

# **Data Integration and Knowledge Discovery using Biological Networks**

by

**Gaurav Kumar**

*Master of Science (Biology)*

*Tata Institute of Fundamental Research (TIFR), India*

A thesis submitted in fulfilment of the requirements for the degree of  
**Doctor of Philosophy**

Department of Chemistry and Biomolecular Sciences  
Macquarie University  
Sydney, Australia

January 2012



**Dedicated to**

*My grandparents (Dada and Dadi) for their endearing love and dream for me,  
who loved me enough to let me go*



## **DECLARATION**

I certify that this thesis entitle “Data Integration and knowledge Discovery Using Network approach”, is a bonafide record of research work carried out by me under the guidance of Professor Shoba Ranganathan during the year 2008-2011 for the degree of Doctor of Philosophy. The results presented in this thesis have not previously formed the basis for award of any degree, fellowship or other recognition. The particulars given in the thesis are true to best of my knowledge.

Gaurav Kumar

January 2012

## ACKNOWLEDGEMENTS

Early in the process of completing this project, it became quite clear to me that a researcher cannot complete a Ph.D. thesis alone. Although the list of individuals I wish to thank extends beyond the limits of this format, I would like to thank the following persons for their encouragement and support:

- ❖ My supervisor, Professor Shoba Ranganathan for providing me an opportunity to work in her group for my PhD thesis and her support as an early career researcher.
- ❖ My co-supervisor Professor Helena Nevalainen for her support throughout my PhD tenure.
- ❖ Macquarie University, for the award of iMQRES research scholarship for pursuing PhD degree and PGRF funding for carrying out short-term research project at Memorial Sloan Kettering Cancer Center, New York, USA.
- ❖ My special thanks to Professor Chris Sander for inspiration I have drawn from their scientific maneuver and agility.
- ❖ Past and current lab colleagues - Dr. Adrian P. Cootes, Dr. Shivshankar Nagaraj, Dr. Jitendra Giakwad, Dr. Javed Khan, Dr. Varun Khanna, Mrs. Ranjeeta Menon, Mrs. Elsa Chacko, Mr. Gagan Garg and Mr. Muralidhara Maradana.
- ❖ Ms. Catherine Wong, Ms. Maria Hyland, Ms. Jane Yang, Mr. Michael Baxter, for administrative and IT support.
- ❖ My mentors, Professor R. Sowdhamini and Dr. Anita Kamra for introducing me to the field of Bioinformatics and Biotechnology, respectively. They have constantly supported and inspired me to do independent research.
- My acknowledgement will be incomplete without thanking my Dadi (grandmother) and all my family members. I would like to remember my 10<sup>th</sup>, 12<sup>th</sup>, B.Sc. and M.Sc. friends and batchmates for sharing their time and experience with me. It is always fun to grow together and cherish the past and present moments with them.

## TABLE OF CONTENTS

<i>Declaration</i>		<i>i</i>
<i>Acknowledgements</i>		<i>ii</i>
<i>Table of Contents</i>		<i>iii</i>
<i>List of Abbreviations</i>		<i>v</i>
<i>Symbols</i>		<i>viii</i>
<i>List of Figures</i>		<i>ix</i>
<i>List of Tables</i>		<i>xi</i>
<i>List of Publications included in this thesis</i>		<i>xii</i>
<i>Abstract</i>		<i>xiii</i>
<b>CHAPTER 1</b>	<b>Introduction and literature Survey</b>	<b>1</b>
1.1	<b>Overview</b>	<b>1</b>
1.2	<b>Experimental methods to elucidate Protein-Protein Interactions (PPIs)</b>	<b>2</b>
1.2.1	<b>Fragment complementation assays</b>	<b>3</b>
1.2.2	<b>Affinity purification methods</b>	<b>4</b>
1.2.3	<b>Complex <i>versus</i> binary protein interactions</b>	<b>8</b>
1.2.4	<b>Experimental studies of protein-protein interactions is incomplete</b>	<b>8</b>
1.3	<b>Biological Data Resources/Databases</b>	<b>9</b>
1.3.1	<b>Protein Interaction Databases</b>	<b>9</b>
1.3.2	<b>Metabolic Databases</b>	<b>13</b>
1.3.3	<b>Subcellular Localisation Database</b>	<b>15</b>
1.3.4	<b>Protein Sequence Databases</b>	<b>16</b>
1.4	<b>Graph-based network models in biology</b>	<b>18</b>
	<i>Publication 1</i>	<i>19</i>
1.5	<b>Data integration using network models</b>	<b>45</b>
	<i>Publication 2</i>	<i>47</i>
1.6	<b>Objectives</b>	<b>69</b>
<b>CHAPTER 2</b>	<b>Methods and Applications</b>	<b>71</b>

<b>CHAPTER 3</b>	<b>Network analysis of human protein location</b>	<b>72</b>
3.1	<b>Summary</b>	72
	<i>Publication 3</i>	73
3.2	<b>Conclusions</b>	89
<b>CHAPTER 4</b>	<b>Dissecting the organization of human and yeast interactomes: network relationships from biological process and molecular function</b>	<b>90</b>
4.1	<b>Summary</b>	90
	<i>Publication 4</i>	91
4.2	<b>Conclusions</b>	127
<b>CHAPTER 5</b>	<b>Identification of ovarian cancer associated genes using an integrated approach in a Boolean framework</b>	<b>128</b>
5.1	<b>Summary</b>	128
	<i>Publication 5</i>	129
5.2	<b>Conclusions</b>	161
<b>CHAPTER 6</b>	<b>Conclusions and future directions</b>	<b>162</b>
6.1	<b>Summary</b>	162
6.2	<b>Conclusions</b>	163
6.3	<b>Innovations</b>	164
6.4	<b>Significance and contributions</b>	164
6.5	<b>Future directions</b>	165
<b>References</b>		<b>167</b>

## **ABBREVIATIONS**

<b>3H</b>	Three-hybrid
<b>ABA</b>	Abscisic acid
<b>AD</b>	Activation domain
<b>B2H</b>	Bacterial two-hybrid
<b>BA model</b>	Barabasi-Albert model
<b>BIND</b>	Biomolecular Interaction Network Database
<b>BioGRID</b>	Biological General Repository for Interaction Datasets
<b>BOND</b>	Biomolecular Object Network Databank
<b>BP</b>	Biological Process
<b>BRENDA</b>	BRAunschweig ENzyme DATabase
<b>CC</b>	Cellular component
<b>co-IP</b>	Co-immunoprecipitation
<b>CYGD</b>	Comprehensive Yeast Genome Database
<b>DAG</b>	Directed acyclic graph
<b>DBD</b>	DNA binding domain
<b>DDBJ</b>	DNA Databank of Japan
<b>DNA</b>	Deoxy-Ribonucleic Acid
<b>DORs</b>	Dense overlapping regulons
<b>EMBL</b>	European Molecular Biology Laboratory
<b>ER model</b>	Erdős-Rényi model
<b>ES</b>	Enrichment Score
<b>FBA</b>	Flux balance analysis
<b>FC</b>	Fragment complementation
<b>FFLs</b>	Feedforward loops
<b>GO</b>	Gene Ontology
<b>GRAM</b>	Genetic regulatory module
<b>GSEA</b>	Gene Set Enrichment Analysis

<b>GST</b>	Glutathione S-transferase
<b>GWAS</b>	Genome wide association studies
<b>HPRD</b>	Human Protein Reference Database
<b>HTFN</b>	Human transcription factor network
<b>IC</b>	Information content
<b>IgG</b>	Immunoglobulin G
<b>IMEx</b>	International Molecular Exchange Consortium
<b>KEGG</b>	Kyoto Encyclopedia of Genes and Genomes
<b>KLR</b>	Kernel logistic regression
<b>M2H</b>	Mammalian two-hybrid
<b>MCL</b>	Multiple component loop
<b>MF</b>	Molecular Function
<b>MIMIx</b>	Minimum Information required to report a Molecular Interaction Experiment
<b>MINT</b>	Molecular INTERaction Database
<b>MIPS</b>	Munich Information Center for Protein Sequences
<b>MLPI</b>	Metabolite-linked protein interaction
<b>MRF</b>	Markov random field
<b>mRNA</b>	<i>messenger</i> -Ribonucleic Acid
<b>MS</b>	Mass Spectrometry
<b>NCBI</b>	National Center for Biotechnology Information
<b>OMIM</b>	Online Mendelian Inheritance in Man
<b>PCN</b>	Percentage of common neighbours
<b>PDB</b>	Protein Data Bank
<b>PGDB</b>	Pathway/Genome Database
<b>PID</b>	Primary immunodeficiency
<b>PIR</b>	Protein Information Resource
<b>PLCP</b>	Paried localisation correlation profile

<b>PPI</b>	Protein-protein interaction
<b>PSI-MI</b>	Proteomics Standards Initiative - Molecular Interactions
<b>RefSeq</b>	Reference Sequence Database
<b>REM-TrEMBL</b>	Remaining TrEMBL
<b>SCL</b>	Subcellular localisation
<b>SDS-PAGE</b>	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
<b>SGA</b>	Synthetic genetic arrays
<b>SIB</b>	Swiss Institute of Bioinformatics
<b>SIMs</b>	Single-input motifs
<b>SIR</b>	Susceptible-infective-removed
<b>SOMs</b>	Self-organizing maps
<b>SP-TrEMBL</b>	Swiss-Prot TrEMBL
<b>STD</b>	Sexually transmitted disease
<b>SVM</b>	Support Vector Machine
<b>TCM</b>	Traditional Chinese Medicine
<b>TFs</b>	Transcription Factors
<b>TrEMBL</b>	Translation of EMBL Nucleotide Sequence Database
<b>UMBBD</b>	University of Minnesota Database of Biocatalysis and Biodegradation
<b>UniMES</b>	UniProt Metagenomic and Environmental Sequences
<b>UniParc</b>	UniProt Archive
<b>UniProtKB</b>	Universal Protein Resource Knowledge Database
<b>UniRef</b>	UniProt Reference Clusters
<b>WWW</b>	World wide web
<b>XML</b>	Extensible Markup Language
<b>Y2H</b>	Yeast two-hybrid

## SYMBOLS

$\chi^2$	Chi-square
$C_i$	Clustering-coefficient
$\langle C_k \rangle$	Average clustering coefficient
$d(u,v)$	Geodesic distance
$E(G)$	Edge set of a graph
$\gamma$	Power-law exponent
$\langle l \rangle$	Average path length
$M_r$	Molecular weight
$P(C_i / C_j)$	Paired Localisation Conditional Probability (PCLP) in compartments $i$ and given that interacting protein neighbour in compartment $j$
$P(k)$	Power-law distribution
$r_\Phi$	Phi-correlation
$V(G)$	Vertices of a graph
$Z(C_i, C_j)$	Z-score correlation between the two given compartments $i$ and $j$

## LIST OF FIGURES

- Figure 1.1      **Key players in biomolecular interactions.** A gene is translated to protein *via* RNA, followed by folding of the protein product to acquire a unique 3D-structure. Protein structure determines its function. Some proteins also regulate gene transcription and translation. Functional complexity of cellular system is determined by the interconnectedness of genes and proteins within the cellular network. 2
- Figure 1.2      **Schematic representation to detect PPIs in the Y2H system.** Expression of the genetically engineered reporter gene, (*His3* is monitored in the presence/absence of interaction. Yeast cells grow under the histidine-deficient condition in the presence of *His3* expression. Proteins X and Y are fused with domain DBD and AD respectively. (A) *Gal4* TF is used to monitor the interaction between X and Y indirectly *via His3* gene expression. (B) In the absence of interaction, the functional form of *Gal4* TF is not formed to express the *His3* gene. Thus, growth of yeast cell is not seen in the histidine-deficient condition. (C) In the presence of interaction *His3* gene expression allows the growth of *His3* gene under histidine-deficient condition. 4
- Figure 1.3      **Principle of GST-pulldown method.** The bait protein is fused with Glutathione S-transferase (GST). GST has a high affinity for glutathione which is fixed on the matrix sepharose-beads. Cell extracts containing proteins are allowed to pass through the fixed GST fusion protein. Prey proteins present in the cell-extract bind to the GST fusion protein and unbound proteins are removed through subsequent elution. 6

Figure 1.4      **Principle of co-immunoprecipitation (Co-IP).** Protein A is fixed      7  
on the sepharose bead which recognises the constant chain of the  
antibody, IgG. The variable regions (i.e. epitope) of the antibody  
are specific to the bait protein. The prey protein binds to the bait  
protein if an interaction exists.

Figure 1.5      **A workflow representation of protein interaction data collation.**      10  
The two central boxes illustrate the general levels of detail collected  
by the databases, whereas, peripheral boxes show various source of  
experimental data related to protein interactions.

## LIST OF TABLES

Table 1.1	<b>Various forms of protein complementation techniques/assay systems.</b> Y2H (Yeast two-hybrid), B2H (Bacterial two-hybrid), M2H (mammalian two-hybrid), FC (fragmentation complementation) and 3H (three-hybrid).	5
Table 1.2	<b>Commercially available affinity tags</b>	7
Table 1.3	<b>List of KEGG resources</b>	12

## LIST OF PUBLICATIONS INCLUDED IN THIS THESIS

The following publications are presented in their published form in this thesis and are referred to from this point onward as listed in respective sections of the thesis.

1. **Kumar G.**, Cootes AP. and Ranganathan S: Untangling Biological Networks Using Bioinformatics: *Algorithms in Computational Molecular Biology: Techniques, Approaches and Application*. Edited by Mourad Elloumi and Albert Y. Zomaya. John Wiley & Sons, New Jersey, Wiley Series in Bioinformatics; 2011:867-888. ISBN: 0-470-505192.  
Contribution to (i) concept: GK 60%, APC 10%, SR 30%; (ii) data gathering: GK 100%; and (iii) writing: GK 60%, APC 10%, SR 30%
2. **Kumar G and Ranganathan S:** Biological Data Integration using Network Model: *Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data*. Edited by Mourad Elloumi and Albert Y. Zomaya. John Wiley & Sons, New Jersey, Wiley Series in Bioinformatics; (in press).  
Contribution to (i) concept: GK 80%, SR 20% (ii) data gathering: GK 100%; and (iii) writing: GK 70%, SR 30%
3. **Kumar G** and Ranganathan S: Network Study of Human Protein Location: *BMC Bioinformatics*; 2010; 11 Suppl 7:S9  
Contribution to (i) concept: GK 80%; SR 20% (ii) data gathering: GK 100%; (iii) data analysis: GK 90%, SR 10%; and (iv) writing: GK 70%, SR 30%
4. **Kumar G**, Cootes AP and Ranganathan S: Dissecting the organisation of Human and Yeast interactome: A detail network relationships on biological process and molecular function. *Manuscript under preparation*.  
Contribution to (i) concept: APC 60% GK 30%; SR 10% (ii) data gathering: GK 100%; (iii) data analysis: APC 50% GK 50%; and (iv) writing: GK 40%, APC 20% SR 40%
5. **Kumar G** and Ranganathan S: Identification of ovarian cancer associated genes using an integrated approach in a Boolean framework. *Manuscript under preparation*.  
Contribution to (i) concept: GK 100%; (ii) data gathering: GK 100%; (iii) data analysis: GK 90%, SR 10%; and (iv) writing: GK 70% SR 30%

## ABSTRACT

The overall objective of this thesis is to analyse and understand the intricate network of protein interactions inside the cell. Proteins are molecular machines, which interact and communicate to perform different cellular functions. Research effort in molecular and cellular biology enables the detection of molecular interactions on a large scale. The experimental results generated by high-throughput studies are archived in various public databases. In this study, statistical and computational approach is used to integrate information from relative inhomogeneous data sources (public databases) derived from high-throughput experiments. Further, the integrated approach is used to explore the relationships within the interacting protein pairs. Graph-based network model is used to determine the protein relationships based on gene ontology (GO) biological processes, molecular functions and cellular components.

Network approach has enabled researchers to study the pervasive nature of protein interactions in systems biology. Moreover, different computational methods have been developed to analyse networks and their topological properties. Foremost among them are the methods for analysing direct/indirect protein interactions networks by integrating with the other types of *-omic* data. This thesis demonstrates the statistical significance of protein interaction networks for the study of subcellular localisation, biological processes and molecular functions. It also suggests the significance of network in biological studies.

The protein-protein interaction (PPIs) network was created by integrating binary protein interactions deposited in various public databases. Similarly, the metabolic network was created by linking proteins *via* metabolites, i.e. indirect protein interactions. Both PPIs and metabolic networks were analysed to show the difference in network topologies. Further, we compared and contrasted the subcellular localisation of human proteins using PPIs and metabolic networks. The statistical significance of human protein localisation is demonstrated through statistical measures such as Chi-square ( $\chi^2$ ) test, protein co-localisation correlation profile and Z-score. These statistical methods are significant to illustrate the cross-talk among various subcellular compartments and highlight the importance of metabolite-linked protein interaction i.e. functional/indirect association in addition to direct physical interaction of proteins.

Statistical analyses were extended further for human and yeast proteomes to show the influence of protein degree for determining protein relationships for biological process and molecular function. This analysis demonstrates the tendency of proximal proteins in a network to have the same relationships to depend strongly on their degree/connectivity. Comparison of real networks with that of randomized networks i.e. permutation testing, suggests the significance of such relationships at a network distance less than three. Networks are randomized using an edge swapping method and the distance in a network is calculated for the shortest path between each protein pair, using the Floyd-Warshall algorithm. The significance of the network distance less than three holds true up to six levels of depth from the root node (i.e. zero level) in the hierarchy of gene ontology (GO) terms.

Application of the network study is further demonstrated using ovarian tumour samples. Gene Expression data from the TCGA (The Cancer Genome Atlas) dataset were collected to encode the functional attributes in a Boolean logic framework for the identification of potential genes in the prognosis and therapy risk assessment in the human diseased condition. The differentially expressed genes were then validated in a co-expression network derived from the ovarian samples deposited in the GEO (Gene Expression Omnibus). A set of 17 differentially expressed genes were identified at the high probability score suggesting their importance in the ovarian cancer diseased condition. Three of these have never been reported before as significant for ovarian cancer.