ISSN:0975-9646

N.Sevugapandi et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2838-2840

# Data Driven Data Mining Model for Biological Pathways

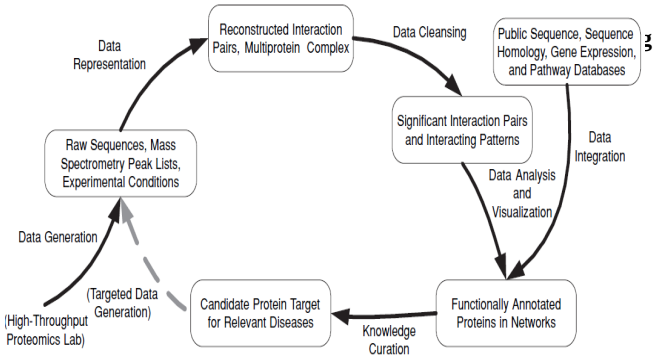
N.Sevugapandi[1] and Dr.C.P.Chandran[2]

[1]*Research Scholar, Ph.D Part-Time, Category –B, Research And Development Center, Bharathiar University, Coimbatore, Tamilnadu*

[2]*Associate Professor of Computer Science, Ayya Nadar Janaki Ammal College, Sivakasi, Tamilnadu*

***Abstract:*** **Advancement in biological data capturing devices such as microarray devices, sequence analyzer helps biologist to systematically monitor cell growth and diseases by elucidating molecular data with genomics, proteomics etc . However making use of these Bio datas is a challenging task to experimentally conclude with any growth in cell or diseases. Data mining techniques in Computing helps algorithmically to solve the above stated problem. Here we propose a new model to store the micro array device data in systematical organization in the perception of computer science. The data mining model addresses data driven knowledge discovery process along with the use of the expression of gene profiles in the context to biological pathway. Here data mining process helps to identify the significant pattern from the consolidated RAW pattern acquired from the RAW data generated by microarray devices. The proposed data mining model can help clinicians to easily identify the cell growth or diseases in early stages with the handy reference to the early stage detection in case of any hazardous tissue growth.**

## INTRODUCTION




The above figure explains the data-mining stages in context of protein interactomics studies. Corresponding to each distinct stage, the challenges for protein interactomics studies are:

1) Data generation challenge: Protein interaction data generated from different high-throughput platforms are often complementary and do not have significant overlap.
2) Data representation challenge. Little progress has been made in representing various types of protein interaction data or their attributes in detail.
3) Data cleansing challenge. Many high-throughput protein interactome data-capturing devices and methods are still in their early development stage; therefore, an uncharacterized level of "noise" exists in all protein interactome data sets.
4) Data integration challenge. Protein interactome data sets have not yet been evaluated in the context of existing sequence homology, protein functional description, and gene expression data at system scales.
5) Data analysis and visualization challenge. Brute-force analysis and the display of protein interaction networks yield little insight into the structure of biological pathways because of a lack of means to reduce noise, data inconsistency, and complexity.
6) Knowledge curation challenge. The protein interaction knowledge must be correlated With the supporting evidence extracted from existing biomedical literature, which is difficult to process due to its free format and rapidly expanding volume. We derived these challenges from our real-world data-mining experience with experimental protein interactome data derived from human brain tissues.

### *A. Data Generation*

We primarily study experimental protein interaction which yields high throughput with 2-hybrid yeast method, to map the proteins interactomes. BD or DNA vector constructs and encodes its DNA binding molecular domain (BD) and also encodes its transcription activation modular domain (AD). These two DNA can be fused subsequently with protein coding techniques. Hybrids of this kind are transformed into different haploid yeasts. Y2H search. Interaction between bait and prey which are protein, activates genes that inducts growth on color reaction or particular media. Y2H can be automated to high gain of protein interactions on a genome-wide scale, which is for viruses like bacteriophage.

If a pair of bait and prey protein interacted in yeast, BD and AD switches on transcription of a yeast reporter gene, in a selective media yeast cells will survive which consequently grow into positive colonies. However BD and AD fail to switch while mated yeast cells dies. Hybrid DNA fragments are extracted for formation and sequencing for the yeast positive colonies which are surviving, interaction of bait and proteins can be identified by biologists. A large number of Y2H searches is performed by many biologists in a library of lone baits

containing yeast colonies simultaneously. Array plates and robots are used to handle this yeast cultures. A thousand of protein pairs per week are discovered by many biologists [1]. This data complement does not compulsorily overlap with those generated from other methods like immune precipitation and pulldown spectrometry [2]

*B. Data Representation*

A protein interaction network is represented by a common method by using a simple graph. In this method, a protein which is identified as a simple bait or prey is represented as node. Representation of data in graph cannot adequately help to accomplish discovery goals. An improved strategy for data representation should take several parameters into consideration which are as follows

- ➔ Different abstract level interactions: three-dimensional (3-D) structural domain level, peptide fragment level, whole protein level, protein-encoding gene level.
- ➔ Reported interaction's observed frequency, directionality of interaction measurement, experimental platform and interaction strengths.
- ➔ Complex interaction types representation; like indirect interactions, multibody N-array interactions generation and mass spectrometry methods.
- ➔ Descriptive network parameter representation like simple node count, simple edge count, the connectivity of nodes, clustering coefficient for nodes.
- ➔ Discovered molecular network knowledge representation, like regulatory chain of interacting proteins.

*C. Data Cleansing*

With the use of database and statistical model the data is audited which detects anomalies and contradictions, this automatically gives an indication of the location and characteristics of anomalies. Even on large set of data workflow should be efficient which inevitably poses a trade-off because the execution of a data-cleansing operation can be computationally expensive.

Uncertainty and ambiguity: poorly sequenced regions or splice variant regions are attacked by BLASTN sequence which results in ambiguity, incorrect or ambiguous identification of proteins. If the target genome BLASTN is not available it becomes uncertain where sequence failed to get identified.

**False Positives:** Y2H positive colonies may grow because of some BD-bait hybrids which are also called as self-activators they can switch on yeast reporting transcription when AD hybrid is absent. AD-prey hybrids choose BD-bait hybrids to interact. Certain "promiscuous proteins" are tough to get removed by applying basic filtering thresholds, which has "<100 interacting partners" these proteins may play the cellular role of "interaction network hubs" due to emerging evidence.

**False Negatives:** In Y2H interaction searches the interacting protein fragment may not be observable because the protein fragment does not fold in correct 3-D structure, which needs specific neighboring structure context.

**Incomplete knowledge:** The functional description of interacting fragments such as Gx and Gy and their domain features that are documented may not be computed or obtainable.

*D. Data Integration*

Collected protein interactome are integrated with genomics which are integrated with functional genomics, proteomics data sets and existing genomics. The protein interactoms results can only be comprehended in the existing context which is protein function's biological knowledge and molecular functions. Correlation of biological observation for protein pair can strengthen confidence for interaction, it can serve as an in silico protein interaction data validation method. At the end data integration prelude high-quality interference, which is basis for hypothesis formulation and testing.

*E. Data Analysis and Visualization*

To gain insight into novel protein functions, anonymous protein and biological processes. Protein and genetic data, analyst annotates protein interaction network through which biologists can traverse via interaction links and explore function knowledge. Functions are assigned to unannotated protein like, guilt-by-association rule can be applied as functional label is assigned to protein by pooling the most infamous functional label in its immediate neighbors.

*F.Knowledge Curation*

Culmination of the entire protein interactome in a data mining process is called as knowledge curation. A new molecular pathway models is constructed by biological knowledge curators. The strength of the gene-disease relationships are interpreted which depicts the degree of variant pathogenicity. Standards are developed for gene curation that defines the amount and type of evidence needed to assist an association between a gene and a particular disease. First multiprotein interaction for a protein pulldown spectrometry method is produced.

## CONCLUSIONS

Computational challenges for protein interactomics are analyzed and presented based on working experience and characterizing human protein interactome data. The variation within system-scale human protein interactome data to collect protein interaction and its interacting domain levels, apply pragmatic noise filters, sequence homology, navigate protein interaction network with 80000 interaction pairs to identify target candidate in disease pathway. The number of available protein interactomes increases along with the interactome data that grows.

## REFERENCES

[1] Hassan, D.; Aickelin, U.; Wagner, C., "Comparison of Distance metrics for hierarchical data in medical databases," *Neural Networks (IJCNN), 2014 International Joint Conference on* , vol., no., pp.3636,3643, 6-11 July 2014.

[2] Aouf, M.; Liyanage, L.; Hansen, S., "Critical Review of Data Mining Techniques for Gene Expression Analysis," *Information and Automation for Sustainability, 2008. ICIAFS 2008. 4th International Conference on* , vol., no., pp.367,371, 12-14 Dec. 2008.

[3] Papadimitriou, S.; Jimeng Sun, "DisCo: Distributed Co-clustering with Map-Reduce: A Case Study towards Petabyte-Scale End-to-End Mining," *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on* , vol., no., pp.512,521, 15-19 Dec. 2008.

[4] Hadzic, F.; Dillon, Tharam S.; Sidhu, A.S.; Chang, E.; Tan, H., "Mining Substructures in Protein Data," *Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on* , vol., no., pp.213,217, Dec. 2006

[5] Aouf, M.; Lyanage, L.; Hansen, S., "Review of data mining clustering techniques to analyze data with high dimensionality as applied in gene expression data (June 2008)," *Service Systems and Service Management, 2008 International Conference on* , vol., no., pp.1,5, June 30 2008-July 2 2008.