**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**

**Importance of Protein-Protein Interactions**

Several cellular processes are understood better by studying the system and the various components it is made of. [1] Protein-protein interactions play a major role in understanding how the system functions and how their components are interlinked. [1] Networks of protein–protein interactions (PPI networks) quantify the functioning and relative importance of particular proteins in cell function. [1] Network analysis can be used to infer novel functions, to quantify positional importance and to support predictions in pathogenesis studies. [1] The systems perspective and network analysis can be of particular importance in studying diseases with complex molecular processes. [1]

Protein-Protein Interactions can be defined as specific physical contact between two proteins, characterized by a molecular docking between them that serves a purpose entirely different from the generic functions of these biomolecules. [2] Computational studies have facilitated the analysis of interactions, revealing that they are more frequent among adjacent proteins in metabolic pathways and that regulatory proteins potentially involve a high proportion of interactions. [3]

**Computer Models**

Computers can serve as virtual laboratories that advance biomedical knowledge in areas such as infectious disease spread and drug interactions in the body. With virtual labs, scientists can perform experiments that are difficult to do in actual labs. [6] They also can quickly identify factors that are important to include in lab-based experiments. Researchers use computer simulations to track biological processes in cells and research organisms. This allows them to computationally test, for example, the possible effects of drugs on those processes. [6]

The drugs that seem the most promising can then be studied in living cells or organisms. These computer models generate millions of different possible outcomes. No single set of results or single computer model can accurately predict an outcome. Therefore, researchers often ask the same questions using different models. When multiple models yield similar results, scientists have more confidence in the predictions. [6] But, computer models have limits. They have been created based on what they already know about a process or disease. Therefore, information and data gained from real-world experiments are used to enhance computer models and predictions to help design additional experiments. Thus, computer modeling and lab experiments go hand in hand—both are needed to advance our understanding of health. [6]

**Previous Studies on *Yeast Saccharomyces cerevisiae***

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. [4] 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.

The 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. [5] 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. [4] 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. [4] The over-represented sequence signature pairs by comparing their observed frequency to those expected at random were identified. The log odds ratios ranged between -1.95 and 12.16 with the information content of 2.48bits. Therefore, it was concluded 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. [4] With a given threshold; >=2 or >= 5, where around 40 sequence-signatures whose log odds were found.

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

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. [4] The excluded protein pair was predicted to be interacting if the log-odds value 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. [4] 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. [4]

**Bacteria as the Research Organism**

In this study we use bacteria as our research organism. Bacteria is a commonly used research organism and there is vast pool of information already present in a number of databases today that serve as a reference, making it easy for the scientists to experiment and expect the predicted results. Therefore, bacteria can serve as a proxy for understanding the biology of humans. [6] Many aspects of these organisms’ biology are similar to ours, and much is already known about their genetic makeup, therefore, studying bacteria helps us learn more about human health. [6] For example, the natural course of a disease in humans may take dozens of years, whereas, a model organism such as bacteria can quickly develop a disease or some of its symptoms. [6] That allows researchers to learn about the disease in a much shorter time. Therefore, allowing the scientists find out what that gene does in a model organism linking the genes to the human disease. [6] This information can provide important clues about what causes a disease.

Studying bacteria can also help researchers develop potential diagnostic tests and treatments that are later tested in clinical trials. [6] For example, research with bacteria has revealed how all living things pass on their genes to offspring. This work detailed the ways cells copy DNA and repair mistakes made during the copying process. [6] Studies with bacteria have identified the basic components of circadian clocks, which drive daily biological rhythms. The research revealed connections between these clocks and sleep deprivation, obesity, diabetes, depression, and other human health conditions. [6]

The data for present study 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.

**The Sprinzak and Margalit Algorithm**

Experimentally deciphering the complete interactomes of each genome that has been sequenced becomes a daunting task. Several computational methods have been developed for PPI prediction. Insights gained through these predictions could potentially be used in the design of experimental PPI studies. It can also be used to predict pathways in the cell, identify novel disease genes and potential targets for therapeutic interventions. The interactions with known proteins will also aid in predicting the functions of novel proteins.

Computational methods of PPI prediction have employed statistical analysis and machine learning techniques, utilizing sequence, domain, 3D-structure, and evolutionary information. Sequence based methods have incorporated gene fusion information (Rosetta stone approach) [10], gene order [11] and phylogenetic profiles [12] to predict PPI. Protein domains are discrete functional and evolutionarily conserved units of interaction and many algorithms utilize this information for prediction. These include the association method which is based on co-occurrence of domain pairs in interacting protein pairs [7], domain pair profile method [13], integrative approaches [14], Maximum Likelihood Estimation approach [8,9], Domain Pair Exclusion Analysis  and Bayesian networks [16]. Yet another statistical method assigns p-values for SCOP Superfamily pairs based on pooled dataset of interactions [17]. The degree of co-evolution of interacting and non-interacting domains was analyzed and it was found that in an interaction, interacting domain pairs exhibit a higher level of co-evolution compared to non-interacting domain pairs [18]. Machine learning techniques for PPI prediction include Support Vector Machine (SVM) based method [19], sequence based kernel approach [20], Random Forest Decision Trees [8], a Genetic Algorithm based method Domain GA [22] and others. While most prediction methods concentrate on domain pairs and their association scores, multidomain co-operativity was addressed by [24] using a Linear Programming Algorithm with multi-domain pairs and an Association Probabilistic Method with multi-domain pairs. Another such algorithm is to use the conditional random field approach for predicting PPI using mutual information between domains. [23]

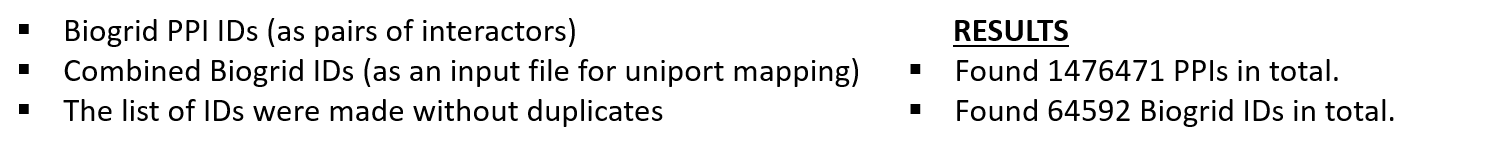
Majority of PPI prediction methods, use experimentally derived PPIs from yeast [7], as the training set while some others also use PPI information from C. elegans and D. melanogaster [8] or from multiple organisms as encoded in DIP and IntAct databases. Most methods derive single domain pair associations while some address multi-domain associations. The success of many existing methodologies for PPI prediction depends on the availability of statistically significant number of PPI which in turn encapsulate specific DDA.

In this study we present a matrix based method for the prediction of PPI based on Pfam domain composition. We use “all against all” i.e., all the proteins in the dataset are tested against each other for protein interactions, experimental PPI datasets from bacteria for finding domain associations.**METHODS**

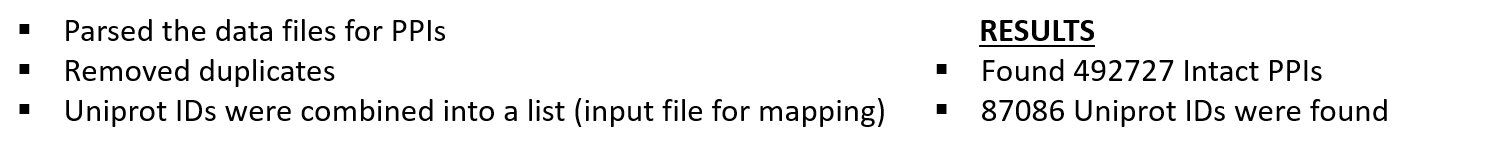
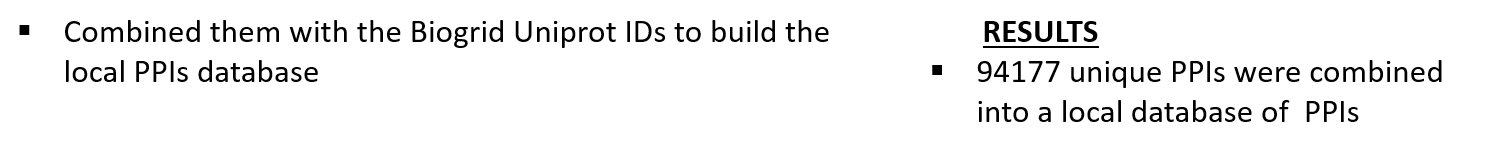
**Parsing and Mapping the Data**

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 uniprot\_sprot\_bacteria.dat.gz file [25] which was the data from Swissprot crosslinked with Uniprot taxonomy 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.

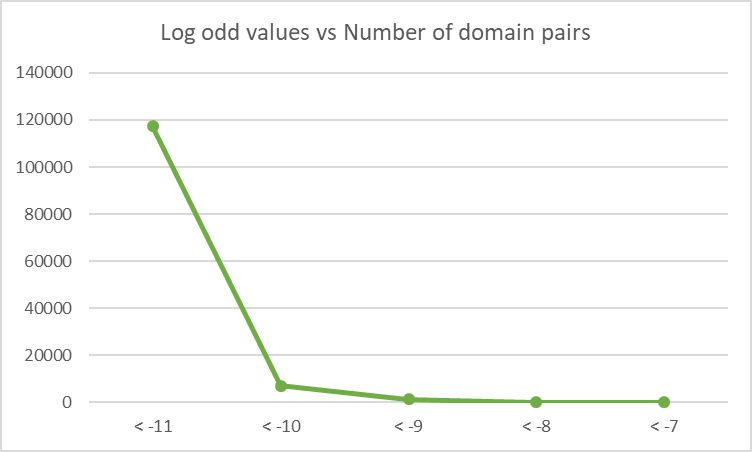
**Figure 2.1: Flowchart describing the steps taken to filter the bacterial protein interactions**

**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, Distribution of log odds values for bacterial PPI-derived sequence signature pairs** 

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. Ferenc Jordán, Thanh-Phuong Nguyen, Wei-chung Liu; Studying protein–protein interaction networks: a systems view on diseases, Briefings in Functional Genomics, Volume 11, Issue 6, 1 November 2012, Pages 497–504, https://doi.org/10.1093/bfgp/els035
2. The study of protein–protein interactions in bacteria, Roberto Velasco-García, Rocío Vargas-Martínez, Canadian Journal of Microbiology, 2012, 58:1241-1257, https://doi.org/10.1139/w2012-104
3. Huthmacher, C., Gille, C., and Holzhütter, H.G. 2008. A computational analysis of protein interactions in metabolic networks reveals novel enzyme pairs potentially involved in metabolic channeling. J. Theor. Biol. 252(3): 456 – 464. doi:10.1016/j.jtbi.2007.09.042. PMID:17988690.
4. 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.
5. 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
6. https://www.nigms.nih.gov/Education/Pages/modelorg\_factsheet.aspx
7. Sprinzak and Margalit, 2001; Deng et al., 2002; Liu et al., 2005; Jansen et al., 2003; Nye et al., 2004; Bock and Gough, 2001; Chen and Liu, 2005; Singhal and Resat, 2007;[24]
8. Y. Liu, N. Liu, H. Zhao Inferring protein–protein interactions through high-throughput interaction data from diverse organisms Bioinformatics, 21 (2005), pp. 3279-3285
9. M. Deng, S. Mehta, F. Sun, T. Chen Inferring domain–domain interactions from protein–protein interactions Genome Res., 12 (2002), pp. 1540-1548
10. E.M. Marcotte, M. Pellegrini, H. Ng, D.W. Rice, T.O Yeates, D. Eisenberg, Detecting protein function and protein–protein interactions from genome sequences Science, 285 (1999), pp. 751-753
11. Overbeek R., M. Fonstein, M. D’Souza, G.D. Pusch, N. Maltsev, The use of gene clusters to infer functional coupling, Proc. Natl. Acad. Sci. USA, 96, 1999, 2896-2901.
12. Pellegrini M., E.M. Marcotte, M.J. Thompson, D. Eisenberg, T.O. Yeates, Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, Proc. Natl. Acad. Sci. USA, 96, 1999, 4285-4288.
13. J. Wojcik, V. Schachter, Protein–protein interaction map inference using interacting domain profile pairs, Bioinformatics, 17 (2001), pp. S296-S305
14. S. Ng, Z. Zhang, S. Tan, Integrative approach for computationally inferring protein–domain interactions, Bioinformatics, 19 (2003), pp. 923-929
15. R. Riley, C. Lee, C. Sabatti, D. Eisenberg, Inferring protein domain interactions from databases of interacting proteins, Genome Biol., 6 (2005), p. R89
16. R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, M. Gerstein, A bayesian networks approach for predicting protein–protein interactions from genomic data, Science, 302 (2003), pp. 449-453
17. T.M.W. Nye, C. Berzuini, W.R. Gilks, M.M. Babu, S.A. Teichmann, Statistical analysis of domains in interacting protein pairs, Bioinformatics, 21 (2004), pp. 993-1001
18. R. Jothi, P.F. Cherukuri, A. Tasneem, T.M. Przytycka, Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions, J. Mol. Biol., 362 (2006), pp. 861-875
19. J.R. Bock, D.A. Gough, Predicting protein–protein interactions from primary structure, Bioinformatics, 17 (2001), pp. 455-460
20. A. Ben-Hur, W.S. Noble, Kernel methods for predicting protein–protein interactions, Bioinformatics, 21 (2005), pp. i38-i46
21. X. Chen, M. Liu, Prediction of protein–protein interactions using random decision forest framework, Bioinformatics, 21 (2005), pp. 4394-4400
22. M. Singhal, H. Resat, A domain-based approach to predict protein–protein interactions, BMC Bioinformatics, 8 (2007), p. 199
23. M. Hayashida, M. Kamada, J. Song, T. Akutsu, Conditional random field approach to prediction of protein–protein interactions using domain information, BMC Syst. Biol., 5 (2011), pp. S1-S8
24. R. Wang, Y. Wang, L. Wu, X. Zhang, L. Chen, Analysis on multi-domain cooperation for predicting protein–protein interactions, BMC Bioinformatics, 8 (2007), p. 391
25. ftp://ftp.uniprot.org/pub/databases/uniprot/current\_release/knowledgebase/taxonomic\_divisions/uniprot\_sprot\_bacteria.dat.gz