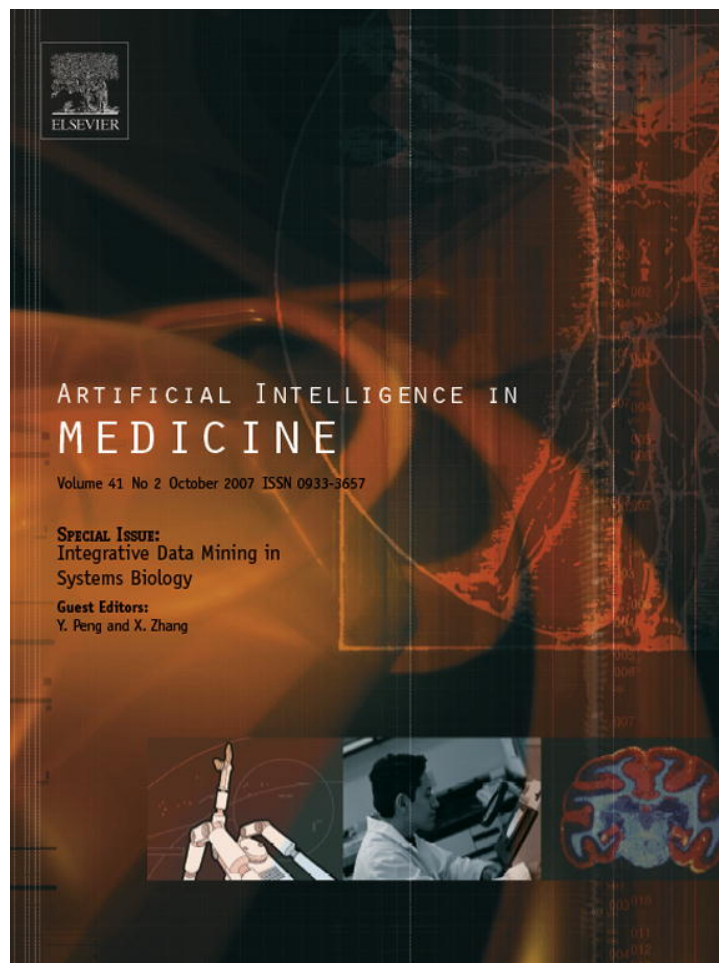


Provided for non-commercial research and education use.  
Not for reproduction, distribution or commercial use.



This article was published in an Elsevier journal. The attached copy is furnished to the author for non-commercial research and education use, including for instruction at the author's institution, sharing with colleagues and providing to institution administration.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



# A multi-layered approach to protein data integration for diabetes research

Ken McGarry<sup>a,\*</sup>, James Chambers<sup>b</sup>, Giles Oatley<sup>b</sup>

<sup>a</sup> School of Pharmacy, University of Sunderland, Wharnccliffe Street, Sunderland SR1 3SD, UK

<sup>b</sup> School of Computing and Technology, University of Sunderland, St. Peters Campus, Sunderland SR6 0DD, UK

Received 1 December 2006; received in revised form 26 July 2007; accepted 26 July 2007

## KEYWORDS

Protein interactions;  
Graph theory;  
Erdos–Renyi

## Summary

**Objective:** Recent advances in high-throughput experimental techniques have enabled many protein–protein interactions to be identified and stored in large databases. Understanding protein interactions is fundamental to the advancement of science and medical knowledge, unfortunately the scale of the requires an automated approach to analysis. We describe our graph-mining techniques to identify important structures within protein–protein interaction networks to aid in human comprehension and computerised analysis.

**Methods and materials:** We describe our techniques for characterizing graph type and associated properties which is constructed from data collated from the Human Protein Reference Database. Using random graph rewiring comparative techniques and cross-validation with other identification methods a further analysis of the identified essential proteins is presented to illustrate the accuracy of these measures. We argue for using techniques based upon graph structure for separating and encapsulating proteins based upon functionality.

**Results:** We demonstrate how rational Erdos numbers may be used as a method to identify collaborating proteins based solely upon network structure. Further, by using dynamic cut-off limit it demonstrates how collaboration subgraphs can be generated for each protein within the network, and how graph containment can be used as a means of identifying which of many possible graphs are likely to be actual protein complexes. The demonstration protein interaction network built for diabetes is found to be a scale-free, small-world graph with a power-law degree distribution of interactions on nodes. These findings are consistent with many other protein interaction networks.

Crown Copyright © 2007 Published by Elsevier B.V. All rights reserved.

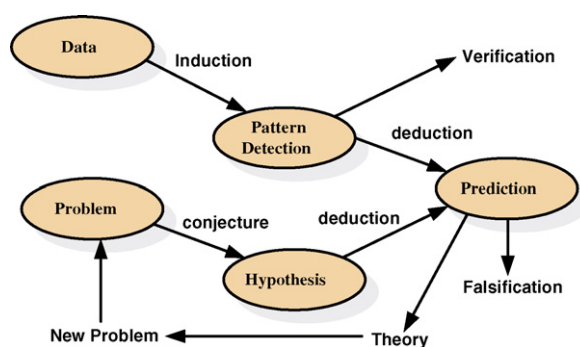
\* Corresponding author. Tel.: +44 191 5153784; fax: +44 191 5153805.  
E-mail address: [ken.mcgarry@sunderland.ac.uk](mailto:ken.mcgarry@sunderland.ac.uk) (K. McGarry).

## 1. Introduction

The objective of systems biology is to develop predictive models at the level of molecular pathways, cells, organs and ultimately at the level of the whole organism. A range of computational techniques has been used to model the high complexity of biological interactions between entities and their structures. However, it is becoming increasingly obvious that the overall complexity of biological systems cannot be understood by the analysis of the individual elements alone [1]. The difficulties of complexity encountered by the pharmaceutical industry when developing the necessary assays for drug discovery have proved this conclusively, that there is no simple or direct link from genome to drugs [2].

The complexity issues are related to the level of detail and number of interacting elements that must be modeled, which rises exponentially. Therefore, the computational tractability of any proposed algorithmic/information processing solution must be taken into account. The use of GRID empowered computing is now perceived as an essential tool to provide medium-to-high complexity simulations within a reasonable timescale. Other computational factors required for developing a successful strategy for systems biology include the database and ontological integration of the many sources disparate data [3]. Markup languages specifically designed for systems biology (SBML) and Laboratory Information Management Systems (LIMS) can help here but they are not by any means a total solution [4].

Computational modeling and simulation of biological processes allow scientists to investigate particular scenarios when given the appropriate probabilistic and stochastic methods to model the necessary genetic and kinetic biochemical pathways required by the biologists. The models can provide useful insights into the processes involved at metabolic, cellular and higher levels. In Fig. 1, the relationships and the links between hypothesis generation and testing and induction from data are



**Figure 1** The integration of hypothesis driven and data driven science [4].

shown, in the age of data driven science we can build models that combine these complementary techniques. Much of the theory and mathematics required is obtained from systems engineering which was one of the first disciplines to take a holistic approach to modeling complex processes. Recent advances in the complexity and quality of computational modeling have led to some interesting hypothesis regarding the targets for molecular therapy of diabetes mellitus, for example the model proposed by Pollard et al. uses over 210,000 molecular relationships [5]. Furthermore, the complexity of modeling is increased when characteristics such as spatial and temporal events are included in the model [6].

It is possible to classify modeling techniques as either qualitative or quantitative, this is depending on the complexity of the approach taken [7]. For example, kinetic models and metabolic flux are essentially quantitative techniques, while Boolean networks and static networks such as protein–protein interaction networks (PPI) are qualitative in nature. Furthermore, it is a vital prerequisite of biology that explanatory models are understandable, it is often the case that data driven experiments can produce counter-intuitive results [8].

We focus primarily on PPI and the prediction of new interactions [9]. Although, such datasets contain errors, recent work has highlighted techniques for assessing their reliability [10–12]. Several bioinformatic models have been proposed that account for some of the characteristics of PPI networks [13]. Scientists use biological ontologies for several tasks, the most notable has been the recent use of the gene ontology for the annotation of new gene products based on similarities [14]. Other, equally important functions are heterogeneous data integration and cross-database querying [15]. Furthermore, semantic specification is an important factor to consider in ontology development [16].

The remainder of this paper is structured as follows; Section 2 gives a detailed treatment of our data sources and an introduction to the factors involved with insulin resistance, Section 3 describes our graph-based mining technique, Section 4 discusses the results and the biological implications, Section 5 mentions some related work and finally Section 6 presents the conclusions.

## 2. Problem domain and protein data sources

Our own particular research area is that of diabetes, in particular the effects of insulin resistance on protein expression and insulin regulated protein

trafficking in fat cells. In recent years there has been a dramatic worldwide increase of those suffering with diabetes [17]. In the year 2000, there were 171 million cases and by 2030 it is predicted there will be 366 million people suffering from this condition ([www.who.int/diabetes/facts](http://www.who.int/diabetes/facts)) (accessed December 2006). This data is for diagnosed cases but the undiagnosed cases are estimated by the World Health Organisation (WHO) at 14.6 million alone for the US.

## 2.1. Diabetes and insulin resistance

Diabetes mellitus is a metabolic disease characterized by persistent hyperglycemia (high blood sugar levels) requiring medical diagnosis, treatment and lifestyle changes. There are three main forms of diabetes: Type 1, Type 2 and gestational diabetes (or Type 3, occurring during pregnancy). Since the first therapeutic use of insulin in 1921 diabetes has been a treatable but chronic condition, its main health risks being long-term complications.

Insulin is the principal hormone that regulates uptake of glucose into cells from the blood (primarily muscle and fat cells), deficiency of insulin or the insensitivity of its receptors in cells plays a central role in all forms of diabetes. Insulin is released into the blood by  $\beta$  cells in the pancreas in response to rising levels of blood glucose. Insulin enables cells to absorb glucose from the blood for use as fuel, for conversion to other needed molecules, or for storage. Insulin is also the principal control hormone responsible for conversion of glucose to glycogen for internal storage in liver and muscle cells. When glucose levels fall, reduced insulin levels result both in the reduced release of insulin from the  $\beta$  cells and in the conversion of glycogen back into glucose in the liver [18].

Type 2 diabetes mellitus is due to a combination of defective insulin secretion and defective responsiveness to insulin. In the early stages, hyperglycemia can be reversed by a variety of measures and medications that improve insulin sensitivity or reduce glucose production by the liver, but as the disease progresses the impairment of insulin secretion worsens and therapeutic replacement of insulin often becomes necessary. Type 2 diabetes comprises of 90% or more of cases of diabetes in the developed world. There is a strong, but not exclusive, association with obesity, aging and with family history, although in the last decade it has increasingly begun to affect children and adolescents. In Type 2 diabetes insulin levels are initially normal or even elevated, but peripheral tissues lose responsiveness to insulin (becoming *insulin resistant*). The cause of this is almost certainly involving the insulin receptor

in cell membranes, the interplay between the genes and the proteins they produce is complex and improperly understood [19]. Unfortunately, insulin resistance increases the risk factors for a number of chronic disorders such as heart disease, kidney failure, nerve damage, failing eyesight and peripheral vascular problems [20].

Although the exact causes of Type 2 diabetes are unknown, it is thought arise due to defects both in the  $\beta$  cells (reducing insulin production) and in the insulin's ability to stimulate the uptake of glucose in tissues (producing insulin resistance) notable proteins involved are GLUT4 and the IRS family of proteins [21]. A causal link with central obesity (fat concentrated around the waist in relation to abdominal organs) is known to predispose insulin resistance, possibly due to its secretion of a group of hormones called adipokines that impair glucose tolerance. By modeling the PPI network of insulin as a mathematical graph it may be possible to gain some insight into potential causes for either defective insulin production and insulin resistance [22].

## 2.2. Protein-to-protein interaction data

The information used to generate the hypothesis approach is derived from the body of available knowledge from ontological sources of protein-to-protein interactions (PPI), and raw experimental data from proteomic/genomic repositories. The process of hypothesis formulation and testing is based upon objective driven approaches but also relies on data driven methods. In fact, hypothesis driven and inductive driven reasoning techniques should be viewed as complementary, rather than competitive approaches. However, the major challenge is to identify how and where they should be applied in any given situation, the principled approach outlined in this paper addresses this problem.

There are numerous public databases of PPI data available for use, e.g. MIPS [23], BIND [24], DIP [25], BioGRID [26], some are described in Table 1. However, we focus on information drawn from the Human Protein Reference Database (HPRD). All the information in HPRD has been manually extracted from literature by expert biologists who read, interpret and analyze the published data [27]. However, part of the annotation problems can be alleviated by the authors themselves by making the submission of their work directly to the PPI databases and thus keep the data up to date, encourage standardization of data formats and remove a source of error [28]. Although XML based data systems enable cross-platform integration and communication they are not as straightforward as flat text

**Table 1** PPI datasets and resources (sites accessed December 2006)

Data base	Objectives and URL address
BIND	Biomolecular Interaction Network Database is hand curated set of small experiments but some taken from the literature of approximately 8500 interactions ( <a href="http://bind.ca">http://bind.ca</a> )
BioGRID	General Repository for Interaction Datasets, contains 116,000 interactions from <i>Saccharomyces cerevisiae</i> , <i>Caenorhabditis elegans</i> , <i>Drosophila melanogaster</i> and <i>Homo sapiens</i> ( <a href="http://www.thebiogrid.org/">http://www.thebiogrid.org/</a> )
DIP	Database of Interaction Proteins, generally from yeast. Location of gene products ( <a href="http://www.dip.doe-mbi.ucla.edu">http://www.dip.doe-mbi.ucla.edu</a> )
HPRD	Human Protein Reference Database contains approximately 20,000 proteins and 30,000 interactions ( <a href="http://www.hprd.org">http://www.hprd.org</a> )
MIPS	Mammalian Protein-Protein Interaction Database consists of manually curated high-quality, small-scale experiments and also from yeast ( <a href="http://mips.gsf.de">http://mips.gsf.de</a> )

files to manage. The root element of a PSI MI XML file is the *entrySet*, as shown in Fig. 2. An *entrySet* contains one or more entries, each a self-contained unit describing one or more PPI.

Though important, the administrative field's source, *availabilityList* and *experimentList* are unused in the diabetes system, for completeness a short description of them:

The source element describes the source of the entry, usually the organisation which provides it. It also contains a release number and a release date.

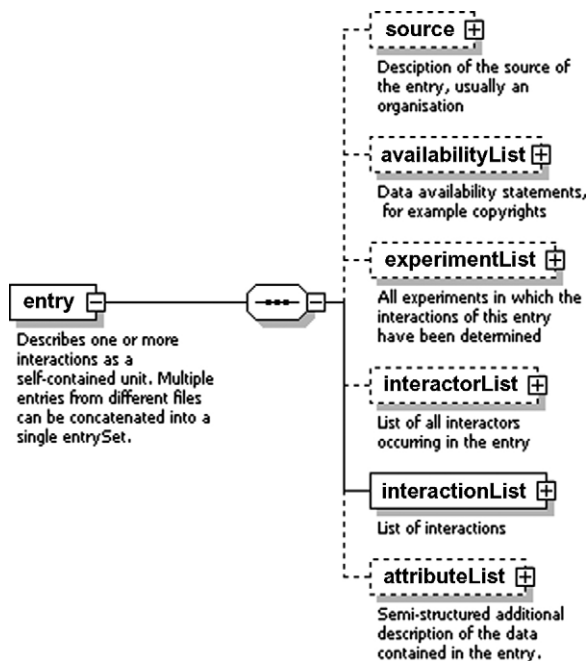
- The *availabilityList* provides statements on the availability of the data, usually copyright statements.
- The *experimentList* contains *experimentDescriptions*. Each *experimentDescription* describes one

set of experimental parameters, usually associated with a single publication.

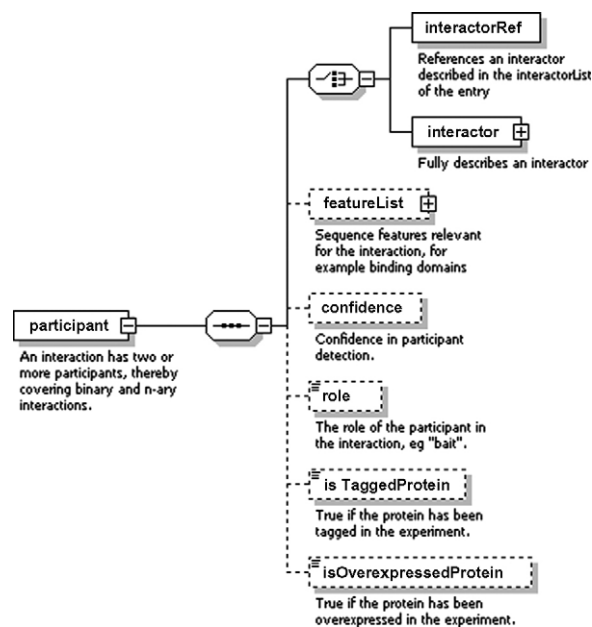
- The *interactorList* describes a set of interactors participating in an interaction. In the current version of the PSI MI all interactors are proteins.

Fields which are used in the diabetes system include:

- The interactor element describes the "normal" form of a protein, consisting of data like protein name and cross-references, and organism and amino acid sequence.
- The *interactionList* contains one or more interaction elements (see Fig. 3 for further details of this structure). Each interaction contains a description of the data availability and a descrip-



**Figure 2** The entry top level element, from proteomics standards initiative (2002).



**Figure 3** Participant element, taken from proteomics standards initiative (2002).

tion of the experimental conditions under which it has been determined.

Every interaction contains two or more participants, which are the molecules participating in the interaction. Each participant element also contains a description of the molecule in its “normal” form, by reference to an element of the *interactorList*, or directly in an interactor element. The first entry on each row is the HPRD id (and filename) of the protein, all other entries on that row are (HPRD id's of) proteins with which it interacts with (taken from the *interactionList*). Note, that at this stage self interactions are allowed.

This textual index file was loaded in the Matlab and processed into what a binary connections matrix and a name vector (the software can be made available upon request to the corresponding author). A binary connections matrix essentially stores the same information as the index file but it is a more compact form that is more suitable for automated processing. The name vector stores the name (extracted from the HPRD:MI file) of each column and row and a value of one on the connections matrix indicates an interaction between those proteins.

All 613 protein XML files were converted into a Matlab native cell structure using functions to translate XML files. This produced an initial graph with 613 vertices and 3422 edges. Prior to performing the analysis, any isolated proteins or disconnected subgraphs containing three or fewer members within the network must be removed. Removing the isolated proteins is a trivial task, as they will have no connections to any other proteins recorded in the connections matrix, identifying disconnected subgraphs is possible because they will not have a path (i.e. a path length of infinity) to most other proteins within the network. Removing these isolated proteins will save computational time later on in the analysis as they are unlikely to have any (known or identifiable) effect upon the proteins in the main network. Once this task was completed the final PPI network is composed of 584 nodes representing proteins and 3117 edges representing interactions between proteins.

### 3. Graph based data mining

Graphical data mining techniques are increasingly used to model systems which have an inherent network structure such as transport networks, ecological webs, biochemical pathways and electronic circuits have all been found to possess *motifs* or patterns of interconnections that are of significance

[29]. The networks generated from PPI data tend to have special properties based upon pairwise interactions (links) between the components (proteins). The graphs are typically very large but important subgraphs or *motifs* can be extracted [30].

#### 3.1. Centrality measures

The first stage in the analysis is the identification of essential proteins, i.e. the “hub”, within the network. Genome-wide studies show that deletion of a hub protein is more likely to be lethal than deletion of a non-hub protein, a phenomenon known as the centrality-lethality rule [31]. The actual reason why the absence of these central, highly connected proteins causes death in their host-organism is still being disputed. One of the initial reasons put forward is that PPI networks are typically small world networks with power law degree (number of edges per vertices) distributions which can be sensitive to the removal of specific nodes [32,33].

The concept of the *shortest path* is important to centrality measures and can be defined as when two vertices  $i$  and  $j$  are connected if there exists a sequence of edges that connect  $i$  and  $j$ . The length of a path is its number of edges. The distance  $l(i, q)$  between  $i$  and  $j$  is the length of the shortest path connecting  $i$  and  $j$ . The Dijkstra algorithm was used for calculating the shortest path between two vertices.

Watts and Strogatz found that when  $p = 0$ , that is the probability of randomly reconnecting the edges of the regular coupled network regular ring-like graph structures are formed, which have high clustering coefficients and long path lengths [34]. As the value for  $p$  increases, more randomness in the structure is introduced and the resulting graphs tending to have low path lengths and high clustering coefficients. Watts and Strogatz termed these graphs *small world networks*. When the value for  $p$  reaches unity, the graph becomes identical to the standard Erdos–Renyi random graph with its small path lengths and low clustering coefficients.

Therefore, by randomly rewiring the generated PPI network for diabetes is possible to determine whether or not it is a small world network, as the randomly rewired Erdos–Renyi networks should on average produce networks with a significantly lower clustering coefficient. The results of this analysis are shown in Table 2.

**Degree centrality:** The simplest of all measures is degree centrality (DC).  $DC(i)$  is the number of edges present upon node  $i$ , i.e. the number of other proteins that protein interacts with. **Closeness centrality:** This measure is the closeness centrality (CC). The closeness centrality of protein  $i$  is the

**Table 2** Erdos numbers

	Average shortest path	Clustering coefficient
Actual diabetes PPI network	3.296	0.288
Erdos–Renyi PPI network (mean values over 100 runs)	3.410	0.078

sum of graph-theoretic distances from all other proteins in the PPI network, where the distance  $d(v_i, v_j)$  from one protein  $i$  to another  $j$  is defined as the number of links in the shortest path from one to the other.

The closeness centrality of protein  $i$  in a PPI network is given by the following expression:

$$CC(v_i) = \frac{N-1}{\sum_j d(v_i, v_j)} \quad (1)$$

**Betweenness centrality:** Is a measure of the degree of influence a protein has in facilitating communication between other protein pairs and is defined as the fraction of shortest paths going through a given node. If  $p(v_i, v_j)$  is the number of shortest paths from protein  $i$  to protein  $j$ , and  $p(v_i, v_k, v_j)$  is the number of these shortest paths that pass through protein  $k$  in the PPI network, then the BC of node  $k$  is given by:

$$BC(v_k) = \sum_i \sum_j \frac{p(v_i, v_j, v_k)}{p(v_i, v_j)}, \quad i \neq j \neq k \quad (2)$$

### 3.2. Combining the centrality measures

A simple method for combining the centrality measures is to perform three-dimensional clustering of the different values. It would be expected then that a separate cluster of points with high values should be found and that these points would indicate the hubs. It is important to remember however that a protein may be essential to the system whilst only scoring highly on one of the centrality measures. For this reason we use a cut-off measure, if the centrality value for a protein is greater than the mean plus standard deviation it is flagged as being central according to that given centrality measure. This allows a human user to make judgment on the importance and role of the protein by presenting them with all of the available facts, e.g. “This protein is *degree* and *betweenness central*”, rather than a simple “yes/no” answer to whether the protein is important.

**Clustering coefficient:** Though technically not a centrality measure, the clustering coefficient measures the interconnectedness of a node’s neighbours

in the network and can also be used to determine whether or not the graph is a small world network. A high cluster coefficient indicates a high level of interconnection between members of a node’s neighbouring nodes. The clustering coefficient for undirected graphs can be defined as:

$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i-1)} : v_j, v_k \in N_i, e_{ij} \in E \quad (3)$$

where  $V = v_1, v_2, \dots, v_n$  are a set of  $n$  vertices and  $E$  a set of edges, where  $e_{ij}$  denotes an edge between vertices  $v_i$  and  $v_j$ ,  $k_i$  refers to the vertex neighbours. The neighbourhood  $N_i$ , for a vertex  $v_i$ , is its immediately connected neighbours as follows:

$$N_i = \{v_j\} : e_{ij} \in E \quad (4)$$

The degree  $k_i$  of a vertex is the number of vertices in its neighbourhood  $|N_i|$ . Making the clustering coefficient  $C_i$  for a vertex  $v_i$  the proportion of links between the vertices within its neighbourhood divided by the number of links that could possibly exist between them. In addition undirected graphs have the property that  $e_{ij}$  and  $e_{ji}$  are considered identical. Therefore, if a vertex  $v_i$  has  $k_i$  neighbours, only the following edges could exist among the vertices within the neighbourhood:

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

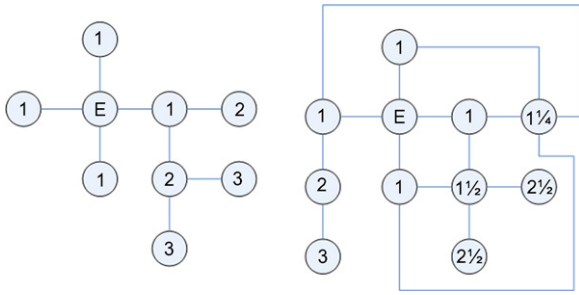
These measures return a value of one if every neighbour connected to  $v_i$  is also connected to every other vertex within the neighbourhood  $n$ , and zero if no vertex that is connected to  $v_i$  connects to any other vertex that is connected to  $v_i$ . The clustering coefficient for the whole system is given by Watts and Strogatz as the average of the clustering coefficient for each vertex:

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

The clustering coefficient measure can be used to verify whether proteins identified as likely to be essential are in fact essential [35]. By calculating and comparing the clustering coefficients between proteins identified as hubs, those identified as not hubs and the system as a whole it would be expected that the proteins identified as hubs should have a higher clustering coefficient in comparison to the others.

### 3.3. Erdos numbers and collaboration graph generation

Groups of proteins interacting together can be viewed as discreet cellular machinery operating



**Figure 4** Simple Erdos numbers—node collaboration numbers increase by one as they move away from Erdos due to there being only single interactions. Rational Erdos numbers—here multiple interactions between nodes produce rational collaboration numbers that increase by  $1/p$  where  $p$  is the number of joint collaborations with nodes.

for a combined purpose. Identifying collaborations between groups of proteins and being able to generate smaller graphs that are more suitable human analysis should allow for a greater understanding of the system as a whole. To achieve this, a collaboration distance measure must be calculated for each protein pair, preferably one which is based solely upon the network structure. Such a collaboration measure should indicate how much influence one protein has on another, or from another perspective, how much its influence on another protein is diluted by other intermediate proteins present between them (Fig. 4).

Collaboration graphs were generated, one for every protein within the diabetes dataset. These graphs were then compared against each other to see which (if any) of the graphs could be contained in other larger graphs. The reasoning behind this comparison is, as previously mentioned, that if the group of proteins collaborate, i.e. interact, with

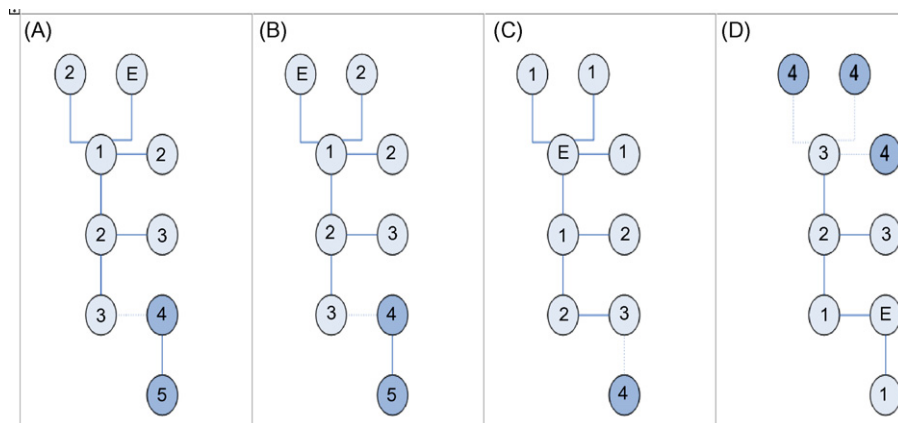
each other extensively it can be expected then that they have similar or even identical collaboration graph structures. Not only is it likely that some graphs will be identical to others, but it is also very likely that some of the graphs can be contained within other larger graphs.

The Erdos number, is a way of describing “collaborative distance” and can be applied to any small world network where the  $6^\circ$  of separation tends to hold true. The analysis uses the “rational Erdos numbers” technique of calculating collaboration distance between proteins. Simple Erdos numbers (see Fig. 5A) can be generated by first selecting a network node to represent Erdos and assigning this node a value of zero, then each node that collaborates (interacts) with this node is given a value of one. This procedure continues, with further nodes being given a value of their predecessor plus one, until every node on the graph has a collaboration distance with respect to Erdos.

The collaboration measure is based on the Watts–Strogatz random model, which uses two parameters,  $p$  the probability of re-arrangement and  $k$ , the number of “ring” neighbours:

- (1) Start with a regular shape ring or lattice graph.
- (2) Each node is connected to  $k$  successive neighbours.
- (3) For each edge in turn.
- (4) Flip a  $p$ -biased coin.
- (5) If heads, replace edge with a random edge from one of its nodes to a random other node.

When  $p = 0$ , regular ring-like graph structures are formed, which have high clustering coefficients and long path lengths. As the value for  $p$  increases,



**Figure 5** This example uses a collaboration limit of 3 (i.e. disconnecting any nodes labeled 4 and 5). Though both A and B use different nodes as Erdos, the graphs that are produced are structurally identical. It is also evident that both A and B are subgraphs (subsets) of C as they can be contained completely within it. D does not fit fully within C however, D can be said to be 83% contained within C (one node is absent from C).

more randomness in the structure is introduced and the resulting graphs tending to have low path lengths and high clustering coefficients. Watts and Strogatz termed these graphs 'small world networks'. When the value for  $p$  reaches 1, the graph becomes identical to the standard Erdos–Renyi random graph with its small path lengths and low clustering coefficients.

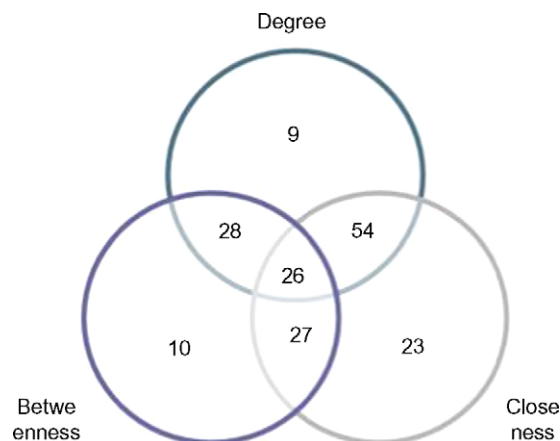
For simple graphs where only single interactions are permitted this technique is adequate, however on complex real world graphs there are typically more than one interaction between nodes (see Fig. 5B), the assignment of numbers in these cases are not fully rational, as a node that interacts with two predecessors of value 1 should be given a value of 1 and not 2. Generalizing the idea, a node collaborating with Erdos  $p$  times should be assigned Erdos number  $1/p$ . In Fig. 5C, both A and B are subgraphs (subsets) of C as they can be contained completely within it. In Fig. 5D the graph does not fit fully within C however, D can be said to be 83% contained within C (one node is absent from C).

The more duplicates of a given graph produced, the more significant that graph is and it increases the possibility that it represents an actual protein complex, as the repeated production of this graph is indicative of a highly collaborative area of the graph (based upon its structure) [36]. Yet the graphs produced need not even be absolutely identical, graphs which can fit completely inside another graph can be considered duplicates and also increases the likelihood of the graph being an actual protein complex.

Our software for the PPI analysis system assigned an equal probability of each collaborative graph generated for each protein being a correct identification of protein complex ( $1/\text{number of proteins in the database}$ ). If the graph was then found to fit inside some other completely its probability of correctness was added to the probability of the containing graph and then set to zero. This simple technique allows for rapid calculation of the most probable graphs, as each possible super graph will be scored a cumulative probability of correctness that is dependent upon the number of supporting subgraphs.

## 4. Results

One indicator as to the accuracy of the hub identification techniques used is the consistency of overlap between the sets of proteins that were highlighted as hubs by each of the centrality measures. There are a total of 99 hubs making up 17% of all nodes and as Fig. 6 shows, many nodes that were identified as central using one measure were also



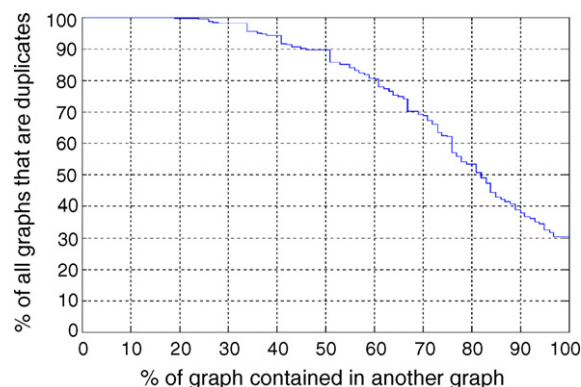
**Figure 6** Showing the distribution and overlap between the various centrality measures.

identified as central by at least one other measure. There were 177 hub proteins out of which only 26 were identified by all three centrality measures.

We generated the collaboration graphs, one for every protein within the diabetes dataset, and then compared against each other to see which (if any) of the graphs could be contained within other larger graphs. Graphs which could contain many other generated graphs were assigned a higher likelihood of importance.

As illustrated in Fig. 7 it can be seen that with 100% containment (on the x-axis), i.e. the entire source graph fits within a destination graph, 30% of all graphs are redundant (on the y-axis). This means that probability values totalling one third will be distributed amongst the other subgraphs, increasing their likelihood of correctness.

There are no hard and fast rules for identifying the "optimum" cut-off limit for the Erdos neighbourhood and it must also be noted that the generated PPI



**Figure 7** The percentage of all graphs which are duplicates (contained) depending upon the percentage of any given graph being contained within some other a graph.

**Table 3** Proteins identified as hubs on average have a higher clustering coefficients than non-hubs

Protein type	Clustering coefficient
Hub	0.29
Non-hub	0.22

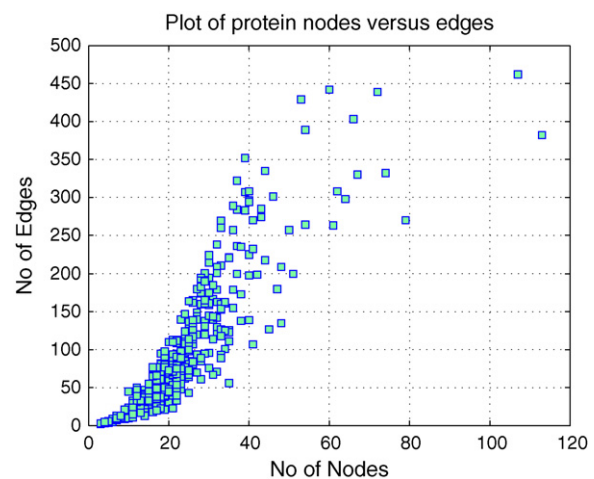
system does not use the full rational Erdos/electrical resistance network system. It is quite possible that both these factors can introduce small inaccuracies which when combined could lead to the incorrect isolation of one or two proteins from a collaboration network; this can be corrected somewhat by allowing a percentage of containment rather than absolute 100%. However, recent work by de Silva et al. has revealed some interesting properties of the sampling process itself and that care must be taken when interpreting the biological significance of the subnetworks [37]. Similarly, investigations by Stumpf et al. have produced results to show that *subnets* of scale-free networks are not scale-free within themselves [38]. However, in Table 3 we can still determine that hub proteins on average have a higher clustering coefficients than non-hubs. In Table 4 we compare the number of detected interactions between essential proteins (IBEPs) in the actual PPI network against the Erdos–Renyi PPI graphs. The ratio of nodes to edges for each of the proteins is shown in Fig. 8, it can be seen that the majority have 20–40 nodes with 50–200 edges. The top 12 ranking proteins are displayed in Fig. 9, these values are based on the three centrality measures for each of the 584 proteins (colored cyan), with the highest ranking protein is the MAPK1 (number 200 on the graph).

Utilising these verification techniques, a significant proportion of the essential proteins within the network have been accurately identified. Information gained from the identification of these proteins can prove invaluable in furthering the understanding of both the collaboration graph results and the network as a whole. They provide biologists with a starting point for their investigations into the operation and possible malfunction of a PPI network.

Care must be taken when selecting the containment thresholds as it can be seen that selecting a value very close to 100% tends not to produce many more results. This is due to the fact that many of the subgraphs generated have relatively few members,

**Table 4** IBEP comparison between PPI network and random rewired network

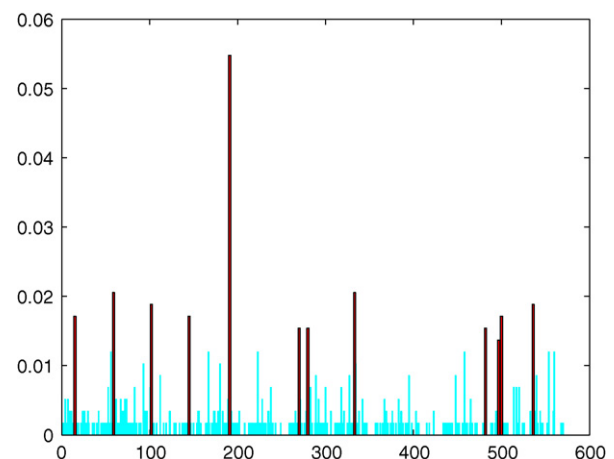
Network type	IBEP's
Actual diabetes PPI network	89
Erdos–Renyi PPI network (mean value over 100 runs)	54.32



**Figure 8** Plot of the number of nodes and edges for each of the complexes.

making the weighting of each member as a percentage of the graph quite high. However, as shown in Fig. 7, it is apparent that if the containment threshold is set to 80%, then 50% of all graphs have been contained within some other graph. This is a significant amount and would drastically reduce the complexity and increase the readability of a PPI graph, if all of the identified subgraphs were each encapsulated by a single node.

With the containment threshold set to 80% the likelihood of certain collaboration graphs being correct increases significantly and in particular one graph becomes more prominent. The collaboration graph centred on the protein MAPK1 (mitogen-activated protein kinase 1) is consistently referenced by a



**Figure 9** Top 12 ranking proteins (dark color) based on the three centrality measures for each of the 584 proteins (colored cyan), the highest ranking protein is the MAPK1 (number 200 on the graph). The y-axis refers to the joint centrality measure and the x-axis refers to the protein id number.

significant proportion of subgraphs and the protein itself has been marked as an essential protein, all of which would seem indicate a very important structure.

A brief description of the MAPK1 protein from the National Centre for Biotechnology Information (NCBI) is:

“The protein is a member of the MAP kinase family. MAP kinases, also are also known as extracellular signal-regulated kinases (ERKs), act as an integration point for multiple biochemical signals, and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development.”

Indeed the representation of the MAPK1 protein within the diabetes PPI network reflects this description with protein having multiple expression sites and many interactions. That the Erdos collaboration measure identified this protein is significant, from the written description that is apparent that indeed this protein is responsible for many interactions with many different types of other proteins. However by examining the collaboration graph generated it is easy to see why this specific protein contained so many subgraphs.

On average, each subgraph contains only 21 members, the MAPK1 graph however, contains 107 members making it by far the largest collaboration graph in the system. The reason for this is that the MAPK1, as the above text states, collaborates extensively with many other proteins which themselves extensively collaborate within each other.

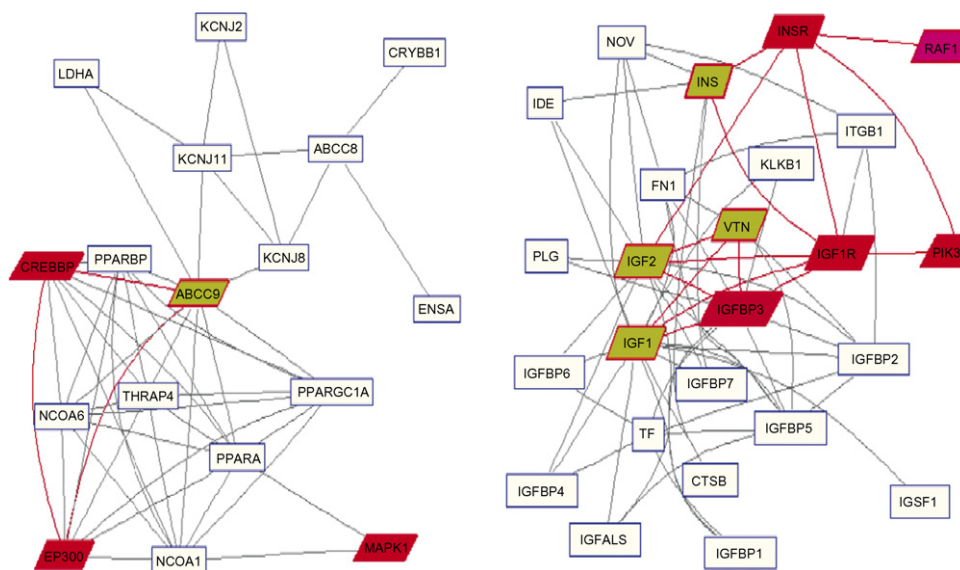
Effectively the MAPK1 protein interacts with so many other highly interactive proteins that many nodes have very low Erdos numbers with respect to it, and accordingly its collaboration graph is very large and tends to dominate and envelope many of the other smaller graphs.

This finding may seem counter productive, as the goal of using the Erdos numbers was to isolate small easily analyzable graphs, however produces a large graph which is difficult to comprehend. The MAPK1 protein indeed is a highly collaborative protein and the Erdos collaboration system highlight this fact. Though the goal of the Erdos numbers is to generate smaller graphs, it is inevitable that some systems found in PPI networks will not be small, and it is entirely natural that there should be found very large collaborative systems.

#### 4.1. Size adjusted importance graphs

The most interesting graphs generated are those with a high likelihood of being correct but also of modest size. One such protein which fits these criteria is the collaboration graph centred on KCNJ11 (Fig. 10). Although this protein network is composed of only 17 members it also has a high likelihood rating. Referring this network to the NCBI database we find that:

“Potassium channels are present in most mammalian cells, where they participate in a wide range of physiologic responses. Mutations in this gene are a cause of familial persistent hyperinsulinemic hypoglycemia of infancy, an autosomal recessive disorder



**Figure 10** Interaction diagrams for KCNJ11 (left) and IGF1 (right), proteins identified as essential using centrality measures are represented as colored parallelograms with nonessential proteins represented as squares.

characterized by unregulated insulin secretion. Defects in this gene may also contribute to autosomal dominant non-insulin-dependent diabetes mellitus type II.”

Similarly the collaboration graph centred on the IGF1 protein also possesses a high likelihood and relatively small size, the NCBI database record for this protein reads:

“The somatomedins, or insulin-like growth factors, comprise a family of peptides that play important roles in mammalian growth and development. IGF1 mediates many of the growth-promoting effects of growth hormone.”

It is readily apparent that these results seem much more relevant to the subject matter of diabetes, the other systems that are picked up are largely generic systems concerned with basic operation of the cell. By dividing a graphs calculated likelihood of importance by its size, a list of graphs sorted upon size adjusted importance can be generated.

At the top of this new list are the collaboration graphs centred around the KCNJ11 and IGF1 proteins, however some other interesting proteins are also highlighted, in particular GCKR and RAMP2, this subnetwork is shown in Fig. 11. The NCBI database records for RAMP2 is:

“RAMP2 and calcitonin receptor-like receptor (CRLR) can function as an ADM receptor. To investigate whether ADM has implications as a pathophysiologic substance in pregnancy-induced hypertension, Makino et al. (2001) measured the changes of expression of RAMP2 and CRLR in fetomaternal tissues in

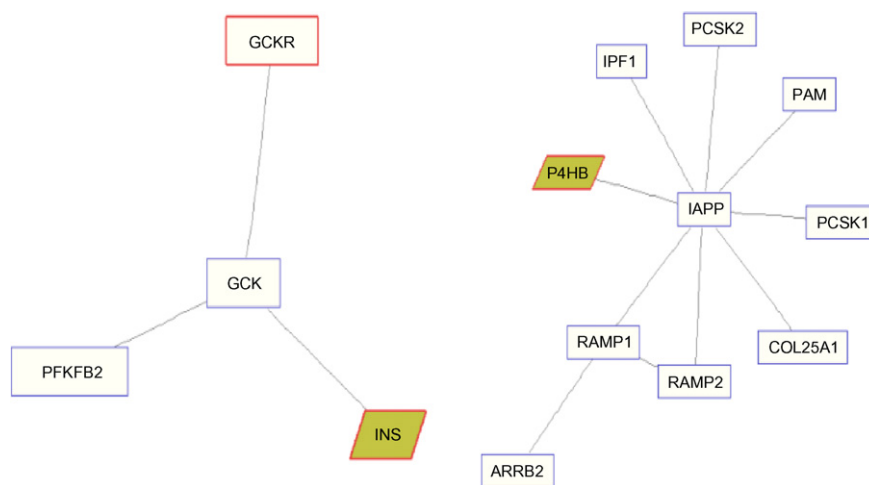
normotensive pregnant women and pregnancy-induced hypertensive women”

Some of the key complexes are presented in Table 5, the first column describes the important hub protein complex, the next columns describe the number of proteins involved with it and the interactions that occur. The remaining columns describe the biochemical role and the major effects (location) of the complex within the body (usually cell nucleus).

The identification of complexes involved in other diseases is particularly interesting because of the known disfunctions associated with diabetes. The diabetes (insulin resistance) hub proteins interact with other hub proteins important for other cellular functions. Should these hub proteins dysfunction for any reason then we may reasonable expect a cascade effect. The identification of cross-talk between signalling pathways is a difficult area to conduct experimentally, but can reveal the multi-purpose role of hub proteins in large networks of interaction [39].

## 4.2. Biological significance of network architecture

The biological implications of this network architecture are that high-degree nodes, i.e. the hub proteins, are involved in regulating or at least facilitating communication between many others [32] making these nodes more important to the overall network function than a low degree nodes. Thus, analyzing the functional importance of a node with respect to the overall network structure can be an accurate means of identifying essential proteins and if the magnitude of changes in global network



**Figure 11** Interaction diagrams for GCKR (left) and RAMP2 (right), proteins identified as essential using centrality measures are represented as colored parallelograms with nonessential proteins represented as squares.

**Table 5** Connectivity and function for discovered biological complexes

Complex	Number of proteins	Interactions	Role	Localisation
MAPK1	107	462	Extracellular signalling	Nucleus
RAMP2	10	10	ADM receptor-hypertension related	Umbilical artery
GCKR	4	3	Glucose regulation	Pancreas
GRB2	66	403	Epidermal growth factor	Nucleus
IGF1	25	64	Insulin growth factor	Nucleus
CREBBP	62	308	Related to Rubinstein–Taybi syndrome	Nucleus
ISL1	36	155	Development of islets of Langerhans	Pancreas
MAD2L1	74	332	Chromosome development	Nucleus
AKT1	45	127	Cell apoptosis	Nucleus
ESR1	67	330	Estrogen receptor	Nucleus
P4HB	31	67	Protein encoding	Nucleus
CCDC6	113	382	Involved with thyroid cancers	Thyroid gland
PCGF2	33	89	Tumor suppressor	Nucleus
TRHDE	7	7	Hormone signalling	Brain, heart, liver
LRDD	7	7	p53 tumor apoptosis	Nucleus
GGA1	18	23	Trafficking between golgi and lysosome	Golgi apparatus
SLC2A4	19	67	Glucose transporter	3T3-L1 adipocytes
KCNJ11	17	49	Membrane construction—defects cause infantile diabetes	Cell wall

structure caused by the removal of a node defines that node's importance, then other network feature measures can also be used to identify further essential proteins.

It is important to note that the discussion so far has focused upon the proteins themselves and not the interactions, recent work [31] suggests that the centrality-lethality rule may be unrelated to the network architecture; rather it is the interactions (edges) between some proteins that are essential. That hubs tend to be essential is explained by the simple fact that they have large numbers of PPI's, and therefore higher probabilities of engaging in essential PPI's. Irrespective of the reasons why hubs are essential, we point out the practicality of such measures, that any centrality measure is capable of identifying almost twice as many indispensable proteins as random selection. By combining multiple centrality measures it is logical that the accuracy of identifying such indispensable proteins should increase.

Not all proteins are ubiquitously expressed in all parts of the body. Some proteins are expressed only in one specific site and may interact with other proteins in other the parts of body indirectly, via some intermediate proteins. Allowing the user to specify the expression site (with essential proteins highlighted) produces PPI graphs that are much more readable and relevant to the information they require [40]. Is evident that drawing boundaries such as these can be very helpful in narrowing down the set of interesting proteins it is also trivial task for further boundaries to be drawn, such as separa-

tion based upon whether a protein is expressed intra or extra cellular. Eventually however a limit will be reached upon which there are no more separable criteria easily available to partition the network.

Segregating the network like this, convenient as it is, also runs the risk of introducing artificial boundaries, as the proteins which are expressed, and thus interact, in multiple expression sites may not be interacting equally in both expression sites. A protein which seems insignificant in expression site A might actually be critical to the functioning of the organism in expression site B. The true 'holistic' functionality of a protein that is expressed in site A and B may not be evident by examining that protein in isolation regarding its operation within either one of those given sites. This begs the question of should proteins with multiple expression sites be analyzed only in site A, only in site B, only with other proteins that are in both sites or should the system be analyzed as a whole without regard to expression sites? A method for partitioning the network based solely upon functionality implied from network structure would be incredibly useful. Such a partitioning system could simplify these complex PPI graphs by grouping together multiple proteins into a single amalgam representative of its component parts. A further advantage of identifying functional subgraphs within the PPI system is that it will allow biologists to focus their efforts in understanding the operation of the identified protein complexes rather than attempting the daunting task of trying to understand the operation of each individual protein within the system as a whole.

## 5. Related work

Investigating PPI datasets has attracted much attention and several novel methods have been proposed to tackle the challenges of extracting meaningful biological knowledge [41–43]. A few approaches use graph based model building integrated with machine learning techniques such as fuzzy logic and neural networks to build richer more flexible models [44–46]. A few approaches have adapted existing graph based algorithms to tackle specific problems posed by PPI networks, or have developed novel graph techniques for predicting protein function [47,48], or to detect frequent subgraphs [49,50].

Several approaches such as PathSys and others were developed with the needs of systems biology in mind and therefore have facilities for heterogeneous data integration, often using pathway information and often incorporating kinetic parameters as the nodes in the network [51–54]. The bioPIXIE systems of Myers is of interest as the authors discuss cross-talk and interference in biological networks [55]. Our work focuses on protein complexes while bioPIXIE integrates data from several sources additionally, bioPIXIE benefits from a well designed, web enabled user interface. Bader and Hogue developed an automated technique to map out the molecular complexes found in yeast [56]. So far few attempts have been made in the literature that address human protein databases, especially relating this work to specific diseases.

## 6. Conclusion

The diabetes PPI network appears to be a small world network with a power law degree distribution, as is common with many PPI networks. It is possible that diabetes is caused by the removal of a hub protein (as small world networks are prone to targeted removal of nodes with high degrees) or disruption to an essential PPI. Highlighting essential proteins and their interactions is invaluable in assisting humans to understand the interaction graphs produced. Though the method shown in this paper is sufficient for illustration of the graphs, a commercial system would require a much higher degree of accuracy and an alternative to the clustering and cut-off values used on the centrality measures would be to produce a probability of each protein being central according to the different measures.

However even with the essential proteins highlighted it has been shown that these PPI interaction graphs are often very complicated and there are

limits and drawbacks to segregating them based solely upon individual protein attributes. Proteins function as networks and examining the properties of one isolated protein does not necessarily explain the complete 'holistic' functionality of the group of proteins that interacts with.

Using rational Erdos numbers as a graph data mining technique it is possible to segregate such complex graphs by identifying smaller subsystems contained within based solely upon possible functionality implied by network structure. The use of this technique however is undirected and its result often highlights many of the different systems operating within the network. This however is not necessarily a bad thing, as it is to be expected that there are indeed many subsystems present. The accurate identification of these protein complexes and the identification of proteins central to those can provide an invaluable starting point for biologists when attempting to understand a new system where many proteins have been identified and their interactions are known but their end functionality is not. By selecting only proteins that are likely to be involved in the system the user wishes to study (for example, by selecting those expressed in the Islets of Langerhans or those that interact with insulin) it is possible to focus the process to a degree.

Future work will include the integration of additional sources of PPI data such as BIND, DIP, InAct, MIPS, etc. Even with the inherent redundancy of interactions that exist, it is hoped that we may draw more reliable and meaningful conclusions from these extra sources. Furthermore, we intend to explore a neglected aspect of PPI modeling, namely the subtle effects of cross-talk that occurs between the different pathways that coordinate signalling and regulatory communications.

## Acknowledgments

We wish also to thank the anonymous reviewers for their helpful comments for improving the paper. This work was part supported by a Research Development Fellowship funded by HEFCE and the Biosystems Informatics Institute (Bii). We also acknowledge the use of the MatlabBGL Boost Graph Library package written by David Gleich.

## References

- [1] OMalley M, Dupre J. Fundamental issues in systems biology. *Bioessays* 2005;27(12):1270–6.
- [2] Butcher E, Berg E, Kunkel E. Systems biology in drug discovery. *Nat Biotechnol* 2004;22(10):1253–9.

- [3] Caragea D, Pathak J, Bao J, Silvescu A, Andorf C, Dobbs D, et al. Information integration and knowledge acquisition from semantically heterogeneous biological data sources. In: Ludäscher B, Raschid L, editors. Proceedings of the 2nd international workshop on data integration in life sciences (DILS'05). 2005. p. 128–39.
- [4] Blake J, Bult C. Beyond the data deluge: data integration and bio-ontologies. *J Biomed Inform* 2006;39:314–20.
- [5] Pollard J, Butte A, Hoberman S, Joshi M, Levy J, Pappo J. A computational model to define the molecular causes of type 2 diabetes mellitus. *Diabetes Technol Ther* 2005;7(2):323–36.
- [6] Bonneau R, Reiss D, Shannon P, Facciottit M, Hood L, Baliga N, et al. The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Gen Biol* 2006;7(R36).
- [7] Kahlem P, Birney E. Dry work in a wet world: computation in systems biology. *Mol Syst Biol* 2006;2(40):1–4.
- [8] Langley P, Shiran O, Shrager J, Todorovski L, Pohorille A. Constructing explanatory process models from biological data and knowledge. *Artif Intell Med* 2006;37(3):191–201.
- [9] Nikolsky Y, Nikolskaya T, Bugrim A. Biological networks and analysis of experimental data in drug discovery. *Drug Discov Today* 2005;10(9):653–62.
- [10] Lin N, Wu B, Jansen R, Gerstein M, Zhao H. Information assessment on predicting protein–protein interactions. *BMC Bioinformatics* 2004;5(154).
- [11] Chen J, Hsu W, Lee M, Ng S. Discovering reliable protein interactions from high-throughput experimental data using network topology. *Artif Intell Med* 2005;35(1/2):37–47.
- [12] Suthram S, Shlomi T, Ruppin E, Sharan R, Ideker T. A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics* 2006;7(360).
- [13] Deeds E, Ashenberg O, Shakhnovich E. A simple physical model for scaling in protein–protein interaction networks. *Proc Natl Acad Sci* 2006;103(2):311–6.
- [14] Letovsky S, Kasif S. Predicting protein function from protein–protein interaction data: a probabilistic approach. *Bioinformatics* 2003;19(1):197–204.
- [15] McGarry K, Garfield S, Morris N. Recent trends in knowledge and data integration for the life sciences. *Expert Syst: J Knowl Eng* 2006;23(5):337–48.
- [16] Yeh I, Karp P, Noy N, Altman R. Knowledge acquisition, consistency checking and concurrency control for gene ontology. *Bioinformatics* 2003;19(2):241–8.
- [17] Halban P, Ferrannini E, Nerup J. Diabetes research investment in the European Union. *Nat Med* 2006;12(1):70–2.
- [18] Baltrusch S, Tiedge M. Glucokinase regulatory network in pancreatic  $\beta$  cells and liver. *Diabetes* 2006;55(2):55–64.
- [19] Döhr S, Klingenhoff A, Maier H, de Angelis MH, Werner T, Schneider R. Linking disease-associated genes to regulatory networks via promoter organization. *Nucleic Acids Res* 2005;33(3):864–72.
- [20] Bergman R. Editorial: insulin action and distribution of tissue blood flow. *J Clin Endocrinol Metab* 2006;88(10):4556–8.
- [21] Cheatham B. GLUT4 and company: SNAREing roles in insulin-regulated glucose uptake. *Trends Endocrinol Metab* 2000;11(9):356–61.
- [22] Perera H, Clarke M, Morris N, Hong W, Chamberlain L, Gould G. Syntaxin 6 regulates glut4 trafficking in 3T3-L1 adipocytes. *Mol Biol Cell* 2003;14:2946–58.
- [23] Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, et al. The MIPS mammalian protein–protein interaction database. *Bioinformatics* 2005;21(6):832–4.
- [24] Bader G, Betel D, Hogue C. BIND: the biomolecular interaction network database. *Nucleic Acids Res* 2003;31(1):248–50.
- [25] Salwinski L, Miller C, Smith A, Pettit F, Bowie J, Eisenberg D. The database of interacting proteins: 2004 update. *Nucleic Acids Res* 2004;32:449–51.
- [26] Stark C, Breitkreutz B, Reguly T, Boucher L. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;34:535–9.
- [27] Mishra S, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, et al. Human protein reference database—2006 update. *Nucleic Acids Res* 2006;34:D411–4.
- [28] Mathivanan S, Periaswamy B, Gandhi T, Kandasamy K, Suresh S, Mohmood R, et al. An evaluation of human protein–protein interaction data in the public domain. *BMC Bioinformatics* 2006;7(5).
- [29] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. *Science* 2002;298:824–7.
- [30] Kashtan N, Itzkovitz S, Milo R, Alon U. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics* 2004;20(11):1746–58.
- [31] He X, Zhang J. Why do hubs tend to be essential in protein networks? *PLoS Genet* 2006;2:826–34.
- [32] Hu X. Mining and analysing scale-free protein–protein interaction network. *Int J Bioinformatics Res Appl* 2005;1(1):81–101.
- [33] Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K, Kanaya S. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics* 2006;7(205):1–13.
- [34] Watts D, Strogatz S. Collective dynamics of small world networks. *Nature* 1998;393:440–2.
- [35] Huang X, Lai J, Jennings F. Maximum common subgraph: some upper bound and lower bound results. *BMC Bioinformatics* 2006;7(4).
- [36] Wernicke S. Efficient detection of network motifs. *IEEE/ACM Trans Comput Biol Bioinformatics* 2006;3(4):347–59.
- [37] de Silva E, Thorne T, Ingram P, Agrafioti I, Swire J, Wiuf C, et al. The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Bioinformatics* 2006;4(39).
- [38] Stumpf M, Wiuf C, May R. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci* 2005;102(12):4221–4.
- [39] Papin J, Hunter T, Palsson B, Subramaniam S. Reconstruction of cellular signalling networks and analysis of their properties. *Mol Cell Biol* 2005;6:99–111.
- [40] Fraser A, Marcotte E. A probabilistic view of gene function. *Nat Genet* 2004;36(6):559–64.
- [41] Ideker T, Ozier O, Schwikowski B, Siegel A. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 2002;18(1):233–40.
- [42] Zotenko E, Guimaraes K, Jothi R, Przytycka T. Decomposition of overlapping protein complexes: a graph theoretical method for analyzing static and dynamic protein associations. *Algorithms Mol Biol* 2006;1(7).
- [43] Rachlin J, Cohen D, Cantor C, Kasif S. Biological context networks: a mosaic view of the interactome. *Mol Syst Biol* 2006;2(66):285–308.
- [44] Scholtens D, Vidal M, Gentleman R. Local modeling of global interactome networks. *Bioinformatics* 2005;21(17):3548–57.
- [45] Klipp E, Liebermeister W. Mathematical modeling of intracellular signalling pathways. *BMC Neuroscience* 2006;7(1).
- [46] Bosl W. Systems biology by the rules: hybrid intelligent systems for pathway modeling and discovery. *BMC Syst Biol* 2007;13(1).
- [47] Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph theoretic analysis of interaction maps. *Bioinformatics* 2005;21(1):302–10.

- [48] Shafer P, Isganitis T, Yona G. Hubs of knowledge: using the functional link structure in Biozon to mine for biologically significant entities. *BMC Bioinformatics* 2005;7(71).
- [49] Koyuturk M, Grama A, Szpankowski W. An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics* 2004;20(1):200–7.
- [50] Klamt S, Saez-Rodriguez J, Lindquist J, Simoeni L, Gilles E. A methodology for the structural and functional analysis of signalling and regulatory networks. *BMC Bioinformatics* 2006;7(56).
- [51] Baitaluk M, Qian X, Godbole S, Raval A, Ray A, Gupta A. Pathsys: integrating molecular interaction graphs for systems biology. *BMC Bioinformatics* 2006;7(55).
- [52] Leibermeister W, Klipp E. Bringing metabolic networks to life: integration of kinetic, metabolic and proteomic data. *Theor Biol Med Model* 2006;3(42).
- [53] Yeang C, Vingron M. A joint model of regulatory and metabolic networks. *BMC Bioinformatics* 2006;332(7).
- [54] Schaub M, Henzinger T, Fisher J. Qualitative networks: a symbolic approach to analyze biological signaling networks. *BMC Syst Biol* 2007;4(1).
- [55] Myers C, Robson D, Wible A, Hibbs M, Chiriac C, Theesfeld C, et al. Discovery of biological networks from diverse functional genomic data. *Gen Biol* 2005;6(R114).
- [56] Bader G, Hogue C. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003;4(2):1–27.