the food supply for the Brazilian population. We also point out that the pipeline and the proposed web system have a much broader coverage than described, as they can support genomes of any kind of organisms and their transcripts in any situation. Plants and abiotic stresses were chosen as a "case study".