PROTEIN-PROTEIN INTERACTIONS IN BACTERIA

Shwetha Hara Sridhar, Master of Science in Bioinformatics - December 2017

Virginia Commonwealth University Department of Life Sciences

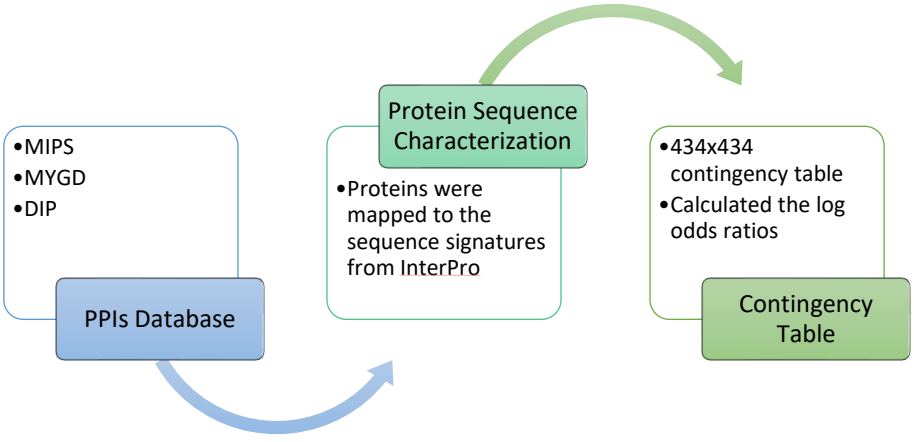
Center for Study of Biological Complexity, Richmond, Virginia

INTRODUCTION

Proteins control all biological systems in a cell, and while many proteins perform their functions independently, the vast majority of proteins interact with others for proper biological activity. [1] Identifying protein–protein interactions by analysis is critical to understand protein function and the biology of the cell and can serve as evidence for good research. Previous research focused on protein interactions found in *Yeast Saccharomyces cerevisiae.* Correlated sequence signatures were identified in the interacting proteins and conclusions were drawn to direct experimental interaction screens. [2]

Interpro is a new database that has an integrated documentation for protein families, domains and functional sites, it unifies databases and creates a nonredundant database from Prosite, Pfam, Prints and Prodom. It also includes all the sequence signature characterization for all Swissprot and Trembl sequences. [2]

Figure 1.1: Steps taken for the Data Analysis in the Previous Studies



A contingency table of sequence-signatures (signatures by signature) was constructed based on the results of characterizing the protein sequences by their sequence-signatures based on InterPro classification. These proteins were mapped to the sequence signatures from InterPro database. The previous study used around 3502 different Interpro signatures and 2908 proteins were mapped to the database from the nonredundant database made from the three other databases, the number of PPIs reduced to 1274 pairs. In order to prevent bias, duplicates were removed and the count resulted in 434 signature pairs combinations. Therefore, constructed a 434x434 contingency table with 2286 entries in the table in which there were 1433 entries that were zero.

Identified over-represented sequence signature pairs by comparing their observed frequency to those expected at random. The calculated the log odds ratios for each cell with observed frequency by the expected frequency ranged between -1.95 and 12.16. And the information content was found to be 2.48bits in this case. Therefore, it can be said that the information about a characteristic sequence-signature in one protein reduces the uncertainty about its potential interacting partners in 2.48 bits of information on average. These log odds values were classified based on a given threshold; >=2 or >= 5, where around 40 sequence-signatures found within this threshold. 1141 lesser or equal to 2 as their log odds, therefore were removed from consideration in threshold.

From Interpro, the parent (super-families) and the daughter (sub families) were identified for clustering. Finally resulting in 185 signature clusters. By using correlated signatures, the search space of putative protein pairs was narrowed down and the probability to detect the interaction increased in such pairs. This approach helped exclude one pair at a time from dataset and reconstruct the contingency table and calculate the log odds values. The excluded protein pair was predicted to be interacting if the log odds was above a given threshold, enabling an estimation of sensitivity (number of correct predictions from unknown interactions). However, sensitivity can only be confirmed after several experimentation and validation. Thus, this approach is useful only because a comparison and removal of the ones that surely

would not give us expected results (unguided search) can be done. Finally, this method resulted in 185 are good for experimental testing which can be applied for genome interest for genome analysis and finding regulatory sequences.

The present study focuses on protein-protein interactions in bacterial species. The data was obtained from BIOGRID and INTACT. Python was used for the analysis, parsing and clustering. A reference database was built using the bacterial Swissprot data from the Uniprot cross-reference database.

METHODS

Parsing and Mapping the Data

The data retrieved from BIOGRID and IntACT, was cleaned and parsed for the interactor genes identifiers. BIOGRID data had BIOGRID identifiers whereas the IntACT data had Uniprot identifiers. Both the identifiers were parsed in pairs (protein-protein interactions) and were mapped from BIOGRID and Uniprot databases respectively to the Uniport database. The mapped proteins in uniport were saved in the text format with selective columns (Entry, Entry name, Protein names, Organism, Cross-reference (Pfam) and Cross-reference (BioGrid or IntACT)). In this mapping process, the data was reduced since the unmapped proteins were filtered out of the data.

Figure 2.1: Flowchart describing the steps taken to parse BIOGRID data and combine the PPIs

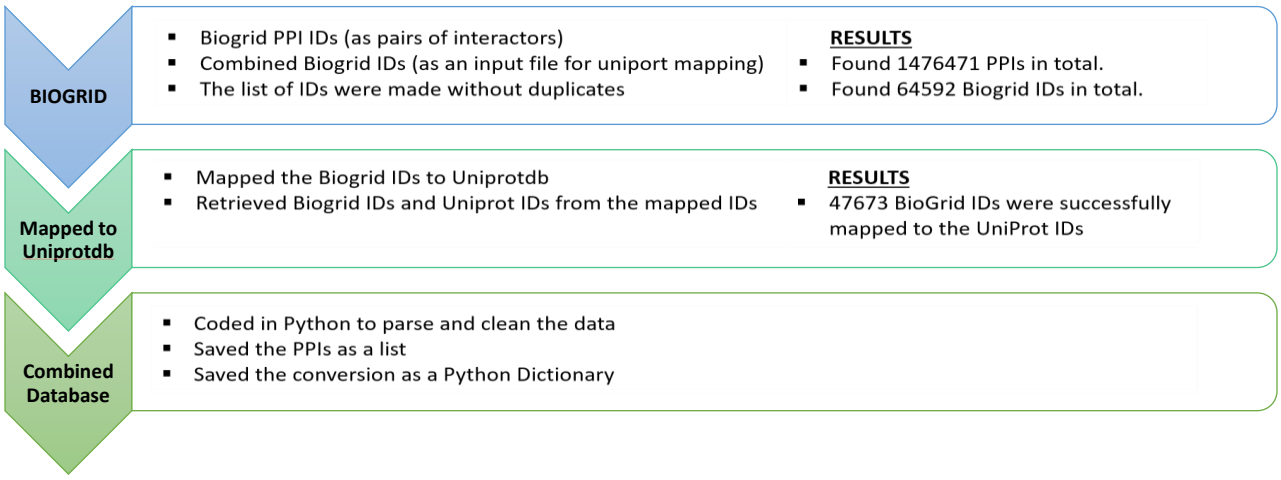
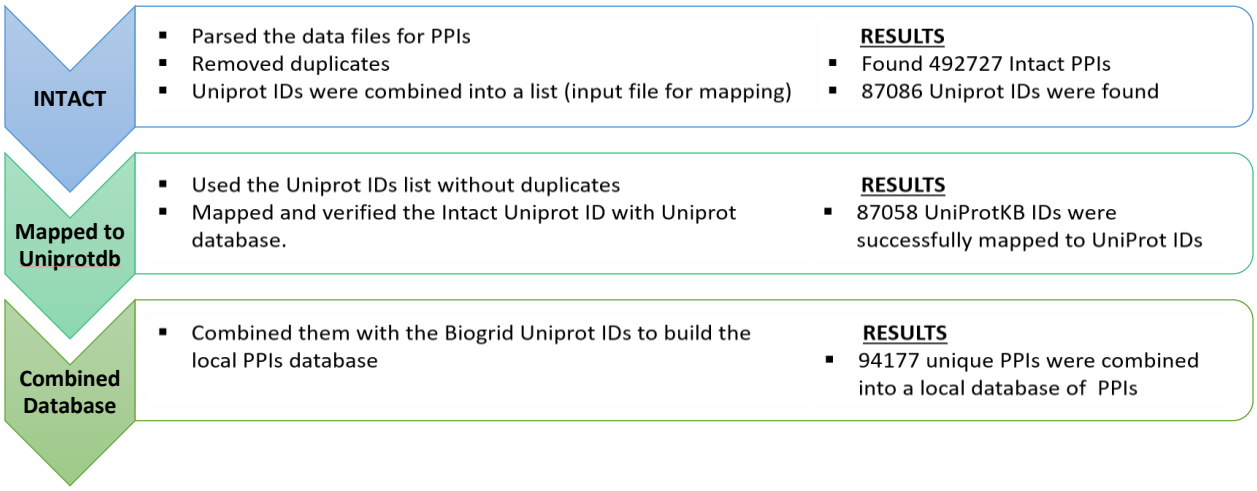


Figure 2.2: Flowchart describing the steps taken to parse IntACT data and combine the PPIs



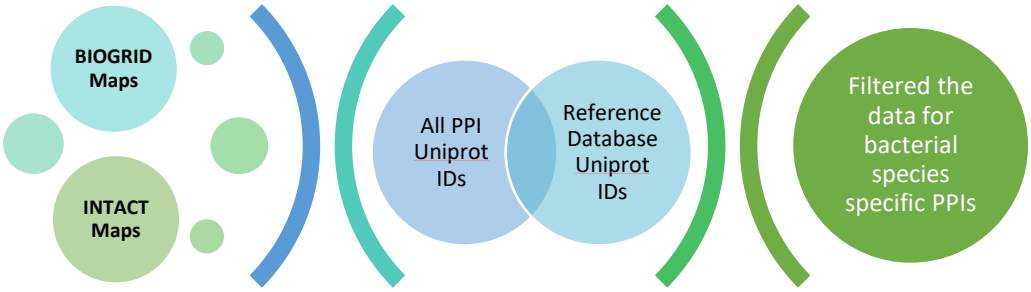
Listing Uniprot Interactions

The mapping files were used to convert the Protein-Protein Interactions with BIOGRID identifiers into Uniprot identifiers. A file was made with the Biogrid IDs and their respective Uniprot IDs, which was used to identify the PPIs with their uniprot identifiers. Similarly, since IntACT had uniprot identifiers already, we only filtered out the unmapped protein IDs. An output file was written with all the Uniprot protein-protein interactions (from IntACT and BIOGRID).

Filtering for Bacterial Protein Interactions

The bacterial protein data was obtained from the Swissprot database that was crosslinked with Uniprot database. This data was bacterial species specific and is ideal to be used as a reference database. Since the protein-protein interactions found so far were not organism specific and contains all the organisms in general, we could filter for bacterial species-specific protein-protein interactions using the reference Swissprot database. The reference database is a list of Uniprot identifiers. The interactions were filtered by checking if both of the interactor IDs were present in the reference database. The output file contained the Uniprot IDs of interactor A and interactor B that were present in bacterial species.

Figure 2.1: Flowchart describing the steps taken to filter the bacterial protein interactions



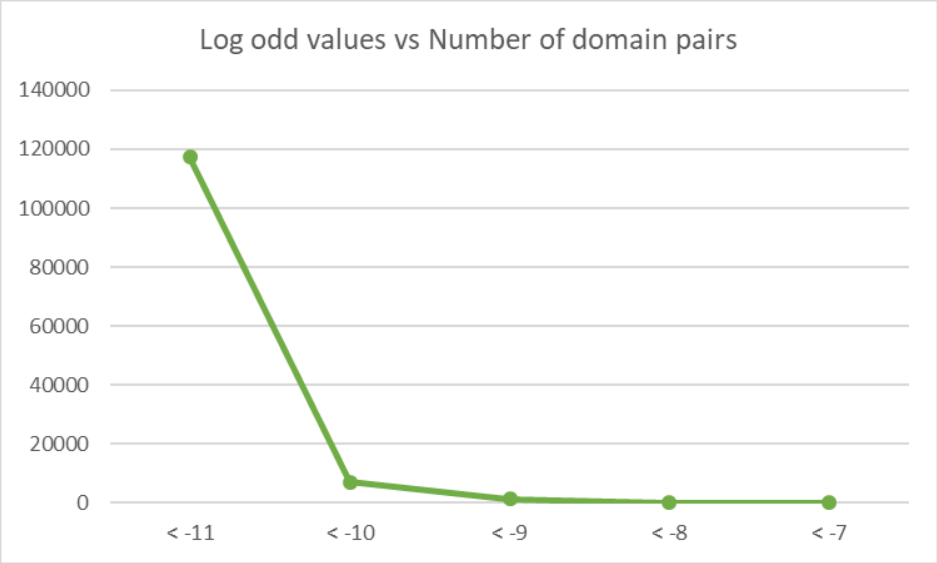
Contingency Table

These PPIs were further converted into domain-specific PPIs using their Pfam identifiers. This narrowed the interactions further. A contingency table was drawn with the domain-specific interactions with the frequencies entered into the table cells. Their log odd values were calculated and the zero values were excluded. A threshold was set to xxxx and the values that fell into the threshold was considered for further interpretations.

RESULTS

Mapping the data to the Uniprot database resulted in 47673 out of 64592 BioGrid identifiers being successfully mapped to 47569 UniProtKB IDs and 87058 out of 87086 UniProtKB AC/ID identifiers being successfully mapped to 58955 UniProtKB IDs. The combined list of Uniprot protein interactions found were 1927842 in number. After filtering for Bacterial Protein Interactions based on the protein IDs obtained from the Swissprot database, we found 192024 protein interactions in total. These protein interactions were converted into domain pairs using their Pfam IDs retrieved from the Uniprot mapping results and we found 171879 domain interactions in total. The contingency table resulted in 125952 nonzero values and no zero values.

Figure 3.1, Frequency of the log odd values found in the domain interactions



From figure 3.1, the highest count was found to be 75 for domain pairs PF07690, whose family was found to be Major Facilitator Superfamily, and PF00873, whose family was found to be AcrB/AcrD/AcrF family. The log odds value for this domain pair was -7.43.

ACKNOWLEDGEMENTS

This research was supported by Dr. Peter Uetz, Associate Professor at Department of Life Sciences, Center for Study of Biological Complexities - Bioinformatics. I would like to thank Dr. Uetz, for his help in providing an insight that greatly assisted the research. I would also like to acknowledge Dr. Norman Goodacre, for his patience in checking my scripts and for helping me improvise the programs to execute faster, so that they give me an efficient output. Dr. Uetz and Dr. Goodacre have been the major support to guide me in interpreting and understand parts of the results that were unknown to me in this study.

I would also like to acknowledge Dr. Allison Johnson, for keeping an eye on my progress. The push and encouragement that she has given me, has helped me stay focused and motivated throughout my master's program.

Last but not the least, I would like to thank my family for supporting me financially, for being there when I needed emotional support and for believing that I could achieve anything.

REFERENCE

[1]. Harper JW, Adami GR, Wei N, Keyomarsi K, Elledge SJ. 1993. The p21 Cdkinteracting protein Cip1 is a potent inhibitor of G1 cyclin-dependent kinases. Cell 75: 805–816.

[2]. Mapping Protein–Protein Interactions Using Yeast Two-Hybrid Assays, Jitender Mehla, J. Harry Caufield, and Peter Uetz Cold Spring Harb Protoc; 2015; doi:10.1101/pdb.prot086157