

Chapter 1: Introduction and literature Survey

1.1 Overview

Living organisms are composed of lifeless molecules. When these molecules are isolated and examined individually, they conform to all physical and chemical laws that describe the behaviour of an inanimate matter. The chemical constituents of living organism are mainly composed of Carbon, Oxygen, Nitrogen, Phosphorous and Sulphur. The special chemical bonding of Carbon allows the formation of wide variety of organic molecules. Organic compounds of molecular weight (M_r) less than about 500 daltons, such as amino-acids, nucleotides and monosaccharides, serve as monomer subunit for proteins, nucleic acids and polysaccharides, respectively.

The genome of an organism is composed of DNA (deoxy-ribonucleic acid), encoding information for survival. Moreover, genomes are dynamic entities; they evolve over time by acquiring cumulative effects of mutation, recombination and selection. Decades of research in molecular biology has shown that genes (i.e. genetic information encoded in DNA) translate to proteins *via* mRNAs (messenger-ribonucleic acids). These proteins then fold to acquire unique three dimensional structures to perform specialized functions (Figure 1.1). Almost everything that occurs in the cell involves one or more proteins. Inside cell, genes are tightly regulated by proteins, such as transcription factors. Thus, proteins play a central role as molecular machines to carry-out everyday tasks inside the cell.

The advent of the computer age and recent advances in high-throughput experimental methods has provided an opportunity to understand the cellular complexity within the inter-connectivity of biomolecules (Figure 1.1). Biological interactions happen at the many different levels of detail, from the atomic interactions in a folded protein structure to the relationship of organisms in a population or ecosystem that can be modelled as biological network [1]. Complex biological networks have received a tremendous amount of attention in the past decades [2-4]. These studies have changed our view of the molecular cellular systems in a fundamental way.

Many biological functions cannot be predicted from a single information source. However, there has been progress in this quest through inductive reasoning, biological data integration and modelling efforts. In the explosion of high-throughput experimental data,

one needs computational and statistical techniques for integrating heterogeneous datasets to decipher the interactions between different biological objects and thereby creating new relationships to find novel insights.

This thesis demonstrates the utility of data integration and protein interaction network to understand subcellular localisation (SCL), biological processes, molecular function and gene co-expression of interacting protein pairs. The specific aims of this thesis and how they have been addressed forms the rest of the thesis, followed by conclusions and future direction.

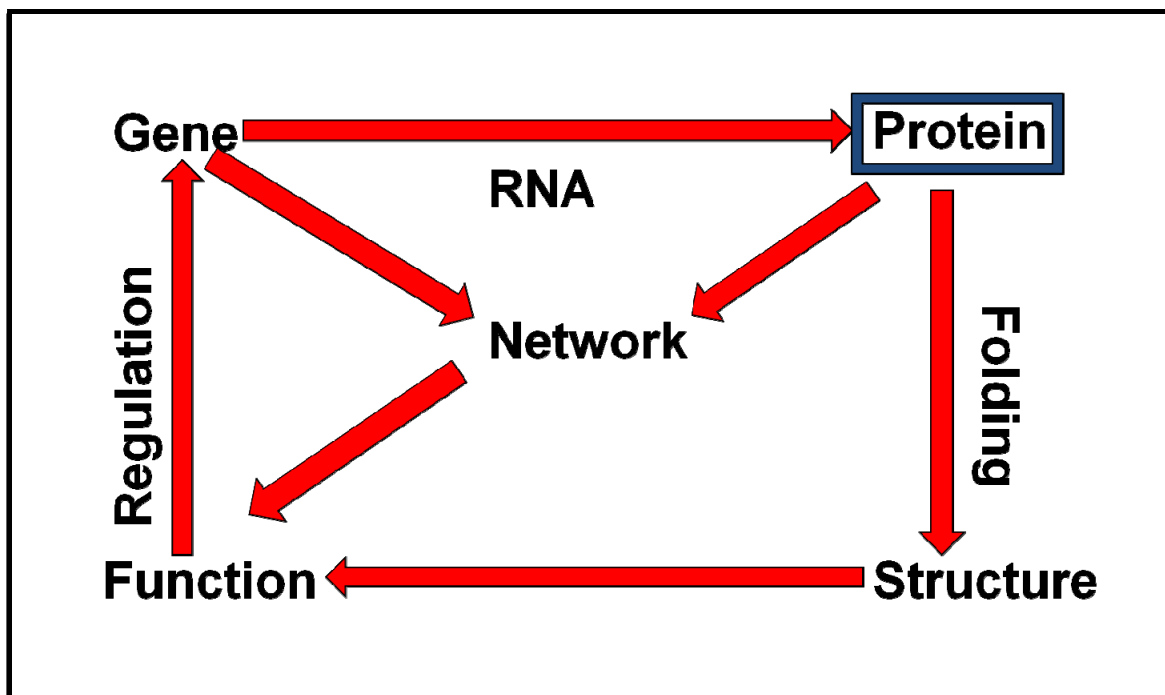


Figure 1.1: Key players in biomolecular interactions. A gene is translated to protein *via* RNA, followed by folding of the protein product to acquire a unique 3D-structure. Protein structure determines its function. Some proteins also regulate gene transcription and translation. Functional complexity of cellular system is determined by the interconnectedness of genes and proteins within the cellular network.

1.2 Experimental methods to elucidate Protein-Protein Interactions (PPIs)

There are many methods to detect PPIs, which falls under two broad categories. The first category comprises Fragment Complementation (FC) assays, which is based on the

principle of split proteins that are functionally reconstituted by the fusion of interacting proteins [5-7]. The second category includes affinity purification methods for structural determination and identification of proteins in complexes. FC assays and protein complex purification methods combined with MS (Mass Spectrometry) have been used in large scale studies of PPIs in model organisms and in human. These two approaches fundamentally represent two different sources of interactions and hence it is important to understand their strength and weaknesses in terms of PPI detection. The FC assay system usually determines binary interactions, whereas a complex purification method detects all the components of complexes.

1.2.1 Fragment complementation assays

The yeast two-hybrid (Y2H) is a popular fragment complementation genetic assay widely used for detecting PPI. It was originally developed by Field and Song inside a living yeast cell [5]. The Y2H system employs a transcription factor, *Gal4*, which activate a reporter gene (i.e. gene whose phenotypic expression is easy to monitor) when its DNA-binding domain (DBD) and its transcription activation domain (AD) are linked. When both DBD and AD domains are separated from each other, *Gal4* lacks capability to activate the expression of reporter gene. To identify whether proteins X and Y interact, each protein is fused to one of these transcription factor domains, AD and DBD, respectively (Figure 1.2 b). Protein X, which binds to DBD domain of transcription factor, is referred as the bait protein. Whereas, protein Y binds to the AD domain is referred as the prey protein. Interaction between X and Y protein indirectly brings DBD and AD domains together to reconstitute the functional form of the transcription factor which then activate the reporter gene (Figure 1.2 c). The expression of reporter gene confirms the true interaction between proteins. In the absence of interaction reporter gene does not express itself.

Several other FC methods are developed for the detection of protein interactions based on the co-expression of two-hybrid fusion proteins, as shown in Table 1.1. All the methods have been proven to work with a selected set of protein interactions. There is lack of systematic approach to compare the quality and methodology biases of these approaches.

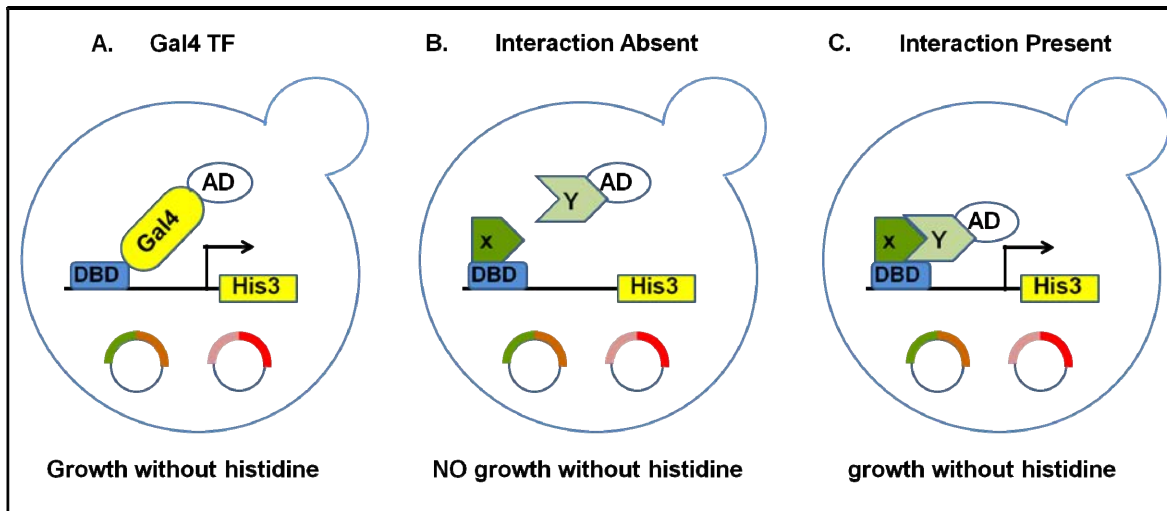


Figure 1.2: Schematic representation to detect PPIs in the Y2H system. Expression of the genetically engineered reporter gene, (*His3* is monitored in the presence/absence of interaction. Yeast cells grow under the histidine-deficient condition in the presence of *His3* expression. Proteins X and Y are 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.

1.2.2 Affinity purification methods

While protein complementation techniques are usually used *in vivo*, affinity purification requires that the interacting proteins to be purified from cell and then identified *in vitro* (even though the interaction takes place *in vivo*) [8]. Historically, GST pulldown and co-immunoprecipitation (co-IP) have been the most popular methods. These methods have been supplemented by refined high-throughput methods that use mass spectrometry for protein identification. However, all these methods are based on the principle that interactions involved affinity-tagged proteins formed *in vivo* are preserved during biochemical purification steps.

Table 1.1: Various forms of protein complementation techniques/assay systems. Y2H (Yeast two-hybrid), B2H (Bacterial two-hybrid), M2H (mammalian two-hybrid), FC (fragmentation complementation) and 3H (three-hybrid).

<i>Class</i>	<i>Method</i>	<i>Principle</i>	<i>Reference</i>
Y2H	Classical Y2H	Reconstitution of active transcription factor (e.g. <i>Gal4</i>)	[5]
Y2H	LexA-based Y2H	Reconstitution of active transcription factor, based on <i>LexA</i> (DBD) and <i>VP16/Gal4</i> (AD)	[9, 10]
Y2H	SOS recruitment system	Activation of Ras signalling pathway made dependent on interaction	[11]
B2H	Split adenylate cyclase	Reconstitution of adenylate cyclase	[6]
B2H	RNA Polymerase recruitment	Activation of reporter gene by RNA polymerase recruitment	[12]
M2H	MAPPIT	Activation of cytokine signaling	[13]
M2H	Reconstitution of active transcription factor	Reconstitution of active transcription factor	[14]
FC	Split ubiquitin (split-UB)	Protein fragment complementation: analysis of membrane proteins	[15, 16]
FC	Split-TEV protease	Protein fragment complementation: flexible choice of reporter gene	[7]
FC	biFC	Protein fragment complementation: fluorescent protein (allow to localize in interaction)	[17]
3H	Three hybrid/kinase co-expression	Classical Y2H with kinase co-expression (detects phosphorylation dependent interactions)	[18]

1.2.2.1 GST-Pulldown

Using glutathione S-transferase (GST) as a tag is a standard approach in *in-vitro* interaction studies of proteins [19-21]. Traditionally, GST-pulldowns have been used to cross-check interactions that were found in the two-hybrid assays and other screening

procedure. GST fusion proteins can be easily expressed and purified from *Escherichia. coli* [22] by running a cell extract through a matrix of glutathione-coated beads. Only GST fusion proteins along with few cellular glutathione-binding proteins bind to the matrix. Non-specific bound proteins can be easily eluted out by salt solutions such as PBS. A second protein solution is allowed to incubate with that of the fusion protein on the matrix. Proteins from this solution then will bind to the GST fusion proteins. The protein fused to GST is known as the “bait” and the protein which binds to it is known as the “prey”. The bait proteins are often radio-labelled, so that interactions can be identified through SDS-PAGE (sodium dodecyl sulphate polyacrylamide gel electrophoresis) separation followed by Western blotting or by mass spectrometry. Subsequent washing of the GST fusion protein (i.e. bait) and the bound interacting protein (prey) is retained. Figure 1.3; shows a schematic representation of GST-pulldown principle.

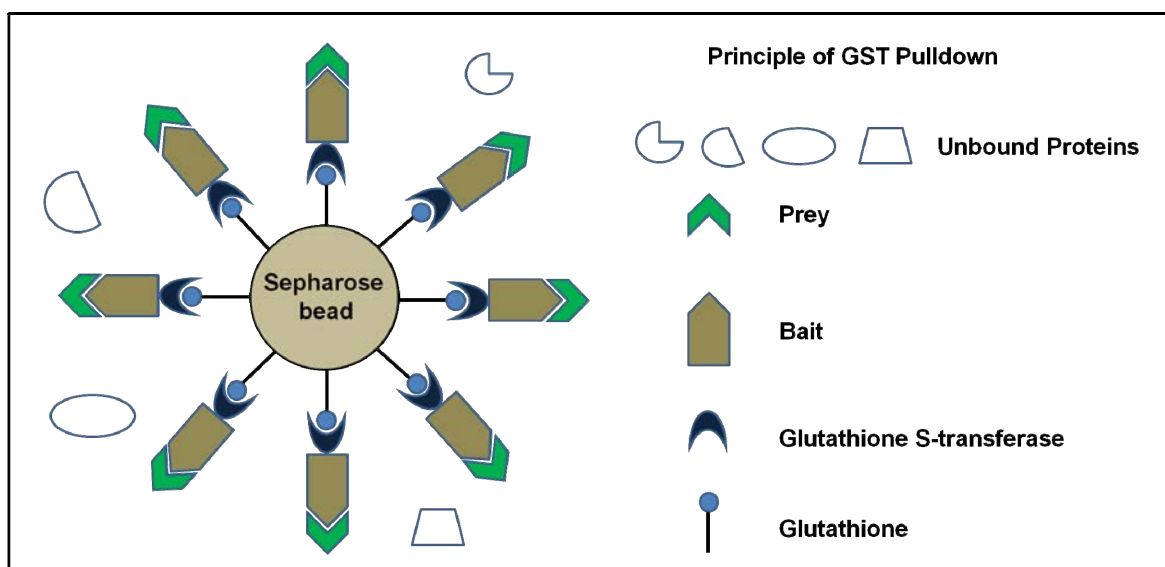


Figure 1.3: Principle of GST-pulldown method. The bait protein is fused with Glutathione S-transferase (GST). GST has a high affinity for glutathione which is fixed on the matrix sepharose-beads. Cell extracts containing proteins are allowed to pass through the fixed GST fusion protein. Prey proteins present in the cell-extract bind to the GST fusion protein and unbound proteins are removed through subsequent elution.

1.2.2.2 Co-immunoprecipitation

Co-immunoprecipitation (Co-IP) has a similar principle in comparison with that of GST-pulldown. It uses Protein A (isolated from *Staphylococcus aureus*) in place of GST.

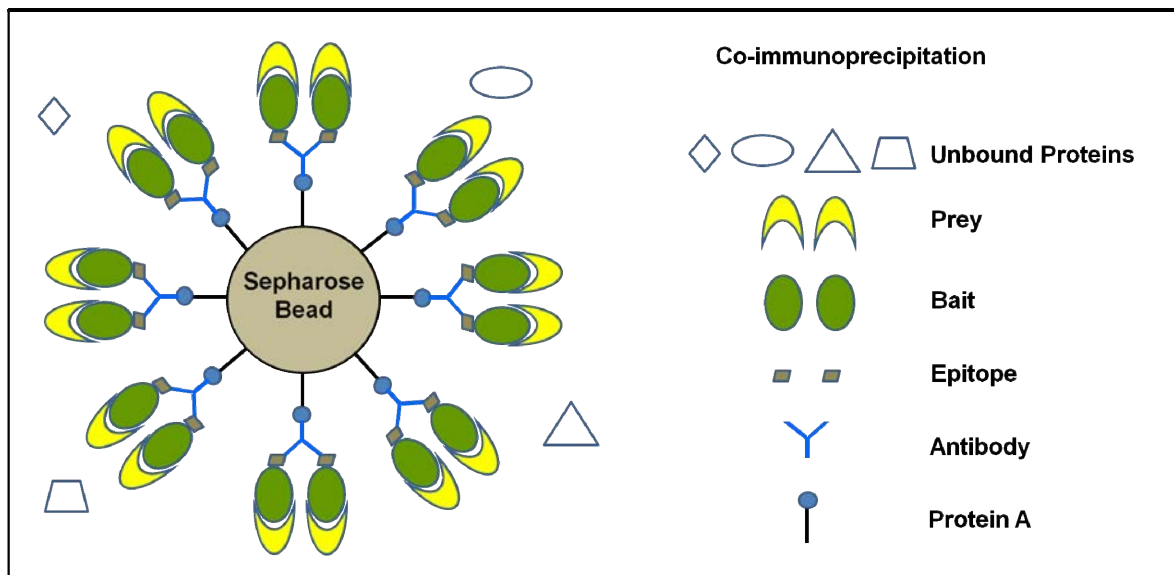


Figure 1.4: Principle of co-immunoprecipitation (Co-IP). Protein A is fixed on the sepharose bead which recognises the constant chain of the antibody, IgG. The variable regions (i.e. epitope) of the antibody are specific to the bait protein. The prey protein binds to the bait protein if an interaction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

Table 1.2: Commercially available affinity tags.

<i>Affinity tag</i>	<i>Capture reagent</i>	<i>Sequence</i>
FLAG	Monoclonal antibody	DYKDDDDK
c-Myc	Monoclonal antibody	EQKLISEEDL
S-tag	S-fragment of RNAase A	KETAAAKFERQHMS
Strep II	Streptavidin variant	WSHPQFEK
Poly-His	Ni ²⁺ -NTA	HHHHHHHH
Poly-Arg	Cation exchange media	RRRRR
Calmodulin-binding domain	Calmodulin	KRRWKKNFIAVSAANRFKKISSGAL

Protein A has a high binding affinity to the constant chains of immunoglobulin G (IgG) antibodies. This allows sepharose-protein A columns to be easily coated with IgG of any specificity. Such a matrix is then incubated with the cell extract. All proteins that are recognised by the antibody, bind to the matrix. Unbounded proteins are then removed

through washing with a buffer. Bound proteins can then be identified *via* separation on SDS-page and Western blotting or mass spectrometry (Figure 1.4). One of the major limitations of Co-IP is the availability of specific antibodies. In the near future though, it will perhaps be possible to produce antibodies against all proteins of a genome. Commercially available peptide affinity tags are available to overcome the limitation of Co-IPs (Table 1.2).

1.2.3 Complex *versus* binary protein interactions

It is important to note that fragment complementation assays detect binary interactions, whereas affinity tag purification methods detect components of complexes. Complex data are often interpreted as if the proteins that co-purify are interacting in a particular manner, consistent with either a spoke or a matrix model. In the spoke model all proteins are assumed to be interacting with the bait protein only, while in a matrix model, all proteins interact with each other. However, a combination of both methods is usually not sufficient to establish the precise topology, as some interactions may be too weak to be detected individually. X-ray crystallography can provide a detailed model of the proteins in a complex. It should be noted that crystallized complexes often lack additional weakly associated proteins that do not co-crystallize and thus may not provide a complete picture.

1.2.4 Experimental studies of protein-protein interactions is incomplete

Proteins work together to carry out various biochemical functions inside the cell. To understand the dynamics of interacting proteins, it is important to know their subcellular location, precise concentration, stability and how the genes of the interacting protein's partners or components are regulated. We also have a limited understanding of post-translation modification and how it affects the assembly of protein complex formation. Keeping these facts in mind, experimental methods only provide a qualitative way to catalogue protein-protein interactions without paying too much attention to the quantitative and dynamic aspects. This will change as we approach complete catalogues of all protein interactions for major model systems and archive these in public repositories for a comprehensive understanding of the biological system. Recent studies estimate that we

have identified only 50% of all yeast interaction and only 10% of all human interaction [23]. Such estimates are not known for other species.

1.3 Biological Data Resources/Databases

Databases are an efficient means to archive, query, retrieve and integrate diverse biological information. Integrating information reduces complexity and provides an efficient way for the dissemination of collated data. The quality of any database relies on its completeness, accuracy and accessibility. The most important and commonly used specialized and general databases for the study of protein interactions are discussed below with their implication for the studies presented in this thesis.

1.3.1 Protein Interaction Databases

As a consequence of experimental and bioinformatics approaches, data on interacting proteins are made available through large scale genome- and proteome-wide analyses. Several research groups have made important efforts in designing and setting up databases that include computer-controlled information about protein interactions or “interactomes”. Intuitively, protein interaction means physical contact between the surfaces of proteins. Figure 1.5 shows a workflow representation of different data source to define protein interaction. Protein relationships can be defined in many ways such as inclusion in a multiprotein complex, common metabolic pathways, common cellular compartments, co-expression, genetic regulation or even molecular co-evolution. These multiple types of protein relationships represent a confounding data landscape. One of the major role of protein interaction database is to provide a way of defining biological entities with a clear definition by extensively linking the data object to several biological characterizations. Moreover, interactome databases not only incorporate new collections of interacting proteins but also curate the definition and annotation of protein interaction included in each case. There has also been a huge effort to display the interaction data in an interactive customizable way. Described below are the important protein interaction databases used to define protein interaction networks.

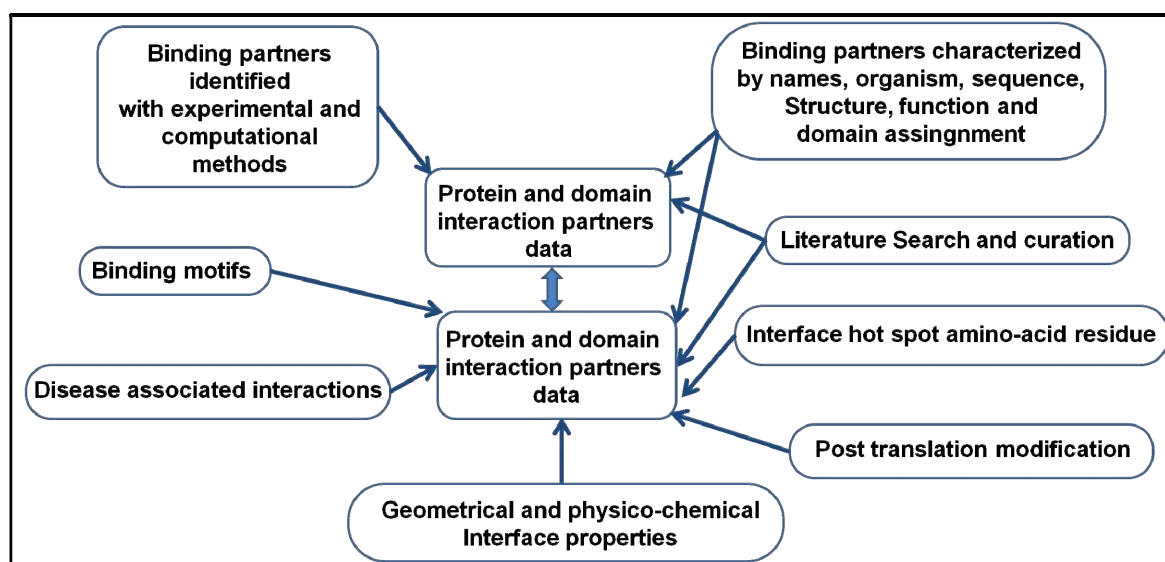


Figure 1.5: A workflow representation of protein interaction data collation. The two central boxes illustrate the general levels of detail collected by the databases, whereas, peripheral boxes show various source of experimental data related to protein interactions.

1.3.1.1 HPRD

The Human Protein Reference Database (HPRD; <http://www.hprd.org>) contains protein interactions and pathways information of human proteins from literature by manual curation [24]. Annotation in this database contains a high level of detail, such as post-translation modification, subcellular localisation, expression, protein-domain architecture and disease association with OMIM (Online Mendelian Inheritance in Man) database. It is connected well with other web resources on proteins.

1.3.1.2 MINT

The Molecular INTeraction Database (MINT; <http://mint.bio.uniroma2.it/mint>) is a relational database, archiving experimentally determined protein-protein interaction [25]. The whole interaction dataset is freely available as XML documents according to PSI-MI (Proteomics Standards Initiative-Molecular Interactions) Level 1 and 2.5 standards [26], MITAB formatted files (a tab-delimited format defined by PSI-MI group where all complexes are represented as binary interactions) and as a simplified tab-delimited file where all participants of an interaction are represented by a single line. This database

provides searching over the web, exploring the interaction network using the MINT Viewer, submitting interaction data to MINT and downloading the interaction dataset in the MITAB format described above.

1.3.1.3 BOND

The Biomolecular Object Network Databank, formerly known as BIND (Biomolecular Interaction Network Database), (BOND; (<http://bond.unleashedinformatics.com>), collates high-throughput experimental data, curated by in-house team of curators along with other data sources including protein complexes from PDB (Protein Data Bank) [27-29]. This database is developed to handle various types of protein-protein interaction data, protein-small molecule interactions and protein-nucleic acid interactions. BOND uses a grammar of unique icons to distinguish functional types of interactions in displays. BOND's web interface also uses specialized text query builder for searching the database.

1.3.1.4 MPact/MIPS

The Protein Interaction and Complex Database (MPact; <http://mips.helmholtz-muenchen.de/genre/proj/mpact/>) is the common access point to the protein interaction resource in the MIPS (Munich Information Center for Protein Sequences) Comprehensive Yeast Genome Database (CYGD) [30-32]. It contains two separate sets of yeast protein interactions, one with manually curated interaction and other set generated *via* high-throughput experimentation. It is very comprehensive and high-quality interactome dataset, commonly referred to as the gold-standard in the analysis of protein interaction network.

1.3.1.5 DIP

The Database of Interaction Proteins (DIP; <http://dip.doe-mbi.ucla.edu>) is one of the best known repositories for experimentally-determined protein-protein interactions including a subset of interactions which have passed a quality assessment [33]. Sources for this interaction data range from literature and PDB, to high-throughput methods like Y2H, protein microarrays and TAP/MS analysis of protein complexes. The database makes use

of several assessment methods to determine the quality of existing interaction data and to check user-specified interaction sets. DIP can also be accessed *via* a plugin in Cytoscape to view molecular interaction networks and links to several related databases including LiveDIP and Prolinks. For proteins in a biological interaction, LiveDIP records information about their states and any state changes upon binding, such as covalent modification, conformations or cellular location [34, 35]. Protlinks employs four methods of functional association: phylogenetic profiles, Rosetta Stone, gene neighbour and gene clusters [36].

1.3.1.6 IntAct

The IntAct Molecular Interaction Database (<http://www.ebi.ac.uk/intact>) provides experimentally determined interactions of biomolecules through deep curation model to capture a high level of experimental details from peer-reviewed research articles. Their interaction data is provided to user which complies with the International Molecular Exchange Consortium (IMEx) guideline and the Minimum Information required to report a Molecular Interaction Experiment (MIMIx) standard [37, 38]. Moreover, interaction data has high-level details to describe experiments, binding-sites, protein tags and mutation *via* PSI-MI ontology. Gene Ontology is used to describe the subcellular location or molecular functions. Interacting molecules are systematically mapped to stable identifiers from public databases such as UniProtKB for proteins, ChEBI [39] for small molecules, Ensemble [40] for genes and DDJB/EMBL/GeneBank [41-43] nucleotide database for nucleic acids sequence. Binding site of protein is linked *via* InterPro database [44] and maps to the protein's sequence/structure.

1.3.1.7 BioGrid

The Biological General Repository for Interaction Datasets (BioGRID; <http://thebiogrid.org>) provides genetic and protein interactions for human and model organisms [45]. This database contains complete coverage for budding yeast (*Saccharomyces cerevisiae*), fission yeast (*Saccharomyces pombe*) and thale cress (*Arabidopsis thaliana*). The interaction datasets are freely available for public usage in standard format such as PSI-MI [46] and tab-delimited text file. BioGRID 3.0 supports

nearly 17 million systematic names, aliases, official symbol and external identifiers from Ensemble [40], UniProtKB [47], NCBI, RefSeq [48], Entrez-Gene, GeneBank [43], SGD [49], WormBase [50], FlyBase [51], MGD [52] and TAIR [53], amongst other sources. BioGRID interaction data is supported by network visualization tools such as Osprey [54], Cytoscape [55], GeneMania [56] and ProHits [57].

1.3.2 Metabolic databases

Metabolic databases integrate molecular information derived from genome sequencing projects into a higher-level of biological organisation to represent cellular complexity. It defines a functional unit to represent cellular processes. These functional units correspond to different levels of molecular organisation such as pathways or operons for metabolic or regulatory networks. The collated information in such databases is of particular interest for two principle reasons: firstly, the topology of metabolic networks provides the basis for detail experimentation, such as drug design for key enzymes [58]; and secondly, the availability of thermodynamic, kinetic and regulatory information allow to achieve a desirable phenotype through simulation and optimization of specific pathways [59]. One of the oldest and the largest projects on the recording of metabolic enzymes, reaction and pathways is the EMP project [60]. Some of the important metabolic databases are described below in detail.

1.3.2.1 KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG; <http://www.genome.jp/kegg>) provides a reference knowledgebase for linking genomes to biological systems (listed in Table 1.3), categorized as building blocks in the genomic space (KEGG GENES) and the chemical space (KEGG LIGAND) [61]. It also provides a deep overview of interaction networks and reaction networks (KEGG PATHWAY). The KEGG PATHWAY database contains pathway maps for molecular systems in both normal and perturbed state. The disease database of KEGG represents a list of known genes, any known environmental factor at the molecular level, diagnostic markers and therapeutic drugs, which may reflect the underlying molecular system. KEGG DRUG is a repository of chemical structures and components used as drugs in USA, EUROPE and TCM (Traditional Chinese Medicine).

Their new resource KEGG MEDICUS provides a reference point for the computational analysis of molecular network by integrating large-scale experimental dataset.

Table 1.3: List of KEGG resources.

Database/content	URL
KEGG PATHWAY	http://www.genome.jp/kegg/pathway.html
KEGG GENES	http://www.genome.jp/kegg/genes.html
KEGG LIGAND	http://www.genome.jp/kegg/ligand.html
KEGG BRITE	http://www.genome.jp/kegg/brite.html
KEGG DRUG	http://www.genome.jp/kegg/drug/
KEGG GLYCAN	http://www.genome.jp/kegg/glycan/
KEGG REACTION	http://www.genome.jp/kegg/reaction/
KEGG EXPRESSION	http://www.genome.jp/kegg/expression/
KEGG ANNOTATION	http://www.genome.jp/tools/kaas/
KEGG DISEASE	http://www.genome.jp/kegg/disease/
KEGG ORTHOLOGY	http://www.genome.jp/kegg/ko.html

1.3.2.2 BioCyc

The BioCyc Collection of Pathway/Genome Databases (<http://www.biocyc.org>) comprises metabolic pathways and genomes databases (PGDBs) of sequenced organisms [62]. As of now, it provides an electronic reference to 1129 genomes. Moreover, it includes computational inference procedures applied to these genomes, including prediction of gene codes for missing enzymes in metabolic pathways and predicted operons. The databases within BioCyc are organised into different tiers according to the amount of manual review and updates they have received.

1. Tier 1 PGDBs are created through manual efforts and receive continuous updates. EcoCyc [63], MetaCyc [62], AraCyc [64] and YeastCyc belong to tier 1.
2. Tier 2 PGDBs contain computationally generated pathway information by the PathoLogic program [65]. The databases undergo a moderate amount of review and update. HumanCyc [66] belongs to this tier. In total, BioCyc contains Tier 2 metabolic data for 33 organisms.

3. Tier 3 PGDBs contain predicted metabolic pathways, operons (for bacteria only) [67], pathway hole fillers [68] and transport reactions [69]. There are a total of 1084 Tier 3 PGDBs.

1.3.2.3 BRENDA

The BRAunschweig ENzyme DAtabase (BRENDA; <http://www.brenda-enzymes.org>) contains information on metabolic enzymes and their ligands, including information on inhibitors, kinetics, thermodynamics and links to organisms for which the relevant reactions have been characterized [70]. Information is obtained through manual curation from the primary literature source. It is one of the most comprehensive resources on enzymes.

1.3.2.4 UMBBD

The University of Minnesota Database of Biocatalysis and Biodegradation (UMBBD; <http://umbbd.msi.umn.edu>) is more specialized database focusing on biodegradation [71]. This database contains information on xenobiotic compounds interconversions, of particular interest for industrial applications and bioremediation projects.

1.3.3 Subcellular Localisation Database

Annotating the subcellular localisation (SCL) of proteins is important in implicating the function and their possibility of interaction with other proteins inside the eukaryotic cell. In prokaryotic cell system, only three locations are possible. Prokaryotic proteins can be present either inside, outside or within the plasma membrane. However, protein localisation is complicated in eukaryotic systems due to the presence of various subcellular compartments. SCL is one of the main biological features that involve protein-protein interaction inside cell. It has been suggested that 76% of yeast [72] and 52% of human [73] proteins tend to co-locate in the same subcellular compartments. In this thesis we have used SCL information from LOCATE database (described below).

1.3.3.1 LOCATE

The Mammalian Protein Localization Database (LOCATE; <http://locate.imb.uq.edu.au>) is a curated database containing information of membrane organisation and SCL for the mouse and human proteins [74]. It uses protein sequences based on the transcripts generated from the direct sequencing of full-length transcripts, generated by the RIKEN FANTOM consortium [75]. At present, it archives a total of 58128 and 64637 proteins for mouse and human respectively. This warehouse has both literature curated and experimentally determined information on the SCL of mammalian proteins.

1.3.4 Protein Sequence Databases

The wealth of sequence information obtained from genome projects lack structural and functional annotations for various gene products, i.e. proteins. A biological insight could only be acquired through proper annotation. Therefore, systematic application of computational tools aided with expert opinion assists in identification and association of genomic sequences with well-characterised proteins. Organising the plethora of information through curation and maintenance of sequence databases is essential for performing efficient biological analysis. Protein sequence databases arise from the translation of nucleic acid sequences obtained from sequencing projects.

1.3.4.1 Entrez Protein Database

The National Center for Biotechnology Information (NCBI)'s Entrez (<http://www.ncbi.nlm.nih.gov/protein>) contains one of the most comprehensive and exhaustive repositories for protein sequences [43, 76]. It contains sequence data translated from nucleotide sequences sourced from DDBJ (DNA Databank of Japan) [41], European Molecular Biology Laboratory (EMBL) [42], GenBank [43], as well as sequences from Swiss-Prot (detailed in the next section), Protein Information Resource (PIR) [77], RefSeq [48] and Protein Data Bank (PDB) [78].

1.3.4.2 Swiss-Prot

The ExPaSY Protein Sequence Database (Swiss-Prot) is a protein sequence database, which from its inception in 1986, was produced collaboratively by the Department of Medical Biochemistry at the University of Geneva and the EMBL. After 1994, the collaboration moved to EMBL's UK outstation, EBI. In April 1998, further changes saw a move to the Swiss Institute of Bioinformatics (SIB), hence SIB and EBI/EMBL now maintain the database collaboratively. The database provides high-level protein annotations, which includes functional descriptions, domain architecture, post-translational modification, splice variants and so on. Swiss-Prot aims to minimize redundancy and is interlinked to many other resources. The structure of the database, and the quality of its annotation, set Swiss-Prot apart from other protein sequence resource and have made it the database of choice for most research purpose [79, 80].

1.3.4.3 TrEMBL

The Translation of EMBL Nucleotide Sequence Database (TrEMBL) was created in 1996 as a computer-annotated supplement to Swiss-Prot [81]. The database benefits from Swiss-Prot format and contains translations of all coding sequences (CDs) in EMBL. TrEMBL has two main sections, designated as SP-TrEMBL (Swiss-Prot TrEMBL) and REM-TrEMBL (Remaining TrEMBL). SP-TrEMBL contains protein entries to be incorporated into Swiss-Prot. Whereas REM-TrEMBL contains protein entries which are not included into Swiss-Prot, such as sequences which are synthetic, truncated, pseudogenes, fragments or of immunoglobins and T-cell receptors.

1.3.4.4 UniProtKB

The Universal Protein Resource Knowledgebase (UniProtKB; <http://www.uniprot.org>) is a comprehensive catalogue of information on proteins [77, 82]. It is a central repository of protein sequence and function created by merging information contained in Swiss-Prot, TrEMBL and PIR. UniProt comprises four components, each optimised for different usage as shown below:

1. The UniProt Knowledgebase is the central access point for extensive curated protein information, including function, classification and cross-reference [47].
2. The UniProt Reference Clusters (UniRef) databases combine closely related sequence into a single record to speed searches [83].
3. The UniProt Archive (UniParc) is a comprehensive repository, reflecting the history of all protein sequences [84]. All new and updated protein sequences are collected and loaded daily into UniParc for full coverage.
4. The UniProt Metagenomic and Environmental Sequences (UniMES) contains metagenomic and environmental data.

1.3.4.5 RefSeq

The Reference Sequence Database (RefSeq; <http://www.ncbi.nlm.nih.gov/RefSeq>) is curated and maintained at the NCBI [48] to provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts and proteins. This database is very useful for medical, functional and diversity studies. It contains sequence information of 297 fungal, 6424 microbial, 216 plants, 146 protozoan, 351 mammalian vertebrates, 1292 non-mammalian vertebrates and 2727 viral genomes.

1.4 Graph-based network models in biology

A comprehensive overview of graph-based network models in biological systems has been reviewed in publication 1 below. Here, we provided a brief description of a graph as a mathematical object, used for modelling complex biological networks and highlighting some of the simple rules to understand a small set of organising principles. This review also provides the reasons underlying the common topological properties shared by many biological networks and the properties of these networks, in order to understand the cellular complexity. We have also discussed some of the main biological network models to understand the cellular behaviour at the molecular level of genes and proteins.

Due to copyright reasons, the selected article has been omitted from this thesis (the article appears from page 19-44):

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.

1.5 Data integration using network models

Following a comprehensive overview of network models in biology, we also reviewed the different data resources available for network integration and their implication for human disease conditions. Here, we have highlighted the importance of combining biological data from relatively heterogeneous data sources from multiple experiments for meta-analysis at the level of cellular systems. We briefly discuss gene and protein interaction databases, gene ontology and gene-expression, followed by a review of the network modelling efforts to integrate data for predicting proteins and genes interactions and their usage in modelling to further the understanding of human diseases. For details, see publication 2 below.

Due to copyright reasons, the selected article has been omitted from this thesis (the article appears from page 47-68):

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

1.6 Objectives

Currently, there is a need to analyse and understand the intricate network of protein interactions inside the cell, as proteins interact and communicate to perform various cellular functions and since experimental determination of protein interactions using high-throughput techniques have made it possible to understand protein connectivity at the cellular level. An integrated approach is needed to combine heterogeneous experimental data resources to explore protein relationships in interaction networks. 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 a network study for gene expression data. The above objectives were subdivided into specific aims listed below, and addressed in detail in five publications presented in this thesis:

1. Review the current status of biological networks and their potential application in describing the cellular complexity using modelling approaches (Publications 1 and 2).
2. Analyse protein-protein interaction and metabolic networks with respect to network topologies and compare and contrast the subcellular localisation of human proteins, using these networks.
3. Analyse PPI networks for human and yeast proteomes to show the influence of network distances and the level of abstracts in GO (gene ontology) hierarchy in determining protein relationships for biological processes and molecular functions.
4. Demonstrate the application of the networks using ovarian tumour samples, with gene expression data from the publically available databases superimposed with gene/protein functional attributes to characterise the diseased state.

Chapter 2: Methods and Applications

Methods and applications that were developed and used in this study are summarised in Table 2.1. The ensuing publications have also been listed and included in the relevant chapter.

Table 2.1: Methods, applications and publications

Methods/Applications	Chapter	Refer to Publication
Untangling Biological Networks Using Bioinformatics	1	1
Biological Data Integration using Network Model	1	2
Network analysis of human protein location	3	3
Dissecting the organisation of human and yeast interactomes: network relationships from biological process and molecular function	4	4
Identification of ovarian cancer associated genes using an integrated approach in a Boolean framework	5	5

Chapter 3: Network analysis of human protein location

3.1 Summary

Proteins are the fundamental molecular machines involved in various cellular processes *via* pairwise or multivalent interactions with other proteins. Discovering and modelling the protein-protein interactions (PPIs) have been a goal of today's systems biology. To interact, proteins must have spatial constraints (i.e. same subcellular compartments or SCL) or have to be at the interface transiently or conditionally between physically interacting SCL. Both experimental and computational methods helped identifying proteins location within various SCL inside the cell. Experimental approaches are expensive and bound by artefact with limitation to understand the biological complexity [85], whereas computational methods allow faster and general description of protein SCL with less accuracy and needs experimental or empirical validation [86]. Recent studies suggest that interacting protein pairs tend to co-locate inside the same subcellular compartments in human [87-89], fly [90] and yeast [72, 73, 91].

In this chapter, empirical study has been carried out to demonstrate the tendency of interacting protein pairs to have same subcellular compartments in human. Statistical analysis was done to compare and contrast the differences in SCL properties using PPI and metabolic networks model. Protein-protein interactions (PPIs) network was created by integrating binary protein interactions deposited in various public databases. Similarly, metabolic network was created by linking proteins *via* metabolites, i.e. indirect protein interactions or functional linkage. 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 (standard normal distribution or Z-distribution). 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.

PROCEEDINGS

Open Access

Network analysis of human protein location

Gaurav Kumar¹, Shoba Ranganathan^{1,2*}

From Asia Pacific Bioinformatics Network (APBioNet) Ninth International Conference on Bioinformatics (InCoB2010)
Tokyo, Japan. 26-28 September 2010

Abstract

Background: Understanding cellular systems requires the knowledge of a protein's subcellular localization (SCL). Although experimental and predicted data for protein SCL are archived in various databases, SCL prediction remains a non-trivial problem in genome annotation. Current SCL prediction tools use amino-acid sequence features and text mining approaches. A comprehensive analysis of protein SCL in human PPI and metabolic networks for various subcellular compartments is necessary for developing a robust SCL prediction methodology.

Results: Based on protein-protein interaction (PPI) and metabolite-linked protein interaction (MLPI) networks of proteins, we have compared, contrasted and analysed the statistical properties across different subcellular compartments. We integrated PPI and metabolic datasets with SCL information of human proteins from LOCATE and GOA (Gene Ontology Annotation) and estimated three statistical properties: Chi-square (χ^2) test, Paired Localisation Correlation Profile (PLCP) and network topological measures. For the PPI network, Pearson's chi-square test shows that for the same SCL category, twice as many interacting protein pairs are observed than estimated when compared to non-interacting protein pairs ($\chi^2 = 1270.19$, $P\text{-value} < 2.2 \times 10^{-16}$), whereas for MLPI, metabolite-linked protein pairs having the same SCL are observed 20% more than expected, compared to non-metabolite linked proteins ($\chi^2 = 110.02$, $P\text{-value} < 2.2 \times 10^{-16}$). To address the issue of proteins with multiple SCLs, we have specifically used the PLCP (Pair Localization Correlation Profile) measure. PLCP analysis revealed that protein interactions are majorly restricted to the same SCL, though significant cross-compartment interactions are seen for nuclear proteins. Metabolite-linked protein pairs are restricted to specific compartments such as the mitochondrion ($P\text{-value} < 6.0\text{e-}07$), the lysosome ($P\text{-value} < 4.7\text{e-}05$) and the Golgi apparatus ($P\text{-value} < 1.0\text{e-}15$). These findings indicate that the metabolic network adds value to the information in the PPI network for the localisation process of proteins in human subcellular compartments.

Conclusions: The MLPI network differs significantly from the PPI network in its SCL distribution. The PPI network shows passive protein interaction, possibly due to its high false positive rate, across different subcellular compartments, which seem to be absent in the MLPI network, as the MLPI network has evolved to maintain high substrate specificity for proteins.

Background

The eukaryotic cell consists of many different subcellular compartments or organelles. Most of the cellular functions critical to the cell's survival are performed by proteins inside the cell. A typical cell thus contains a large number of protein molecules that are resident in

specific compartments or organelles, referred to as "subcellular locations" (SCL). The major compartments, according to the Gene Ontology Consortium, are: cell surface, chromosome, cytoplasm, cytoskeleton, cytosol, endosome, endoplasmic reticulum, extracellular region, Golgi apparatus, membrane, mitochondria, nucleus, spliceosome, ribosome, vacuoles and organelle lumen [1]. These subcellular compartments are further refined into more specific compartments.

* Correspondence: shoba.ranganathan@mq.edu.au

¹ARC Centre of Excellence in Bioinformatics and Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney NSW, Australia
Full list of author information is available at the end of the article

The functions of proteins are determined by specific physico-chemical environment present inside various compartments or organelles. Therefore, it is important to identify the SCL of each protein, for understanding its functional and cellular role. While protein SCL can be determined by biochemical experimentation, with the growing number of new protein sequences in the post-genomic era, experimental characterization of SCL is available for only 11.1% of the total protein sequences present in the UniProt Knowledge Base (version 57.9) [2]. For human proteins, the number is slightly better, with 34.1% having SCL annotations (Table 1). There is thus a huge gap between protein sequences with and without SCL annotation, necessitating computational approaches to predict the SCL from sequence information.

Early computational methods were restricted to specific subcellular compartments and depended on sequence information alone [3]. Protein sequence information comprises amino-acid composition, their physico-chemical properties (such as molecular weight, hydrophobicity, side-chain mass and amino-acid propensity), protein motifs, signal peptides and functional domain composition. However, given the variety of accepted subcellular locations that are functionally essential to completely characterize a protein, novel approaches such as machine learning and text mining have improved SCL predictability [3,4]. A machine-learning method relies on the recognition of patterns that are best characterized on the set of proteins whose localisation are known. A few studies use a systems biology approach for the prediction of a protein's SCL [5], adopting an integrated methodology of high-throughput proteomic data such as protein-protein interaction (PPI) networks and protein motifs to understand and predict the SCL of a eukaryotic protein [5,6].

The use of PPI network to predict function relies on the principal assumption that the interacting protein pairs are likely to collaborate for a common purpose and have to be in close proximity in order to interact. Schwikowski *et al.* [7] were the first to show that the

Saccharomyces cerevisiae PPI network could be used to classify protein SCL based on the idea of "guilt by association or neighbouring count method". Their approach correctly identifies 76% of the interacting protein pairs as occurring within the same SCL. A similar approach was used in a comparative study to show that 52% of the interacting protein pairs in humans tend to have same SCL [8]. Lee *et al.* [9] extended the network-based approach by complementing the classification with a 'Divide and Conquer k-Nearest Neighbour' (DC-kNN) approach, with increased SCL predictive ability in yeast. Previous researchers have shown the importance of highly connected metabolites in the evolution of biochemical pathways which govern the flow of mass and energy in an organism [10,11]. To the best of our knowledge, the metabolite-linked network has only been used by Wagner and Fell [11] to report a positive correlation between the evolutionary age of metabolites and their degree of connectivity. Oron *et al.* [12] used constraint-based modelling on the metabolic network for predicting enzyme SCL, specifically considering the cross-membrane metabolite transporters (i.e. proteins). Thus, metabolic network information has not been implemented for predicting protein SCL, compared to data from PPI networks. As a first step towards developing such a prediction methodology, we have carried out large-scale statistical analysis of the SCL information contained in PPI and metabolite-linked networks.

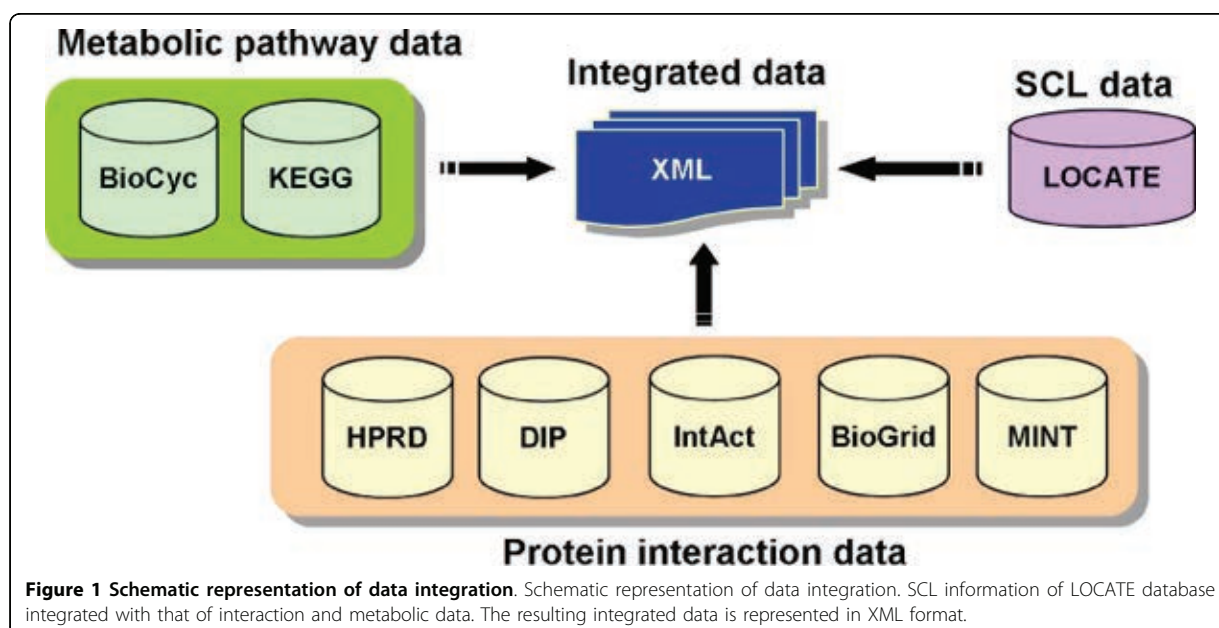
The availability of a large number of protein interaction and metabolic datasets from multiple databases has motivated us to conduct a statistical study to benchmark the predictive ability of localisation of human proteins, with respect to the various subcellular compartments. In this study, we collated PPI interaction and metabolite-linked protein interaction (metabolic information) from seven major databases and integrated these with the high quality SCL information present in the LOCATE database [13] (Figure 1; see Materials and Methods for details), to critically analyze the PPI and metabolic datasets for the SCL assignment of human proteins. Using experimentally validated physical interaction and

Table 1 Summary of SCL annotation in UniProtKB.

Items	Description	No. of Protein Sequences	Dataset Size	%
A	Proteins with SCL annotation in UniProt database	274730	494762	55.52
B	Proteins in A with experimentally known SCL	55079	494762	11.13
C	Proteins in A with uncertain terms such as potential/probable/similarity	219651	494762	44.39
D	Proteins with GO annotation	461365	494762	93.24
E	Protein with SCL annotation in GO database	337762	494762	68.26
F	UniProt human entries with experimentally known SCL	6923	20274	34.14
G	UniProt human entries with uncertain terms such as potential/probable/similarity	7486	20274	36.92

Distribution of 494762 protein entries from UniProtKB/Swiss-Prot* database (version 57.9) according to their SCL annotation and GO database reference.

* The original number of UniProt protein entries was 510076. Of these, 15314 were annotated as "fragment" or contained less than 50 amino acids residues, hence, were removed from further consideration, i.e. 494762. Similarly, we considered only 20274 human protein entries out of 20334 sequences.



metabolic datasets archived in various databases, we compared SCL annotations assigned by LOCATE with that of the Gene Ontology (GO) assignment for major subcellular compartments: cytoplasm (GO:0005737), cytoplasmic vesicle (GO:0016023), extracellular (GO:0005576), endoplasmic reticulum (GO:0005783), endosomes (GO:0005767), Golgi apparatus (GO:0005794), lysosomes (GO:0005764), mitochondria (GO:0005739), nucleus (GO:0005634), plasma membrane (GO:0005886) and tight junction (GO:0005923). Our results provide an estimate of the reliability of SCL predictive ability of human proteins in the absence of sequence and structural features using the high-throughput protein interaction and metabolic dataset.

Results

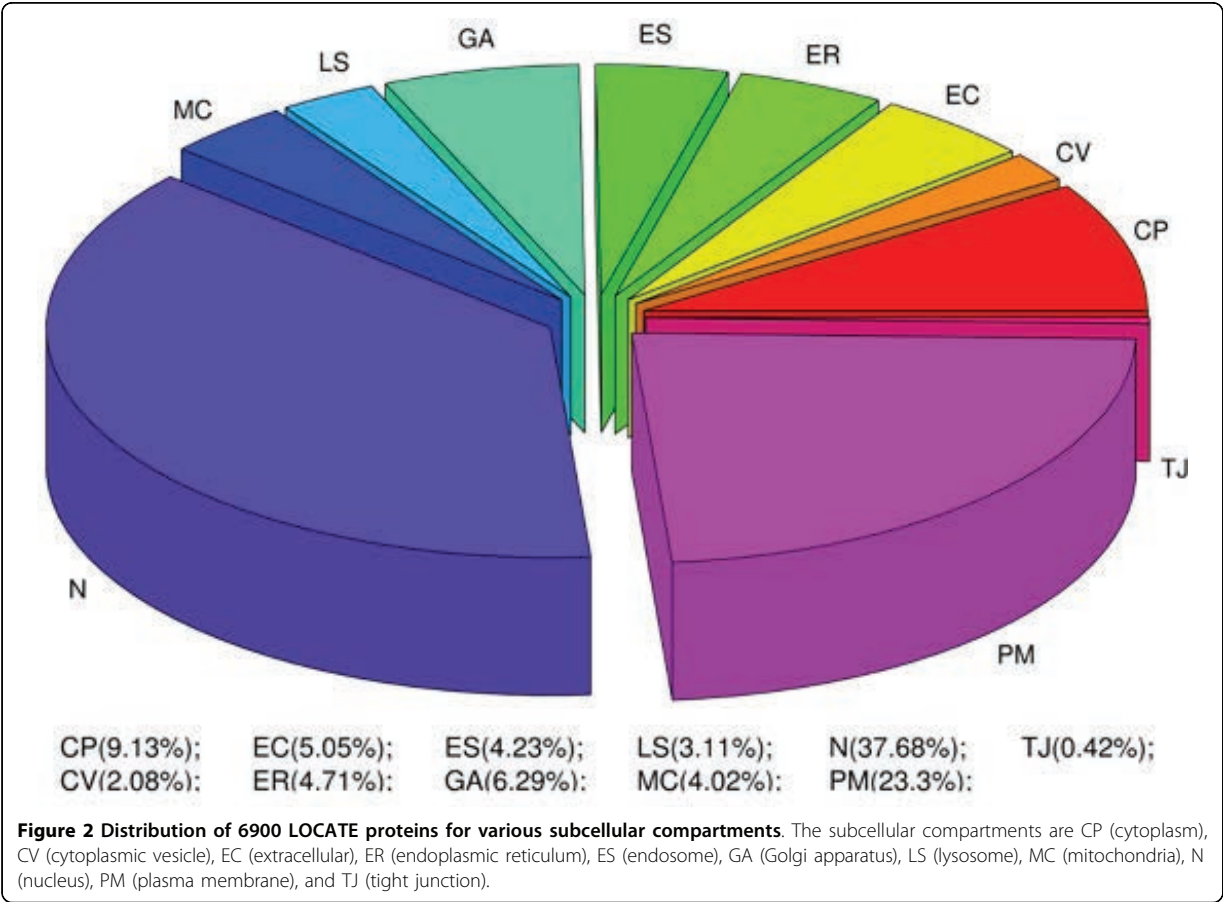
As there is no specific database which combines protein interaction, metabolic and SCL information, we integrated data from independent individual databases containing pertinent information. The SCL data from LOCATE [13], PPI data from five interaction databases and metabolic data from two databases (Figure 1; details in materials and methods section) were integrated. LOCATE contains literature-curated SCL information for about 6900 human proteins (Figure 2) in various subcellular compartments. The distribution of proteins is not homogeneous across the various subcellular compartments, with proteins from some compartments such as the nucleus and the plasma membrane being over-represented. Therefore, we have carefully normalized the dataset, while measuring the statistical properties of

our networks, to remove any bias toward specific SCL compartments.

Overall, 1,718 and 1036 proteins, respectively from the LOCATE dataset contain PPI and metabolic interactions. These reduced datasets were used for further analysis by considering the consistency of proteins across different databases and removal of the duplicate and redundant entries. For comparing the SCL assignment, we carefully merged low-level SCL annotation with that of the high-level SCL annotation mentioned in the GO hierarchy (see Additional file 1 for the merged GO-IDs). We used the same hierarchical level of SCL annotation for comparing LOCATE and GO annotations. Also, we will refer to the metabolite-linked protein interaction network as the metabolic network or MLPI, and the gene ontology annotation as GOA.

Categorical analysis of protein pairs

In order to test, how protein pairs are localized within the same subcellular compartments, Pearson's χ^2 (chi-square) test was performed. This statistical test shows that $\chi^2 = 1270.19$, $P\text{-value} < 2.2 \times 10^{-16}$ for physically interacting protein pairs and $\chi^2 = 110.02$, $P\text{-value} < 2.2 \times 10^{-16}$ for metabolite-linked protein pairs (Tables 2 and 3). Thus, the incorporation of PPI and metabolic data dramatically improve the significance of SCL prediction, while the confidence level in SCL predictions with PPI information is much higher than that with metabolic information. The contingency table for metabolic interaction revealed that the observed frequency of metabolite-linked protein pairs with the



same SCL is 20.94% more compared to the expected value, whereas the same observation seem to be twice as much (93.35%) for physically interacting protein pairs. The number of interacting protein pairs having the same or different SCL is observed to be nearly the same as in the PPI network. However, the metabolic network has fewer metabolite-linked protein pairs with the same SCL compared to that with different SCL. From Tables 2 and 3, we have extracted 4136 physically interacting protein pairs from 1156 proteins and 4551 metabolically linked pairs from 509 proteins for network analysis.

Interaction between various subcellular compartments
We measured the statistical significance of SCL correlation profile based on the Paired-Localisation Conditional Probability (PLCP; see Methods section for details), for both the LOCATE (manually curated from the literature) data as well as the GOA assigned SCL (excluding electronic annotation, which is automatically-assigned evidence code). Figure 3 shows significant correlation along the diagonals suggesting that the interacting protein pairs tend to co-localize in the same compartment. Comparing the LOCATE-assigned SCL (Figure 3A), we observe a strong correlation for physically interacting

Table 2 Chi-square test for physically interacting protein pairs.			
	Pairs with same SCL	Pairs with different SCL	Row total
Physical interaction present	2081 (1076.26)	2055 (3059.74)	4136
Physical interaction absent	381716 (382720.74)	1089051 (1088046.26)	1470767
Column total	383797	1091106	1474903
Chi-square (χ^2) Value: 1270.192	P-Value: < 2.2 × 10 ⁻¹⁶		
A 2 × 2 contingency table, showing the distribution of direct physical interaction of protein-pairs, as the observed number of pairs and the expected values (assuming independence) shown in parenthesis.			

Table 3 Chi-square test for the metabolite-linked protein pairs.

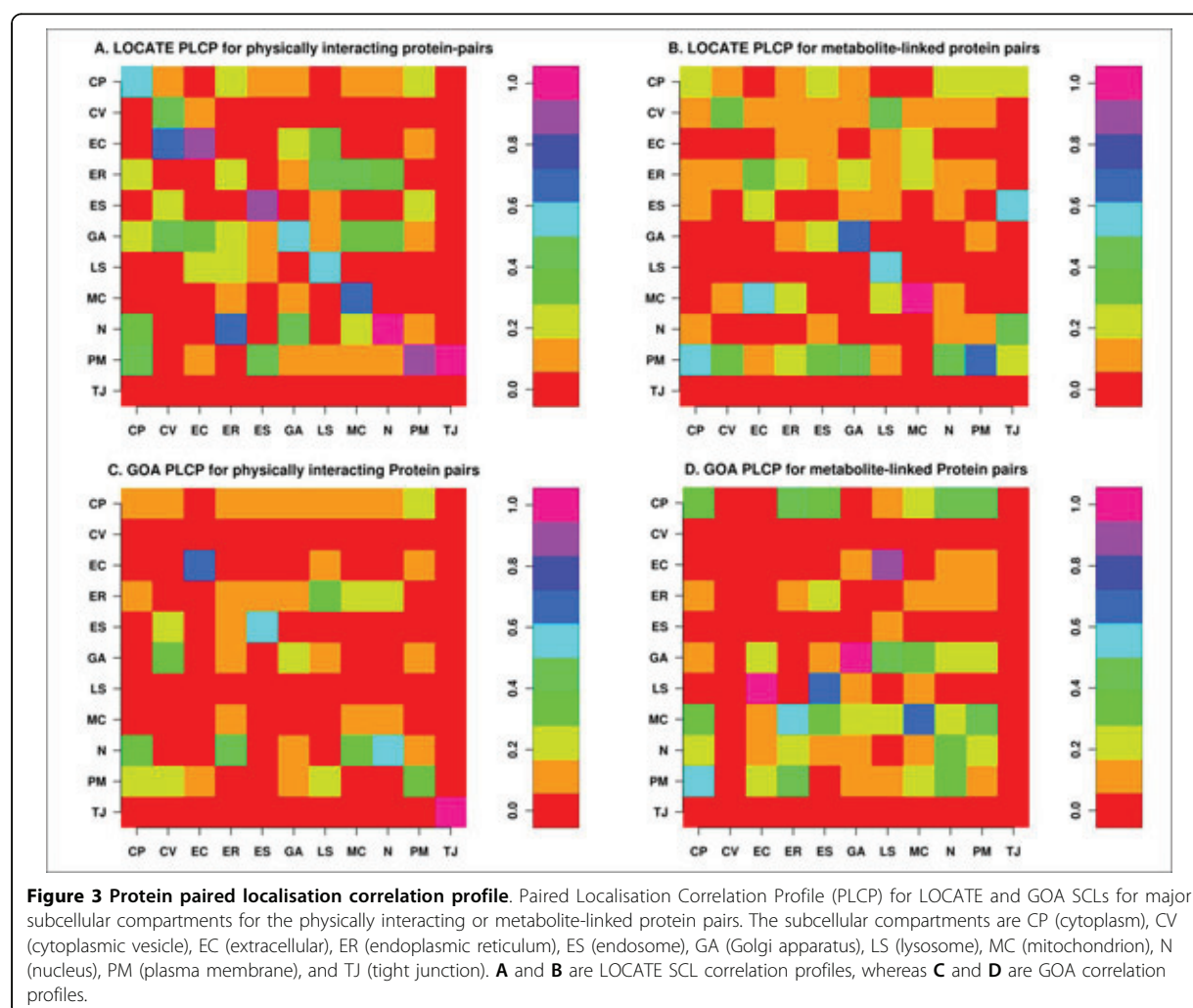
	Pairs with same SCL	Pairs with different SCL	Row total
Metabolite-linked Pairs	1465 (1158.12)	3086 (3392.88)	4551
Non-metabolite-linked Pairs	132345 (132651.88)	388929 (388622.12)	521274
Column total	133810	392015	525825
Chi-square (χ^2)- Value: 110.02	P-Value: < 2.2×10^{-16}		

A 2×2 contingency table, showing the distribution of metabolite-linked protein pairs, as the observed number of pairs and the expected values (assuming independence) in parenthesis.

protein pairs to occupy the same compartment in the cytoplasm (CP), cytoplasmic vesicles (CV), extracellular (EC), endosomes (ES), Golgi apparatus (GA), lysosome (LS), mitochondrion (MC), nucleus (N) and plasma membrane (PM). The same comparison on the GOA SCL (Figure 3C) shows conservation for EC, ES, GA, MC, N, PM and TJ. We also observed significantly strong correlation of nuclear proteins (Figures 3A and

3C) to interact with proteins found in cytoplasm, ER and Golgi for the LOCATE dataset and the cytoplasm, ER and mitochondrion for the GOA dataset. Similarly, plasma membrane proteins show significant interaction with the proteins in the several other subcellular compartments (Figures 3A and 3C).

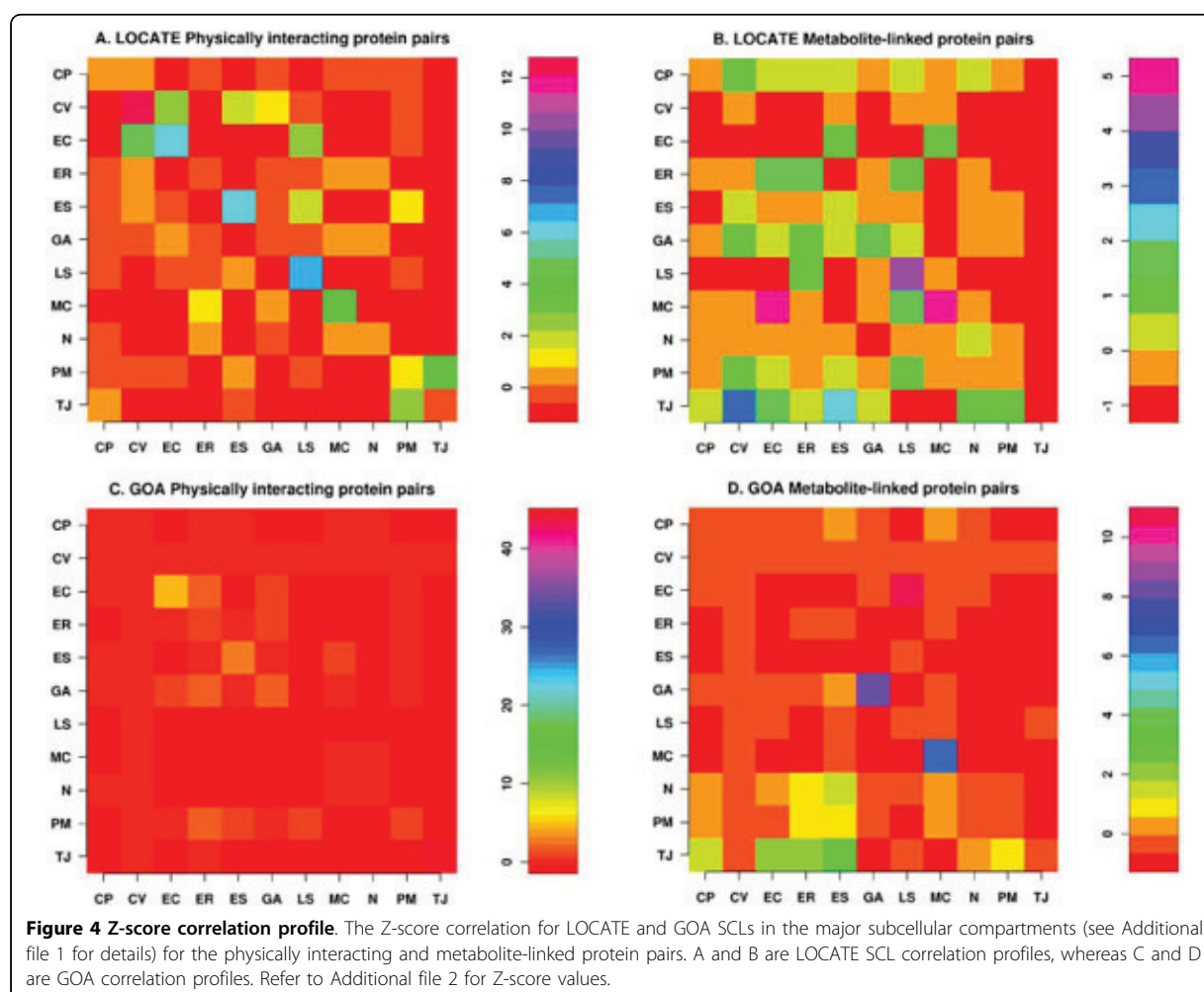
The MLPI profile shows strong correlation of interacting protein pairs to have same SCL for GA, LS and MC.



LOCATE data suggests significant correlation of metabolite-linked interaction of PM proteins with those in other compartments. Overall, the GOA dataset shows significant interaction across compartments in comparison to that of the LOCATE dataset (Figures 3B and 3D).

We further tested the hypothesis of whether the network of interacting protein pairs is different from a random network, by calculating the Z-score between the given compartments (described in the Methods section). The random network was simulated by rewiring the network such that the degree associated with each node in the real network remains the same [14]. The *P*-value can then be obtained by comparing the Z-score to a standard normal distribution. Comparing with a “properly” randomized network ensemble (1000 in our case) allows us to concentrate on those statistically significant localisation patterns of these complex interaction networks that are likely to reflect the conserved interaction pairs across different subcellular

compartments. The statistical significance of correlation profiles were calculated for PPI and metabolic networks for each paired compartments. The Z-score profile scales differently for the physically interacting and metabolite-linked protein pairs (Figure 4). The PPI network Z-score (Figures 4A, C) suggest that compared to random networks, the number of interacting protein pairs co-locating in the same compartment is significant for EC (*P*-value < 9.8 e-10), MC (*P*-value < 3.7 e-05), LS (*P*-value < 4.5 e-12), ES (*P*-value < 1.8 e-09) and CV (*P*-value < 1.9 e-35) for the LOCATE dataset (Figure 4A and Additional file 2). We also observed a significant correlation for CV proteins to interact with EC proteins (*P*-value < 5.4 e-06) but not otherwise i.e. EC proteins do not interact with CV proteins at a significant *P*-value < 0.01. Similarly, TJ proteins are more likely to interact with that of the PM proteins (*P*-value < 4.3e-05), whereas the likelihood of PM proteins to interact with TJ proteins is



less significant (P -value ~ 0.01). GOA SCL assignment (Figures 4C) suggests that statistically significant protein pair interactions occur within TJ (P -value ~ 0) and EC (P -value $< 1.36\text{e-}07$). Proteins pairs within the ES compartment seems to have a weak interaction (P -value ~ 0.0007). Similar weak interactions have been noticed between the proteins in the ER compartment with those of the GA (P -value ~ 0.007) (Additional File 2).

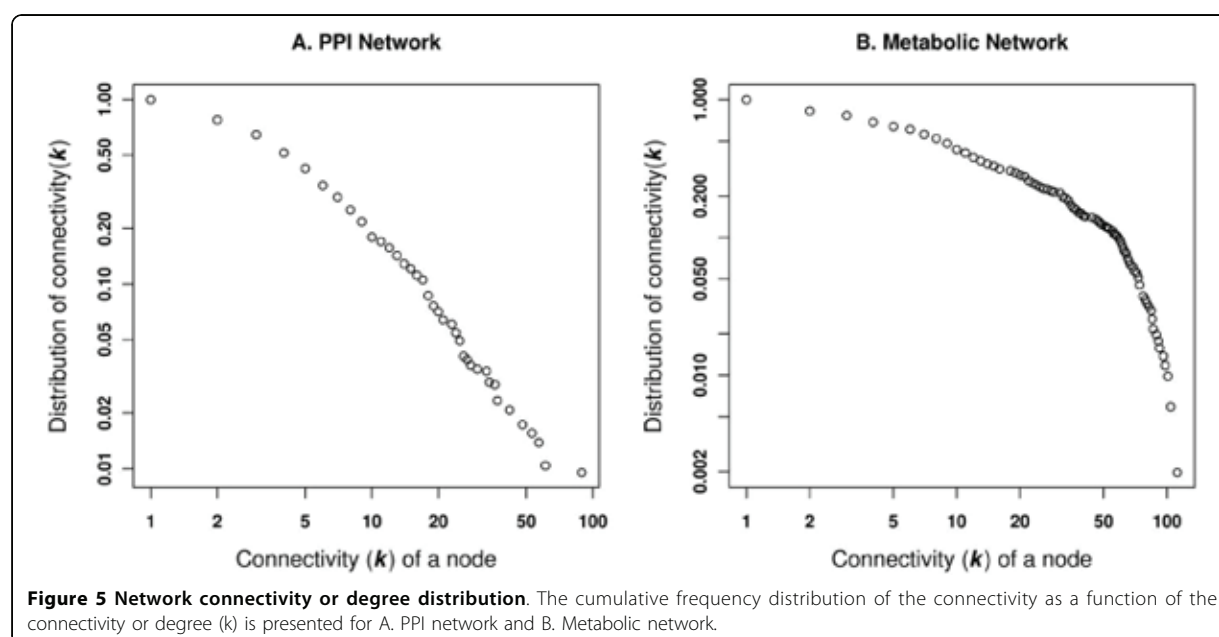
The metabolic Z-score correlation profile suggests a strong correlation of metabolite-linked protein pairs to have the same SCL within MC (P -value $< 6.0\text{e-}07$) and LS (P -value $< 4.7\text{e-}05$) in the LOCATE dataset (Figure 4B), while the GOA SCL (Figure 4D) assignment suggests the same for GA (P -value $< 1.0\text{e-}15$) and MC (P -value $< 1.3\text{e-}10$). A statistically significant proportion of EC proteins interacts with MC proteins (P -value $< 1.0\text{e-}05$) for the LOCATE SCL (Figure 4B). In the GOA dataset, LS proteins interact with EC proteins (P -value $< 1.1\text{e-}26$; Figures 4D). The detailed description of paired-compartment Z-scores and calculated P -values are available from Additional File 2.

Analysis of PPI and Metabolic Networks

To track the variation in structural topology between PPI and metabolic networks, we analyzed their topological properties of both the networks for human proteins in integrated dataset (Figure 1). The interaction network used in this study consists of 4136 direct physical interactions between 1156 human proteins (Table 2), whereas the metabolic network consists of 4551

interactions between 509 proteins (Table 3). This suggests that the metabolic network is denser with more edges between the protein nodes. Both the protein interaction network and the MLPI network belong to the class of scale-free networks, suggesting that both networks evolved by adding new nodes to existing highly connected nodes. In these networks, the number of nodes with a given number of neighbours (connectivity, K), scales as $P(K) \propto 1/K^\gamma$. The plot of the connectivity can be fitted by a power law, where $\gamma = 1.52$ and $\gamma = 1.34$, respectively for the physically interacting and metabolite-linked protein pairs (Figure 5A and 5B).

The connectivity probability of nodes and its nearest neighbours are the same compared to the connectivity of any of the nodes chosen randomly, in a random network. On the other hand, a real network comprises an ordered lattice which is extended as the network grows, i.e. some order is achieved depending on how the coordinates of each new node are added, with respect to that node's neighbours (clusters) and independent of the total number of nodes present in the network [15]. Therefore, we have calculated the average clustering coefficient ($\langle C_k \rangle$) associated with the given degree in PPI and metabolic networks, to study the global network topology. The PPI network shows random but gradual decrease of larger values of $\langle C_k \rangle$ associated with the high degree protein nodes. This simply means that the highly connected protein nodes are not connected, i.e. protein hubs are not connected, which is a specific signature for the non-modular nature of any real network (Figure 6A) [16]. The metabolic network, on the other



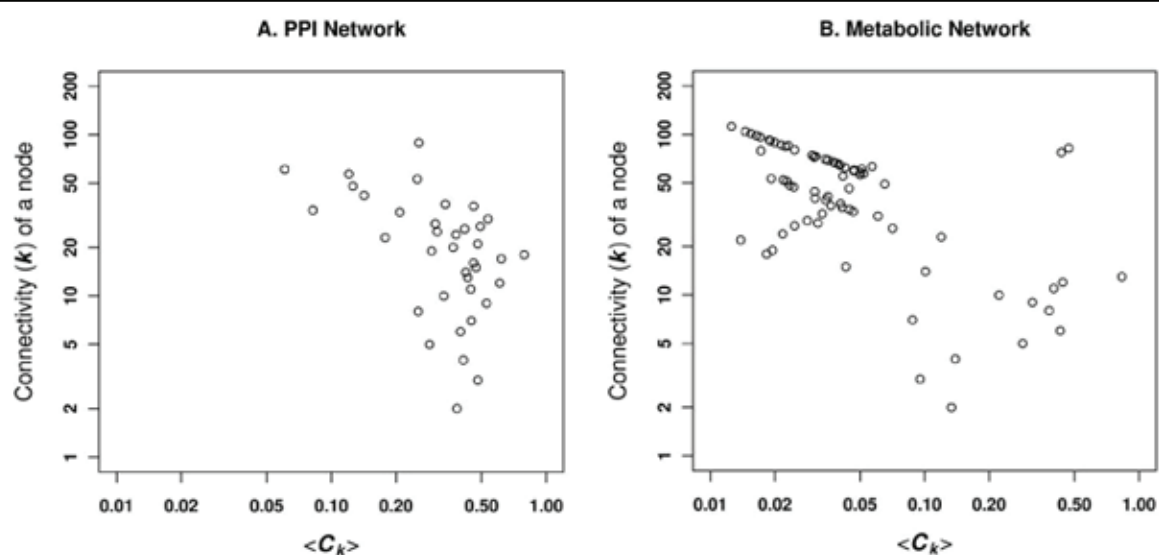


Figure 6 Average clustering against the node degree. The average clustering coefficient ($\langle C_k \rangle$) for each degree/connectivity, showing the probability that the adjacent neighbouring nodes of a node are connected is plotted as a function of the node degree in A. PPI network and B. Metabolic network.

hand, shows linear variation of highly connected nodes for the lower range of $\langle C_k \rangle$ associated with the higher degree nodes, implying the existence of hierarchical or modular structures (Figure 6B) [16,17].

Assortativity measures the collaboration of similar entities to achieve a single goal, whereas a disassortative nature suggests the association of different entities to achieve the same goal. Therefore, to observe the

assortative or disassortative nature of human PPI and metabolic networks, we calculated the average degree of the neighbouring proteins as a function of the each nodes degree [18]. For the PPI network, Figure 7A shows an increase in the neighbouring node degrees associated with higher degree nodes. This topological behaviour is the characteristic signature of the assortative network, thus suggesting that PPI is an assortative

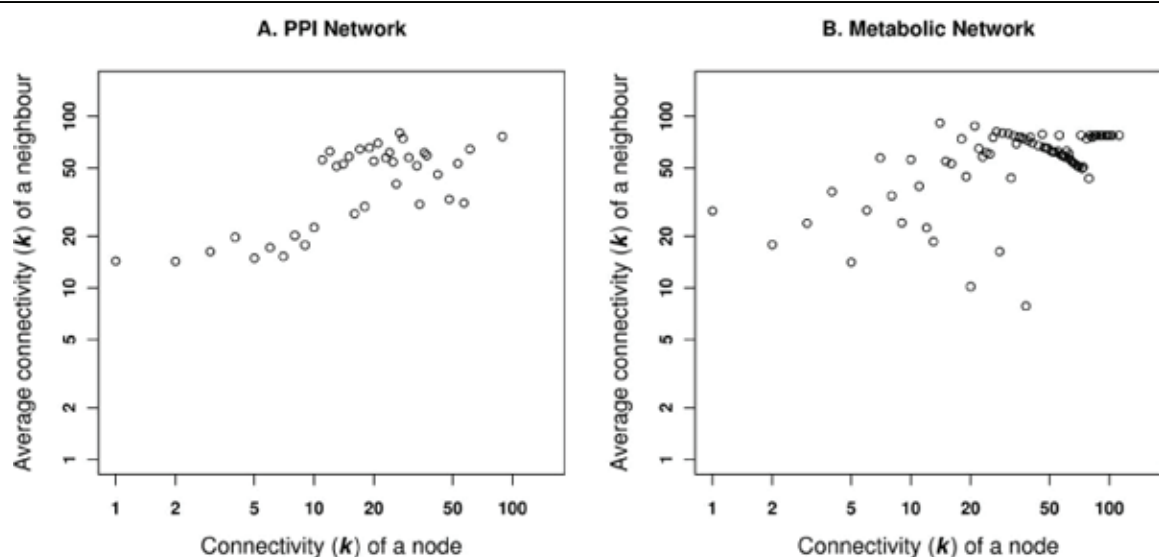


Figure 7 Average connectivity of a neighbouring nodes. Correlation in the connectivity of neighbours, with respect to a specific node of a given degree in A. PPI network and B. Metabolic network.

network. This observation is absent in the metabolic network (Figure 7B), where there is a decrease in the association with the high degree neighbours for the high degree nodes, i.e. nodes with the high degree k tend to be disconnected on an average, to others of lower degree. The power-law exponents (γ) for the degree assortativity are 1.2 and 1.1 in PPI and metabolic networks, respectively.

We have also calculated the betweenness centrality, to measure the load in our PPI and metabolic networks [19]. This measurement is commonly used in sociology to quantify the influence of a person in a society. In our case, it helps to quantify the information carrying capacity of a specific protein in the network. The PPI network shows a linear behaviour of the centrality measure associated with the connectivity of a node (k), whereas the metabolic network has a non-linear, random behaviour (Figure 8).

Figures 6 and 7 together indicate that the metabolic networks can be characterized with high degree nodes interconnecting highly connected subgraphs, but with no or few connections among nodes in different subgraphs. This implies that the metabolic pathways are inter-connected via substrates between different compartments. Table 4 provides data on other topological features of the networks.

Network-based neighbours for example proteins

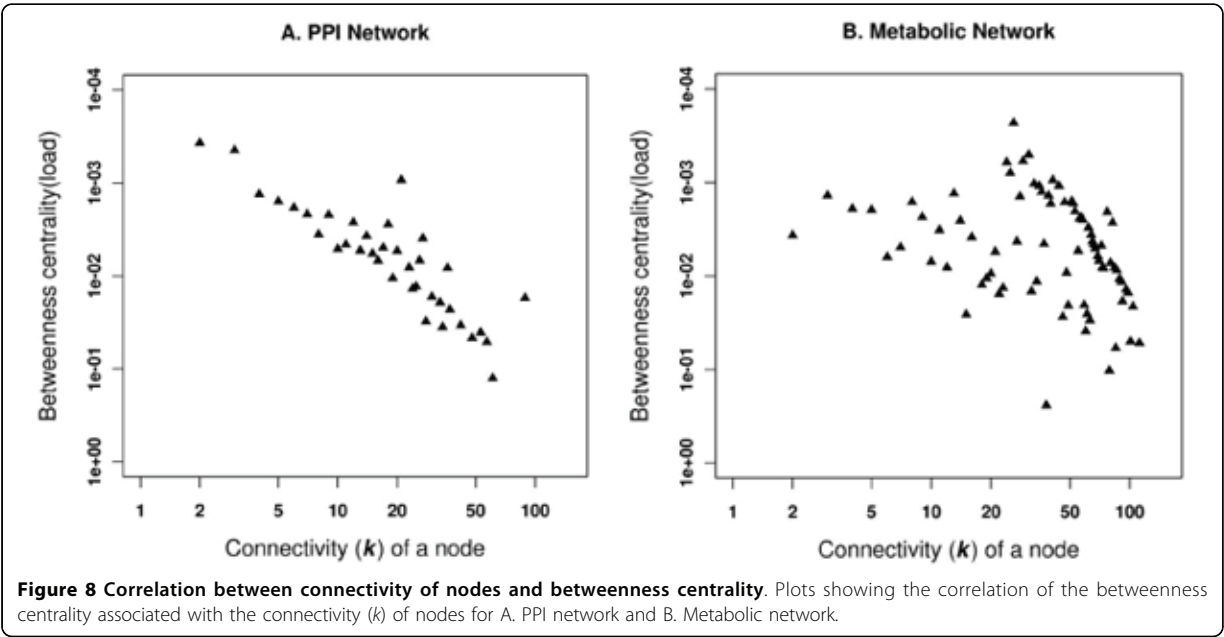
From the normalized datasets that we have studied, of the many biologically relevant proteins, we have presented two specific examples. The first example is of a

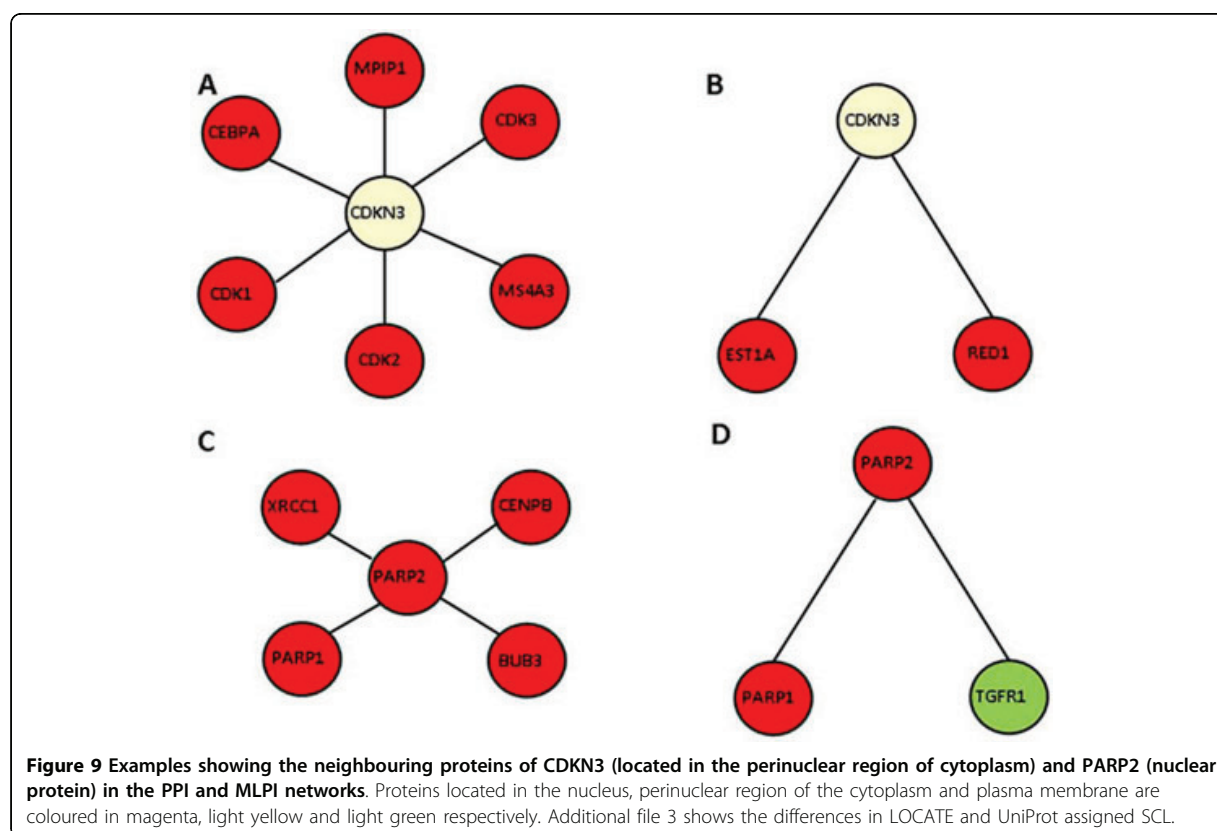
Table 4 Topological characteristics of PPI and metabolic networks.

	Protein interaction network	Metabolic network
Number of nodes	1156	509
Number of edges	4136	4551
Clustering coefficient	0.29	0.05
Average clustering coefficient	0.40	0.16
Average path length	4.77	4.09
Diameter	13	14

protein which specifically interacts with proteins co-located in the same SCL, while the second protein has interaction partners in different SCLs.

We examined the neighbouring proteins of human cyclin-dependent kinase inhibitor 3, CDKN3, in our PPI and MLPI networks (Figure 9). We note that this protein has been assigned the perinuclear region of the cytoplasm as SCL in UniProt, for a normal cell [20] (data available from Additional file 3). We found that CDKN3 is linked to double-stranded RNA-specific editease 1, RED1 and telomerase-binding protein, EST1A in our metabolic network, both interaction partners being located in the nucleus (Figure 9B). In the PPI network (Figure 9A), the same protein, CDKN3 is observed to interact with six proteins located in the nucleus: CDK2 (cell division protein kinase 2), MS4A3 (protein modulator of G1-phase to S-phase cell cycle transition), CDK3 (cell division protein kinase 3), MPIP1 (phosphatase protein inducer of mitotic





progression), CEBPA (DNA-binding protein) and CDK1 (cell division protein kinase 1, required for the progression of S-phase and mitosis). As early as 1993, Gyuris *et al.* [21] have reported that CDKN3 is expressed at the G1-phase to S-phase transition during the cell division process and is known to form a stable complex with CDK2. Our network analysis clearly supports CDKN3 being located in the periplasmic space and interacting with neighbouring proteins in the nucleus due to the porous nature of the nuclear membrane (Figure 9A and 9B) and is consistent with our PLCP analysis results on the interaction, which show that the nuclear proteins seem to interact with proteins of the cytoplasm (Figure 3).

Subsequently, we examined the neighbouring proteins of human poly [ADP-ribose] polymerase 2 (PARP2) (Figure 9C and 9D). In the MLPI (Figure 9D), one of the interacting partners of PARP2 is TGF-beta receptor type-1 (TGFRI), which is a signalling molecule located in the plasma membrane. The other interacting neighbour is PARP1 (poly [ADP-ribose] polymerase 1) located inside the nucleus, which interaction alone is preserved in the PPI network (Figure 9C). Considering the integrated network approach of combining different networks, we can thus infer not only the SCL of the

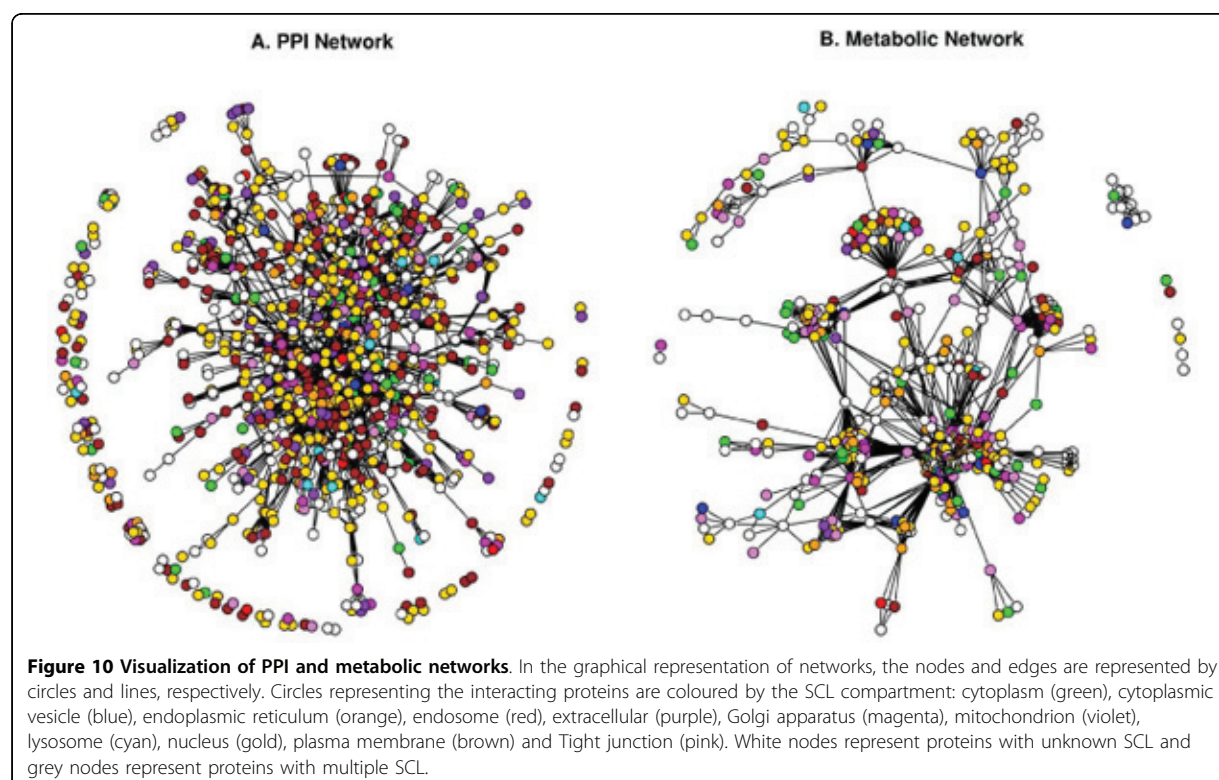
interacting proteins but also the biochemical signal *via* the plasma membrane, to identify the exact biological function of this polymerase, which is in accord with the earlier findings of Sharan and Ideker [22].

We have analyzed the SCL annotation of the 15 proteins in the above interacting pairs to determine the correlation of SCL assignment between LOCATE and UniProt databases (available in Additional file 3). We note that UniProt has no annotation for four proteins (27%), while two (13%) of the proteins have SCL assignments different from those in LOCATE. The remaining nine proteins have the same SCL assignments in both databases. These results support the use of experimentally determined SCL annotations from LOCATE for this analysis, over UniProt SCL assignments.

Discussion

Based on the topological comparison of networks, we were able to gain more insights into the structural differences in the PPI and metabolic networks of human proteins. Having shown that PPI and metabolic networks are scale-free, we further showed that the metabolic network is not assortative and modular (Figure 10).

The PPI network can be viewed as a network model where proteins collaborate on the number of cellular



processes a single protein can handle at any time. This network model is evident from network behaviour with a power-law distribution $P(k) \sim k^{-\gamma}$ where $\gamma = 1.5$ [23]. A similar observation is noted in the PPI network for passive interaction across subcellular compartments with $\gamma = 1.52$, due to the high false-positive rate. PPI data is known to have a high false-positive rate, i.e. the reliability of the possible observed interaction is questionable as with the high coverage rate. If a given protein interacts with a large number of other proteins, it is most likely a sticky protein and the observed interactions associated with this protein do not have a real functional association. Therefore, the passive interaction defines the unreliability of the observed interaction, which could happen by chance. The linear behaviour of betweenness centrality against the connectivity of node (k) in PPI network further suggests the presence of non-localized behaviour of interactions across compartments, compared to localized metabolite linkages among proteins inside the same subcellular compartments. This observation is also evident from the χ^2 statistics where the number of interacting protein pairs having the same localization is nearly the same as in different subcellular compartments (Table 2). We compared LOCATE assigned SCL with that of the GOA for the protein pairs across the different subcellular

compartments, considering the multiple localisation for proteins. This comparison suggests significant differences among the annotation process (Figure 3A and 3C). The correlation profile (PLCP) suggests a strong correlation of interacting protein pairs within the same subcellular compartments. There is statistically significant cross-interaction among proteins in the nucleus with those of other cellular compartments. This is attributed to the fact that the nucleus has a porous cell membrane, which facilitates free diffusion and interaction between proteins across compartments. Subcellular compartments such as the Golgi apparatus, the endoplasmic reticulum and the lysosome indicate weak but significant correlation, which is in accord with the fact that the Golgi apparatus and the endoplasmic reticulum are inter-linked subcellular compartments for the translocation of proteins to various other compartments after the translation of mRNA to protein on the ribosome. The Z-score correlation profile for the PPI network shows that while interactions are conserved within compartments (along the diagonal, Figure 4A and 4C) with respect to the random network, there is also significant interaction of protein pairs across other subcellular compartments.

The metabolic network has an evolutionary constraint where only a few proteins are linked through common

metabolites to maintain high substrate specificity in the higher eukaryotes [24]. Hence proteins are distributed in various subcellular compartments unlike prokaryotic proteins which contain co-evolving protein domains to carry out multiple tasks. Moreover, eukaryotic metabolic pathways are optimized via cross connections across subcellular compartments. This is revealed in the χ^2 statistics where few protein pairs have the same subcellular compartments compared with pairs from different compartments. PLCP suggest that protein pairs are not conserved for the compartments such as cytoplasm, cytoplasmic vesicles, endoplasmic reticulum and endosome (Figure 3B and 3D). This is due to the fact that the numbers of metabolite-linked protein-pairs are less and secondly, there are lots of dynamics happens among these compartments, as number of cellular pathway are distributed across compartments, hence it makes difficult to capture from our static picture of PLCP calculation. Even though the dynamics of some compartments are difficult to capture through the statistical measures, it is very useful to see how cellular processes are tightly controlled inside the subcellular systems such as mitochondrion and lysosome. The Z-score correlation profile of LOCATE and GOA SCL suggests that the metabolite-linked protein pairs seems to be more conserved across diagonals compare to that of randomized network and hence metabolite-linked interactions are tightly regulated within the same compartments (Figure 4B and 4D).

Conclusions

The network analysis showed that there is significant difference between the topological properties measured in the human PPI and metabolic networks. Network comparison indicates the usefulness of metabolite-linked protein interaction (metabolic network) that can be used for the prediction of protein's SCL in the compartments such as mitochondria and lysosome. Our results lead to the observation that proteins in PPI network interact passively, whereas metabolic network evolve under evolutionary constrain to maintain substrate specificity. The series of analysis presented in this study suggests the applicability of metabolic (metabolite-linked protein interaction) network to explain the empirical data. The integrated network approach of using PPI and MLPI data developed here will provide a robust basis for predicting SCL for higher eukaryotes, along with the comparative network studies across species.

Methods

Data integration and construction of database

In the absence of a specialized database combining protein interaction, metabolic and SCL information, we have integrated data from independent individual

databases. The LOCATE database contains SCL information from human and mouse proteins collected from both literature and direct experiment [13]. SCL data on human proteins from LOCATE database were integrated with the interaction data deposited in the PPI databases: HPRD [25], DIP [26], MINT [27], BioGRID [28] and IntAct [29]. Similarly, metabolic data (MD) were collected from the databases, KEGG [30] and HumanCyc [31] and integrated with the SCL data of the human proteins with the LOCATE database. This integrated dataset is recorded in XML format (Figure 1 and Additional file 4). LOCATE data contains 64,637 human proteins with known or predicted SCL information. Our integrated database contains 6,900 proteins with known SCL information curated from the literature (Figure 2). We used UniProt-ids and RefSeq-ids for consistent mapping across the three different datasets (i.e. SCL, PPI and MD).

Identification and removal of inconsistency and redundancy

The LOCATE protein database [13] contains references to sequence databases such as UniProtKB [2] and RefSeq [32]. Protein entries with secondary accession were mapped to their primary identifiers mentioned in the protein sequence databases. RefSeq identifiers were used to extract UniProt identifiers where LOCATE entries contain RefSeq identifier but not the UniProt accession number. This allows consistent one-to-one mapping of protein entries across various databases. Duplicate entries of known protein interactions mentioned in PPI databases were carefully removed while analyzing interaction information in each LOCATE entry.

The metabolic linkage between proteins was established by considering only those compounds which occur in less than 50 reactions per compound in a given metabolic database. This ensures the removal of ubiquitous compounds such as ATP, NADH, H₂O, H⁺ etc. (see Additional files 5 and 6 for the lists of ubiquitous compounds). Ambiguous metabolites were removed, for example, HumanCyc reaction: GLUTATHION + RX <=> [S-Substituted-Glutathione] + HX, where RX and HX are ambiguous metabolites. Only those metabolites which contain unique compound-ids, were further considered for linking proteins, while those with generalized descriptions were omitted. E.g. General-Protein-Substrates and General-Phos-Protein-Substrates were not considered as linking metabolites shown in a reaction: |**General-Protein-Substrates**| + ATP <=> |**General-Phos-Protein-Substrates**|.

For the current study 1,718 and 1036 LOCATE proteins out of 6900 (literature curated), were linked *via* direct physical and metabolite-linked protein

interactions, respectively. In the topological studies of PPI and metabolic networks, we considered 1156 and 509 proteins with 4136 and 4551 interactions respectively.

Construction of networks

All LOCATE protein entries were linked *via* interactions (either physical or through a common metabolite) and the data were recorded in xml format (available from Additional file 4). This dataset was used to build the undirected networks using the R igraph package [33]. We used *degree* and *transitivity* functions for calculating the degree distribution and clustering coefficient in our networks. Random networks were generated by using the *rewire* function of the R igraph package.

SCL analysis of the protein pairs

Correlation profiles were created using Paired-Localisation Conditional Probability (PLCP) for both PPI and metabolic networks [9]. This measure shows how the interacting protein pairs are distributed across various subcellular compartments. For a given protein in the compartment C_i having an interacting partner in compartment C_j , PLCP is defined as

$$P(C_i | C_j) = \frac{C_{ij}}{\sum_k C_{jk}}, \quad (1)$$

where C_{ij} is the normalized number of interactions between protein pairs spanning compartments C_i and C_j . C_{ij} is defined as:

$$C_{ij} = \frac{\sum_{x \in C_i, y \in C_j (x \neq y)} \lambda(x, y)}{N(C_i) + N(C_j)} \quad (2)$$

where, $\lambda(x, y)$ is 1 if there is an interaction between proteins x and y , otherwise, 0. $N(C_i)$ is the number of proteins in compartment C_i and $N(x)$ is the number of localisations known for protein x .

The Z-score correlation profiles were analyzed between interacting protein pairs from the real and random networks as given by:

$$Z(C_i, C_j) = \frac{N(C_i, C_j)_{real} - \langle N(C_i, C_j)_{random} \rangle}{\sigma(C_i, C_j)_{random}} \quad (3)$$

where, $N(C_i, C_j)_{real}$ and $\langle N(C_i, C_j)_{random} \rangle$ represent numbers of physically interacting or metabolite-linked protein pairs in real and random networks respectively. $\sigma(C_i, C_j)_{random}$ represents the standard deviation in the ensemble of a 1000 random networks.

Statistical validation of networks

We analyzed the topological property of PPI and metabolic network calculating the most significant network features, namely clustering coefficient, betweenness centrality, average path length, degree distribution and correlation profile calculation. For a graph G with u and v as two vertices, the path from u to v will pass sequentially through vertices v_1, v_2, \dots, v_k , with $u = v_1$ and $v = v_k$, such that for $i = 1, 2, \dots, k-1$: (i) $(v_i, v_{i+1}) \in E(G)$ i.e. the edges set and (ii) $v_i \neq v_j$ for $i \neq j$. The path length is then said to be $(k-1)$. The simple *geodesic distance*, $d(u, v)$ from u to v is the length of the shortest path from u to v in the graph G . The average path length, $\langle l \rangle$, of such a graph is defined as the average of values taken over all the possible pairs of nodes connected by at least one path:

$$\langle l \rangle = \frac{2}{N(N-1)} \sum_{u,v=1}^N l_{uv} \quad (4)$$

where, N is the number of nodes and l_{uv} is the distance between two nodes, u and v . The diameter of the network is defined as the maximum distance between two nodes of a graph G , i.e. $D = \max\{d_{uv} | u, v \in N\}$, where N is the total number of nodes in the graph or network.

The clustering coefficient is another characteristic of a network which is unrelated to the degree distribution. It is a quantitative measure to the proximity of the neighbourhood of each node to form a complete subgraph (clique) and thus defines a measure of the local behaviour of the small world network [34]. The clustering coefficient is defined as,

$$C_i = \frac{2K}{k_i(k_i-1)} \quad (5)$$

where, K denotes the sum of the neighbouring pairs among the k_i nodes connected to the node i . Similarly, one can define an average clustering coefficient as,

$$\langle C \rangle = \frac{1}{K} \sum_{i=1}^K C_i \quad (6)$$

Centrality is one of the key structural aspects of the nodes in a network and is a measure of the relative influence of each node on the network. We calculated betweenness centrality, which is the fraction of shortest paths between all the pairs of nodes that passes through a given node [19].

Additional file 1: Merged list of subcellular compartments for the LOCATE and GOA SCL. This contains the list of compartment at the lower-level of GO hierarchy which were merged with that of the higher level of GO cellular compartments for the analysis of major subcellular compartments.

Additional file 2: List of Z-score values for the paired SCL. This contains the Z-score values and their calculated P-values for the paired compartments in the PPI and metabolic dataset, as described in Figure 3.

Additional file 3: SCL assignment of example proteins in Figure 9. The LOCATE SCL information compared to SCL annotations from the UniProt database. For each protein, the description, HGNC gene name and UniProt identifier are also provided.

Additional file 4: Integrated data. This contains the LOCATE proteins with SCL information integrated with that of the PPI and metabolic dataset, as described in Figure 1.

Additional file 5: List of KEGG compounds per reaction. A list of compounds from the KEGG database [30] with the number of known reaction.

Additional file 6: List of HumanCyc compounds per reaction. A list of compounds from the HumanCyc database [31] with the number of known reactions.

Acknowledgements

This research was supported by Macquarie University Research Scholarship (MQRES) to GK and the ARC Centre of Excellence in Bioinformatics grant (CE0348221) to SR. We thank Dr. Adrian P Cootes and Dr. Antonio Reverter for valuable discussions and for their constructive comments on the statistical analysis. Dr. Rohan Teasdale for providing LOCATE database. This article has been published as part of BMC Bioinformatics Volume 11 Supplement 7, 2010: Ninth International Conference on Bioinformatics (InCoB2010): Bioinformatics. The full contents of the supplement are available online at <http://www.biomedcentral.com/1471-2105/11?issue=S7>.

Author details

¹ARC Centre of Excellence in Bioinformatics and Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney NSW, Australia.

²Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore.

Authors' contributions

GK designed the experiment, analysed the data and wrote the first draft of the manuscript. SR directed this study and finalized the manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 15 October 2010

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology.** The Gene Ontology Consortium. *Nat Genet* 2000, **25**(1):25-29.
2. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E: **Infrastructure for the life sciences: design and implementation of the UniProt website.** *BMC Bioinformatics* 2009, **10**:136.
3. Nakai K, Horton P: **Computational prediction of subcellular localization.** *Methods Mol Biol* 2007, **390**:429-466.
4. Nair R, Rost B: **Protein subcellular localization prediction using artificial intelligence technology.** *Methods Mol Biol* 2008, **484**:435-463.
5. Shin CJ, Wong S, Davis MJ, Ragan MA: **Protein-protein interaction as a predictor of subcellular location.** *BMC Syst Biol* 2009, **3**:28.
6. Scott MS, Calafell SJ, Thomas DY, Hallett MT: **Refining protein subcellular localization.** *PLoS Comput Biol* 2005, **1**(6):e66.
7. Schwikowski B, Uetz P, Fields S: **A network of protein-protein interactions in yeast.** *Nat Biotechnol* 2000, **18**(12):1257-1261.
8. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, et al: **Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets.** *Nat Genet* 2006, **38**(3):285-293.
9. Lee K, Chuang HY, Beyer A, Sung MK, Huh WK, Lee B, Ideker T: **Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species.** *Nucleic Acids Res* 2008, **36**(20):e136.
10. Morowitz HJ: **A theory of biochemical organization, metabolic pathways and evolution.** *Complexity* 1999, **4**:39-53.
11. Wagner A, Fell DA: **The small world inside large metabolic networks.** *Proc Biol Sci* 2001, **268**(1478):1803-1810.
12. Mintz-Oron S, Aharoni A, Ruppin E, Shlomi T: **Network-based prediction of metabolic enzymes' subcellular localization.** *Bioinformatics* 2009, **25**(12):i247-252.
13. Sprenger J, Lynn Fink J, Karunaratne S, Hanson K, Hamilton NA, Teasdale RD: **LOCATE: a mammalian protein subcellular localization database.** *Nucleic Acids Res* 2008, **36** Database: D230-233.
14. Maslov S, Sneppen K: **Specificity and stability in topology of protein networks.** *Science* 2002, **296**(5569):910-913.
15. Albert R, Barabasi AL: **Statistical mechanics of complex networks.** *Rev Mod Phys* 2002, **74**(1):47-97.
16. Soffer SN, Vazquez A: **Network clustering coefficient without degree-correlation biases.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2005, **71**(5 Pt 2):057101.
17. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297**(5586):1551-1555.
18. Newman MEJ, Park J: **Why social networks are different from other types of networks.** *Physical Review E* 2003, **68**(3):036122.
19. Goh KI, Oh E, Jeong H, Kahng B, Kim D: **Classification of scale-free networks.** *Proc Natl Acad Sci USA* 2002, **99**(20):12583-12588.
20. Lee SW, Reimer CL, Fang L, Iruela-Arispe ML, Aaronson SA: **Overexpression of kinase-associated phosphatase (KAP) in breast and prostate cancer and inhibition of the transformed phenotype by antisense KAP expression.** *Mol Cell Biol* 2000, **20**(5):1723-1732.
21. Gyuris J, Golemis E, Chertkov H, Brent R: **Cdi1, a human G1 and S phase protein phosphatase that associates with Cdk2.** *Cell* 1993, **75**(4):791-803.
22. Sharan R, Ideker T: **Modeling cellular machinery through biological network comparison.** *Nat Biotechnol* 2006, **24**(4):427-433.
23. Vazquez A, Oliveira JG, Dezzo Z, Goh KI, Kondor I, Barabasi AL: **Modeling bursts and heavy tails in human dynamics.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2006, **73**(3 Pt 2):036127.
24. Zhu D, Qin ZS: **Structural comparison of metabolic networks in selected single cell organisms.** *BMC Bioinformatics* 2005, **6**:8.
25. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, et al: **Human protein reference database-2006 update.** *Nucleic Acids Res* 2006, **34** Database: D411-414.
26. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Res* 2004, **32** Database: D449-451.
27. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular Interaction database.** *Nucleic Acids Res* 2007, **35** Database: D572-574.
28. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bahler J, Wood V, et al: **The BioGRID Interaction Database: 2008 update.** *Nucleic Acids Res* 2008, **36** Database: D637-640.
29. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, et al: **IntAct-open source resource for molecular interaction data.** *Nucleic Acids Res* 2007, **35** Database: D561-565.
30. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic Acids Res* 2004, **32** Database: D277-280.
31. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD: **Computational prediction of human metabolic pathways from the complete human genome.** *Genome Biol* 2005, **6**(1):R2.
32. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, **35** Database: D61-65.

33. Csárdi, G, Nepusz, T: **The igraph software package for complex network research.** *InterJournal* 2006, *Complex Systems*.
34. Watts DJ, Strogatz SH: **Collective dynamics of 'small-world' networks.** *Nature* 1998, **393**(6684):440-442.

doi:10.1186/1471-2105-11-S7-S9

Cite this article as: Kumar and Ranganathan: Network analysis of human protein location. *BMC Bioinformatics* 2010 **11**(Suppl 7):S9.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



Additional files for this publication are available from the journal's website:

Additional file 1 – Merged list of subcellular compartments for the LOCATE and GOA SCL (*.xls).

This contains the list of compartment at the lower-level of GO hierarchy which were merged with that of the higher level of GO cellular compartments for the analysis of major subcellular compartments

(<http://www.biomedcentral.com/content/supplementary/1471-2105-11-s7-s9-s1.xls>)

Additional file 2– List of Z-score values for the paired SCL (*.xls)

This contains the Z-score values and their calculated P-values for the paired compartments in the PPI and metabolic dataset, as described in Figure 3.

(<http://www.biomedcentral.com/content/supplementary/1471-2105-11-s7-s9-s2.xls>)

Additional file 3 – SCL assignment of example proteins in Figure 9 (*.doc)

The LOCATE SCL information compared to SCL annotations from the UniProt database.

(<http://www.biomedcentral.com/content/supplementary/1471-2105-11-s7-s9-s3.doc>)

Additional file 4 – Integrated data (*.xml)

This contains the LOCATE proteins with SCL information integrated with that of the PPI and metabolic dataset, as described in Figure 1.

(<http://www.biomedcentral.com/content/supplementary/1471-2105-11-s7-s9-s4.xml>)

Additional file 5– List of KEGG compounds per reaction (*.xls)

A list of compounds from the KEGG database [92] with the number of known reaction.

(<http://www.biomedcentral.com/content/supplementary/1471-2105-11-s7-s9-s5.xls>)

Additional file 6 – List of HumanCyc compounds per reaction (*.xls)

A list of compounds from the HumanCyc database [66] with the number of known reactions.

(<http://www.biomedcentral.com/content/supplementary/1471-2105-11-s7-s9-s6.xls>)

3.2 Conclusions

Statistical analysis using PPI and metabolic networks suggest that the tendency of interaction protein pairs to have same SCL is twice for the physically interacting protein pairs. However, the tendency of functional linkage of proteins in metabolic network has only 20% more than expected. Protein co-localisation correlation profile suggests that proteins annotated with experimentally determine SCL deposited in the LOCATE database show much better localisation of interacting protein pairs within same SCL compare to GO annotation. Moreover, this measure also suggests a significant cross-talk between SCL compartments such as plasma membrane and nucleus among physically interacting proteins. However, metabolite-linked protein interaction or functional linkage has significant cross-talk among various SCL compartments both of LOCATE and GO annotated proteins. Z-distribution or standard normal distribution suggests that functional linkage of proteins in metabolic network has statistically significant tendency to be in compartments such as mitochondrion, lysosome and Golgi apparatus. Moreover, the significant cross-interaction between nucleus and other compartments is due to a porous cell membrane, which facilitates free diffusion [93]. Thus, it highlights the importance of metabolic network in addition to PPI network in the prediction of SCL.

Chapter 4: Dissecting the organisation of human and yeast interactomes: network relationships from biological processes and molecular functions

4.1 Summary

Protein-protein interaction network data has enabled researchers to understand how proteins with given molecular functions or biological processes organize to enact the complex behaviour of cellular diversity [94, 95]. Conversely, interaction networks have allowed the molecular functions of uncharacterized proteins to be inferred from their relationships with other proteins [96]. Experimental work studying the co-expression of hubs has suggested that they fall into two main categories called “date” and “party” hubs [97]. Date hubs are those that tend not to be co-expressed with their interacting partners and are thought to act as communicators between different functional modules. Party hubs are those that do tend to be co-expressed with interacting partners and are thought to act as part of a single functional module. Further work has challenged this interpretation [98] and resolving this issue will be important, not only for understanding functional organisation in the cell, but also for more effectively using network information to determine function of individual proteins. This study does not aim to add to this debate directly, but rather determine how proteins of differing degree influence functional/process relationships within the network generally. Instead of classifying protein degree in a binary fashion (hubs and non-hubs), we consider the behaviour of proteins over a large number of degree categories.

Dissecting the organisation of human and yeast interactomes: network relationships from biological processes and molecular functions

Gaurav Kumar^{1*}, Adrian P. Cootes^{1*} and Shoba Ranganathan^{1,2 §}

¹ARC Centre of Excellence in Bioinformatics and Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney NSW 2109, Australia.

²Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597.

*These authors contributed equally to this work

§Corresponding author

Email addresses:

GK: gaurav.kumar@mq.edu.au

APC: adrian.cootes@yahoo.com

SR: shoba.ranganathan@mq.edu.au

Abstract

Background

The tendency of proximal proteins in interaction networks to have the same molecular function and biological process is often exploited to predict cellular functions. Highly connected interaction hubs have been shown to be important in the functional organisation of the cell. However, a detailed study of the influence of protein degree on determining biological processes and molecular functional similarity in proximal proteins has not been performed to date.

Results

Our analysis examines the tendency of proximal proteins in high-confidence *H. sapiens* and *S. cerevisiae* protein-protein interaction networks. We demonstrate that the tendency of proximal network proteins to have the same functions and processes depends strongly on their degrees and the degrees of the proteins separating them. Removal of the highest interacting proteins, i.e. hubs, decreases the tendency of interacting proximal proteins to have the same gene ontology terms for molecular function and biological process in higher proportion. This is consistent with the view that many hubs interact with different functional modules. Furthermore, we showed that the types of paths connecting proteins with the same function and process differ with the level of GOA (gene ontology annotation) hierarchy being examined in this study at the different level of abstraction.

Conclusions

Proximal proteins are more likely to have same molecular functions and biological processes for the network distances 1, 2 and 3 at GOA level 2 and 3, respectively. The molecular functions and biological processes of uncharacterised proteins can thus be inferred from their relationships with other proteins in the interaction network, upto GOA levels 2 and 3, respectively. Moreover, the tendency of interacting protein pairs in a network to have the same functions and processes decreases with increasing network

distance *via* the shortest connecting path passing through hubs. Our network analysis results extend and complement the existing knowledge of the interactome.

Background

Protein-protein interaction (PPI) networks provide a wealth of new insights into the organisation of proteins in biomolecular systems. PPI network data has enabled researchers to understand how proteins with a given molecular function or biological process, are organised into teams, leading to the complex behaviour of cellular diversity. Conversely, interaction networks have allowed the molecular functions of uncharacterised proteins to be inferred from their relationships with other proteins. For example, 76% [1] and 52% [2] of interacting protein pairs in yeast and human, respectively, have been reported to show the same subcellular localisation. A network-based method has consequently been developed to predict protein function independently of sequence [3].

Subsequent studies of functional organisation in networks have focused on complex network features involving many interactions acting in concert. In particular, network clustering often occurs when proteins form part of a functional module or protein complex [4]. Thus, proteins of unknown function in a highly clustered region of the network are likely to have the same molecular function as other proteins in the cluster. Many protein function prediction techniques explicitly search for significant local clustering to infer functional relationships between highly interconnected proteins. Similarly, studies of signalling/pathway networks suggest that proteins work in collaboration to achieve common biological processes [5, 6].

There has been much interest in the scale-free nature of interaction networks and how this affects functional organisation [7]. Highly-connected proteins in the network (known as hubs) are relatively rare and have been suggested to play special functional roles. Experimental work studying the co-expression of hubs has suggested that they fall into two main categories called “date” and “party” hubs [8]. Date hubs are those that tend not to be

co-expressed with their interacting partners and are thought to act as communicators between different functional modules. Party hubs are those that do tend to be co-expressed with interacting partners and are thought to act as part of a single functional module. Further work has challenged this interpretation [9]. Resolving this issue will be important, not only for understanding functional organisation inside the cell, but also for more effectively using network information to determine the function of individual proteins. This study does not aim to add to this debate directly, but rather to determine how proteins of differing degree (i.e. proteins node connectivity with their neighbouring proteins) influence functional/process relationships within the network in general. Instead of classifying proteins based on their network degree in a binary fashion (hubs and non-hubs), we consider the behaviour of proteins over a larger number of degree categories.

In addition, a protein's functional role can be described at various levels of abstraction. For example, a protein might be known to be a type of kinase or, in more detail, known to be a tyrosine kinase. The successful inference of an unknown protein's functional role from its network context may depend on the level of gene ontology (GO) detail at which the function of its neighbours is known. Moreover, a recent study suggests that the central interactome acts as a platform characterized by biological process, to exchange information through protein interactions [10]. Functional annotation of a gene or protein from GO basically characterizes three parts: cellular component, biological process and molecular function. The three-fold characterization of GO terms has different stages or levels of definition (see Figure 1 for an example). We have previously focussed on the network analysis of human protein location [11], comparing the subcellular localization definitions from gene ontology annotation (GOA) with experimentally determined locations. This study has now been extended to explore the ability of network relationships to infer the extent of conservation of molecular function and biological process at different levels of GOA.

In this study, we analysed the influence of protein degree on GOA biological process and molecular functional relationships in the *Saccharomyces cerevisiae* and *Homo sapiens* protein interaction networks or interactomes. We examined the dependence of functional and process similarity of protein pairs on the connectivity properties of the shortest path linking them. We also investigated how the functional inference from protein interactions depends on the GOA level of detail at which function or process is being studied. Here, we have shown the tendency of interacting protein pairs to have same molecular function and biological processes over network distances of 1, 2 and 3 is more likely compared to random chance at GOA level 2 and 3, respectively. Removal of hubs from the network over this range of protein degree suggests a decrease in the tendency of interacting protein pairs to have the same molecular function for the network distances 1, 2 and 3 at different GOA levels. This observation holds for biological processes as well, although the decrease is less prominent. Moreover, interacting protein pairs through the shortest path via hubs suggests a decrease in the percentage of common neighbours having the same GOA molecular function and biological process, at network distances 2 and 3, compared to hubless interaction, reconfirming that highly connected nodes or hubs play the crucial role of linking different function or process modules inside the cell.

Methods

Construction of the Protein Interaction Networks

Protein interactions were downloaded from BIND, BioGrid, DIP, HPRD, MINT, IntAct and MIPS databases [12-18]. HPRD and MIPS datasets were considered exclusively for human and yeast, respectively. High-confidence (HC) human protein interactions were created by merging interactions from the above mentioned databases and by removing self-interacting, redundant and false-positive binary interactions.

We considered interaction among protein pairs to be true as follows:

1. if interaction is known to be present in more than one database;

2. PSI-MI (Proteomics Standards Initiative - Molecular Interactions) [19] identifiers were considered to verify true interactions, shortlisting interactions that have been confirmed by more than one biochemical, biophysical, imaging techniques and/or protein complementation assays.
3. Interacting domain pairs were considered from iPFAM [20] and 3DID [21] databases, as interaction among proteins are mediated through independent folding modular domains [22].
4. PMIDs (PubMed identifiers) were used to verify true interactions protein pairs as interaction is validated by more than one independent study..

Similarly, a yeast-HC interaction dataset was created by merging interactions from BIND, BioGrid, DIP, MINT, IntAct and MIPS databases. UniprotKB/Swiss-Prot identifiers were used for consistent mapping of proteins across different databases [23].

Protein Biological Process and Molecular Function

Each protein in the interaction network was mapped to its experimentally verified Gene Ontology (GO) biological process(es) and molecular function(s), taken from the GOA database [24]. All GO with the GOA codes: EXP: Inferred from Experiment, IDA: Inferred from Direct Assay, IPI: Inferred from Physical Interaction, IMP: Inferred from Mutant Phenotype, IGI: Inferred from Genetic Interaction, IEP: Inferred from Expression Pattern, TAS: Traceable Author Statement and IC: Inferred by Curator, were considered.

In this study, each protein assigned a given GO function/process from experiment, was also assigned each of its ancestral functions/processes by tracing *via* every possible path to the top of the hierarchy. The distance of each path defines the level of that GO function/process in the hierarchy. For this study, we labelled the associated GO function/process at a given level by considering one to six steps away from the top or root of the GO hierarchy. This is in accordance with the earlier study done by Duan *et al.* [25], suggesting maximal grouping of biological processes and molecular functions occurs at

level 6 and level 4, respectively in the gene ontology hierarchy based on sequence similarity search.

Network paths

The shortest path between each protein pair in the interaction network was calculated using the Floyd-Warshall algorithm [26]. Different characteristics of paths were considered based on the properties of proteins within that path, their relationships with one another and with the rest of the network. A detailed definition of path characteristics in a interaction network is available from Kumar *et al.* [27].

Firstly, the paths were characterised by their length at distances one, two and three. Secondly, we considered the average degree of proteins along each path, including the terminal proteins. Thirdly, we considered specific combinations of protein degree along the path. Fourthly, we considered the extent to which proteins along the path share interacting neighbours with adjacent path proteins.

The average number of interactions in a path is simply the total number of unique interactions in which the protein members of the path participate, over a number of protein nodes in the path. Statistics were collected for paths with average numbers of interactions in the following ranges: 1-5, 6-10, 11-20, 21-50, 51-100 and >100. For the sake of brevity, each range was labelled in this study by the number immediately below that range i.e. 0, 5, 10, 20, 50 and 100, respectively.

For the study of specific degree combinations of paths, similar protein degree ranges were considered to that above. However, due to the relatively large number of degree combinations and the relative scarcity of proteins with large numbers of interactions, we considered degree ranges 1-5, 6-10, 11-20, 21-50 and >50 for each protein in the path, the corresponding degree ranges were labelled by the numbers 0, 5, 10, 20 and 50 respectively. Paths were then labelled by the sequence of degree labels of proteins in each path,

accounting for path symmetry e.g. a path of length two with degree sequence 0-5-10 is identical to a path of sequence 10-5-0.

Percentage of paths with a common function or process

The extent to which path proteins shared interacting neighbours with adjacent path proteins were assessed by the percentage of common neighbours (*PCN*), defined as follow:

$$PCN = 100 \times \left(\frac{\sum_{i=1}^{N-1} \sum_{j=1}^{D_i} \sum_{k=1}^{D_{i+1}} \delta(I_i^j, I_{i+1}^k)}{\sum_{i=1}^{N-1} \min(D_i, D_{i+1})} \right)$$

where N is the number of proteins in the path, D_i is the degree of the protein in the path I_{ij} is the j^{th} interactor of the i^{th} protein in the path. The value of $\min(D_1, D_2)$ is the smallest value of D_1 and D_2 . The value of $\delta(I_1, I_2)$ is 1 if I_1 is identical to I_2 and 0 otherwise.

In this study, paths were categorised according to their *PCN* values, and statistics collected for *PCN* ranges 0-20, 20-40, 40-60, 60-80 and 80-100. The probability of each average interaction/specific degree sequence path category having a *PCN* in each of these ranges was also calculated. Given that higher *PCN* categories were sparsely populated for some types of path, *PCN* range categories were added to the next lower category if their count was <25 in an iterative fashion, starting with the highest *PCN* category (80-100).

Hubbed and Hubless networks

To evaluate the influence of highly connected proteins (hubs) on functional and biological relationships in the *H. sapiens* and *S. cerevisiae* networks, the consequences of their removal from the network was studied. Several definitions of hubs were used. Table 1 gives the number of proteins present in the different ranges of node degrees. As there are relatively few proteins with >100 interactions, these were merged with proteins >50 interactions. Then, proteins with >10, >20 and >50 interactions were removed in turn and paths in the resulting hubless networks examined.

The participation of hubs in network paths is two-fold: hubs may be one of the path's terminal proteins, whose relationships we are studying, or a mid-path protein, facilitating communication between the path's terminal proteins. Therefore, the removal of hubs from an interaction network not only removes relationships between hubs and other proteins but may also remove short paths between non-hubs. To differentiate between these two separate influences of hubs, we considered paths in hubless networks, with no hubs as terminal or mid-path proteins and also paths from the entire network excluding those with hubs as terminal proteins (i.e. with no terminal protein hubs). We compared the distance profile of the overall network ("with hubs") with the hubless networks ("without hubs") and terminal-hubless paths ("without (via) hubs"), illustrated in Figure 2.

Results

Protein interactions and the associated biological process and molecular function data were collected for *S. cerevisiae* and *H. sapiens* networks. The shortest paths between all pairs of proteins in each interaction network were calculated and subsequently characterised according to the various criteria (as detailed in the Methods section). Paths were assessed for their tendency to connect proteins with the same GO function or process, measured by their percentage common neighbour (*PCN*), measured as the percentage of paths whose terminal proteins have at least one function/process in common.

Path distance and level of GOA functional and process detail

Proximal proteins in interaction networks have been shown to have the same function significantly more often than expected [28]. However, function can be defined in different ways and described in varying levels of detail. Proximal proteins may have some aspects of function in common but not others. For example, if two proteins were known to be different types of kinases, a simple description of function might have these proteins in the same category whereas a more detailed description would put these proteins into different categories.

Here, we examined the molecular functions and biological processes of proximal proteins in the *S. cerevisiae* and *H. sapiens* interaction networks at various levels of detail by exploiting the hierarchical structure of GO. A GO category may have one or more descendant functional or process terms describing more specific aspects of that term. Conversely, a GO characteristic may have ancestral terms with more general descriptions of that function or process. A protein associated with a given GO term must also be associated with its ancestral (more general) terms in the GO hierarchy but not necessarily with its descendant (more specific) terms.

Each protein in the interaction network with a given GO function or process term assigned from the experiment was also assigned its ancestral GO functions/processes (see Methods). Functions or processes were assigned at GO levels according to their distances from the root term of the GO hierarchy tree. Thus, function/process terms at GO level 1 were the simplest with least detailed descriptions, with function/process detail increasing with increasing GO level. In this study, we considered GO levels 1 through 6.

For each GO level, we studied the function/process similarity of proteins at distances up to 3 in the *S. cerevisiae* and *H. sapiens* interaction networks and randomised networks. This was done by calculating the percentage of common neighbours (*PCN*) with respect to molecular function and biological process for each path distance, where *PCN* is the percentage of paths whose terminal proteins have at least one function/process in common. Results for GO levels 1 to 6 are shown in Figure 3 and 4.

Proteins at distances 1 and 2 generally had functions/processes in common more often than expected, in line with other studies [28]. However, there is little difference between real and random networks for paths of distance 3. Paths of distance 1 and 2 were more likely to have functions or processes in common at higher GO levels relative to random networks. A very high percentage of paths had the same function for GO levels 1 and 2, whereas for biological process, this is true up to level 3, however, at this level, the percentage is also

very close to that in random networks. There was generally better discrimination between real and random networks for more detailed function and process descriptions, with random networks consistently showing higher *PCN* values than real networks. However, there was a slightly higher tendency (relative to random) to have the same function at distance 3 for low GO levels than high.

Removal of hubs from the network

To evaluate the influence of highly-connected proteins (hubs) on functional or process relationships in the *S. cerevisiae* and *H. sapiens* networks, they were removed from the network. Several different definitions of hubs were used. Proteins with >10, >20, >50 and >100 interactions were each removed in turn and paths in the resulting hubless network were examined in a similar fashion to the previous section. The numbers of proteins with each range of degrees considered can be found in Table 1.

The participation of hubs in network paths is two-fold: hubs may be one of the path terminal proteins, whose functional or process relationships we are studying, or a mid-path protein, facilitating functional communication between the path's terminal proteins. Therefore, the removal of hubs from an interaction network not only removes relationships between hubs and other proteins but may also remove short paths between non-hubs. To differentiate between these two separate influences of hubs, we considered paths in hubless networks (with no hubs as terminal or mid-path proteins) and also paths from the entire network excluding those with hubs as terminal proteins (with no hubs as terminal proteins). We compared the distance profile of the overall network (with hubs) with the hubless networks (without hubs) and terminal-hubless paths (without (via) hubs). Distance profiles are shown for molecular function in Figure 5 for hubs defined as having >10, >20 and >50 interactions at GO level 3, whereas Figure 6 shows the same for biological process. There is little difference between profiles for hubs with >100 interactions (data not shown).

Distance profiles are also shown for hubs with >10 interactions at GO levels 1, 2, 4, 5 and 6 in Additional file 1 and 2.

Clearly, at GO level 3, the distance profiles of the hubless and terminal-hubless paths were more distinct with a lower interaction threshold for hubs. When hubs were removed entirely, paths in hubless networks were more likely to have the same function or processes at end-points at distances one and two. When hubs were excluded as terminal path proteins but act as conduits for paths between non-hubs, the behaviour was the same at a distance of 1 by construction (there are no intervening hubs in direct interactions). However, at a distance of 2, the paths were less likely to connect proteins with the same function if hubs were retained as a means of communication between non-hubs. This is consistent with a model of hubs communicating between different function modules in the network. Similar trends were seen with other GO levels of 3 and greater (Supplementary figures), with greater tendencies of hubless networks to have similar functions at greater GO levels. At GO levels 1 and 2, there was even a slight tendency for terminal-hubless networks to have fewer paths connecting similar functions and processes than unaltered networks, even for direct interactions, thus confirming the importance of hubs in biological networks.

Discussion

Protein interaction networks offer useful insight into the functional/process organisation of the cell at the molecular systems level. Proteins at distances of less than three in the network are significantly more likely to have the same function/process than would be expected in random networks. However, the tendency of proximal proteins to have the same function/process depends heavily on the connectivity properties of the path joining them. Assignment of an unknown protein's function and process is thus reliable upto GOA levels 2 and 3, respectively in a protein interaction network. Highly connected proteins (hubs) have been the focus of many studies of network organisation in recent times. Here, we have shown that hubs are less likely to have the same function/process as their near

neighbours. Removing hubs from networks also generally results in a higher likelihood of remaining paths connecting proteins of the same function/process. More generally, paths are increasingly less likely to connect proteins of the same function/process with increasing degree of path proteins. These results are consistent with the view that hubs tend to act as conduits between different functional units in the network. Complicating this view is the observation that removing hubs from networks does not greatly affect functional/process relationships for more basic function/process properties. Proteins connected via hubs are relatively more likely to retain the same function/process at low GO levels than is the case for high GO levels. While hubs may connect proteins whose functions/processes are different at high levels of detail, their more basic underlying function may still be the same.

It is perhaps too simplistic to discuss the functional organisation of networks in terms of two discrete species of protein: hub and non-hub. The behaviour of proteins in networks appears to change continuously over a range of degrees. High and low interacting proteins are generally less likely to have the same function/process than their proximal proteins. However, proteins of low degree have relatively high similarity with their neighbours for functions/processes at low GO levels. Proteins of middle-ranking degrees are more likely to have the same function as their neighbours. In all cases, functional/process relationships depend on the degrees of all path members. Proteins of similar degree are more likely to share functions, particularly mid-ranking degrees.

Functional relationships between paths of differing degree sequence can be partially understood in terms of their relative levels of local clustering. The functional similarity of interacting pairs with different degree combinations varies consistently with the level of clustering. However, clustering does not explain the behaviour of paths of length 2. The most similar path types of length 2 are also not always conjunctions of the most similar

interaction types. These observations all hint at an additional functional role for hubs in networks.

Conclusions

In this paper, we have presented a detailed statistical analysis of interacting protein pairs to determine conservation of GO molecular functions or biological processes at various network distances and GOA detailed annotations, described up to six levels of abstraction. We have shown the tendency of proximal proteins to have same function more likely than chance at GOA level 1 and 2. Whereas, this tendency hold true upto GOA level 3 for biological processes. Remove of hub decreases the tendency of proximal protein to have same function and process. However, a decrease in *PCN* values with increasing number of network paths connecting proximal protein through hubs clearly suggests that hubs connect different functional modules inside the cell. Our analysis can help other researchers to prioritise protein characterization based on GOA hierarchy and serve as background when analysing focused datasets.

Authors' contributions

APC and GK conceived and designed the experiment, GK and APC collected the data and performed the analysis. The manuscript was initially drafted by APC and finalised by GK and SR.

Conflict of interest: none declared.

Acknowledgements

This research was supported by the Macquarie University Research Scholarship (MQRES) to GK and the ARC Centre of Excellence in Bioinformatics grant (CE0348221) to SR.

References

1. Schwikowski B, Uetz P, Fields S: **A network of protein-protein interactions in yeast.** *Nat Biotechnol* 2000, **18**(12):1257-1261.
2. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B *et al*: **Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets.** *Nat Genet* 2006, **38**(3):285-293.
3. Sharan R, Ulitsky I, Shamir R: **Network-based prediction of protein function.** *Mol Syst Biol* 2007, **3**:88.
4. Spirin V, Mirny LA: **Protein complexes and functional modules in molecular networks.** *Proc Natl Acad Sci U S A* 2003, **100**(21):12123-12128.
5. Bhalla US, Iyengar R: **Emergent properties of networks of biological signaling pathways.** *Science* 1999, **283**(5400):381-387.
6. Francesconi M, Remondini D, Neretti N, Sedivy JM, Cooper LN, Verondini E, Milanesi L, Castellani G: **Reconstructing networks of pathways via significance analysis of their intersections.** *BMC Bioinformatics* 2008, **9 Suppl 4**:S9.
7. Barabasi AL, Albert R: **Emergence of scaling in random networks.** *Science* 1999, **286**(5439):509-512.
8. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP *et al*: **Evidence for dynamically organized modularity in the yeast protein-protein interaction network.** *Nature* 2004, **430**(6995):88-93.
9. Batada NN, Hurst LD, Tyers M: **Evolutionary and physiological importance of hub proteins.** *PLoS Comput Biol* 2006, **2**(7):e88.
10. Burkard TR, Planyavsky M, Kaupe I, Breitwieser FP, Burckstummer T, Bennett KL, Superti-Furga G, Colinge J: **Initial characterization of the human central proteome.** *BMC Syst Biol* 2011, **5**:17.
11. Kumar G, Ranganathan S: **Network analysis of human protein location.** *BMC Bioinformatics* 2010, **11 Suppl 7**:S9.
12. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bahler J, Wood V *et al*: **The BioGRID Interaction Database: 2008 update.** *Nucleic Acids Res* 2008, **36**(Database issue):D637-640.

13. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular INTERaction database**. *Nucleic Acids Res* 2007, **35**(Database issue):D572-574.
14. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R *et al*: **IntAct--open source resource for molecular interaction data**. *Nucleic Acids Res* 2007, **35**(Database issue):D561-565.
15. Prasad TS, Kandasamy K, Pandey A: **Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology**. *Methods Mol Biol* 2009, **577**:67-79.
16. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update**. *Nucleic Acids Res* 2004, **32**(Database issue):D449-451.
17. Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V: **MPact: the MIPS protein interaction resource on yeast**. *Nucleic Acids Res* 2006, **34**(Database issue):D436-441.
18. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E *et al*: **The Biomolecular Interaction Network Database and related tools 2005 update**. *Nucleic Acids Res* 2005, **33**(Database issue):D418-424.
19. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C *et al*: **The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data**. *Nat Biotechnol* 2004, **22**(2):177-183.
20. Finn RD, Marshall M, Bateman A: **iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions**. *Bioinformatics* 2005, **21**(3):410-412.
21. Stein A, Ceol A, Aloy P: **3did: identification and classification of domain-based interactions of known three-dimensional structure**. *Nucleic Acids Res* 2011, **39**(Database issue):D718-723.
22. Boxem M, Maliga Z, Klitgord N, Li N, Lemmens I, Mana M, de Lichtenvelde L, Mul JD, van de Peut D, Devos M *et al*: **A protein domain-based interactome network for C. elegans early embryogenesis**. *Cell* 2008, **134**(3):534-545.
23. Magrane M, Consortium U: **UniProt Knowledgebase: a hub of integrated protein data**. *Database (Oxford)* 2011, **2011**:bar009.

24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**(1):25-29.
25. Duan ZH, Hughes B, Reichel L, Perez DM, Shi T: **The relationship between protein sequences and their gene ontology functions.** *BMC Bioinformatics* 2006, **7 Suppl 4**:S11.
26. Sedgewick R: **Algorithms in C++ Part 5: Graph Algorithms**, 3rd edn: Addison-Wesley Professional; 2001.
27. Kumar G, Cootes AP, Ranganathan S: **Untangling Biological Networks Using Bioinformatics.** In: *Algorithms in Computational Molecular Biology, Approaches and Applications*. Edited by Elloumi MaZ, A.Y.: Wiley-Blackwell; 2011: pp. 867-888.
28. Schwikowski B, Uetz P, Fields S: **A network of protein–protein interactions in yeast.** *Nature Biotechnology* 2000, **18**:1257-1261.

Figures

Figure 1 – Illustration of GOA abstraction levels

A schematic representation of GOA levels in a directed acyclic graph (DAG) is presented. Root nodes are defined as cellular component, biological process and molecular function. GOA level increases with the depth of GOA hierarchy moving from away from root node, shown here for biological process.

Figure 2 – Schematic representation of networks with hubs, without hubs and without-(via)-hubs.

Considering the white node as the hub under investigation, networks A. with the hub, B. without the hub and C. bypassing the hub (without-(via)-hub) are shown.

Figure 3 - Variation of molecular function with network distance and GO level

PCN values for molecular function at network distances 1-3 and GO levels 1-6 are shown, with H-real, H-rand, Y-real and Y-rand representing *H. sapiens* real, *H. sapiens* random, *S. cerevisiae* real and *S. cerevisiae* random interaction networks. Error bars represent standard error.

Figure 4 - Variation of biological process with network distance and GO levels

PCN values for biological processes at network distances 1-3 and GO levels 1-6 are shown, with H-real, H-rand, Y-real and Y-rand representing *H. sapiens* real, *H. sapiens* random, *S. cerevisiae* real and *S. cerevisiae* random interaction networks. Error bars represent standard error.

Figure 5 - Variation of molecular function in hubless networks at GOA level 3.

PCN values for molecular function for the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 and GO level 3 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10, >20 and >50 connections are removed (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”). Data for GO levels 1, 2, 4, 5 and 6 are shown in Supplementary File 1: Figures S1-S5.

Figure 6 - Variation of biological process in hubless networks at GOA level 3

PCN values for biological process with the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 and GO level 3 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10 (A, B), >20 (C, D) and >50 (E, F) connections are removed from the interaction network (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”). Data for GO levels 1, 2, 4, 5 and 6 are shown in Supplementary File 2: Figures S6-S10.

Tables

Table 1 - Node distribution for yeast and human interactomes

Number of proteins in a given range of node degrees is presented.

Additional files

Additional file 1 – Variation of molecular function in hubless networks at GOA levels 1, 2, 4, 5 and 6.

PCN values for molecular function for the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10, >20 and >50 connections are removed (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”). Data for GO levels 1, 2, 4, 5 and 6 are shown in Figures S1-S5, respectively.

Additional file 2 – Variation of biological process in hubless networks at GOA levels 1, 2, 4, 5 and 6.

PCN values for biological process with the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10 (A, B), >20 (C, D) and >50 (E, F) connections are removed from the interaction network (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”). Data for GO levels 1, 2, 4, 5 and 6 are shown in Figures S6-S10, respectively.

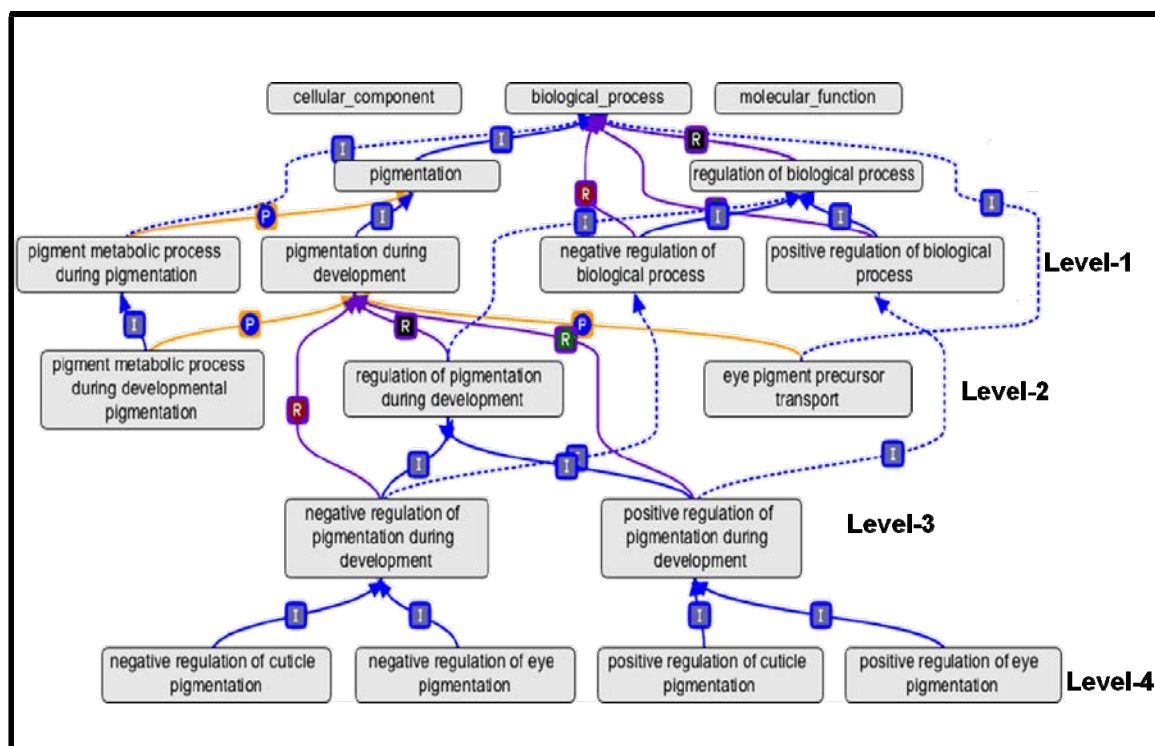


Figure 1 – Illustration of GOA abstraction levels

A schematic representation of GOA levels in a directed acyclic graph (DAG) is presented. Root nodes are defined as cellular component, biological process and molecular function. GOA level increases with the depth of GOA hierarchy moving from away from root node, shown here for biological process.

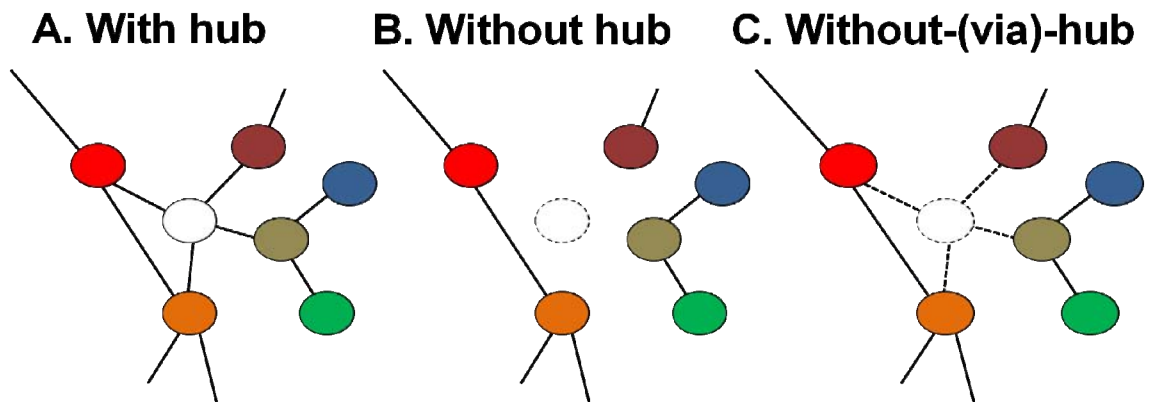


Figure 2 – Schematic representation of networks with hubs, without hubs and without-(via)-hubs.

Considering the white node as the hub under investigation, networks A. with the hub, B. without the hub and C. bypassing the hub (without-(via)-hub) are shown.

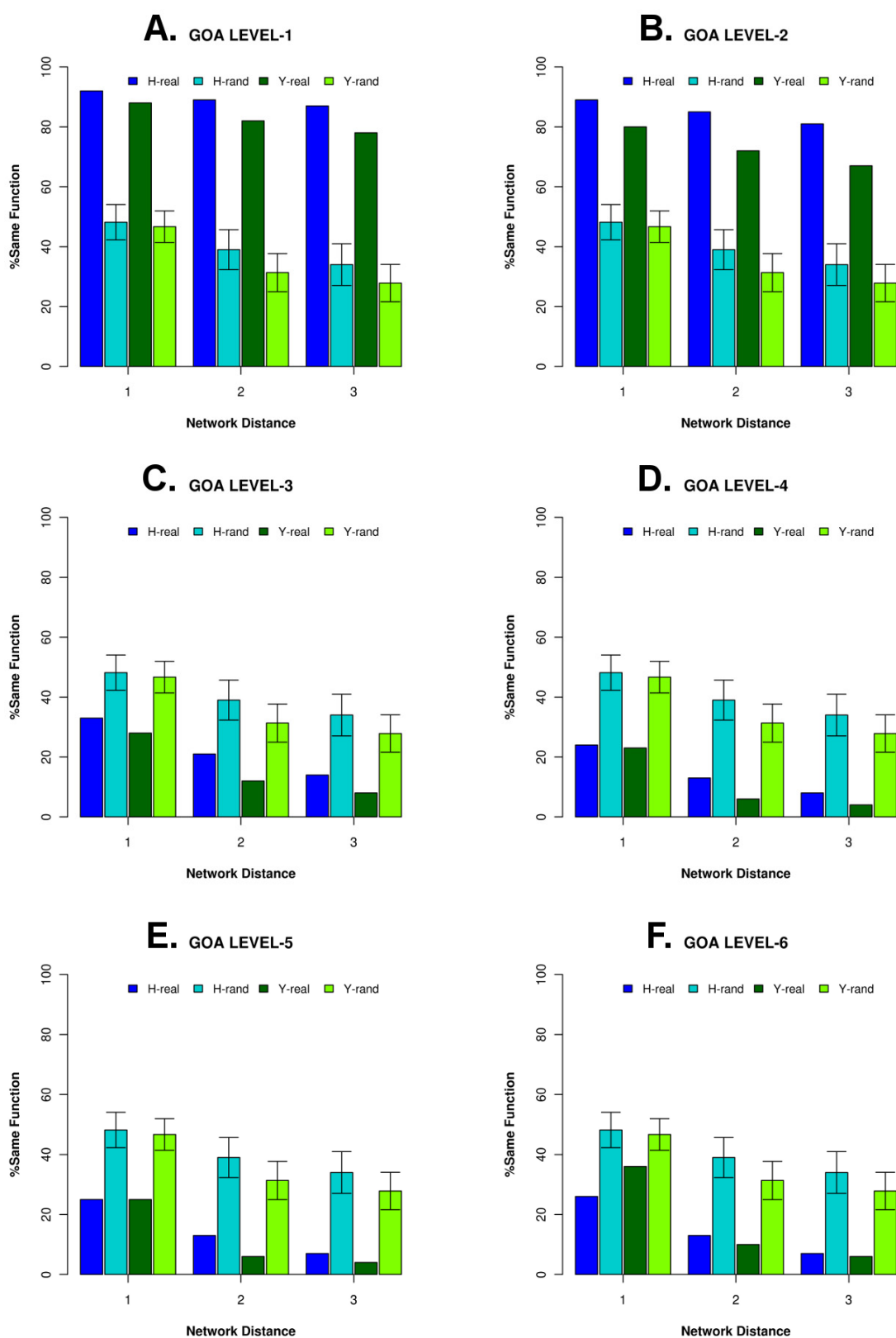


Figure 3 - Variation of molecular function with network distance and GO level

PCN values for molecular function at network distances 1-3 and GO levels 1-6 are shown, with H-real, H-rand, Y-real and Y-rand representing *H. sapiens* real, *H. sapiens* random, *S. cerevisiae* real and *S. cerevisiae* random interaction networks. Error bars represent standard error.

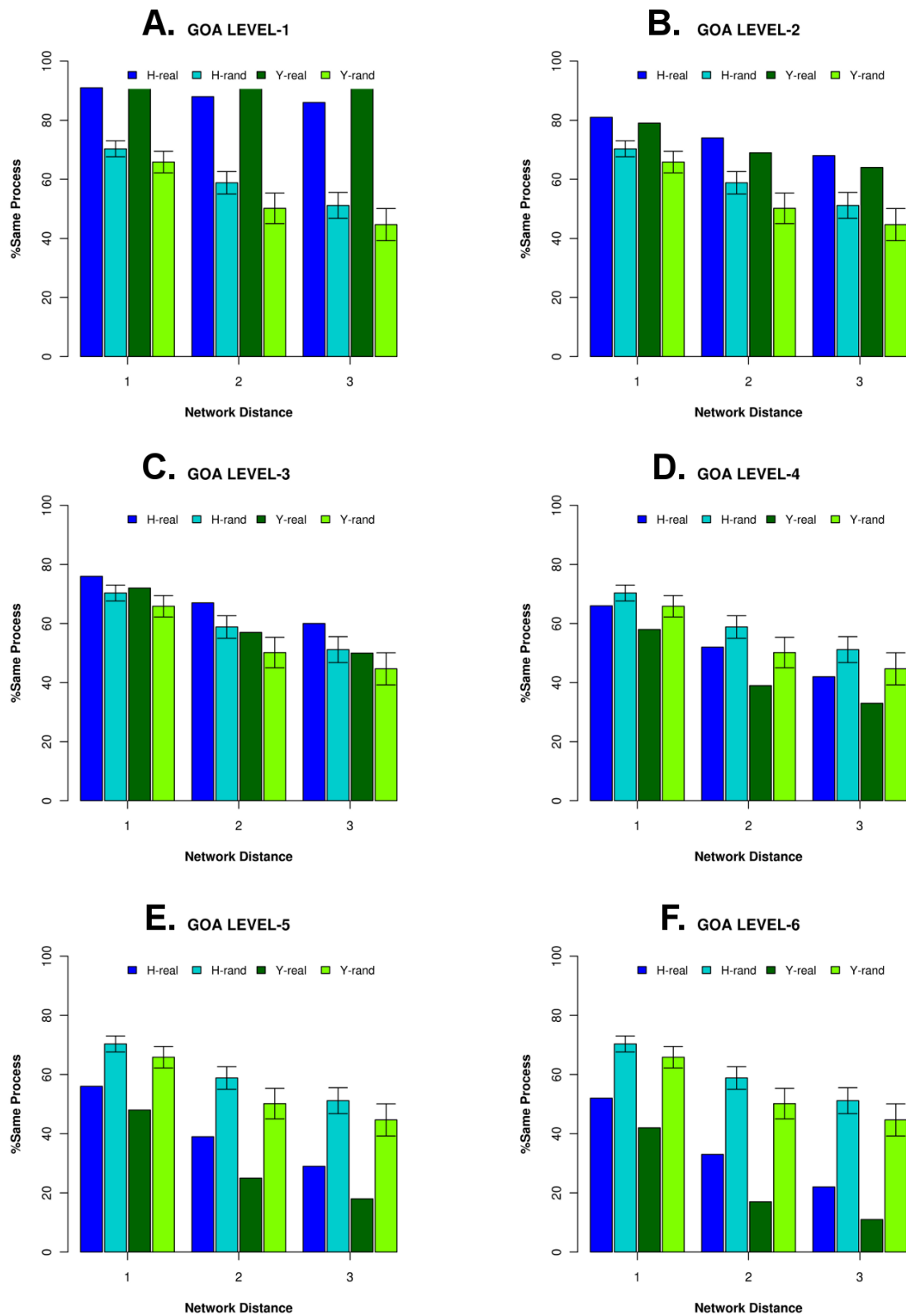


Figure 4 - Variation of biological process with network distance and GO levels

PCN values for biological processes at network distances 1-3 and GO levels 1-6 are shown, with H-real, H-rand, Y-real and Y-rand representing *H. sapiens* real, *H. sapiens* random, *S. cerevisiae* real and *S. cerevisiae* random interaction networks. Error bars represent standard error.

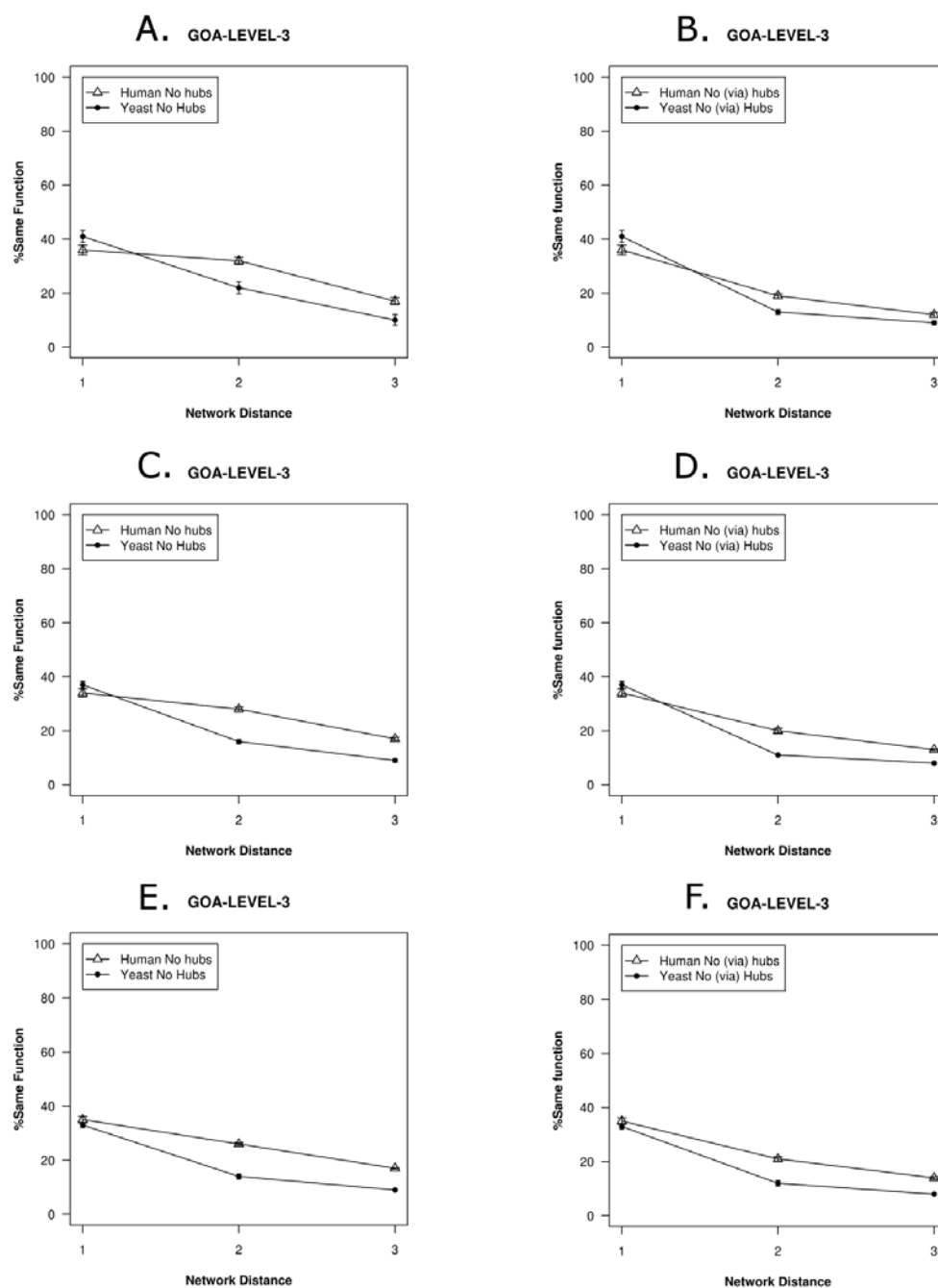


Figure 5 - Variation of molecular function in hubless networks at GOA level 3.

PCN values for molecular function for the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 and GO level 3 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10, >20 and >50 connections are removed (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”). Data for GO levels 1, 2, 4, 5 and 6 are shown in Supplementary File 1: Figures S1-S5.

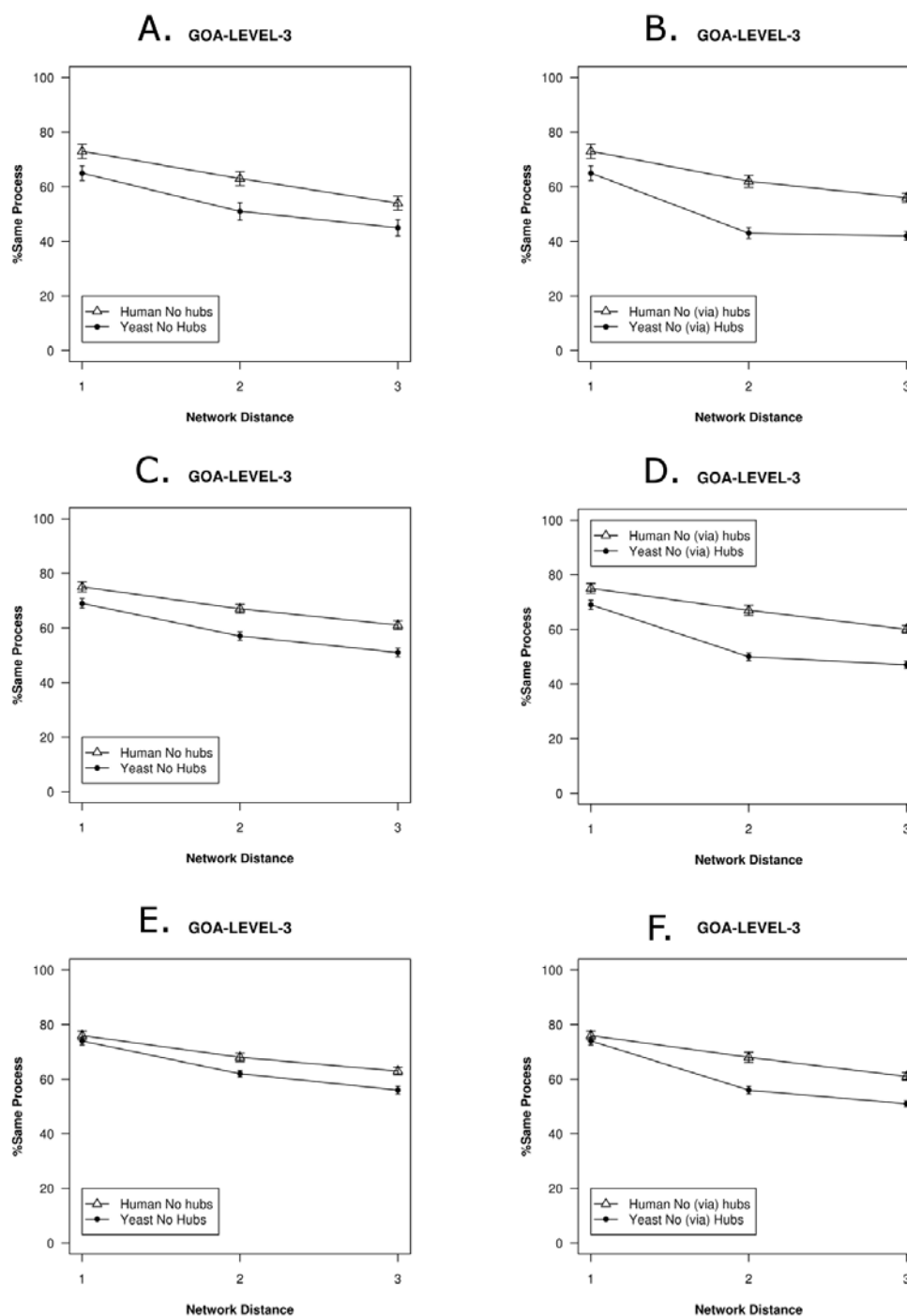


Figure 6 - Variation of biological process in hubless networks at GOA level 3

PCN values for biological process with the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 and GO level 3 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10 (A, B), >20 (C, D) and >50 (E, F) connections are removed from the interaction network (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”). Data for GO levels 1, 2, 4, 5 and 6 are shown in Supplementary File 2: Figures S6-S10.

Table 1 - Node distribution for yeast and human interactomes

Number of proteins in a given range of node degrees is presented.

Degree	Number of yeast proteins	Number of human Proteins
All	5177	7390
>10	1403 (27%)	778 (10%)
>20	637 (12%)	257 (3%)
>50	110 (2%)	62 (1%)
>100	18 (0.003%)	19 (0.002%)

Additional file 1 –Variation of molecular function in hubless networks at GOA levels 1, 2, 4, 5 and 6.

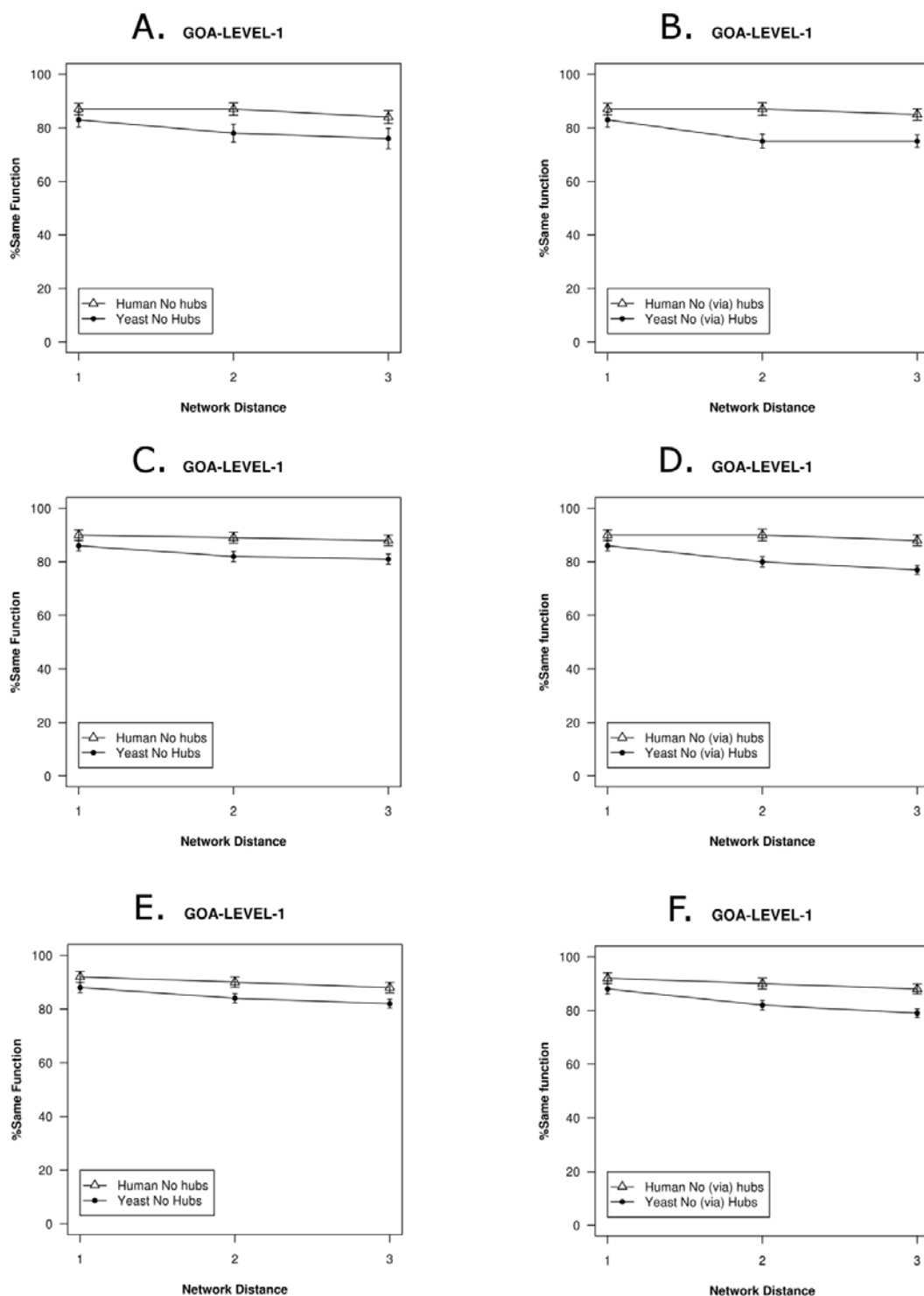


Figure S1 - Variation of molecular function in hubless networks at GOA level 1

PCN values for molecular function for the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 and GO level 1 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10, >20 and >50 connections are removed (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”).

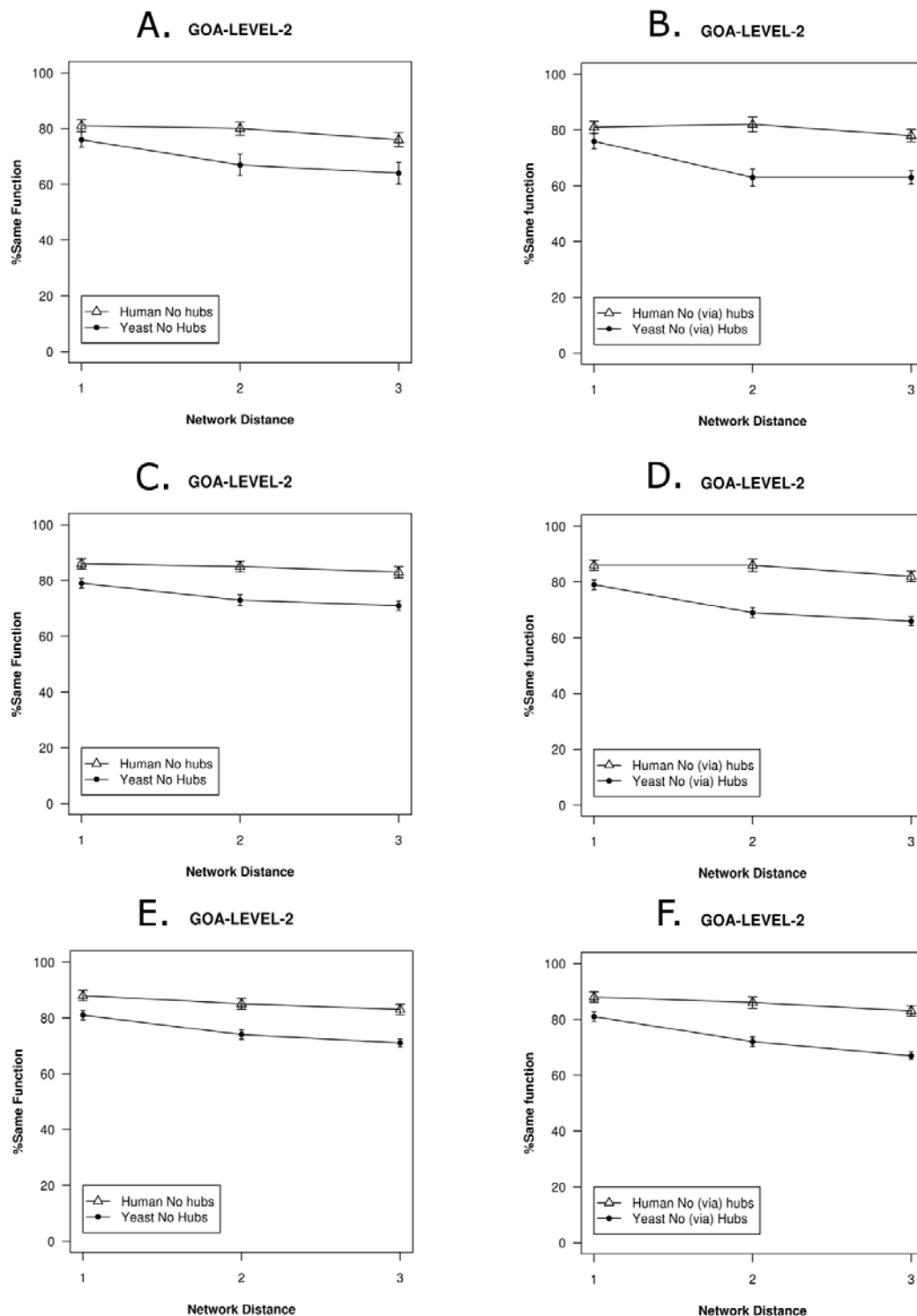


Figure S2 - Variation of molecular function in hubless networks at GOA level 2

PCN values for molecular function for the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 and GO level 2 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10, >20 and >50 connections are removed (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”).

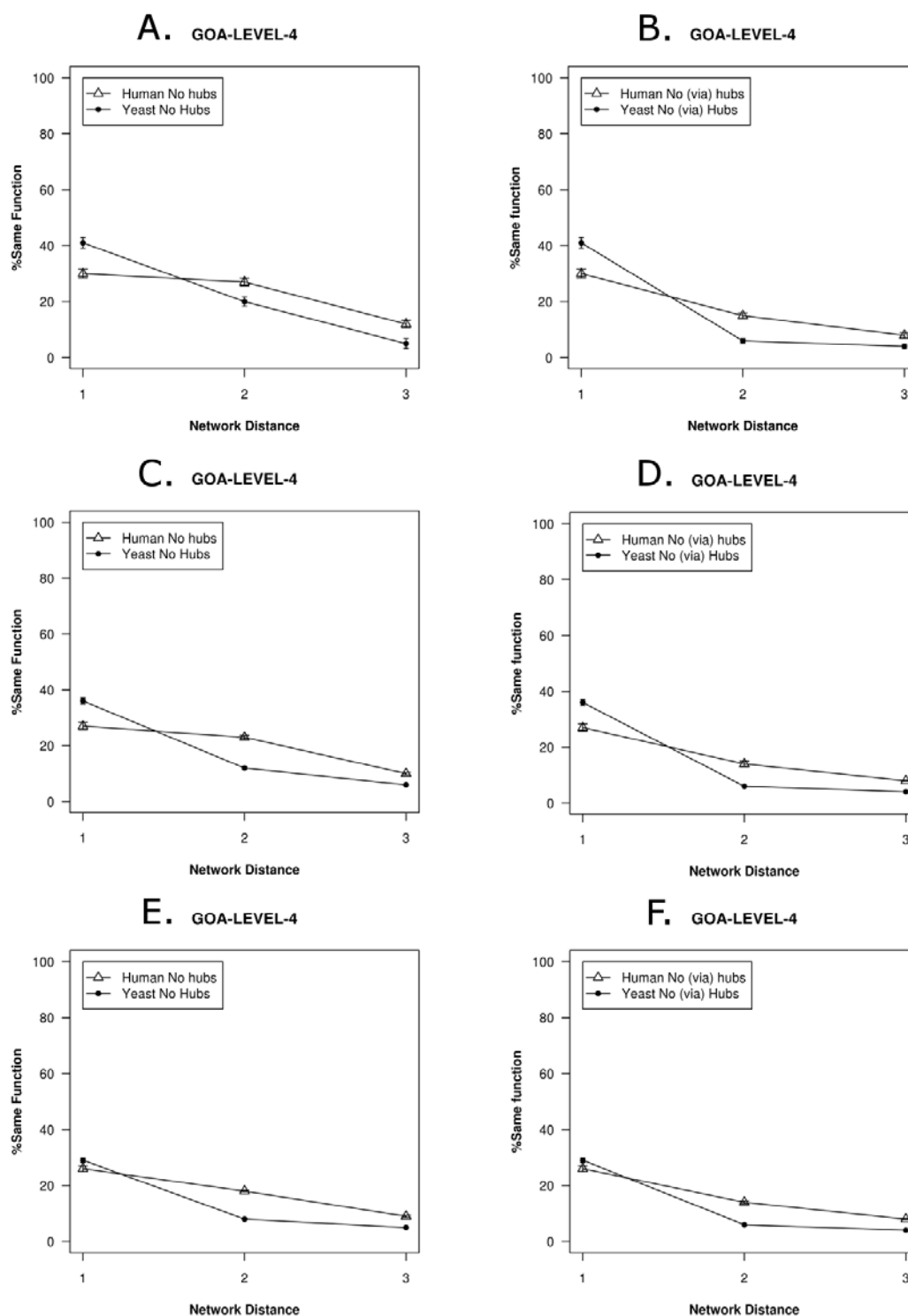


Figure S3 - Variation of molecular function in hubless networks at GOA level 4

PCN values for molecular function for the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 and GO level 4 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10, >20 and >50 connections are removed (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”).

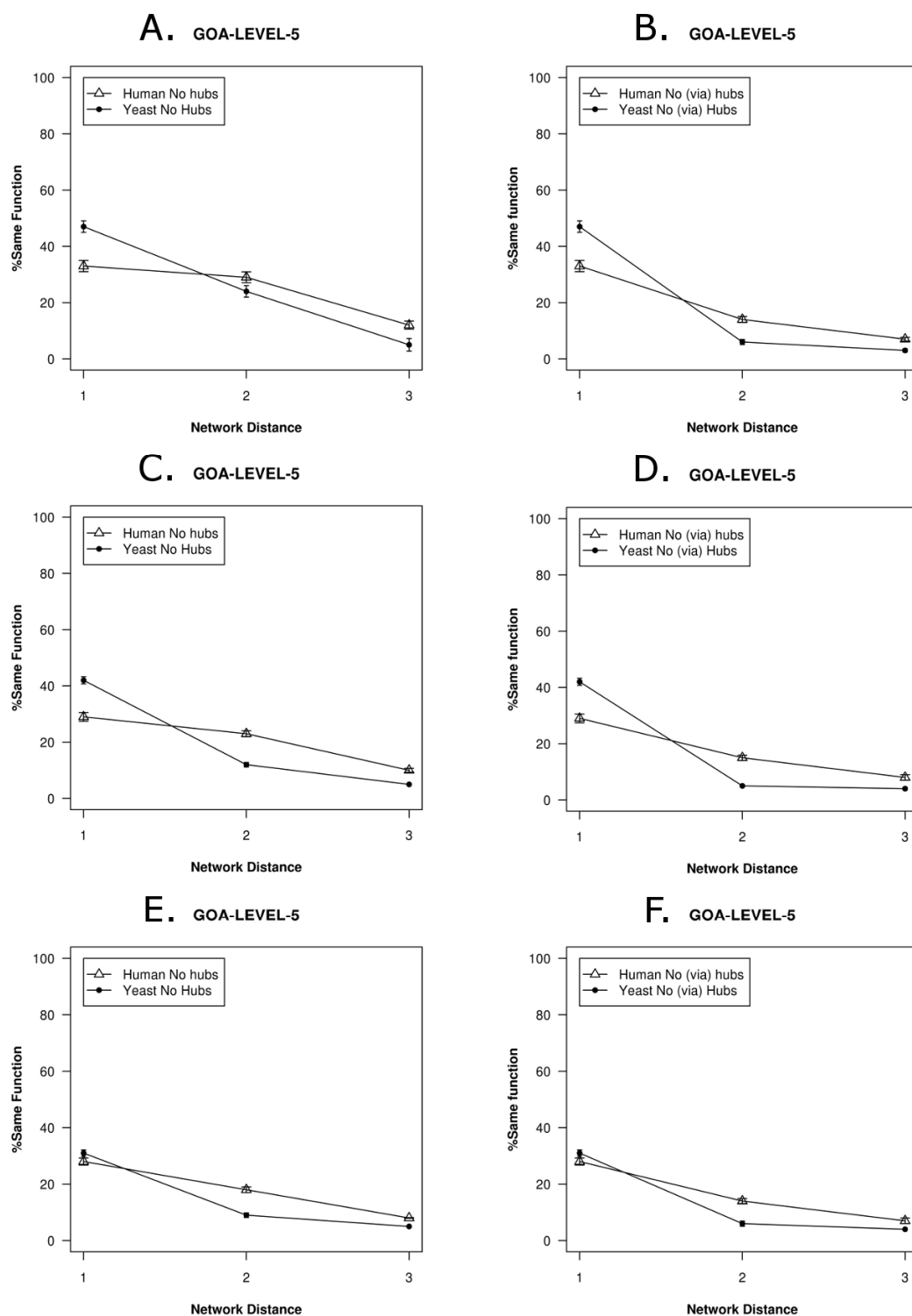


Figure S4 - Variation of molecular function in hubless networks at GOA level 5

PCN values for molecular function for the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 and GO level 5 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10, >20 and >50 connections are removed (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”).

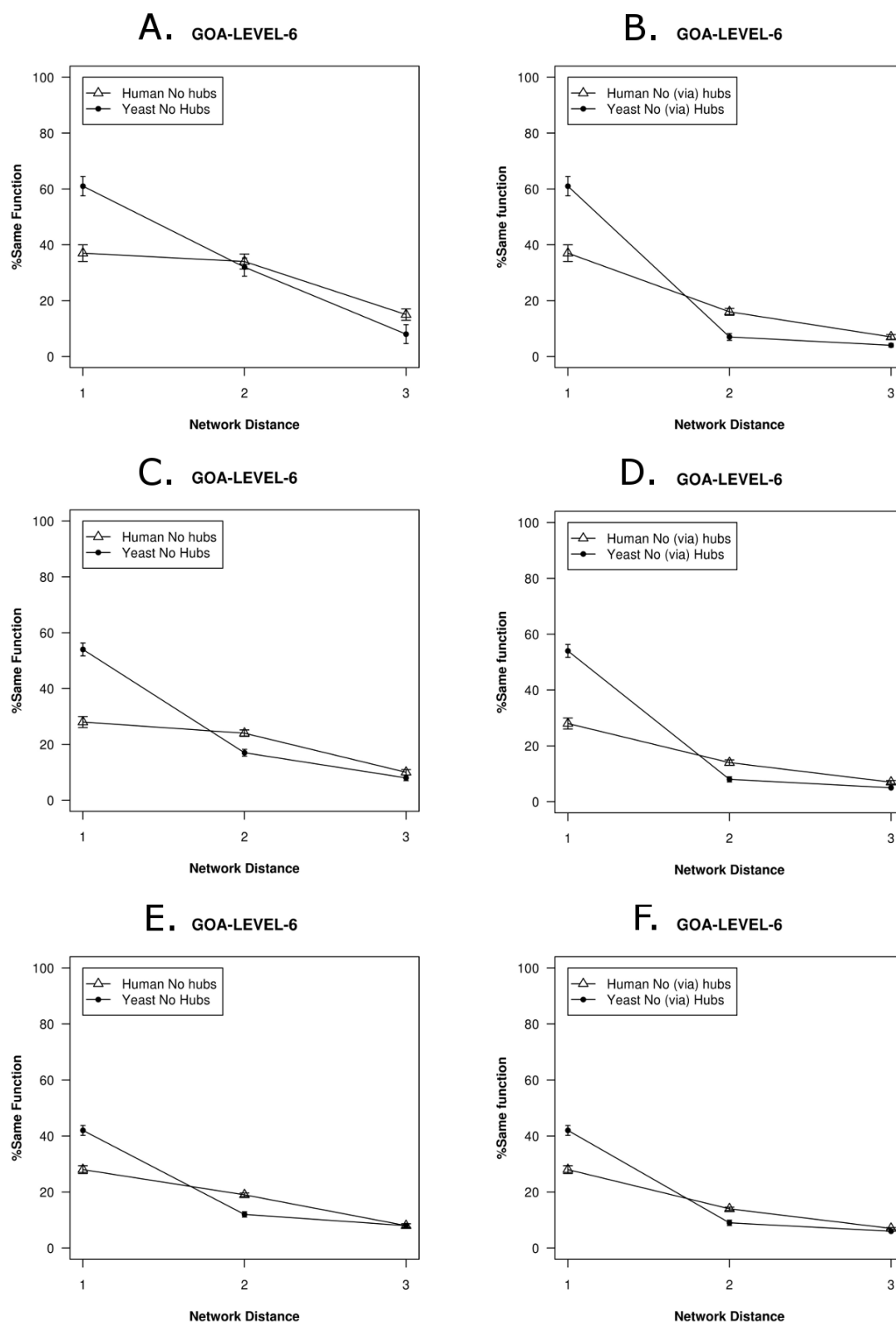


Figure S5 - Variation of molecular function in hubless networks at GOA level 6

PCN values for molecular function for the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 and GO level 6 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10, >20 and >50 connections are removed (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”).

Additional file 2 –Variation of biological process in hubless networks at GOA levels 1, 2, 4, 5 and 6.

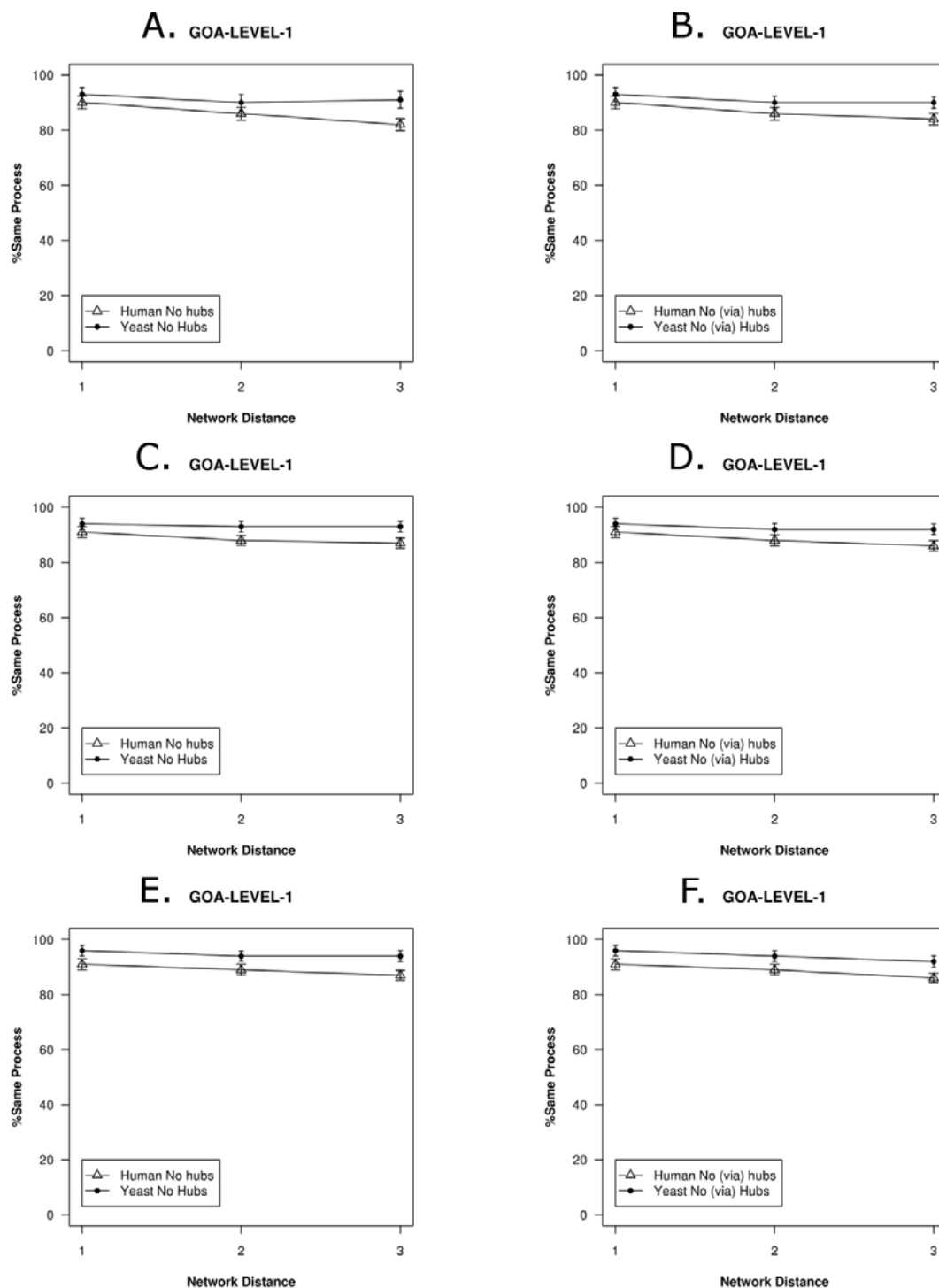


Figure S6 - Variation of biological process in hubless networks at GOA level 1

PCN values for biological process with the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 and GO level 1 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10 (A, B), >20 (C, D) and >50 (E, F) connections are removed from the interaction network (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”).

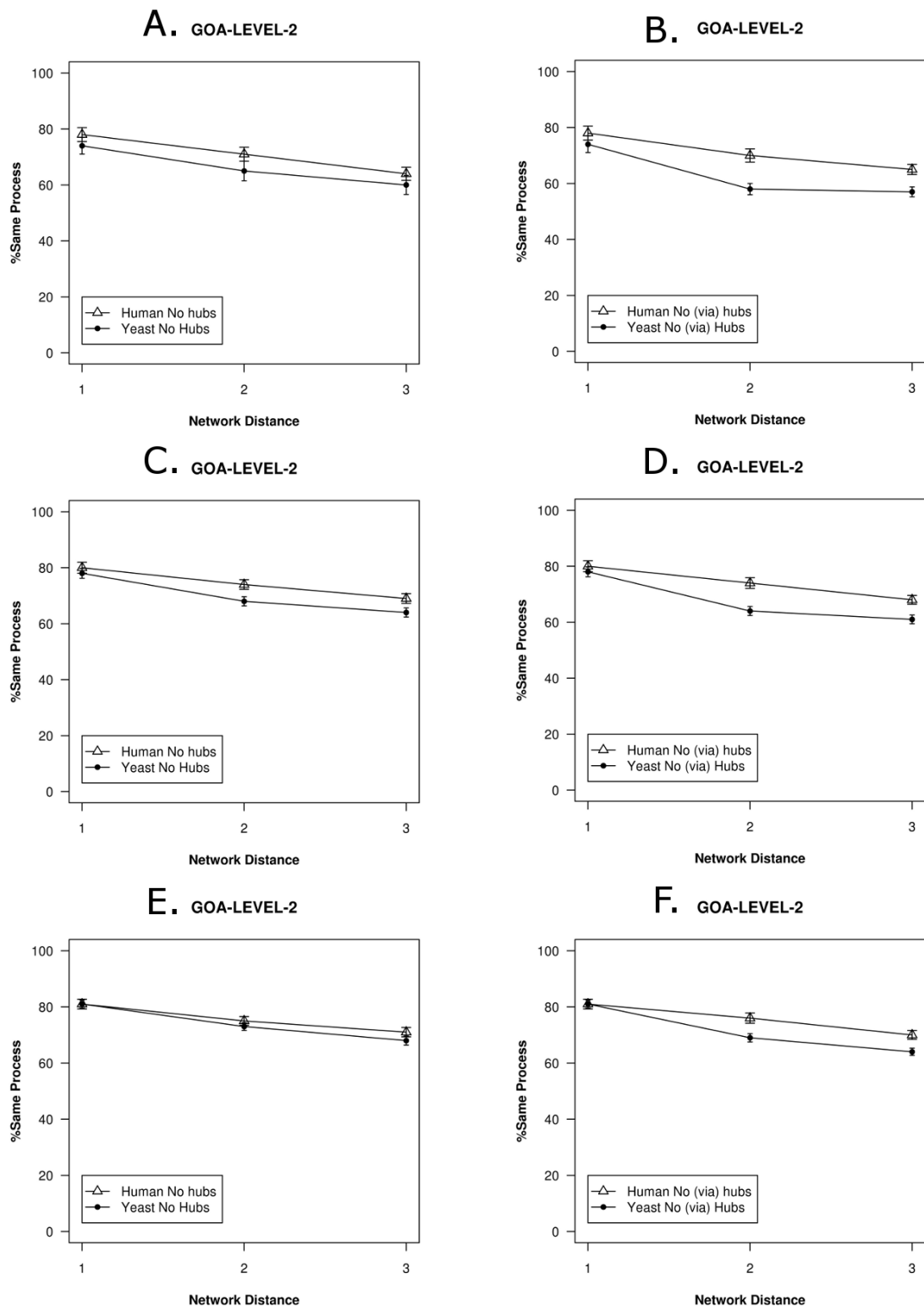


Figure S7 - Variation of biological process in hubless networks at GOA level 2

PCN values for biological process with the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 and GO level 2 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10 (A, B), >20 (C, D) and >50 (E, F) connections are removed from the interaction network (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”).

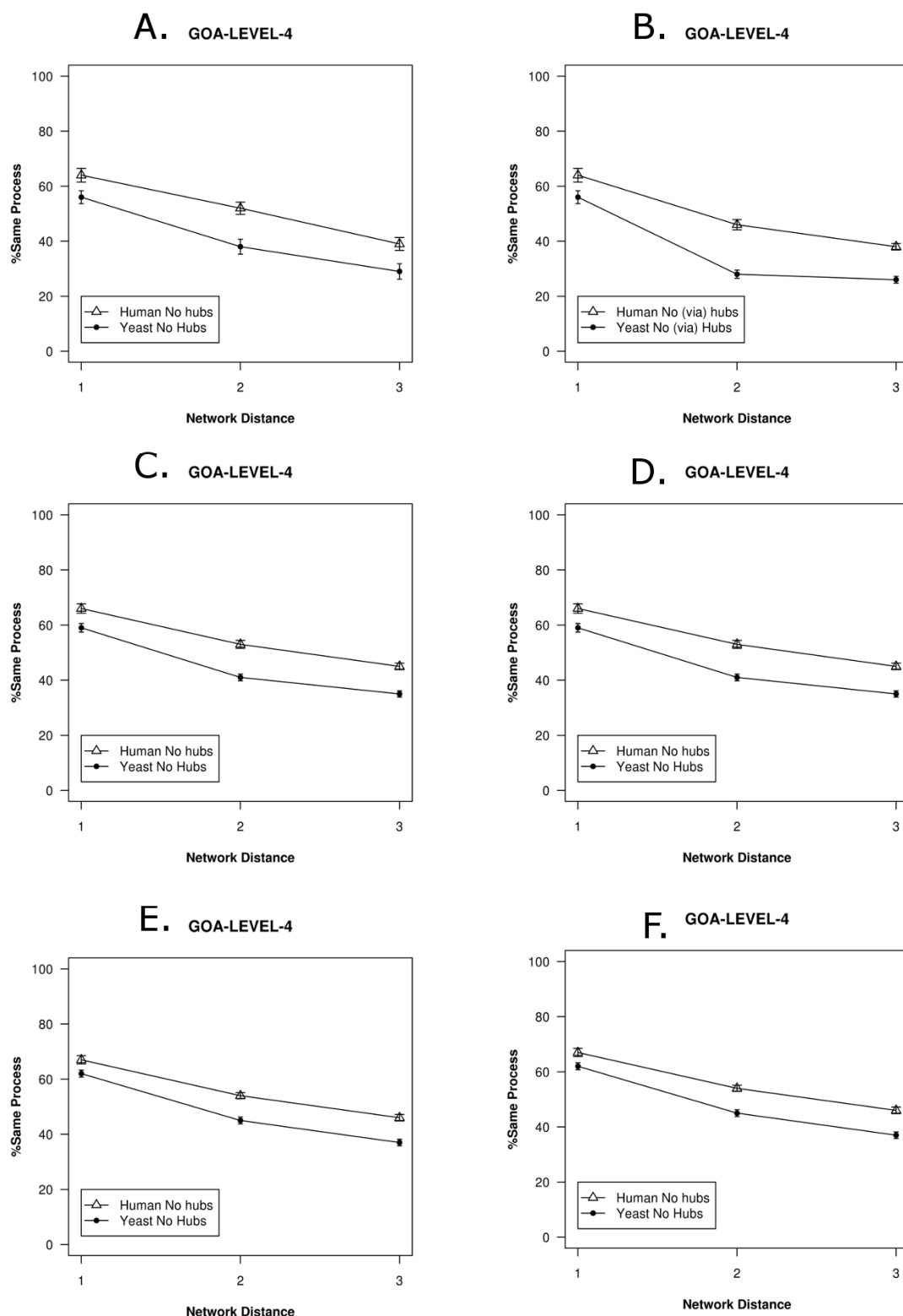


Figure S8 - Variation of biological process in hubless networks at GOA level 4

PCN values for biological process with the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 and GO level 4 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10 (A, B), >20 (C, D) and >50 (E, F) connections are removed from the interaction network (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”).

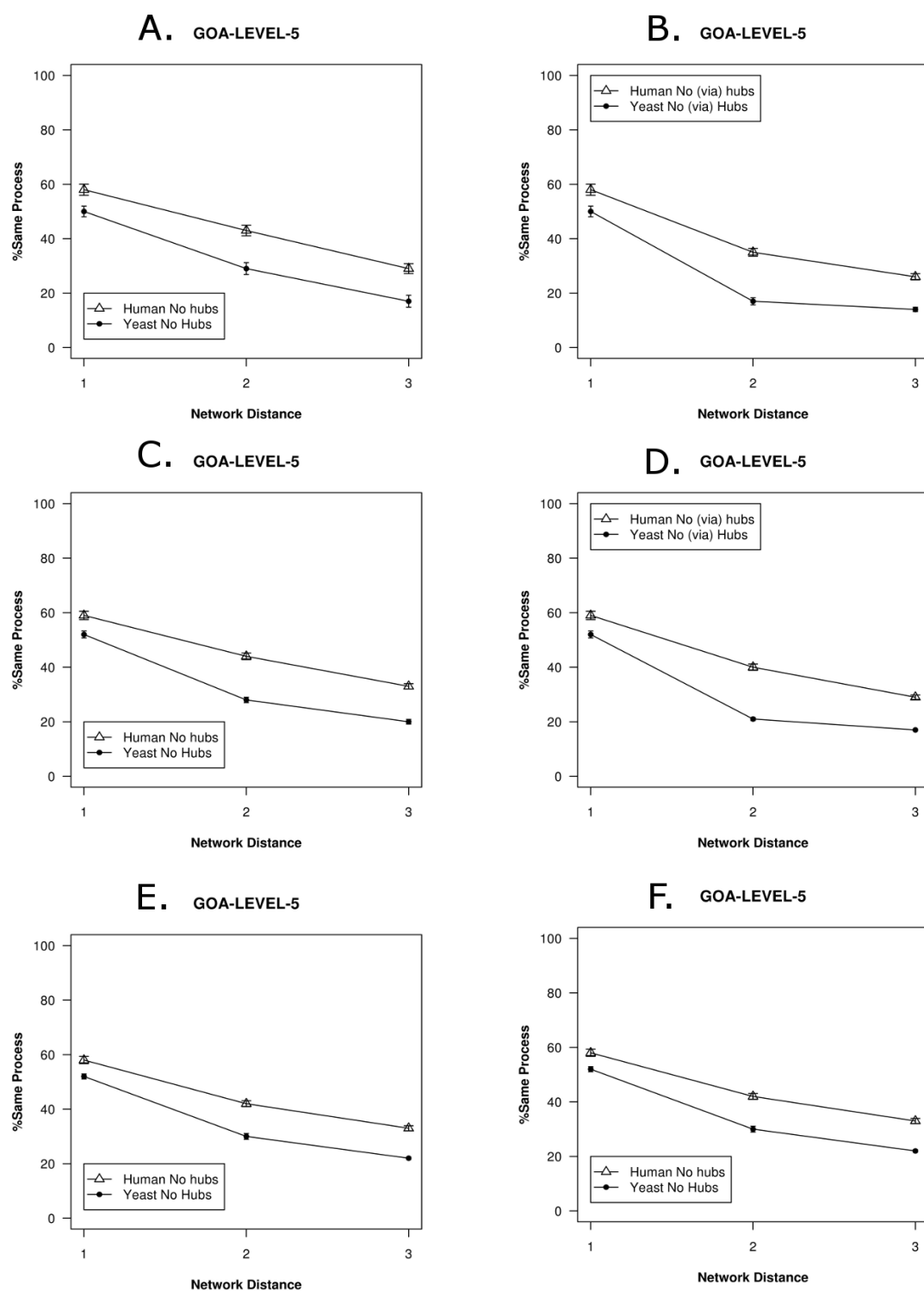


Figure S9 - Variation of biological process in hubless networks at GOA level 5

PCN values for biological process with the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 and GO level 5 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10 (A, B), >20 (C, D) and >50 (E, F) connections are removed from the interaction network (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”).

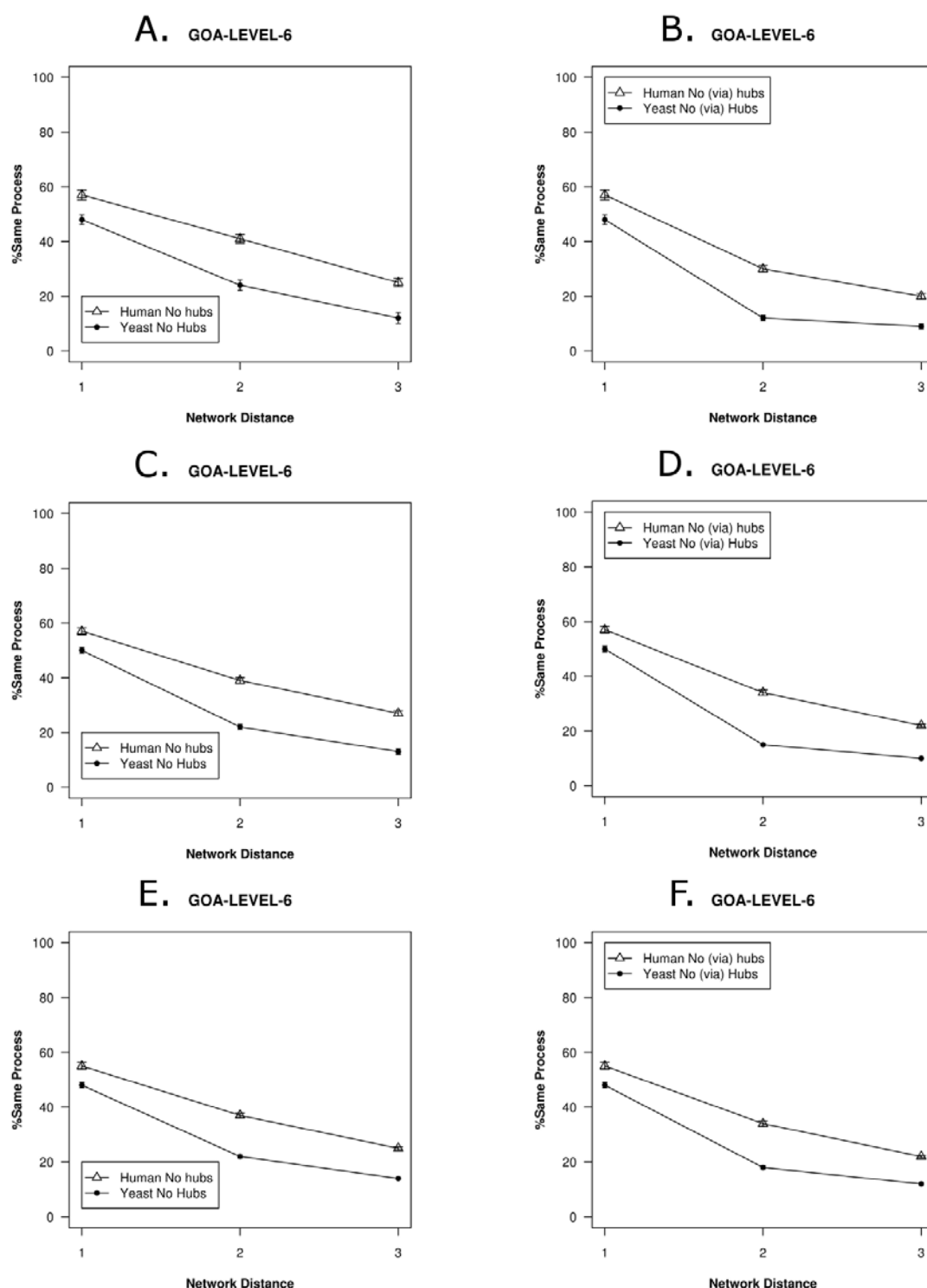


Figure S10 - Variation of biological process in hubless networks at GOA level 6

PCN values for biological process with the removal of hubs, i.e. no hubs and without (via) hubs at network distances 1-3 and GO level 6 are shown for *H. sapiens* and *S. cerevisiae* interaction networks, when hubs with >10 (A, B), >20 (C, D) and >50 (E, F) connections are removed from the interaction network (A, C and E: “without hubs”), or bypassed if they are terminal (B, D, F: “without (via) hubs”).

4.2 Conclusions

In this study, we analysed the influence of protein degree on GOA biological process and molecular functional relationships in the *S. cerevisiae* and *Homo sapiens* interaction networks. We examined the dependence of functional and process similarity of interacting protein pairs on the connectivity properties of the shortest path connecting them. We also investigated how functional inference from protein interactions depends on the level of GOA detail at which function is being studied.

Our results indicate that the tendency of interacting protein pairs in a network to have the same functions and processes decreases with increasing network distance *via* the shortest connecting path passing through hubs. Our network analysis results extend and complement the existing knowledge of the interactome.

Proximal proteins are most likely to have same molecular functions and biological processes for the network distances 1, 2 and 3 at GOA level 2 and 3, respectively. The molecular functions and biological processes of novel, uncharacterised proteins can thus be inferred from their relationships with other proteins in the interaction network, upto GOA levels 2 and 3, respectively..

Chapter 5: Identification of ovarian cancer associated genes using an integrated approach in a Boolean framework

5.1 Summary

Here, we have show application of interactome data along with other functional attributes in Boolean logic framework on gene expression dataset. Boolean logic has computational advantage for searching the sample space quickly and repeatedly. However, it falls short of arriving at the results where parameters for the sample search are not implicit, which happens often in the case of biological scenarios where cause and effect cannot not always be inferred directly. We have therefore selected the functional attributes based on observation between cancerous and non-cancerous genes reported from literature and weighted them suitably. This weighing schema is then encoded in the Boolean logic framework to rank differentially expressed genes. We have identified 17 genes to be differentially expressed, where ten genes are reported to be down-regulated *via* epigenetic inactivation and seven genes are up-regulated. Here, we report for the first-time that the over-expressed genes, *IRAK1*, *CHEK1* and *BUB1* may play an important role in ovarian cancer. Cancer is a complex disease which needs systems approach by integrating diverse biological information for the prognosis and therapy risk assessment using mechanistic approach to understand gene interaction in pathways, network and functional attributes to unravel the biological behaviour of tumours.

Identification of ovarian cancer associated genes using an integrated approach in a Boolean framework

Gaurav Kumar¹ and Shoba Ranganathan^{1,2, *}

¹ARC Centre of Excellence in Bioinformatics and Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney NSW 2109, Australia

²Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597

*Corresponding author

Email address:

GK: gaurav.kumar@mq.edu.au

SR: shoba.ranganathan@mq.edu.au

Abstract:

Background

Cancer is a complex disease where molecular mechanism remains elusive. A systems approach is needed to integrate diverse biological information for the prognosis and therapy risk assessment using mechanistic approach to understand gene interactions in pathways and networks and functional attributes to unravel the biological behaviour of tumors.

Results-

We weighted the functional attributes based on various functional properties observed between cancerous and non-cancerous genes reported from literature. This weighing schema is then encoded in the Boolean logic framework to rank differentially expressed genes. We have identified 17 genes to be differentially expressed from a total of 11,173 genes, where ten genes are reported to be down-regulated *via* epigenetic inactivation and seven genes are up-regulated. Here, we report for the first time that the overexpressed genes *IRAK1*, *CHEK1* and *BUB1* may play an important role in ovarian cancer.

Conclusion

We provided a workflow using Boolean logic schema for the identification of differential expressed genes by integrating diverse biological information. Using integrated approach resulted in the identification of genes as potential biomarker in ovarian cancer.

Background

The development of gene expression microarrays more than a decade ago has led the study of changes in the mRNA transcripts in disease-related tissues. These transcriptomic analyses under the microarrays experiments serves as the proxy for protein expression, and thereby revealed important properties of gene sets related to tissue-specificity [1, 2]. It has also facilitated the understanding of living cell at a systemic level by linking molecules to biological functions and thus bridging the genotype-to-phenotype gap *via* understanding the organisation of biological pathways [3] and network of protein interactions [4]. In a seminal review by Hanahan and Weinberg [5, 6], they introduce the six “*hallmarks of cancer*” and the seventh hallmark (*stemness*) of cancer was concluded through the gene expression analysis [7, 8]. In recent years researchers have made an effort to provide their microarray experiments for further studies through freely available public repositories such as Gene Expression Omnibus (GEO) [9] and ArrayExpress [10].

The knowledge acquired over the years of research suggests that the cancer cells harbour genetic defects that alter the balance of cell proliferation and cell death [11]. This has led to the compilation of cancer gene list, which is increasing steadily over the last two decades. It is also a highly variable disease with multiple heterogeneous genetic and epigenetic changes which makes it ideal to study by integrating data from multiple experiments to understand the causes at the cellular level. Therefore, identification and characterisation of susceptible genes associated with cancer is one of the greatest challenges in today’s biology and medical research. This challenge is partly due to the limitation of statistical methods on which a hypothesis about the value of statistical parameter is made for the detection of genes effects and their interactions, as multiple biological components work in the concerted fashion. Moreover, biological systems are highly enriched with examples of combinatorial regulation and influence as molecules in signalling pathway and gene regulatory pathway jointly influence the cellular state [12]. In

order to explore the combinatorial influence of multiple factors Boolean-based logic is a popular approach for the SNP association studies [13, 14] and in cancer [12, 15, 16] .

In this study, integrated systems approach is used to identify diseased-associated genes that are either not reported or poorly characterized in the ovarian tumor samples. We have estimated weights for the functional attributes associated with the known cancer gene list. These weights are then combined in Boolean logic schema to calculate the probability based rank associated with the differentially and non-differentially expressed genes. Finally, we have mapped high scoring ranks of differentially expressed genes on the co-expression gene interaction network to validate disease-associated genes (Figure 1). This study suggests that of the 17 shortlisted genes flagged as significant, the overexpressed genes *IRAK1*, *CHEK1* and *BUB1* may play an important role in ovarian cancer.

Methods

Identification of differential expressed genes

We extracted and analysed TCGA (The Cancer Genome Atlas) level 3 (Batch 9) ovarian serous cystadenocarcinoma data from Affymatrix platform [17]. TCGA gene expression data are normalised, annotated and validated for the expression variation relevant with the tissue types and with the type of array platforms, thus increasing the robustness in analysing expression data. Rather than a fold-change, we have calculated the differential expression of each gene by considering percentage of false prediction (pfp) $\leq 5\%$ using the RankProd R package [18]. RankProd uses the rank product non-parametric method to identify up/down-regulated genes under one condition against the other (in our case tumor vs. normal ovarian samples). This is based on the null hypothesis that the order of all items is random and the probability of finding a specific item among the top r of n items in a list is $p = r/n$. Multiplying these probability leads to the identification of the rank product $RP = \prod_i \frac{r_i}{n_i}$, where r_i is the rank of the item and n_i is the total number of items in

the i^{th} list. The smaller the RP value, the smaller the probability that the observed placement of the item at the top of the list is due to chance.

Relevant functional attributes in the diseased-condition

Although microarray measures the relative abundance of mRNA transcripts, their translated proteins are likely to be differentially present in diseased tissue. Therefore, we considered a total of six proteins functional attributes to capture Hanahan and Weinberg's "hallmarks of cancer" [5, 6], i.e. tissue specificity (TS), transcription factors (TF), post-translation modifications (PTM), protein kinases (PK), secreted proteins (SP) and Hub proteins in interactome (node connectivity >4) and the gene attribute of methylation (METH), in cancer vs. non-cancer associated genes.

Data integration from multiple experiments

We extracted functional attributes *via* a text-mining approach. The cancer gene list was obtained by combining data from the Atlas of Genetics and Cytogenetics in Oncology and Haematology [19] and Futreal *et al.* [20], with information related to secreted proteins, tissue-specificity and protein's post-translation modifications obtained from HPRD [21]. Human protein kinases were extracted from the Human Kinome [22]. Transcription factors were extracted from TRED [23], HPRD [21] and TargetMine [24] databases. Gene methylations in ovarian samples were extracted from the studies done by Mankoo *et al.* [25]. We considered the presence/absence of interaction in our high-confidence (HC) interactome dataset (detailed below) for differentially expressed genes, as biological pathways and networks of protein interactions are key paradigms to link molecules to biological functions. Therefore, interaction data were collected from BIND [26], BioGrid [27], DIP [28], HPRD [21], IntAct [29] and MINT [30] databases and merged into a single coherent interaction set after removing duplicate entries. Human protein interactions network were further analysed to create a HC (high-confidence) dataset by considering true interaction protein pairs as follow:

1. If binary interaction among proteins is known to be present in more than one databases.
2. Interacting protein pairs are true, if interaction is verified from more than one detection methods such as biochemical, biophysical, imaging techniques and/or protein complementation assay (PCA).
3. If interacting protein pairs have know protein domain interaction mentioned in 3did [31] and iPfam [32] databases.
4. PMIDs [33] were used as a proxy to support true interaction confirmed by more than one independent study.

These filters were used previously by us (Kumar, Cootes and Ranganathan, unpublished results) to define HC protein interaction set to study the network properties of molecular functions and biological processes of interacting proteins. In this study, scoring schema for interactions were considered for those protein nodes with more than four interactions, as this is empirical value of hubs suggested in gene co-expression stability in the analysis of protein interaction networks [34]. Therefore we weighted such highly connected protein nodes encoded by the known cancerous genes.

Weighting schema for Boolean-based probability calculation

We used phi-correlation (r_ϕ) as a measure of association between functional attributes in the cancerous genes. It is one of the powerful methods to detect the association strength between two categorical data having binary values. Moreover, computationally it is related with the chi-square (χ^2) as:

$$r_\phi = \sqrt{\frac{\chi^2}{N}}, \text{ where } N \text{ is the total number of genes.}$$

Scoring schema on the weighted functional attributes for ranking genes

We used the guilt-based-association algorithm proposed by Nagaraj and Reverter [15] for ranking the differentially expressed genes in ovarian samples, with our own set of Boolean variables representing relevant functional attributes in the diseased-condition. The particular combination across the seven Boolean variables i.e. functional attributes for a given differentially and non-differentially expressed genes, were decomposed into its root. For example, if a given gene has four known functional attributes, then 2^4 Boolean states are known to exist containing (2^4-1) roots, i.e. all possible combination of Boolean states at the positions of known functional attributes, excluding the Boolean values with all zero status. The probability of each root is simply the average sum of all the weights associated with known functional attributes calculated *via* r_ϕ . These root probabilities are then used to rank the differentially and non-differential expressed genes by summing up all the probability values associated with individual roots to rank the genes.

Validation set

We retrieved the raw expression data for 153 ovarian tumor samples from the Gene Expression Omnibus entry GSE1349, containing samples in four tumor stages [9]. Raw expression values for each probe were transformed to log-scale with base 2. Probe IDs were converted to Entrez Gene IDs using AILUN [35]. For genes with multiple probes, the probes with the highest variance across the samples were used to describe the expression value for the genes. Probes with multiple or without Gene IDs were removed from the analysis. Pearson's correlation coefficients were calculated to describe the pairwise gene co-expressions. We have taken a Pearson's coefficient >0.5 to represent a defined link between the co-expressed genes.

Results and Discussion

We used systems biology approach to integrate diverse data resources as described methods. 2157 genes were identified to be differentially expressed in tumor condition

using the RankProd R package at a $pfp \leq 5\%$. A total of 11173 genes were considered in the TCGA expression set. This analysis suggested that 1353 and 804 genes were up-regulated and down-regulated respectively (Figure 2). An estimation of weight was carried out *via* simple observation of known functional attributes present between cancerous and non-cancerous genes. Table 1 lists different functional attributes used as weights in this study. An odds-ratio analysis of differentially and non-differentially expressed genes showed no apparent differences (Supplementary Table 1). Therefore, it suggests that no single functional attribute can be selected alone in the classification of genes as a potential biomarker for the prognosis of the ovarian tumor condition. Moreover, cancer is well established as a disease model where the cellular system is abnormal leading to an uncontrolled cell division. Hence, a synergistic approach is needed to encapsulate the various functional attributes together for the understanding of the cancerous state. Figure 1 illustrates the workflow used for ranking genes. A Boolean framework for measuring unknown interactions between different biological entities and for the classification of genes in disease conditions have been reported by earlier studies [12, 15].

In this study seven functional attributes, such as epigenetic inactivation (CpG gene methylation), protein's post-translation modification, protein kinase, secreted protein, tissue-specificity, transcription factor and hub proteins in an interactome (protein node connectivity >4) were considered for the classification in the Boolean logic framework. We defined the Boolean logic for each gene, corresponding to the selected functional attributes (Table 2 and supplementary Table 2). These Boolean values were then decomposed to their roots to calculate the overall probability based on their functional attributes weights (see, section 2.6 of methods for detail). An empirical probability score greater than 0.5 was used as a cut-off to identify differential and non-differential gene expression as potential biomarkers. At this cut-off value, we were able to identify 17 differentially expressed genes (Table 2), whereas non-differential expression is noted for

48 genes (supplementary Table 2). In the TCGA expression dataset, we found seven (*IRAK1*, *STC2*, *CDC7*, *CHEK1*, *KLK6*, *BUB1* and *CHEK2*) and ten (*IGF1R*, *DAB2*, *IGFBP7*, *FOXL2*, *LCN2*, *CLU*, *LYN*, *PGR*, *AR* and *VIM*) genes to be up-regulated and down-regulated, respectively, using RankProd analysis. Figure 3 compares the known functional attributes present in proteins encoded by differentially and non-differentially expressed genes. Moreover, we have shown the verified the importance of these differentially expressed genes by mapping to their biological pathways (Supplementary Table 3).

Protein kinases

Protein kinases are important regulators of cell function and belong to a functionally diverse gene family. They affect the activity, localisation and overall function of other proteins by adding a phosphate group and thereby control the activity of cellular processes. Kinases are particularly important in signal transduction and co-ordination of complex functions such as cell cycle and pathological conditions. Identification of *IRAK1* as a differentially expressed gene in ovarian cancer suggests its important role in this disease. It is a putative Ser/Thr kinase known to partially interact with transcription factor, NF- κ B. Activation of NF- κ B leads to cell proliferation, survival and migration [36]. Over-expression of this gene suggests indirect cell survival and proliferation in the ovarian tumor condition. Similarly, *IGF1R* is a receptor with tyrosine kinase activity, which binds an insulin-like growth factor. It is over-expressed in most malignant tissue, acting as an anti-apoptotic agent by enhancing cell survival [37, 38]. *LYN* is a non-receptor tyrosine kinase, phosphorylating caspase 8, rendering it inactive and thereby assisting apoptosis of the inflammatory cell [39]. In the absence of the normal expression of *LYN*, active caspase 8 may prevent the tumor cells from undergoing apoptosis.

Other important kinases in cell survival and proliferation during tumorigenesis are associated with key cell cycle proteins. *CDC7* (cell-division cycle 7 homolog of *S.*

cerevisiae) and *BUB1* (budding uninhibited by benzimidazoles 1 homolog of *S. cerevisiae*) encode protein kinases which induce G1/S transition and are involved with the spindle checkpoint function, respectively during cell mitosis. *CDC7* is known to be overexpressed in the epithelial ovarian carcinoma, resulting in tumor progression, genomic instability and accelerated cell division [40]. On the other hand, *BUB1* overexpression induces aneuploidy and tumor formation [41]. *CHEK1* (checkpoint kinase 1) is another important cell-cycle molecule of Ser/Thr protein kinase family mediating signals from ATM and ATR cell cycle proteins involved in the DNA damage response and associated with chromatin in the meiotic prophase I. The importance of this protein in tumor invasiveness has been suggested by researchers in lung, bladder, liver, prostate, gastric, brain, cervical and colorectal cancers and B-cell lymphoma [42-44]. *CHEK2* (checkpoint kinase 2) is yet another important cell cycle protein which regulated key proteins during cell division. It interacted with *BRCA1* (- breast cancer 1) to restore survival in response to DNA damage with known association with endometrial cancer risk [45]. We observed overexpression of *IRAK1*, *BUB1*, *CDC7*, *CHEK1* and *CHEK2* genes in TCGA-samples at a high Boolean probability score of 0.607561, together with the co-expression of other key cell-cycle molecules in an independent validation expression set GSE1349 suggesting their association in ovarian cancer (Figure 4).

Serine proteases

Serine proteases are proteolytic enzymes, hydrolysing the peptide bond of protein substrates *via* a nucleophilic serine residue in the active site [46]. Serine proteases play diverse roles in human health, from non-specific digestion to highly regulated functions like embryonic development, immune response and blood coagulation. Moreover, insufficient or excess protease activity can promote significant pathologies like cancer, inflammation, hemophilia, heart attack, stroke, pancreatitis and parasite infection [47]. We suggest the potential use of *KLK6* (Kallikrein-related peptidase 6) as a potential biomarker

for ovarian cancer based on its high Boolean probability score (0.697808). *KLK6* is a serine protease with diverse functional roles inside the cell. It has been suggested that overexpression of this protein leads to the loss of cell-cell adhesion in skin cancer (melanoma) [48]. Moreover, a recent study reports the up-regulation of *KLK6* in colon cancer and its use as a potential biomarker and therapeutic agent [49].

Secreted proteins

Secreted proteins are secreted from the cell into the extracellular space and have important biological regulatory roles with the potential for therapeutics. *STC2* (Stanniocalcin 2) is a secreted homodimeric glycoprotein that is expressed in a variety of tissues. *STC2* is known to promote the epithelial-mesenchymal transition and invasiveness in human ovarian cancer under inadequate oxygen supply to the tissue [50]. Our results show that *STC2* is a significant up-regulated gene, promoting ovarian cancer. On the other hand, *CLU* (clusterin) and *LCN2* (lipocalin2) are down-regulated genes in our analysis. *CLU* encodes a protein which is secreted under stress conditions, that functions as a strong anti-migratory and anti-invasive agent by inducing the destruction of the actin cytoskeleton inside the cell [51]. The decreased expression of *CLU* thus promotes the cancerous diseased condition. *LCN2* encodes a 25kDa secretory protein involved with iron-transportation and contributes to endometrial carcinoma [52]. Moreover, it is a key molecule in various signalling pathways (Supplementary Table 3). Down-regulation of *LCN2* due to epigenetic inactivation may lead to ovarian carcinoma.

Other types of proteins

We observed down-regulation of genes with high probability associated with phosphoproteins, transcription factors and receptors due to epigenetic inactivation. Phosphoprotein *DAB2* is a mitogen-responsive agent, acting as tumor suppressor in normal ovarian epithelial cells and down-regulation of this gene modulates the TGF- β signalling pathway [53]. *FOXL2* (forkhead box L2) encodes a transcription factor which helps in the

normal development of ovarian tissue -. *IGFBP7* (insulin-like growth factor binding domain) is known as the tumor suppressor gene, leading to lung cancer due to the epigenetic inactivation [55]. *PGR* (progesterone receptor) encodes a protein playing a central role in the reproductive system by maintaining progesterone levels and ensuring normal pregnancy (Table 3). *AR* (progesterone receptor) encodes a protein which functions as a steroid hormone-activated transcription factor and has been shown to be involved in prostate cancer [56]. *VIM* (vimentin) encodes a protein that is responsible for maintaining cell shape, integrity of the cytoplasm and stabilizing cytoskeleton interaction. Thus, the decreased expression of these genes could be indicative of ovarian cancer.

Conclusions

We have statistically integrated gene expression and protein interaction data by combining weights in a Boolean framework to identify high scoring differentially expressed genes in ovarian tumor samples. This has resulted in the identification of important genes associated with critical biological processes. We identified 17 differentially expressed genes from a dataset of 11,173 genes, where seven and ten genes were up- and down-regulated, respectively with significant probability score in a Boolean logic schema. We report three genes (*IRAK1*, *CHEK1* and *BUB1*) to be significant in ovarian tumor samples for the first time, to the best of our knowledge. Our results demonstrate the significance of multiple data types and knowledge-guided integration of diverse biological information to understand the molecular mechanisms associated in ovarian cancer and their application in the discovery of biomarkers.

Authors' contributions

Conceived and designed the experiment: GK. Data collected and analysed: GK.

Manuscript has been written and finalised by GK and SR.

Acknowledgements

GK is thankful to Drs. Nagaraj and Reverter, CSIRO for providing a copy of the Boolean algorithm. This research was supported by the Macquarie University Research Scholarship (MQRES) to GK and the ARC Centre of Excellence in Bioinformatics grant (CE0348221) to SR.

References

1. Muller FJ, Laurent LC, Kostka D, Ulitsky I, Williams R, Lu C, Park IH, Rao MS, Shamir R, Schwartz PH *et al*: **Regulatory networks define phenotypic classes of human stem cell lines.** *Nature* 2008, **455**(7211):401-405.
2. Walker JR, Su AI, Self DW, Hogenesch JB, Lapp H, Maier R, Hoyer D, Bilbe G: **Applications of a rat multiple tissue gene expression data set.** *Genome Res* 2004, **14**(4):742-749.
3. Magwene PM, Kim J: **Estimating genomic coexpression networks using first-order conditional independence.** *Genome Biol* 2004, **5**(12):R100.
4. Kar G, Gursoy A, Keskin O: **Human cancer protein-protein interaction network: a structural perspective.** *PLoS Comput Biol* 2009, **5**(12):e1000601.
5. Hanahan D, Weinberg RA: **The hallmarks of cancer.** *Cell* 2000, **100**(1):57-70.
6. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation.** *Cell* 2011, **144**(5):646-674.
7. Wong DJ, Segal E, Chang HY: **Stemness, cancer and cancer stem cells.** *Cell Cycle* 2008, **7**(23):3622-3624.
8. Glinsky GV: **"Stemness" genomics law governs clinical behavior of human cancer: implications for decision making in disease management.** *J Clin Oncol* 2008, **26**(17):2846-2853.
9. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM *et al*: **NCBI GEO: archive for functional genomics data sets--10 years on.** *Nucleic Acids Res* 2011, **39**(Database issue):D1005-1010.

10. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A *et al*: **Oncogenic pathway signatures in human cancers as a guide to targeted therapies.** *Nature* 2006, **439**(7074):353-357.
11. Zhang W, Laborde PM, Coombes KR, Berry DA, Hamilton SR: **Cancer genomics: promises and complexities.** *Clin Cancer Res* 2001, **7**(8):2159-2167.
12. Mukherjee S, Pelech S, Neve RM, Kuo WL, Ziyad S, Spellman PT, Gray JW, Speed TP: **Sparse combinatorial inference with an application in cancer biology.** *Bioinformatics* 2009, **25**(2):265-271.
13. Kooperberg C, Ruczinski I: **Identifying interacting SNPs using Monte Carlo logic regression.** *Genet Epidemiol* 2005, **28**(2):157-170.
14. Schwender H, Ruczinski I: **Logic regression and its extensions.** *Adv Genet* 2010, **72**:25-45.
15. Nagaraj SH, Reverter A: **A Boolean-based systems biology approach to predict novel genes associated with cancer: Application to colorectal cancer.** *BMC Syst Biol* 2011, **5**:35.
16. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: **Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer.** *Am J Hum Genet* 2001, **69**(1):138-147.
17. **Integrated genomic analyses of ovarian carcinoma.** *Nature* 2011, **474**(7353):609-615.
18. Breitling R, Armengaud P, Amtmann A, Herzyk P: **Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.** *FEBS Lett* 2004, **573**(1-3):83-92.
19. Huret JL, Dessen P, Bernheim A: **Atlas of Genetics and Cytogenetics in Oncology and Haematology, year 2003.** *Nucleic Acids Res* 2003, **31**(1):272-274.

20. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes.** *Nat Rev Cancer* 2004, **4**(3):177-183.
21. Prasad TS, Kandasamy K, Pandey A: **Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology.** *Methods Mol Biol* 2009, **577**:67-79.
22. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S: **The protein kinase complement of the human genome.** *Science* 2002, **298**(5600):1912-1934.
23. Jiang C, Xuan Z, Zhao F, Zhang MQ: **TRED: a transcriptional regulatory element database, new entries and other development.** *Nucleic Acids Res* 2007, **35**(Database issue):D137-140.
24. Chen YA, Tripathi LP, Mizuguchi K: **TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery.** *PLoS One* 2011, **6**(3):e17844.
25. Mankoo PK, Shen R, Schultz N, Levine DA, Sander C: **Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles.** *PLoS One* 2011, **6**(11):e24709.
26. Gilbert D: **Biomolecular interaction network database.** *Brief Bioinform* 2005, **6**(2):194-198.
27. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X *et al*: **The BioGRID Interaction Database: 2011 update.** *Nucleic Acids Res* 2011, **39**(Database issue):D698-704.
28. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Res* 2004, **32**(Database issue):D449-451.

29. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R *et al*: **IntAct--open source resource for molecular interaction data**. *Nucleic Acids Res* 2007, **35**(Database issue):D561-565.
30. Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G: **MINT, the molecular interaction database: 2009 update**. *Nucleic Acids Res* 2010, **38**(Database issue):D532-539.
31. Stein A, Ceol A, Aloy P: **3did: identification and classification of domain-based interactions of known three-dimensional structure**. *Nucleic Acids Res* 2011, **39**(Database issue):D718-723.
32. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2010, **38**(Database issue):D211-222.
33. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C *et al*: **The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data**. *Nat Biotechnol* 2004, **22**(2):177-183.
34. Patil A, Nakai K, Kinoshita K: **Assessing the utility of gene co-expression stability in combination with correlation in the analysis of protein-protein interaction networks**. *BMC Genomics* 2011, **12** (Suppl 3):(S19).
35. Chen R, Li L, Butte AJ: **AILUN: reannotating gene expression data automatically**. *Nat Methods* 2007, **4**(11):879.
36. Papanikolaou V, Iliopoulos D, Dimou I, Dubos S, Kappas C, Kitsiou-Tzeli S, Tsezou A: **Survivin regulation by HER2 through NF-kappaB and c-myc in irradiated breast cancer cells**. *J Cell Mol Med* 2011, **15**(7):1542-1550.

37. Shiratsuchi I, Akagi Y, Kawahara A, Kinugasa T, Romeo K, Yoshida T, Ryu Y, Gotanda Y, Kage M, Shirouzu K: **Expression of IGF-1 and IGF-1R and their relation to clinicopathological factors in colorectal cancer.** *Anticancer Res* 2011, **31**(7):2541-2545.
38. Mejia W, Castro C, Umana A, de Castro C, Riveros T, Sanchez-Gomez M: **[Insulin-like growth factor receptor I signaling in a breast cancer cell line].** *Biomedica* 2010, **30**(4):551-558.
39. Jia SH, Parodo J, Kapus A, Rotstein OD, Marshall JC: **Dynamic regulation of neutrophil survival through tyrosine phosphorylation or dephosphorylation of caspase-8.** *J Biol Chem* 2008, **283**(9):5402-5413.
40. Kulkarni AA, Kingsbury SR, Tudzarova S, Hong HK, Loddo M, Rashid M, Rodriguez-Acebes S, Prevost AT, Ledermann JA, Stoeber K *et al*: **Cdc7 kinase is a predictor of survival and a novel therapeutic target in epithelial ovarian carcinoma.** *Clin Cancer Res* 2009, **15**(7):2417-2425.
41. Ricke RM, Jeganathan KB, van Deursen JM: **Bub1 overexpression induces aneuploidy and tumor formation through Aurora B kinase hyperactivation.** *J Cell Biol* 2011, **193**(6):1049-1064.
42. Thorsen K, Schepeler T, Oster B, Rasmussen MH, Vang S, Wang K, Hansen KQ, Lamy P, Pedersen JS, Eller A *et al*: **Tumor-specific usage of alternative transcription start sites in colorectal cancer identified by genome-wide exon array analysis.** *BMC Genomics* 2011, **12**:505.
43. Hoglund A, Nilsson LM, Muralidharan SV, Hasvold LA, Merta P, Rudelius M, Nikolova V, Keller U, Nilsson JA: **Therapeutic implications for the induced levels of chk1 in myc-expressing cancer cells.** *Clin Cancer Res* 2011, **17**(22):7067-7079.

44. Mazumder ' Indra D, Mitra S, Singh RK, Dutta S, Roy A, Mondal RK, Basu PS, Roychoudhury S, Panda CK: **Inactivation of CHEK1 and E124 are associated with the development of invasive cervical carcinoma: Clinical and prognostic implications.** *Int J Cancer* 2010.
45. O'Mara TA, Ferguson K, Fahey P, Marquart L, Yang HP, Lissowska J, Chanock S, Garcia-Closas M, Thompson DJ, Healey CS *et al*: **CHEK2, MGMT, SULT1E1 and SULT1A1 polymorphisms and endometrial cancer risk.** *Twin Res Hum Genet* 2011, **14**(4):328-332.
46. Di Cera E: **Serine proteases.** *IUBMB Life* 2009, **61**(5):510-515.
47. Page MJ, Di Cera E: **Serine peptidases: classification, structure and function.** *Cell Mol Life Sci* 2008, **65**(7-8):1220-1236.
48. Rezzè GG, Fregnani JH, Duprat J, Landman G: **Cell adhesion and communication proteins are differentially expressed in melanoma progression model.** *Hum Pathol* 2011, **42**(3):409-418.
49. Kim JT, Song EY, Chung KS, Kang MA, Kim JW, Kim SJ, Yeom YI, Kim JH, Kim KH, Lee HG: **Up-regulation and clinical significance of serine protease kallikrein 6 in colon cancer.** *Cancer* 2010.
50. Law AY, Wong CK: **Stanniocalcin-2 promotes epithelial-mesenchymal transition and invasiveness in hypoxic human ovarian cancer cells.** *Exp Cell Res* 2010, **316**(20):3425-3434.
51. Moretti RM, Mai S, Montagnani Marelli M, Rizzi F, Bettuzzi S, Limonta P: **Molecular mechanisms of the antimetastatic activity of nuclear clusterin in prostate cancer cells.** *Int J Oncol* 2011, **39**(1):225-234.
52. Miyamoto T, Asaka R, Suzuki A, Takatsu A, Kashima H, Shiozawa T: **Immunohistochemical detection of a specific receptor for lipocalin2 (solute**

carrier family 22 member 17, SLC22A17) and its prognostic significance in endometrial carcinoma. *Exp Mol Pathol* 2011, **91**(2):563-568.

53. Hannigan A, Smith P, Kalna G, Lo Nigro C, Orange C, O'Brien DI, Shah R, Syed N, Spender LC, Herrera B *et al*: **Epigenetic downregulation of human disabled homolog 2 switches TGF-beta from a tumor suppressor to a tumor promoter.** *J Clin Invest* 2010, **120**(8):2842-2857.
54. Kuo FT, Bentsi-Barnes IK, Barlow GM, Pisarska MD: **Mutant Forkhead L2 (FOXL2) proteins associated with premature ovarian failure (POF) dimerize with wild-type FOXL2, leading to altered regulation of genes associated with granulosa cell differentiation.** *Endocrinology* 2011, **152**(10):3917-3929.
55. Chen Y, Cui T, Knosel T, Yang L, Zoller K, Petersen I: **IGFBP7 is a p53 target gene inactivated in human lung cancer by DNA hypermethylation.** *Lung Cancer* 2011, **73**(1):38-44.
56. Zhu LY, Zhong KB, Lu SX, He LY: **[Vinculin and the androgen receptor in prostate cancer: expressions and correlations].** *Zhonghua Nan Ke Xue* 2010, **16**(9):794-798.

Figures

Figure 1: Schematic representation for ranking genes in a Boolean logic framework.

Schematic representation of the workflow used to rank genes in a Boolean framework for identifying potential biomarkers in ovarian cancer.

Figure 2: Differential gene expression in TCGA ovarian dataset

Affymatrix TCGA gene expression dataset in ovarian tumor samples (class 1) vs. normal samples (class 2). RankProd analysis of differential gene expression at percentage of false prediction (pfp) $\leq 5\%$ is shown.

Figure 3: Functional attributes presented in various proteins encoded by differential/non-differential gene expression in TCGA data.

Histogram representing functional attributes such as Meth (Methylation), PK (Protein-Kinase), TF (Transcription Factor), TS (Tissue-specificity), PTM (Post-translation modification), SP (secreted-proteins) and Hub (Protein interaction where node connectivity > 4) presented in protein encoded by differential/non-differential expressed genes.

Figure 4: Co-expression of four up-regulated genes

Schematic representation of co-expressed of four up-regulated genes. Edges are colour-coded to highlight the range of pearson's correlation coefficient in co-expression network: pink (0.05-0.55), green (0.55-0.60), red (0.60-0.65), blue (0.65-0.70) and black (> 0.70).

Tables:

Table-1: Phi-correlation (r_ϕ) weights calculated for the functional attributes such as methylation, post-translation modification, protein kinase, secretory proteins, tissue-specificity, protein interaction nodes with connectivity >4 and transcription-factor in cancerous vs. non-cancerous genes associated with ovarian cancerous tumor samples.

Table 2: Boolean-based probability score for ranking 17 differentially expressed genes.

Supplementary data:

Supplementary Table 1: Differential/Non-differential gene expression for various functional attributes.

Supplementary Table 2: Boolean-based probability score for ranking 48 non-differentially expressed genes.

Supplementary Table 3: Statistically significant pathway analysis from the NCI-nature *PID* (Pathway Interaction Database) of the 17 differentially expressed genes in various biological pathways.

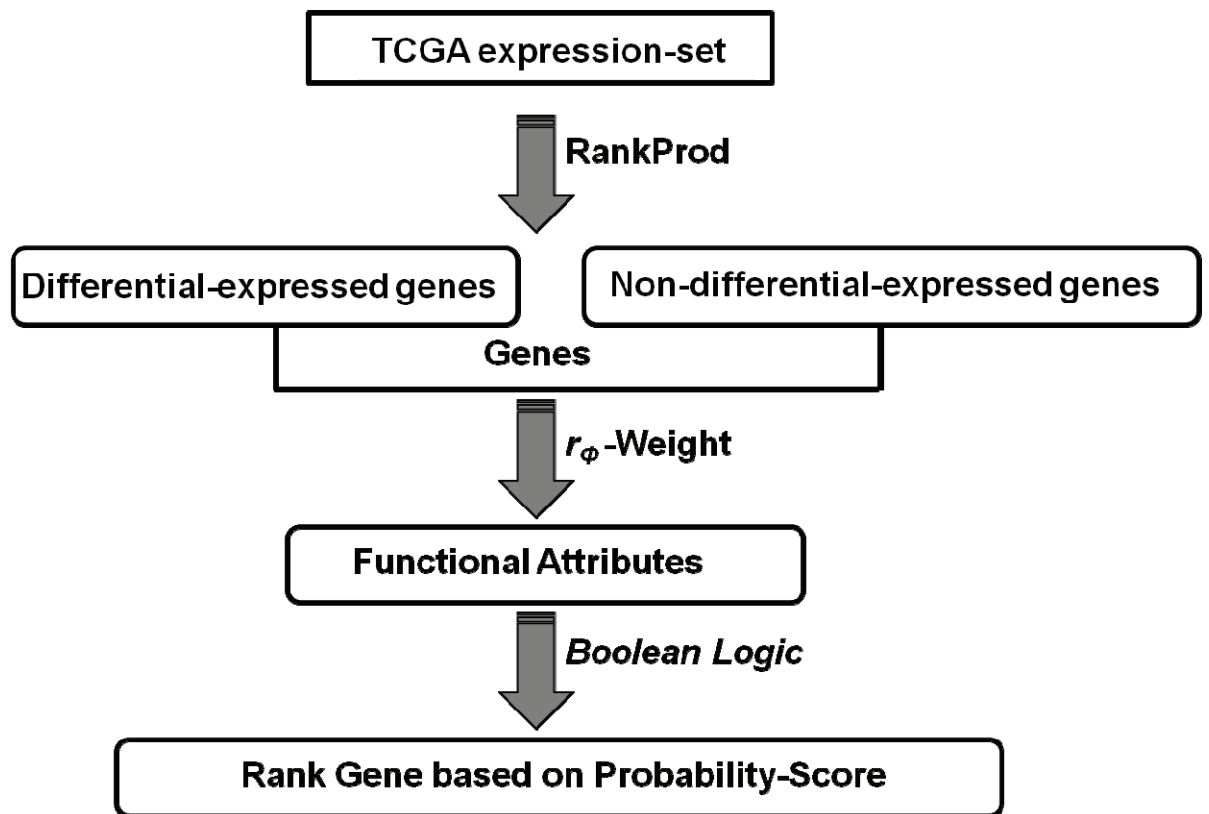


Figure 1: Schematic representation for ranking genes in a Boolean logic framework. Schematic representation of the workflow used to rank genes in a Boolean framework for identifying potential biomarkers in ovarian cancer.

Figure 2: Differential gene expression in TCGA ovarian dataset

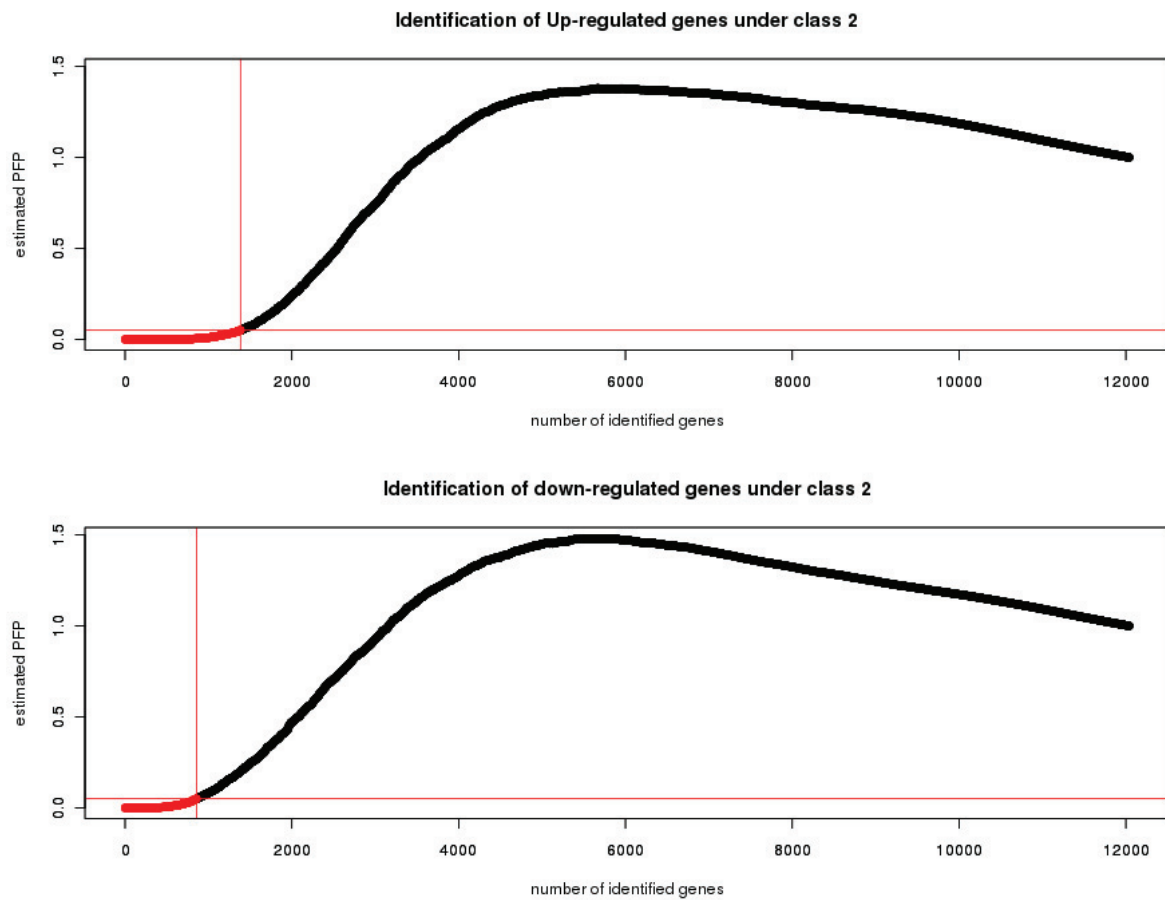


Figure 2: Differential gene expression in TCGA ovarian dataset

Affymatrix TCGA gene expression dataset in ovarian tumor samples (class 1) vs. normal samples (class 2). RankProd analysis of differential gene expression at percentage of false prediction (pfp) $\leq 5\%$ is shown.

Distribution of proteins encoded by differential/Non-differential expressed genes

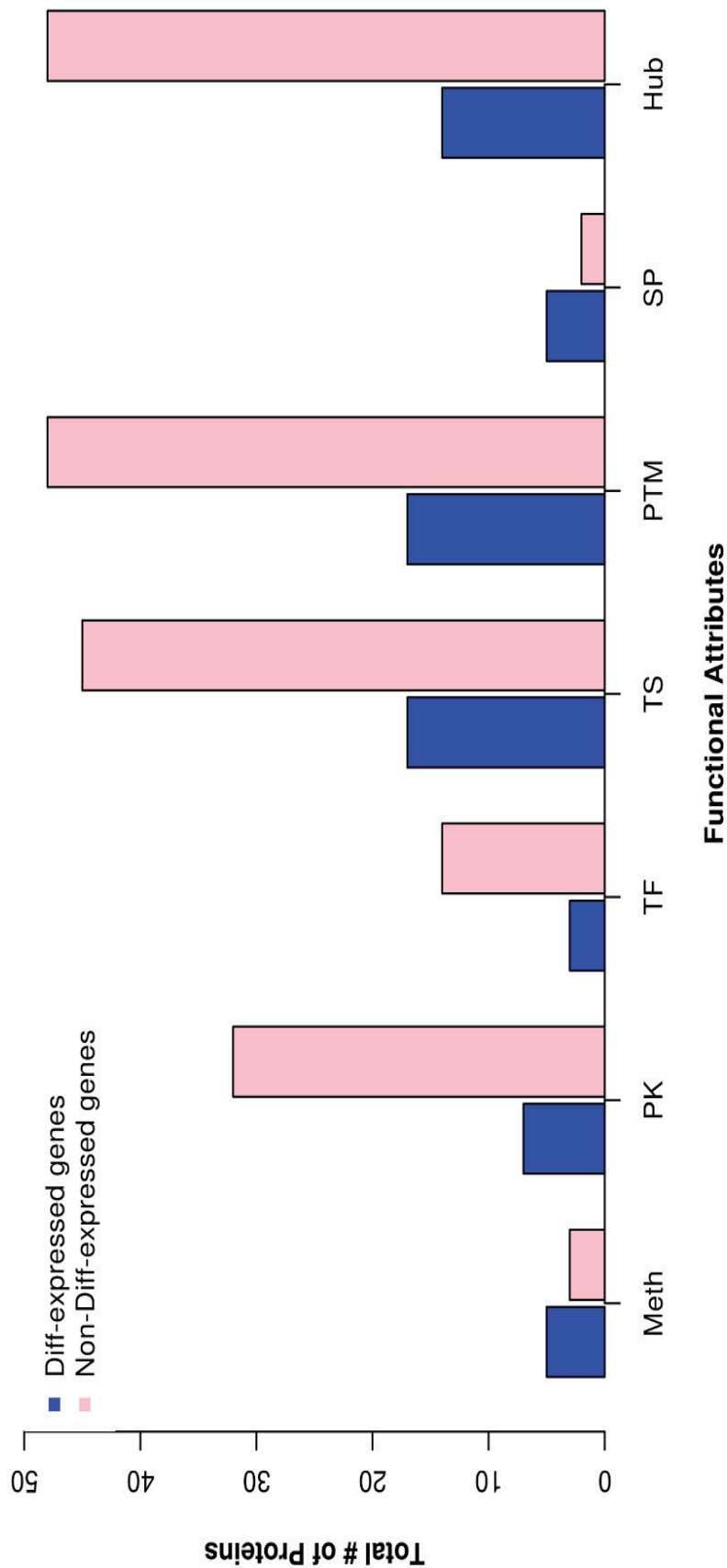
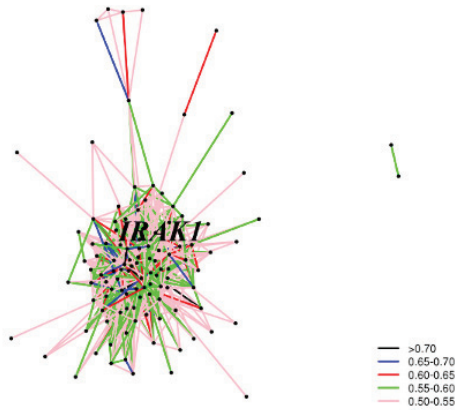


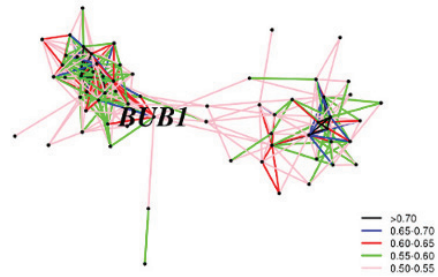
Figure 3: Functional attributes presented in various proteins encoded by differential/non-differential gene expression in TCGA data.

Histogram representing functional attributes such as Meth (Methylation), PK (Protein-Kinase), TF (Transcription Factor), TS (Tissue-specificity), PTM (Post-translation modification), SP (secreted-proteins) and Hub (Protein interaction where node connectivity > 4) presented in protein encoded by differential/non-differential expressed genes.

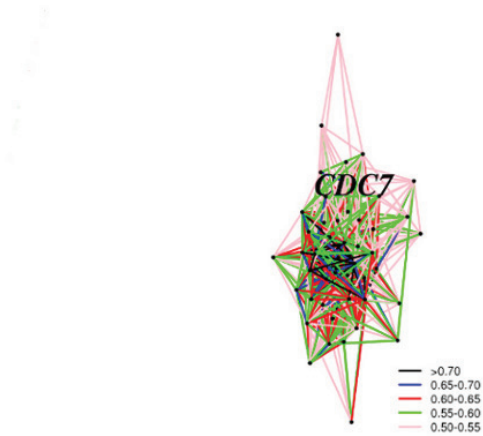
IRAK1 co-expression network



BUB1 co-expression network



CDC7 co-expression network



CHEK1 co-expression network

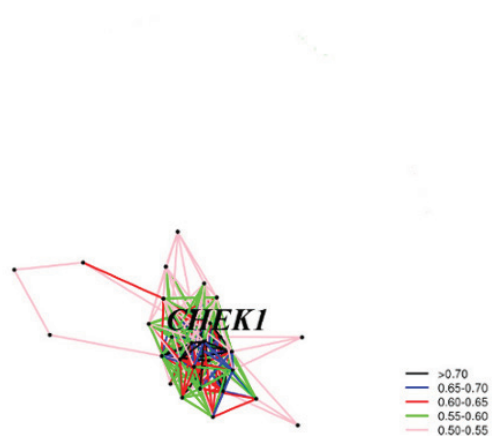


Figure 4: Co-expression of four up-regulated genes

Schematic representation of co-expressed of four up-regulated genes. Edges are colour-coded to highlight the range of pearson's correlation coefficient in co-expression network: pink (0.05-0.55), green (0.55-0.60), red (0.60-0.65), blue (0.65-0.70) and black (> 0.70).

Table 1: Phi-correlation (r_ϕ) weights calculated for the functional attributes such as methylation, post-translation modifications, protein kinase, secretory proteins, tissue-specificity, protein interaction nodes with connectivity >4 and transcription factor in cancerous vs. non-cancerous genes associated with ovarian cancerous tumor samples.

Functional Attributes	Phi-correlation value	<i>P-value</i>
Methylation	0.021944	0.0803
Post-translation modifications	0.046598	0.0004
Protein kinase	0.037870	0.0030
Secretory proteins	0.036727	0.0026
Tissue specificity	0.038675	0.0019
Interactome (node connectivity >4)	0.072986	0.0001
Transcription factor	0.048745	0.0002

Table 2: Boolean-based probability score for ranking 17 differentially expressed genes.

<i>Gene symbol</i>	<i>Gene ID</i>	<i>Up</i>	<i>Down</i>	<i>Boolean values</i>	<i>Rank</i>
<i>KLK6</i>	5653	1	0	1011001	0.697808
<i>IRAK1</i>	3654	1	0	0111010	0.607561
<i>CDC7</i>	8317	1	0	0111010	0.607561
<i>CHEK1</i>	1111	1	0	0111010	0.607561
<i>BUB1</i>	699	1	0	0111010	0.607561
<i>CHEK2</i>	11200	1	0	0111010	0.607561
<i>STC2</i>	8614	1	0	1011010	0.584684
<i>DAB2</i>	1601	0	1	0011011	0.743532
<i>VIM</i>	7431	0	1	0011011	0.743532
<i>FOXL2</i>	668	0	1	0011101	0.735481
<i>LNC2</i>	3934	0	1	1011001	0.697808
<i>PGR</i>	5241	0	1	0011110	0.644578
<i>AR</i>	367	0	1	0011110	0.644578
<i>IGF1R</i>	3480	0	1	0111010	0.607561
<i>LYN</i>	4067	0	1	0111010	0.607561
<i>IGFBP7</i>	3490	0	1	1011010	0.584684
<i>CLU</i>	1191	0	1	1011010	0.584684

Supplementary Table 1: Differential/Non-differential gene expression for various functional attributes.

Functional Attributes		Differential gene expression				Non-differential gene expression			
		Average Rank	Min.	Max.	Odd-Ratio	Average Rank	Min.	Max.	Odd-Ratio
Secretome	+	0.093	0.000	0.698	2.111	0.051	0.000	0.585	1.246
	-	0.044	0.000	0.744		0.041	0.000	0.776	
Protein kinase	+	0.216	0.037	0.608	5.194	0.196	0.037	0.739	5.515
	-	0.042	0.000	0.744		0.036	0.000	0.776	
Tissue specificity	+	0.112	0.038	0.744	3.647	0.102	0.038	0.744	3.768
	-	0.031	0.000	0.374		0.027	0.000	0.766	
Post-translation modification	+	0.131	0.039	0.744	6.389	0.126	0.039	0.766	6.517
	-	0.021	0.000	0.372		0.019	0.000	0.393	
Transcription Factor	+	0.172	0.047	0.735	3.909	0.169	0.047	0.776	4.584
	-	0.044	0.000	0.744		0.037	0.000	0.744	
Protein-interaction (nodes >4)	+	0.157	0.049	0.744	5.744	0.146	0.049	0.776	6.597
	-	0.027	0.000	0.735		0.022	0.000	0.369	
Methylation	+	0.187	0.073	0.744	4.388	0.148	0.000	0.776	3.807
	-	0.043	0.000	0.645		0.039	0.000	0.645	

Supplementary Table 2: Boolean-based probability score for ranking the 48 non-differentially expressed genes.

<i>Gene symbol</i>	<i>Gene-ID</i>	<i>Boolean values</i>	<i>Rank</i>
<i>ESR1</i>	2099	0001111	0.776265
<i>PRLR</i>	5618	0011011	0.743532
<i>WEE1</i>	7465	0101011	0.739246
<i>SMAD2</i>	4087	0011110	0.644578
<i>CDC5L</i>	988	0011110	0.644578
<i>HIF1A</i>	3091	0011110	0.644578
<i>BRCA1</i>	672	0011110	0.644578
<i>JUNB</i>	3726	0011110	0.644578
<i>RUNX1</i>	861	0011110	0.644578
<i>BACH1</i>	571	0011110	0.644578
<i>MAX</i>	4149	0011110	0.644578
<i>GTF2A1</i>	2957	0011110	0.644578
<i>SREBF1</i>	6720	0011110	0.644578
<i>TP73</i>	7161	0011110	0.644578
<i>CDK7</i>	1022	0101110	0.640294
<i>AKT1</i>	207	0111010	0.607561
<i>HIPK2</i>	28996	0111010	0.607561
<i>ERBB2</i>	2064	0111010	0.607561
<i>KIT</i>	3815	0111010	0.607561
<i>JAK3</i>	3718	0111010	0.607561
<i>TBK1</i>	29110	0111010	0.607561
<i>PAK4</i>	10298	0111010	0.607561
<i>MAP3K14</i>	9020	0111010	0.607561
<i>BRD4</i>	23476	0111010	0.607561
<i>TRIM28</i>	10155	0111010	0.607561
<i>LATS1</i>	9113	0111010	0.607561
<i>MAPK14</i>	1432	0111010	0.607561
<i>STK11</i>	6794	0111010	0.607561
<i>TEC</i>	7006	0111010	0.607561
<i>FGFR1</i>	2260	0111010	0.607561
<i>STK16</i>	8576	0111010	0.607561
<i>MAP3K5</i>	4217	0111010	0.607561
<i>MAP3K7</i>	6885	0111010	0.607561
<i>IKBKB</i>	3551	0111010	0.607561
<i>PTK2</i>	5747	0111010	0.607561
<i>PTK2B</i>	2185	0111010	0.607561
<i>FGFR3</i>	2261	0111010	0.607561
<i>JAK2</i>	3717	0111010	0.607561
<i>ATR</i>	545	0111010	0.607561

<i>Gene symbol</i>	<i>Gene-ID</i>	<i>Boolean values</i>	<i>Rank</i>
<i>FLT1</i>	2321	0111010	0.607561
<i>FGFR2</i>	2261	0111010	0.607561
<i>DYRK1A</i>	1859	0111010	0.607561
<i>PRKCD</i>	5580	0111010	0.607561
<i>ERBB4</i>	2066	0111010	0.607561
<i>SRC</i>	6714	0111010	0.607561
<i>SERPINA1</i>	5265	1011010	0.584684
<i>SMAD1</i>	5265	1011010	0.584684
<i>F2</i>	2147	1011010	0.584684

Table 3: Statistically significant NCI-Nature Pathway Interaction Database analysis of the 17 differentially expressed genes. PID is a database with biomolecular interactions and cellular processes assembled into authoritative human signalling pathways.

<i>Gene symbol*</i>	<i>Gene ID</i>	<i>PID-Pathways and associated P-Value</i>
<i>AR</i>	367	Regulation of nuclear SMAD2/3 signaling (1.88e-04); Nongenotropic Androgen signaling (2.80e-04); FOXA1 transcription factor network (1.06e-03); Coregulation of Androgen receptor activity (3.84e-03); Notch-mediated HES/HEY network (1.33e-02); Regulation of Androgen receptor activity (1.62e-02); Regulation of nuclear beta catenin signaling and target gene transcription (3.70e-01)
<i>BUB1</i>	699	p73 transcription factor network (1.25e-03); Aurora B signaling (5.66e-02); PLK1 signaling events (7.34e-02)
<i>CHEK1</i>	1111	p53 pathway (1.56e-07); p73 transcription factor network (1.25e-03); Circadian rhythm pathway (1.16e-02); Fanconi anemia pathway (1.33e-02); ATR signaling pathway (5.89e-02)
<i>CHEK2</i>	11200	p53 pathway (1.56e-07); ATM pathway (5.03e-03); PLK3 signaling events (6.08e-02); FOXM1 transcription factor network (6.37e-02)
<i>CLU</i>	1191	Validated targets of C-MYC transcriptional repression (8.87e-04)
<i>DAB2</i>	1601	TGF-beta receptor signaling (1.78e-02)
<i>IGF1R</i>	3480	IGF1 pathway (2.80e-04); Plasma membrane estrogen receptor signaling (8.91e-04); Integrins in angiogenesis (1.19e-03); SHP2 signaling (2.12e-02); Posttranslational regulation of adherens junction stability and disassembly (7.84e-02); Stabilization and expansion of the E-cadherin adherens junction (2.98e-01)
<i>IRAK1</i>	3654	IL1-mediated signaling events (5.88e-03); Endogenous TLR signaling (2.84e-02); p75(NTR)-mediated signaling (3.18e-02)
<i>KLK6</i>	5653	Alpha-synuclein signaling (1.78e-05)
<i>LYN</i>	4067	Glypican 1 network (7.49e-06); Alpha-synuclein signaling (1.78e-05); LPA receptor mediated events (5.22e-05); Signaling events mediated by PTP1B (1.89e-04); Signaling events mediated by Stem cell factor receptor (c-Kit) (2.25e-04); Thromboxane A2 receptor signaling (2.67e-04); EPO signaling pathway (4.47e-04); CXCR4-mediated signaling events (6.26e-04); BCR signaling pathway (6.46e-04); Fc-epsilon receptor I signaling in mast cells (3.43e-03); Ephrin B reverse signaling (3.56e-03); Regulation of p38-alpha and p38-beta (3.90e-03); GMCSF-mediated signaling events (5.88e-03); amb2 Integrin signaling (8.35e-03); IL5-mediated signaling events (9.01e-03); IL8- and CXCR1-mediated signaling events (3.44e-02); PDGFR-beta signaling pathway (3.72e-02); EPHA forward signaling (4.51e-02); IL8- and CXCR2-mediated signaling events (4.73e-02); Class I PI3K signaling events (8.09e-02)
<i>PGR</i>	5241	Validated nuclear estrogen receptor alpha network (2.98e-02); Cellular roles of Anthrax toxin (1.88e-01)
<i>VIM</i>	7431	Caspase cascade in apoptosis (1.55e-02); Aurora B signaling (5.66e-02)

* *CDC7*, *FOXL2*, *IGFBP7*, *LCN2* and *STC2* do not map to statistically significant pathways and therefore are not listed above.

5.2 Conclusions

We statistically integrated genomic and proteomic data by combining weights in a Boolean framework to identify high scoring differential expressed genes in the ovarian tumour samples. This lead to the identification of important genes associated with the critical biological processes. Our results demonstrate the significance of multiple data type and knowledge guided integration of diverse biological information to understand the molecular mechanism associated with ovarian cancer and its application in the discovery of biomarker.

Chapter 6: Conclusions and future directions

6.1 Summary

This thesis is divided into six chapters. Chapter 1 provides a detailed literature survey on network concepts in biology. A brief introduction to experimental methods for determining protein-protein interactions is followed by some of the major public databases archiving such interactions along with other important biological resources. We also provided the basic mathematical framework for characterizing the network properties of biomolecular connectivity, illustrating different types of biological networks and their topological properties in the understanding of fundamental cellular processes and in human diseased conditions. Chapter 2 lists the publications included in this thesis and the respective chapters they are included in as a table for cross reference purpose.

Chapter 3 provides the detailed network analysis of subcellular localisation (SCL) or cellular compartmentalisation of human proteins using protein-protein interaction and metabolite-linked protein interaction networks. We compared and contrasted the above networks using rigorous statistical methods to understand the human protein localisation in various subcellular compartments using large-scale protein SCL information from LOCATE and GOA databases. This is followed by chapter 4, highlighting the statistical analysis of human and yeast high confidence interactome datasets to examine the tendency of proximal interacting protein pairs to have same molecular functions or biological processes. We carried out this analysis by measuring the shortest paths between interacting protein pairs at various GOA hierarchy levels (from one to six) and at network distances upto three. Rather than use simplistic definitions of hub proteins as date hubs and party hubs, we have investigated the behaviour of proteins over a larger number of degree categories. The dissection of the interactome in terms of interacting protein pairs characterized by their cellular components, molecular functions and biological functions is extended to the proteins implicated in diseased states from gene expression analysis. In chapter 5, an integrative approach is used for the identification of differentially expressed genes in ovarian tumour samples using seven functional attributes in a Boolean logical framework, and further validated using the human interactome.

Chapter 6 highlights the innovation, significance and contributions of this thesis, drawing conclusions from the interactome analysis for the understanding of proteins cellular components, molecular functions and biological processes and the usage of interaction network in addition to other six functional attributes in identification of genes in ovarian tumour samples. This chapter also discusses future directions. The work presented in this thesis has been published as book chapters and journal articles highlighting the importance of large scale interactome analysis.

6.2 Conclusions

This thesis highlights the statistical studies carried out on interactome datasets for the better understanding of eukaryotic proteins with respect to their subcellular localisation, molecular functions and biological processes. An integrative approach is used to combine interactome information along with other specific functional attributes in the identification of ovarian cancer genes that are differentially expressed in tumour samples. Several novel aspects are presented in the thesis. The following inferences can be drawn:

1. A detail network study is done using rigorous statistical methods to show the underlying differences in network properties of physically interacting protein pairs with respect to metabolite-linked protein pairs or functional association. We have shown the importance of metabolite-linked protein pairs in understanding the localisation of human proteins in subcellular compartments such as mitochondrion, lysosome and Golgi apparatus. Chapter 3 describes the statistical measures and methods for comparing and contrasting the localisation of human protein in the above two networks.
2. We studied the tendency of proximal proteins in the human and yeast interactome datasets in detail, upto the GOA abstraction level six. We conclude in chapter 4 that the tendency of proximal interaction protein pairs to have same molecular functions or biological processes hold true upto the network distance of two and three, respectively, with little difference in the tendency of interaction protein pairs measured *via* percentage common neighbours at the network distance of three, after which there is a marked decrease in function/process conservation.
3. Chapter 5 highlights the importance of high interacting proteins (interactions >4) i.e. hubs of interactome, selected by imposing additional functional attributes in the

identification of cancer associated genes in ovarian tumour samples using a Boolean logic framework.

6.3 Innovations

This thesis highlights the original finding and application of protein interaction networks to the study of protein subcellular localisation, molecular function and biological process. In addition to this fundamental understanding of eukaryotic proteins, we have applied the interactome in conjunction with other functional attributes of proteins for understanding the human disease condition, using ovarian cancer as a disease model.

This is, to the best of my knowledge, the first study of its kind, where an interactome dataset has been used for the detailed analysis of eukaryotic proteins with respect to their cellular components, molecular functions and biological processes. Rigorous statistical methods have been used in the studies of above mentioned proteins characteristics. We have also provided a meaningful way of integrating the interactome with other functional attributes in the human diseased condition.

6.4 Significance and contributions

This work reiterates the inherent importance of statistical methods for an integrated systems approach in the understanding of fundamental of cellular processes encoded in the interconnectivity of protein-protein interaction networks. The significant findings and contributions of this thesis are listed below.

1. This thesis presents the importance of the metabolite-linked protein interaction network i.e. functional association in the understanding of protein localisation, using human interactome data (chapter 3).
2. We have shown the importance of network distances in interaction network to attributes the same functional or process association of interacting protein pairs (chapter 4).

3. We have also introduced a novel approach to create a high confidence protein interaction set by introducing the filters to generate gold standard positive set for analysis (chapter 4).
4. The results presented here offer a compelling insight into protein's functional attributes in the characterisation of human cancer gene association (chapter 5).
5. We have outlined the rationale behind the data integration to combine diverse biological information in a statistical framework, for the characterisation of genes as potential biomarkers (chapter 5).

6.5 Future directions

The studies presented in this thesis could lead to an advancement in many directions for the better understanding of fundamental cellular processes and human diseased condition by integrating diverse biological information under robust statistical framework. The network comparison suggested in chapter 3 can be used to automate high-throughput identification of SCL in model organisms. This fully automated identification of SCL using PPI and metabolic/functional association networks can then be implemented as a research tool or a web application that provides services to the scientific community.

The statistical analysis shown in Chapter 4 on GOA molecular function and biological process paves way for refining statistical models in the prediction of protein's functional association. Future developments will include combining all the three GOA components i.e. cellular component, molecular function and biological process simultaneously for the prediction of a given protein from its known neighbours by incorporating the spatial constraints (subcellular compartmentalisation), functional diversity and process complexity, respectively. Also, I have not attempted to discriminate between the hubs with simultaneous binding of several proteins (forming a large multi-protein complex) and the other hub proteins with the ability to bind to a number of proteins but one at a time. This could be addressed by measuring the network centrality associated with each of the two kinds of hubs, to confirm the essentiality associated with such proteins in the overall cell survival. It has been shown in 2006 that multi-interface hub proteins are twice more likely to be essential on an average compare to the single-interface proteins [99]. It would be

interesting to see whether these conclusions are still valid, as more and more data on proteins in interaction networks becomes available.

The analysis done in chapter 5 has revealed the importance of functional attributes in the identification of disease causing genes. This can be extended further by incorporating gene features such as copy number variation and mutation. Moreover, there is a scope to include gene regulatory network in a Boolean framework to capture disease causing events, where proteins regulate the gene expression, followed by experimental validation.

Furthermore, this study can be extended to the domain analysis of interacting protein pairs in an interaction network and mapped with that of the co-expression similarity of interacting neighbours from the COXPRESdb [100], to convert a protein-protein interaction network to a domain-domain interaction network, to focus on the functional domain level interactions within a protein interaction network.

References

1. Alm E, Arkin AP: **Biological networks**. *Curr Opin Struct Biol* 2003, **13**(2):193-202.
2. Vogelstein B, Lane D, Levine AJ: **Surfing the p53 network**. *Nature* 2000, **408**(6810):307-310.
3. Ideker T, Sharan R: **Protein networks in disease**. *Genome Res* 2008, **18**(4):644-652.
4. Vidal M, Cusick ME, Barabasi AL: **Interactome networks and human disease**. *Cell* 2011, **144**(6):986-998.
5. Fields S, Song O: **A novel genetic system to detect protein-protein interactions**. *Nature* 1989, **340**(6230):245-246.
6. Karimova G, Pidoux J, Ullmann A, Ladant D: **A bacterial two-hybrid system based on a reconstituted signal transduction pathway**. *Proc Natl Acad Sci U S A* 1998, **95**(10):5752-5756.
7. Wehr MC, Laage R, Bolz U, Fischer TM, Grunewald S, Scheek S, Bach A, Nave KA, Rossner MJ: **Monitoring regulated protein-protein interactions using split TEV**. *Nat Methods* 2006, **3**(12):985-993.
8. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM *et al*: **Functional organization of the yeast proteome by systematic analysis of protein complexes**. *Nature* 2002, **415**(6868):141-147.
9. Vojtek AB, Hollenberg SM, Cooper JA: **Mammalian Ras interacts directly with the serine/threonine kinase Raf**. *Cell* 1993, **74**(1):205-214.
10. Golemis EA, Khazak V: **Alternative yeast two-hybrid systems. The interaction trap and interaction mating**. *Methods Mol Biol* 1997, **63**:197-218.
11. Aronheim A: **Improved efficiency sos recruitment system: expression of the mammalian GAP reduces isolation of Ras GTPase false positives**. *Nucleic Acids Res* 1997, **25**(16):3373-3374.
12. Joung JK, Ramm EI, Pabo CO: **A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions**. *Proc Natl Acad Sci U S A* 2000, **97**(13):7382-7387.
13. Eyckerman S, Verhee A, der Heyden JV, Lemmens I, Ostade XV, Vandekerckhove J, Tavernier J: **Design and application of a cytokine-receptor-based interaction trap**. *Nat Cell Biol* 2001, **3**(12):1114-1119.
14. Luo Y, Batalao A, Zhou H, Zhu L: **Mammalian two-hybrid system: a complementary approach to the yeast two-hybrid system**. *Biotechniques* 1997, **22**(2):350-352.

15. Johnsson N, Varshavsky A: **Split ubiquitin as a sensor of protein interactions in vivo.** *Proc Natl Acad Sci U S A* 1994, **91**(22):10340-10344.
16. Stagljar I, Korostensky C, Johnsson N, te Heesen S: **A genetic system based on split-ubiquitin for the analysis of interactions between membrane proteins in vivo.** *Proc Natl Acad Sci U S A* 1998, **95**(9):5187-5192.
17. Hu CD, Kerppola TK: **Simultaneous visualization of multiple protein interactions in living cells using multicolor fluorescence complementation analysis.** *Nat Biotechnol* 2003, **21**(5):539-545.
18. Marti F, Xu CW, Selvakumar A, Brent R, Dupont B, King PD: **LCK-phosphorylated human killer cell-inhibitory receptors recruit and activate phosphatidylinositol 3-kinase.** *Proc Natl Acad Sci U S A* 1998, **95**(20):11810-11815.
19. Einarson MB: **Detection of Protein-Protein Interactions Using the GST Fusion Protein Pulldown Technique.** In: *In Molecular Cloning: A Laboratory Manual*. 3rd edn; 2001: pp.55-59.
20. Einarson MB, Orlinick JR: **Identification of Protein-Protein Interactions with Glutathione S-Transferase Fusion Proteins.** In: *In Protein-Protein Interactions: A Molecular Cloning Manual*. Cold Spring Harbor Laboratory Press; 2002: pp.37-57.
21. Vikis HG, Guan K-L: **Glutathione-S-Transferase-Fusion Based Assays for Studying Protein-Protein Interactions.** In: *Protein-Protein Interactions: Methods and Applications (Methods in Molecular Biology)*. Edited by Fu H. Totowa, New Jersey: Humana Press; 2004: pp.175-186.
22. Ron D, Dressler H: **pGStag--a versatile bacterial expression plasmid for enzymatic labeling of recombinant proteins.** *Biotechniques* 1992, **13**(6):866-869.
23. Hart GT, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** *Genome Biol* 2006, **7**(11):120.
24. Prasad TS, Kandasamy K, Pandey A: **Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology.** *Methods Mol Biol* 2009, **577**:67-79.
25. Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G: **MINT, the molecular interaction database: 2009 update.** *Nucleic Acids Res* 2010, **38**(Database issue):D532-539.
26. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C *et al*: **The HUPPO PSI's molecular interaction format--a community standard for the representation of protein interaction data.** *Nat Biotechnol* 2004, **22**(2):177-183.
27. Gilbert D: **Biomolecular interaction network database.** *Brief Bioinform* 2005, **6**(2):194-198.

28. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E *et al*: **The Biomolecular Interaction Network Database and related tools 2005 update**. *Nucleic Acids Res* 2005, **33**(Database issue):D418-424.
29. Bader GD, Hogue CW: **BIND--a data specification for storing and describing biomolecular interactions, molecular complexes and pathways**. *Bioinformatics* 2000, **16**(5):465-477.
30. Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V: **MPact: the MIPS protein interaction resource on yeast**. *Nucleic Acids Res* 2006, **34**(Database issue):D436-441.
31. Mewes HW, Frishman D, Mayer KF, Munsterkotter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stumpflen V: **MIPS: analysis and annotation of proteins from whole genomes in 2005**. *Nucleic Acids Res* 2006, **34**(Database issue):D169-172.
32. Guldener U, Munsterkotter M, Kastenmuller G, Strack N, van Helden J, Lemer C, Richelles J, Wodak SJ, Garcia-Martinez J, Perez-Ortin JE *et al*: **CYGD: the Comprehensive Yeast Genome Database**. *Nucleic Acids Res* 2005, **33**(Database issue):D364-368.
33. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update**. *Nucleic Acids Res* 2004, **32**(Database issue):D449-451.
34. Duan XJ, Xenarios I, Eisenberg D: **Describing biological protein interactions in terms of protein states and state transitions: the LiveDIP database**. *Mol Cell Proteomics* 2002, **1**(2):104-116.
35. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D: **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions**. *Nucleic Acids Res* 2002, **30**(1):303-305.
36. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D: **Prolinks: a database of protein functional linkages derived from coevolution**. *Genome Biol* 2004, **5**(5):R35.
37. Orchard S, Kerrien S, Jones P, Ceol A, Chatr-Aryamontri A, Salwinski L, Neroth J, Hermjakob H: **Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition**. *Proteomics* 2007, **7 Suppl 1**:28-34.
38. Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stumpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P *et al*: **The minimum information required for reporting a molecular interaction experiment (MIMIx)**. *Nat Biotechnol* 2007, **25**(8):894-898.
39. de Matos P, Alcantara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C: **Chemical Entities of Biological Interest: an update**. *Nucleic Acids Res* 2010, **38**(Database issue):D249-254.

40. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L *et al*: **Ensembl 2009**. *Nucleic Acids Res* 2009, **37**(Database issue):D690-697.
41. Sugawara H, Ikeo K, Fukuchi S, Gojobori T, Tateno Y: **DDBJ dealing with mass data produced by the second generation sequencer**. *Nucleic Acids Res* 2009, **37**(Database issue):D16-18.
42. Cochrane G, Akhtar R, Bonfield J, Bower L, Demiralp F, Faruque N, Gibson R, Hoad G, Hubbard T, Hunter C *et al*: **Petabyte-scale innovations at the European Nucleotide Archive**. *Nucleic Acids Res* 2009, **37**(Database issue):D19-25.
43. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Res* 2009, **37**(Database issue):D26-31.
44. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L *et al*: **InterPro: the integrative protein signature database**. *Nucleic Acids Res* 2009, **37**(Database issue):D211-215.
45. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X *et al*: **The BioGRID Interaction Database: 2011 update**. *Nucleic Acids Res* 2011, **39**(Database issue):D698-704.
46. Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, Bader GD, Xenarios I, Wojcik J, Sherman D *et al*: **Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions**. *BMC Biol* 2007, **5**:44.
47. Magrane M, Consortium U: **UniProt Knowledgebase: a hub of integrated protein data**. *Database (Oxford)* 2011, **2011**:bar009.
48. Pruitt KD, Tatusova T, Klimke W, Maglott DR: **NCBI Reference Sequences: current status, policy and new initiatives**. *Nucleic Acids Res* 2009, **37**(Database issue):D32-36.
49. Engel SR, Balakrishnan R, Binkley G, Christie KR, Costanzo MC, Dwight SS, Fisk DG, Hirschman JE, Hitz BC, Hong EL *et al*: **Saccharomyces Genome Database provides mutant phenotype data**. *Nucleic Acids Res* 2010, **38**(Database issue):D433-436.
50. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, De La Cruz N, Davis P, Duesbury M, Fang R *et al*: **WormBase: a comprehensive resource for nematode research**. *Nucleic Acids Res* 2010, **38**(Database issue):D463-467.
51. Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R *et al*: **FlyBase: enhancing Drosophila Gene Ontology annotations**. *Nucleic Acids Res* 2009, **37**(Database issue):D555-559.
52. Blake JA, Bult CJ, Kadin JA, Richardson JE, Eppig JT: **The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics**. *Nucleic Acids Res* 2011, **39**(Database issue):D842-848.

53. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L *et al*: **The Arabidopsis Information Resource (TAIR): gene structure and function annotation.** *Nucleic Acids Res* 2008, **36**(Database issue):D1009-1014.
54. Breitkreutz BJ, Stark C, Tyers M: **Osprey: a network visualization system.** *Genome Biol* 2003, **4**(3):R22.
55. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B *et al*: **Integration of biological networks and gene expression data using Cytoscape.** *Nat Protoc* 2007, **2**(10):2366-2382.
56. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT *et al*: **The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.** *Nucleic Acids Res* 2010, **38**(Web Server issue):W214-220.
57. Liu G, Zhang J, Larsen B, Stark C, Breitkreutz A, Lin ZY, Breitkreutz BJ, Ding Y, Colwill K, Pasculescu A *et al*: **ProHits: integrated software for mass spectrometry-based interaction proteomics.** *Nat Biotechnol* 2010, **28**(10):1015-1017.
58. Banerjee D, Mayer-Kuckuk P, Capiaux G, Budak-Alpdogan T, Gorlick R, Bertino JR: **Novel aspects of resistance to drugs targeted to dihydrofolate reductase and thymidylate synthase.** *Biochim Biophys Acta* 2002, **1587**(2-3):164-173.
59. Stephanopoulos G, Sinskey AJ: **Metabolic engineering--methodologies and future prospects.** *Trends Biotechnol* 1993, **11**(9):392-396.
60. Selkov E, Jr., Grechkin Y, Mikhailova N, Selkov E: **MPW: the Metabolic Pathways Database.** *Nucleic Acids Res* 1998, **26**(1):43-45.
61. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Res* 2010, **38**(Database issue):D355-360.
62. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M *et al*: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.** *Nucleic Acids Res* 2010, **38**(Database issue):D473-479.
63. Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, Muniz-Rascado L, Bonavides-Martinez C, Paley S, Krummenacker M, Altman T *et al*: **EcoCyc: a comprehensive database of Escherichia coli biology.** *Nucleic Acids Res* 2011, **39**(Database issue):D583-590.
64. Mueller LA, Zhang P, Rhee SY: **AraCyc: a biochemical pathway database for Arabidopsis.** *Plant Physiol* 2003, **132**(2):453-460.
65. Dale JM, Popescu L, Karp PD: **Machine learning methods for metabolic pathway prediction.** *BMC Bioinformatics* 2010, **11**:15.

66. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD: **Computational prediction of human metabolic pathways from the complete human genome.** *Genome Biol* 2005, **6**(1):R2.
67. Romero PR, Karp PD: **Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases.** *Bioinformatics* 2004, **20**(5):709-717.
68. Green ML, Karp PD: **A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases.** *BMC Bioinformatics* 2004, **5**:76.
69. Lee TJ, Paulsen I, Karp P: **Annotation-based inference of transporter function.** *Bioinformatics* 2008, **24**(13):i259-267.
70. Schomburg I, Chang A, Schomburg D: **BRENDA, enzyme data and metabolic information.** *Nucleic Acids Res* 2002, **30**(1):47-49.
71. Ellis LB, Hou BK, Kang W, Wackett LP: **The University of Minnesota Biocatalysis/Biodegradation Database: post-genomic data mining.** *Nucleic Acids Res* 2003, **31**(1):262-265.
72. Schwikowski B, Uetz P, Fields S: **A network of protein-protein interactions in yeast.** *Nat Biotechnol* 2000, **18**(12):1257-1261.
73. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B *et al*: **Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets.** *Nat Genet* 2006, **38**(3):285-293.
74. Sprenger J, Lynn Fink J, Karunaratne S, Hanson K, Hamilton NA, Teasdale RD: **LOCATE: a mammalian protein subcellular localization database.** *Nucleic Acids Res* 2008, **36**(Database issue):D230-233.
75. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C *et al*: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309**(5740):1559-1563.
76. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2003, **31**(1):23-27.
77. **Ongoing and future developments at the Universal Protein Resource.** *Nucleic Acids Res* 2011, **39**(Database issue):D214-219.
78. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD *et al*: **The RCSB Protein Data Bank: redesigned web site and web services.** *Nucleic Acids Res* 2011, **39**(Database issue):D392-401.
79. Bairoch A, Boeckmann B, Ferro S, Gasteiger E: **Swiss-Prot: juggling between evolution and stability.** *Brief Bioinform* 2004, **5**(1):39-55.

80. Boeckmann B, Blatter MC, Famiglietti L, Hinz U, Lane L, Roechert B, Bairoch A: **Protein variety and functional diversity: Swiss-Prot annotation in its biological context.** *C R Biol* 2005, **328**(10-11):882-899.
81. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999.** *Nucleic Acids Res* 1999, **27**(1):49-54.
82. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E: **Infrastructure for the life sciences: design and implementation of the UniProt website.** *BMC Bioinformatics* 2009, **10**:136.
83. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH: **UniRef: comprehensive and non-redundant UniProt reference clusters.** *Bioinformatics* 2007, **23**(10):1282-1288.
84. Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R: **UniProt archive.** *Bioinformatics* 2004, **20**(17):3236-3237.
85. Davis TN: **Protein localization in proteomics.** *Curr Opin Chem Biol* 2004, **8**(1):49-53.
86. Sprenger J, Fink JL, Teasdale RD: **Evaluation and comparison of mammalian subcellular localization prediction methods.** *BMC Bioinformatics* 2006, **7 Suppl 5**:S3.
87. Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom-Cerajewski L, Robinson MD, O'Connor L, Li M *et al*: **Large-scale mapping of human protein-protein interactions by mass spectrometry.** *Mol Syst Biol* 2007, **3**:89.
88. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N *et al*: **Towards a proteome-scale map of the human protein-protein interaction network.** *Nature* 2005, **437**(7062):1173-1178.
89. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S *et al*: **A human protein-protein interaction network: a resource for annotating the proteome.** *Cell* 2005, **122**(6):957-968.
90. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E *et al*: **A protein interaction map of Drosophila melanogaster.** *Science* 2003, **302**(5651):1727-1736.
91. Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, Piccirillo S, Umansky L, Drawid A, Jansen R, Liu Y *et al*: **Subcellular localization of the yeast proteome.** *Genes Dev* 2002, **16**(6):707-719.
92. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic Acids Res* 2004, **32**(Database issue):D277-280.

93. Paine PL, Moore LC, Horowitz SB: **Nuclear envelope permeability.** *Nature* 1975, **254**(5496):109-114.
94. Lee TY, Bo-Kai Hsu J, Chang WC, Huang HD: **RegPhos: a system to explore the protein kinase-substrate phosphorylation network in humans.** *Nucleic Acids Res* 2011, **39**(Database issue):D777-787.
95. Krishnadev O, Srinivasan N: **A data integration approach to predict host-pathogen protein-protein interactions: application to recognize protein interactions between human and a malarial parasite.** *In Silico Biol* 2008, **8**(3-4):235-250.
96. Boden M, Dellaire G, Burrage K, Bailey TL: **A Bayesian network model of proteins' association with promyelocytic leukemia (PML) nuclear bodies.** *J Comput Biol* 2010, **17**(4):617-630.
97. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP *et al*: **Evidence for dynamically organized modularity in the yeast protein-protein interaction network.** *Nature* 2004, **430**(6995):88-93.
98. Batada NN, Hurst LD, Tyers M: **Evolutionary and physiological importance of hub proteins.** *PLoS Comput Biol* 2006, **2**(7):e88.
99. Kim PM, Lu LJ, Xia Y, Gerstein MB: **Relating three-dimensional structures to protein networks provides evolutionary insights.** *Science* 2006, **314**(5807):1938-1941.
100. Obayashi T, Kinoshita K: **COXPRESdb: a database to compare gene coexpression in seven model animals.** *Nucleic Acids Res* 2011, **39**(Database issue):D1016-D1022.