Bioinformatics approach for Transcriptomics: Integration of gene expression data based on RNA-seq with tissue localization

– Trabajo Fin de Grado –

Computer Science and Engineering

submitted by

Ignacio Marcos Serrano

Jaunary 2023

Supervisors

Iban Latasa Zudaire
Estibaliz Larrainzar Rodriguez

E.T.S de Ingeniería Industrial, Informática y de Telecomunicaciones UPNA

Abstract

The big challenge in current systems biology research comes when different types of data must be accessed from different sources, metrics and visualization and analysis tools. The challenge increases when it comes to genetics and transcriptomics, in which, for most experiments, gene nomination has its different versions, data normalization its different protocols and data compilation its different techniques.

Accordingly, there is a need for an analytic-visual tool for scientific support which works as a research platform that incorporates all compiled data and provides integrated search, analysis and visualization features within a single platform, providing the latest and best data vailable.

In this thesis, we present MtView; https://mtview.herokuapp.com/. A web tool for exploring multiple research levels of the leguminosae *Medicago truncatula* and its nitrogen-fixing symbiotic interaction with the bacterium *Sinorhizobium meliloti* through a simple user interface. The app connects with various public available web services containing gene transcriptome, genome, proteome, interactome, tissue localization and 3D molecular structure for around the 70.000 genes expressed in the aforementioned leguminosae.

Keywords

Bioinformatics
Transcriptomics
Genomics
Proteomics
Interactomics
RNA-seq
Symbiosis
Tissue localization
Medicago truncatula
Sinorhizobium meliloti
Normalization
Web development
Python
HTML
JavaScript
CSS
R

Contents

1 Introduction **3**

1.1 Premise . . . 3
1.2 State of the art . . . 4
1.3 Objectives . . . 6
1.4 Development proposal . . . 6

2 Biological basics **7**

2.1 *Medicago truncatula* . . . 7
2.2 Taxonomy . . . 8
2.3 Genomics . . . 8
2.3.1 Gene expression . . . 8
2.3.2 RNA-seq . . . 8
2.3.3 TMM Normalization . . . 8
2.4 Transcriptomics . . . 8
2.4.1 Basics . . . 8
2.4.2 Tissue localization . . . 8
2.5 Proteomics . . . 8
2.5.1 Basics . . . 8
2.5.2 3D molecule structure . . . 8
2.6 Nitrogen-fixing symbiosis . . . 8
2.6.1 *Sinorhizobium meliloti* . . . 8
2.6.2 Nodules . . . 8
2.6.3 KNO_3 treatment . . . 8

3 Analysis and design of the application **9**

3.1 Analysis . . . 9
3.2 Design . . . 9

4 Application development **10**

4.1 Taxonomy . . . 10
4.2 Expression value . . . 10
4.3 eFP . . . 10
4.4 Molecule viewer . . . 10
4.4.1 3D structure . . . 10
4.4.2 PAE . . . 10

5 Conclusions **11**

5.1 Performance analysis . . . 11
5.2 Application validation . . . 11
5.3 Future lines . . . 11

References **12**

1 Introduction

1.1 Premise

With the arrival of the digitalization and information era came the use of large-scale databases as a resource for biological research, tools for compiling all those data are becoming increasingly significant. Currently, several databases of Medicago gene expression data are acccessible, being one such example MtExpress [1], which collects data from both microarray and RNA-seq techniques.

Some data analysis tools have been developed to provide visual representation of that gene expression data, which is usually not easily visible to perform hypothesis generation. Still, biological researchers must go through multiple platforms to explore that data at different levels of analysis (Figure 1). Each platform with their own user interface, gene nomination versions and methods for collecting and categorizing information, such as its normalization protocols, among other things.

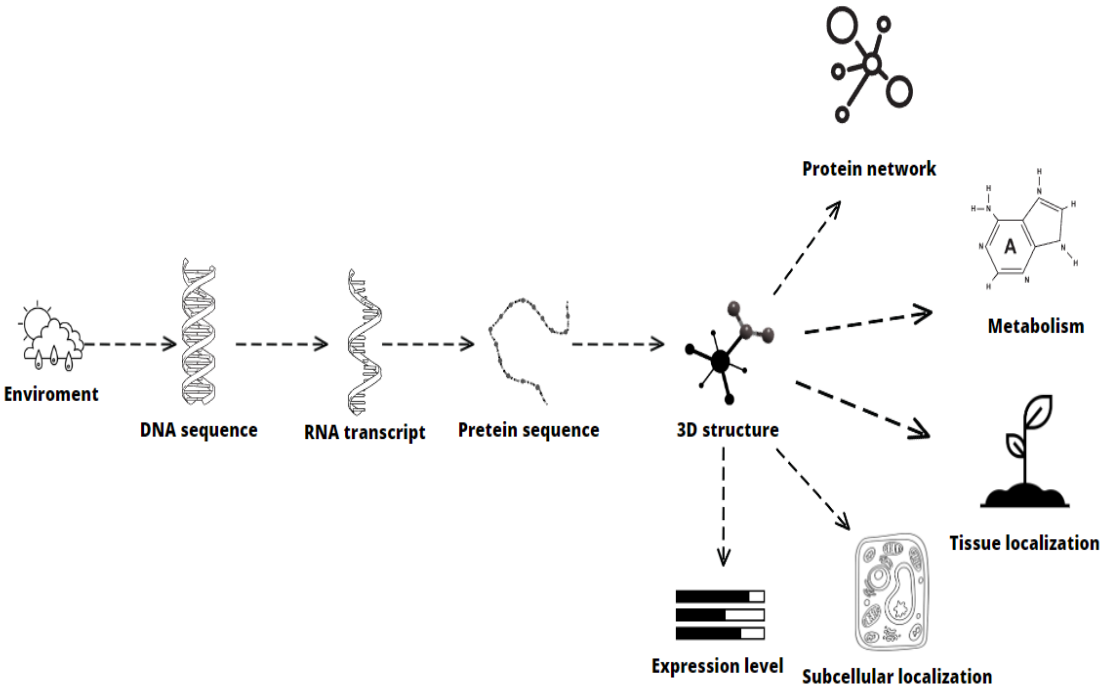


Figure 1: Some of the different levels of biological analysis.

For an expert, investigating a single gene would require going over all those different levels, from information about the annotation of that gene, its sequence, expression level, tissue location, subcellular location, protein structure, its protein-protein interactions and countless other levels. That means visiting an endless number of platforms with its own aforementioned peculiarities. This is a huge challenge when it comes to collect and interpret all those data for hypothesis generation, which leads to a significant decreased productivity.

These problems could be improved with the development of an integrated software platform. In this project we address this challenge by combining different data visualization and analysis tools within a single platform, compiling data from several sources and processing them into the same nomenclature, normalization and all of them gathered with the latest and reliable methods available. We believe that providing biological researchers an analytic-visual user-centred tool for exploring multiple levels of the plant data should improve their ability to extract information from that data for hypothesis generation to gain a deeper understanding of the biological processes they are researching of, wasting as little time as possible.

1.2 State of the art

Legumes are unique among cultivated plants for their ability to establish a nitrogen-fixing symbioses with rhizobial bacteria, this process promotes the development of nitrogen-fixing root nodules. The legume *Medicago truncatula* (*Mt*) is a primary model organism used for the study of legume biology mainly due to its small genome, which around the 94% of it have been sequenced by researchers.

Thereby, there is a good amount of studies related to th *Mt* genome, such as (C de Bang et. al.) [2] for root and nodule development, (Roux et. al.) [3] for symbiotic root nodules using laser-capture microdissection or (Larrainzar et. al.) [4] for interaction between nodulation factor and ethylene signals. However, despite the model importance, there are not many analytic-visual tools for *Mt* genetic research at its different levels of analysis.

One example could be the *Medicago eFP Browser*. This platform creates an Electronic Fluorescent Pictograph (eFP) representation of a gene (Figure 2). The expression data is based on the Benedito et. al. (2008) Gene Expression Atlas of the Model Legume *Medicago truncatula* [5].

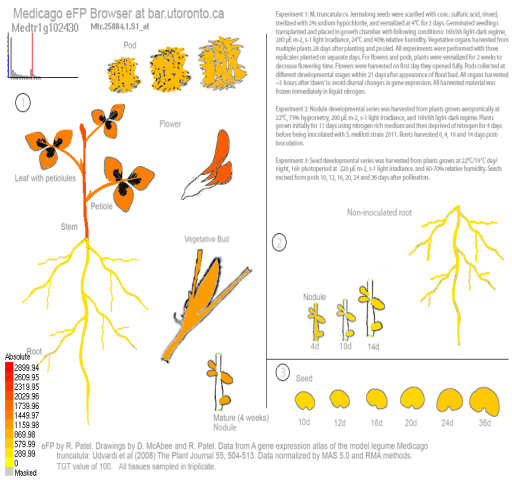


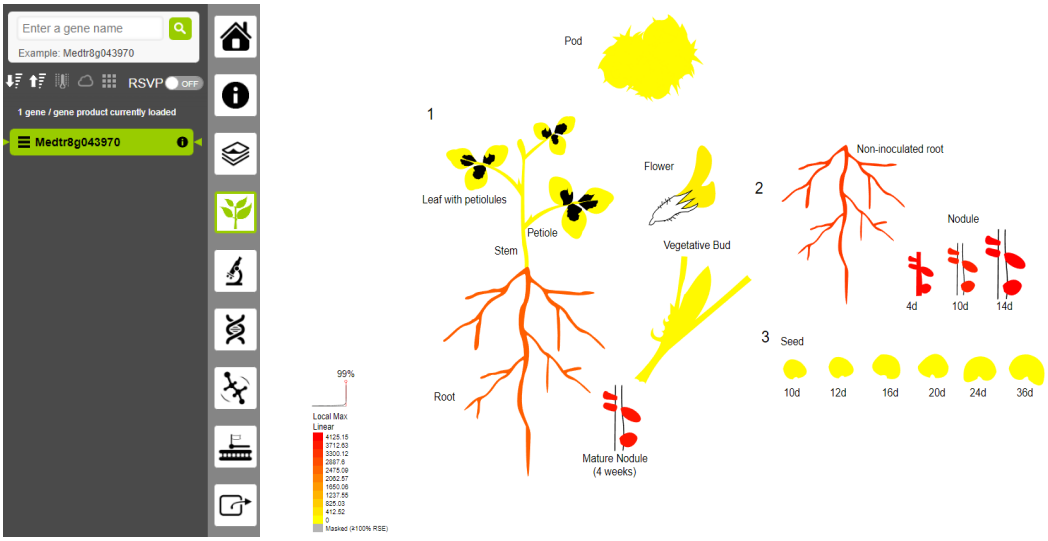
Figure 2: Medicago eFP Browser
http://bar.utoronto.ca/efpmedicago/cgi-bin/efpWeb.cgi

These gene expression values are based on hybridization-based microarrays, the prior method of of RNA-seq for measuring gene expression. The issues with microarrays are the cross-hybridization artifacts, poor quantification of lowly and highly expression genes and needing to previously know the sequence [6].

Due to this issues, transcriptomics were forced to evolve into sequencing-based methods, such as RNA-seq.

Another related platform could be *ePlant Medicago*. In addition of being a more well-looking platform, it connects to some publicly available web services to download the genome and transcripome data from the *Mt* genes, offering a few more tools (Figure 3a).

Despite the addition of those tools, that platform has the same issue of the aforementioned *Medicago eFP Browser*, the gene expression values are based on hybridization-based microarrays and gene nomination are only based on the outdated 4.0 version.



(a) Tool selector (b) eFP tool

Figure 3: eFP
http://bar.utoronto.ca/eplant_medicago/

A better approach for that platform is *ePlant* [7], based on the flowering plant *Arabidopsis thailiana* (*At*). Those data are downloaded connecting to several publicly available web services and displayed with a set of visualization tools through different analysis levels.

Therefore, our aim is to create an *ePlant*-inspired platform focused on the model legume *Medicago truncatula* to help biologist researchers visualize the connections between DNA sequences, polymorphisms, gene expression patterns, tissue localization, molecular structures and protein-protein interactions by combining different data visualization tools within a single platform.

1.3 Objectives

With all the above, the objective of this work wold be the creation of a web-based analitic-visual platform for biological research support, providing analysis and visualization tools for exploring multiple research levels of the model legume *Medicago truncatula*, which connects with several publicly available web services containing the latest genomics, transcriptomics, proteomics, interactomics, RNA-seq based gene expression and tissue localization data available.

1.4 Development proposal

To develop this platform we propose a collection of programs written in Python, Flask, JavaScript, HTML, CSS, jQuery and R. It is Python 3.8 compliant and runs within a web browser on most laptops, desktops and some tablets.

2 Biological basics

In this section we explain the basics of the biological approach of the project to gain a better understanding of the biological processes of which the project is focused in. First, we will introduce the model plant on which the project is focused. Then, we will describe a brief basics of each biological concepts concerning to the project, and its implications with the model plant. We will finally introduce the nitrogen-fixing bacterium with which the plant symbioses.

2.1 *Medicago truncatula*

Medicago truncatula (Figure 4) or barrelclover is a small legume native to the Mediterranean. It is a low-growing, clover-like plant of the Fabaceae familly. [8]

This model legume is famous for its ability to enter an important plant-microbe symbioses with soil microbes which promotes the development of nitrogen-fixing root nodules. This nitrogen-fixing symbiosis is caused by the bacterium *Sinorhizobium meliloti* (2.6.1), which, due to their symbiotic relationship, fixes atmospheric nitrogen into ammonia for the model legume while the plant provides it with sugar and proteins. This makes *M. truncatula* an important tool for studying these processes, leading, among other things, to the decrease of nitrogen use in plant fertilizers.

This species is also used as model organism for legume biology mainly due to its small genome, which around 94% of it has been already sequenced by researchers. Moreover, the model legume is self-fertile, is receptive to genetic transformation and has a rapid generation time. [9]

This legume is the cornerstone of the project, which will mainly focus on its gene expression based on RNA-seq.

Figure 4: *Medicago truncatula* https://jwp-nme.public.springernature.app/en/nmiddleeast/article/10.1038/nmiddleeast.2011.161

2.2 Taxonomy

2.3 Genomics

2.3.1 Gene expression

2.3.2 RNA-seq

2.3.3 TMM Normalization

2.4 Transcriptomics

2.4.1 Basics

2.4.2 Tissue localization

2.5 Proteomics

2.5.1 Basics

2.5.2 3D molecule structure

2.6 Nitrogen-fixing symbiosis

2.6.1 *Sinorhizobium meliloti*

2.6.2 Nodules

2.6.3 KNO_3 treatment

3 Analysis and design of the application

3.1 Analysis

3.2 Design

4 Application development

4.1 Taxonomy

4.2 Expression value

4.3 eFP

4.4 Molecule viewer

4.4.1 3D structure

4.4.2 PAE

5 Conclusions

5.1 Performance analysis

5.2 Application validation

5.3 Future lines

eFP for all MtExpress experiments

References

[1] Sebastien Carrere. "Plant and Cell Physiology". In: 62.9 (2021), pp. 1494–1500. URL: https://academic.oup.com/pcp/article/62/9/1494/6318817.

[2] Thomas C de Bang. "Genome-Wide Identification of Medicago Peptides Involved in Macronutrient Responses and Nodulation". In: 175.4 (2017), pp. 1669–1689. URL: https://pubmed.ncbi.nlm.nih.gov/29030416/.

[3] Brice Roux. "An integrated analysis of plant and bacterial gene expression in symbiotic root nodules using laser-capture microdissection coupled to RNA sequencing". In: 77.6 (2014), pp. 817–37. URL: https://pubmed.ncbi.nlm.nih.gov/24483147/.

[4] Estíbaliz Larrainzar. "Deep Sequencing of the Medicago truncatula Root Transcriptome Reveals a Massive and Early Interaction between Nodulation Factor and Ethylene Signals". In: 169.1 (2015), pp. 233–65. URL: https://pubmed.ncbi.nlm.nih.gov/26175514/.

[5] Vanger A Benedito. "A gene expression atlas of the model legume Medicago truncatula". In: 55.3 (2008), pp. 504–13. URL: https://pubmed.ncbi.nlm.nih.gov/18410479/.

[6] Kimberly R. Kukurba. "RNA Sequencing and Analysis". In: 11 (2015), pp. 951–969. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4863231/.

[7] Jaime Waese. "ePlant: Visualizing and Exploring Multiple Levels of Data for Hypothesis Generation in Plant Biology". In: 29.8 (2017), pp. 1806–1821. URL: https://academic.oup.com/plcell/article/29/8/1806/6100398?login=false.

[8] Helge Küster. "Brenner's Encyclopedia of Genetics (Second Edition)". In: (2013), pp. 335–337. URL: https://www.sciencedirect.com/science/article/pii/B9780123749840009153.

[9] Nevin D. Young. "The Medicago Genome Provides Insight into the Evolution of Rhizobial Symbioses". In: (2011), pp. 520–524. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3272368/.

Statutory Declaration:

I declare that I have developed and written the enclosed TFG completely by myself, and have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked. The TFG was not used in the same or in a similar version to achieve an academic grading or is being published elsewhere.

Pamplona, January 2023 Signature