Protein Interactions for functional Genomics, In:
Biological Data Mining in Protein Interaction Networks
Pp.223-238. Eds. Xiao-Li Li, See-Kiong Ng
Idea Group Inc (IGI), 2009
Hershey, USA

Chapter XIII

Protein Interactions for Functional Genomics

Pablo Minguez[1,2] & Joaquin Dopazo[1,2,3]

1 Department of Bioinformatics and Genomics, 2 CIBER de Enfermedades Raras (CIBERER), ISCIII, Spain, 3 Functional Genomics Node (INB), Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain.

Abstract

Here we will revise the state of the art in the use of protein-protein interactions (ppis) within the context of the interpretation of genomic experiments. We will report the available resources and methodologies used to create a curated compilation of ppis introducing a novel approach to filter interactions. Especial attention will be paid in the complexity of the topology of the networks formed by proteins (nodes) and pairwise interactions (edges). These networks can be studied using graph theory and a brief introduction to the characterization of biological networks and definitions of the more used network parameters is also given.
Also a report on the available resources to perform different modes of functional profiling using ppi data is provided along with a discussion on the approaches that have typically been applied into this context. We will also introduce a novel methodology for the evaluation of networks and some examples of its application.

Introduction

The available data for protein-protein interactions (ppis) has increased enormously in the last few years with the emergence of high-throughput techniques that can report thousands of ppis in a short time span. The most used techniques in this field are: yeast two hybrid (y2h), tandem affinity purification (TAP) and high-throughput Mass Spectrometry techniques (MS). Reviews on these and related methodologies can be found in Drewes and Bouwmeester (2003), Cho *et al*. (2003), Falk *et al*. (2007) and Berggard *et al*. (2007).

The reliability of this data is not exempt of controversy. Studies comparing resulting data from several experiments demonstrate that the overlap between them is not as complete as desirable. This can be due to the fact that some methodologies do not reach the saturation point (Bader & Hogue, 2002) or because of the lack of accuracy and coverage on some of them (von Mering *et al*., 2002). A conventional large-scale experiment can cover only 3-9% of the total interactome, so limited overlap should be expected (Han *et al*., 2005). False positives are also a problem: in y2h these represent up to 50% of the total data (Ito *et al*., 2001; Mrowka *et al*., 2001). Moreover, there is a bias in the functional categories of the ppis each technique detects, e.g. y2h fails in detecting proteins involved in translation (von Mering *et al*., 2002).

Beyond discussions about accuracy and coverage of this kind of experiments, the relevance of ppis in the cellular machinery has fostered an unprecedented interest in the exploration of the interactome of model organisms such as *Saccharomyces cerevisiae* (Uetz *et al*., 2000; Ito *et al*., 2001), *Drosophila melanogaster* (Gio et al., 2003; Formstecher *et al*., 2005), *Caenorhabditis elegans* (Li *et al*., 2004) or human (Stelzl *et al*., 2005, Rual *et al*., 2005), just to cite a few examples.

Actually, after years of intensive study, there is a high-quality, literature curated set of ppis free from false positives that probably represents the complete yeast interactome (Reguly *et al*., 2006). In the case of human, the scenario is still far away from this degree of detail. The estimated size of the human interactome is of 650,000 ppis (Stumpf *et al*., 2008). None of the public databases contain more than 10% of this number of ppis, and a compilation of all the known ppis would only cover about 10% of the interactions.

The interactome is an abstract scaffold that does not provide information about particular conditions, cell developmental stage or cell type in which a particular ppi occurs (if any). To infer a case-specific interactome it is necessary to integrate other types of data that provide information that allows inferring the active ppis at a particular condition. To achieve this, the transcriptome, defined as the set of transcripts that are expressed at a given moment in a particular cell type, can be used. An integrative study of the interactome filtered by the transcriptome will provide valuable information on the active ppis in a given cell state.

Actually, ppis play a central role at almost every level of cell activity: they are involved in the structure of organelles (structural proteins), transport machinery (nuclear pore importins), response to stimulus (signalling cascades), regulation of gene expression (transcription factors), protein modification (kinases) among many other processes. The proper use of this type of information is of crucial importance in order to understand cell behaviour.

However, the conventional methodologies used to understand the functional basis of the cell behaviour are almost restricted to functional profiling methods. Such methods exploit the differences observed in the comparison of transcriptomes among different experimental conditions to find over-representations of predefined functional modules of genes (see Dopazo 2006 for a recent review). Classically, standard annotations like Gene Ontology

(GO) terms (Ashburner et al., 2000), KEGG pathways (Kanehisa et al., 2004) or Mesh terms, have been used to define such modules. Nevertheless, ppis have not extensively used for such puroses.

The combination of expression and interactions has been used to infer gene function (Ideker *et al*., 2001), to extract signatures to predict disease phenotypes (Camargo & Azuaje, 2007; Lee *et al*., 2007; Liu *et al*., 2007; Chuang *et al*., 2007) as well as to detect possible drug targets by inferring topological features of particular classes of genes (Wachi *et al*., 2005; Johsson and Bates, 2006).

It is widely accepted that there are very few processes that can be explained by the action of a single protein. On the contrary, the units of activity involved in cellular processes seem to be modules composed by several interacting molecules (Hartwell *et al*., 1999; Barabasi and Oltvai, 2004). Apart from classical definitions of these modules, such as proteins that share a GO term or proteins integrating the same biological pathway, ppi data is also used to define modules, as representative of units of action characterized by the interaction of their components.

Interactome-related definitions

Ppi resources

In this new era of massive production of biological data, an important challenge is its storage in a standardised format with the proper annotation. This facilitate further queries, as simple as possible, to retrieve relevant information from the databases. Data from high-throughput technologies, such as DNA sequences or microarray experiments have developed structured formats to submit the data to the databases with annotations following an ontology-based vocabulary. Learning from those experiences, the Proteomic Standards Initiative (PSI) of the Human Proteome Organization (HUPO) has established a Molecular Interaction (MI) group to develop a standard format to interchange information called PSI-MI (Hermjakob *et al*., 2004).

At the time of writing this revision, there is not a common repository that stores all the ppis. Contrarily to other genomic data such as sequences, microarrays, protein structures, etc., ppi data are spread through several databases, among which a small overlapping exists. Moreover, there are strong differences in the type and depth of ppi annotations among the databases. The major repositories are the Human Protein Reference Database (HPRD, Peri *et al*., 2003), IntAct (Kerrien *et al*., 2006), the Bimolecular Interaction Network Database (BIND, Bader *et al*., 2003), the Database of Interacting Proteins (DIP, Salwinski *et al*., 2004), BioGRID (Breitkreutz *et al*., 2008) and the Molecular INTeractions database (MINT, Chatr-aryamontri *et al*., 2006); see Resources section for web addresses. Reviews on the resources dedicated to store and annotate ppis can be found in Xenarios and Eisenberg (2001) and Mathivanan *et al*. (2006).

Therefore, it is not a trivial task for the end user to obtain a reasonably complete and curated set of ppis to work with. Several methodologies have been proposed to solve this

problem; see Methods for a small revision on them.

Recently, an ambitious initiative has been proposed by FEBS Letters journal (Ceol *et al*., 2008), which consists on linking scientific manuscripts with protein interactions databases through a structured summary with controlled vocabulary that has to be filled by the authors. These kinds of approaches are going to be crucial in the quality and accessibility of biological data.

Ppis as networks (The Interactome)

Whichever the set of ppis chosen, a subset of the interactome can be defined as a compilation of pairwise relationships that, taken all together, represent a network where the nodes are the proteins and the edges the interaction events. Apart from the elements of the network (nodes and edges), the topology of the networks is also of crucial importance when trying to understand their role in a cellular process (Yeger-Lotem *et al*., 2004).

Graph theory has helped biology to study these networks and established the bases for their description. One of the first discoveries brought about by graph theory was that biological networks are scale-free networks (Barabasi & Albert, 1999; Barabasi & Bonabeu, 2003) instead of random networks. Scale-free networks are defined by a connections degree, number of connections of a node, and a distribution that approximates to a power law $P(k) = k^{-\gamma}$, being $\gamma < 3$. This indicates that the network has a low number of highly connected nodes called hubs. In other words, there are a few proteins, the hubs, which connect at long distance much of the whole network. Indeed, identifying hubs is a hot topic in functional analysis (Batada *et al*., 2006, He & Zhang, 2006, Sporns *et al*., 2007).

Apart from degree, which identifies hubs, there are other network parameters that help to describe properties of these systems (Barabasi & Oltvai, 2004). The Betweenness Centrality of a node v (C_B (v)) is a parameter that accounts for the centrality of the node V within the graph. It is obtained from the expression:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

being $\sigma_{st}(v)$, the number of shortest paths through a node and σ_{st}, the total number of shortest paths in the graph. Relative betweenness centrality (rC_B (v)) is calculated as:

$$rC_B(v) = \frac{2 * C_B(v)}{n^2 - (3n + 2)}$$

being n the total number of nodes in the graph.

A node with high betweenness centrality is a protein which has many shortest paths between any two other nodes passing through it. The action of removing that particular node from the network would cause a strong disconnection of the network. By this

description, central nodes seem to be crucial in the global compactness of the network and in the definitions of boundaries of sub-networks seen as modules of action (Girvan and Newman, 2002; Wilkinson and Huberman, 2004; Joy *et al*., 2005). Betweenness has also been shown to be bigger in modules formed by proteins associated with cancer (Hernandez *et al*., 2007).

Clustering coefficient of a node ν (C(ν)) is a measure of connectivity that evaluates how connected is the node neighborhood. This parameter helps to distinguish among the highly connected proteins that form star-shaped sub-networks (classical hub configuration) and the proteins in a more connected area, e.g. complexes. The clustering coefficient is calculated as follows:

$$C(v) = \frac{2e_n}{n_v(n_v - 1)}$$

where e_n is the number of edges among the nodes connected to node ν, and n_V is the number of neighbors of node ν.

Other interesting features in the structure of a network are the concepts of components and bicomponents. A component is a group of nodes connected among them and a bicomponent is a group of nodes connected to another group of nodes by a single edge, which is called the articulation point. When analysing network parameters of a set of proteins related to some phenotype (such as gene/protein signatures of diseases or differentially expressed genes in a two conditions comparison in microarray experiments) components and bicomponents can be considered extreme examples of gene/protein modules if they can be defined as sub-networks with a higher internal connectivity than its connectivity to other modules.

Resources to apply ppi data to Functional Genomics

Historically, in the field of ppis, the majority of the resources available have been focused mainly on visualization aspects rather than in the proper analytical steps. Most of the databases have their own visualization tools that exceptionally provide some applications to carry out very simple analyses. DIP through its satellite project LiveDIP has a tool for finding the path between two proteins. This tool also performs a simple mapping of the proteins selected in microarray experiments onto the interactome. BIND can export directly to Cytoscape (Shannon et al., 2003), a popular visualization tool, and may perform some simple analysis of enrichment in GO clusters called OntoGlyphs. HPRD does not provide visualization although it does have transcriptome information. MINT estimates the Minimal Connected Network (MCN, see Methods for complete explanation) and has a java environment available for visualization purposes. The IntAct database includes an application called MINE that can calculate and represent the MCN. BioGRID does not provide any visualization yet although the application Osprey (Breitkreutz *et al*., 2003) uses it as underlying support database and consequently can be used as an interface to BioGRID. The database STRING (von Mering, *et al*., 2007) includes a visualization tool with a

complete set of options including generating the MCN and displaying co-expression analysis of the set of proteins.

Besides the facilities provided by the databases there are programs, such as Osprey (Breitkreutz *et al*., 2003), Cytoscape (Shannon *et al*., 2003), VisANT (Hu *et al*., 2007) and PATIKA (Dogrusoz *et al*., 2006), that aim to provide a general framework for ppi data management. Cytoscape and VisANT allows the development of plug-ins that can be integrated into them to perform more specific tasks. Cytoscape is probably the most successful application in this field and it has an ample community of users and developers. At the time of writing this revision there were 48 plug-ins available. A good review about visualization and network management packages can be found in Suderman *et al*. (2007).

Other applications like the Agile Protein Interaction DataAnalyzer (APID) (Prieto *et al*., 2006), Genes2Networks (Berger *et al*., 2007) and PIANA (Aragues *et al*., 2006) were developed with the aim of become a common repository for different ppi datasets. APID and Genes2Networks are web-based tools that make the datasets available. PIANA is more orientated to computer scientists as a working framework for ppi data management. It also can predict novel interactions and calculate some topological parameters.

In a more general functional profiling context, ppi data has quite recently been introduced into suites of programs like Babelomics (Al-Shahrour *et al*., 2006, 2007, 2008) and DAVID (Dennis *et al*., 2003) although with different grade of sophistication. DAVID simply reports the interactions associated to the genes of a list and performs a classical enrichment analysis for each of the interactions. Babelomics has included a new module called SNOW (Studying Networks in the Omic World) that calculates the MCN and evaluates the significance of its robustness as functional class comparing its topological parameters versus distributions of same sized lists of random genes or proteins (Al-Shahrour et al., 2008). It also evaluates the presence of hubs, central nodes and highly connected areas in the pre-selected genes or proteins versus a curated interactome. The methodologies that SNOW applies are explained in more detail in the Methods section.

Methods

Interactome generation methodologies

As discussed in the previous section, obtaining a curated set of ppis as complete as possible to work with is not a trivial task. The databases' coverage, the depth and type of annotation and the lack of accuracy of some of the techniques constitute a limitation for this type of analysis. Nevertheless, Reguly *et al*. (2006) established a milestone in this field by generating what probably is the complete yeast interactome, free from false positives, via manual curation. The interactomes of the rest of species are far from this level of completion, basically due to their comparative bigger size. Even small differences in genome sizes can account for drastic increases in the number of ppis, suggesting that the final cause of organism's complexity must be a post-transcriptional event. For instance, human has 21,541 protein coding genes and 650,000 predicted interactions while *Caenorhabditis elegans* has 20,140 genes and three times less predicted interactions (gene

counts taken from ensembl genome browser release 49 and interactome size predictions taken from Stumpf *et al*., 2007).

There is a clear necessity of methodologies to filter ppis given that manual curation is not always a feasible task. Therefore, several approaches have been proposed (review on Badet *et al*., 2004). We will point out some of them:

1. Promiscuity (Uetz *et al*., 2000; Ito *et al*., 2001; Gavin *et al*., 2002; Ho *et al*., 2002), that consists in removing proteins having many interaction partners (known as sticky proteins asd most of them presenting a binding without any biological meaning).

2. Topological criteria (Bader & Hogue, 2002), it is specific to Co-IP experiments, retaining the bait-hit (spoke) rather than the bait-hit and hit-hit (matrix) interactions.

3. Intersection of multiple high-throughput datasets (von Mering *et al*., 2002; Deane *et al*., 2002).

4. Selecting ppis detected with two different techniques (von Mering *et al*., 2002), based on the idea that the combination of two methods will increase coverage and accuracy.

5. Intersection with other type of data, e.g. interacting proteins whose transcripts co-express are more likely to be real (Ge *et al*., 2001; Deane *et al*., 2002; Jansen *et al*., 2002) or inferences of ppis across species due to protein homology (Deane *et al*., 2002).

6. Logistic regression approach (Bader *et al*., 2004), that uses statistical and topological descriptors to predict the biological relevance of protein-protein interactions obtained from high-throughput screens.

Our experience in compiling data to build an accurate set of human ppis showed us that the annotation in the different databases sometimes is not comparable. The approach proposed here is a modification of the point 4. To build a filtered interactome we have taken the six top categories of experimental methods described in the Molecular Interaction (MI) Ontology (Hermjakob *et al*., 2004) plus the categories *in vivo* and *in vitro* from HPRD as reference. HPRD seems to be essential when approaching a human interactome (Mathivanan *et al*., 2006). Every ppi in each of the datasets was annotated with these categories. Ppis verified by at least two of these methods were introduced in the filtered interactome. By using lower levels of depth in the ontology of "techniques annotation" we ensure that ppis extracted with experiments with similar basics (that may have same biases in the detection process) are not selected.

Network features evaluation

A classical experiment in Functional Genomics ends up with the assessment of a functional interpretation of the results of a genome-scale experiment such as a microarray analysis.

Typically, these results are lists of genes or proteins potentially relevant to the case of study because they have a common behaviour in their expression (for instance, they are over or under expressed in a disease or they have a similar pattern of expression through time when a disease is treated with a drug, etc.).

Conventional methods for functional profiling use functional definitions provided by well known repositories (such as GO, KEGG pathways, etc.) and study the enrichment on these modules among the selected list of genes. Ppi data can also be used under a similar philosophy. An interesting analysis than can be conducted on a list of genes or proteins is to check whether it is enriched in any kind of nodes. That is, to check whether it has significantly more hubs, central proteins or proteins in a very connected area compared to the complete interactome. This can be done comparing the distribution of the connections degree, betweenness centrality and clustering coefficients, respectively, versus the distribution of these parameters in the set of ppis used as background. Quite useful information about the set of proteins can be obtained in this way. For example, it has recently been reported that cancer-related proteins exhibit a higher degree of connections and centrality than the nodes in complete interactome not associated to the disease (Johsson *et al*., 2006; Hernandez *et al*., 2007). Nevertheless, the final aim in this kind of experiments is to seek for modules of proteins with a cooperative activity. Therefore, a more holistic approach is needed.

Methodologies to infer a sub-network

The introduction of ppi data into a functional profiling framework requires from different analytical methods due to the nature of this kind of annotation. The way in which ppis are defined do not conform discrete classes (as GOs, KEGGs, etc.) but they have an internal structure in the form of networks where proteins have different roles according to their position in the network and its global shape.

As previously mentioned in the introduction, a module in a network is a sub-network with an internal connectivity higher than its connectivity to other modules. Many attempts have been made to explore the interactome seeking for modules of action, most of them based on the application of clustering methods to weighted matrices. Pereira-Leal *et al*. (2004) proposed the number of experiments that support a given ppi as the index of the general ppi matrix. Rives and Galitski (2003) used the shortest paths among pairs of nodes to measure de relationship between nodes. There are also other approaches based for instance on topological features of the network such as the betweenness (Girvan and Newman, 2002; Wilkinson and Huberman, 2004). Central nodes, with a high betweenness, may define the boundaries of the sub-networks because this implies that many shortest paths pass through them and the action of removing them from the network would lead to the disconnection of some sub-networks.

Modules obtained by these methodologies may be enriched in proteins with related biological functionalities, shown by its significant enrichment in GO terms (Luo *et al*., 2006) or by its co-occurrence within the literature (Wilkinson and Huberman, 2004). Indeed, it has also been shown that there are sub-networks associated to diseases (Badano and Katsanis, 2002; Brunner and van Driel, 2004; Gandhi *et al*., 2006). Gandhi *et al*. (2006)

found in the analysis of the human interactome that proteins encoded by genes mutated in inherited genetic disorders are likely to interact with proteins known to cause similar disorders.

When seeking for modules of action (sub-networks) within lists of genes or proteins, these have not to be considered as a mere collection of nodes but rather as a potential unit of functionality in cell activity, in a similar way than Gene Ontology terms or KEGG pathways are defined. It is not enough obtaining only the interactions associated to each of node and explore the function of the interacting proteins. Thereby, we need to test for specific sub-networks within the set of proteins analysed in order to assign a common putative functionality to the list. In other words, it is essential for the biological interpretation of a gene or protein list to seek for the sub-network that they might configure, or in other words, the module that has been activated.

A common approach to envisage this sub-network is to calculate the called Minimal Connected Network (MCN). The MCN is the minimal network that connects a set of nodes. It is generated by the calculation of the shortest path between any two proteins in the list. When generating the MCN for functional profiling of experiments, the resulting network should be representative of the list of proteins analysed. Therefore not all the paths should be integrated in the final graph but only the ones that connect directly any pair of proteins in the list. Such graphs can be completed using other proteins not contained in the list analysed that might have been missing (a typical problem in proteomics experiments) and could contribute with important connections to the resulting network. The number of proteins outside the list that connects pairs of proteins within the list should be small enough to keep equilibrium between the exploratory capabilities of this methodology and the maintenance of the accuracy of the assignment of a network to the list.

Indeed, proteins no pre-selected by expression profiling has been reported to be related to disease due to its inclusion into a network of ppis (Xu and Li, 2006; Liu et al., 2007; Chuang *et al*., 2007). In microarray analysis it is a common practice to apply a threshold in the p-value (typically 0.05 considering multiple testing adjustment) to the selection of differentially expressed genes. This could be an additional problem in the selection of important genes because strong corrections are applied to the p-values, due to the multiple testing nature of the analysis, and some important genes might not be reported as differentially expressed when they actually are. Moreover, there are observations that point out that important proteins in the networks such as hubs and superhubs may not be differentially expressed (Camargo & Azuaje, 2007).

Methodologies to evaluate a sub-network

When trying to assign a functional interpretation to a list of genes using classical annotations such as GO terms or KEGG pathways a simple inventory of the annotations found does not give significant information. The frequencies of the annotations found have to be compared to the background to test whether the annotations show a significant enrichment or not. There are several methodologies and resources available for functional profiling (review in Dopazo, 2006). The application of ppi data to this field is quite recent so there are not standard methodologies to be applied to the evaluation of the modules

found yet.

A standard approach proposed has been to test if the proteins in a given network are enriched in any functional category (Wilkinson and Huberman, 2004; Luo *et al*., 2006). There are specific tools for doing this task. BINGO (Maere *et al*., 2005) is a java applet that can be integrated into Cytoscape visualization tool (Shannon *et al*., 2003) that performs GO enrichment analysis to the nodes of a network. Although informative, this test does not guarantee in the case of negative results that the network is not a module of action yet unannotated or simply a module of proteins that do not share a functional category but they are indeed carrying out some cooperative function. In fact, functional analysis using ppis do not always overlap with functional label based analysis (Liu *et al*., 2007).

Liu *et al*. (2007) proposed a systems biology oriented approach called Gene Network Enrichment Analysis (GNEA). The method evaluates the association of sub-networks to a determined disease. A limitation of this method is that it must start with a set of pre-defined gene signatures already associated to the disease, each of them with a particular annotation. The gene signatures are assembled and then, the relative expression in a microarray analysis, exploring the case of study, is mapped to a global network of ppis. From the interactomic and transcriptomic information a High Scoring Matrix (HSM) is extracted as a sub-network that is highly transcriptionally affected in the disease. Finally, they evaluate the hypothesis that a particular gene signature is enriched into the sub-network. Basically, ppis in this methodology substitutes the classical differentially expression analysis but it is not taking advantage of the structured data of the biological networks.

For the evaluation of a sub-network we propose to take into account the special topological features of the biological networks. Our hypothesis is that sub-networks associated to a specific cellular activity should have a compact topology. This type of topology can be characterised by having a higher distribution of connections degree and significantly less components than a sub-network integrated by random proteins that do not share any functionality. The distribution of two other parameters, clustering coefficient and betweenness centrality, that are more related to the special features of each functional type of active network, can also be evaluated. Thus, finding statistical significance in the different parameters points towards different possible topologies of the network and the combined study of some of them may reflect the actual shape of the network. In this way, obtaining a significantly higher connection degree but a non significant clustering coefficient would point to a star-shaped network. A significantly low number of components with a significantly high connections degree would reflect a compact network. Only a significant number of connections degrees might point to the presence of a group of small protein complexes. Finally, significance for betweenness centrality but not for connections degree could indicate the presence of a cascade signalling network.

This methodology uses as input a list of proteins selected from a genome scale experiment (e.g. proteins that co-express or that are differentially expressed under certain conditions). The aim is to find the active networks inside these lists and to evaluate if they are important in the cooperative behaviour of the list. In other words, the methodology proposed intends to highlight the sub-networks activated in response to new stimuli and to evaluate their importance within the entity selected as the unit of study, which is the list of proteins.

We first calculate the MCN of the pre-selected proteins or genes and then we test, using the Kolmogorov-Smirnov test, each of the distributions of the MCN node parameters (connections degree, betweenness centrality and clustering coefficient) against their corresponding reference distribution, which we take from generating MCNs of lists of the same size as the query list, populated with random proteins. The number of components of the MCN is compared to a 95% confidence interval generated from the random datasets.

This novel methodology is implemented and available in the SNOW module of Babelomics suite for functional profiling of genome-scale experiments.

Connectivity of conventional functional modules

We performed a massive analysis of lists of human genes and proteins taken from microarray experiments, co-expression modules in cancer, GO terms, KEGG pathways and Biocarta classes with the aim of studying how ppi networks are spread in different types of lists and in some classical sources of annotation normally used for functional profiling. Table 1 shows the percentages of the lists in each category that presented a positive result in one of the analysis performed. The procedure was to take every list and calculate its MCN allowing the inclusion of one non-listed node into the network. We used a curated human interactome as described in the Methods section. The distribution of the connections degree (degree), betweenness centrality (betweenness) and clustering coefficient of the nodes in each of the MCNs generated was compared using a Kolmogorov-Smirnov test versus the distribution of the same parameter in a set of 10,000 MCNs generated from a same size range set of lists populated with random proteins/genes.

	Gene Ontology	Modules	Cancer Lists	Non-Cancer Lists	Up-regulated Lists	Down-Regulated Lists	Biocarta	KEGGs
bdcG	15.11	10.5	6.67	1.9	1.91	5.6	9.27	26.9
bG	36.53	34.9	20.44	19.85	21.02	20	32.59	48.97
dG	71.52	52.1	29.33	31.23	30.57	30	55.91	59.31
cG	22.21	13.4	7.56	2.91	2.87	6.4	12.46	31.03
compL	51.92	38	18.22	18.4	15.61	19.2	33.87	52.41
compL + dG	47.72	34.9	14.67	14.29	13.38	14	32.91	48.28

Table1. Network parameters evaluation in different types of sets of genes. Lists of genes taken from differential expression analysis of microarray experiments (cancer lists, non-cancer lists, up-regulated lists and down-regulated lists), modules of co-expression in cancer also taken from microarray experiments and genes belonging to the same annotation class (GO terms, KEGG pathways and Biocarta pathways). In rows we show the network topological parameters evaluated, bdcG (betweenness, degree and clustering coefficient have a significantly higher value than random sets), bG (betweenness significantly higher than random), dG (degree significantly higher than random), cG (clustering coefficient significantly higher than random), compL (number of components below the 95% confidence interval of random sets) and compL + dG (number of components below the 95% confidence interval of random sets plus degree significantly higher than random) . The value in the cells is the percentage of the set of lists with a p-value less than 0.05 compared to networks generated using same size random lists.

This analysis reveals the global role of networks across biologically meaningful lists. The results show that generally the ppi networks seem to have a wide distribution in classes formed by functionally related proteins. We take a significant result in the compL + dG analysis as indicative of a compact network, where the nodes have more connections and fewer components than a network of random genes, hypothetically not functionally related. Betweenness and clustering coefficient seem to be parameters more related to the shape of the network that may point out its activity, e.g. a signalling cascade network is a network with a low clustering coefficient distribution because its nodes are not in very connected areas. Nevertheless, it should indeed have more connections and fewer components than a network coming from a random list.

Summarising, the results of the analysis indicate that networks are more widely spread in the annotation classes (GO, KEGG, BioCarta and modules of co-expression) than in the up- and down-regulated lists of genes (cancer and non-cancer lists are contained into this categories as well). GO terms and KEGG pathways appear to be the more connected entities. The classification of differentially expressed lists into 4 categories do not show differences among them but in the clustering coefficient comparison where cancer and down-regulated lists show a higher percentage indicating that their networks must have a higher interconnectivity.

Conclusions and future trends

In conclusion, we could say that ppi networks do have an important role in functionally related genes. Even in genes detected to be differentially expressed in an experiment that may not be involved in a single activity but in more than one, we could detect modules of action using ppi data. Therefore the applicability of interactomics as a source of annotation in functional profiling is more than justifiable.

Nevertheless, there are still three main challenges that have to be approached to be able to obtain whole capabilities from this kind of data:

1. High-throughput techniques produce a high proportion of false positives. Besides, there is still a low coverage of the interactome for the majority of the species. For filtering ppis according to their accuracy, literature curation does not seem to be a realistic approach, so new methodologies have to be proposed while the techniques do not overcome this limitation.

2. There is a clear necessity of applying the standard annotation developed by the HUPO for the ppi experiments. This should be enough to encourage the community to take a policy of sharing data to be able to have one or several repositories with all the available ppis.

3. The network nature of the functional classes that the ppis form requires more complex methodologies to study modules enrichment. The topology of the networks should be taken into account as an important parameter of the module. A protein in a network cannot be annotated just as part of the network but as a node with a special

position that affects the rest of nodes. Moreover, the global shape of the network is characteristic of its functional activity.

The introduction of more structured data as the network concept in functional analysis is more and more required. Analysis of regulatory (Yeger-Lotem *et al*., 2004), co-expression (Ghazalpou *et al*., 2006) and genetic (Kelley & Ideker, 2005) networks are some examples of the applicability of graph theory to biological data.

Concluding remarks

In this chapter we have highlighted the relevance of introducing interactomic data into Functional Genomics. Last discoveries show that cell complexity is more associated to post-transcriptional and post-translational events rather than to size of the genome. Hence, to understand cell behaviour is of crucial importance to introduce not only functional gene annotations but also information of post-transcriptional events such as protein-protein interactions.

We have given short but comprehensive review on the databases that store ppi data, methodologies to curate ppi data as well as to infer and evaluate networks within a functional genomics context. In addition, we have analysed the role of ppi networks in different gene modules build up using different functional criteria. Our observations suggests that conventional functional modules (GO, KEGG, Biocarta) seems to be fairly connected. This shows that the introduction of ppi data into Functional Genomics may provide both complementary and reinforcing results in the functional profiling of genome scale experiments.

Resources

General resources

A very complete revision on some available resources related to ppis:
http://www.imb-jena.de/jcb/ppi/jcb_ppi_databases.html

Protein Interaction Databases

Database of Interacting Proteins (DIP):
http://dip.doe-mbi.ucla.edu/
Biomolecular Interaction Network Database (BIND):
http://bond.unleashedinformatics.com/
Human Protein Reference Database (HPRD):
http://www.hprd.org/
Molecular INTeractions database (MINT):
http://mint.bio.uniroma2.it/mint/Welcome.do
Database of protein InterAction data (IntAct):
http://www.ebi.ac.uk/intact/
General Repository for Interaction Datasets (BioGRID):

http://www.thebiogrid.org/
Literature Curated (LC) Saccharomyces cerevisiae ppis database (Reguly *et al*., 2006):
http://www.thebiogrid.org/ & http://yeastgenome.org
STRING
http://string.embl.de/

Visualization tools

Agile Protein Interaction DataAnalyzer (APID):
http://bioinfow.dep.usal.es/apid/index.htm
APID2NET (APID applet for Cytoscape)
http://bioinfow.dep.usal.es/apid/apid2net.html
CYTOSCAPE
http://www.cytoscape.org
Integrative Visual Analysis Tool for Biological Networks and Pathways (VisAnt)
http://visant.bu.edu/
Osprey
http://biodata.mshri.on.ca/osprey/servlet/Index
STRING
http://string.embl.de/

Tools for inferring (and evaluate) networks

MINE (IntAct interface)
http://www.ebi.ac.uk/intact/mine/do/welcome
MINT
http://mint.bio.uniroma2.it/mint/search/search.do
MEDUSA (interface to STRING database)
http://coot.embl.de/medusa/
Genes2Networks
http://actin.pharm.mssm.edu/genes2networks/
PIANA
http://sbi.imim.es/piana/
Gene Network Enrichment Analysis (GNEA), R package
http://genomics10.bu.edu/manwayl/
BINGO (Cytoscape plug-in)
http://www.psb.ugent.be/cbd/papers/BiNGO/
Babelomics (SNOW module)
http://www.babelomics.org

Packages for Functional Genomics analysis

Babelomics
http://www.babelomics.org
DAVID
http://david.abcc.ncifcrf.gov/
OntoTools

http://www.bioconductor.org/packages/2.0/bioc/html/ontoTools.html

Acknowledgements

This work is supported by grants from projects BIO 2005-01078 from the Spanish Ministry of Education and Science and National Institute of Bioinformatics (www.inab.org), a platform of Genoma España. The CIBER de Enfermedades Raras (CIBERER) is an initiative of the ISCIII.

References

Al-Shahrour, F., Díaz-Uriarte, R. & Dopazo, J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20: 578-580

Al-Shahrour *et al*. (2006) BABELOMICS: a systems biology perspectve in the functional annotation of genome-scale experiments. Nucleic Acids Research, 34, Web Server Issue, W472-W476

Al-Shahrour, F., Minguez, P., Tárraga, J., Medina, I., Alloza, E., Montaner, D., & Dopazo, J. (2007). FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interpretation data with microarray experiments. Nucleic Acids Research 35 (Web Server issue): W91-96

Al-Shahrour, F., Carbonell, J., Minguez, P., Goetz, S., Conesa, A., Tárraga, J., Medina, I., Alloza, E., Montaner, D. & Dopazo, J. (2008) Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments. Nucleic Acid Research. 36:341-346

Aragues R., Jaeggi D. and Oliva B. (2006) PIANA: Protein Interactions and Network Analysis. Bioinformatics, 22(8):1015-7

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 25, 25-29.

Badano,J.L. and Katsanis,N. (2002) Beyond Mendel: an evolving view of human genetic disease transmission. Nat. Rev. Genet., 3, 779–789.

Bader G.D. & Hogue C.W.V. (2002) Analyzing yeast protein−protein interaction data obtained from different sources. Nature Biotechnology **20**, 991 - 997

Bader G.D., Betel D. and Hogue C.W.V. (2003) BIND: The Biomolecular Interaction Network Database. Nucleic Acids Research, 31, 1

Bader J.S. et al. (2004) Gaining confidence in high-throughput protein interaction networks. Nature Biotechnology, 22, 1, 78-84

Barabasi, AL., Albert R. (1999) Emergence of scaling in random networks. *Science* 286(5439):509-12.

Barabasi, AL., Bonabeu, E. (2003) Scale-Free Networks. *Scientific American* 288(5):60-9

Barabasi, AL., Oltvai, ZN. (2004) Network Biology: Understanding the cell's functional organization. *Nature Reviews* 5(2):101-13

Batada, N.N., Hurst, L.D., Tyers, M. (2006) Evolutionary and Physiological Importance of Hub Proteins. *Plos Computational Biology*, 2, 7, e88.

Berger S.I., Posner J.M., Ma'ayan A. (2007) Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. BMC Bioinformatics, 8:372

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

Breitkreutz B., Stark C. and Tyers M. (2003) Osprey: a network visualization system. Genome Biology, 4:R22

Breitkreutz B. et al. (2008) The BioGRID Interaction Database: 2008 update. Nucleic Acids Research, 36, Database issue, D637-D640

Brunner,H.G. and van Driel,M.A. (2004) From syndrome families to functional genomics. Nat. Rev. Genet., 5, 545–551.

Camargo A. & Azuaje F. (2007) Linking Gene Expression and Functional Network Data in Human Heart Failure. PloS One, 12, e1347

Ceol A., Chatr-Aryamontri A., Liacata L., Cesareni G. (2008) Linking entries in protein interaction database to structured text: The FEBS Letters experiment. FEBS Lettwers, 585, 1171-1177

Chatr-aryamontri A., Ceol A., Palazzi L.M., Nardelli G., Schneider M.V., Castagnoli L., Cesareni G. (2006) MINT: the Molecular INTeraction database Nucleic Acids Research

Cho S. et al. (2003) Protein-protein Interaction Networks: from Interactions to Networks. Journal of Biochemistry and Molecular Biology, 37, 1, 45-52.

Chuang H., Lee E., Liu Y., Lee, D. and Ideker T. (2007) Network-based classification of breast cancer metastasis. Molecular Systems Biology, 3, 140

Deane G.D. *et al*. (2002) Protein Interactions: two methods for assessment of the realiability of high throughput observations. Mol. Cel. Proteomics, 1, 349-356.

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID:

Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 2003;4(5):P3.

Dopazo, J. (2006) Functional interpretation of microarray experiments. Omics, 10, 398-410.

Drewes G. and Bouwmeester T. (2003) Global approaches to protein-protein interactions. Current Opinion in Cell Biology. 15:199-205

Falk R. *et al*. (2007) Approaches for systematic proteome exploration. Biomolecular Engineering 24, 155-168

Drogrusoz U. *et al*. (2006) PATIKAweb: a Web interface for analyzing biological pathways through advanced querying and visualization. Bioinformatics, 22, 3, 374-375

Formstecher E. *et al*. (2005) Protein interaction mapping: a Drosophila case study. Genome Research, 15, 376-384

Gandhi,T.K. *et al*. (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat. Genet., 38, 285–293.

Gavin A.C. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 415, 141-147

Ge H. *et al*. (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nature Genetics, 29, 482-486

Ghazalpour A. et al. (2006) Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight. Plos Genetics, 2, 8, e130

Giot L. *et al*. (2003) A protein interaction map of Drosophila melagaster. Science, 302, 1727-1736

Girvan M. and Newman M.E.J. (2002) Community structure in socila and biological networks. Proc. Natl. Acad. Sci., 99, 12, 7821-7826

Han J-DJ. *et al*. (2005) ffect of sampling on topology predictions of protein-protein interaction networks. Nature Biotechnology, 23, 839-844.

Hartwell, H., Hopfield, J., Leibler, S. and Murray, AW (1999) From molecular to modular biology. *Nature* 402, C47-C52.

He, X. and Zhang, J. (2006) Why do hubs tend to be essential in protein networks? *Plos Genetics*, 2, 6, e88.

Hermjakob H. *et al*. (2004) The HUPO PSI's Molecular Interaction format-a community standard for the representation of protein interaction data. Nature Biotechnology, 22, 2,

177-183

Hernandez P. *et al*. (2007) Evidence for system-level molecular mechanisms of tumorgenesis. BMC Genomics, 8:186

Hernandez-Toro J., Prieto C. and De Las Rivas J. (2007). APID2NET: unified interactome graphic analyzer. Bioinformatics 23(18): 2495-2497

Ho Y. *et al*. (2002) Systematic identifiaction of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature, 415, 180-183

Hu Z. *et al*. (2007) VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. Nucleic Acids Research, 1-8

Ideker T. *et al*. (2001) Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. Science, 292, 929-933

Ito T. *et al*. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. 98, 4569-4574.

Jansen R. *et al*. (2002) relating whole-genome expression data with protein-protein inetractions. Genome Research, 12, 37-46

Johsson P.F. and Bates P.A. (2006) Global topological features of cancer proteins in the human interactome. Bioinformatics, 22(18), 2291-2297.

Joy M.P., Brock A. Ingber D.E. and Huang S. (2005) High-Betweenness Porteins in the Yeast Protein Interaction Network. Journal of Biomedicine and Biotechnology, 2, 96-103

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res, 32, D277-280.

Kerrien S. *et al*. (2006) IntAct – Open Source Resource for Molecular Interaction Data. Nucleic Acids Research 2006

Kelley R. and Ideker T. (2005) Systematic interpretation of genetic interactions using protein networks. Nature Biotechnology, 23, 5, 561-566

Lee *et al*. (2007) Asingle gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. Nature Genetics, 40, 2, 181-188

Li S. (2004) A map of the interactome network of the metazoan C. elegans. Science, 303, 540-543

Liu, M., Liberzon, A., Kong, S.W., Lai, W.R., Park, P.J., Kohane, I.S., Kasif, S. (2007) Network-Based Analysis of Affected Biological Proceses in Type 2 Diabetes Models. *Plos Genetics*, 3, 6, e96.

Luo F. *et al*. (2006) Modular organization of protein interaction networks. Bioinformatics, 23, 2, 207-214

Maere S., Heymans K., Kuiper M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics, 21(16), 3448-9

Mathivanan S. *et al*. (2006) An evaluation of human protein-protein inetraction data in the public domain. BMC Bioinformatics, 7, 5, S19

Mrowka, R., Patzak, A. & Herzel, H. (2001) Is there a bias in proteome research? Genome Research, 11, 1971-1973

Pereira-Leal J.B. *et al*. (2004) Detection of functional modules from protein-protein interaction networks. Proteins: Struct. Func. Bioinformatics, 54, 49-57.

Peri, S. *et al*. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Research, 13:2363-2371.

Prieto C. and De Las Rivas J. (2006) APID: Agile Protein Interaction DataAnalyzer. *Nucl. Acids Res.* 34: W298-W302

Reguly T. *et al*. (2006) Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. J. Biol 5:11

Rives A.W. and Galitski T. (2003) Modular organization of cellular networks. Proc. Natl. Acad. Sci., 100, 3, 1128-1133

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 20;437(7062):1173-8

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. (2004) The Database of Interacting Proteins: 2004 update. NAR 32 Database issue D449-51

Shannon P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research, 13(11), 2498-504

Sporn, Olaf, Honey, C.J., Kötter, R. (2007) Identification and classification of hubs in brain networks. *Plos One* 10, e1049.

Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M,

Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksöz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE. (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 2005 122(6):957-68

Stumpf M.P.H. *et al*. (2008) Estimating the size of the human interactome. Proc. Natl. Acad. Sci., 105, 19, 6959-6964

Suderman M. and Hallet M. (2007) Tools for Visually Exploring Biological Networks. Bioinformatics, 23(20):2651-9

Uetz, P. *et al*. (2005) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature, 403, 603-627

von Mering C. *et al*. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature, 417, 399-403

von Mering C. *et al*. (2007) STRING 7 – recent developments in the integration and prediction of protein interactions. Nucleic Acids Research, 35, Database issue, D358-D362

Wachi S., Yoneda K. and Wu R. (2005) Interactome-trancriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. Bioinformatics, 21(23), 4205-4208.

Wilkinson D.M. and Huberman B.A. (2004) A method for finding communities of related genes. Proc. Natl. Acad. Sci., 101, 1, 5241-5248

Xenarios I. and Eisenberg D. (2001) Protein interaction databases. Current Opinion in Biotechnology, 12:334-339

Xu J. and Li Y. (2006) Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics, 22, 2800-2805

Yeger-Loten, E., Sattah, S. Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R., Alon, U. Margalit, H. (2004) Networks motifs in integrated cellular networks of transcription-regulation and protein-protein interactions. *PNAS* 101, 16, 5934-5939.

Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, **4**, R28