Complex network based computational techniques for ‘edgetic’ modelling of mutations implicated with human diseases

Ken McGarry, Kirsty Emery, Vithusa Varnakulasingam, Sharon McDonald and Mark Ashton
Department of Pharmacy, Health and wellbeing, Faculty of Applied Sciences, University of Sunderland, UK
email:ken.mcgarry@sunderland.ac.uk

***Abstract*—Complex networks are a graph theoretic method that can model genetic mutations, in particular single nucleotide polymorphisms (snps) which are genetic variations that only occur at single position in a DNA sequence. These can potentially cause the amino acids to be changed and may affect protein function and thus structural stability which can contribute to developing diseases. We show how snps can be represented by complex graph structures, the connectivity patterns if represented by graphs can be related to human diseases, where the proteins are the nodes (vertices) and the interactions between them are represented by links (edges). Disruptions caused by mutations can be explained as loss of connectivity such as the deletion of nodes or edges in the network (hence the term *edgetics*). Furthermore, diseases appear to be interlinked with hub genes causing multiple problems and this has led to the concept of the human disease network or *diseasome*. Edgetics is a relatively new concept which is proving effective for modelling the relationships between genes, diseases and drugs which were previously considered intractable problems.**

***Keywords*—complex networks, hubs, nearness, betweeness.**

I. Introduction

Many human diseases often have a genetic cause, that is to say the gene or genes responsible for encoding a specific biological function have become defective. This may be caused by some sort of single-nucleotide substitution (mutation) that causes the gene to produce a protein that can no longer interact with other proteins and elements in the usual way. Protein interactions are key to the majority of functions occurring in the cell and also account for several signaling mechanisms for processes external to the cell. The connectivity of interacting proteins (*interactome*) when mapped as a network reveals a complex web of relationships. Some proteins have many connections while others are sparsely connected, however applying computational techniques such as clustering can also reveal the modular nature of proteins as they cooperate in various activities [1], [2]. Researchers have modified graph theoretic methods to tackle the issues inherent with protein interaction networks, or have created novel statistical methods able to predict protein function [3], [4], or to model subgraphs using a mixture clustering and classification methods [5], [6], [7]. These computational techniques are essential to unraveling the complex nature of genes, proteins and their relationship with diseases.

Our interest in protein interactions is concerned with their network structure and how this is related to human diseases, which can be explained as loss of connectivity such as the deletion of nodes or edges in the network (hence edgetics). Furthermore, diseases appear to be interlinked with hub genes that can cause multiple problems when they become defective, the human disease network or *diseasome* [8] is now receiving attention as a means of understanding how diseases occur. The goal is to develop new drug products to tackle and combat diseases and perhaps reposition existing drugs to new targets [9], [10], [11].

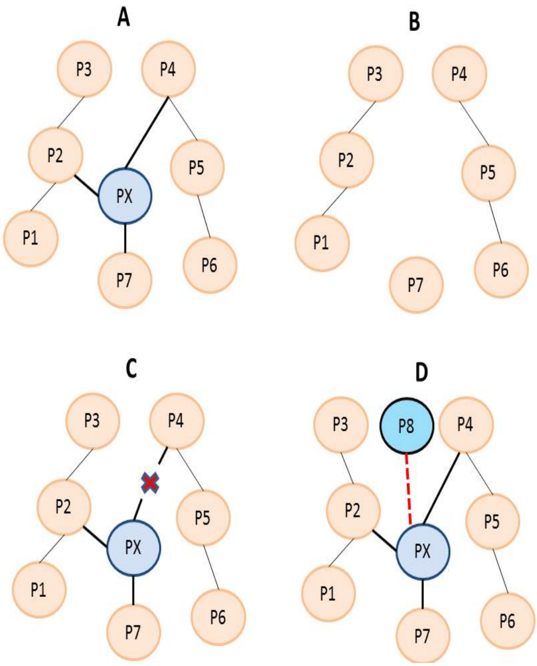


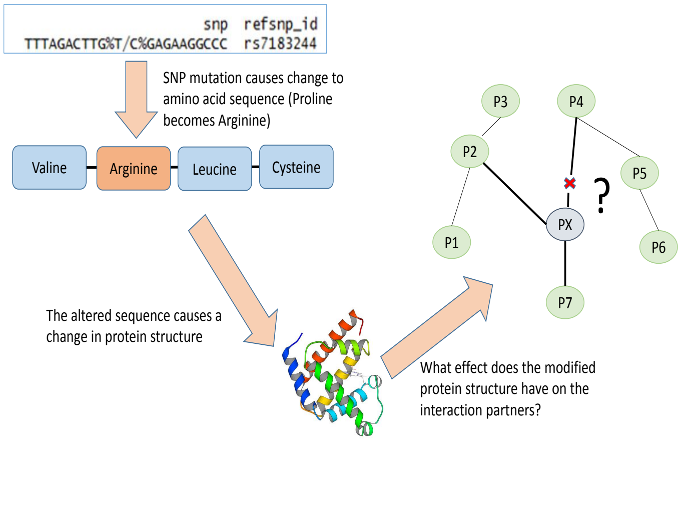
Fig. 1. Example of mutation causing node and edgetic perturbations, showing in A the wild type or normal protein interactions between protein X and it’s partners, in B we have node deletion of protein X with complete loss of interactions. In C we have the edgetic removal of a link between protein X and P4-P5-P6 with partial or perhaps even complete loss of functionality. In D we have a gain-of function through an additional edge and a new interacting protein P8.

Edgetic analysis was first proposed by Zhong as a method of explaining certain disease causing mutations by a loss of network connectivity on key genes[12], [13]. The aim is to improve our understanding of the genotype-to-phenotype relationship, that is from genes to the physical shape and wellbeing of the individual (phenotype). However, from table **??** can be seen the complexity of the situation, sometimes it is not a one-gene to one-function relationship and hence one-disease. Variants of the same gene can cause different functional defects (allelic heterogeneity), however the same disease can be caused by mutations in different genes (genetic heterogeneity) [14], [15]. The benefits from this analysis would hopefully provide the knowledge for developing new treatments [16], [17], [18] and the potential for repositioning existing drugs to other diseases[19], [20].

A. *The role of mutations on disease*

The role of genetics dominates almost all of human diseases, even those where the environment does play a significant factor. Many diseases such as cystic fibrosis can be identified with a single defective gene and as such are described as Mendelian because they follow Mendel's law of inheritance. Other diseases do not follow Mendel's law and are the result of the interaction between several genes and environmental factors, however these are classed as complex diseases and do not fall into our area of study.

In this study we are interested in Single Nucleotide Polymorphisms (SNP) that affect protein structure, this where a single letter of the genetic code, called nucleotides (consisting of four letters: A, C, G, T) are changed by a mutation. For example an A can be changed to a C, this may or may not results in a mutation. The genetic code operates using triplets of letters (codons) that code for a specific amino acid, proteins are constructed out of long chains of amino acids, e.g. the triplet CGT encodes for Arginine. The genetic code is redundant so many mutations (especially in the 3rd position of codons) will not change the encoded amino acid, so keeping with our Arginine example CGT, CGC, CGA, CGG will all encode this amino acid. However, changing CGG to CCG will result in Proline substituted for Arginine, this may or may not result in damaging changes to the protein. There are 20 amino acids which form the majority of proteins and enzymes in the human body. In figure 2 we show the sequence of events, from mutation to change of protein structure.



```
                       snp  refsnp_id
TTTAGACTTG%T/C%GAGAAGGCCC  rs7183244
```

SNP mutation causes change to amino acid sequence (Proline becomes Arginine)
Valine
Arginine
Leucine
Cysteine
The altered sequence causes a change in protein structure
What effect does the modified protein structure have on the interaction partners?
P3
P4
P2
P5
?
PX
P1
P6
P7

Fig. 2. Predicting the effects of structural change on protein-to-protein interaction partners. An SNP changes from a ”T” to a ”C”, encoding for a different amino acid.

SNPs are classified based on the region of the gene they are located. The introns or non-coding region of the DNA, do not generally affect the protein as this sequence is not translated (the most prevalent type of SNP). The exon or coding region of a gene is more problematic as these sequences are transcribed into protein, they are identified as either:

- Synonymous SNPs, the change to the codon will still code for the same amino acid, protein remains unchanged.
- Non-synonymous SNPs, the change to the codon results in amino acid substitution and therefore changes the protein. Non-synonymous SNPs can also subdivided by the type of mutation they generate:
  a) Missense mutations, the SNP will code for a different amino acid.
  b) Nonsense mutation, the SNP encodes a “stop” command, thus prematurely ending the production of further protein. Nonsense mutations can be particularly damaging if located at the beginning or middle of a gene.

The aim of this work is to identify the type of SNP's, their frequency of occurrence on disease and non-disease genes and how this will effect their protein interactions, thus increasing the risk of contracting cardiovascular disease (CVD). We also aim to determine any trends or patterns which can indicate those SNPs likely to cause a greater risk of cardiovascular events. The remainder of this paper is structured as follows; section two describes our methods, indicating the types of data used such as the SNP short DNA sequences and the formation of protein interaction networks; section three describes how graph theory, clustering and the other computational techniques can manage this data, section four highlights the results and finally section five presents the conclusions and future work.

II. METHODS

In figure 3 we present the overall system operation, along with the flow of information and its transformation. We extracted the SNP data from the esembl database which contains the sequences and point information. The known disease proteins and their interaction partner proteins were downloaded from the STRING database [21]. The STRING database contains approximately six million known protein interactions generated by text-mining, annotation by experts and through statistical prediction. Each protein pair contains a confidence score based on the interaction source, text-mining for example has a much lower score than annotation by experts.

The system was implemented using the R language with the RStudio programming environment, on an Intel Xenon CPU, 64-bit with dual processors (3.2GHz) and 128 GB of RAM. The following R packages were used: BiomaRT to download the required data from ensembl database [22]; SeqInR was used to convert DNA sequences into amino acid chains [23]. Other packages including, ggplot2, dplyr, tidyr, igraph, stringr and notably Peptides for assessing the chemical properties of the proteins. The R code was not compiled or optimized, as a general rule, R is generally quite slow compared with a compiled language. Our R code and data files are freely available on GitHub for download: https://github.com/kenmcgarry/Edgetics

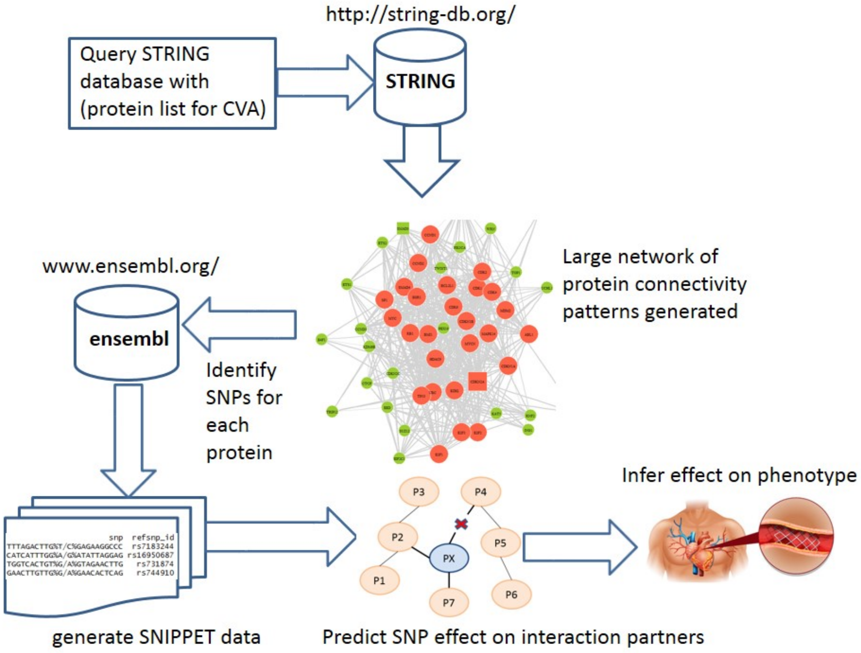

http://string-db.org/
Query STRING database with (protein list for CVA)
STRING
Large network of protein connectivity patterns generated
www.ensembl.org/
ensembl
Identify SNPs for each protein
snp refsnp_id
TTTAGACTTG%T/C%GAGAAGGCCC rs7183244
CATCATTTGG%A/G%ATATTAGGAG rs16950687
TGGTCACTGT%G/A%GTAGAACTTG rs731874
GAACTTGTTG%G/A%GAACACTCAG rs744910
P1
P2
P3
P4
P5
P6
P7
PX
Infer effect on phenotype
generate SNIPPET data
Predict SNP effect on interaction partners

Fig. 3. The overall system operation, data sources and transformations.

A. SNIPPETS

The basic data unit we manipulate is a 20 length DNA sequence upstream and downstream of the SNP location, this will encode around the SNP the amino acids. We observe the normal (wildtype) DNA and compare with the SNP (mutated) version. We call these small lengths of DNA, SNIPPETS. In table I the snp, the allele (the mutation) are shown for four SNPs.

Thus, we investigate if single point SNP mutations cause changes to the amino acid sequence, and if these transformations cause a change in protein function and hence interactions with other proteins. Protein interactions are essentially stable or transient, and both types are characterized as either strong or weak. Stable interactions are linked with proteins that are formed from multi-unit complexes. Transient interactions are involved in the majority of cellular processes that govern the cell. They are short lived and often depend on a set of initial conditions that trigger the interaction. Proteins interact or bind to each other through various methods such as van der Waals forces, hydrophobic bonding and salt bridges at specific locations or domains on each protein. The sites can be small binding clefts or large surfaces and can be just a few amino acids long or comprised of several hundred, the connection strength of the binding is moderated by the size of the binding domain.

B. Limitations of the study

The study of protein interactions and the analysis of such databases is a highly dynamic landscape, new protein interactions are continuously identified along with potential disease causing mutations. Bias is also present because disease genes tend to be more studied than others, for example the gene TP53 (notorious for involved in cancer as well as the study presented in this paper) has 1098 protein interactions, 2962 SNP's identified and 11,283 scientific papers written about it (as of April 2016).

C. Related work

The current algorithms fall into either the machine learning camp using classified data or use heuristics based on theoretical models. The MutationTaster [24] and PolyPhen [25] algorithms are examples of the machine learning based approach and use examples of known SNP mutations which are damaging or benign . They take the sequences of the known mutations as training data. The SIFT [26] and MutationAssessor [27] algorithms create models using scoring matrices based on the mutations position They use sequence alignment and score the mutations based on how well the position is conserved for such criteria as polarity, charge and other chemical properties.

III. The Relationship between Graph Theory and Interactome Modelling

Graph theory or complex networks as it is more commonly now called is a set of mathematical principles that describe the structure, relationships and topology of many real-world situations. We find that social networks such as Twitter and FaceBook, ecological networks of predator-prey situations and econometric networks such as supply and demand all have similar properties, such as a graph-like structure that can be described with a formal language. The graph network is simply a set of nodes or entities connected by links that define the relationships and hierarchy between the nodes. The frequency of node connectivity, either their sparseness or abundance provides useful information as to their importance or redundancy.

TABLE I. SNP'S FOR THE GENE, SMAD3 AND SNIPPET INFORMATION.

	snp	refsnpid	chrom start	allele	chrome name
1	TATGCGCCCTTTTAGACTTG%T/C%GAGAAGGCCCCTTGGACTTC	rs7183244	67168973	T/C	15
2	TGAAGAAACTCATCATTTGG%A/G%ATATTAGGAGATGCTTGAAA	rs16950687	67171675	A/G	15
3	GCAGAGCACATGGTCACTGT%G/A%GTAGAACTTGCAGTGAGACC	rs731874	67154493	G/A	15
4	CACACTTACAGAACTTGTTG%G/A%GAACACTCAGGAAACTCAGC	rs744910	67154447	G/A	15

To manage networks in a formal way we use graph theoretic methods which can be applied to any network of interacting entities are linked together through various relationships. Graph creation and inferencing is usually through matrix algebra, edge lists are converted into connectivity matrices, we can define a graph G = (V, E) where the nodes also called vertices V containing links called edges E. The links can be either directed, that is to say information implying direction or causality in the relationship is available in the sense that A causes B. Links can also be undirected, implying that we only know there is a relationship between the nodes but are unable to specify the sequence of events or perhaps this information is unnecessary. In our application, we determine the relevance of protein connectivity patterns using criteria from the graph theoretic centrality statistics, see equations 1-5 [28], [29], [30].

A. Identification of hub proteins with centrality measures

Identifying "hub" proteins within the network is often the first task, hubs tend to have important connections with other key proteins so the deletion/disruption of a hub protein may be more problematic than deletion of a non-hub protein [31]. The hub protein may participate in several cellular functions and this observation is confirmed in many protein networks which are typically small world networks. This characteristic is called a power law degree distribution, which manifests itself as a susceptibility to the removal of certain proteins [32], [33], thus proteins implicated in lung cancer for example typically have twice as many interaction partners as non disease related proteins [34].

We use a number graph based measures to evaluate the protein networks. The closeness statistics provides a measure of how close each node is to every other node in the network. Some nodes may be more prominent than others due to their topology.

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

The betweeness statistic calculates the extent that a node is located between pairs of other nodes in the network, such that a path between the other nodes has to go through that node.

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

The clustering coefficient provides a measure of modularity of the network in terms of shared components.

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

Where $V = v1, v2...vn$ define the n vertices or nodes and E the collection of edges or connections, where e_{ij} indicates an edge (E) connecting vertices v_i, the term v_j k_i represents the vertex (V) neighbourhood. The neighbours N, for a given vertex (V) v_i, is its closely connected neighbours :

$$N_i = \{v_j\} : e_{ij} \in E \tag{4}$$

The in-degree and out-degree of each the vertices (in undirected graphs it is just overall degree) k_i corresponds to the number of vertices in its near neighbourhood $|Ni|$. Calculating the clustering coefficient C_i for each vertex v_i the ratio of links between the vertices within its near neighbourhood partitioned by the total links that potentially could occur between them. Furthermore, graphs that are undirected have the characteristic that e_{ij} and e_{ji} are considered equivalent.

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

A value of unity is returned when all vertices connected to v_i are also linked to all other vertices within the neighbourhood n, and returns zero when no vertex connected to v_i links to all other vertices that connect to v_i.

IV. RESULTS

A list of 13 known proteins involved in CVD was derived from the literature, these protein names were uploaded to STRING database and their interaction partner proteins were downloaded. The Biomart database was searched for SNPs attributed to each protein (along with mutation type including upstream and downstream sequence data). Graph network statistics were calculated for each important protein.

We first calculated some metrics to assess bias in our data, the disease causing proteins are more likely to be reported in the literature than those with no apparent association. Hence these are expected to have more interaction partners identified, figures 4 and 5 indicate this.

Overall graph statistics for the combined network is presented in table II. One statistics to note is the *connected* parameter, this implies that the network of disease proteins and their interacting protein partners are not fully connected - that is to say CVD group of proteins are not fully connected there are seven isolated subnetworks. The largest contains 60% of the nodes (79 out of 130), while the others each individually account for 8%.

TABLE II. OVERALL GRAPH STATISTICS MEASURES FOR THE ENTIRE NETWORK

modularity	avepath	nedges	nverts	transit	avedegree	diam	connect
0.47	2.32	2594.00	225	0.54	23.06	7.00	FALSE

In table III the statistics for each individual protein in the network is presented. The information in this table is sorted

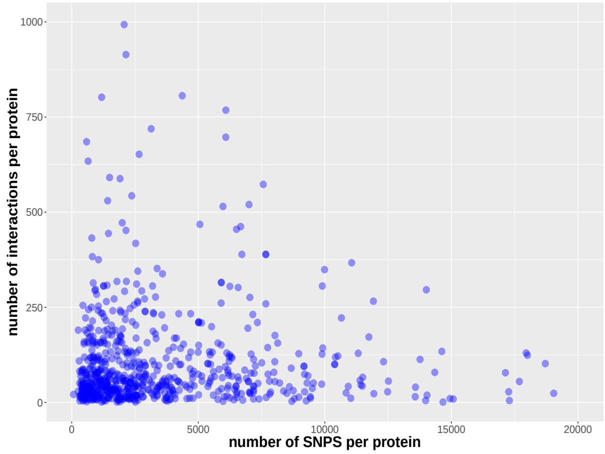


Fig. 4. Number of SNPs versus protein interactions

according to the *hubness* criteria, therefore protein SMAD3 is the highest scoring and highly connected protein. Followed by CDKN2A, however out of the disease proteins there are only two to three that may be considered hubs.

TABLE III. GRAPH STATISTICS MEASURES FOR KNOWN DISEASE PROTEINS

	hubness	closeness	betweenness	authority
CXCL12	0.72	5.11E-05	597.174	0.7370
SMAD3	0.05	2.76E-05	2525.578	0.0620
CDKN2A	0.05	4.85E-05	231.734	0.0029
CDKN2B	0.01	2.19E-05	27.651	0.0223
PDGFD	0.00	3.62E-05	0.000	-0.0000
PLTP	0.00	2.04E-05	0.000	-0.0000
LIPA	0.00	1.98E-05	0.000	0.0003
CETP	0.00	1.98E-05	3.119	0.0002
FMN2	0.00	1.98E-05	0.000	0.0000
HSPE1	0.00	1.98E-05	169.577	0.0035

In table IV we have the highest scoring proteins based on hubness, these range from 1.0 to 0.92. The cutoff point for hubness is 0.8 and we have 16 proteins that match this criteria, these are also displayed in figure 6.

TABLE IV. GRAPH STATISTICS MEASURES FOR TOP RANKING PROTEINS BASED ON HUBNESS

	hubness	closeness	betweenness	authority
CCR3	1.00	9.40E-05	41.842	0.1107
MTRNR2L2	1.00	1.03E-04	32.477	0.0000
S1PR2	0.99	1.00E-04	238.597	0.0378
HCAR3	0.97	6.34E-05	0.000	0.1486
TAS2R43	0.96	6.25E-05	0.000	0.1852
CASR	0.96	6.08E-05	28.856	0.2574
CXCL2	0.95	6.17E-05	0.000	0.2214
RGS6	0.94	9.59E-05	0.000	0.0753
ADCY5	0.93	6.00E-05	42.392	0.2582
CHRM4	0.92	5.92E-05	1.500	0.3288

In figure 6 we display the entire network of 225 proteins and the 2,594 interactions between them. In diagrams such as these it is impossible to get any sense of the connection patterns for anything more then a dozen nodes. Instead we concentrate on the hubness parameter, those proteins with a value of 0.8 are classed as hubs and are coloured in orange, all other less than this cutoff point are light green. The 12 disease proteins are squared shaped while all others are circles, not all disease proteins are hubs.

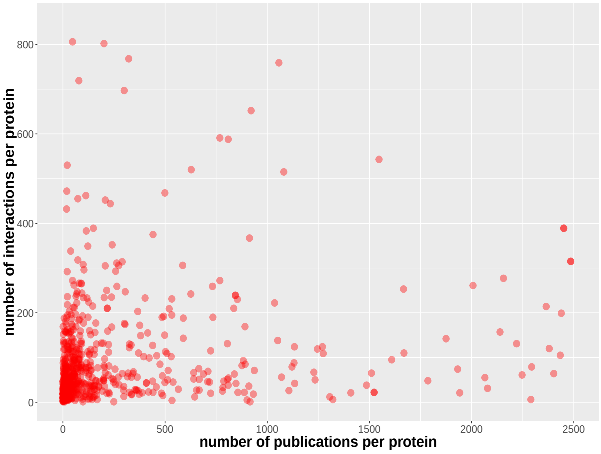


Fig. 5. Number of publications versus protein interactions

CDKN2A in visceral tissues relates to the increase in atherosclerosis. The gene plays a role in various other processes including, signal transduction (FoxO signaling, TGF-beta signaling), cellular processes (cell growth and death) and human diseases (cancer pathways, viral carcinogenesis and small cell lung cancer).

Other proteins that we know about but our graph based analysis failed to uncover are FMN2 and MTHFD1L. The FMN2 (Formin-2) is a gene located on chromosome 1q43 at position 240282296 , which has been reported in literature as part of a wide association study, where it is believed to be linked to coronary heart disease. FMN2 is involved in other pathways or processes including, organismal systems. Isolating the FMN2 from the main super network and recalculating the statistics enabled further insights to be gained. Figure 7 shows the connectivity.

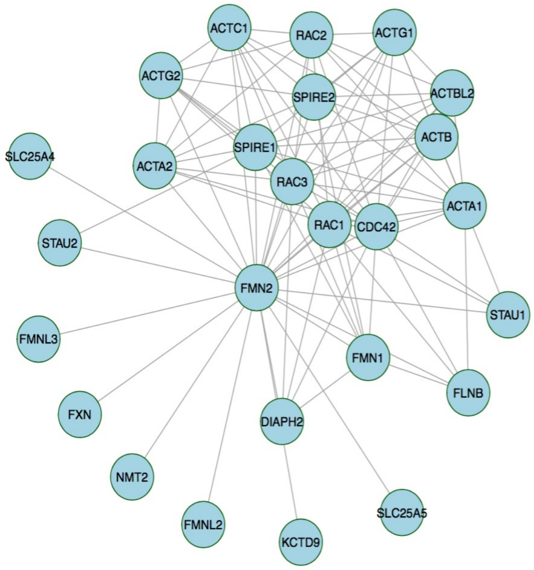


Fig. 7. The FMN2 subnetwork.

Based on the statistics gained from table V we can deter-

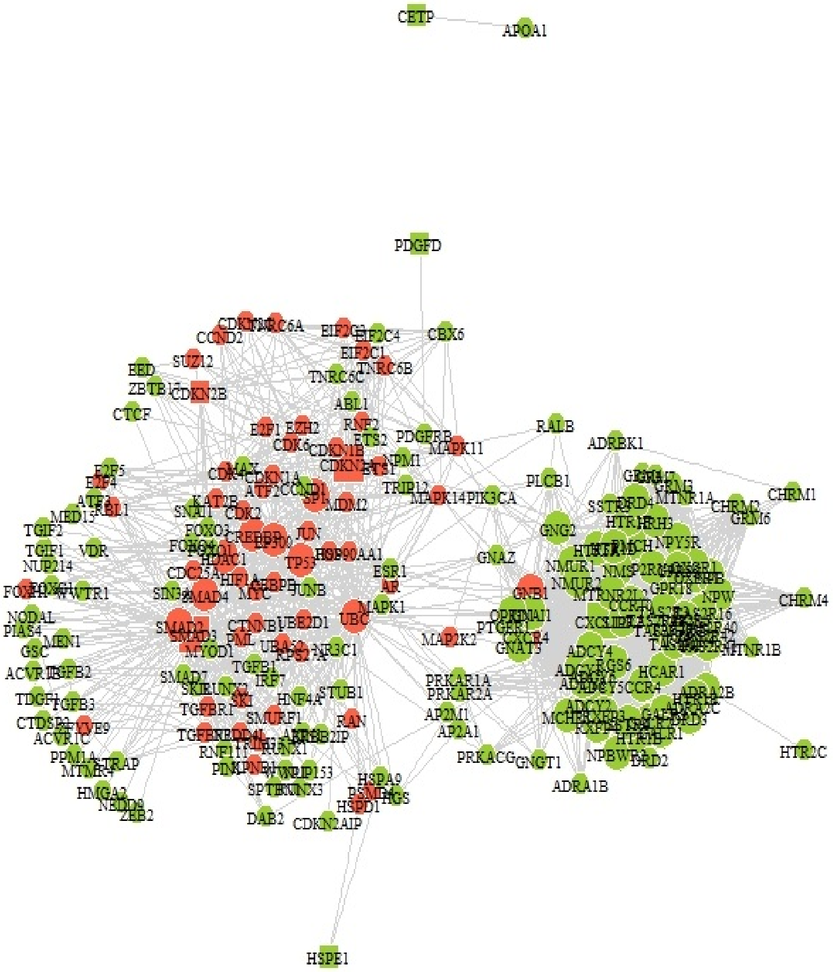


Fig. 6. Interactions between the networks 217 proteins.

mine that FMN2 is an important protein in its own right and must participate in other signal cascades.

TABLE V. PROTEIN INTERACTION NETWORK OF FMN2

	Closeness	Betweenness	Hubness	Authority
ACTA1	0.025	2.77777778	0.626466	0.626466
RAC1	0.02777778	10.57738095	0.778112	0.778112
SPIRE1	0.025	6.911071429	0.553964	0.553964
DIAPH2	0.02222222	0.08333333	0.35453	0.35453
FMN2	0.04	173.9384921	1	1
ACTB	0.025	3.31944444	0.626993	0.626993

Also, the MTHFD1L (methylenetetrahydrofolate dehydrogenase) is a gene located on chromosome 6 involved in the synthesis of tetrahydrofolate (THF) in the mitochondrion and metabolism of cofactors and vitamins (11). Single nucleotide polymorphisms (SNPs) in MTHFD1L, including the lead polymorphism (rs69222269), are known to be implicated with coronary heart disease (CAD). The functional effect of the leading polymorphism (rs69222269) is unknown, a likely mechanism for the association to CAD, may be related to the effects on the folate metabolic pathway.

The next stage is to examine the effect of the SNPs on protein compositions, we need to monitor the chemical properties based on the DNA to amino acid changes. Examining the CDKN2A protein we use the following SNPs. In table VI we have highlighted in grey those parameters for the Physiochemical properties that have changed between the wildtype and mutant.

TABLE VI. PHYSIO-CHEMICAL PROPERTIES FOR CDKN2A PROTEIN (SNP RS11552822)

Physiochemical properties	Wildtype	Mutant
Aliphatic Index	22.271293	22.271293
Boman Index	1.351167	1.340442
Charge	16.057535	16.057535
Instability Index	78.448428	76.272327
Peptide Length	317	317

The Aliphatic index is the relative volume occupied by aliphatic side chains, found on alanine, valine, isoleucine and leucine. It is a positive factor for thermostability of globular proteins, so if the aliphatic index is decreased the thermostability is reduced. Less stability could lead to changes in structure (denaturation) meaning different interactions.

The Boman Index indicates the binding potential of a protein, and can be used to predict multifunctionality. The higher the boman index the more like a protein is to interact with other proteins.

The Theoretical Net Charge can account for protein-protein repulsion of attraction, (more negative more repulsion). Changes in charge may affect protein-protein interactions based on charge/charge interactions. Charge calculations based on primary amino acid sequence do not factor in the 3D structure of proteins, in which some amino acids can be buried or exposed in the center of the protein. The pKas of amino acids within proteins are also influenced by different interactions which are highly protein dependent, and not account for in net charge calculations, so this may not be the best indicator of a chemical property that can affect protein-protein interactions.

The instability index can be used as a measure of the in-vivo half-life of a protein. A value of >40 means a half-life of <5h, meaning less stability, whereas <40 indicates a higher stability with half-lives of >16h. Longer lifetimes for partially folded intermediates may influence the aggregation of intermediates as there is a greater chance of interaction between proteins, and more exhaustion of molecular chaperones.

In table VII we have highlighted in grey those amino acid parameters that have changed between the wildtype and mutant.

TABLE VII. AMINO ACID COMPOSITION FOR CDKN2A PROTEIN (SNP RS11552822)

	Number	% Mole	Number	% Mole
Tiny	77	24.29	76	23.975
Small	102	32.177	101	31.861
Aliphatic	35	11.041	35	11.041
Aromatic	11	3.47	11	3.47
Non-Polar	76	23.975	76	23.975
Polar	83	26.183	82	25.686
Charged	34	10.726	34	10.726
Basic	27	8.517	27	8.527
Acidic	7	2.208	7	2.208

Differences in the amino acid composition (table VII) could lead to changes in protein structure, leading to new/different interactions partners. The position of change could be involved in ligand binding, disulphide bridging or other protein-protein interactions site, causing changes to such interactions. Changes in polar or hydrophobic residue containing amino acids would be expected to be less harmful to interactions as these are usually tucked in the center of the protein, although could still change the 3D conformation of the protein.

Most of the changes for our chosen proteins are fairly negligible in terms of these statistics, if there are any at all. Could not say whether they would affect the protein-protein interactions or cause any loss of functions.

Of the many challenges in predicting protein-protein interaction parameters, are the surface areas that are involved in many such interactions which typically are of the order of several hundred Angströms and hence it is often the case that a single amino acid change has limited impact on the global binding interaction energy.

TABLE VIII. PHYSIO-CHEMICAL PROPERTIES FOR CDKN2A PROTEIN (SNP RS100586)

Physiochemical properties	Wildtype	Mutant
Aliphatic Index	32.9968454	32.9968454
Boman Index	0.57091478	0.5883912
Charge	6.9662124	6.9662124
Instability Index	89.5566038	91.108805
Peptide Length	317	317

TABLE IX. AMINO ACID COMPOSITION FOR CDKN2A PROTEIN (SNP RS100586)

	Wildtype		Mutant	
	Number	% Mole	Number	% Mole
Tiny	57	17.981	57	17.981
Small	111	35.016	110	34.7
Aliphatic	43	13.565	43	13.565
Aromatic	9	2.839	9	2.839
Non-Polar	105	33.123	104	32.808
Polar	54	17.035	55	17.35
Charged	22	6.94	22	6.94
Basic	16	5.047	16	5.047
Acidic	6	1.893	6	1.893

V. CONCLUSION

In recent years we have all become accustomed to the discovery of new genes and their alleged roles in causing specific diseases. All humans have these genes, however it is the mutated version that actually causes the problems. Thus the role of mutations is pivotal in defining how the disease will manifest itself and how the individual may be affected. The various computational methods now being explored to study edgetics will play an important role in understanding the etiology of diseases. Our ongoing work is exploring the impact of interacting proteins being located on the same chromosome, along locality in the same tissues and organs. Locality and modularity may make protein interactions more likely to occur, giving more information to improve prediction of type and severity of diseases.

ACKNOWLEDGMENT

The authors would like to thank Lewis Bingle for helpful discussions on SNP issues and David Tang for help with the biomaRt package. This work was funded by the University of Sunderland's Health Sciences and Wellbeing Research Beacon.

REFERENCES

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

[2] K. McGarry, "Discovery of functional protein groups by clustering community links and integration of ontological knowledge," *Expert Systems with Applications*, vol. 40, no. 13, pp. 5101–5112, 2013.

[3] 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*, vol. 21, no. 1, pp. 302–310, 2005.

[4] 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*, vol. 7, no. 71, p. , 2006.

[5] A. Lee, L. Ming-Chih, and C. Hsu, “Mining dense overlapping subgraphs in weighted proteinprotein interaction networks.” *BioSystems*, vol. 103, pp. 392 – 399, 2011.

[6] M. Koyuturk, A. Grama, and W. Szpankowski, “An efficient algorithm for detecting frequent subgraphs in biological networks,” *Bioinformatics*, vol. 20, no. 1, pp. 200–207, 2004.

[7] 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*, vol. 7, no. 56, p. , 2006.

[8] 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*, vol. 4, no. 11, p. e8090, 11 2009.

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

[10] G. Hu and P. Agarwal, “Human disease-drug network based on genomic expression profiles,” *PLoS ONE*, vol. 4, no. 8, p. e163, 2009.

[11] K. McGarry, A. Rashid, and H. Smith, “Computational methods for drug repositioning,” *Drug Target Review*, vol. 3, pp. 31–33, 2016.

[12] N. Sahni, Y. Song, Q. Zhong, N. Jailkhani, B. Charloteaux, M. Cuisick, and M. Vidal, “Edgotype: a fundamental link between genetype and phenotype.” *Genetics and Development*, vol. 23, pp. 649–657, 2013.

[13] Q. Zhong and N. Simonis, “Edgetic perturbation models of human inherited disorders.” *Molecular Systems Biology*, vol. 5, no. 321, pp. 1–10, 2009.

[14] A. Bauer-Mehren, M. Bundschus, M. Rautschka, M. Mayer, F. Sanz, and L. Furlong, “Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases,” *PLoS ONE*, vol. 6, no. 6, p. e20284, 06 2011.

[15] M. Constanzo, A. Baryshnikova, C. Nislow, B. Andrews, and C. Boone, “You too can can play with an edge,” *Nature Methods*, vol. 6, no. 11, pp. 797 – 798, 2009.

[16] A. Pujol, R. Mosca, J. Farrs, and P. Aloy, “Unveiling the role of network and systems biology in drug discovery,” *Trends in Pharmacological Sciences*, vol. 31, no. 3, pp. 115 – 123, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165614709002041

[17] X. Ma, L. Gao, and K. Tan, “Modeling disease progression using dynamics of pathway connectivity,” *Bioinformatics*, vol. 30, no. 16, pp. 2343–2350, 2014. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/30/16/2343.abstract

[18] M. Vazquez, A. Valencia, and T. Pons, “Structure-ppi: a module for the annotation of cancer-related single-nucleotide variants at proteinprotein interfaces,” *Bioinformatics*, vol. 31, no. 14, pp. 2397–2399, 2015. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/31/14/2397.abstract

[19] K. McGarry and U. Daniel, “Data mining open source databases for drug repositioning using graph based techniques,” *Drug Discovery World*, vol. 16, no. 1, pp. 64–71, 2015.

[20] K. McGarry, N. Slater, and A. Amaning, “Identifying candidate drugs for repositioning by graph based modeling techniques based on drug side-effects,” in *The 15th UK Workshop on Computational Intelligence, UKCI-2015*, University of Exeter, UK, 7th-9th September 2015.

[21] K. Michael, D. Szklarczyk, A. Franceschini, C. von Mering, L. Jensen, L. Juhl, and P. Bork, “Stitch 3: zooming in on proteinchemical interactions,” *Nucleic Acids Research*, vol. 40, no. D1, pp. D876–D880, 2012.

[22] S. Durinck, P. Spellman, E. Birney, and W. Huber, “Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package,” *Nature Protocols*, vol. 4, pp. 1184–1191, 2009.

[23] D. Charif and J. Lobry, “SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis.” in *Structural approaches to sequence evolution: Molecules, networks, populations*, ser. Biological and Medical Physics, Biomedical Engineering, U. Bastolla, M. Porto, H. Roman, and M. Vendruscolo, Eds. New York: Springer Verlag, 2007, pp. 207–232, ISBN : 978-3-540-35305-8.

[24] I. Adzhubei, S. Schmidt, and L. Peshkin, “A method and server for predicting damaging missense mutations.” *Nature Methods*, vol. 7, no. 4, pp. 248–249, 2010.

[25] J. Schwarz, C. Rdelsperger, M. Schuelke, and D. Seelow, “Mutationtaster evaluates disease-causing potential of sequence alterations,” *Nature Methods*, vol. 7, pp. 575–576, 2010.

[26] P. Ng and S. Henikoff, “Sift: predicting amino acid changes that affect protein function.” *Nucleic Acids Research*, vol. 31, no. 13, pp. 3812–3814, 2003.

[27] B. Riva, Y. Antipin, and C. Sander, “Predicting the functional impact of protein mutations: application to cancer genomics.” *Nucleic Acids Research*, vol. 39, no. 17, pp. 1–14, 2011.

[28] L. Freeman, “Centrality in social networks I: Conceptual clarification.” *Social Networks*, vol. 1, pp. 215–239, 1979.

[29] R. Albert and A. Barabasi, “Statistical mechanics of complex networks,” *Rev Mod Physics*, vol. 74, no. 1, pp. 450–461, 2002.

[30] A. Barabasi and Z. Oltvai, “Network biology: understanding the cell’s functional organization,” *Nat Rev Genet*, vol. 5, pp. 101–113, 2004.

[31] X. He and J. Zhang, “Why do hubs tend to be essential in protein networks?” *PLoS Genetics*, vol. 2, pp. 826–834, 2006.

[32] X. Hu, “Mining and analysing scale-free protein-protein interaction network,” *International Journal of Bioinformatics Research and Applications*, vol. 1, no. 1, pp. 81–101, 2005.

[33] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya, “Development and implementation of an algorithm for detection of protein complexes in large interaction networks,” *BMC Bioinformatics*, vol. 7, no. 205, pp. 1–13, 2006.

[34] S. Wachi, K. Yoneda, and R. Wu, “Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues,” *Bioinformatics*, vol. 21, no. 23, pp. 4205–4208, 2005.