PaIntDB: Visualizing Protein-Protein and Protein-Metabolite Interaction Networks in *Pseudomonas aeruginosa*

Javier J. Castillo-Arnemann (javiercastilloar@gmail.com)

1. Introduction

Pseudomonas aeruginosa is a multi-drug resistant pathogen involved in cystic fibrosis and other diseases. In the same way as all other pathogenic bacteria, the misuse and overuse of antibiotics has lead to resistant phenotypes that will eventually make antibiotics obsolete. These resistance mechanisms, like any other biological process, are the result of many genes and the complex interactions between them. Therefore, a systems-level approach, focused on group of genes instead of individual genes, is a powerful aid to identify new potential drug targets and elucidate how this resistance is achieved and how to combat it. Examples of this approach are differential expression experiments that test the change in expression for every gene under different conditions and return a list of genes that show statistically significant changes. However, these lists are often very long (>1000 genes for *Pseudomonas*) and therefore hard to interpret by themselves.

PaIntDB (**P**seudomonas **a**eruginosa **Int**eractions **D**ata**B**ase) contains 157,427 protein-protein and protein-metabolite interactions in *P. aeruginosa* strains PAO1 and PA14. It will be a web application where a researcher can upload a list of genes and generate a network showing the interactions between them. This allows the fast identification of co-expressed gene modules and other interesting structures that would be impossible to detect by just looking at the list. These networks can be used to obtain insight about important genes in any given process (not only antibiotic resistance) and generate new hypotheses for further experiments.

This project was started last summer by an undergraduate student who built the database and was continued by me starting this April. It will be the main component of my Master's thesis project. All of the basic functions and a basic GUI to generate the networks are already implemented and working properly, so one of the next steps is to build a visualization module.

2. Data and Tasks

2.1 Domain

The domain for this project is biology, specifically microbiology and systems biology.

2.2 Data

The data are the undirected networks generated by the application. Every network is static after being generated but is dynamic in the sense that you will always get a different network depending on the queried genes and other user-selected parameters. Most networks have between 500 and 1000 thousand nodes. Every node represents a protein or metabolite and edges represent their biophysical interaction in the cell.

PaIntDB currently creates 3 different network types with different attributes, described below. Table 1 specifies which attributes are included in each network type. There are other attributes associated with each individual node, (such as description, accession numbers, p-values for DE genes) but these do not need to be visually encoded and will be shown in a separate table view after selecting the node(s) of interest.

- **BioNetwork**: Basic network that includes the interactions between the queried genes but no experimental information.
- **DENetwork**: DE stands for Differential Expression. Includes all the information from a BioNetwork but has additional attributes for handling differential expression experimental data.

Network Type	Attributes
BioNetwork	- Location - Type - NodeDegree
DENetwork	- Log2FoldChange
CombinedNetwork	- SourceofInterest

Table 1: Attributes corresponding to each network type. They are cumulative, so all the attributes in a BioNetwork are included in a DENetwork, and the same with a DENetwork and a CombinedNetwork.

- **CombinedNetwork**: TnSeq (Transposon Sequencing) is another recent high-throughput technique to identify genes of interest under certain experimental conditions in bacteria. Combined networks include all the information from a DENetwork but have an additional attribute indicating the experiment where the gene was identified.

2.2.3 Data Attributes

Categorical

- Localization:
  - Location of the expressed protein in the cell (cytoplasm, membrane, etc.).
  - 12 levels.
- Type:
  - Indicates if a node is a protein or a metabolite.
  - 2 levels.
- SourceOfInterest:
  - Indicates if the gene was identified through RNASeq (differential expression), TnSeq, or both.
  - 3 levels.

Ordered

- NodeDegree:
  - Number of direct connections to other nodes.
  - Quantitative, sequential.
  - Range depends on the network, but usually 1 to 100.
- Log2FoldChange:
  - Statistic that estimates the change in gene expression between conditions.
  - Quantitative, divergent.
  - Range depends on the input experimental data, but usually -4 to 4.

2.3 Tasks

The tool is designed to allow biologists without a computational background to explore the results of high-throughput experiments and identify interesting groups of genes to generate hypotheses for new experiments. This high-level task can be broken down and abstracted as follows:

- Analyze, consume, discover: The user will explore the networks to find interesting network regions and nodes, and generate hypotheses about the experimental conditions and how they effect the biological process of interest.
- Search, locate: The user can find a specific gene of interest in the network and see its interactions,.
- Search, browse: The user can find genes that are up-regulated, down-regulated, located in a specific part of the cell, etc.

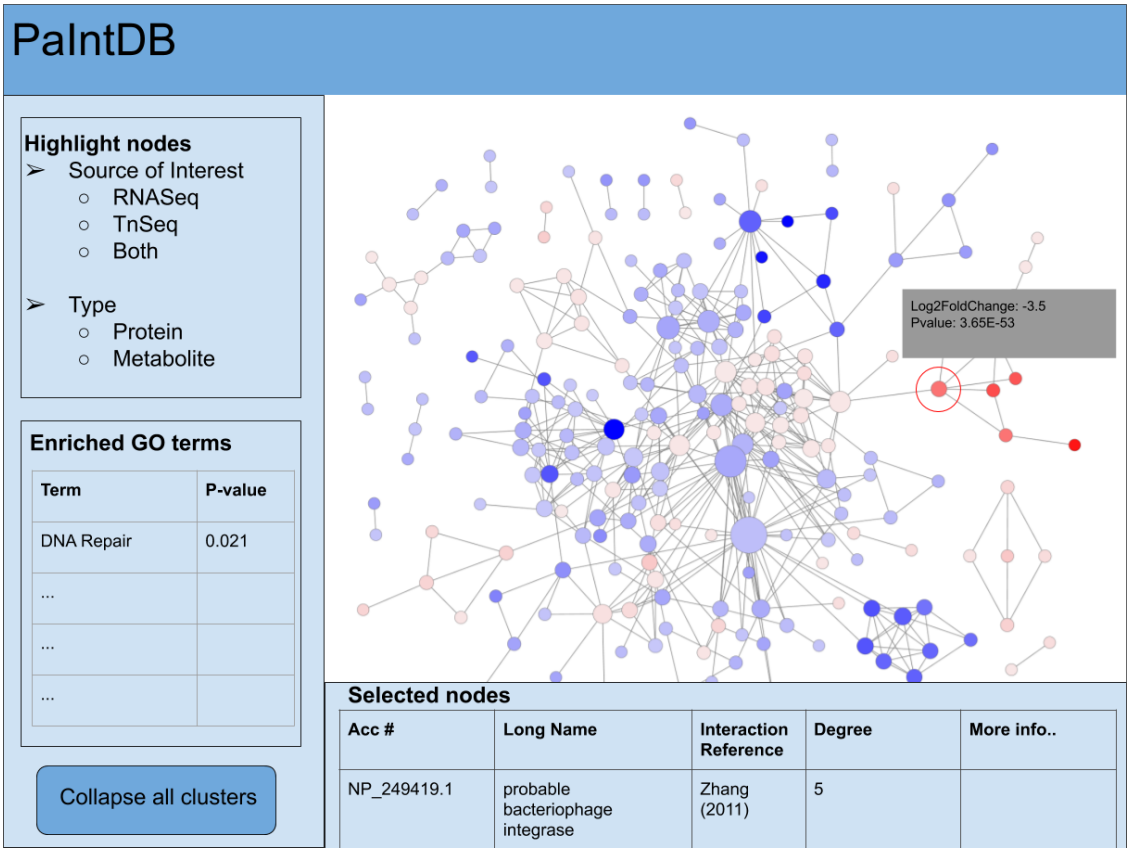

Figure 1: PaIntDB visualization module mock-up. The red circle represents a highlighted node.

- Search, explore: The user can look at the whole network to find interesting groups of genes, either by expression or by network topology.
- Query, identify: The user can select gene(s) of interest and query the database to get all the information on that protein or metabolite, including the reference for the interaction, links to other databases, etc.

3. Proposed solution

I have been using Cytoscape to visualize and explore the networks generated by PaIntDB, so I have experimented with different idioms and encodings for the attributes. So far, this is what I have:

- Map node size to node degree, to identify highly connected nodes.
- Map hue and saturation to Log2FoldChange in DEnetworks and CombinedNetworks to identify up- and down-regulated genes.
- Use luminance to interactively highlight nodes based on SourceOfInterest or other attributes.
- If time permits, implement a layout based on cell location, something similar to Cerebral[1].

In addition to this, the tool will have many interactive features, including geometric zooming and node highlighting based on node attributes to facilitate network exploration. As the networks get bigger, the hairball problem appears, so PaIntDB will also include a clustering algorithm based on network topology, to reduce the number of nodes and expand/collapse them when the user clicks them. As mentioned above, when clicking on an individual node, the user will get a table view with all of the node's information from the database. When hovering over a node, a pop-up table will show some relevant data (p-values for differentially expressed genes, for example).

Milestone	Time (h)	Deadline	Description
Pitch	5	Oct. 8	Make slides, rehearse
Proposal	10	Nov. 4	Write proposal, summarize previous work, create mock-up.
Viz module skeleton	5	Nov. 7	Make non-functional GUI with mock-up as guide.
Import networks 1	5	Nov. 11	Write script to generate Cytoscape Dash networks out of Networkx objects.
Import networks 2	10	Nov. 15	Explore and use included layouts and polish exploratory functions (zooming, panning, etc.)
Design stylesheets	5	Nov. 15	Create stylesheets to visually encode network attributes.
Peer Review	5	Nov. 19	Prepare slides and polish prototype.
Left Pane Func.	10	Nov. 25	Add highlighting and selection functions.
Bottom Pane Func.	5	Nov. 29	Query database to show information on the selected nodes.
Clustering	10	Dec. 5	Implement network topology clustering.
Custom layouts (Maybe)	10		Explore and implement custom network layouts.
Final Presentation	10	Dec. 10	Prepare presentation, finish prototype and practice demo.
Final Report	15	Dec. 13	Finish writing final report.

Table 2: Milestones schedule.

The Gene Ontology (GO) initiative aims to assign functional information to every gene, called GO terms. GO term enrichment is an approach to find statistically overrepresented terms in a list of genes to characterize their function and help interpret large gene lists. PaIntDB will also perform this test on the networks, and have the option to highlight and select the nodes associated with the enriched terms.

3.1 Usage Scenario

A microbiologist has a dataset of differentially expressed genes for *P. aeruginosa* treated with antimicrobial peptide DJK5. After uploading her list of genes with associated expression values, PaIntDB generates a network and she sees a red cluster of down-regulated genes to the right (Fig. 1), and clicks it to see what they are. With the information provided by the database, she finds that all of those genes are bacteriophage Pf1 genes. With this in mind, she does a literature review and finds that Pf1 genes have been linked to increased virulence and chronic infections, so now she can devise a new experiment knocking out or over-expressing these genes to further investigate their biological role and see if the resistant phenotype changes.

4. Implementation

At this time, all of PaIntDB is implemented in Python. The back-end uses the sqlite3 library to query the database, pandas to handle the queried data, and networkx to create and manage the network objects. The GUI to generate the networks was made using Dash, a library to build interactive web applications in Python that abstracts all of the JavaScript, React.js and HTML code underneath. For the visualization module, I will use the Dash Cytoscape library, which was also developed recently and again, abstracts the JavaScript

code into easy-to-use Python components. However, I will probably have to modify the underlying JavaScript code to add custom functionality needed for PaIntDB.

5. Milestones and Schedule

Table 2 describes the project milestones, their estimated time commitment and the respective deadlines.

6. Previous Work

PaIntDB was inspired by similar tools developed in the Hancock Lab for systems-levels analyses: NetworkAnalyst[2] and InnateDB[3]. These tools have a much larger scope than PaIntDB, since they support multiple organisms and have many more features. However, this leads to a higher learning curve and feature overload. Because PaintDB has a more specific and focused objective, the goal is to develop an intuitive tool that any *Pseudomonas* researcher can pick up and use with minimal effort.

References

1: Barsky, A., Gardy, J. L., Hancock, R. E. W., & Munzner, T. (2007). Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. Bioinformatics, 23(8), p. 1040–1042.

2: Zhou, G., Soufan, O., Ewald, J., Hancock, R. E. W., Basu, N., & Xia, J. (2019). NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic Acids Research, 47(W1). p. 234-241.

3: Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., Lo, R., ... Lynn, D. J. (2012). InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. Nucleic Acids Research, 41(D1). p. 1228-1233.