3

Protein Interactome and Its Application to Protein Function Prediction

Woojin Jung[1], Hyun-Hwan Jeong[1,2], and KiYoung Lee[1]
[1]Department of Biomedical Informatics, Ajou University School of Medicine
[2]Department of Computer Engineering, Ajou University
Republic of Korea

1. Introduction

Diverse molecules interact with proteins to produce a biological function. Proteins exhibit many interactions with other molecules including other proteins, nucleic acids, carbohydrates, lipids, minerals, metabolites, and chemical compounds, resulting in diverse roles within and/or between cells. Some of these proteins locate in subcellular organelles, where they modulate biochemical reactions, and some other proteins locate in membranes mediating various stimuli to signaling pathways. Cellular systems can be represented as complex networks. We may consider the molecules as nodes and the associations among the molecules as edges in the network. In this network, all kinds of the molecular interactions can be referred to as an interactome. Even though all kinds of the interactome are important, we here focus on protein-protein interactions (PPIs) since they are fundamental in cellular systems. To function correctly, a protein should interact with other proteins in the context of complex formation, signalling pathways and biochemical reactions. To perform a specific biological function, these interactions need to be specifically formed with proper interacting partners at the right time and locations.

Given the knowledge of genome sequencing on model organisms including human, we have elucidated a large number of unknown molecular structures and interactions within nucleic acids. In the post-genomic era, functional genomics is an emerging area of research that seeks to annotate every bit of information of the genome structure with relevant biological function. Still, many proteins (or genes) remain functionally unannotated (Apweiler et al, 2004; Sharan et al, 2007). These missing links between structures and functions need to be resolved to understand complex biological phenomena including human diseases, development and aging.

Protein function is widely defined in several different ways. It is highly context- and condition-dependent, which means that proteins participate in most biological processes. There have been various attempts to categorize the protein functions (Bork et al, 1998). One of them categorized the protein function into three parts: molecular function, cellular function and phenotypic function. First, the molecular function is defined as biochemical reactions performed by proteins. Second, the cellular function is defined as various pathways associated with proteins. Lastly, the phenotypic function is defined as an integration of all physiological subsystems to environmental stimuli.

Aside from the conceptual definition, many annotation efforts on protein function have been undertaken (Table 1). One of these efforts, the Gene Ontology (GO) consortium (Ashburner et al, 2000), made a standard and multi-labelled hierarchical annotation on proteins in the category of biological process, molecular function and cellular component. The GO consortium is regularly accumulating annotations on proteins according to GO category in open databases. In this chapter, we consider the three kinds of GO terms in annotation of protein function.

Many experimental techniques are available for discovering the protein function, such as gene knockout and transcript knockdown, but these approaches are low-throughput and time-consuming. In recent decades novel high-throughput techniques have been developed, and we are now able to analysis genome-wide data, which is broadening our biological insights. Computational methods are necessary for analysing the massive quantity of data and they are complementary with the low- and high-throughput experimental methods.

In this chapter, we first introduce PPI data available through public databases and compare the contents of major databases. We also describe PPI detection methods by experimental and computational approaches. Next, network- and non-network-based computational methods for the identification of protein function are described. Finally, computational prediction methods of protein subcellular localization, especially by exploiting PPI data, are shown.

Databases	Description
GO	The Gene Ontology project/consortium
COGs	Clusters of Orthologous Groups of proteins
ENZYME	A repository of information relative to the nomenclature of enzymes
Pfam	A database of protein families that includes their annotations and multiple sequence alignments
PROSITE	Database of protein domains, families and functional sites
HAMAP	High-quality Automated and Manual Annotation of microbial Proteomes
UniProt	The Universal Protein Resource
FunCat	MIPS (Munich Information Center for Protein Sequences) Functional Catalogue
DAVID	The Database for Annotation, Visualization and Integrated Discovery
FANTOM	A database for functional annotation of the mammalian genome
ANNOVAR	Functional annotation of genetic variants from high-throughput sequencing data
EFICAz	A genome-wide enzyme function annotation database
KEGG	Kyoto Encyclopedia of Genes and Genomes

Table 1. Databases for the functional annotation of genes and proteins.

2. PPI data

PPI can be considered as one kind of protein interactome. Proteins mutually interact in the biological context for specific functions. Given the knowledge of a single gene, expressing distinct transcripts and protein isoforms, a protein also interacts with other proteins including itself to give specific function. PPIs are defined as physical interactions between protein pairs (Bonetta, 2010). There are also non-physical interactions such as genetic and functional interactions. Genetic interaction is typically defined as when two genes are simultaneously perturbed, with the quantitative phenotype being more or less than expected (Mani et al, 2008). Functional interaction between two proteins is a much broader concept than other experiment-derived interactions. It may include any functionally associated gene/protein pairs which are integrated and predicted from heterogeneous data. We will explain these computational prediction methods later in this section.

The physical interactions between protein pairs also can be either direct or indirect. Binary interaction is an example of a direct interaction while indirect interaction includes subunits of protein complex. To give a specific function, proteins often form a large complex including direct and indirect interaction among the participant proteins. These interactions are also separable according to their binding lifetime. Some interactions between protein pairs are transient, with the interactions associating and dissociating under particular physiological conditions. On the other hand, some of proteins form stable complexes where the participants in the complexes permanently interact with each other. Various PPI types are defined in standard and annotated across many PPI databases (Cote et al, 2010; Kerrien et al, 2007).

2.1 PPI databases

Currently, there are 132 PPI databases indexed by the Pathguide (Bader et al, 2006; accessed 23 Dec 2011). The quantity of physical interactions to date is 386,495 across all species when integrated among major 11 databases by the iRefWeb (Turner et al, 2010; accessed 23 Dec 2011). The PPI data derived from both high- and low-throughput experiments are altogether deposited into any of primary databases which manually curate experimental results. These primary databases include not only physical interactions but also genetic interactions and annotate standard minimal information about a molecular interaction (MIMIx) (Orchard et al, 2007). There is an inconsistency problem related to the literature curation across different databases (Turinsky et al, 2010). Turinsky et al. confirmed that the agreement between curated interactions from 15,471 papers shared across nine databases was only 42% for interactions and 62% for proteins. This result was averaged between any two databases curated from the same publication. Some of the primary databases altogether formed a consortium called IMEx (The International Molecular Exchange) to enhance the quality of literature curation efforts.

Since we have plenty of primary databases, comprehensive integration of those primary databases has become an intriguing research field. Such meta-databases minimize redundancy and inconsistency that are limitations of the primary databases (Turinsky et al, 2010). Moreover, functional interaction databases consist of both experimentally-detected and computationally-predicted data. Sometimes, these predicted and experimental PPIs need to be distinguished for the degree of confidence. They both give useful information but should be separated according to the relevant evidence codes. There are also species-specific functional interaction databases (Lee et al, 2011; Lee et al, 2010a).

Type	Name	Description	URL
Primary databases	BioGRID	Physical and genetic interaction	http://thebiogrid.org
	MINT	Physical interaction	http://mint.bio.uniroma2.it
	IntAct	Physical interaction	http://www.ebi.ac.uk/intact
	DIP	Physical interaction	http://dip.doe-mbi.ucla.edu
	BIND	Physical and genetic interaction	http://bond.unleashedinformatics.com
	Phospho-POINT	A human kinase interactome resource	http://kinase.bioinformatics.tw
	PIG	Host-Pathogen interactome	http://pig.vbi.vt.edu
	SPIKE	A database of highly curated human signaling pathways	http://www.cs.tau.ac.il/~spike
	MPPI	The MIPS mammalian PPI database	http://mips.helmholtz-muenchen.de/proj/ppi
	HPRD	Human physical interaction	http://www.hprd.org
	CORUM	Mammalian protein complexes	http://mips.helmholtz-muenchen.de /proj/corum
Meta-databases	APID	Agile Protein Interaction DataAnalyzer	http://bioinfow.dep.usal.es/apid
	MiMi	Michigan Molecular Interactions	http://mimi.ncibi.org
	UniHI	Unified Human Interactome	http://www.mdc-berlin.de/unihi
	iRefWeb	Interaction Reference Index	http://wodaklab.org/iRefWeb
	DASMI	Distributed Annotation System for Molecular Interactions	http://dasmi.de/dasmiweb.php
	HIPPIE	Human Integrated Protein-Protein Interaction rEference	http://cbdm.mdc-berlin.de/tools/hippie
	HAPPI	Human Annotated and Predicted Protein Interaction database	http://bio.informatics.iupui.edu/HAPPI
Functional databases	STRING	Search Tool for the Retrieval of Interacting Genes/Proteins	http://string-db.org
	Gene-MANIA	Multiple Association Network Integration Algorithm	http://genemania.org
	Functional-Net	Species-specific functional gene networks	http://www.functionalnet.org

Table 2. List of PPI databases.

Contents	BIND	BioGRID	DIP	HPRD	IntAct	MINT	MPPI
Biological role (PSI-MI)					O	O	
Experimental role (PSI-MI)					O	O	
Taxonomy ID					O	O	
Interaction category					O	O	
Interaction title							
Interaction type (Text)		O					
Interaction type (PSI-MI)			O		O	O	
Interactor type (peptide, protein)					O	O	
Detection method (Text)				O			O
Detection method (PSI-MI)		O	O		O	O	
Evidence (PMID or doi-number)						O	
PubMed ID	O	O	O	O	O		O
BioGRID ID		O					
HPRD ID				O			
NCBI Gene ID	O	O		O			
Protein ID					O	O	O
ID type					O	O	O
Protein accession number	O			O			
UniProt ID			O				
Link to source ID	O	O	O			O	
Description	O				O		
Confidence score		O				O	

Table 3. Contents of primary PPI databases. Available contents are colored in grey with "O" shape.

We have listed some of the major primary databases, meta-databases, and functional databases in Table 2. Comparisons among the primary databases are shown in Table 3. We compared various features including interaction types, detection methods, references, and biological and experimental roles. This information would be valuable for researchers when they need to select and integrate various PPI data bases.

2.2 Methods for PPI identification

There are two major ways to determine PPIs. One is an experimental detection and the other is computational prediction. The former method is more reliable and well-established in both small and large scales while the latter method is based on the characteristics of accumulated protein interactions. In this section, we will briefly describe both approaches.

2.2.1 Experimental detection methods for PPIs

Experimental detection of interactions between protein pairs is achieved by various methods. Here, we describe only two representative methods: yeast two-hybrid (Y2H) (Suter et al, 2008) and mass spectrometry (MS) (Berggard et al, 2007). These methods both detect physical PPIs but the type of PPIs is different. As previously stated, direct and binary PPIs are distinct from protein groups in a complex and this type of PPI is detected only by Y2H method. This method uses a transcription factor found in yeast which consists of two other domains. Y2H method relies on an artificial insertion of a protein coding sequence to one of the domains and another protein inserted on the other domain using a plasmid. PPI can be assessed by confirming phenotype of the target gene of the transcription factor. The Y2H method can detect PPIs in large-scale and the sensitivity is high, enabling detection of even weak transient PPIs. But, since the experiment is done only in the nucleus, the real location information of such PPIs is hard to annotate, which obscures the detailed biological interpretation. Moreover, Y2H detects only binary interactions and results in a high rate of false positive, which are noteworthy limitations.

Another method in this category is based on mass-spectrometry (MS). The MS analyzes the mass of molecules rapidly and accurately. If the weight of all proteins in question is known, this information can be linked to the specific protein. This method is powerful when protein co-complexes are examined. Although it cannot provide details on the direct-level of interactions, the grouping of the proteins in a complex can be revealed. For this method, one protein (“bait”) and all of interacting partners in a complex are pulled out and separated by electrophoresis. Finally, all the constructs derived from electrophoresis are used for MS. This method yields many false positive results when the sampling strategy is thoroughly different. This sampling might include fake interactions resulting in a high rate of false readings. There are many strategies related to this problem (Bousquet-Dubouch et al, 2011; Gingras et al, 2007). The experimental results obtained with MS-based methods are different from those obtained with binary methods (Y2H). Data derived from co-complex experiments cannot directly assign a binary interpretation. An algorithm is needed to translate group-based observations into pairwise interactions.

2.2.2 Computational prediction methods for PPIs

While recent reviews (Lees et al, 2011; Pitre et al, 2008; Shoemaker & Panchenko, 2007; Skrabanek et al, 2008; Xia et al, 2010) have discussed computational prediction methods for PPIs in details, we here briefly introduce some of approaches that are widely used. Although the amounts of experimental resources of PPIs are growing rapidly, proteome-wide PPIs information is still lacking and mostly limited on several model organisms. Given wide types of indirect but genome-wide resources, we can enhance our understanding of overall protein interactome. Methods in prediction of direct physical PPIs are less investigated than those of functional association between protein pairs. These functional association methods of PPIs can give information of which protein pairs have same biological process and potential physical interactions.

The first data used in these prediction methods is genomic sequences. Co-occurrence-based methods use assumption that if gene pairs are co-inherited across evolutionary processes (i.e. species), they are considered as functionally associated (Barker & Pagel, 2005; Bowers et al, 2004; Pellegrini et al, 1999). These methods applied to microorganisms and successfully discovered novel participants of known pathway (Carlson et al, 2004; Luttgen et al, 2000). Other similar methods based on this genomic sequence use the information of gene fusion events (Marcotte et al, 1999; Reid et al, 2010; Zhang et al, 2006) and gene neighbourhood (Ferrer et al, 2010; Itoh et al, 1999; Koonin et al, 2001). Another type of data used is amino acid (AA) sequences and the interface of interacting protein pairs are composed of specific AA residues (Tuncbag et al, 2008; Tuncbag et al, 2009). This knowledge is reflected in the co-evolution of specific interface residues between interacting proteins and by alignments of multiple sequences, the results are highly correlated with physical PPIs (Pazos et al, 2005). Commonly occurring domain pairs are also considered in this context (Eddy, 2009; Finn et al, 2010; Stein et al, 2009; Yeats et al, 2011) and simple AA sequence such as 3-mers of interacting residues can be used (Ben-Hur & Noble, 2005). Another well-known information is homology of PPIs across different species. Methods on this information simply find PPIs which are conserved across species, called interologs (Matthews et al, 2001). Here, any known PPIs regarded as query to find conserved interactions across species using an ortholog database. There are many algorithms which follow this approach (Kemmer et al, 2005; Persico et al, 2005). Aside from the sequence-level data, structural information is also a valuable resource to predict PPIs, especially a protein 3D structure. (Aloy & Russell, 2003; Ezkurdia et al, 2009; Hosur et al, 2011; Shoemaker et al, 2010; Singh et al, 2010; Zhang et al, 2010). A huge amount of genome-wide gene expression profiles are another useful data to predict PPIs and they are investigated to define gene co-expression patterns of any pairs and consider higher correlation degree as higher probability of PPIs (Grigoriev, 2001; Lukk et al, 2010; Stuart et al, 2003). As shown in the earlier section, there are many literature-curated PPI databases. While those approaches are based on the manual inspection, such PPIs information can be automatically extracted using a text-mining algorithm (Blaschke et al, 2001; Szklarczyk et al, 2011; Tikk et al, 2010).

3. Computational prediction methods for protein function

Even before the prevalence of genome sequencing technologies, typical experimental identification on a protein function has been executed. Such identification has focused on a specific target gene or protein, or a small set of protein complexes. Gene knockout, knockdown of gene expression, and targeted mutations are some methods for protein function identification (Recillas-Targa, 2006; Skarnes et al, 2011). Such low-throughput experiments were replaced by high-throughput experiments including genome sequencing and determination of the protein interactome. Computational methods followed by massively archived data have been developed for better analysis. Based on the assumption that structural similarity correlates with functional similarity, homology-based functional annotation across organisms has now become a trivial approach (Aloy et al, 2001; Gaudet et al, 2011).

3.1 Non-network based approaches

Classical computational methods use features from only a single protein in prediction of protein function (Bork et al, 1998). These approaches use a set of features like amino acid sequences, genome sequences, protein structures (2D and 3D), phylogenetic data, and gene expression data. PSI-BLAST (Altschul et al, 1997) and FASTA (Mount, 2007) are popular sequence alignment tools used to reveal homologous proteins between known and unknown (query) proteins. Proteins with similar sequences are assumed to have similar functions. Moreover, protein folding patterns are also preserved enough to identify homologs (Huynen et al, 1998; Sanchez-Chapado et al, 1997). The comparative genomics across different species is a powerful approach for analysing functional annotation of proteins. In fact, it has been suggested that correlation of sequence-structure is much stronger than that of sequence-function (Smith et al, 2000; Whisstock & Lesk, 2003). So many approaches take the sequence to structure to function route for protein function prediction (Fetrow & Skolnick, 1998).

Likewise, these data are showing only single aspect of functional features conserved during evolution. Data derived from different sources can be inter-connected it should be integrated to analyse simultaneously (Kemmeren & Holstege, 2003). We next show that PPI networks potentially enrich functional relationship between protein pairs that may not be detectable from other genomic data such as primary or higher level sequence structure.

3.2 Network-based approaches

As we mentioned in the Introduction, biological function is never achieved by a single protein. Rather, proteins dynamically interact with each other and the interacting partners adopt similar performances for specific functions. With a plethora of data being generated by high-throughput proteomic experiments, it has become possible to use proteome-wide PPI patterns in protein function prediction. Among a broad type of protein interactome, a PPI network generates well-known data that is invaluable in prediction of protein function. It is possible to annotate the function of undefined proteins according to its neighbours that are functionally annotated. This assumption is based on simple idea called “guilt-by-association”, and we consider an association by possible physical interaction in any condition and, sometimes, functional association are given with relevant evidence score.

Here, we review the general network-based approaches in predicting protein functions. These approaches are categorized into two methods for better description. The first one is a straightforward method of inferring protein function based on the topological structure of a PPI network. The other method first identifies distinct sub-networks from a whole PPI network. These sub-networks are also referred to as functional modules since they perform specific biological functions such as protein complexes, and metabolic and signalling pathways. Functional modules are detected by a broad variety of clustering algorithms and, thereafter, each module is annotated with appropriate functional association. In this section, basic concepts and pioneering studies on this corresponding approaches are introduced.

3.2.1 Direct annotation of protein function using PPI network

3.2.1.1 Neighbourhood approaches

Direct functional annotation considers the correlation of the network distance between two proteins, which means the closer the two proteins are in the network the more similar are their functions. One of the earliest studies extrapolated only adjacent neighbours within an entire PPI network. This simple approach used information of the immediate neighbourhood and took the most common functions up to three among its neighbours. In spite of the effectiveness, accuracy was achieved by 72% (Schwikowski et al, 2000). However, this method lacked significance values for each association and the full network topology was not considered in the annotation process. A strategy was proposed to tackle the first problem of assigning statistical significance (Hishigaki et al, 2001). This was done by using χ^2 .-like scores and, instead of using the immediate neighbours, the *n*-neighbourhood of a protein that consists of proteins with distance of *k*-links to the protein is considered. Simply put, the neighbours of adjacent neighbours are taken into account with the frequencies of all the distance of in this neighbourhood. For an unknown protein, the functional enrichment in its *n*-neighbourhood in identified with χ^2 test, and the top ranking functions are assigned to the unknown protein. In another approach, the shared neighbourhood of a pair of proteins are considered besides from the neighbourhood of the protein of interest. Chua et al. investigated the correlation between functional similarity and network distance (Chua et al, 2006). They developed a functional similarity score, called the FS-weight measure, which gives different weights to proteins depending on their network distance from the query protein. This approach showed higher accuracy when employing indirect interactions and its functional association.

3.2.1.2 Global optimization approaches

Although the neighbourhood approach is very attractive and effective by its simplicity, shortcomings arise when there is not enough number of protein neighbours and sufficiently annotated proteins. To overcome this issue, several approaches that utilize the entire topology of the network have been proposed. These global approaches attempt to optimize annotation of function-unknown protein using the topology of a whole network. One of the first studies that took this approach used the theory of Markov random fields, which determines the probability of a protein having a certain function (Deng et al, 2004). This theory is then used to determine the joint probability of the whole interaction network regarding to a certain function. This formulation is transformed to that of the conditional probability of a protein having a certain function given the annotations of its interaction partners. After that, the Gibbs sampling technique is iteratively applied to determine the stable values of this probability for each protein. This approach resulted in higher performance than those of neighbourhood-based approaches (Chua et al, 2006; Hishigaki et al, 2001; Schwikowski et al, 2000) when utilized to the yeast PPI data.

Additional attempts according to this approach had been followed. Here, the objective function is defined for the whole network, which is a sum of the following variables (Vazquez et al, 2003).

1. The number of neighbours of a protein having the same function as itself.
2. The number of neighbours of a protein having the function under consideration.

Thus, this function estimates the number of pairs of interacting proteins with no common functional annotation. Since a high value of this function is biologically undesirable, it is minimized using a simulated annealing procedure. As expected, this approach outperformed the majority rule-based strategy on the *Saccharomyces cerevisiae* interaction data (Schwikowski et al, 2000), since the latter tried to optimize only the second factor above. An additional advantage of this approach was that multiple annotations of all proteins were obtained in one shot, unlike earlier approaches which ran independent optimization procedures for different functions.

The above discussion shows that a wide variety of approaches based on principles of global optimization have been proposed in the literature and many more are in the pipeline. The most accurate results in the field of function prediction from PPI networks have also been achieved by these approaches, which is intuitively acceptable since they extract the maximum benefit from the knowledge of the structure of the entire network.

3.2.2 Indirect annotation of protein function

This approach uses a protein interaction network, not directly for annotation, but identifies functional modules first and then assigns functions to unknown proteins based on their membership in the functional modules. This is based on the assumption that most biological networks are organized as distinct sub-networks to give specific functions (Hartwell et al, 1999). We assume that proteins in the same module participate in a similar biological process. Modular patterns and dense regions are found in the PPI network (Gavin et al, 2006).

3.2.2.1 Distance-based clustering approaches

To find biologically significant modules, clustering algorithms can be applied efficiently. Clustering is a popular unsupervised learning algorithm that does not use any prior information about the class label. There are two widely-used ways of clustering: topology-based or distance-based. The key procedure in distance-based clustering is to select the similarity measure between two proteins to detect modules. The distance between two proteins (also called as nodes) in a network is usually defined as the number of interactions (also called as edges) on the shortest path between them. However, there is a serious problem in this hierarchical clustering, known as the 'ties in proximity' problem (Arnau et al, 2005). This means that the distance between many protein pairs are identical.

To solve this problem, a network clustering method was developed to identify modules in the biological network based on the fact that each node has a unique pattern of shortest path lengths to every other node. But for a specific module in the network, the nodes/members of the module shared similar pattern of shortest path lengths (Rives & Galitski, 2003). Another study used the hierarchical clustering method with the shortest path length between proteins as a distance measure to overcome the 'ties in proximity'. This was achieved by exploiting equally valid hierarchical clustering solution with a random select when ties are met (Arnau et al, 2005). Although many methods in the similarity measures have been proposed, a single validation for such methods is insufficient. For this, two evaluation schema are suggested, which are based on the depth of a hierarchical tree and width of the ordered adjacency matrix (Lu et al, 2004). Furthermore, there are various types of cellular network with distinct modular patterns, and so network-specific methods should be investigated in the future.

3.2.2.2 Graph-based clustering approaches

Dissecting functional modules in a large PPI network is the same problem of graph partitioning and clustering. One of the pioneering method using this network topology-based concept was the MCODE (molecular complex detection algorithm) (Bader & Hogue, 2003). This method predicts complexes in a large PPI network consisting of three processes. First, the nodes of the network are weighted by their core clustering coefficients (the density of the largest *k*-core of its adjacent neighbourhood), and then densely connected modules are identified in a greedy fashion. The use of this coefficient instead of a standard clustering coefficient was proposed, as it increases the weights of densely interconnected graph regions while giving small weights to the less connected nodes. The next step is to filter or add proteins based on the connectivity criteria. This method was applied to large-scale PPI networks and given as a plug-in for the Cytoscape (Kohl et al, 2011).

Another similar study to find complexes and functional modules is based on super paramagnetic clustering. This method used an analogy to the physical features of a heterogeneous ferromagnetic model to detect densely connected clusters in a large graph (Spirin & Mirny, 2003). There is also an algorithm called the restricted neighbourhood search clustering (RNSC), which starts with an initial random cluster assignment and then proceeds by reassigning nodes to maximize the partition's score. Here, the score represents an intra-connectivity in the cluster, not an inter-connectivity across other clusters. The RNSC algorithm is known to perform better than the MCODE algorithm (King et al, 2004). The Markov clustering algorithm (MCL) is another fast and scalable clustering algorithm based on simulation of random walks on the underlying graph (Pereira-Leal et al, 2004). This algorithm has an assumption that a random walker in natural clusters (i.e. dense region of the graph) sparsely goes from one to another natural cluster. Such clusters in a whole graph are structurally identified by the MCL algorithm. It starts by measuring the probabilities of random walks through the graph to build a stochastic "Markov" matrix, by alternating two operations: expansion and inflation. The expansion takes the squared power of the matrix while the inflation takes the Hadamard power of a matrix, followed by a re-scaling. Therefore the resulting matrix is remained as stochastic. Clusters are detected by alternation of expansion and inflation until the graph is partitioned into distinct subsets where no paths between these subsets are available. This algorithm can be efficiently implemented to weighted and large dense graphs. Various PPI networks were applied using the MCL algorithm to find functional modules such as protein complex (Krogan et al, 2006).

It is true that a protein might have multiple functions and this characteristics of a protein leads to overlap of different modules. That means graph partitioning in a strict manner might not be reasonable for the PPI network. However, most current methods are based on the hard-partition algorithms, meaning that each protein can belong to only one specific module. To handle this limitation, a clustering algorithm based on the information flow was suggested. This algorithm efficiently identified the overlapping clusters in weighted PPI network by integrating semantic similarity between GO function terms (Cho et al, 2007). Since the common proteins in the overlapping modules are interpreted as a connecting bridge across the different modules, biologically significant and functional sub-networks could be identified. Still, there are few clustering methods identifying such overlapping modules. Novel clustering methods for this theme are required with enhancement of prediction accuracy.

4. Prediction of protein subcellular localization

4.1 Introduction

Proteins should move to specific locations after synthesis to work in our body correctly. Thus, knowing subcellular localization of proteins is important to understand their own functions. Unicellular organisms like budding and fission yeasts can find systematic protein localization by experimental studies. However, such studies could not be performed well in higher eukaryotes such as *Caenorhabditis elegans*, *Drosophila melanogaster*, or mammals because of large-scale proteome sizes and technical difficulties associated with protein tagging.

Therefore, bioinformatical approaches to develop efficient methods are required instead of wet experiments. Actually, many computational methods to predict subcellular localization of protein have been proposed over several decades. A considerable number of computational classification methods have been developed for this purpose. Typically these algorithms input list of features and output subcellular localizations of target proteins. The features contain various characteristics of the proteins. Molecular weight, amino acid content and codon bias can be the features. Input features for prediction of subcellular localization can be broadly categorized into four categories: protein sorting signals, empirically correlated characteristics, sequence homology with known answer sets, and other sources (Imai & Nakai, 2010).

During the training phase, in the methods, learning utilizes a set of gold-standard proteins whose localizations are well known. This set consists of the feature vectors. After the training phase, a model is constructed to recognize those features or patterns of features that are useful and then predicts the subcellular localization of proteins whose localization is unknown. Various algorithms have been used to construct a model for prediction of sub-cellular localization.

In the field of bioinformatics, there are several problems to resolve for predicting subcellular localization of proteins. First, there are generally too many classes (localization). According to Huh et al, 22 distinct localizations exist in budding yeast. Next, one protein may have multiple different localizations (Huh et al, 2003). This is referred to a multi-label classification problem and traditional classification algorithms have a limit on handling the multi-label problem well. Another problem is that there may be a higher dimensional feature space for prediction. More than tens of thousands features exist in some cases.

Another issue is that data for each localization is too imbalanced. All these characteristics make the prediction difficult. More importantly, the localization prediction is sometimes difficult to achieve sufficient performance when we use information of single proteins only. Recently, large-scale protein-protein interaction networks have been elucidated in yeast, fly, worm, and human. To interact physically, two proteins should localize to the same or adjacent subcellular localization. That means we can get useful information of a protein from its interacting neighbours. Thus, we can improve the localization prediction performance particularly using PPI networks.

4.2 Computational prediction of protein subcellular localization

4.2.1 Single-protein feature based localization predictions

Table 4 summarizes previous studies that have used the features of single proteins. The studies for prediction of subcellular localization have the following trends. The first is an increase in the number of predicting localizations. At first, Nakashima & Nishikawa predicted localization of a protein that is inter-cellular or extra-cellular using Amino Acid (AA) and Pair coupled Amino Acid (PairAA) (Nakashima & Nishikawa, 1994). After their study, many studies tried to increase the number of distinct localizations to predict. For example, Gardy et al predicted five distinct subcellular localization including ‘cytoplasmic’, ‘inner membrane’, ‘periplasmic’, ‘outer membrane’ and ‘extra-cellular’ (Gardy et al, 2003). Nair & Rost predicted ten distinct subcellular localizations (Nair & Rost, 2003). Also, Chou & Cai predicted 22 distinct subcellular localizations that experimentally identified localization of Huh et al. (Chou & Cai, 2003).

The second trend is handling of a multi-label problem. A protein can localize to several sub-cellular locations. However, most of these studies did not consider multiple localization property, but rather assumed that a protein has a single representative localization. Also, the accuracy of prediction is lower when the number of distinct localizations for a protein is increased. Some researchers have been tried to address this issue (Lee et al, 2006).

Another tendency is the development of a classification algorithm for an elaborate and efficient model construction. Least distance algorithm, artificial neural network, a nearest neighbour approach, a Markov model, a Bayesian network approach, and support vector machine (SVM) were used to archive the goal. Some studies mixed several algorithms. Lee et al. developed an algorithm that reflects of property of the prediction task (Lee et al, 2006). They developed an extended Density-induced Support Vector Data Description (D-SVDD) classification algorithm to handle well the issues related to class imbalance, higher dimensionality, multi-label, and many distinct classes. The classical D-SVDD algorithm can handle only one-class classification tasks. Thus, Lee et al. extended it to handle multi-label classification tasks.

4.2.2 Network-based localization prediction

As mentioned earlier, two proteins that localize to same or adjacent subcellular localization have a tendency to interact with each other. That means two proteins can be a tag protein to one other for subcellular localization. Therefore, if a molecular network such as PPIs is available, we may take advantage of the PPI network for the prediction. Several studies tried to predict subcellular localization using network data. This section consists two parts: first one is a brief explanation of the study by Lee et al. (Lee et al, 2008), which is the cornerstone of the network-based approach for location prediction using PPI network. We describe a methodology to generate of feature vectors for a protein in the aforementioned study and introduce a DC-kNN classifier for the prediction. The second part is a summary of the network-based approaches from the work of Lee et al. to the present.

Author(s)	Method(s)	Feature(s)	# Classes	Multi-label	Imbalanced
(Nakai & Kanehisa, 1991)	Expert Systems	SignalMotif	4	X	X
(Nakai & Kanehisa, 1992)	Expert Systems	AA, SingalMotif	14	X	X
(Nakashima & Nishikawa, 1994)	Scoring System	AA, diAA	2	X	X
(Cedano et al, 1997)	LDA using Mahalanobis distance	AA	5	X	X
(Reinhardt & Hubbard, 1998)	ANN Approach	AA	3, 4	X	X
(Chou & Elrod, 1999)	CDA	AA	12	X	X
(Yuan, 1999)	Markov Model	AA	3, 4	X	X
(Nakai & Horton, 1999)	k-NN approach	SignalMotif	11	X	X
(Emanuelsson et al, 2000)	Neural network	SignalMotif	4	X	X
(Drawid & Gerstein, 2000)	CDA	Gene Expression Pattern	8	X	X
(Drawid & Gerstein, 2000)	Bayesian Approach	SignalMotif, HDEL motif	5, 6	X	X
(Cai et al, 2000)	SVM	AA	12	X	X
(Chou, 2000)	Augumented CDA	AA, SOC factor	5, 7, 12	X	X
(Chou, 2001)	LDA using various distance measures	pseuAA	5, 9, 12	X	X
(Hua & Sun, 2001)	SVM	AA	4	X	X
(Chou & Cai, 2002)	SVM	SBASE-FunD	12	X	X
(Nair & Rost, 2002)	Nearest Neighbor Approach	functional annotation	10	X	X
(Cai et al, 2003)	SVM	SBASE-FunD, pseuAA	5	X	X
(Cai & Chou, 2003)	Nearest Neighbor Approach	GO, InterProFunD, pseuAA	3, 4	X	X
(Chou & Cai, 2003)	LDA using various distance measures	pseuAA	14	X	X
(Pan et al, 2003)	Augumented CDA	pseuAA with filler	12	X	X
(Park & Kanehisa, 2003)	SVM	AA, diAA, gapAA	12	X	X
(Zhou & Doctor, 2003)	Covariant discrinant algorithm	AA	4	X	X
(Cai et al, 2003)	SVM	SBASE-FunD, pseuAA	5	X	X

Author(s)	Method(s)	Feature(s)	# Classes	Multi-label	Imbalanced
(Gardy et al, 2003)	SVM, HMM, Baysian	AA, motif, homlogy analysis	5	X	X
(Reczko & Hatzigerrorgiou, 2004)	ANN Approach	AA, SingalMotif	3	X	X
(Huang & Li, 2004)	fuzzy k-NN	diAA	11	X	X
(Cai & Chou, 2004)	Nearest Neighbor Approach	GO, InterProFunD, pseuAA	3, 4	X	X
(Chou & Cai, 2005)	Nearest Neighbor Approach	FunDC(5875D), pseuAA	3, 4	X	X
(Bhasin & Raghava, 2004)	SVM	AA, diAA	4	X	X
(Lee et al, 2006)	PLPD	AA, diAA, gapAA, InterProFunD	22	O	O
(Chou & Shen, 2007)	Nearest Neighbor Approach	GO, InterProFunD, pseuAA	22	O	X
(Shatkay et al, 2007)	SVM	SignalMotif, AA, text-based feature	11	X	X
(Garg et al, 2009)	k-NN, PNN	AA, sequence order, physicochemical properties	11	X	X
(Zhu et al, 2009)	SVM	AA, PSSM	14	O	X
(Shen & Burger, 2010)	SVM	AA, groupedAA, gapAA,, GO	4	X	X
(Mei et al, 2011)	SVM	AA, diAA, gapAA, GO	10	O	X
(Wang et al, 2011)	Frequent Pattern Tree	Motif, Overall-sequence	12	X	X
(Mooney et al, 2011)	N-to-1 Neural Network	BLAST	5	X	O
(Tian et al, 2011)	PCA, WSVM	PesAA	20	X	X
(Pierleoni et al, 2011)	SVM	AA, ChemAA, protein length, GO	3	X	X

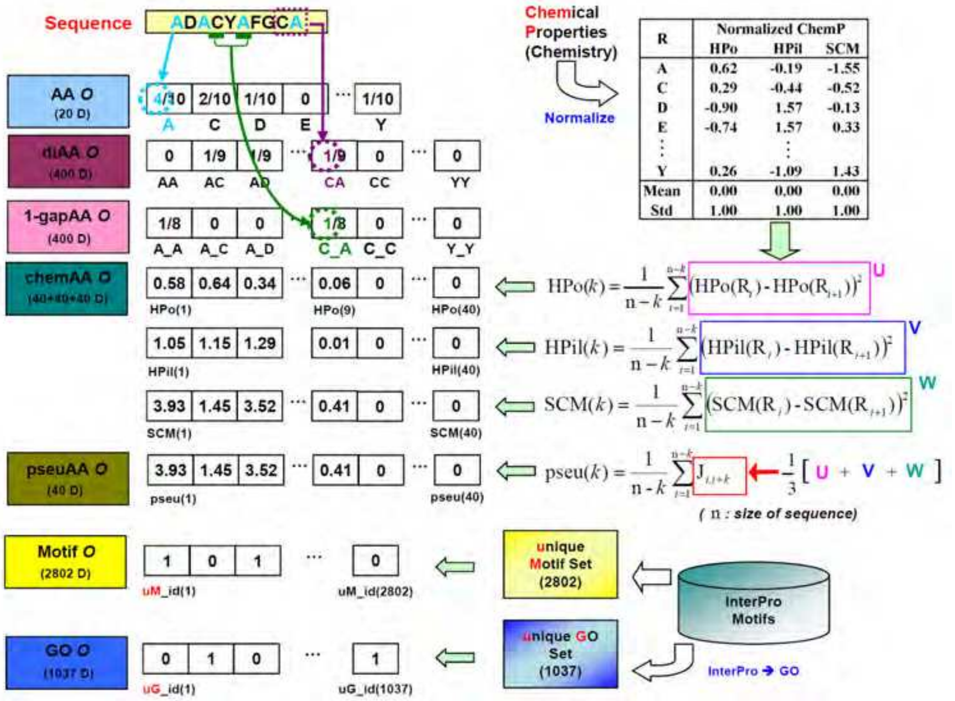
Table 4. Summary of previous methods for prediction of protein subcellular location.

4.2.2.1 Generation of feature vectors

Lee et al. used three types of feature to predict the localization and integrated these features (Lee et al, 2008). These are single protein features (***S***) and two kinds of network neighbourhood features (***N*** and ***L***).

Seven ***S*** features were based on a protein's primary sequence and its chemical properties. Amino acid composition frequencies (AA), adjacent pair amino acid frequencies (diAA) and pair-wise amino acid frequencies with a gap which is length of 1 (gapAA) from a protein's primary sequence were used. Also, three kinds of chemical amino acid compositions (chemAA) were generated from normalized hydrophobicity (HPo), hydrophilicity (HPil), or side-chain mass (SCM). Also, they combined these chemical properties into pseudo-amino acid composition (pseuAA), which is another ***S*** feature vector. Occurrences of known signalling motifs in the primary protein sequence (Motif) are also used as one of the ***S*** features. The last ***S*** feature encoded functional annotations of the protein from Gene Ontology (GO) (Ashburner et al, 2000). Figure 1 provides an example.

N network features are summary of ***S*** features from neighbourhood of a protein. Knowledge for neighbours of a protein comes from PPI data, which are pooled from various databases such as BioGRID (Stark et al, 2011), DIP (Salwinski et al, 2004) and SGD (Engel et al, 2010). ***L*** network features are summary of location distribution of interacting neighbours. Figure 2A shows a relationship among the three PPI databases. It shows that a single protein interaction database covers a different part of the whole reported interactions. The diagonal pattern in Figures 2B-D shows that interacting protein pairs share similar localization information. For example, a protein in an "ER to Golgi" tends to interact with other proteins which localized in the "ER to Golgi" more than other localizations.



R	Normalized ChemP		
	HPo	HPil	SCM
A	0.62	-0.19	-1.55
C	0.29	-0.44	-0.52
D	-0.90	1.57	-0.13
E	-0.74	1.57	0.33
⋮		⋮	
Y	0.26	-1.09	1.43
Mean	0.00	0.00	0.00
Std	1.00	1.00	1.00

$$\mathrm{HPo}(k)=\frac{1}{n-k}\sum_{i=1}^{n-k}\underbrace{(\mathrm{HPo}(R_i)-\mathrm{HPo}(R_{i+1}))^2}_{U}$$

$$\mathrm{HPil}(k)=\frac{1}{n-k}\sum_{i=1}^{n-k}\underbrace{(\mathrm{HPil}(R_i)-\mathrm{HPil}(R_{i+1}))^2}_{V}$$

$$\mathrm{SCM}(k)=\frac{1}{n-k}\sum_{i=1}^{n-k}\underbrace{(\mathrm{SCM}(R_i)-\mathrm{SCM}(R_{i+1}))^2}_{W}$$

$$\mathrm{pseu}(k)=\frac{1}{n-k}\sum_{i=1}^{n-k}\boxed{J_{i,i+k}}\leftarrow\frac{1}{3}[U+V+W]$$

(n : size of sequence)

Fig. 1. Summary of feature generation scheme for a single protein (adapted from Lee et al, 2008).

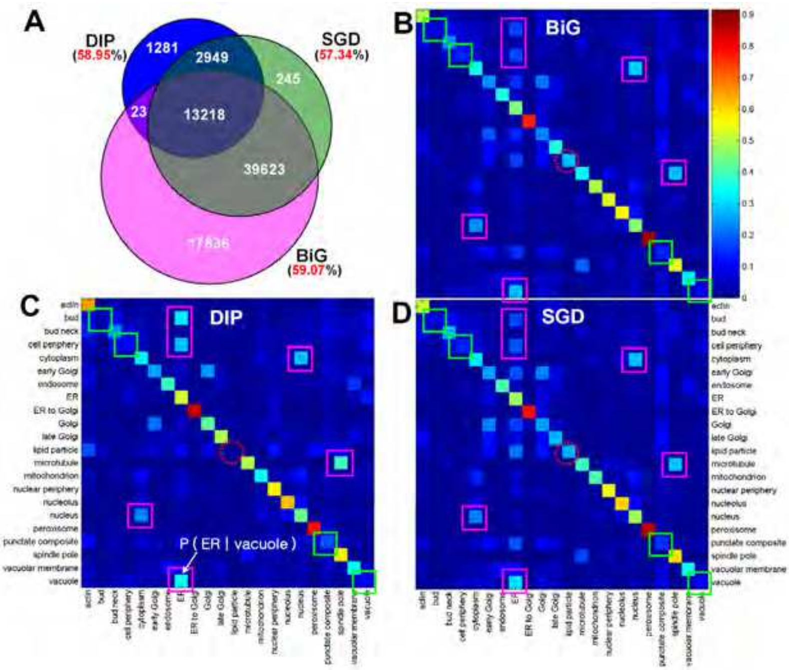


Fig. 2. Correlation between known localizations and protein interactions of yeast proteins. (A) The number of interactions (inside the circles) and the fraction of interactions whose proteins share localization information (outside the circles) of three interaction databases: BiG, DIP and SGD. (B-D) They show that interacting protein pairs have similar localization information in DIP, BiG and SGD (adapted from Lee et al, 2008).

4.2.2.2 Divide-and-Conquer k-Nearest Neighbour (DC-kNN) Classifier

After generating feature vectors, large-scale feature vectors with a high order may generate. A high dimensional feature vectors generally cause some problems like *curse-of-dimensionality*. In other words, data from higher dimensional feature vectors usually require a corresponding amount of inputs and it, sometimes, causes an over-fitting problem to a given dataset (Guyon et al, 2002). Also some feature vectors may be useless in constructing a model for a specific localization. Thus, individual model for different subcellular localizations may require different sets of useful feature sets. Therefore, extraction for feasible feature vectors for individual localizations may be needed to construct robust and reliable prediction models.

To construct a prediction model, Lee et al. proposed a DC-kNN classifier which is a variety of a k-Nearest Neighbours classification algorithm. A DC-kNN classifier tackles high-dimensional features in a divide-and-conquer manner. Briefly mentioning, a DC-kNN has three main steps (Figure 3): dividing, choosing, and synthesizing. In the dividing step, the full feature vector is divided into ***m*** meaningful subsets. After the dividing step, the k-nearest neighbours are chosen for each protein and for each subvector. In the synthesizing step, results of kNNs of individual ***m*** sets are synthesized to produce confidence scores using an average of Area under the ROC curve (AUC) for each localization. DC-kNN finds a feasible combination of feature sub-vectors for each label (localization) based on a feature forward selection approach.

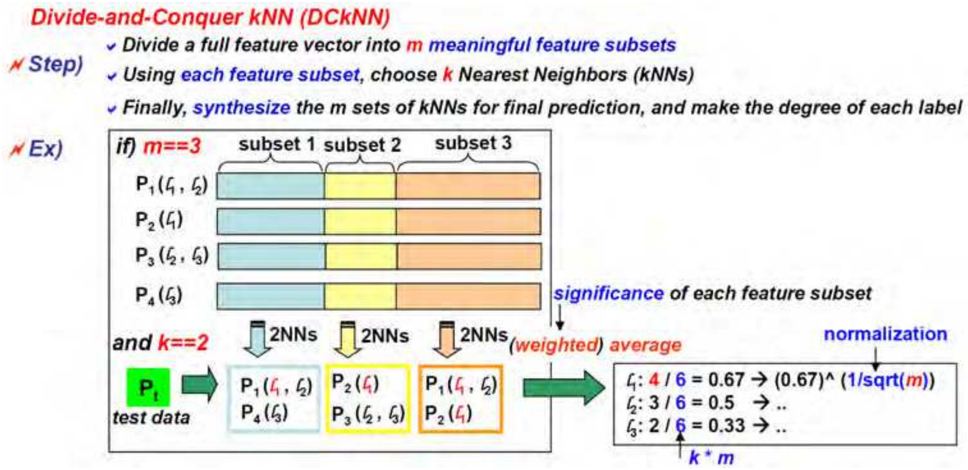

Divide-and-Conquer kNN (DCkNN)
Step)
Divide a full feature vector into m meaningful feature subsets
Using each feature subset, choose k Nearest Neighbors (kNNs)
Finally, synthesize the m sets of kNNs for final prediction, and make the degree of each label
Ex)
if) m==3
subset 1
subset 2
subset 3
P1 (L1, L2)
P2 (L1)
P3 (L2, L3)
P4 (L3)
significance of each feature subset
and k==2
2NNs
2NNs
2NNs
(weighted) average
normalization
Pt
test data
P1 (L1, L2)
P4 (L3)
P2 (L1)
P3 (L2, L3)
P1 (L1, L2)
P2 (L1)
L1: 4 / 6 = 0.67 → (0.67)^ (1/sqrt(m))
L2: 3 / 6 = 0.5 → ..
L3: 2 / 6 = 0.33 → ..
k * m

Fig. 3. Brief description of a DC-kNN (adapted from Lee et al, 2008).

4.2.3 Results of location prediction

Lee et al. first compared prediction performance of a DC-kNN for localization prediction with different feature sets: ***S*** features only, ***N*** features only, ***L*** features only, all features together (***S***+***N***+***L***), and random guesses. ***N*** and ***L*** features are generated using DIP (Salwinski et al, 2004). Performance of each case was evaluated by the technique of leave-one-out cross-validation (LOOCV). Proteins of *Saccharomyces cerevisiae* (n=3914) (Huh et al, 2003) were used for the LOOCV. They used three different performance metrics: Top-K, Total, and Balanced. These metrics were used to summarize the results of 3914 LOOCV runs. Top-K measurement considers as correct if at least one of the real localization of a protein is in the top-K predictions. Total measurement counts all the correctly predicted localizations based on the number of real localizations of test data. Balanced measure calculates the averaged fraction of correctly predicted proteins in each localization. As a result, every classifier showed clearly better performance than random guess (Figure 4A), and combination of ***S***, ***N***, and ***L*** features showed the highest performance.

Figures 4A and 4B inform that information of neighbourhood acquired from a PPI database improves prediction performance. However, Figure 4C illustrates that acquiring more information does not always contribute to an improvement of performance. On the contrary, additional information can decrease prediction performance. To find the necessary feature vectors for each localization, Lee et al. used a DC-kNN and found feasible subsets using the prepared feature vectors for individual localizations (Figures 4C and 4D). Using the selected features for individual localizations, the average of the AUC values was 0.94.

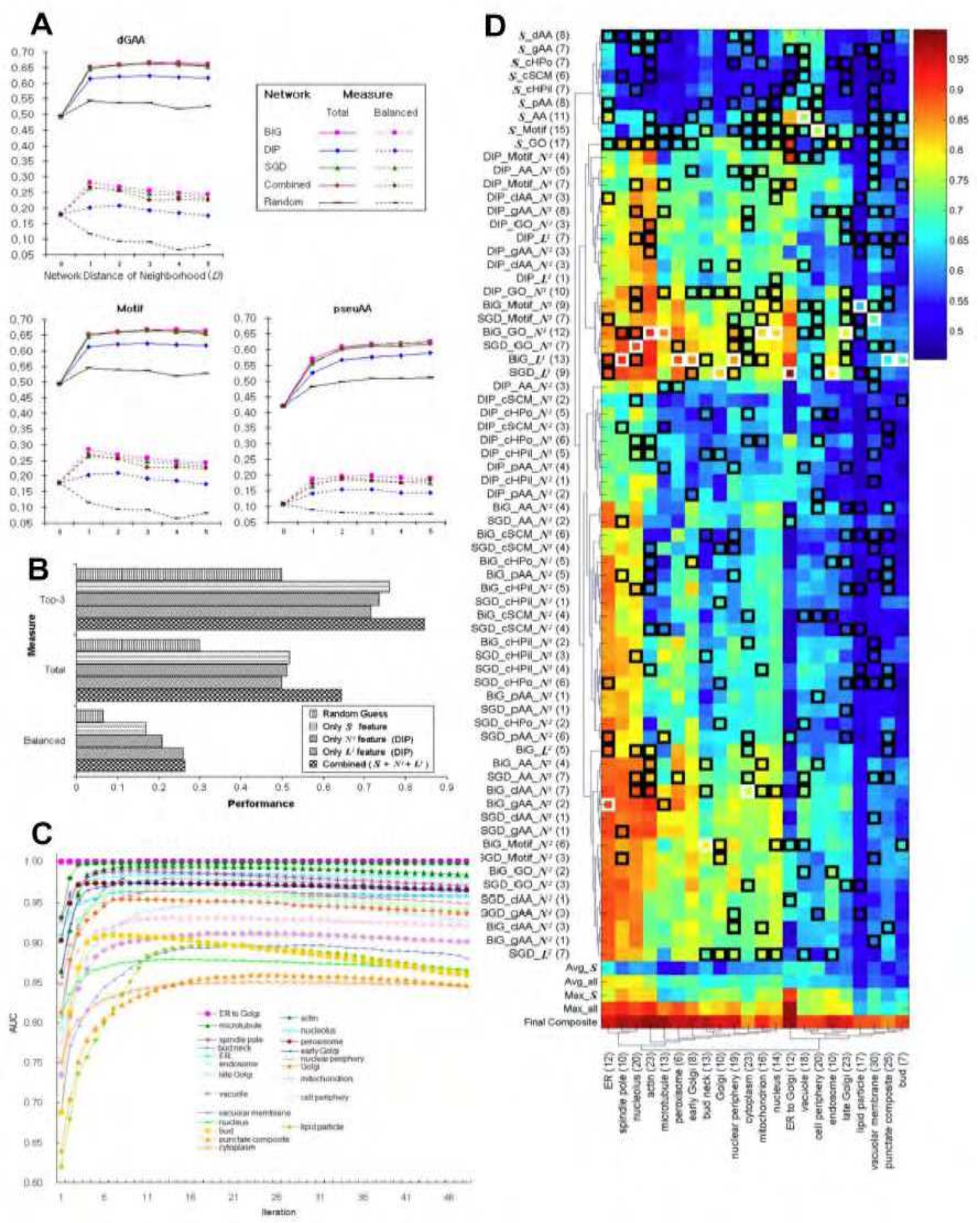


Fig. 4. (A) Shows performance of the classifiers by input from various kinds of feature. (B) shows performance for combination of feature vectors. (C) shows averaged AUC of the classifier for each localization based on feature selection using a DC-kNN. (D) shows selected feature sets for each of 22 localizations in yeast (adapted from Lee et al, 2008).

Based on the methodology, Lee et al. applied their method to the prediction of the localizations of genome-wide yeast proteins. Surprisingly, they also validated novel localizations of 61 proteins. For example, Huh et al. reported that Noc4/Ypr144c and Utp21/Ylr409c were localized in the nucleus (Huh et al, 2003). However, the proposed method developed by Lee et al. predicted the localization of the both proteins as the nucleolus. They revaluated for both proteins using new experiments and finally confirmed the previous results of Huh et al. had errors (Figures 5A and 5B). The correct prediction mainly owes to the fact that Lee et al. combined evidence from multiple interacting partners. For example, Noc4 interacts with many other proteins known to exist in the nucleolus, so we can assume that Noc4 localizes nearby or directly in the nucleolus. They confirmed the assumption by the network neighbours (Lee et al, 2008) (Figure 5C).

The number of localizations and known PPIs for yeast proteins are larger than those for other organisms. In other words, some organisms have less information on known localization and protein interaction, which might make the location prediction difficult based on a PPI network. Lee et al. evaluated their method using yeast data with some random missing information (Lee et al, 2008). As a result relatively robust results were obtained with less information. For example, the average number of neighbours of a protein in yeast is 27 and the number in worm is three. Decrement in the number of neighbours from yeast to worm was 9-fold. However, the average of AUC value decreased from 0.94 (yeast) to 0.87 (worm) (Figure 6). In other words, their method can be easily applied, not only to yeast but to other species with less known localization and/or interaction information. Actually they predicted subcellular localization of fly, human, and Arabidopsis (Lee et al, 2008; Lee et al, 2010b) using protein interactions. The results of both works showed that the prediction worked well for the other organisms and could find real localizations of some unknown proteins (Figures 6-7).

They also compared a DC-kNN with two previous popular methods, ISort (Chou & Cai, 2005) and PSLT2 (Scott et al, 2005). ISort is a comprehensive sequence-based machine learning method. ISort can predict more than 15 compartments. PSLT2 is a previous method that used a protein interaction network to predict subcellular localizations. They compared to DC-kNN with ISort and PSLT2 using both total and balanced measures. As illustrated in Figure 8, DC-kNN outperformed both methods in total and balanced measurement.

4.2.4 Other network-based methods

After the study of Lee et al. in 2008, several studies based on network-based approaches tried to predict subcellular localization. Mintz-Oron et al. used a constraint-based method for predicting subcellular localization of enzymes based on their embedding metabolic network, relying on a parsimony principle of a minimal number of cross-membrane metabolite transporters (Mintz-Oron et al, 2009). They showed that their method outperformed pathway enrichment-base methods. Another group constructed a decision tree-based meta-classifier for identification of essential genes (Acencio & Lemke, 2009). Their method relied on network topological features, cellular localization and biological process information for prediction of essential genes. Tung & Lee integrated various biological data sources to get information of neighbour proteins in a probabilistic gene-network (Tung & Lee, 2009). They predicted the subcellular localization using a Fuzzy k-nearest neighbour classifier. Lee et al. curated IntAct *Arabidopsis thaliana* PPI dataset

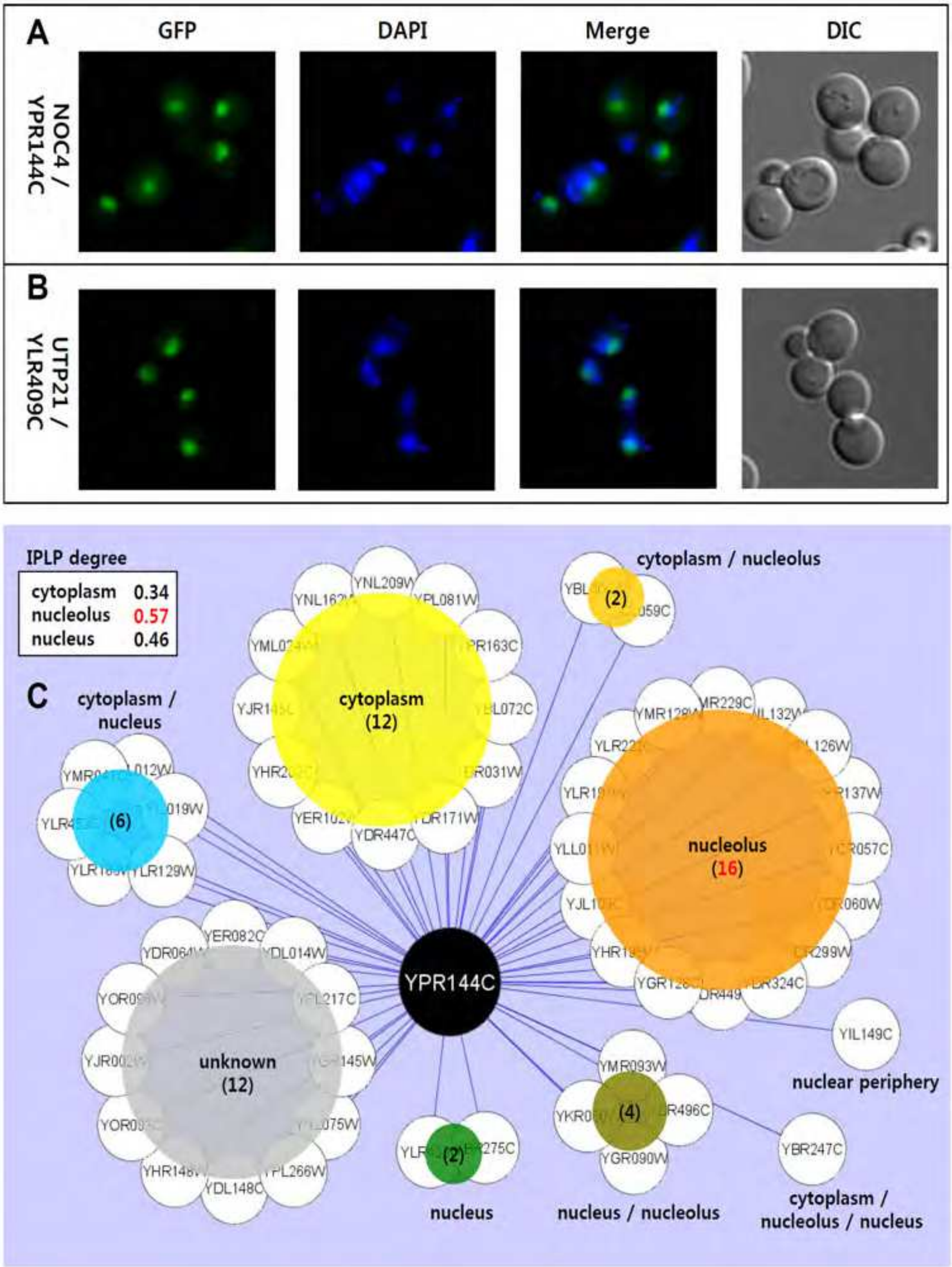


Fig. 5. (A, B) represent results of new experiments for Noc4/Ypr144c and Utp21/Ylr409c. (C) shows the interacting neighbours of Ypr144c (adapted from Lee et al, 2008).

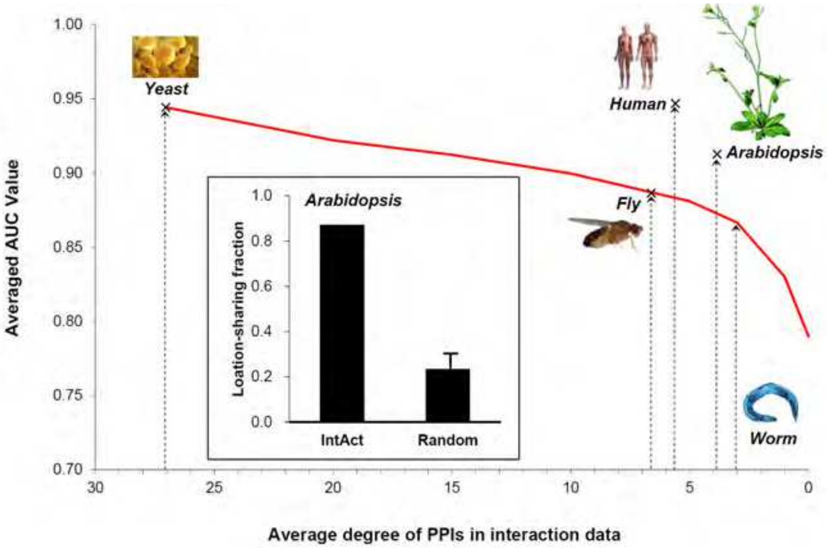


Fig. 6. Averaged AUC values across different organisms (adapted from Lee et al, 2010b).

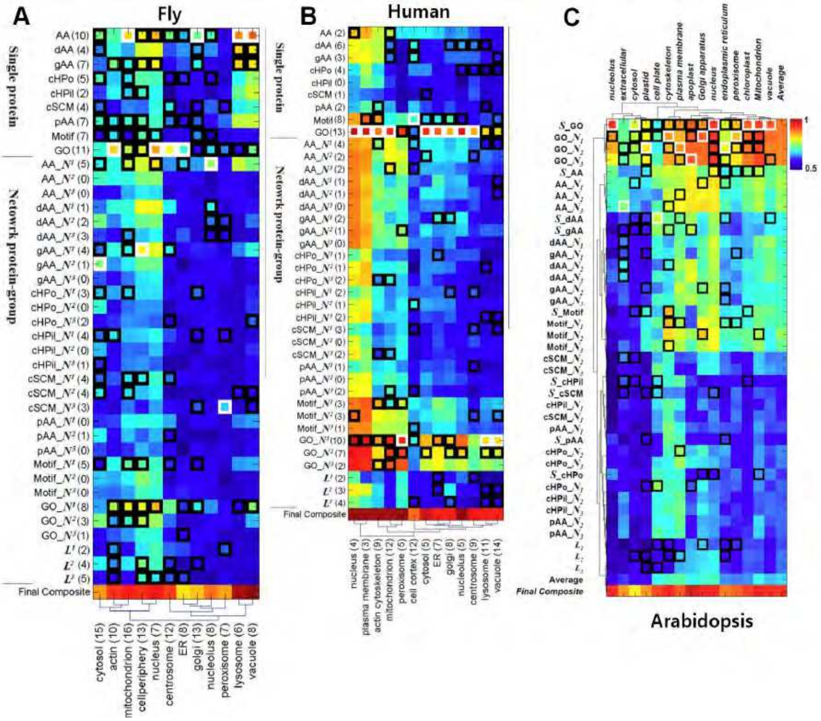


Fig. 7. Generated models for the location prediction for Fly (A), Human (B), and Arabidopsis (C) (adapted from Lee et al, 2008 and Lee et al, 2010b).

Lee et al. curated IntAct *Arabidopsis thaliana* PPI dataset (Aranda et al, 2010) using the DC-kNN method, which was proposed before and which showed good performance (Lee et al, 2010b). They also showed that the DC-kNN is applicable to other organisms. Kourmpetis et al. predicted a function of proteins in *Saccharomyces cerevisiae* based on network data, such as PPI data (Kourmpetis et al, 2010). They took a Bayesian Markov Random field analysis method for prediction and predicted the functions of 1170 un-annotated *Saccharomyces cerevisiae* proteins.

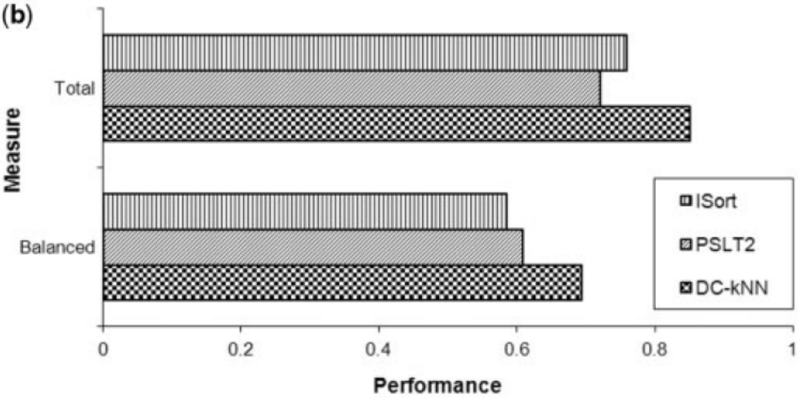


Fig. 8. Performance comparison of Isort, PSLT2 and DC-kNN (adapted from Lee et al, 2008).

5. Conclusions

We reviewed on PPI databases and the methods for detection of PPIs. Then, the computational methods of protein function prediction were briefly reviewed. We finally discussed that the prediction of protein function, especially the subcellular localization, shows outstanding performance when using PPIs data. This is because real biological functions are maintaining through a cascade of PPIs. Moreover, the computational approaches are very much promising when compared to the experimental identification especially for the false reading corrections. Functional genomics is an ongoing field in systems biology and this must be done well to drive further progress. We are facing other issues concerning the lack of conditional protein interactomes. We have identified and accumulated only static information at the molecular level in cells to make a scaffold of cellular systems. Computational methods should be applied to this conditional analysis when sufficient data become available and the next field of utilization would be personalized medicines, such as the early diagnosis with specific markers and treatments with specific drug targets.

6. Acknowledgement

This work was supported by Basic Science Research Programs through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0022887 and 2010-0018258).

7. References

Acencio ML, Lemke N (2009) Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC bioinformatics* 10: 290

Aloy P, Querol E, Aviles FX, Sternberg MJ (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *Journal of molecular biology* 311: 395-408

Aloy P, Russell RB (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* 19: 161-162

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25: 3389-3402

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2004) UniProt: the Universal Protein knowledgebase. *Nucleic acids research* 32: D115-119

Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, Kerssemakers J, Leroy C, Menden M, Michaut M, Montecchi-Palazzi L, Neuhauser SN, Orchard S, Perreau V, Roechert B, van Eijk K et al (2010) The IntAct molecular interaction database in 2010. *Nucleic acids research* 38: D525-531

Arnau V, Mars S, Marin I (2005) Iterative cluster analysis of protein interaction data. *Bioinformatics* 21: 364-378

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25: 25-29

Bader GD, Cary MP, Sander C (2006) Pathguide: a pathway resource list. *Nucleic acids research* 34: D504-506

Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics* 4: 2

Barker D, Pagel M (2005) Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS computational biology* 1: e3

Ben-Hur A, Noble WS (2005) Kernel methods for predicting protein-protein interactions. *Bioinformatics* 21 Suppl 1: i38-46

Berggard T, Linse S, James P (2007) Methods for the detection and analysis of protein-protein interactions. *Proteomics* 7: 2833-2842

Bhasin M, Raghava GP (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic acids research* 32: W414-419

Blaschke C, Hoffmann R, Oliveros JC, Valencia A (2001) Extracting information automatically from biological literature. *Comparative and functional genomics* 2: 310-313

Bonetta L (2010) Protein-protein interactions: Interactome under construction. *Nature* 468: 851-854

Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y (1998) Predicting function: from genes to genomes and back. *Journal of molecular biology* 283: 707-725

Bousquet-Dubouch MP, Fabre B, Monsarrat B, Burlet-Schiltz O (2011) Proteomics to study the diversity and dynamics of proteasome complexes: from fundamentals to the clinic. *Expert review of proteomics* 8: 459-481

Bowers PM, Cokus SJ, Eisenberg D, Yeates TO (2004) Use of logic relationships to decipher protein network organization. *Science* 306: 2246-2249

Cai YD, Chou KC (2003) Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. *Biochemical and biophysical research communications* 305: 407-411

Cai YD, Chou KC (2004) Predicting 22 protein localizations in budding yeast. *Biochemical and biophysical research communications* 323: 425-428

Cai YD, Liu XJ, Xu XB, Chou KC (2000) Support vector machines for prediction of protein subcellular location. *Molecular cell biology research communications : MCBRC* 4: 230-233

Cai YD, Zhou GP, Chou KC (2003) Support vector machines for predicting membrane protein types by using functional domain composition. *Biophysical journal* 84: 3257-3263

Carlson BA, Xu XM, Kryukov GV, Rao M, Berry MJ, Gladyshev VN, Hatfield DL (2004) Identification and characterization of phosphoseryl-tRNA[Ser]Sec kinase. *Proceedings of the National Academy of Sciences of the United States of America* 101: 12848-12853

Cedano J, Aloy P, Perez-Pons JA, Querol E (1997) Relation between amino acid composition and cellular location of proteins. *Journal of molecular biology* 266: 594-600

Cho YR, Hwang W, Ramanathan M, Zhang A (2007) Semantic integration to identify overlapping functional modules in protein interaction networks. *BMC bioinformatics* 8: 265

Chou KC (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochemical and biophysical research communications* 278: 477-483

Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 43: 246-255

Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *The Journal of biological chemistry* 277: 45765-45769

Chou KC, Cai YD (2003) Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *Journal of cellular biochemistry* 90: 1250-1260

Chou KC, Cai YD (2005) Predicting protein localization in budding yeast. *Bioinformatics (Oxford, England)* 21: 944-950

Chou KC, Elrod DW (1999) Protein subcellular location prediction. *Protein engineering* 12: 107-118

Chou KC, Shen HB (2007) Recent progress in protein subcellular location prediction. *Analytical biochemistry* 370: 1-16

Chua HN, Sung WK, Wong L (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics* 22: 1623-1630

Cote R, Reisinger F, Martens L, Barsnes H, Vizcaino JA, Hermjakob H (2010) The Ontology Lookup Service: bigger and better. *Nucleic acids research* 38: W155-160

Deng M, Chen T, Sun F (2004) An integrated probabilistic model for functional prediction of proteins. *Journal of computational biology : a journal of computational molecular cell biology* 11: 463-475

Drawid A, Gerstein M (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *Journal of molecular biology* 301: 1059-1075

Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. *Genome informatics International Conference on Genome Informatics* 23: 205-211

Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of molecular biology* 300: 1005-1016

Engel SR, Balakrishnan R, Binkley G, Christie KR, Costanzo MC, Dwight SS, Fisk DG, Hirschman JE, Hitz BC, Hong EL, Krieger CJ, Livstone MS, Miyasato SR, Nash R, Oughtred R, Park J, Skrzypek MS, Weng S, Wong ED, Dolinski K et al (2010) Saccharomyces Genome Database provides mutant phenotype data. *Nucleic acids research* 38: D433-436

Ezkurdia I, Bartoli L, Fariselli P, Casadio R, Valencia A, Tress ML (2009) Progress and challenges in predicting protein-protein interaction sites. *Briefings in bioinformatics* 10: 233-246

Ferrer L, Dale JM, Karp PD (2010) A systematic study of genome context methods: calibration, normalization and combination. *BMC bioinformatics* 11: 493

Fetrow JS, Skolnick J (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *Journal of molecular biology* 281: 949-968

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A (2010) The Pfam protein families database. *Nucleic acids research* 38: D211-222

Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, Simon I, Hua S, deFays K, Lambert C, Nakai K, Brinkman FS (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic acids research* 31: 3613-3617

Garg P, Sharma V, Chaudhari P, Roy N (2009) SubCellProt: predicting protein subcellular localization using machine learning approaches. *In silico biology* 9: 35-44

Gaudet P, Livstone MS, Lewis SE, Thomas PD (2011) Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefings in bioinformatics* 12: 449-462

Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M et al (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440: 631-636

Gingras AC, Gstaiger M, Raught B, Aebersold R (2007) Analysis of protein complexes using mass spectrometry. *Nature reviews Molecular cell biology* 8: 645-654

Grigoriev A (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucleic acids research* 29: 3513-3519

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* 46: 389-422

Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. *Nature* 402: C47-52

Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T (2001) Assessment of prediction accuracy of protein function from protein--protein interaction data. *Yeast* 18: 523-531

Hosur R, Xu J, Bienkowska J, Berger B (2011) iWRAP: An interface threading approach with application to prediction of cancer-related protein-protein interactions. *Journal of molecular biology* 405: 1295-1310

Hua S, Sun Z (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics (Oxford, England)* 17: 721-728

Huang Y, Li Y (2004) Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics (Oxford, England)* 20: 21-28

Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK (2003) Global analysis of protein localization in budding yeast. *Nature* 425: 686-691

Huynen M, Doerks T, Eisenhaber F, Orengo C, Sunyaev S, Yuan Y, Bork P (1998) Homology-based fold predictions for Mycoplasma genitalium proteins. *Journal of molecular biology* 280: 323-326

Imai K, Nakai K (2010) Prediction of subcellular locations of proteins: where to proceed? *Proteomics* 10: 3970-3983

Itoh T, Takemoto K, Mori H, Gojobori T (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Molecular biology and evolution* 16: 332-346

Kemmer D, Huang Y, Shah SP, Lim J, Brumm J, Yuen MM, Ling J, Xu T, Wasserman WW, Ouellette BF (2005) Ulysses - an application for the projection of molecular interactions across species. *Genome biology* 6: R106

Kemmeren P, Holstege FC (2003) Integrating functional genomics data. *Biochemical Society transactions* 31: 1484-1487

Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, Bader GD, Xenarios I, Wojcik J, Sherman D, Tyers M, Salama JJ, Moore S, Ceol A, Chatr-Aryamontri A, Oesterheld M, Stumpflen V, Salwinski L, Nerothin J, Cerami E et al (2007) Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions. *BMC biology* 5: 44

King AD, Przulj N, Jurisica I (2004) Protein complex prediction via cost-based clustering. *Bioinformatics* 20: 3013-3020

Kohl M, Wiese S, Warscheid B (2011) Cytoscape: software for visualization and analysis of biological networks. *Methods Mol Biol* 696: 291-303

Koonin EV, Wolf YI, Aravind L (2001) Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach. *Genome research* 11: 240-252

Kourmpetis YA, van Dijk AD, Bink MC, van Ham RC, ter Braak CJ (2010) Bayesian Markov Random Field analysis for protein function prediction based on network data. *PloS one* 5: e9293

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B et al (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* 440: 637-643

Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research* 21: 1109-1121

Lee I, Lehner B, Vavouri T, Shin J, Fraser AG, Marcotte EM (2010a) Predicting genetic modifier loci using functional gene networks. *Genome research* 20: 1143-1153

Lee K, Chuang HY, Beyer A, Sung MK, Huh WK, Lee B, Ideker T (2008) Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic acids research* 36: e136

Lee K, Kim DW, Na D, Lee KH, Lee D (2006) PLPD: reliable protein localization prediction from imbalanced and overlapped datasets. *Nucleic acids research* 34: 4655-4666

Lee K, Thorneycroft D, Achuthan P, Hermjakob H, Ideker T (2010b) Mapping plant interactomes using literature curated and predicted protein-protein interaction data sets. *The Plant cell* 22: 997-1005

Lees JG, Heriche JK, Morilla I, Ranea JA, Orengo CA (2011) Systematic computational prediction of protein interaction networks. *Physical biology* 8: 035008

Lu H, Zhu X, Liu H, Skogerbo G, Zhang J, Zhang Y, Cai L, Zhao Y, Sun S, Xu J, Bu D, Chen R (2004) The interactome as a tree--an attempt to visualize the protein-protein interaction network in yeast. *Nucleic acids research* 32: 4804-4811

Lukk M, Kapushesky M, Nikkila J, Parkinson H, Goncalves A, Huber W, Ukkonen E, Brazma A (2010) A global map of human gene expression. *Nature biotechnology* 28: 322-324

Luttgen H, Rohdich F, Herz S, Wungsintaweekul J, Hecht S, Schuhr CA, Fellermeier M, Sagner S, Zenk MH, Bacher A, Eisenreich W (2000) Biosynthesis of terpenoids: YchB protein of Escherichia coli phosphorylates the 2-hydroxy group of 4-diphosphocytidyl-2C-methyl-D-erythritol. *Proceedings of the National Academy of Sciences of the United States of America* 97: 1062-1067

Mani R, St Onge RP, Hartman JLt, Giaever G, Roth FP (2008) Defining genetic interaction. *Proceedings of the National Academy of Sciences of the United States of America* 105: 3461-3466

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* 285: 751-753

Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome research* 11: 2120-2126

Mei S, Fei W (2010) Amino acid classification based spectrum kernel fusion for protein subnuclear localization. *BMC bioinformatics* 11 Suppl 1: S17

Mei S, Fei W, Zhou S (2011) Gene ontology based transfer learning for protein subcellular localization. *BMC bioinformatics* 12: 44

Mintz-Oron S, Aharoni A, Ruppin E, Shlomi T (2009) Network-based prediction of metabolic enzymes' subcellular localization. *Bioinformatics (Oxford, England)* 25: i247-252

Mooney C, Wang YH, Pollastri G (2011) SCLpred: protein subcellular localization prediction by N-to-1 neural networks. *Bioinformatics (Oxford, England)* 27: 2812-2819

Mount DW (2007) Using a FASTA Sequence Database Similarity Search. *CSH protocols* 2007: pdb top16

Nair R, Rost B (2002) Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics (Oxford, England)* 18 Suppl 1: S78-86

Nair R, Rost B (2003) Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins* 53: 917-930

Nakai K, Horton P (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in biochemical sciences* 24: 34-36

Nakai K, Kanehisa M (1991) Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins* 11: 95-110

Nakai K, Kanehisa M (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 14: 897-911

Nakashima H, Nishikawa K (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *Journal of molecular biology* 238: 54-61

Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stumpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P, Salama JJ, Moore S, Wojcik J, Bader GD, Vidal M, Cusick ME, Gerstein M, Gavin AC, Superti-Furga G, Greenblatt J et al (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature biotechnology* 25: 894-898

Pan YX, Zhang ZZ, Guo ZM, Feng GY, Huang ZD, He L (2003) Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *Journal of protein chemistry* 22: 395-402

Park KJ, Kanehisa M (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics (Oxford, England)* 19: 1656-1663

Pazos F, Ranea JA, Juan D, Sternberg MJ (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *Journal of molecular biology* 352: 1002-1015

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America* 96: 4285-4288

Pereira-Leal JB, Enright AJ, Ouzounis CA (2004) Detection of functional modules from protein interaction networks. *Proteins* 54: 49-57

Persico M, Ceol A, Gavrila C, Hoffmann R, Florio A, Cesareni G (2005) HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC bioinformatics* 6 Suppl 4: S21

Pierleoni A, Martelli PL, Casadio R (2011) MemLoci: predicting subcellular localization of membrane proteins in eukaryotes. *Bioinformatics (Oxford, England)* 27: 1224-1230

Pitre S, Alamgir M, Green JR, Dumontier M, Dehne F, Golshani A (2008) Computational methods for predicting protein-protein interactions. *Advances in biochemical engineering/biotechnology* 110: 247-267

Recillas-Targa F (2006) Multiple strategies for gene transfer, expression, knockdown, and chromatin influence in mammalian cell lines and transgenic animals. *Molecular biotechnology* 34: 337-354

Reczko M, Hatzigerrorgiou A (2004) Prediction of the subcellular localization of eukaryotic proteins using sequence signals and composition. *Proteomics* 4: 1591-1596

Reid AJ, Ranea JA, Clegg AB, Orengo CA (2010) CODA: accurate detection of functional associations between proteins in eukaryotic genomes using domain fusion. *PloS one* 5: e10908

Reinhardt A, Hubbard T (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic acids research* 26: 2230-2236

Rives AW, Galitski T (2003) Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America* 100: 1128-1133

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The Database of Interacting Proteins: 2004 update. *Nucleic acids research* 32: D449-451

Sanchez-Chapado M, Angulo JC, Ibarburen C, Aguado F, Ruiz A, Viano J, Garcia-Segura JM, Gonzalez-Esteban J, Rodriquez-Vallejo JM (1997) Comparison of digital rectal examination, transrectal ultrasonography, and multicoil magnetic resonance imaging for preoperative evaluation of prostate cancer. *European urology* 32: 140-149

Schwikowski B, Uetz P, Fields S (2000) A network of protein-protein interactions in yeast. *Nature biotechnology* 18: 1257-1261

Scott MS, Calafell SJ, Thomas DY, Hallett MT (2005) Refining protein subcellular localization. *PLoS computational biology* 1: e66

Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. *Molecular systems biology* 3: 88

Shatkay H, Hoglund A, Brady S, Blum T, Donnes P, Kohlbacher O (2007) SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics (Oxford, England)* 23: 1410-1417

Shen YQ, Burger G (2010) TESTLoc: protein subcellular localization prediction from EST data. *BMC bioinformatics* 11: 563

Shoemaker BA, Panchenko AR (2007) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS computational biology* 3: e43

Shoemaker BA, Zhang D, Thangudu RR, Tyagi M, Fong JH, Marchler-Bauer A, Bryant SH, Madej T, Panchenko AR (2010) Inferred Biomolecular Interaction Server--a web server to analyze and predict protein interacting partners and binding sites. *Nucleic acids research* 38: D518-524

Singh R, Park D, Xu J, Hosur R, Berger B (2010) Struct2Net: a web service to predict protein-protein interactions using a structure-based approach. *Nucleic acids research* 38: W508-515

Skarnes WC, Rosen B, West AP, Koutsourakis M, Bushell W, Iyer V, Mujica AO, Thomas M, Harrow J, Cox T, Jackson D, Severin J, Biggs P, Fu J, Nefedov M, de Jong PJ, Stewart AF, Bradley A (2011) A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* 474: 337-342

Skrabanek L, Saini HK, Bader GD, Enright AJ (2008) Computational prediction of protein-protein interactions. *Molecular biotechnology* 38: 1-17

Smith KL, DeVos V, Bryden H, Price LB, Hugh-Jones ME, Keim P (2000) Bacillus anthracis diversity in Kruger National Park. *Journal of clinical microbiology* 38: 3780-3784

Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America* 100: 12123-12128

Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, Reguly T, Rust JM, Winter A, Dolinski K, Tyers M (2011) The BioGRID Interaction Database: 2011 update. *Nucleic acids research* 39: D698-704

Stein A, Panjkovich A, Aloy P (2009) 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic acids research* 37: D300-304

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302: 249-255

Suter B, Kittanakom S, Stagljar I (2008) Two-hybrid technologies in proteomics research. *Current opinion in biotechnology* 19: 316-323

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research* 39: D561-568

Tian J, Gu H, Liu W, Gao C (2011) Robust prediction of protein subcellular localization combining PCA and WSVMs. *Computers in biology and medicine* 41: 648-652

Tikk D, Thomas P, Palaga P, Hakenberg J, Leser U (2010) A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS computational biology* 6: e1000837

Tuncbag N, Gursoy A, Guney E, Nussinov R, Keskin O (2008) Architectures and functional coverage of protein-protein interfaces. *Journal of molecular biology* 381: 785-802

Tuncbag N, Kar G, Keskin O, Gursoy A, Nussinov R (2009) A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Briefings in bioinformatics* 10: 217-232

Tung TQ, Lee D (2009) A method to improve protein subcellular localization prediction by integrating various biological data sources. *BMC bioinformatics* 10 Suppl 1: S43

Turinsky AL, Razick S, Turner B, Donaldson IM, Wodak SJ (2010) Literature curation of protein interactions: measuring agreement across major public databases. *Database : the journal of biological databases and curation* 2010: baq026

Turner B, Razick S, Turinsky AL, Vlasblom J, Crowdy EK, Cho E, Morrison K, Donaldson IM, Wodak SJ (2010) iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database : the journal of biological databases and curation* 2010: baq023

Vazquez A, Flammini A, Maritan A, Vespignani A (2003) Global protein function prediction from protein-protein interaction networks. *Nature biotechnology* 21: 697-700

Wang J, Li C, Wang E, Wang X (2011) An FPT approach for predicting protein localization from yeast genomic data. *PloS one* 6: e14449

Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. *Quarterly reviews of biophysics* 36: 307-340

Xia JF, Wang SL, Lei YK (2010) Computational methods for the prediction of protein-protein interactions. *Protein and peptide letters* 17: 1069-1078

Yeats C, Lees J, Carter P, Sillitoe I, Orengo C (2011) The Gene3D Web Services: a platform for identifying, annotating and comparing structural domains in protein sequences. *Nucleic acids research* 39: W546-550

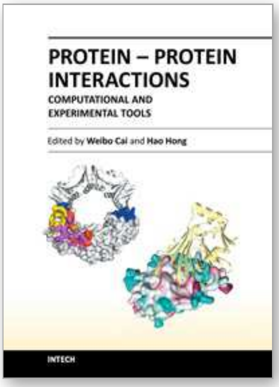
Yuan Z (1999) Prediction of protein subcellular locations using Markov chain models. *FEBS letters* 451: 23-26

Zhang QC, Petrey D, Norel R, Honig BH (2010) Protein interface conservation across structure space. *Proceedings of the National Academy of Sciences of the United States of America* 107: 10896-10901

Zhang Z, Sun H, Zhang Y, Zhao Y, Shi B, Sun S, Lu H, Bu D, Ling L, Chen R (2006) Genome-wide analysis of mammalian DNA segment fusion/fission. *Journal of theoretical biology* 240: 200-208

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. *Proteins* 50: 44-48

Zhu L, Yang J, Shen HB (2009) Multi label learning for prediction of human protein subcellular localizations. *The protein journal* 28: 384-390



PROTEIN – PROTEIN INTERACTIONS
COMPUTATIONAL AND EXPERIMENTAL TOOLS
Edited by Weibo Cai and Hao Hong
INTECH

Protein-Protein Interactions - Computational and Experimental Tools

Edited by Dr. Weibo Cai

ISBN 978-953-51-0397-4
Hard cover, 472 pages
Publisher InTech
Published online 30, March, 2012
Published in print edition March, 2012

Proteins are indispensable players in virtually all biological events. The functions of proteins are coordinated through intricate regulatory networks of transient protein-protein interactions (PPIs). To predict and/or study PPIs, a wide variety of techniques have been developed over the last several decades. Many in vitro and in vivo assays have been implemented to explore the mechanism of these ubiquitous interactions. However, despite significant advances in these experimental approaches, many limitations exist such as false-positives/false-negatives, difficulty in obtaining crystal structures of proteins, challenges in the detection of transient PPI, among others. To overcome these limitations, many computational approaches have been developed which are becoming increasingly widely used to facilitate the investigation of PPIs. This book has gathered an ensemble of experts in the field, in 22 chapters, which have been broadly categorized into Computational Approaches, Experimental Approaches, and Others.

Woojin Jung, Hyun-Hwan Jeong and KiYoung Lee (2012). Protein Interactome and Its Application to Protein Function Prediction, Protein-Protein Interactions - Computational and Experimental Tools, Dr. Weibo Cai (Ed.), ISBN: 978-953-51-0397-4, InTech, Available from: http://www.intechopen.com/books/protein-protein-interactions-computational-and-experimental-tools/protein-interactome-and-its-application-to-protein-function-prediction

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821