Complex network theory for the identification and assessment of candidate protein targets

Ken McGarry[a], Sharon McDonald[b,1]

[a] *Faculty of Health Sciences and Well-being, University of Sunderland, City Campus, Sunderland, SR1 3SD, UK*
[b] *Faculty of Computer Science, University of Sunderland, St Peters Campus, Sunderland, SR6 ODD, UK*

Abstract

In this work we use complex network theory to provide a statistical model of the connectivity patterns of human proteins and their interaction partners. Our intention is to identify important proteins that may be predisposed to be potential candidates as drug targets for therapeutic interventions. Target proteins usually have more interaction partners than non-target proteins, but there are no hard-and-fast rules for defining the actual number of interactions. We devise a statistical measure for identifying hub proteins, we score our target proteins with gene ontology annotations. The important drugable protein targets are likely to have similar biological functions that can be assessed for their potential therapeutic value. Our system provides a statistical analysis of the local and distant neighborhood protein interactions of the potential targets using complex network measures. This approach builds a more accurate model of drug-to-target activity and therefore the likely impact on treating diseases. We integrate high quality protein interaction data from the HINT database and disease associated proteins from the DrugTarget database. Other sources include biological knowledge from Gene Ontology and drug information from DrugBank. The problem is a very challenging one since the data is highly imbalanced between target proteins and the more numerous nontargets. We use undersampling on the training data and build Random Forest classifier models which are used to identify previously unclassified target proteins. We validate and corroborate these findings from the available literature.

Keywords: Complex network theory, link-clustering, protein interactions, ontologies

1. Introduction

Protein interactions play a key role in the majority of activities occurring in the cell and participate in communications between cells [24]. The connectivity patterns of the interacting proteins can be modeled by complex network theory (graph theory) which can provide a statistical explanation of these activities and processes [21]. Integrating clustering methods with complex networks has enabled further insights, revealing the modular nature of proteins [28]. Proteins are often cooperate in modules and may be shared between several different cellular activities. Those proteins with a large number of verified interactions are classed as hub proteins. If they are implicated in one disease it is possible they may be participating in other disorders [23]. It should be noted that high connectivity (degree) or hubness does not necessarily imply that a given protein is important in some way with respect to disease. In this work we investigate the degree of protein connectivity patterns and also the location of a proteins position in the local network with respect to its predisposition to be a drug target.

The majority of disease causing genes are generally implicated with a single or small number of disorders although there are striking exceptions. The tumor suppressor gene TP53 appears to be involved with up to ten related diseases [12]. This gives credence to the disease network theory which is providing a new insights regarding how diseases occur [5]. Some diseases are more difficult to resolve, often a module of cooperating proteins can compensate for malfunctions of individual proteins. Consequently, making the identification of the faulty biological process more difficult to identify [17]. The idea of structural motifs may be a good candidate to help resolve the challenges such as cellular organization [3]. We can improve our knowledge and understanding of the mechanisms of disease based on a better understanding of protein targets and non-targets and may suggest alternative therapeutic interventions [13, 22].

However, any potential for a protein to be drug target implies it must possess a particular shape that can bind/interact with drug-like molecules i.e. it must con-

Email address: ken.mcgarry@sunderland.ac.uk (Ken McGarry)

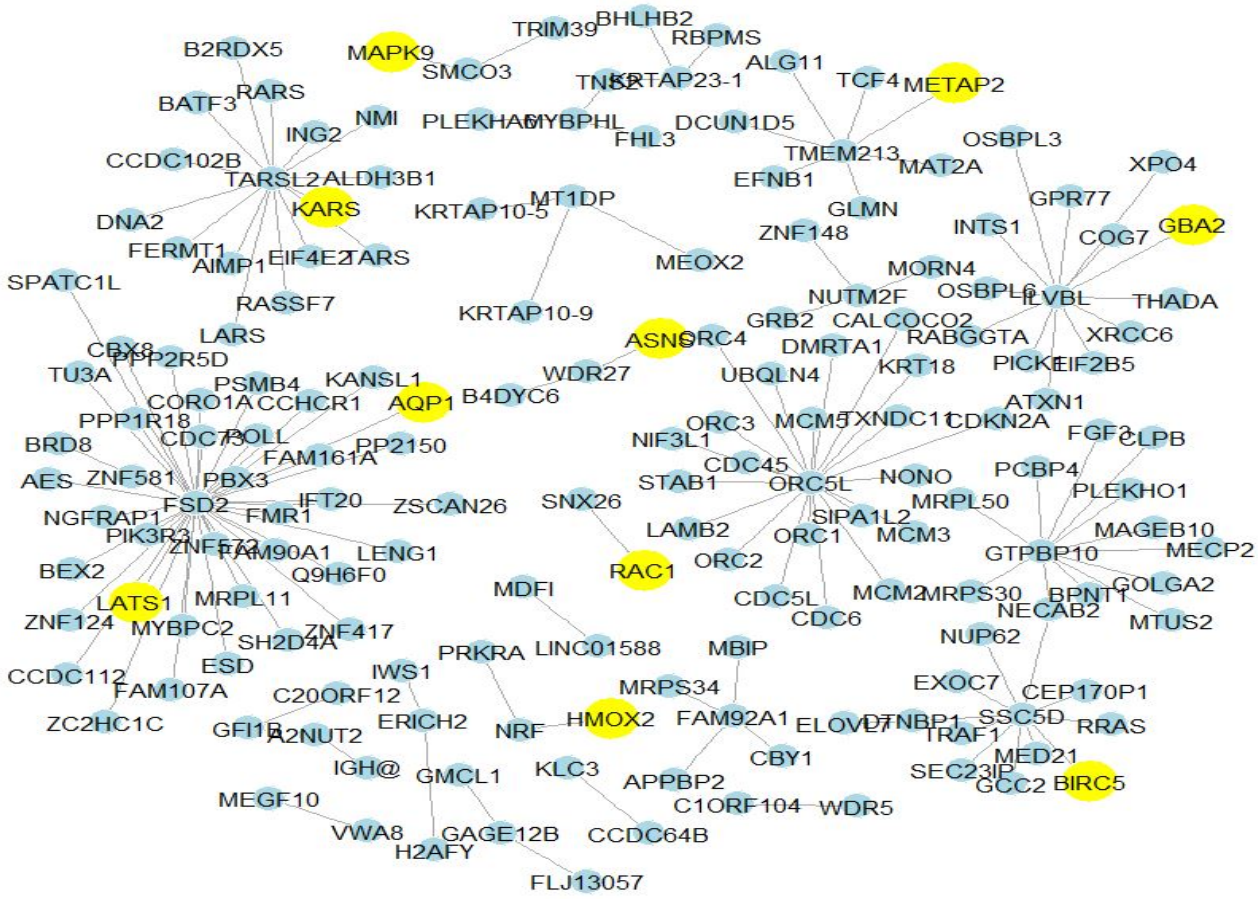


Figure 1: Small fraction of the protein network with drug targets colored yellow and slightly larger in size, non-targets are colored light blue. However, based on their connectivity patterns their biological and complex network statistics some of the non-targets may be prove to be viable drug targets.

tain a binding site. Recent research has investigated the role of the types of proteins such as G-protein coupled receptors, ion channels and kinases [7]. This work determined that a proteins relationship to the membrane and it's hydrophobicity may play an important role. However, it does beg the question, how many potential protein targets are out there? [25]. One analysis suggested there may be between 2000-3000 proteins that are potentially druggable candidates [26]. Another approach was able to identify 668 proteins that are currently not drug targets but that have target-like potential [2]. Some proteins may be completely undruggable, while others can only be perturbed by targeting their network neighborhood proteins. Currently, complete knowledge of the proteome and interaction targets is some years away from completion [4].

1.1. Related work

The technique developed by Yu et al, considers the problem as one of module distance estimation with the understanding that the human interactome is still incomplete and with all the uncertainty inherent [31]. Yu's ultimate goal was concerned with repositioning drugs for different diseases. The modules are composed of drug-protein pairs and all are involved with cancer specific functions. The disease module distance metric was able to identify several candidate drugs. The MBiRW method developed by Luo et al uses a bi-random walk to measure similarity of drugs and diseases [20]. MBiRW uses novel similarity measures and is well validated against gold standard data but lacks target information and biologically relevant information. The CommWalker algorithm devised by Luecken uses a random walk approach to sample the proteins assigned to functional modules [19]. For robustness, the modules are formed by three different link analysis procedures and an average walk will produce a goodness of fit value. The walks are terminated when they have approach a critical value. At each step the functional GO annotation is averaged out to calculate the module homogeneity, scores are then combined to enable each module to be ranked on its biological plausibility.

The closest work to ours tackles the challenges and opportunities of integrating biological knowledge in the form of annotations from gene ontology (GO). For example, Hsing *et al* used GO to build classifiers to identify hub pro-

teins which are highly connected proteins with many interaction partners [14]. However, the classifiers performed badly on some proteins through lack of suitable annotations. Work by Zhang *et al* explored the issues of identifying protein interaction partners through use of GO terms [32]. Support Vector Machine classifiers were constructed on the GO annotated PPI data and good accuracy was achieved on predicting the likely interaction partners. Research by Fu *et al* explored the likelihood that intrinsic disorder proteins will form highly interconnected hubs and potentially drug targets [11]. Again, the usefulness of GO was employed to annotate and analyze the relationships.

We extend all this previous work by adding novel analysis of the hub and target proteins using complex network theory and community structure of the protein interactions. Furthermore, we annotate the protein interactions with GO terms to help identify novel protein targets. Thus we are able to generate target protein candidates for use by other researchers to conduct biological experiments in the lab. The remainder of this is paper is structured as follows: section two discusses the methods including the architecture of our system and the sources of data and knowledge, section three describes the experimental results, section four presents the discussion and finally section five presents the conclusions and future work.

2. Materials and methods

2.1. Data and knowledge sources

The candidate drugs have known on-target and off-target proteins, this knowledge is augmented by accessing protein-to-protein interactions found in the HINT database [8]. This database contains high-quality protein-protein interactions from 8 interactome resources (BioGRID, MINT, iRefWeb, DIP, IntAct, HPRD, MIPS and the PDB). The database contains 12,429 unique proteins with 59,128 interactions between them. http://hint.yulab.org/. This data was further augmented by high quality protein interactions from the BioPlex database maintained by Harvard Medical School (http://bioplex.hms.harvard.edu/) [15].

Drug and protein target data was obtained from DrugCentral, this is a comprehensive drug information resource for FDA drugs and drugs approved outside USA. The resources can be searched using: drug, target, disease, and pharmacological action. The information resource is created and maintained by the Division of Translational Informatics at University of New Mexico. (as of 8th January 2018, http://drugcentral.org/). The data is particularly suited for our purposes, because we use the drugname, protein targets and protein type [27].

The data is highly imbalanced between the target proteins (1,449) and the nontargets (10,567). The data are randomly divided into three groups, table 1 indicates the split and objective of each data set. The training data consist of 724 targets and 5,284 nontargets, the test data is similarly unbalanced. The Exploratory data consists of 1,000 nontargets reserved for identifying potential targets.

Table 1: Composition of data

Data id	No targets	No nontargets	Purpose
T	724	5,284	Training
TE	725	5,283	Testing
NT	0	1,000	Exploratory data

We select the best method of down sampling on the training data based on the accuracy and other statistics gained from the random forest classifiers.

Gene ontology (GO) provides useful biological information and it is recognized as the *de facto* standard for gene product annotation [1]. This enables an assessment to be made regarding the biological plausibility of the interacting proteins and to observe the extent they actually cooperate in viable biological functions rather than random or spurious associations. The GO terms are organized to represent the three main aspects of biology: molecular functions (MF), cellular components (CC) and biological process (BP). For each protein, enrichment was performed using the GO-slim version of the gene ontology (GO). The enrichment method is based on similarity measures using information content techniques [9]. We mapped our low-level annotations to several high-level GO-slim terms, in our generic version 149 terms are available.

All proteins without GO terms are removed and the remaining proteins are annotated the GO-slim version (http://www.geneontology.org/page/go-slim-and-subset-guide) simplifies the terms. GO-slims are subsets of terms in the ontology that give a broad overview of the ontology content. GO-slims do this without including the detail of the specific fine grained terms and therefore simplify computational model building. We discovered that accuracy is not greatly compromised since we obtain accuracy's that are comparable to other work that uses the full GO terms [9].

The GO-slim annotated proteins are used to build a Random Forest classifier [6]. Random Forests are a supervised ensemble classification technique requiring class labels, where a group of weak models combine to form a powerful model. Several decision trees are created (hence a forest) with random sampling used as the attribute to split on, there is a direct correlation between the number of trees in the forest and the accuracy. Another important advantage of the Random Forest classifier is that the predictor variables are ranked according to importance using the Gini impurity measure and is used for the calculation of splits during training.

The Random Forest was assessed using standard classifier metrics:

Accuracy = (TP + TN) / (TP+ TN + FP + FN)
Sensitivity = TP / (TP + FN)
Specificity = TN / (TN + FP)
PPV = TP / (TP + FP)

NPN = TN / (TN + FN)

Where: TP is a True Positive, TN is a True Negative, FP is a False Positive and FN is a False Negative. PPV is the Positive Predictive Value and the NPV is the Negative Predictive Value. ROC and PR curves were plotted using these values.

2.2. Software implementation and availability

The analysis was conducted using the R language (version 3.4.4) with the RStudio programming environment. We used an Intel Xenon CPU, 64-bit (3.2GHz) with 128 GB of RAM. The R software written by other researchers included the igraph package by Kolaczyk [16] along with other packages for data manipulation, transformations and plotting include: ggplot2, dplyr and ontologySimilarity. The RandomForest package by Liaw allowed the building and testing of the classifier models [18]. The R code that created the data, diagrams, tables and charts described in this paper are freely available on GitHub for download: https://github.com/kenmcgarry/ComplexNetworks

2.3. Complex network theory

Graph theoretic methods are suitable to any application 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 recent discoveries. It is likely that the essential genes and the disease genes encode the hubs [28] and that gene network topology is unlikely to encode the information to deduce disease modules [12]. In algorithm 1, we detail our computational methods of discovering the disease modules.

Closeness centrality (CC) of protein i is the sum of graph-theoretic distances from all other proteins in the 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, where N is the number of all proteins in the network. The closeness centrality of protein i in a network is given by the equation:

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

Betweenness centrality assesses the extent of influence a protein has in facilitating communication between pairs of proteins and is defined as the fraction of shortest paths going through a given node [10]. In the PPI network, then the CB of a node v is given by:

$$CB(v) = \sum_{s \neq t \neq \nu \in V} \frac{\sigma(s,t,|v)}{\sigma(s,t)} \tag{2}$$

Where: $\sigma(s,t,|v)$ is the total number of distinct shortest paths from node s to node t passing through v, $\sigma(s,t)$ is the total number of shortest paths between node s and node t irrespective of whether they pass through node v or not.

Complex networks (graph theory) can generate very large structures in real-world applications, often consisting of thousands of nodes and tens of thousands of connections. This results in a common task to query if two particular nodes are connected and if so how distant are they are. Detection and assessment of the *shortest path* is important to centrality measures and can be defined as when two nodes i and j are connected if there exists a sequence of connections that connect i and j. The length of a path is the number of connections between them, denoted by d_{ij}. In biology, two proteins may not interact directly but may communicate through a signaling cascade of other proteins. Small path lengths from tightly coupled networks tend to give high clustering coefficients as derived by equation 3, random networks have low clustering coefficients compared with real-world networks.

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

Where $V = v1, v2...vn$ 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, for a vertex v_i, is its immediately connected neighbors as follows:

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

The degree k_i of a vertex is the number of vertices in its neighborhood $|Ni|$. Making the clustering coefficient C_i for a vertex v_i the proportion of links between the vertices within its neighborhood 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 neighbor, only the following edges could exist among the vertices within the neighborhood.

The clustering coefficient is used to estimate the density of the immediate neighborhood of each vertex and formally, for undirected graph $G = (V : E)$, define the neighbor set of $v_i \in V$, denoted by N_{v_i} . A high cluster coefficient indicates a high level of interconnection between members of a node's neighboring nodes. These measures return a value of one if every neighbor connected to v_i is also connected to every other vertex within the neighborhood 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 entire network is the average of the clustering coefficient for each vertex:

$$C = \frac{1}{n}\sum_{i=1}^{n} C_i \tag{5}$$

We therefore expect that hub proteins on average have a higher clustering coefficients than non-hubs. This was confirmed by our experimental work but the important proteins are not identified by connectivity numbers alone, the location of a key protein is also by its position (topology) in the network.

Algorithm 1
Build Random Forest classifier and complex network

```
1:  procedure BUILD RF CLASSIFIER(HINT, DrugTarget, GO − slim)
2:      do initialize
3:          protein list ← annotate with target/nontarget labels
4:          protein list ← annotate with GO-Slim
5:          Remove proteins with < 1 GO-Slim annotation
6:          NT (unseen data) ← randomly select 1,000 nontargets
7:          T (train data) ← 50/50 split target/nontarget
8:          TE (test data)← 50/50 split target/nontarget
9:          proteins matrix ← convert list to binary matrix
10:     end initialize
11:
12:      PPInetwork ← igraph builds complex network
13:      NetStatistics ← calculate k-coreness for all nodes
14:
15:     while
16:         accuracy≤accuracy from last iteration do
17:             RF parameters← (Num Trees, Cutoff )
18:             RF accuracy← test data (TE)
19:             modify RF sampling ← classifier accuracy
20:     end while
21:
22:         Candidate list← RFmodel(unseen data (NT))
23:     return PPInetwork, NetStatistics, RandomForest, CandList
24: end procedure
```

Referring to algorithm 1, lines 2-9 perform the initialization of key values. Based on the available information on protein targets we label proteins as either a target or nontarget. All proteins are then annotated with GO-slim, any protein without an annotation of some form is removed from the study. A binary matrix for training the random forest is made which is then split 50/50 for train/test data, the matrix consists of a series of 1's' for presence of a GO-slim term and 0's if it is not for each protein with the class label for target/nontarget identify.

Lines 12 and 13 build the protein to protein interaction network (PPI) and calculate a number of statistics, the most important is the k-coreness for each protein. Lines 15-20 handle the training of the Random Forest classifier, we modify the number of tree's, the cutoff (the winning class for an observation is the one with the maximum ratio of proportion of votes to cutoff) to obtain the best model based on confusion matrix and PR/ROC curves. When the RF sampling has provided a useful classifier, line 22 passes the unseen data (NT) through the trained RF classifier and generates a list of candidate targets.

A graph is k-connected if every pair of vertices is connected through at least k distinct paths, that do not share edges. This property is related to the strength of a network, a k-connected network remains connected whenever fewer than k links are removed or broken. The k-core of graph is a maximal subgraph in which each vertex has at least degree k. The coreness of a vertex is k if it belongs to the k-core but not to the (k+1) core. The cores represent an important subgraph that contain functions of biological significance. The formation of a hub node is the result of the scale-free property of a network but the definition of hubness is also a consequence of the size of the network. Hubs are to found in most real-world networks and have a significant impact on the topology, but they highly unlikely to be formed in a random network.

3. Results

The first stage was to assess the impact of study bias of the known protein connectivity patterns. The data are the result of scientific experiments that may have concentrated on the better known or more 'glamorous' proteins. That is to say proteins with few connections may be the result of neglect on the part of the scientific community and not because these proteins do not have an important role.

In figure 2 the frequency counts of the types of protein targets is displayed, the most predominant type is GPCR. We removed any protein type with fewer than 50 records as they were deemed unlikely to produce successful classifiers.

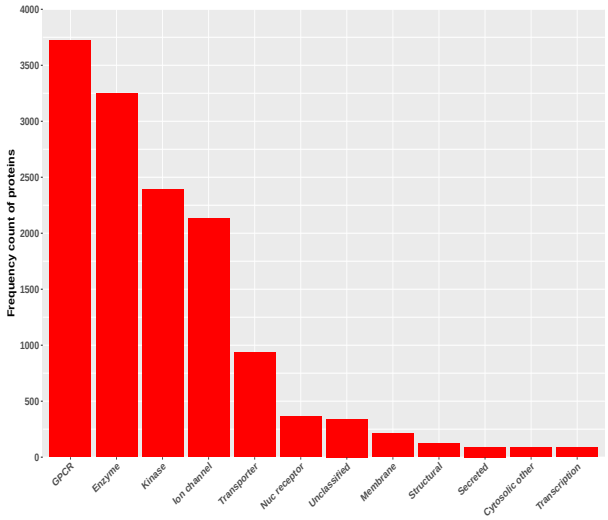


Figure 2: Breakdown of protein types (13,735 unique proteins) from DrugCentral database.

We created the protein networks using the available proteins. In figure 3 the degree of the network is plotted against the cumulative distribution function and it is evident that a powerlaw exists in the network. That is we have many proteins with few connections but there are a small number of proteins with several hundred connections.

Complex networks were constructed using the protein interaction data and annotated each node with information regarding its status as a drug target, disease protein, hub protein, essential protein or function simply unknown.

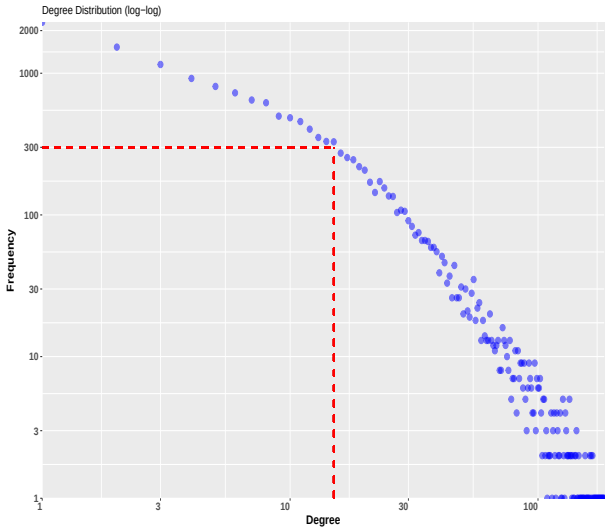


Figure 3: Protein network degree with power law evident indicating the non-random network structure. The dotted line represents the 90th percentile cut-off point for determining if a protein is a hub. In this network, a degree of 17 is required giving 300 hubs.

The overall network statistics based on connectivity patterns of 15726 nodes with 109,953 links gave: Modularity of 0.37, Average path length of 3.81, Transitivity of 0.053, Diameter of 11 and is fully connected with no disconnected (isolates) proteins. Table 2 shows the top 20 proteins ordered according to the *betweenness* statistics, this produces a very large value and is related to the *degree* of each individual node, these nodes have the largest number of connections to other proteins. The *closeness* statistic is reasonably constant for these 20 proteins and implies that they have similar characteristics. The *hubness* value also varies little with these proteins and suggests that connectivity patterns are similar. The *centrality* measure is computed using the alpha centrality method, which is less likely to encounter problems with asymmetric matrices. The *community* measure indicates that these 20 proteins generally cooperate with few other communities or modules. A community is usually interpreted as a biological function or process. However, we must be careful not to infer too much from the network statistics alone the next stage is to annotate the network proteins with gene ontology (GO) terms. Determining biological plausibility through GO provides a more valid interpretation.

Table 3 gives the protein statistics organized by *hubness*, the top 20 proteins are dominated by ribosomal proteins, these are the organelles that catalyze protein synthesis. They form 45 tightly clustered communities.

Not all of the important proteins can be classed as hub-like, a number of proteins with low degree but high betweenness may account for 30% of the available proteins in a PPI network. These would not be highly rated as potential drug targets under the usual statistical criteria but some have the potential to be considered as targets. We employed GO-slim annotations (http://www.geneontology.org/page/go-slim-and-subset-guide) rather than apply all known GO annotations to the proteins, we wished to apply a general level of terms that provide a computationally tractable solution. The generic GO-slim term set provides 149 terms across the Cellular Component, Biological Process and Molecular Function databases.

The data was highly imbalanced between the non-targets (10,000 proteins) and the targets (1,449 proteins), this adversely affected the initial classifiers with accuracies dropping to 45% since the more numerous non-target class tended to predominate the classifiers. The nature of the data implied we could not upsample using SMOTE or similar techniques, therefore we down sampled the training protein data to give more balanced training set. This allowed a Random Forest based classifier to be successfully constructed using the 149 ontology terms as variables. This gave a matrix 149 x 6000: 149 terms (independent variables) and 6,000 proteins unevenly divided into target and non-target classes. The overall accuracy of the ontol-

Table 2: Highest scoring 20 proteins ordered according to *betweenness*

	close	degree	between	hub	central	comm
HSP90AB1	5e-08	454	1353987	0.093	15	13.0
GRB2	6e-08	452	1228333	0.12	18	18.0
GOLGA2	5e-08	340	632816	0.13	36	8.0
ATXN1	4e-08	276	593291	0.051	29	11.0
ZDHHC17	3e-08	249	502844	0.053	33	11.0
KDM1A	3e-08	224	399557	0.069	32	11.0
TRIM27	4e-08	258	393388	0.12	65	8.0
KRT40	3e-08	333	379048	0.14	58	8.0
MAPK6	3e-08	201	363744	0.058	47	11.0
UBE2I	4e-08	202	351061	0.063	58	11.0
HSPA8	3e-08	199	347256	0.038	37	11.0
EGFR	4e-08	196	339368	0.05	24	11.0
LZTS2	4e-08	244	338341	0.094	12	8.0
TERF1	3e-08	218	332496	0.074	39	11.0
TRAF2	4e-08	214	332495	0.074	59	8.0
MEOX2	3e-08	210	313349	0.094	62	8.0
CRK	3e-08	216	307830	0.045	36	11.0
UBQLN4	3e-08	177	305193	0.042	26	11.0
REL	3e-08	197	295228	0.056	2	8.0
RBPMS	3e-08	220	287819	0.065	17	8.0

Table 3: Highest scoring 20 proteins ordered according to *hubness*

	close	degree	between	hub	central	comm
RPL37A	1e-08	138	30867	1.00	9.03	45.0
RPL18	1e-08	143	39136	0.98	9.81	45.0
RPS8	1e-08	132	34027	0.96	41.24	45.0
RPS3A	1e-08	106	38479	0.93	13.31	45.0
RPL18A	2e-08	138	47331	0.93	20.43	45.0
RPL14	1e-08	117	25628	0.90	61.55	45.0
RPL30	1e-08	129	32356	0.89	3.81	45.0
RPS2	2e-08	129	65550	0.88	0.44	45.0
RPL7A	1e-08	82	9141	0.85	17.27	45.0
RPL5	1e-08	87	13180	0.84	18.64	45.0
RPS13	1e-08	76	2923	0.83	11.48	45.0
RPL4	1e-08	79	13453	0.83	28.97	45.0
RPL10A	1e-08	83	12913	0.82	10.64	45.0
RPS3	1e-08	86	17144	0.81	13.83	45.0
RPL6	1e-08	99	21401	0.80	13.30	45.0
RPL7	1e-08	94	13661	0.80	13.28	45.0
RPS16	1e-08	73	3639	0.80	6.59	45.0
RPS6	1e-08	77	14896	0.78	36.98	45.0
RPS14	1e-08	115	35086	0.77	37.49	45.0
RPS4X	1e-08	77	13443	0.77	16.60	45.0

ogy based classifier is 74%, proving it has reasonable accuracy at discriminating protein targets from non-targets using ontological terms, 12 re-samples were used and the forest was composed of 5000 trees. Recalling table 1, the split in data was 50% assigned to the training set and 50% of data in the test/validation set. The ROC curves for test data are shown in figure 4a.

The receiver operating characteristic curve (ROC) is based on evaluating the tradeoffs between specificity and sensitivity. Specificity is the probability of predicting a target given the true state is in fact a target, whereas sensitivity is the probability of predicting a non-target given the true state is a non-target. The classifier statistics are:

Accuracy = 74%
Sensitivity = 0.86%
Specificity = 0.72%
Negative Predictive Value = 0.97%
Positive Predictive Value = 0.30%

The confusion matrix for the test set data is given below:

	0	1
0	3849	99
1	1434	626

The random forest has misclassified 1,434 nontargets as targets and 99 targets as nontargets. It has correctly identified 626 targets and 3,849 nontargets. The accuracy for the targets is much higher than for nontargets because the sampling procedure we used placed a higher value on correctly identifying targets. Since the the data is highly imbalanced with some overlap this was a necessary tradeoff.

The RandomForest classifier uses Gini index to rank each GO predictor term for it's contribution to discriminating between protein targets and non-targets. This information is displayed in table 4 with the highest scoring 15 predictors out of 149.

In figure 4 the precision-recall curves are presented, this shows the precision values for corresponding sensitivity (recall) values. Similar to the ROC plot, the PR plot provides a model-wide evaluation. However, precision is directly influenced by class imbalance and does take the TN (true negatives) into consideration. The precision-recall curve highlights the tradeoff between precision and recall, where a high area under the curve represents both high recall and high precision. High precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both criteria, indicate that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall). Any classifier with high recall but low precision returns many results, but the majority of its predicted labels are incorrect when compared to the training labels. A classifier with high precision but with low recall will return very few results, but most of its predicted labels are correct when compared to the training labels.

Examining figure 4, there appears little difference between the classifiers with perhaps the purple line (no sampling) having the best performance. This is because not using sampling will produce a classifier that will allow the majority class to dominate, hence a greater accuracy is produced by always predicting it. Together with the ROC curve the precision-recall curve indicate the performance level for the random forest.

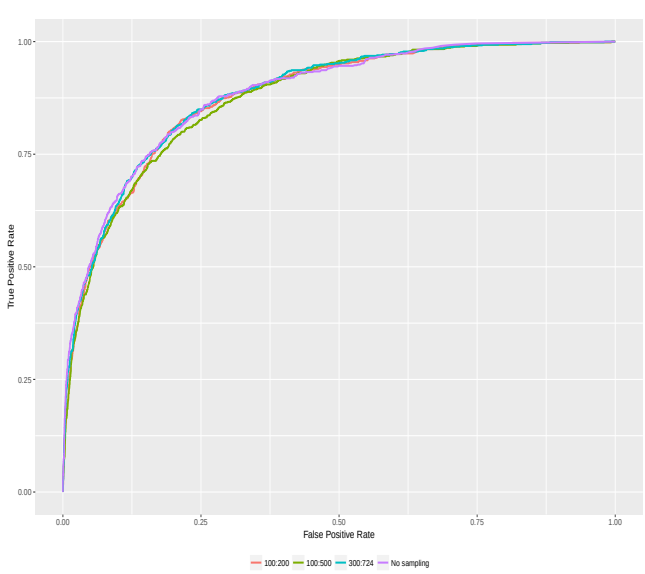
Table 4: Top 15 GO Term predictor importance ranked on Gini score generated from the RandomForest classifier

GOterm	Type	Description	Value
GO:0016301	MF	kinase activity	86.37
GO:0044281	BP	small molecule metabolic.process	59.43
GO:0043167	MF	ion binding	30.58
GO:0004871	MF	signal transducer activity	29.86
GO:0005886	CC	plasma membrane	27.04
GO:0055085	BP	transmembrane transport	20.54
GO:0022857	MF	transmembrane transporter activity	20.17
GO:0016491	MF	oxidoreductase activity	18.89
GO:0007165	BP	signal transduction	14.61
GO:0008233	MF	peptidase activity	13.84
GO:0006464	BP	cellular protein modification process	12.03
GO:0005615	CC	extracellular space	11.58
GO:0006950	BP	response to stress	10.87
GO:0042592	BP	homeostatic process	9.04
GO:0006810	BP	transport	8.20

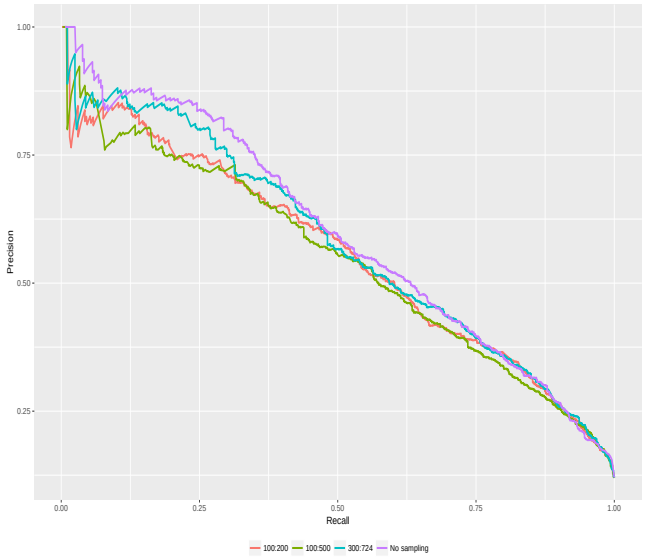
The variable importance is calculated by the 'impurity' from the Gini index for splitting (deciding when to make splits on the input variables and when to stop growing the trees). At each split the importance of the variable is accumulated for every tree for that variable. For classification tasks such as this one, the Random Forest keeps a tally of each tree's output and then takes a majority vote to decide the overall class output. Further understanding of the classifier was gained when a breakdown of the GO terms giving frequency counts between targets and non-targets was made.

The term GO:0016301 for kinase activity was the strongest variable with a score of 86.37, the second key variable was term GO:0044281 for small molecule metabolic processes at 59.43. The third highest scoring variable was GO:0043167 which codes for ion binding with a value of 30.58. Examining the frequency counts of the differences between the presence or absence of an ontological term is presented in figure 5. The frequency counts for the presence (red bars) or absence (cyan) of a GO term for targets and non-targets. The *present'* category is generally the most predominant, the *absent* category is more frequent for the target proteins and thus discriminates between the non-targets.

The k-core of a graph relates to the maximal connected subgraph, the vertices of which must have at least of degree k within the subgraph. It is a useful technique for examining network connectivity where it is important to detect community and clusters. However, it can be quite interesting to visualize the coreness structure of small or medium networks. In figure 6 we show an example candidate protein and its neighbors. As a statistic, the numerical value of the k-coreness cannot identify potential



(a) The ROC curves for test data accuracy



(b) The Precision and Recall curves for test data accuracy

Figure 4: ROC and PR for several sampling strategies. The overall accuracy on test data is 74%

targets. We examined the medians and IQR for the two groups and for targets the median=2, IQR=6, for nontargets median=5 and IQR=8. The k-coreness is too over dispersed to discriminate between targets and nontargets but is useful for further analysis when targets are identified by the classifier.

Examining the MIPS protein complex database it would appear that most of the k-plex cores are part of protein complexes. We examined the most significant GO terms among k-plex core members are ranked and again we examined the protein types associated with these k-cores (between targets and non-targets). The conclusion is that the location of a node is more important than the number of its connections or the number of connections of near neighbors. The k-coreness criteria is a better indicator of a nodes influence on biological function (target value) than degree.

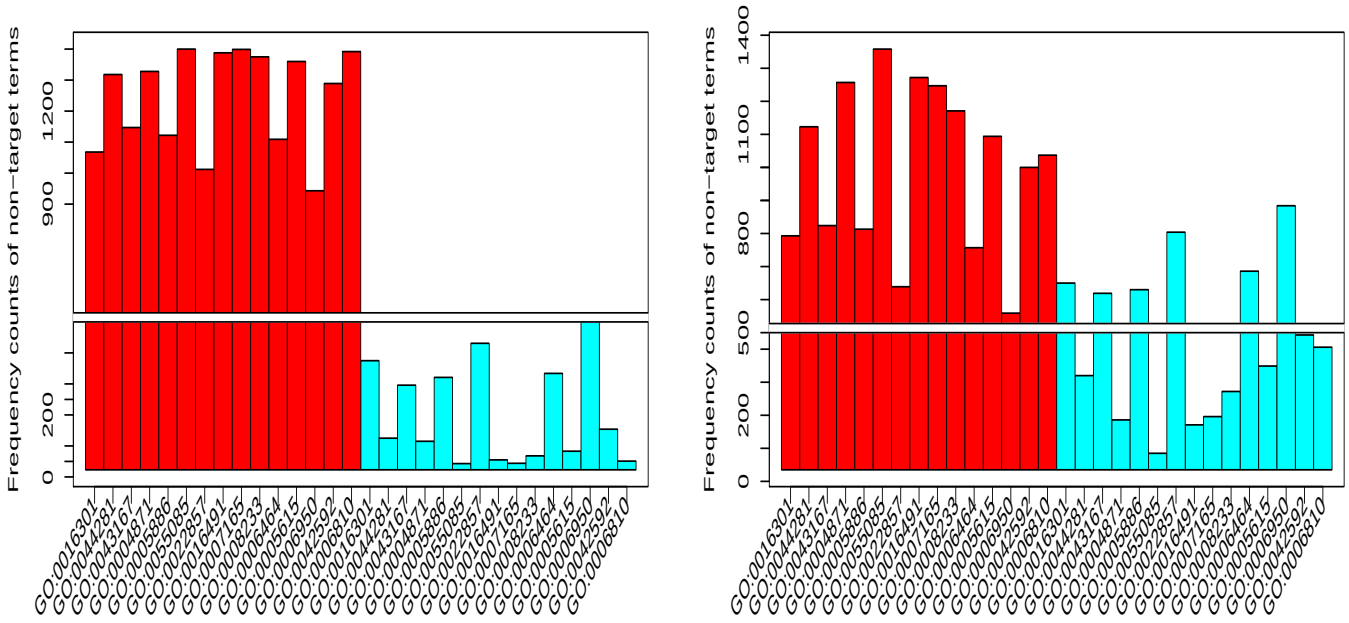
The next stage of the analysis required generating a third list of proteins that were not drug targets but had GoSlim ontology annotations. The first and second lists of proteins refer to the targets and non-targets used to build and test the RandomForest classifier. Third list proteins that were identified as target-like by the classifier were examined at the k-coreness level to determine network neighborhood and likely importance.

In table 5 the list of potential protein targets identified by our system are displayed. They are ordered according to their probability score, the likelihood they are target proteins. The first column gives the protein type and the second column identifies the protein name. The third column identifies the protein as a hub or non hub (based on our criteria) and the fourth column gives the probability score of the protein as a likely target (derived from the Random Forest outputs). The fifth column provides the k-coreness score, the sixth column provides evidence found from literature or other sources that the protein has been proposed as a drug target.

The final stage of the analysis takes the generated a list of candidate target proteins along with their characteristics as identified in table 5 and to place them them in the context of available drugs. Previous work by Yamanishi considered four possible approaches to this problem [30, 29]: (i) new drug candidate compounds versus known target proteins, (ii) known drugs versus new target candidate proteins, (iii) new drug candidate compounds versus known target candidate proteins, (iv) known drugs versus known target candidate proteins. We tackle class (ii), because we are introducing previously unknown/untargeted proteins.

Based on the protein class (e.g. enzyme, GPCR, Transporter etc) of our candidates we assessed the chemical structure and pharmacological properties of drugs targeting proteins of these types. We filled in missing protein class data from other sources. The characteristics of GPCR, Ion Channel, Transporter and enzyme class proteins were profiled based on the chemical structure and pharmacological similarity scores of the drugs targeting them.

Examining table 5 reveals that the highest scoring k-coreness proteins are generally hubs. That is to say they have over 17 connections to other proteins. However, their k-coreness score indicates they are centrally placed in the network and is not an indicator for “targetness”. The majority of candidates are non-hubs with fewer connections although their k-coreness score is between 12 and 2. We examine the top scoring candidate target (APLP1) which



(a) GO term frequencies for non-targets

(b) GO term frequencies for targets

Figure 5: The frequency counts for the presence (red bars) or absence (cyan) of a GO term for targets and non-targets. The y-axis is broken to highlight the differences between them. The *present'* category is generally the most predominant. The *absent* category is more frequent for the target proteins and thus discriminates between the non-targets

is an amyloid beta precursor like protein 1, it is involved in synaptic maturation during cortical development and has been associated with late-onset of Alzheimer's disease. It has a k-coreness rank of 16 and a neighborhood of 45 interconnected proteins, four of which have 13 drugs targeted at them. The APLP1 protein is located in a reasonably active part of the drugome. Experimental evidence indicates that Copper and also Herapin (Heparin has anti-clotting properties) binds to this protein. A similarity search based on amino acid sequence reveals it is similar to APLP2 and APP, both are linked to Alzheimer's. Referring to table 6 the drugs targeted at the four APLP1 partners are displayed.

The protein PNP (Purine Nucleoside Phosphorylase) has been implicated in adenosine deaminase deficiency and lesch-nyhan syndrome. The main clinical indication is recurrent infections due to severe T-cell immunodeficiency. Other patients may also suffer from neurologic impairment. The TK1 (Thymidine Kinase 1) is a cytosolic enzyme that catalyzes several components and is used as a biomarker in a number cancers (colorectal cancer, lung and leukemia). CDK4 (Cyclin Dependent Kinase 4) is a catalytic component of the protein kinase complex that is important for cell cycle G1. Mutations in this protein are also associated with a variety of cancers. The SAT1 (Spermidine/Spermine N1-Acetyltransferase 1) protein is a rate-limiting enzyme in the catabolic pathway of polyamine metabolism. It is implicated in Keratosis Follicularis which is a rare, inherited skin condition disease.

We accessed the STITCH database which contains experimental evidence and also predicted information for the ALPL1 protein. It is centrally placed at the locus of several diseases, drugs and other proteins thus may be playing an as yet unknown role. We examined the role of the top 10 proteins for evidence of potential use as drug targets, these are displayed in table 7.

The NUDT18 (Nudix Hydrolase 18) protein functions to eliminate potentially toxic nucleotide metabolites from the cell and regulate concentration levels of cofactors and signaling molecules. CCNB1 (Cyclin B1) is a regulatory protein involved in mitosis and is involved in several cancers such as nasopharyngeal carcinoma and adrenal carcinoma. It has several drugs targeted such as Temozolomide, Nocodazole and Purvalanol. The PIP4K2A (Phosphatidylinositol-5-Phosphate 4-Kinase Type 2 Alpha) protein and is involved in the regulation of secretion, cell proliferation, differentiation and motility. It is implicated in leukemia and acute lymphoblastic tumors. It is currently undergoing investigation with Adenosine triphosphate (ATP) as a potential intervention.

The GALNS (Galactosamine (N-Acetyl)-6-Sulfatase) protein is involved in mucopolysaccharidosis which an au-

Table 5: Potential target proteins identified by RandomForest classifier and measures calculated for k-coreness criteria. Keeping proteins with probability (>0.8) and sorted by k-coreness (> 1), we have approximately 50 target protein candidates.

	TargetClass	Gene	myhubs	prob	core
1	Unknown	APLP1	hub	0.85	16
2	Enzyme	NUDT18	hub	0.95	15
3	Unknown	CCNB1	hub	0.81	14
4	Kinase	PIP4K2A	hub	0.94	14
5	Enzyme	GALNS	hub	0.88	13
6	Kinase	PBK	hub	0.97	13
7	Unknown	NPPA	hub	0.88	13
8	Unknown	LRRC8A	hub	0.84	13
9	Unknown	KIR3DS1	hub	0.85	13
10	Unknown	PSME2	hub	0.82	13
11	Kinase	STK40	hub	0.94	12
12	Kinase	PRKAG1	hub	0.86	12
13	Unknown	LRRC8E	hub	0.81	12
14	Kinase	PRKACG	hub	0.95	12
15	Unknown	NME2	hub	0.95	11
16	Kinase	KSR1	hub	0.89	10
17	Kinase	MAP3K14	non-hub	0.95	10
18	Enzyme	ADSL	hub	0.82	10
19	Kinase	PSKH2	hub	0.85	10
20	Enzyme	BLVRA	hub	0.81	9
21	Unknown	RASSF2	non-hub	0.85	9
22	Enzyme	GAD1	non-hub	0.83	8
23	Kinase	CKM	hub	0.88	8
24	Unknown	ERLIN2	non-hub	0.83	8
25	Enzyme	FAM20C	non-hub	0.89	7
26	Kinase	CDK20	non-hub	0.92	6
27	Enzyme	DHTKD1	non-hub	0.82	6
28	Unknown	C8G	non-hub	0.87	6
29	Kinase	TSSK6	non-hub	0.91	6
30	Enzyme	NLN	non-hub	0.88	5
31	Enzyme	NUDT9	non-hub	0.84	5
32	Unknown	FGF10	non-hub	0.90	5
33	Enzyme	SDHAF2	non-hub	0.86	5
34	Transporter	SLC6A6	non-hub	0.94	5
35	Kinase	PIKFYVE	non-hub	0.87	5
36	Unknown	CNNM4	non-hub	0.87	4
37	Enzyme	SULT1B1	non-hub	0.82	4
38	Unknown	PEX2	hub	0.82	3
39	Enzyme	GOT1L1	non-hub	0.82	3
40	Enzyme	AMDHD2	non-hub	0.85	3
41	Unknown	GART	non-hub	0.81	3
42	Enzyme	UQCC2	non-hub	0.86	3
43	Transporter	SLC2A3	non-hub	0.93	3
44	IC	FXYD7	non-hub	0.86	3
45	Kinase	PRPS1L1	non-hub	0.89	2
46	Transporter	SLC16A2	non-hub	0.96	2
47	Unknown	FDX1L	non-hub	0.95	2
48	Enzyme	HAO1	non-hub	0.82	2
49	Enzyme	CBR4	non-hub	0.85	2
50	Enzyme	HAGH	non-hub	0.81	2

Table 6: The drug targeted proteins in the APLP1 module (k-coreness) network.

	DrugName	TargetClass	Gene
2	Mercaptopurine	Enzyme	PNP
3	Sorafenib	Unclassified	PNP
4	Brivudine	Kinase	TK1
5	Idoxuridine	Kinase	TK1
6	Zidovudine	Kinase	TK1
7	Broxuridine	Kinase	TK1
8	Sunitinib	Kinase	CDK4
9	Quercetin	Kinase	CDK4
10	Ruboxistaurin	Kinase	CDK4
11	Ceritinib	Kinase	CDK4
12	Nintedanib	Kinase	CDK4
13	Ribociclib	Kinase	CDK4
14	Pentamidine	Enzyme	SAT1

Table 7: Evaluation of the first 10 candidate target proteins. The larger the k-core, then usually more drugs will be involved in the network

Gene	k-core size	Drugs	Target	Evidence
APLP1	16	22	Yes	(Simons, 2002)
NUDT18	15	2	Tentative	(Takagi, 2012)
CCNB1	14	17	Yes	(Momeny, 2017)
PIP4K2A	14	12	Tentative	(Bekker-Mndez, 2017)
GALNS	13	2	Yes	(Hiramatsu, 2017)
PBK	13	1	Yes	(Yang, 2017)
NPPA	13	55	Yes	(Lynch, 2012)
LRRC8A	13	162	Tentative	(Platt, 2017)
KIR3DS1	13	1	Yes	(Pyo, 2013)
PSME2	13	10	Yes	(Gomes-Alves ,2010)

tosomal recessive lysosomal storage disease caused by build up of keratan sulfate. The main features include short stature and skeletal dysplasia. Several drugs such as elosulfase are undergoing trials. The PBK (PDZ Binding Kinase) protein has recently been associated with malignant process of cancers and may be involved in the activation of lymphoid cells and various testicular functions. The NPPA (Natriuretic Peptide A) protein is involved in the control of extracellular fluid volume and electrolyte homeostasis. It is implicated in various cardiac and heart congestion

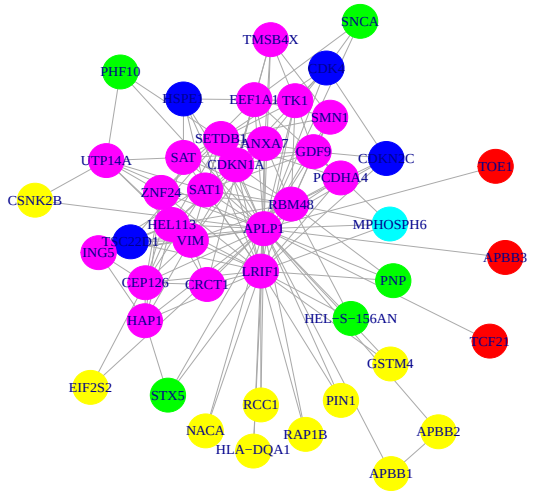


Figure 6: The protein APLP1 and its subnetwork indicating k-Coreness. The colors indicate the centrality of the subnetworks: red or yellow nodes are on the periphery, green and cyan represent the nearest neighbors while the central nodes are blue or purple. The plotting algorithm attempts to obtain a circular pattern where possible to highlight the k-cores

problems, a number of drugs such as Chlorthalidone are targeted at NPPA.

The LRRC8A (Leucine Rich Repeat Containing 8 VRAC Subunit A) is part of a family of proteins that are involved in many biological processes, such as cellular trafficking and cell adhesion. It is implicated in several immunodeficiency problems highlighted by low or absent serum antibodies and low or absent circulating B-cells. The KIR3DS1 (Killer Cell Immunoglobulin Like Receptor, three Domains and Short Cytoplasmic Tail) protein are transmembrane glycoproteins formed by killer cells and T cells. Killer-cells are highly polymorphic and this causes problems developing treatments and are thus an research active area. The PSME2 (Proteasome Activator Subunit 2) protein is distributed throughout eukaryotic cells at a high concentration and cleave peptides in an ATP/ubiquitin-dependent process in a non-lysosomal pathway. It is implicated in problems such as ulceroglandular tularemia and cystic fibrosis, which are targeted by Bortezomib, Carfilzomib and other drugs.

Examining the k-core statistic, we find that the larger the k-core module the more likely there will be drugs already targeting the disease implicated proteins in the network neighborhood. This is confirmed by the information in table 7 and also table 5. It must be recalled that in the databases used in our study there is no mention of these proteins as targets. The Random Forest classifier has identified these proteins based on their gene ontology annotations. Some of the classifier discovered targets are in fact already targets, unfortunately this information was not available in our database. However, it has demonstrated the reliability and validity of our technique.

4. Conclusions

The method described in this paper presents a novel approach for considering the viability of previously unrecognized/untargeted proteins as potential drug targets. We stress that full validation can only be achieved through using in-vitro, animal tissue and ultimately human tissue experiments. Our method provides an in-silico approach for generating lists of candidate targets that have rankings on several criteria such as biological similarities with known targets, protein type and complex network statistics describing the protein network neighborhood. The latter may play a role in the patient developing unwanted side-effects. We determined that both structure and connectivity patterns are necessary for identification of target proteins but this information needs to be coupled with biological knowledge. The use of Gene Ontology provides a biologically plausible method of protein annotation and assessment for further consideration as a target. In terms of limitations, we can say nothing about drug overall efficacy or long term effects on the predicted targets. Nor do we consider the mode of drug action or pharmacological action. Future work will address the issues of drug repositioning in a principled way using some of the methods described here.

Conflict of interests

We declare that we have no conflicts of interest.

Acknowledgments

We would like to thank the anonymous reviewers for their suggestions and comments which have improved the quality of the paper. We would also like to thank Kolaczyk and Csardi for developing the igraph package and Liaw for the RandomForest package.

References

[1] M. Ashburner. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[2] T. Bakheet and A. Doig. Properties and identification of human protein drug targets. *Bioinformatics*, 25(4):451–457, 2009.

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

[4] A. Barabasi, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nat Rev Genet*, 12:56–68, 2011. doi: 10.1038/nrg2918.

[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. doi: 10.1371/journal.pone.0008090.

[6] L. Breiman. Random forests. *Machine Learning*, 45(1): 5–32, Oct 2001. doi: 10.1023/A:1010933404324. URL `https://doi.org/10.1023/A:1010933404324`.

[7] S. Bull and A. Doig. Properties of protein drug target classes. *PLoS ONE*, 10:1–44, 2015. doi: 10.1371/journal.pone.0117955.

[8] J. Das and H. Yu. High-quality protein interactomes and their applications in understanding human disease. *BMC Systems Biology*, 6(1):92, 2012. doi: 10.1186/1752-0509-6-92.

[9] M. Davis, M. Sehgal, and M. Ragan. Automatic, context-specific generation of gene ontology slims. *BMC Bioinformatics*, 11(498):1–13, 2010.

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

[11] Y. Fu, Y. Guo, Y. Wang, J. Luo, X. Pu, and M. Li. Exploring the relationship between hub proteins and drug targets based on GO and intrinsic disorder. *Computational Biology and Chemistry*, 66:41–48, 2015.

[12] S. Ghiassian, J. Menche, and A. Barabasi. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Computational Biology*, 11(4), 2015. doi: 10.1371/journal.pcbi.1004120.

[13] J. Heritage, S. McDonald, and K. McGarry. Integrating association rules mined from health-care data with ontological information for automated knowledge generation. In *The 17th UK Workshop on Computational Intelligence, UKCI-2017*, University of Cardiff, UK, 6th-8th September 2017.

[14] M. Hsing, K. Byler, and A. Cherkasov. The use of Gene Ontology terms for predicting highly-connected hub nodes in protein-protein interaction networks. *BMC Systems biology*, 2(80):1–14, 2008.

[15] E. Huttlin, L. Ting, and R. Bruckner. The BioPlex network: A systematic exploration of the human interactome. *Cell*, 162: 425–440, 2015.

[16] E. Kolaczyk and G. Csardi. *Statistical Analysis of Network Data with R*. Springer, 2014.

[17] D. Lee, J. Park, K. Kay, N. Christakis, Z. Oltvai, and A. Barabasi. The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences*, 105(29): 9880–9885, 2008. doi: 10.1073/pnas.0802208105. URL `http://www.pnas.org/content/105/29/9880.abstract`.

[18] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

[19] N. Luecken, M. Page, A. Crosby, S. Mason, and G. Reinert. CommWalker: correctly evaluating modules in molecular networks in light of annotation bias. *Bioinformatics*, 1-7, 2017. doi: 10.1093/bioinformatics/btx706.

[20] Huimin Luo, Jianxin Wang, Min Li, Junwei Luo, Xiaoqing Peng, Fang-Xiang Wu, and Yi Pan. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics*, 32(17):2664–2671, 2016. doi: DOI:https://doi.org/10.1093/bioinformatics/btw228.

[21] 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.

[22] K. McGarry and E. Assamoha. Data integration with self-organising neural network reveals chemical structure and therapeutic effects of drug atc codes. In *The 17th UK Workshop on Computational Intelligence, UKCI-2017*, University of Cardiff, UK, 6th-8th September 2017.

[23] J. Menche, A. Sharma, M. Kitsak, S. Ghiassian, M. Vidal, J. Loscalzo, and A. Barabasi. Uncovering disease-disease relationships through the incomplete human interactome. *Science*, 347(6224), 2015. doi: 10.1126/science.1257601.

[24] 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.

[25] J. Overington, B. AlLazikani, and A. Hopkins. How many drug targets are there? *Nature Reviews Drug Discovery*, 5:993–996, 2006. doi: doi:10.1038/nrd2199.

[26] A. Russ and S. Lampel. The druggable genome: an update. *Drug Discovery Today*, 10(23/24):1607–1610., 2005.

[27] O. Ursu, J. Holmes, J. Knockel, C. Bologna, J. Yang, S. Mathias, S. Nelson, and O. Tudor. Drugcentral: online drug compendium. *Nucleic Acids Res*, 45:D932D939, 2017. doi: 10.1093/nar/gkw993.

[28] M. Vidal, ME. Cusick, and A. Barabasi. Interactome networks and human disease. *Cell*, 144(6):986–998, 2011. doi: doi:10.1016/j.cell.2011.02.016.

[29] Y.C. Wang, S. Chen, N. Deng, and Y. Wang. Network predicting drug's anatomical therapeutic chemical code. *Bioinformatics*, 29(10):1317–1324, 2013.

[30] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, 2008. doi: 10.1093/bioinformatics/btn162.

[31] L. Yu, B. Wang, and L. Gao. The extraction of drug-disease correlations based on module-distance in incomplete human interactome. *BMC Systems biology*, 10(111), 2016. doi: 10.1186/s12918-016-0364.

[32] S. Zhang and Q. Tang. Protein-protein interaction inference based on semantic similarity of Gene Ontology terms. *Journal of Theoretical Biology*, 401:30–37, 2016.