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 10th, 12th, 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

ABBREVIATIONS

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

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

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

SYMBOLS

χ^2	Chi-square
C_i	Clustering-coefficient
$<C_k>$	Average clustering coefficient
$d(u,v)$	Geodesic distance
$E(G)$	Edge set of a graph
γ	Power-law exponent
$<l>$	Average path length
M_r	Molecular weight
$P(C_i \mid C_j)$	Paired Localisation Conditional Probability (PCLP) in compartments *i* and given that interacting protein neighbour in compartment *j*
$P(k)$	Power-law distribution
r_Φ	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

LIST OF TABLES

List of Publications included in this Thesis

The following publications are presented in their published form in this thesis and are referred to form 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: Prepossessing, 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 (χ^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.