Computational Techniques for Identifying Networks of Interrelated Diseases

Ken McGarry and Ukeme Daniel
Department of Pharmacy, Health and wellbeing, Faculty of Applied Sciences, University of Sunderland, UK
Email:ken.mcgarry@sunderland.ac.uk

***Abstract*—Recently there has been a lot of interest in using computational techniques to build networks of protein-to-protein interactions, interacting gene networks and metabolic reactions. Many interesting and novel discoveries have been made using graph based structures using links and nodes to represent the relationships between proteins and genes. Analysis of graph networks has revealed that genes and proteins cooperate in modules performing specific functions and that there is crosstalk or overlap between modules. In this paper we take these ideas further and build upon current knowledge to build up a network of human related diseases based on graph theory and the concept of overlap or shared function. We explore the hypothesis that many human diseases are linked by common genetic modules, therefore a defect in one of any of the cooperating genes in a module may lead to a specific disease or related symptom. We build our networks using data and information extracted from several online databases along with supporting knowledge in the form of biological ontologies.**

***Keywords*—Bipartite network, hub, protein, biological pathway.**

I. Introduction

Global health problems and associated costs are increasing and causing concern, notably diseases such as diabetes. It is estimated that there will be approximately 380 million individuals world-wide suffering from diabetes by 2025 with all the associated problems of treatment and long-term support [32]. This is a particularly worrying trend since those suffering from diabetes have a doubled relative risk of developing dementia and Alzheimer disease [1]. A better understanding of the causes of disease such as the underlying genetic make up of individuals will eventually lead to improved treatments. This has only recently been made possible by the development of advanced genomic and proteomic techniques which are able to provide detailed and accurate data on individual cellular processes. We are now able to determine which genes (and proteins) interact together and form related functional groups that modify the behavior and ultimately the health of the cell [24].

Recent studies based on modeling cooperating modular groups of genes have suggested that diseases themselves are in fact network like [11], [5]. The concept of the human disease network (HDN) or *diseasome* is relatively new and is now starting to be explored as means of developing new drug products to tackle and combat diseases [7], [13]. Figure 1 shows only a small part of the connectivity relating to diabetes.

The view taken is that the diseasome is best represented as a bipartite graph formed by two sets of disjoint nodes, with the disease state forming one set and the other comprising

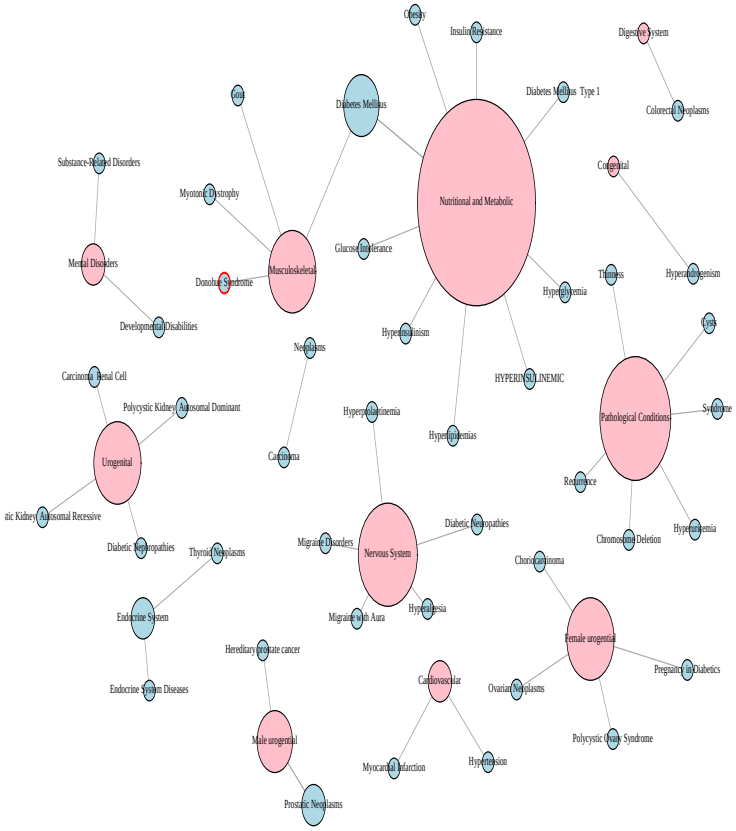

Fig. 1. Diabetes related diseases and their class. The diseases are denoted by light blue nodes and the class of disease by pink nodes, the size of the node is determined by the number of connections and hence it's influence. Several diseases appear unconnected with diabetes but in fact they all have two genes in common.

the disease causing nodes [16]. The first diseasome network was constructed using data from the OMIM database, the links between the two sets of nodes were generated when mutations in a gene were implicated in a disorder[11].

There are many challenges to overcome when building accurate networks of disease, most notably is the scale of the problem; the human genome consists of approximately 25,000 protein coding genes of which about 10% have some linkage or association to a specific disease [3]. Furthermore, there are approximately 1,000 known metabolites and chemicals that can interact with and modify the behavior of the protein

978-1-4799-5538-1/14/$31.00

products [25]. Additional complexity of the disease network comes in part from cooperating clusters of genes participating in more than one cellular function, loss of gene function can contribute to several disorders in humans. For example Zellweger syndrome, where long chains of fatty acids can accumulate in the tissues causing problems with the central nervous system has so far been linked to defects in any one of 11 genes [33].

For other diseases, the issue however is not so straightforward as the other genes in a particular cluster may compensate for the malfunction of one gene, this is not so easy to identify [21]. However, recent work has identified an underlying set of organizing principles that can be used to assess and identify the structures involved, we discuss these criteria in section two. A good overview of the technologies and challenges such as the cellular organization and the detection of structural motifs involved in network-based approaches to understanding diseases can be found in Barabasi [4].

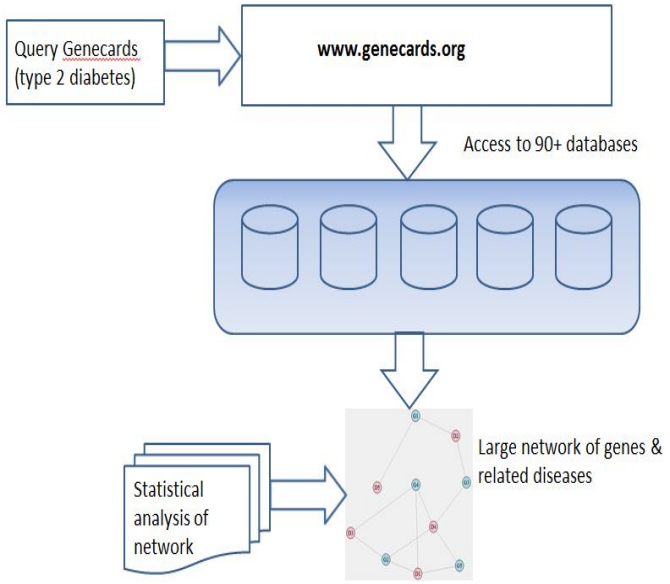

Fig. 2. Overview of system operation, from initial database query to statistical analysis of generated networks

The fundamental building block for creating networks of diseases are the data held in the protein to protein interaction (PPI) databases [20]. There now exists a great body of work on PPI datasets and several novel methods have been proposed to tackle the challenges of extracting meaningful biological knowledge [14], [35], [28]. 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 [30], [18], [6]. 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 [26], [31], or to detect frequent subgraphs [19], [17].

Several approaches such as PathSys [2] 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 within the nodes in the network [22], [34], [29].

The system we have developed obtains data from a variety of sources, we query the genecards/malacards database with genes that are associated with type 2 diabetes. In figure 2 the overall information flow starts with a web based search on the genecards site where the search results are saved as an excel file. This is loaded into R as an edge list consisting of disease/gene pairs, which is structured into a bipartite graph using several packages. At this stage the network is suitable for data analysis and several metrics relating to its connectivity are calculated.

However, a number of challenges are presented by the nature of the data used in the construction of networks, most notably it breaks a number of statistical assumptions, for example there is a high degree of correlation within the cell of the metabolites and various activity patterns. This invalidates the important statistical assumption of independent variables. Also, the most powerful statistical techniques are parametric but the majority of proteomic data are highly skewed. Considering these limitations our aim is to develop a computational model of the key aspects of the human disease network using graph mining and clustering.

The remainder of this is paper is structured as follows; section two describes the issues specific to graph based computational techniques we use; section three discusses the methods and results including sources of data; finally, section four presents the conclusions.

II. Using network theory to link diseases with genes

Graph theoretic methods can be applied to any discipline where the entities of interest are linked together through various associations or relationships. Quite diverse application areas such as social network analysis and biological networks are particularly suited to the mathematics of graph construction, traversal and inferencing. A graph G = (V, E) consists of a set of nodes often called vertices V and a set of links called edges E, . The links in this case are undirected, that is to say there is no implied direction to the relationship in the sense that A causes B.

The criteria we use to determine the relevance of disease connectivity is based upon well known measures from the graph theoretic literature [10], [4].

- HUBS. Non-essential disease genes which represent the majority of genes tend not to be hubs and segregate at the functional periphery of the interactome.
- LOCAL. Proteins involved in the same disease have an increased tendency to interact with each other.
- MODULARITY. Many cellular components tend to be clustered together and form networks.
- PARSIMONY. Various pathways often coincide with the shortest molecular paths between known disease-associated pathways.
- SHARED COMPONENTS. Genes that share cellular components show phenotypes and co-morbidity.

It is likely that the essential genes and the disease genes encode the hubs.

- CLOSENESS. 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 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:

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

- BETWEENNESS. 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:

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

Bipartite graphs often called two-mode graphs are a special case of the standard graph whereby two sets of different nodes are connected by links. The nodes in the first set are of the same type (in our case proteins), the nodes in the second set are different entities (in our case diseases). The connections or links between these two sets of nodes defines the relationships between them. A graph may be described as bipartite if there is a partition of its vertex set V = S U T such that each edge in E has exactly one end-vertex in S and one end-vertex in T. In figure 3 a bipartite graph is displayed showing G = (S U T, E) the set of vertices S = Disease 1, Disease 2, Disease 3, Disease 4, Disease 5 along with the set of vertices T =Gene1, Gene2, Gene3, Gene4, Gene5.

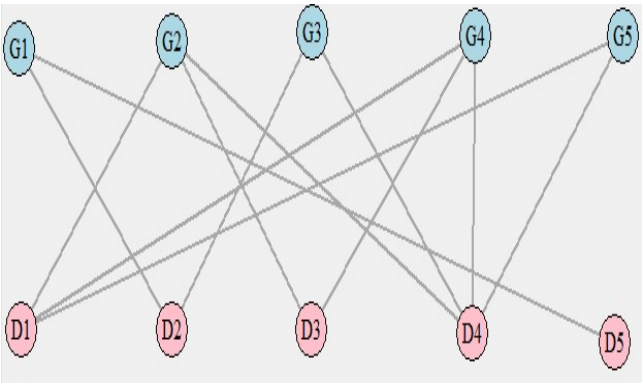

Fig. 3. Simple example of a bipartite graph with the pink nodes representing diseases and the blue nodes represent genes known to be implicated.

The bipartite model for the data consists of genes and diseases, the connectivity of the graph is stored as an k-by-n matrix denoted by M. The structure of the bipartite graph in figure would be represented by the matrix below where the row index refers to the genes and the columns refer to the diseases:

$$M_{k\times n} = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

III. Methods and Results

We obtained our data from the GeneCards online system (http://www.genecards.org/), this repository contains information on human genes, proteins, clinical, and functional information . It integrates data from a wide variety of sources (approx 90) such as HGNC, Ensembl, and NCBI data bases and includes disease related information [9].

The Gene/Mala Cards database assigns a score to each gene, which represents a measure of the confidence based on the p-value, and then by the size of the group of genes associated with the descriptor. The confidence scores range from high scoring multiple experiments that confirm protein to disease associations to lower scoring information gained from text mining.

The Malacards score is calculated by:

$$MCRS = log_{10}(log_{10}(S_{GD}) \sum_{i=1}^{sharedgenes} log_{10}(S_{LR}(i))) + 10 + N_s \tag{3}$$

Where: $S_{GD}$is the rank of the GeneDecks score and $S_{LR(i)}$ is the search engine rank score of a gene and it's association with a disease. We use the MCRS score as a measure for clustering and module detection.

Recent observations confirm that a great degree of modularity occurs in protein networks with overlap or crosstalk of functionality [23]. This is likely to have arisen through evolution since it confers the benefits of robustness and the possibility of increasing the number of cellular functions with the same genetic machinery. It may also lead to co-morbidity which implies the presence of one or more diseases in addition to the primary disease. The most often cited example of co-morbidity is the relationship between obesity and diabetes, where the structural changes in the adipose fatty tissues leads to an irreversible progressive defect in insulin secretion coupled with a progressive rise in insulin resistance.

Figure 4 is a scatter plot of the 'closeness' measure for the entire bipartite network. The graph suggests that a fairly constant level of closeness for most diseases (0.25 to 0.32) in the network until the index score reaches about 900 and the variance becomes somewhat higher. This can be interpreted as the chance or probability of a given protein to be interacting with several other proteins, but with the possibility to be irrelevant for few other proteins.

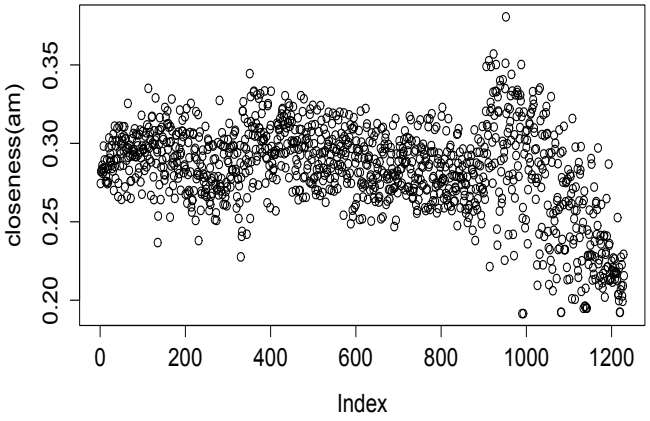

Fig. 4. The closeness measure for the full Bipartite graph

Thus, a protein with high closeness, compared to the average closeness of the network, will be easily central to the regulation of other proteins but with some proteins not influenced by its activity. However, it is also useful to analyze proteins with low closeness, in contrast to the average closeness of the network, as these proteins, although less relevant for that specific network, are possibly behaving as intersecting boundaries with other networks (crosstalk between modules) [23].

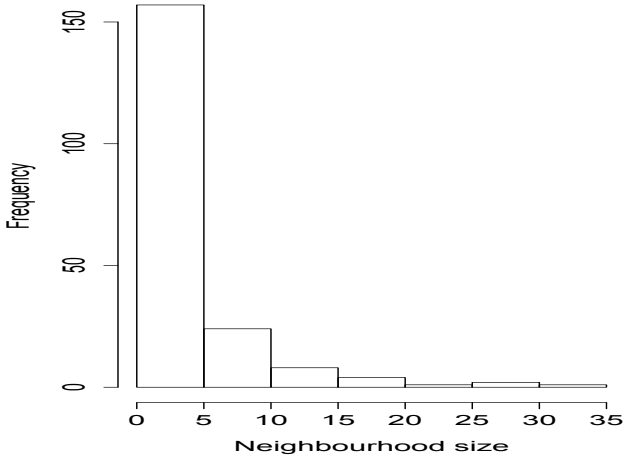

Fig. 5. Distribution of nodes per neighbourhood

Referring to figure 5, the histogram reveals that approx 150 nodes have a connectivity of up to 5 other nodes, while a much smaller number of nodes have a connectivity greater than this. A further analysis using a link connectivity package linkcomm) provided more details [15].

- Number of nodes = 1230 (905 diseases, 325 genes)
- Number of edges = 5430
- Number of communities = 169
- Maximum partition density = 0.01006797
- Number of nodes in largest cluster = 475

Since at the time of download the GeneCards system only allowed six genes per disease, therefore 6 x 905 will account for the 5430 edges or connections in the network. This is a major limitation of the current study and future access to the database will incorporate more genes. Further bias is introduced because disease related genes (proteins) tend to be studied more, the high connectivity experienced by disease genes in general may be a result of this.

TABLE I. DISEASES WITH A COMMONALITY OF AT LEAST TWO GENES

Type II Diabetes	Pancreatitis	Thyroiditis	Fibrosis
PIK3R1	PIK3R1	PIK3R1	PIK3CG
PIK3C2A	PIK3C2A	PIK3C2A	RECK
LRP5	LRP5	PIK3CG	REG1
LRP6	PDCD1LG2	COG2	REN
COG2	PDE3B	RECK	RET
PDE3B	PIK3CG	REN	CNR1

Table I indicates the top closest four diseases based on the shared genes, each diseases has at least two genes in common with diabetes. These common genes are generally referred to as hubs and are central to the idea of the diseasome and mutual dependence.

The gene PIK3R1 stands for Phosphatidylinositol 3-kinase and is an enzyme noted for its role in insulin resistance when it becomes defective. It also affects signaling pathways and this is its likely role in pancreatic cancer. The other possible effect is in thyroid cancer where the mutation will effect cellular processes in the thyroid gland. Fibrosis may occur which in turn can lead on to diabetic retinopathy. Furthermore, recent work has shown that PIK3R1 plays a role in hypertension, breast cancer and Short Syndrome. Therefore, the role of even a single hub gene such PIK3R1 is pivotal in a number of disorders.

The modularity of the disease network is better represented by the grid diagrams 6 to 9 which clearly show the key genes against the clusters of diseases they participate in.

The modules are determined by taking the bipartite weighted graph and applying Newmans modularity measure in a bipartite weighted version to it [8]. It is very memory intensive and sends interim calculations to files on the hard drive, these are deleted after the computation terminates. It should be noted that these grid diagrams present the same connectivity information displayed in the bipartite graph but with the additional information of cluster groupings.

The diseases are listed on the y-axis and the x-axis shows the proteins involved, cross referencing the black oblongs, denotes where a disease is linked to a protein. The clusters are highlighted by a box drawn around the gene/disease cross reference.

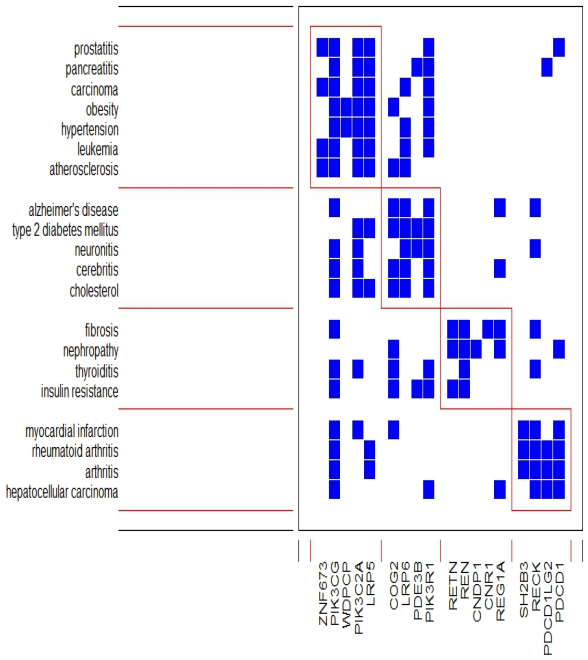

Fig. 6. Disease/gene functional modules with four clusters identified

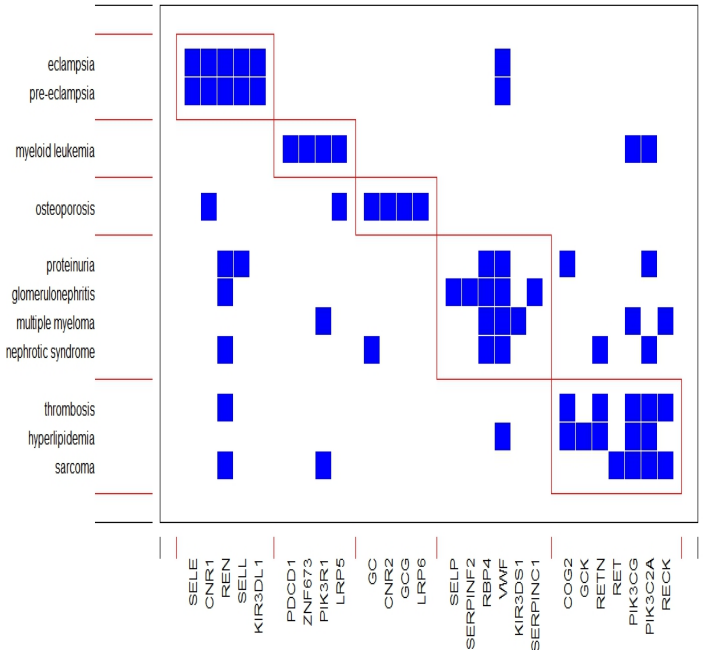

Fig. 7. Disease/gene functional modules with five clusters identified

A further analysis of the clusters revealed the known biological pathways these proteins take part in. For example out of the six proteins forming the top left-hand cluster in figure 6 we can find pathway information (using the NCBI2R package) for three of these six proteins, this is displayed in table II.

TABLE II. PATHWAY INFORMATION FOR COOPERATING PROTEIN CLUSTERS

HPRD ID	Entrez ID	Pathway/functions
LRP5	4041	Reactome Event:Signal Transduction
PIK3C2A	5286	Reactome Event:Membrane Trafficking
PIK3C2A	5286	Reactome Event:Metabolism
PIK3CG	5294	Reactome Event:Hemostasis
PIK3CG	5294	Reactome Event:Signal Transduction
PIK3CG	5294	Reactome Event:Metabolism

Based on the data in table II we are not surprised to see signal transduction and membrane trafficking present given that the associated diseases in this cluster are diabetes/lysosomal storage disease related - that is to say they involve membrane trafficking problems of protein by products not entering or leaving the cell.

We have used another technique to support the cluster analysis, this involved calculating two measures for each node (protein) these are based on the criteria developed by Guimera and Amaral [12] and refined by Olesen *et al* [27]. The criteria define the characteristics of a node (protein) based on empirical cut-off points that can be displayed on a scatterplot.

The actual values of the cut-off points have been tested on a variety of different datasets including proteomic/metabolomic data similar to our application and a wide variety of ecological networks [8]. The location of each node in the scatterplot is determined by the so-called CZ values, the Z-score is calculated for each protein's *within module degree* on the y-axis and the C or participant correlation coefficient is displayed on the x-axis.

The z-score is defined by the following equation where k_i is the number of links of node *i* to the nodes in nearby modules si, and ks_i is the average of k over all nodes in si and $\sigma_{k_{s_i}}$ is the standard deviation of k in s_i :

$$Z_i = \frac{k_i - k_{s_i}}{\sigma_{k_{s_i}}} \tag{4}$$

The z-score allows us to measure how well-connected protein i is to the other proteins within the module. The second parameter called the participation coefficient measures the *role* of a particular protein in other modules, where the participation coefficient P_i of a given node i:

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{k_{i_s}}{k_i} \right)^2 \tag{5}$$

Where k_{i_s} defines the number of links from protein i to proteins in modules s, k_i is the degree of protein i. We can interpret the participation coefficient for each protein by noting that the value is constrained between 1.0 and 0, that is to say 1.0 implies that all of protein i links are evenly distributed between all of the modules and 0 if its links are only within one module.

The scatterplot is divided into four quadrants, the upper left defines those proteins that are acting as *module hubs* based on a cut-off value of z $\geq$ 2.5 and for z $\leq$ 2.5 are *non-hubs*. The coefficient cut-off values for C are $\geq$ 0.62 and $\leq$ 0.62 accordingly, to determine if a node is either a *network hub* or a *connector*. The connector proteins generally link at least two modules.

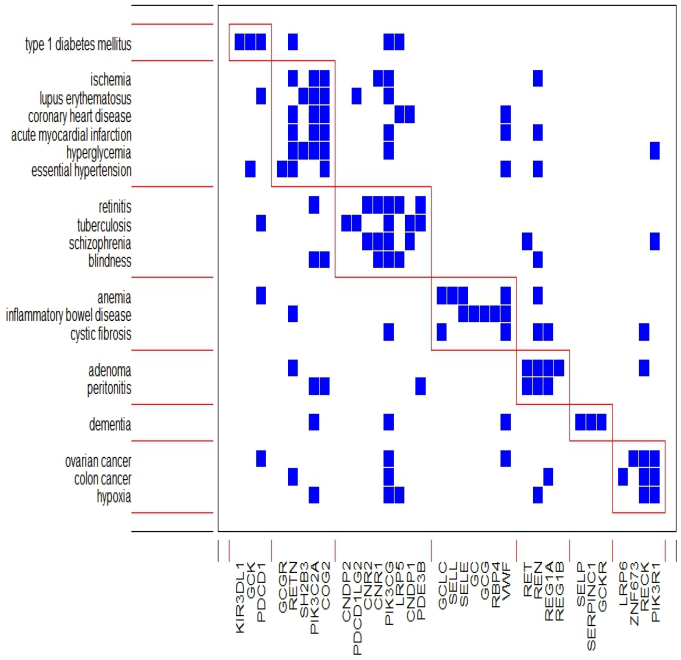

Fig. 8. Disease/gene functional modules with seven clusters identified

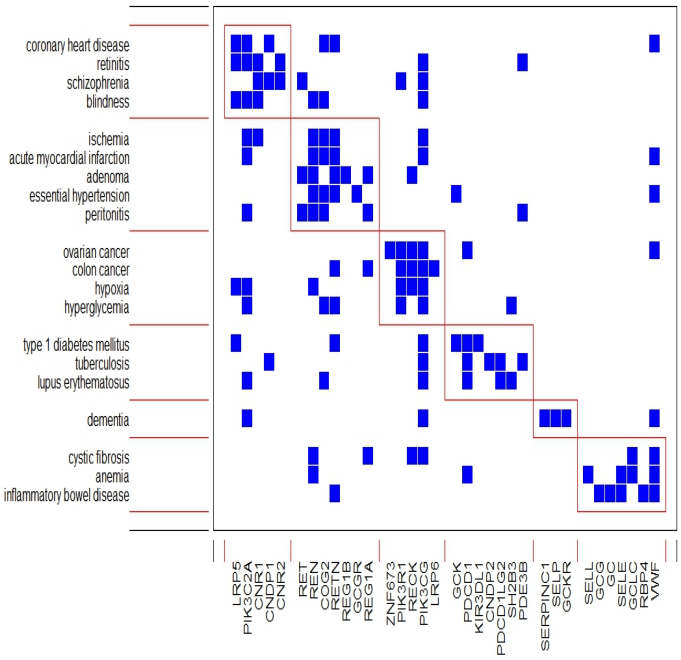

Fig. 9. Disease/gene functional modules with six clusters identified

We were interested in discovering the effects of increasing the number of related diseases on the connectivity of individual proteins in the bipartite graph. We therefore split the number of diseases into three groups of 300, 600, and 900. The number of genes remained the same (n=325).

Examining figure 10 containing the key proteins for 300 diseases we look to the cut off point of z $\geq$ 2.5 and can identify nine proteins that are important module hubs, there are also four further proteins that are borderline according to the definitions used by Olesen *et al* [27]. However, interestingly enough the PIK3C2A and PIK3CG proteins are not the highest scoring proteins and appear not to act as hubs with 300 diseases. Below the z-score cutoff point lie the vast majority of the other proteins which are classed as non-hubs. The protein called VEGFA (vascular endothelial growth factor A) is the highest scoring protein across all datasets, it has numerous roles in many cellular activities and pathways.

Examining figure 11 which contains the 600 diseases, we look to the cut off point of z $\geq$ 2.5 and can identify several hubs that were key proteins present in the first partition of the data. The PIK3C2A and PIK3CG proteins now appear in the quadrant of hub proteins. In figure 12 with 900 (full dataset) diseases the patterns do not change too much from 600 .

Computationally, the three sets of calculations took 40 minutes to compute the clusters for 300 diseases, 1.5 hours for 600 diseases and 18 hours for 900 diseases on a Intel Xenon CPU with dual processors (2.4GHz) and 24 GB of RAM. The R code was not compiled or optimized.

Finally, in table III we ran some statistics on the three bipartite networks, as a further measure we created random networks with the same number of nodes. Since we are examining bipartite networks, the top half of the table refers to the proteins and the bottom part refers to the diseases. We examined

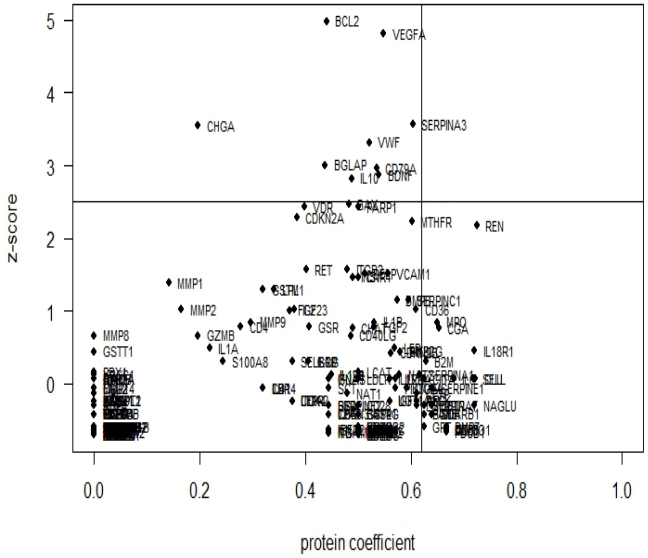

Fig. 10. CZ scatterplot 300 diseases: the upper left defines those proteins that are acting as *module hubs* based on a cut-off value of z $\geq$ 2.5 and for z $\leq$ 2.5 are *non-hubs*. The coefficient cut-off values for C are $\geq$ 0.62 and $\leq$ 0.62 accordingly, to determine if a node is either a *network hub* or a *connector*. 10 proteins have been identified as important hubs within this section of the dataset.

the average number of links per node; the averaged number of shared partner nodes; the clustering coefficient and the degree distribution. Unfortunately, computational difficulties

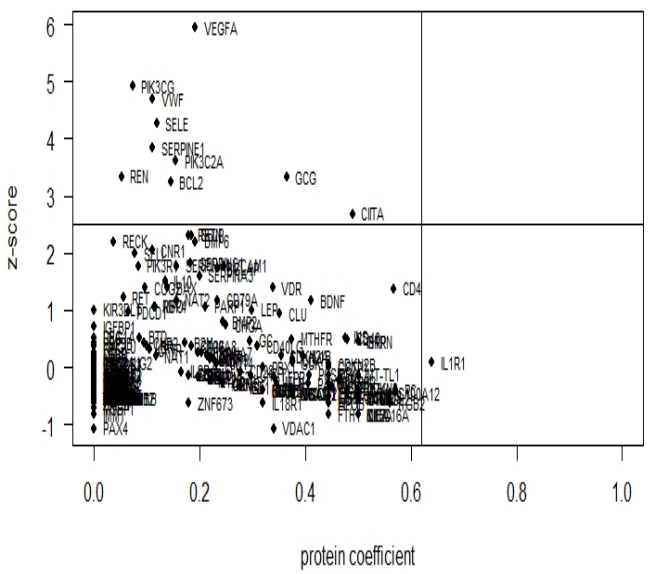

Fig. 11. CZ scatterplot 600 diseases:the upper left defines those proteins that are acting as *module hubs* based on a cut-off value of z $\geq$ 2.5 and for z $\leq$ 2.5 are *non-hubs*. 10 proteins have been identified with an important hub role within 600 diseases.

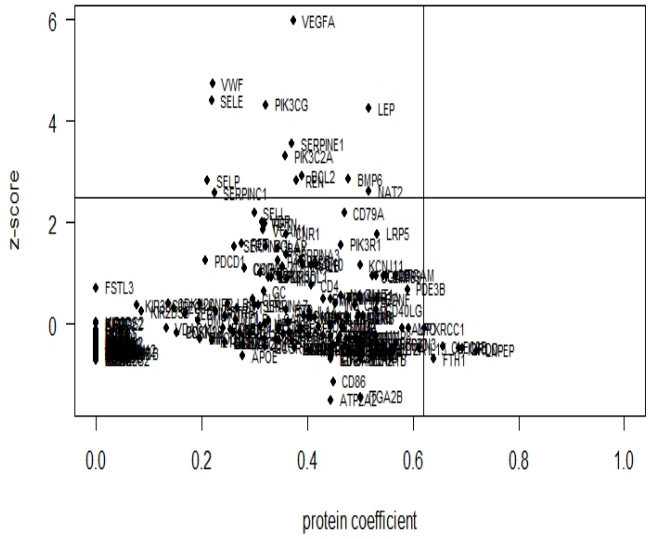

Fig. 12. CZ scatterplot 900 diseases:the upper left defines those proteins that are acting as *module hubs* based on a cut-off value of z $\geq$ 2.5 and for z $\leq$ 2.5 are *non-hubs*. With diseases present in the network, 13 proteins have been identified as being key hubs.

TABLE III. BIPARTITE STATISTICS ON THE THREE GRAPHS (300, 600, 900) AND RANDOM GRAPHS OF THE SAME DIMENSIONS (SAME NUMBER OF NODES PER DISEASE AND PROTEIN)

Measure (Proteins)	300	Ran300	600	Ran600	900
Ave No Links	2.77	197	4.17	438	6.22
Ave No Shared partners	8.57	5.29	1.79	9.37	0.25
Cluster Coeff	7.59	0.65	0.066	0.73	0.69
Degree distribution	1.284e-03	0.261	1.42e-03	0.21	0.013
Measure (Diseases)	300	Ran300	600	Ran600	900
Ave No Links	6.0	226	6.0	208	6.0
Ave No Shared partners	4.36	5.41	0.4	5.48	0.49
Cluster Coeff	0.018	0.67	0.018	0.64	0.018
Degree distribution	0.07	0.27	0.067	0.26	0.034

prevented the random 900 diseases network data from being generated. However, the trend is for a randomly generated network to have more links per node than a real network which are generally but not always sparsely connected.

IV. CONCLUSION

This paper presents our work in building networks of related diseases generated from a database of known protein to protein interactions and protein to diseases relationships. We have used bipartite graphs and clustering techniques to link diseases with their underlying genetic causes, in addition we have used graph structures and cluster diagrams that are useful in describing the modular nature of protein complexes. However, this work represents only a tiny fraction of the known disease network and of course there will be many protein interactions that are unknown and hence their contribution to disease remains uncertain. The landscape of known protein interactions is very dynamic, changing almost daily and our work can only represent a single snapshot. However, we find that in most cases diseases are caused by non-functional proteins and usually they cannot be cured but only alleviated. Sometimes a defective protein may have interacted with several other proteins in different pathways and signaling networks thus causing more than one illness. Additionally, side-effects occasionally have benefits, whereby a drug will unintentionally interact with more than the target protein. There is now an interesting trend by the pharmaceutical companies to reposition drugs to target diseases that they were not initially designed for, the most cited example is the Viagra drug developed by Pfizer. Certain orphan diseases have also benefited from this approach, that is to say diseases that were deemed economically not viable are now receiving attention from the drug companies. Our future work will concentrate on the STITCH and STRING databases which identify interacting proteins with chemicals with a view to indexing the DrugBank database for potential treatments.

ACKNOWLEDGMENT

The authors would like to thank Alex Kalinka for useful discussions regarding the linkcomm package and Solomon Messing for advice regarding the finer points of adjacency matrices. We also wish to thank the anonymous reviewers for their comments for improving the quality of the paper. The R code and data files can be made available upon application to the corresponding author.

REFERENCES

[1] Ott A, Stolk RP, van Harskamp F, Pols HA, Hofman A, and Breteler MM. *Neurology*, 53(9):1937–42, 1999.

[2] M. Baitaluk, X. Qian, S. Godbole, A. Raval, A. Ray, and A. Gupta. Pathsys: integrating molecular interaction graphs for systems biology. *BMC Bioinformatics*, 7(55):, 2006.

[3] AL Barabasi, N Gulbahce, and J Loscalzo. Network medicine: a network-based approach to human disease. *Nat Rev Genet*, 12:56–68, 2011.

[4] AL Barabasi and Z Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5:101–113, 2004.

[5] F. Barrenas, S. Chavali, P. Holme, R. Mobini, and M. Benson. Network properties of complex human disease genes identified through genome-wide association studies. *PLoS ONE*, 4(11):e8090, 11 2009.

[6] W. Bosl. Systems biology by the rules: hybrid intelligent systems for pathway modeling and discovery. *BMC Systems Biology*, 13(1):, 2007.

[7] He D, Liu ZP, and Chen L. Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach. *BMC Genomics.*, 12, 2011.

[8] Carsten F. Dormann and Rouven Strauss. A method for detecting modules in quantitative bipartite networks. *Methods in Ecology and Evolution*, 5(1):90–98, 2014.

[9] Belinky F, Bahir I, Stelzer G, Zimmerman S, Rosen N, Nativ N, Dalah I, Iny Stein T, Rappaport N, Mituyama M, Safran M, and Lancet D. Non-redundant compendium of human ncRNA genes in genecards. *Bioinformatics*, 29:255–261, 2013.

[10] L. Freeman. Centrality in social networks I: Conceptual clarification. *Social Networks*, 1:215–239, 1979.

[11] Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Lszl Barabsi. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[12] R. Guimer and L. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, 2004.

[13] G. Hu and P. Agarwal. Human disease-drug network based on genomic expression profiles. *PLoS ONE*, 4(8):e163, 2009.

[14] T. Ideker, O. Ozier, B. Schwikowski, and A. Siegel. Discovering regulatory and sigalling circuits in molecular interaction networks. *Bioinformatics*, 18(1):233–240, 2002.

[15] Alex Kalinka and Pavel Tomancak. linkcomm: an r package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. *Bioinformatics*, 27(14):2011 – 2012, 2011.

[16] Goh KI and Choi IG. Exploring the human diseasome: the human disease network. *Brief Funct Genomics.*, 11(6):533–42, 2012.

[17] S. Klamt, J. Saez-Rodriguez, J. Lindquist, L. Simoeni, and E. Gilles. A methodology for the structural and functional analysis of signalling and regulatory networks. *BMC Bioinformatics*, 7(56):, 2006.

[18] E. Klipp and W. Liebermeister. Mathematical modeling of intracellular signalling pathways. *Neuroscience*, 7(1), 2006.

[19] M. Koyuturk, A. Grama, and W. Szpankowski. An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics*, 20(1):200–207, 2004.

[20] Martin Lechner, Veit Hohn, Barbara Brauner, Irmtraud Dunger, Gisela Fobo, Goar Frishman, Corinna Montrone, Gabi Kastenmuller, Brigitte Waegele, and Andreas Ruepp. Cider: multifactorial interaction networks in human diseases. *Genome Biology*, 13(7):R62, 2012.

[21] D.-S. Lee, J. Park, K. A. Kay, N. A. Christakis, Z. N. Oltvai, and A.-L. Barabasi. The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences*, 105(29):9880–9885, 2008.

[22] W. Leibermeister and E. Klipp. Bringing metabolic networks to life: integration of kinetic, metabolic and proteomic data. *Theoretical Biology and Medical Modelling*, 42(3):1–15, 2006.

[23] K. McGarry. Discovery of functional protein groups by clustering community links and integration of ontological knowledge. *Expert Systems with Applications*, 40(13):5101–5112, 2013.

[24] K. McGarry, J. Chambers, and G. Oatley. Graph based analysis of protein interaction for diabetes research. *Artificial Intelligence in Medicine*, 41(2):129–144, 2007.

[25] K. Michael, D. Szklarczyk, A. Franceschini, C. von Mering, L. Jensen, Lars Juhl, and P. Bork. Stitch 3: zooming in on proteinchemical interactions. *Nucleic Acids Research*, 40(D1):D876–D880, 2012.

[26] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. Whole-proteome prediction of protein function via graph theoretic analysis of interaction maps. *Bioinformatics*, 21(1):302–310, 2005.

[27] Jens M. Olesen, Jordi Bascompte, Yoko L. Dupont, and Pedro Jordano. The modularity of pollination networks. *Proceedings of the National Academy of Sciences*, 104(50):19891–19896, 2007.

[28] J. Rachlin, D. Cohen, C. Cantor, and S. Kasif. Biological context networks: a mosaic view of the interactome. *Molecular Systems Biology*, 2(66):285–308, 2006.

[29] M. Schaub, T. Henzinger, and J. Fisher. Qualitive networks: a symbolic approach to analyze biological signaling networks. *BMC Systems Biology*, 4(1):, 2007.

[30] D. Scholtens, M. Vidal, and R. Gentleman. Local modeling of global interactome networks. *Bioinformatics*, 21(17):3548–3557, 2005.

[31] P. Shafer, T. Isganitis, and G. Yona. Hubs of knowledge: using the functional link structure in Biozon to mine for biologically significant entities. *BMC Bioinformatics*, 7(71):, 2006.

[32] S van Dieren, JW Beulens, YT van der Schouw, DE Grobbee, and B Neal. The global burden of diabetes and its complications: an emerging pandemic. *Eur J Cardiovasc Prev Rehabil*, 17(Suppl 1):S3–S8, 2010.

[33] R. Wanders. Metabolic and molecular basis of peroxisomal disorders: a review. *American Journal of Medical Genetics*, 126A:355–75, 2004.

[34] C. Yeang and M. Vingron. A joint model of regulatory and metabolic networks. *BMC Bioinformatics*, 332(7):1–5, 2006.

[35] E. Zotenko, K. Guimaraes, R. Jothi, and T. Przytycka. Decomposition of overlapping protein complexes: a graph theoretical method for analyzing static and dynamic protein associations. *Algorithms for Molecular Biology*, 1(7), 2006.